Bayesian Gaussian Processes for Identifying the Deteriorating Patient

Glen Wright Colopy

St Cross College

University of Oxford

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfillment of the requirements for the degree of

Doctor of Philosophy

Hilary Term, 2018

Dedication

This thesis is dedicated to my father, Michael Wright Colopy.

For reading to me, making play-dough sculptures to specification, remaking the play-dough sculptures, cub scouts, statistics posters, boy scouts, checking my homework, setting an example of how a man should act, Philmont, technical writing advice, encouraging me to join the ASA, a weekend in Cologne, reading drafts conference papers, taking me to JSM, reading drafts of journal articles, and reading draft after draft of the chapters in this thesis.

I look forward to writing the next chapters together.

Acknowledgements

Monitoring a patient's health is the coolest topic. The ideas generated and sense of purpose derived from working in this field have brought me great joy. Many thank yous are due:

Prof. David Clifton and Prof. Steve Roberts for proposing a thesis topic that has given me so much fun, happiness, and sense of purpose. Thank you both for your foresight in selecting such a great topic and sending me off on the correct (read "Bayesian nonparametric") path to pursue it. Thank you as well for including me in your excellent research communities, and in particular, the Computation Health Informatics laboratory.

Steve Roberts and Gari Clifford for advocating my selection for the Clarendon Scholarship. The financial assistance of the Clarendon Scholarship is life-changing. (On this account, my family would like to thank you as well!) Thank you to all donors to the Clarendon Fund who make this award possible.

My examiners, Magnus Rattray (University of Manchester) and Maarten De Vos (University of Oxford). This Spring, in the face of deadlines, you demonstrated such speed and flexibility that I suspect you to be ninjas. During this time, I was in the process of moving back home and getting married. I truly appreciate your hard work on my behalf.

My many clinical and industrial collaborators, both within and beyond my DPhil thesis, from whose work I have benefited. These include the Diabetes Trial Unit (Oxford Centre for Diabetes, Endocrinology and Metabolism), the Oxford Centre for Statistics in Medicine, the Oxford Vaccine Group (Department of Paediatrics), Addenbrooke's Hospital (Cambridge University Hospitals), the Oxford Cancer and Haematology Centre, and Pfizer Inc.

My friends with the CDT: in particular Jo Armitage - thank you for your friendship and constant advocacy over 5 years.

The "cool kids on the block" when I first started in the CDT: Marco Pimentel (for early and on-point advice on Gaussian processes), Alistair Johnson

(my favorite machine learning heretic) and Mauro Santos. Mauro - I think you deserve extra-special mention due to the sheer magnitude of time you spent discussing the field with me...from machine learning to practical clinical considerations and the earliest drafts of my publications. You were the viva prep I never had.

Richard Franzese, Sean Ledger, and Oliver Langford: my long-term friends in Oxford. It's amazing how many life events we've now shared together.

Also, Ben. Yes, you. Let the eternal record of the Oxford archive of DPhil theses show that Glen Wright Colopy and Benjamin Villard were friends from the very first day of CDT induction and will continue to be until the very end.

My family, which is most important and without whom success and the rest hardly matter:

Klaus and Heike Weber, thank you for welcoming me into your family and your willingness to surrender your dinner table for the sake of my work.

Pippin, Wicket, Willow, and Zeus, your fuzzy heads were always in demand of a scratch behind the ear.

James Leifeste, who has always been the most devoted of family.

Glen Wright Sr., my grandfather: I have always tried to model your work ethic. Thank you for your many years of support of my education. Kathleen Venter for all of your encouragement in reaching my goals.

And finally my innermost circle: my parents Michael and Alisa Wright Colopy, my brother, Travis Wright Colopy, and my wife, Anika Weber. No words are sufficient to thank you for the past. I look forward to our future.

Abstract

Patients discharged from the ICU will commonly be placed in intermediary care, such as the step-down ward, where the nurse-to-patient ratio is reduced (compared to that of the ICU). Although most of these patients will continue to recover and stabilise, a significant portion will suffer cardiac arrest and/or other clinical emergencies, and readmission into intensive care. Upon readmission, the risk of mortality is significantly higher than that of the general ICU population. Evidence suggests that early detection of deterioration may prevent or alleviate the severity of clinical emergencies. Notable shortcomings of current practices are that they (i) involve manual calculation of risk scores, (ii) depend on heuristic decision criteria, (iii) ignore time-series dynamics of physiological measurements, and (iv) lack patient-specificity.

Gaussian process regression (GPR) models are proposed as a principled, probabilistic method to address the clinical need to continuously monitor patient vital-sign time-series with the flexibility to address the aforementioned weaknesses of current methods. The proposed GPR models focus on the robust forecasting of patient vital-sign time-series and early detection of patient deterioration.

The primary contributions of this thesis describe how:

1. Probabilistic models may be used to identify artefactual measurements from continuously-acquired vital-sign monitoring devices.

2. GP covariance functions may be constructed and regularised for robust modelling, suitable for both patient-cohorts and personalised care.

3. GPR-based methods may quantify erratic physiological time-series and provide useful advanced warning of deterioration events.

Each of the above contributions use the time-series correlation of vital-sign measurements for advantageous clinical inference.

List of Publications

The following lists all publications produced over the course of this DPhil. Numbering corresponds to bibliographical reference.

Thesis Publications

[1] G. W. Colopy, S. J. Roberts and D. A. Clifton, "Bayesian Optimisation of Personalised Models for Patient Vital-Sign Monitoring", IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 2, pp. 301-310, 2018. [**Invited Paper in Special Issue on Informatics for Personalized, Precision and Preventive Healthcare IEEE BHI 2017**]

[2] G. W. Colopy, T. Zhu, L. Clifton, S. Roberts and D. Clifton, "Likelihood-based artefact detection in continuously-acquired patient vital signs", 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017.

[3] G. W. Colopy, M. Pimentel, S. Roberts and D. Clifton, "Bayesian optimisation of Gaussian processes for identifying the deteriorating patient", IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2017.

[4] G. W. Colopy, M. Pimentel, S. Roberts and D. Clifton, "State-space approximations to Gaussian processes for patient vital-sign monitoring in computationally-constrained clinical environments", MEIbioeng, 2016.

[5] G. W. Colopy, M. Pimentel, S. Roberts and D. Clifton, "Bayesian Gaussian processes for identifying the deteriorating patient", 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016.

Thesis Patents

[6] D. A. Clifton, M. A. F. Pimentel, and G. W. Colopy, "System Monitor and Method of System Monitoring", GB 1613318.3, 2016

Related Non-Thesis Publications

[7] G. W. Colopy, J. Jorge, M. Theodorakis, R. Holman, L. Tarassenko, D. A. Clifton, “Vital-Sign Fusion for Novelty-based Detection of Imminent Hypoglycaemia” [**In review**]

[8] G. W. Colopy, D. A. Clifton, “Personalised Volatility Metrics for Temperature Time-Series of Post-Vaccination Infants ” [**In preparation**]

[9] T. Zhu, G. W. Colopy, Y. Yang, C. W. Pugh, and D. A. Clifton, “Modelling Patient-Specific Trajectory Using Hierarchical Bayesian Gaussian Processes” [**Invited Paper in Special Issue on Informatics for AI-Enabled Connected Health Informatics IEEE BHI 2018**] [**In review**]

[10] T. Zhu, G. W. Colopy, C. W. Pugh, and D. A. Clifton, "Identifying Patient-Specific Trajectories in Haemodialysis Using Bayesian Hierarchical Gaussian Processes", IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2018 [**Winner of Conference Paper Competition: 2nd Place.**]

[11] D. Clifton, K. Niehaus, P. Charlton and G. W. Colopy, "Health Informatics via Machine Learning for the Clinical Management of Patients", IMIA Yearbook, vol. 10, no. 1, pp. 38-43, 2015.

Non-Related Publications

[12] T. Rieger, et al, "Improving the Generation and Selection of Virtual Populations in Quantitative Systems Pharmacology Models", Progress in Biophysics and Molecular Biology, 2018. [**In press**]

[13] R. Clifford, et al, "SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage", Blood, vol. 123, no. 7, pp. 1021-1031, 2013.

Contents

List of Figures

List of Tables

List of Abbreviations

1-D, 2-D, etc.	One-dimensional, two-dimensional, etc.
APACHE	Acute Physiology And Chronic Health Evaluation
bpm	Beats per minute or breaths per minute
cdf, CDF	Cumulative density function
CEWS	Centile-based early warning score
CI	Confident interval
DBP	Diastolic blood pressure
EI	Expected improvement
EM	Expectation–maximisation
EVT	Extreme value theorem (also known as the the Fisher–Tippett–Gnedenko or Fisher–Tippett theorem)
EWS	Early warning score
FiO_2	Fraction of inspired oxygen
FPR	False positive rate
GCS	Glasgow Coma Score
GP	Gaussian process
GPR	Gaussian process regression
HR	Heart rate
ICU	Intensive care unit
IID	Independent and identically distributed
KDE	Kernel density estimate
LML	Log marginal likelihood
LOM	Length of measurement
LOS	Length of stay
MAP	Maximum a posteriori

MCMC . . . . Markov chain Monte Carlo

MET . . . . . Medical emergency team

MEWS . . . . Modified early warning score

MLE . . . . . Maximum likelihood estimate

mmHg . . . . . Millimetres of mercury

MODS . . . . . Multiple Organ Dysfunction Score

MPM . . . . . Mortality Prediction Model

MVN . . . . . Multivariate Normal or Multivariate Gaussian

NEWS . . . . National Early Warning Score

NLML . . . . . Negative log marginal likelihood

NZEWS . . . . Wellington Score

pdf, PDF . . . Probability density function

PEWS . . . . . Paediatric Early Warning Score or Paediatric Early Warning Signs

PI . . . . . Probability of improvement

RBF . . . . . Radial basis function

RR . . . . . Respiratory rate

RS . . . . . Random search

SAPS . . . . . Simplified Acute Physiology Score

SBP . . . . . Systolic blood pressure

SDU . . . . . Step-down unit (Step-down ward)

SE . . . . . Squared exponential

SOFA . . . . . Sepsis-related Organ Failure Assessment/ Sequential Organ Failure Assessment

SpO_2 . . . . . Blood oxygen saturation

SVM . . . . . Support vector machine

TEW . . . . . Time of early warning

UPMC . . . . University of Pittsburgh Medical Center

1 Clinical Need

This chapter begins by describing deterioration detection within the setting of preventative care. Two definitions of deterioration detection are described with a brief example. The role of the step-down unit is then described, followed by a documentation of the risk associated with emergency readmission to intensive care unit from the step-down unit. These risks motivate the use of deterioration detection in the step-down unit.

Finally, an outline of the technical requirements of engineering approaches to address the clinical need are described, along with a summary of how this thesis' contributions meet those requirements.

1.1 Overview of Deterioration Detection

Patient deterioration detection is a broad group of medical methods tasked to identify patients who are currently experiencing, or soon to experience, an adverse clinical event. Deterioration detection may be considered a tool in preventative care, since the time-period of early warning presents an opportunity to mitigate or avoid the effects of the adverse event.

The term "deterioration detection" may be apt to conflate two distinct clinical objectives:

1. the detection of a patient who is currently deteriorating (towards an adverse event), and

2. the detection of a patient who has deteriorated (i.e., has experienced the adverse event, and is now suffering the effects, and possibly deteriorating further, towards secondary adverse outcomes.

The distinction between these two forms of deterioration detection is illustrated in the following extended example.

Detection of a patient who is deteriorating It is common practice to monitor older patient cohorts for key risk factors of stroke, such as high blood pressure, smoking, and family history. The detection of deteriorating health (which may, ultimately, lead to stroke) may result in a treatment protocol aimed to prevent the stroke from occurring. This preventative care would benefit the patient by preventing the adverse health outcome, as well as any secondary health outcomes that may follow.

Detection of a patient who has deteriorated We may also consider the goal to identify quickly any patient who is currently suffering a stroke. Since the primary adverse event (stroke) can no longer be prevented, the goal of early detection is how to mitigate the risk or severity of further adverse outcomes (e.g., death or permanent disability).

It may be helpful then to consider both (i) early detection in advance of an adverse outcome, and (ii) early detection upon the onset of an adverse outcome, as two highly-related yet distinct facets of deterioration detection. Since these two facets overlap in both goals and methods, the term "deterioration detection" will be used to refer to either of these clinical goals (with additional clarifications only given where necessary).

Deterioration detection is applied across a wide range of clinical settings including critical care, sepsis detection, and the management of chronic illness such as hypoglycaemia, irritable bowel syndrome [14], mental health, and chronic obstructive

pulmonary disorder. Accordingly, the adverse events of interest may differ between clinical settings, as may the specific definition of the same adverse event. Typically, deterioration detection involves inference on pertinent physiological signals as they they relate to the clinical outcome (or adverse event) of interest.

This thesis addresses the need to detect deteriorating patients in hospital step-down wards.

1.2 Overview of Step-down Units

A step-down unit[1], or SDU, is "a hospital nursing unit providing care intermediate between that of an intensive care unit and a normally-staffed in-patient division" [15].

The SDU manages the recovery of stabilised acutely-ill patients after discharge from the intensive care unit (ICU) but with less staff-intensive monitoring, in accordance with patient condition. The SDU may also receive patients from the general ward who require an escalation in care [16].

In essence, the goal of an SDU is to manage the recovery of patients after discharge from the Intensive Care Unit (ICU) while reducing the staff-intensive burden of monitoring patients who are more stable than acutely-ill ICU entrants.

Although exact definitions of an SDU and ICU vary between countries [17] [18], the ICU accounts for 1.2% [17] [19] of hospital beds in the UK, and 9% - 20% [17] [20] of hospital beds in the US. SDU patients are of a more stable condition than those of 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) [21]. It is common for SDUs to be staffed by nurses trained in critical care, just as in the ICU [22].

Critical-care wards in the US have been estimated to account for about 13.4% [23] to 17.4%-39% [24] of total hospital care expenditures, and therefore it is important to optimise the management of patients in such settings.

[1] Alternatively referred to as a step-down ward or high-dependency unit.

1.3 Clinical Deterioration and Readmission to ICUs

1.3.1 ICU Readmission Rates

It is common for patients who have been discharged from the ICU to require readmission to the ICU. Readmission usually implies that the clinical episode, which caused the initial admission to ICU, may have precipitated the readmission as well. Readmissions within the same hospital stay are more readily documented than readmissions that are separated by time and/or hospital.

Although SDU patients are physiologically stable in general, a significant portion of SDU patients experience a clinical emergency event, or require emergency re-admission to the ICU. Various studies across different hospitals have estimated ICU readmission rates (within the same hospital stay) to be 3.9%-9% (severity-adjusted 4.2% - 7.6%) [25], 8.8% [26], and 0%-18.3% [27].

Due to the varying definitions of an SDU, not all patients in studies [25], [26], and [27] were transferred into an SDU. However, these patients were admitted to other hospital wards intended for escalated care, and are suggestive of the severity of illness of patients received by the SDU.

ICU readmission rates were 5.8% for postoperative (i.e., post-surgical) patients and 6.4% for non-postoperative patients [25]. In light of a large number of acknowledged, but controlled confounding factors, this left the postoperative-status of a patient to be a positive, but somewhat ambiguous, risk factor.

Using similar criteria on the same SDU, [28] and [21] determined that 31% and 34% of SDU patients experienced cardiorespiratory instability during their stay on ward. Mortality rates were 2% and 3% respectively. Neither study described the relation between these outcomes and ICU readmission.

1.3.2 Mortality upon ICU Readmission

Readmission to the ICU has significant implications for patient outcomes: mortality rates for ICU patients readmitted within the same hospital stay have been estimated at 40.2% [26]. Another study estimated readmission mortality to be 24.7% [25]

(in contrast to 4.0% mortality of patients who were not readmitted) and went on to note, "Several studies conducted in single hospitals during the 1980s and early 1990s indicate that roughly 5% to 20% of ICU admissions represent readmission during the same hospitalisation, and that patients readmitted to the ICU have disproportionately high in-hospital mortality rates and lengths of stay." These high levels of mortality motivate the use of principled methods to identify and, ideally, forewarn physiological deterioration.

1.3.3 Causes for ICU Readmission

There is significant heterogeneity within patients who are readmitted to the ICU after discharge. A commonly-cited study [25] found that congestive heart failure (7.6%) and sepsis (4.5%) were the most common causes of readmission. Over 20% of the total number of readmissions were from cardiovascular-related causes. Fewer than 20% of patients had identical reasons for initial admission and readmission to the ICU. In contrast, another study calculated that 49% of ICU patients were readmitted "for the same or related diagnosis" as the diagnosis of the original ICU admission [26]. Cardiovascular-related causes were consistently mentioned as risks for both admission and re-admission to the ICU. It is not surprising that patients who were admitted to the hospital on the account of a (typically chronic) cardiovascular condition would (i) exhibit subsequent cardiovascular events in hospital, and (ii) require readmission due to a cardiovascular condition on a subsequent hospital admission.

A confounding aspect in these studies is that a single "primary" cause for readmission was required, which ignores the correlation across different clinical risks. Despite the absence of information about the correlation between risks, the heterogeneity of causes is clear evidence that it is advantageous to incorporate a variety of factors when deciding whether a patient requires urgent care or readmission. Additionally, it is possible that cardiovascular signals are a bellwether for physiological abnormalities that are otherwise difficult to quantify. When combined with the relative ease of acquiring cardiovascular measurements, it is

unsurprising that cardiovascular signals are a key focus when assessing a patient's risk of readmission.

1.3.4 Time-Frame of Deterioration and ICU Readmission

Several studies suggest that there is sufficient time and precursor-physiology to warrant deterioration detection for many SDU patients prior to ICU readmission.

In [25] (cited previously for examining the causes of ICU admission and readmission), ICU readmission was most common 24-48 hours after ICU discharge (even more common than readmission within 24 hours of discharge). Over half of readmitted patients were readmitted within 72 hours and 22% were readmitted after a week or longer. From this, we way conclude that the time since the initial ICU discharge (i.e., the time-from-admission into SDU) is a consideration when assessing a patient's risk of a clinical event or readmission.

SDU patients exhibit symptoms of deterioration well in advance of the response of medical emergency teams (METs) where such teams of additional bedside support are available (this does not include the UK). One study reported "a mean of 6.3 hours elapsed between the onset of a clinically apparent cardiorespiratory instability and the activation of our rapid response system" [21]. This statistic advocates the presence of an observable period of deterioration, prior to emergency readmission.

A confounding element to the presence of retrospectively-identified periods of abnormal physiology is the presence of physiology that induces false alarms and, subsequently, alarm fatigue among nursing staff. These false alarms may be the result of, for example, (i) poorly-set alarm parameters (e.g., an excessively-high lower threshold on HR for a patient with athletic bradycardia), or (ii) technical failure (e.g., partially-detached probes, or algorithmic failure of heartbeat detectors). It is also plausible that many alarms that detect true physiological abnormality are dismissed on the account of (i) being uninterpretable to clinical staff, or (ii) perennially ill patients providing a constant stream of warranted alarms.

These studies suggests that there is a time-window in advance of a verified clinical event in which a clinical alarm would be justified as a "true positive" alarm.

Any alarms outside such a window, but still preceding a clinical event, would be considered more tenuous, as would alarms that did not precede a confirmed clinical event. This has important implications to the design of several experiments within this thesis.

1.4 Value of Early Deterioration Detection in Critical Care

Studies have also demonstrated that early warning (via deterioration detection) reduces the risk and/or severity of adverse clinical events.

Not all deaths after release from the ICU are preventable [18, 26, 27] but there is evidence that earlier response can reduce mortality rates [26, 29]. Further examples include a study in which 5.2% (52 of 1000) of acute hospital deaths were determined to be preventable [30]. Another study identified over 2000 preventable patient deaths over 29 months [31]. Both [30] and [31] identified the primary cause of most of the preventable deaths to be missed early warning signs and failure to act on evidence of deterioration.

A recent retrospective study [20] by authors of the APACHE system [32] (a multivariable threshold-based risk scoring system, originally proposed in 1981, updated to APACHE II in 1985, APACHE III in 1991, and APACHE IV in 2006) found that from 1988-2012 United States ICUs using the APACHE system saw a drop in mortality rates from 17.3% to 12.4%, adjusting for severity. Notably, "[m]ost of these dramatic relative decreases in hospital mortality rate occurred between 1988 to 1989 and 1993 to 1996." The study did not relate the trend in mortality rate to any confounding factors, such as new technologies or therapies, that would also improve clinical outcomes. Furthermore, mortality improvements varied greatly by diagnosis. These trends were mirrored by a 23% reduction in ICU length-of-stay (LOS) and a 38% reduction in hospital LOS. Patients are increasingly placed into other wards (including SDUs), which likely contributes to these trends. A reduction in overall ICU mortality has been documented in the UK as well [33].

Using the Visensia system (a multi-vital-sign kernel density estimate-based risk score system), a single-site prospective study [34] found that the time of using the early warning system resulted in statistically-significant decreases in the number and duration of cardiorespiratory instabilities per admission, compared to the time period prior to using the system. Mortality decreased from 2% (before Visensia) to 1% (after using Visensia), but a statistical comparison was not made.

There is a notable interplay between earlier ICU readmission and improved survival upon readmission. Patients who require readmission to ICU from other wards are more likely to survive their hospital stay if readmitted earlier [26].

For particular clinical applications, the extent of early warning per se may not matter. For example, a body of research has been dedicated to an early detection of sepsis in critical care [35–37]. The work demonstrates a time period of abnormal vital-signs may be identified prior to severe sepsis. Since the clinically accepted physiological markers of sepsis occur in up to 90% of patients [37], it is ambiguous whether these abnormal physiologies are particular to sepsis, or whether sepsis is simply present in a large portion of deteriorating patients. More importantly, the time scale of early warning in sepsis is of debatable value with opposing views arguing that outcomes are best when (i) sepsis is detected as early as possible, or (ii) that detection need only occur in advance of acute sepsis, which results in imminent organ failure. The changing definition of sepsis (revised as recently as 2016 [38]) over the last several decades may further complicate inference on the value of early detection.

It may be asked, "Why, if advanced warnings of deterioration are well-described in retrospective studies, are these indicators not codified in new monitoring practice?" While there are many reasons for slow change in clinical practice, there are several clear reasons why the medical decision criteria of current practice is only adequate to identify late-stage deterioration. These aspects of current clinical practice will be discussed in the literature review chapter, which catalogs common shortcomings of different monitoring techniques. The operational constraints will be discussed in the next section, as an aspect of clinical need.

1.5 Addressing Clinical Need via Engineering

The remainder of this chapter will describe how the research contributions of this thesis are amenable to the organisational and computational constraints in the critical care setting. These considerations are essential to ensure that clinical adoption of the engineering solutions developed in this thesis are possible. Several essential requirements are described, followed by details on how the proposed engineering approach meets these requirements and works within the confines of current clinical practice.

Pantelopoulos [39] lists 16 features on which the feasibility of a sensor-based health monitoring system should be assessed for clinical feasibility. Aside from the economic and ergonomic features that were listed, many of the features can be evaluated within the process of algorithm development. Such features include computation and storage requirements, ease of use and interface, reliability and robustness to faults, and interpretable context for decision support. Many alternative engineering and non-engineering solutions, in fact, meet these operational requirements as well, on some level. Compared to these alternatives, the advantages of the proposed GP-based methods include improvements in reliability of forecasts and deterioration detection, capacity for patient personalisation, and ability to detect artefactual data. These aspects, which motivate clinical adoption, will not be covered in this section but in Chapters 5, 6, 7, and 8, where adequate detail can be given.

1.5.1 Clinical Requirements of an Engineering Solution

Staff and operational constraints prevent many patient monitoring solutions from being used in practice, regardless of the clear and widely-accepted clinical need described earlier. A non-trivial aspect of the clinical need is for a system that complies with current operational practice on critical care wards.

The clinical need and state of current clinical practice has strong implications of the types of solutions that are feasible to implement into new practice. Engineering solutions must be built to address the needs and specifications of the clinical staff who will use those solutions in critical care.

An engineering solution to deterioration detection in critical care should have, at least, the following attributes:

1. The system must be simple to learn (with appropriate training of staff) and easy to integrate into current clinical practice.

2. The system must not increase the work burden of staff. Ideally, the system should reduce staff burden (e.g., via automation of any time-intensive tasks) where feasible.

3. The system must be transparent in its alarm criteria and, ideally, in the cause or information that precipitated any particular alarm.

4. The system should be robust to worst case performance (e.g., with respect to detecting the adverse event of interest, or minimising false-alarm rates).

5. The system's computational requirements should not exceed the computational resources of the ward.

6. The system should be responsive/adaptable to clinician input.

A more thorough coverage as to how machine learning technology in particular may be applied can be found in [40]. Further detail on these points may be found below.

1.5.2 Compatibility of Research Contribution to Clinical Requirements

This section describes how the proposed GP-based methods address the need to work within the confines of current clinical practice. Each item will be discussed in the order that it was listed above.

The advantages of using GPR-based monitoring systems (over current patient-monitoring systems) will be discussed throughout the thesis, particularly in Chapters 2, 6, and 8.

The system must be simple to learn.

Staff training is required at the introduction of any new technique and equipment in critical care. Continuing eduction is accepted as a common and frequent aspect of critical care work. The GP-based methods produce clinical inference in a manner that can be easily translated into familiar terms, such as a risk score. In essence, the clinical staff may be presented with familiar information, it is merely the background calculation of that information that has changed.

The system must not increase the work burden of staff.

The required data collection will impose no additional burden to the clinical staff. It is already common practice for hospitals to electronically monitor each patient's vital-signs. An electronic system to process, analyse, and forecast vital-signs would be a natural extension of this data acquisition without placing a further burden on the patient and/or nursing staff.

Furthermore, the automation of many of these tasks would reduce the staff-intensive burden, thereby freeing clinical staff to address competing priorities. Computational methods may also accomplish tasks that would be impossible at any level of staffing, for example: computer-based methods can make use of the large amount of continuous data being collected at bedside. Volume of such data would be impossible for an intermittent human observer to synthesise. The exact vital-sign time-series of thousands of previous patients (and their associated clinical outcomes) can also be stored in a computer's memory, giving a very unique approach to "clinical experience". Such methods would also be free from human transcription error and human calculation error. In the absence of computational methods, the staff required to record, error-check and analyse the available data would be prohibitive.

A computer-based continuous monitoring system will alleviate some of the risks associated with the reduced nurse-to-patient ratio. Continuous electronic monitoring will provide attention to all patients between the intermittent observations of the nursing staff and obviate the resultant estimates of risk that only incorporate single points in time. Continuous observation, which allows for clinical warnings in the

absence of clinical staff, will decrease the lag between a patient first exhibiting abnormal physiology and the action undertaken by medical staff. Earlier action should, in turn, be associated with decreased mortality, and an improvement in other clinical outcomes. Furthermore, the focus of the nursing staff can be directed towards those patients demonstrating the greatest risk of an adverse clinical event (for example, those patients with vital-signs that have been forecasted to be in a dangerous range). This would further optimise the use of a hospital's personnel and financial resources.

The system must be transparent in its alarm criteria.

A key component of alarm fatigue is that many devices in critical care (in the eyes of the clinician) seem to generate alarms at random, or for no particular reason. In part, this effect may be aided by (i) a poor understanding of preset factory settings of many devices' alarm parameters or (ii) an underestimation of the prevalence of artefactual data acquired by a device recording critically ill physiology (many devices are not validated on a wide range of critically ill patients, but instead on relatively small set of healthy test subjects). Further confusion may arise when the device is unclear as to the reason of the alarm. In contrast, the simplistic hand-calculated EWS tables (such as the National Early Warning Score [41, 42]) are clear in the cause of alarm, by displaying the thresholds at which a risk score is escalated.

Despite their further technological complexity, GP-based methods can clearly elucidate their cause of alarm. As shown in Figure 1.1, an alarm from 1.1(a) GP-based forecasting, 1.1(b) step-change detection, and 1.1(c) time-series matching could each be accompanied with an intuitive display, to explain the cause of the alarm. This would provide context by which clinical staff may evaluate the merit of the alarm. This transparency would reduce alarm fatigue by allowing both true and false alarms to be better understood. Such transparency is also required by regulatory bodies, when such alarms would influence clinical decisions.

In 1.1(a), this display is achieved by illustrating that the patient's heart rate drop from a "dangerous" 120 bpm to a more healthy 70 bpm likely portends heart

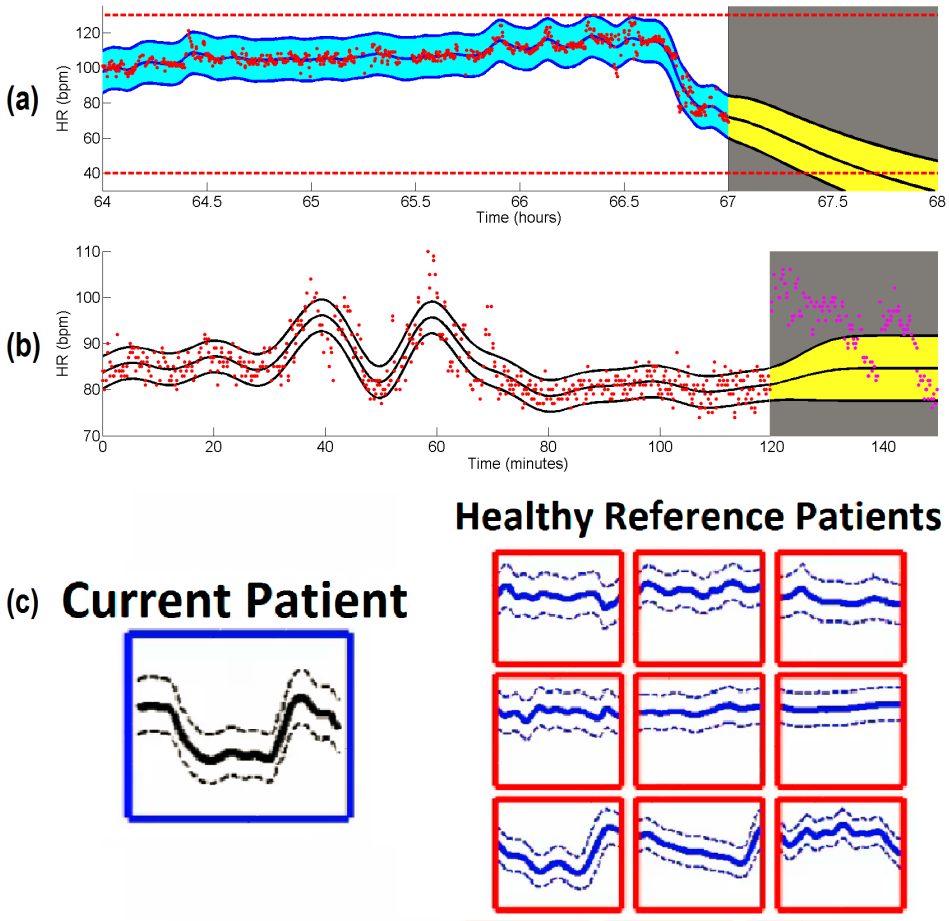


Figure 1.1: Displays of GP-based clinical inference via (a) vital-sign forecasting, (b) step-change detection, and (c) time-series matching. The GP provides three different forms of inference in the presence of unusually large change in the vital-sign time-series. The onus is on the GPR method to demonstrate the validity of its alarm, since the patient's current vital-sign measurement does not suggest deterioration. In (a), the patient's heart rate measurements have recently decreased from 120 bpm to 70 bpm. Since 70 bpm is closer to typical values of healthy patients, this may be mistaken as a sign of improvement, however the GP illustrates that the decrease is rapid and the patient is at risk of surpassing the 40 bpm emergency threshold. (The prior mean of this GP forecast was set to the last observed measurement in the training window.) In (b) the patient's data from minutes 120 to 140 show a decrease from about 110 bpm to 80 bpm, however the GP-based step-change detector shows that these values are highly unusual with respect to the forecasted values at 120 minutes. In (c) a relative bradycardia event has occurred current patient's time-series. The unusual time-series segment is identified by its dissimilarity to a reference set of healthy patient time-series segments.

rate falling below 40 bpm. By showing how the patient's earlier HR measurements inform our current expectations (displayed by the confidence intervals) the GP makes the case that an adjustment of alarm bounds (tailored to the current patient at the current time) is more appropriate than a generic population-based threshold (which would not alarm at a heart rate of 70 bpm).

In 1.1(b), a similar scenario has occurred, in which heart rate is shown to be decreasing towards a more healthy range, from minutes 120 to 140. However, the GPR can illustrate that (i) these dynamics were preceded by a drastic heart rate jump at minute 120, and (ii) the current dynamics from minutes 120 to 140 are highly unusual with respect to the forecasted values at 120 minutes.

In 1.1(c), the current patient's time-series exhibits a period of relative bradycardia, which may be identified by a comparison to a dictionary of healthy patients' time-series. From this, the doctor may see, from previous examples, that while healthy patients may have values similar in magnitude, they rarely have such radical volatility, and instead tend to exhibit trends with relatively constant measurements over short time periods.

Each display could be generated automatically so that a clinician (or regulator) could intuit the added clinical value without a strong technical understanding of the probabilistic GP model. We note that in (a,b,c) an alert would be unlikely under current EWS-based protocols, since the vital-sign has not yet exceeded an extreme threshold in any of the examples.

The system should be robust to worst case performance.

A typical misgiving of new technological solutions is the tendency to aim to improve average or best-case performance. While this may prove useful in many circumstances, in many clinical applications, prevention of worst-case performance is more important. This is because worst-case performance is typically associated with the most severe adverse clinical outcomes. Examples of worst-case performance include (i) late or missed warning on a deteriorating patient, or (ii) vital-sign measurements that fall outside of pre-specified confident bounds on a vital-sign forecast.

The GP-based methods described in this thesis, not only improve performance for the average patients, but also generate significant improvements for patients being completely missed by current methods. As will be described in Chapters 6, 7, and 8, these patients include (i) patients whose vital-signs are most difficult to forecast and (ii) patients with the least early warning in advance of deterioration events. As described in the respective chapters, many of the proposed methods can be tuned or regularised to avoid worst-case performance.

A concrete example of improving worst-case performance is provided in Figure 1.2. GP-based step-change detection (as illustrated in Figure 1.1(b) is compared to a kernel density estimation-based method [43, 44] that was developed in another doctoral thesis [45]. Each monitoring system is compared according to its trade-off between time of early warning (TEW) in advance of deterioration and the false positive alarm rate (FPR).[2] The three performance lines for each method show the 33rd, 50th, and 67th percentile of TEW across a group of patients who deteriorated. The GP-based method shows superior median performance (middle line) compared to the KDE-based method. However, it is more important that the 33rd percentile (lower line) has better performance, since this relates to those patients who would have the least amount of early warning in advance of deterioration. It is more important that all patients have some early warning than it is to continue to improve performance on patients who already had many hours of early warning.

The system's computational requirements should not exceed the computational resources of the ward.

A computational solution is unacceptable in settings that lack the computational resources to implement such methods. In critical care, this may include the need to make computations in the time-sensitive fashion required for clinical intervention. While advances in high-performance and parallel computing have resolved such roadblocks in many applications, the requirements of security, privacy, ownership, and location mean that many wards still prefer, or are required by law, to use local or in-house computing.

[2] These metrics will be discussed in greater detail in Chapters 3, 7, and 8.

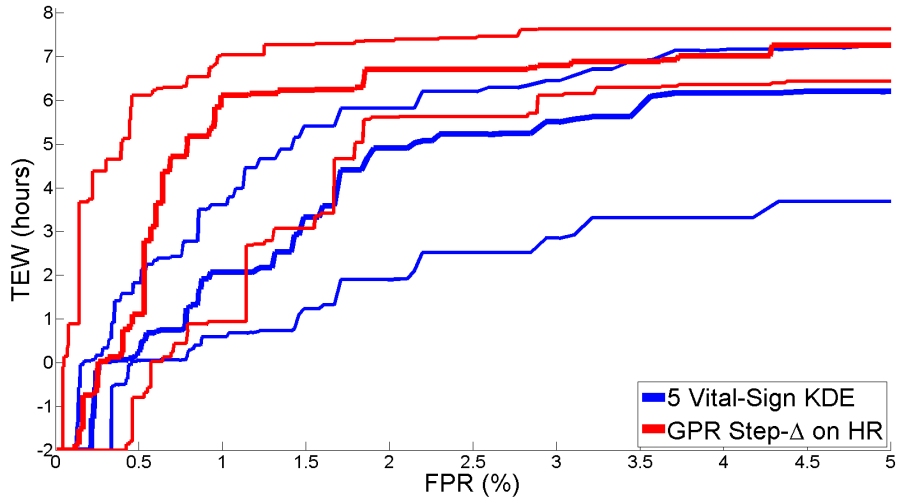


Figure 1.2: Early deterioration detection via GP-based step-change detection compared to a KDE-based alternative. Lines represent the 33rd, 50th, and 67th percentile of TEW (across a group of patients who deteriorated) at the given FPR. The increased TEW creates greater opportunity for clinical intervention, but comes at the cost of higher FPR, which induces alarm fatigue in clinical staff. It is arguably more important for a monitoring system to improve the 33rd percentile line than median performance, since this would improve the worst-case performance.

All methods described in this thesis can run faster than real-time on multiple patients from a single PC.

The system should be responsive/adaptable to clinician input.

The factory settings of many patient-monitoring devices are typically insufficient for the diverse needs of different patients or patient cohorts. Accordingly, adjustments to the device settings are frequently required to accommodate clinical need. The most typical examples include switching off alarm mechanisms or allowing a clinician to adjust the alarm thresholds.

The proposed GP-based methods circumvent the need of many tailored thresholds, for example, by generating alarms already personalised to the patient's physiology (e.g., step-change detection in Chapter 8). Furthermore, GP-based methods can be tuned in direct reference to the number of anticipated alarms, as shown in Chapter 6 and 8.

1.6 Conclusion

This chapter described how advanced warning of deterioration is both necessary and possible for patients in the SDU. Several practical considerations are described of how engineering methods may contribute to addressing this clinical challenge. Finally, a description is given of how the GP-based methods proposed in this thesis can meet those requirements while addressing the clinical need for improved deterioration detection.

2
Literature Review

As described briefly in Chapter 1, the methods, frequency, and physiology of interest in patient deterioration detection varies according to clinical context and the clinical outcome of interest. This chapter will first describe the current literature in critical care monitoring. We propose several useful distinctions between patient monitoring systems based on whether they rely on (i) heuristic decision criteria, as opposed to (ii) empirical decision criteria. The latter includes the technical state-of-the-art. Within the empirical patient monitoring systems, we further distinguish between (i) probabilistic and (ii) non-probabilistic methods. The penultimate section describes current applications of GPR (which is a probabilistic empirical method) to patient monitoring. The final section of the literature review gives a brief overview of the current state-of-the-art in GP-based modelling, in particular, those advancements that are most promising (although not necessarily used) in patient monitoring applications.

2.1 Heuristic vs Empirical Approaches to Patient Monitoring

A multitude of patient monitoring systems are used in ICU, SDU, and other hospital wards. A useful distinction can be drawn between (i) heuristic methods

and (ii) empirical methods which, broadly, describe the means by which clinical decision criteria are derived.

Heuristic methods have been developed since the beginning of specialisation in intensive care, but were not widely comparable between hospitals until the 1980's [46]. Heuristic methods tend to stratify patients according to clinical risk without recourse to explicit representation of the relation between risk factors and the outcome of interest. Heuristic methods are typically concerned with the development of decision rules (typically drawn a consensus of experts) by which clinicians can score, rank, or stratify the current risk-status of a patient (i.e., the current risk of a patient deteriorating or experiencing an adverse clinical event). A useful illustration in medical literature is that it is common to examine (i) how heuristic methods stratify patients according to clinical outcome, but it is rare to see (ii) how the rules of heuristic methods are developed with explicit reference to those same clinical outcomes.

Empirical methods, on the other hand, generally quantify the relationship between a set of patient characteristics and various clinical outcomes. Typical examples are prognostic/diagnostic models, which attempt to regress/classify outcome from a set of risk factors. Alternative empirical methods include novelty detection (1-class classification) [47]. If the relation between the patient characteristics and outcome is described in a probabilistic framework then the risk of a clinical event can then be examined within the context of the probability distribution that the model describes. Not all methods that describe the clinical outcome as a function of predictive variables have an exclusively probabilistic interpretation: support vectors machines, neural networks, and random forests have all been used in the critical care setting to forecast clinical outcomes such as length-of-stay and in-hospital mortality. Probabilistic and non-probabilistic approaches will be described in separate sections.

Empirical methods in literature have typically explicitly formalised several of the following:

1. how data were used to derive the decision rule,

2. inclusion/exclusion criteria of data,

3. experimental design,

4. performance metrics of interest, and

5. validation of model performance.

This review will focus on applications of heuristic and empirical methods in critical care for the purpose of monitoring patients for clinical deterioration.

Distinct from the method of monitoring is the implementation of that method into clinical practice. In general, heuristic methods tend to be implemented via non-technical means (e.g., heuristic EWS are typically hand-calculated as part of routine monitoring). Advantages to such methods include compatibility with routine clinical practice and the direct involvement of clinical staff's discretion when deciding to raise an alarm. Empirical methods, due to computational burdens tend to be implemented via technical means. Advances in m-health and related monitoring technology has helped bridge this gap in technical implementation, particularly in tasks that are easy to achieve such as the automated recording of medical information and electronic calculation of EWSs.

2.2 Heuristic Approaches to Patient Monitoring

2.2.1 Overview of Heuristic Patient Monitoring

The current practice of patient vital-sign monitoring is to observe a patient's current measurements and decide whether the patient's health state is currently at-risk of deterioration. Simplistic methods include rule-based thresholds, for example issuing a warning if a vital-sign exceeds a pre-specified threshold. These thresholds are usually set according to expert clinical experience about the global population of stable patients. Univariate alarm thresholds may be factory-programmed into the device generating the vital-sign measurements [48, 49]. Some literature denotes these systems as "triggering systems" [50], in contrast to "scoring systems" (which are multivariate and graduated, instead of univariate and binary). A review of these

National Early Warning Score Table

Parameter	3	2	1	0	1	2	3
Respiratory Rate (bpm)	≤ 8		9 – 11	12 – 20		21 – 24	≥ 25
SpO_2 (%)	≤ 91	92 – 93	94 -95	≥ 96			
Supplemental Oxygen		Yes		No			
Temperature (C)	≤ 35		35.1 – 36.0	36.1 – 38.0	38.1 – 39.0	≥ 39.1	
Systolic BP (bpm)	≤ 90	91 – 100	101 – 110	111 – 219			≥ 220
Heart Rate (bpm)	≤ 40		41 – 50	51 – 90	91 – 110	111 - 130	≥ 131
Level of Consciousness				Alert			V, P, or U

NEWS Score	Frequency of Monitoring
0	Minimum of every 12 hours
1 – 4	Minimum of every 4 - 6 hours
5 – 6 3 in one vital	Minimum of every hour
≥ 7	Continuous monitoring

Figure 2.1: Scoring table and escalation table of the Nation Early Warning Score (NEWS). NEWS assigns a warning score from 0-3 to each of 7 clinical parameters. A higher aggregate warning score across all parameters requires an increased frequency of patient monitoring. A patient's level of consciousness is classified as alert, reacts to voice (V), reacts to pain (P), and unresponsive (U).

single-parameter systems can be found in [51], and for multi-parameter systems in its companion article [52].

Improvements have been made by considering the simultaneous abnormality of multiple vital-signs *in addition* to abnormality of a single vital-sign [50, 53]. As shown in each of the EWS systems in Figures 2.1, 2.2, and 2.3 such methods typically include (i) a scoring table, which assigns a warning score, (ii) a composite early warning score, which is typically the sum the score assigned to each vital-sign, and (iii) an action or escalation table, assigning an appropriate clinical response to each possible EWS. An example of such a system is the manually-calculated National Early Warning Score (NEWS) [41] in Figure 2.1, which is described along with similar EWS in the next section.

Modified Early Warning Score Table

Parameter	3	2	1	0	1	2	3
Respiratory Rate (bpm)		< 9		9 – 14	15 – 20	21 – 29	≥ 30
SpO_2 (%)							
Supplemental Oxygen							
Temperature (C)		< 35		35 – 38.4		≥ 38.5	
Systolic BP (bpm)	< 70	71 – 80	81 – 100	101 – 110		≥ 220	
Heart Rate (bpm)		< 40	41 – 50	51 – 100	101 – 110	111 – 129	≥ 130
Level of Consciousness				Alert	Voice	Pain	U
Glascow Coma Scale	3	4 – 9	10 – 13	14 – 15			

MEWS Score	Clinical Implication
0	Continue observation of 1 time per day.
1	Observe at minimum 3 times per day.
2	Observe at minimum 3 times per day. Staff informs ward doctor.
3 – 4	Acute supervision by attending doctor. Next observation, within 1 hour.
≥ 5	Acute supervision by attending doctor. Next observation, within ½ hour.
Change ≥ 3	Call MET

Figure 2.2: Scoring table and escalation table of a Modified Early Warning Score (MEWS). MEWS assigns a warning score from 0-3 to each of 6 clinical parameters. A higher aggregate warning score across all parameters requires an increased frequency of patient monitoring and clinical supervision, culmination in an MET call. A patient's level of consciousness is classified as alert, reacts to voice, reacts to pain, and unresponsive (U). Notably, this MEWS incorporates neither SpO_2 nor supplemental oxygen into the decision criterion.

2.2.2 Example Heuristic Methods in Deterioration Detection

Common threshold-based early warning scores include NEWS [41, 42] in Figure 2.1, the Modified Early Warning Score (MEWS) [54] in Figure 2.2, and the Wellington Score (NZEWS) [55, 56] in Figure 2.3. These scores were designed for a generic population, however further modifications have been made to accommodate more specific patient cohorts, such as the Paediatric Early Warning Score (PEWS) [57] in Figure 2.4. With the exception of paediatric scoring systems, the (typical)

Wellington Score Table

Parameter	MET	3	2	1	0	1	2	3	MET
Respiratory Rate (bpm)	< 5	5 - 8		9 - 11	12 – 20		21 – 24	25 – 35	> 35
SpO_2 (%)		≤ 91	92 – 93	94 – 95	≥ 96				
Supplemental Oxygen			Yes		No				
Temperature (C)			< 35	35.0 – 35.9	36.0 – 37.9	38.0 – 38.9	≥ 39.0		
Systolic BP (bpm)	< 70	70 – 89	90 – 99	100 – 109	110 – 219			≥ 220	
Heart Rate (bpm)	< 40		40 – 49		50 – 89	90 – 110	111 – 129	130 – 139	≥ 140
Level of Consciousness					Alert			V or P	U or F

Score	Clinical Interpretation
1 – 5 Any VS in 1	Manage pain, fever, or distress. Increase monitoring frequency.
6 – 7 Any VS in 2	Acute illness or unstable chronic disease Increase monitoring frequency.
8 – 9 Any VS in 3	Likely to deteriorate rapidly. Increase monitoring frequency.
≥ 10 Any VS in MET	Immediately life-threatening illness Call medical emergency team.

Figure 2.3: Scoring table and escalation table of the Wellington Score (NZEWS). NZEWS assigns a warning score from 0-3 to each of 7 clinical parameters. A parameters that exceed the score of 3 require an immediate MET call. A higher aggregate warning score across all parameters or a more extreme single parameter requires an increased frequency of patient monitoring. A patient's level of consciousness is classified as alert, reacts to voice (V), reacts to pain (P), unresponsive (U), and fitting (F).

Paediatric Early Warning Score

Parameter	0	1	2	3
Respiratory	Normal parameters No retractions	>10 above normal parameters Using accessory muscles 30+%FiO2 or 3+liters/min	>20 above normal parameters Retractions 40+%FiO2 or 6+liters/min	≥ 5 below normal parameters Retractions 50+%FiO2 or 8+liters/min
Cardiovascular	Pink Capillary Refill 1-2 sec	Pale or dusky Capillary Refill 3 sec	Grey or cyanotic Capillary Refill 4 sec >20 above normal parameters	Grey or cyanotic and mottled Capillary Refill ≥ 5 sec >30 above normal parameters Bradycardia
Behavior	Playing Appropriate	Sleeping	Irritable	Lethargic / Confused Reduced Pain Response

Age	Heart Rate at Rest	Respiratory Rate at Rest
0 – 1 Month	100 – 180	40 – 60
1 – 12 Months	100 – 180	35 – 40
13 Months – 3 Years	70 – 110	25 -30
4 – 6 Years	70 – 110	21 – 23
7 – 12 Years	70 – 110	19 – 21
13 – 19 Years	55 – 90	16 – 18

Figure 2.4: Scoring table and age-stratified healthy reference ranges of a Paediatric Early Warning Score (PEWS). PEWS assigns a warning score from 0-3 to each of 3 clinical parameters. Notably, the abnormality of heart rate and respiratory rate is stratified by the patient's age, however the normal ranges to not always change between age groups.

differences between many of these EWSs are superficial changes to the vital-sign thresholds [52]. This can be confirmed be examining the various vital-sign thresholds in Figures 2.1, 2.2, and 2.3. Table 2.1 contains the year in which several of these system were first designed and then subsequently updated, along with the number parameters under consideration in the original version that particular EWS system.

Warning Score	Parameters*	Versions
NEWS	7	2012 [41], 2017 [42]
MEWS	5	1999 [54] (Multiple others from disparate sources)
ZNEWS	7	2011, 2015 [56]
PEWS	7	2005 [57] (Multiple others from disparate sources)
APACHE	34	1981 [58], 1985 [59], 1991 [60]**, 2016 [32]**
SAPS	14	1984 [61], 1993 [62]**, 2007 [63]**
Notes:	* = Original EWS	** = Empirical EWS

Table 2.1: Common heuristic early warning score systems

NEWS literature has described many shortcomings, for example the need to include further important information such as age [64], mobility [65], near-patient testing [66], D-dimer levels [67], and laboratory results [68]. Despite this, few studies have suggested alterations to the EWS based on specific data. Even those systems that did alter the thresholds (e.g., changing from NEWS to PEWS, with graduated vital-sign ranges by patient-age) did not specify the particular evidence for selecting one threshold over another. Several studies suggested alterations by providing data-driven thresholds for a single parameter of interest, but left thresholds for other parameters unaltered and, therefore, heuristic.

Exceptions to this include the APACHE and SAPS scores, both of which were originally determined via heuristic expert consensus but later changed to empirical weightings. In particular, when transitioning from APACHE II [59] to APACHE III [60], or from SAPS [61] to SAPS II [62], both scores transitioned to the use of multiple logistic regression (trained on data of patients' in-hospital mortality) to determine weightings.

Since these systems are designed heuristically, explicit descriptions (beyond "expert consensus") of their design process is difficult. However, the performance of

these systems is described extensively in the literature. For example, various MEWS scores have been assessed in a variety of clinical scenarios including for the general ward [66, 69], emergency department [50, 67, 70], ICU, and post-surgical patients. EWS have been evaluated for a range of clinical outcomes, including emergency ICU (re)admission [50], in-hospital cardiac arrest [71], in-hospital mortality [52, 66, 68], n-day mortality [67, 68], and hospital length-of-stay [66]. Such systems are present in the US [16], UK [41], and British Common Wealth countries [55, 68], as well as Western Europe [67, 72].

Instead of considering each threshold simultaneously, tree-based decisions consider a sequence of thresholds. The sequence and type of decision made at each point is usually derived from the clinical application, which adds a measure of specificity. Since decision trees are usually tailored to specific clinical applications, it is more difficult to find instances in which the same decision tree is examined across multiple studies. Decision trees are typically limited to a small number of rules to minimise complexity.

A final example of heuristic monitoring is nurse intuition. Studies have described the predictive value in nurse intuition for over 30 years [73]. For example, Clifton et al [74] noted that nurse observations that incorrectly alert or fail to alert are highly predictive of future alerting/non-alerting observations. This implies that clinical staff are prone to incorporate information that is not currently encoded within EWSs. In [75], a two-fold approach was suggested to improve EWS: (i) to incorporate "nurses' worry or concern" into early scoring systems, and (ii) to identify concrete factors for nurses' worry which would provide further objective criteria in EWSs. One example of nurse intuition that is difficult to codify is their familiarity with the patient. As described in [76], nurses were more capable of interrupting preventable cardiac arrest if they were knowledgeable of the patient's medical history. Unfamiliarity with the patient's medical history, removed that personalised inference and therefore diminished the nurse's ability to interrupt preventable adverse events.

2.2.3 Discussion and Critiques of Heuristic Systems

Critiques of the heuristic early warning systems generally address either (i) the implied the clinical inference or (ii) the clinical implementation of the heuristic method. Heuristic early warning systems have well-established modes of failure for both implementation and inference.

Modes of Failure in Implementation of Heuristic Systems

For implementation, modes of failure generally pertain to either the staff-intensive burden of calculating the risk scores, or human errors associated with hand calculation.

Due to the many time burdens on clinical staff, the time spent hand-calculating heuristic EWSs necessarily takes time from other aspects of care. Staff who are busy recording the vitals-signs and performing computation by hand for one patient, may miss the earliest signs of deterioration in another patient, leaving the deterioration undetected.

Human error is also a significant factor in implementing these heuristic EWSs. The error-rate in manually versus electronically recorded vital-signs has been estimated to be 18.75% versus 0% [77], and 10% versus 5% [78]. In the absence of recording error of the vital-signs, there is also error in the calculation of the EWS. In Smith et al [79], 21.9% of EWS calculations were incorrect. The proportion of scoring errors by vital-sign were respiratory rate, 9.6%, heart rate 5.4%, systolic blood pressure 4.3%, and temperature 3.9%. Presumably, some portion of the remaining 76.8% were due to incorrectly summing of the scores assigned to individual vital-signs. Incorrect scoring was most likely when the true EWS values were higher and those that should have triggered. Clifton et al [74] found that incomplete nurse recordings were twice as likely (7.6% vs 15.1%) to contain vital-sign measurements that should have led to an alerting score.

Both staff burden and human error can be immediately remedied via simple software solutions. The simplest examples are devices with pre-set factory setting to alarm in the presence of extreme vital-sign values. Furthermore, numerous

electronic health record systems automatically record vital-sign measurements from the bedside monitor. These same systems can also automatically calculate correct EWS. Several systems with proprietary EWSs display both their own score and a more traditional score, such as NEWS or MEWS, for clinical comparison.

Modes of failure in clinical inference

Heuristic EWS leave much to be desired, even if the concern of implementation were addressed.

To begin, these heuristic early warning systems are not designed (based on evidence) to optimise a particular measure of clinical performance. Of the EWS papers described, none made an explicit reference to patients' vital-signs or clinical outcome data to inform the creation of the score. In rare cases, a study will describe the individual vital-sign parameters with data-driven thresholds. We can contrast this to the empirical centile-based EWS (described later) for which each EWS threshold is assigned according to the distribution of vital-sign measurements from a patient cohort.

With the exception of the rarely-used trajectory methods [49], these heuristic EWS look at vital-signs at single point in time and ignore time-series dynamics. Steep trajectories or erratic volatilities are not formally implemented in the decision process. Any such inference then, is left to the clinical intuition of nursing staff. Similarly, a patient might demonstrate significant variation, but remain entirely within the pre-specified alarm bounds of the EWS. Furthermore, the described EWSs do not assess whether vital-signs are jointly abnormal, they simply aggregate signals of abnormality for each individual vital-sign.

In terms of patient-specificity, the described heuristic early warning systems are tailored, at most, to a cohort of patients. This results in early warning systems with wide ranges for "normal" vital-sign values where there ought to be precise patient-specific ranges. As shown in Chapters 3, 5, 7, and 8, individual patients tend to occupy a relatively small range of values, compared to the wide range of variability across the population. Inter-patient variability remains even when

patients are stratified by clinical outcome. An example of this will be shown in Chapter 3, when patients have wide individual ranges, even when grouped according to whether they experience cardiorespiratory instability. In this, alarm systems are made highly insensitive to accommodate the range of values within a cohort.

Furthermore, patient-specificity or personalised-risk (as learned by nurses in [76]) may be difficult to codify within the rigid confines of a NEWS-like system, empirical approaches may implicitly codify patient-specific information via (i) comparability to already-observed patients of a similar phenotype, and (ii) acquisition of patient-specific data via continuous monitoring. The prior/likelihood paradigm of Bayesian inference may naturally handle both (i) and (ii).

A further critique of heuristic EWSs is that they ignore the uncertainty in vital-sign measurements (or other physiological parameters of interest). Decisions regarding a measurement's uncertainty are generally left to the nurse, instead of being contained within the decision process. Furthermore, it is unclear how to handle such systems in the presence of missing vital-sign measurements, other than to forgo clinical inference. In contrast, the probabilistic empirical methods, described next, parametrise the uncertainty in vital-sign measurements, when the vital-sign is either missing or present.

A final critique of heuristic EWSs is that the score dichotomises a continuous vital-sign measurement, thereby creating arbitrary delineations between nearly-identical vital-sign values. For example, it makes little sense that two patients with respective systolic blood pressures of 210 mmHg and 220 mmHg would have respective EWSs of 0 and 3, under the Wellington EWS [55]. The troubles associated with dichotomising a continuous entity are a well-described in medical literature [80, 81] for diagnostic and prognostic models. It is worth noting that whereas the dichotomisations described in medical informatics literature [80, 81] are generally tuned to optimise metrics of clinical performance, the heuristic EWSs (which are prognostic tools as well) fall into the trap of dichotomisation, but without reference to a clinical goal. This represents a 2-fold failing.

2.3 Probabilistic Empirical Approaches to Patient Monitoring

2.3.1 Overview of Probabilistic Patient Monitoring

Empirical approaches to patient monitoring attempt to learn more explicit relationships between the patient's predictive features (e.g., vital-signs) and clinical outcomes of interest (e.g., mortality, emergency ICU readmission). Probabilistic methods may incorporate the inherent uncertainty in the predictive features and/or the clinical outcome when formalising this relationship.

There are no probabilistic methods that are not empirical, since probabilistic methods are motivated by the properties of clinical data and estimated via clinical data. The application of Gaussian processes (GPs) to patient monitoring is a specific empirical probabilistic method that will be discussed in a separate section. Methods that are still empirical but do not make considerations of uncertainty are "non-probabilistic", and are covered in the next section.

An advantage in assessing novel empirical models is that their empirical performance compared against alternative methods is usually discussed in the paper proposing the novel method. For example, it is typical to compare the performance of a novel patient monitoring algorithm to (i) common clinical practice, and (ii) more sophisticated alternatives. Performance is usually measured by area under a receiver operating characteristic curve (AUROC) for accurate classification of a deteriorating patient (defined by in-hospital mortality/cardiac arrest, emergency transfer to ICU, etc.).

As described earlier, data-driven methods may add value by using the large amount of available patient data (acquired in routine practice) in a more rigorous fashion. Probabilistic methods are popular because they provide a familiar and principled framework in which to handle uncertainties and correlation in relevant physiological measurements and/or the clinical outcomes of interest.

Many methods that use some form of probability (or are amenable to probabilistic interpretation) have been examined for patient monitoring. We define

several broad methods for probabilistic deterioration detection: (i) regression-based methods, (ii) one-class classification or novelty detection, and (iii) cluster-based classification methods.

2.3.2 Regression-based Methods

Regression-based methods are the first of three broad categories of probabilistic methods.

Regression methods in deterioration detection specify a functional relationship between the patient's measured physiology and particular adverse events of interest. The adverse event of interest can be general risks of the hospital population (e.g., in-hospital mortality) or specific to a ward or patient group (e.g., failure of a particular organ). The scores from such systems are intended to be either (i) static, characterising the patient's status over the duration of stay, or (ii) transient, necessitating subsequent remeasurment and recalculation. Finally, to accommodate manual calculation of a risk score such methods are typically paired with additional decision thresholds to inform clinical action. Although similar in concept to the heuristic EWS described above, strengths of these empirical regression models include (i) clinically-principled model selection, (ii) data selection and experimental design, and (iii) diligent testing/validation on a held-out patient population. This final element is crucial for clinical implementation.

The scoring tables for three logistic regression-based EWS are provided in Figure 2.5 for APACHE II, Figure 2.6 for SAPS II, and Figure 2.7 for LODS. An immediately apparent difference in these scoring tables and those of the heuristic EWS of Figures 2.1, 2.2, 2.3, and 2.4 is that the empirical systems are not constrained by the 0-1-2-3 scoring progression. Instead, the score assigned to different ranges reflects the increased risk of mortality. These methods, however are still subject to the illogic of cut-off thresholds to escalate risk score, described in [80, 81].

Regression methods are popular in clinical literature presumably due to their familiarity and interpretability. Since these models abound in traditional epidemiological, pharmaceutical, and medical research (see, e.g., [80, 82]), they are a

Physiologic Variable	APACHE II Score								
	High Abnormal Range					Low Abnormal Range			
	+4	+3	+2	+1	0	+1	+2	+3	+4
Rectal Temp (C)	≥41	39–40.9		38.5–38.9	36–38.4	34–35.9	32–33.9	30–31.9	≤29.9
Mean Art Pressure	≥ 160	130–159	110–129		70–109		50–69		≤ 49
Heart Rate	≥ 180	140–179	110–139		70–109		55–69	40–54	≤ 39
Arterial pH	≥ 7.7	7.6–7.69		7.5–7.59	7.33–7.49		7.25–7.32	7.15–7.24	< 7.15

Figure 2.5: The APACHE II scoring table. APACHE II is calculated upon entry to ward and remains unchanged. APACHE II assigns a warning score from 0-4 to each of 12 clinical parameters (a selection of which is shown). A further score from 0-6 is assigned for a patient's age. A further score from 0-12 is assigned according to the patient's Glasgow Comma Score (GCS).

Parameter	SAPS II Score								
	Low Abnormal Range					High Abnormal Range			
Heart Rate			< 40 (11)	40-69 (2)	70-119	120-159 (4)	≥ 160 (7)		
SBP			< 70 (13)	70-99 (5)	100-199	≥ 200 (2)			
Temp					< 39.0	≥ 39.0 (3)			
GCS	< 6 (26)	6-8 (13)	9-10 (7)	11-13 (5)	14-15				

Figure 2.6: The SAPS II scoring table. SAPS II is calculated upon entry to ward and remains unchanged. While SAPS II assigns a warning score to clinical parameters according to thresholds, the score assignment is based on the risk calculated by a logistic regression model, as opposed to an arbitrary score escalation from 0-3. A selection of the SAPS II parameters is shown.

common feature in a clinician's formal education. Such models are ubiquitous in technical fields as well, allowing an easy interface between technical experts and medical practitioners.

Two exemplar regression-based methods are described in turn: the Acute Physiology And Chronic Health Evaluation (APACHE) Score and and the Sequential Organ Failure Assessment (SOFA) Score[1].

Acute Physiology And Chronic Health Evaluation (APACHE) Score

The APACHE score was first developed in 1981 [58], and subsequently updated as APAPHE II in 1984 [59], APACHE III in 1991 [60][2], and APACHE IV in 2006 [32]. APACHE and APACHE II were heuristic systems, whereas the weightings of the APACHE III and APACHE IV scores are derived empirically from a logistic regressor. The APACHE score is static: it is intended for calculation upon admission

[1] Frequently called the sepsis-related organ failure assessment score, as well.
[2] APACHE III was updated in 1998 while retaining the same name [32]

		Logistic Organ Dysfunction Score						
System		Low Abnormal Range				High Abnormal Range		
		5	3	1	0	1	3	5
Neurological	GCS	3-5	6-8	9-13	14-15			
Cardiovascular	Heart Rate	< 30			30-139	140		
	Systolic BP	< 40	40-69	70-89	90-239	240-269	≥ 270	
Hematological	Leukocytes (× billion/l)		< 1	1.0-2.4	2.5-49.9	≥ 50.0		
	Platelets (× billion/l)			< 50	≥ 50			

Figure 2.7: Logistic Organ Dysfunction Score (LODS) scoring table. LODS is calculated upon entry to ward and recalculated throughout the stay on ward. While LODS assigns a warning score of 0, 1, 3, or 5 to clinical parameters according to thresholds, the score assignment is based on the risk calculated by a logistic regression model, as opposed to an arbitrary score escalation. A selection of the LODS parameters is shown from three different physiological systems: neurolical, cardiovascular, and haematological.

to the ward (no later than the first 24 hours) and characterises patient-risk for the duration of stay.

Since APACHE is a proprietary system, the weightings of APACHE IV are currently unpublished. However, the derivation of the APACHE IV system, described in [32], would be familiar to those from a medical informatics or machine learning background: The APACHE IV system was created to address the falling predictive accuracy of APACHE III over the previous decade. The data comprised 110,558 patients (who met inclusion criteria) from 45 hospitals. Of these patients, 60% were used to train a logistic regressor, and the remaining 40% as the validation set. The validation set was further broken down to assess calibration and discrimination on specific patient sub-populations. The predictive features were various physiological measurements collected during the first 24-hours on ward, with the set of observations for each patient being the "worst" value recorded within the first 24 hours. The recorded clinical outcomes were (i) length of stay in ICU (in hours), (ii) length of stay in hospital (in hours), (iii) mortality at ICU discharge, and (iv) mortality at hospital discharge. The logistic regression itself was based on mortality at hospital discharge. To avoid the inaccuracies induced via the assumption of linear coefficients, restricted cubic regression splines and restricted cubic spline transformations were applied. However the method for selecting the number and locations of spline knots was not disclosed.

In addition to APACHE's derivation as a probabilistic model, its calibration is assessed probabilistically as well, by comparing the estimated probability of mortality to the actual mortality rates within the patient cohort.

The APACHE score's prediction is frequently compared to the Mortality Probability Model (MPM) and the Simplified Acute Physiology score (SAPS), since each method (in its latest form) is (i) derived from logistic regression estimating probability of mortality, (ii) static after being calculated in the first day of stay on ward, and (iii) frequently considered applicable to stratify risk in non-ICU patient cohorts. It should be noted that while the MPM is calculated at the time of admission, there are further unique models for MPM to characterise the patient at 24, 48, and 72 hours. Like APACHE, these systems use the "worst" observation for each predictive feature to describe the patient's current status. Further information on MPM, MPM II, and MPM III can be found in [83], [84], and [85], respectively.

Sequential Organ Failure Assessment (SOFA) Score

The Sequential Organ Failure Assessment score, also known as the Sepsis-related Organ Failure Assessment score, was designed to monitor the changing patient status due the deterioration of specific organ systems. In particular, SOFA monitors organ failure due to a specific ailment: sepsis. Unlike APACHE, SAPS, or MPM, SOFA is (i) intended for continuous recalculation, and (ii) not explicitly modelled to predict mortality.

The SOFA score combines an assessment on several physiological systems, including kidney, liver and brain function. Although proposed for predicting the morbidity associated with organ failure, it has been further calibrated to correspond to mortality rates [86, 87]. The sequential recalculation of SOFA serves to (i) update the patient's current status, and (ii) allow for trend-based clinical inference. A SOFA score trend carries its own implications for clinical prognosis: regardless of initial SOFA score, patients with a decreasing SOFA score have lower mortality rates than patients with the same initial SOFA score but with stable or increasing SOFA scores thereafter [87].

To further aid manual calculation, the Quick SOFA (qSOFA) [88] uses three simpler criteria, instead of six more complex criteria. For qSOFA, an emergency is triggered if the patient meets two of the three criteria.

Similarities between the SOFA and Multiple Organ Dysfunction Score (MODS) are apparent, including that (i) both use six similar criteria (focused on organ function instead of generic vital-sign measurements), (ii) the scores are intended to be recalculated frequently, and (iii) the scores have been applied to many patient populations and many adverse events (due to their discriminative ability).

2.3.3 Novelty Detection-based Methods

Novelty detection-based methods are the second of three broad categories of probabilistic methods.

It is fortunate for patients that adverse clinical outcomes are in the minority of clinical observations. However, this creates a challenge for empirical modelling, since there is little data on which to build a model to contrast patients with different outcomes. Novelty detection circumvents this problem by generating a specific model only for the class for which there is ample available data. Abnormal cases are then identified by their divergence from the well-defined class. An extensive coverage of novelty detection is provided by Pimentel et al [47], who define novelty detection as "the task of recognising that test data differ in some respect from the data that are available during training." The value of such methods, it is explained, come when the preponderance of training data is from the "normal" class, thereby hindering the creation of a satisfactory explicit model for the "abnormal" class. These methods model a single class of interest, but have no model for other classes of interest. A patient is considered novel according to his divergence from a model of normality, instead of his similarity to abnormal examples (which is the case for regression and cluster-based methods). Novelty detection methods that have been applied in the ward tend to focus on defining the distribution of physiology in healthy patients, and alarming when observations are made that fall into the extreme tails of those distributions.

Examples, of such methods include (i) kernel density estimation, (ii) quantile estimation, and (iii) extreme-value distributions.

Kernel Density Estimation

The kernel density estimate (initially proposed by Rosenblatt [44] and Parzen [43]) is a non-parametric approach to modelling the joint distribution of multiple random variables. This method forms the basis for the current technical state-of-the-art, as described in [45] and is described further in Chapters 3, 4, 7, and 8.

KDE-based novelty detection provided the basis for the first CE marked and FDA-approved data-fusion algorithm for critical care patient monitoring. The method has been associated with commercial names such as BIOSIGN™[89] and Visensia™(OBS Medical) [34].

For patient monitoring, the KDE models the joint distribution of key vital-signs, such as heart rate, respiratory rate, SpO_2, and blood pressure from a cohort of "healthy" patients. New patients' vital-signs can then be compared to this distribution, and alarm on low-likelihood values. The KDE method is classified as novelty detection because the set of training data only includes patients who did not experience cardiorespiratory instability for the duration of their stay on ward. Typical KDE modelling choices (described further in Chapter 4) include selection of the kernel's bandwidth and the handling of finite-domain variables.

KDE-based novelty detection carries several advantages compared to the heuristic EWS described earlier: Besides being empirical, the likelihood-based novelty score allows for gradual escalation of a score, instead of a small number of stepwise escalations. This means that the vital-sign deviations within any range will affect the novelty score. Furthermore, the joint-distribution of the KDE allows for consideration of the simultaneous abnormality of all vital-signs, instead of merely summing univariate abnormalities, as in NEWS.

In essence, KDE methods replace an absolute threshold with a probabilistic threshold that can account for the correlation between vital-signs. Patient risk,

though, is still assessed at a single point in time, thereby losing information from previous measurements, and making the unrealistic assumption of i.i.d. observations.

In addition to deterioration detection, KDE methods have been used to identify artefactual anomalies in vital-sign data.

Centile-based Early Warning Scores (CEWS)

While the KDE method models the joint distribution of all vital-signs, centile-based early warning scores (CEWS) model the univariate vital-sign distributions to inform the familiar thresholds used in common EWS systems. The clinical applications of CEWS methods have been described in [90] and [91].

Unlike the heuristic thresholds described earlier, CEWS' threshold-selection was both transparent and data-driven: A CEWS of 1, 2, and 3 were assigned to the outermost 10, 5, and 1 percent in each tail. For example, a heart rate measurement in either the lowest or highest 1% of heart rate measurements received a CEWS of 3.

The authors explained that optimising the thresholds according to the distribution of vitals is preferable to directly tuning the thresholds to optimise the AUROC. The latter method would bias thresholds (i) in favour of identifying "salvageable" patients who died under the current system, and (ii) against identifying "salvageable" patients who did not die under the current system (since the system would not reward identifying such patients). While this observation is clear in the case of assessing a system's sensitivity and specificity with regard to patients with in-hospital mortality, it is less-clear in the case when "deteriorated" patients are labelled as such due to their need for clinical action (without regard to a resultant clinical action). This latter approach to patient-outcome labelling is used in this thesis, as well as other studies, including [16, 21, 28, 34, 45].

Furthermore, CEWS moves towards an established rate of alarms on ward, since each score is directly interpretable as a proportion of all vital-sign values. This provides nurses with a greater understanding of the abnormality implied by such an alarm.

Further developments of CEWS-like methods include stratifying patients by further characteristics, for example the Age- and Sex-Specific Early Warning Scores (ASEWS). These methods take the same approach of using extreme quantiles of patient data to define thresholds, however the data under consideration is restricted to particular patient cohorts to adapt the EWS to known physiological differences that occur due to sex, ageing, or both.

Extreme Value Distributions

A final novelty-based approach makes use of the Fisher-Tippett Extreme Value Theorem (EVT).[3] Instead of modelling the distribution of all vital-signs, and alarming in the presence of observations in the tails of a distribution, EVT-based methods use the distribution of vital-signs to perform inference on the tail-behaviour of vital-signs. A principled inference of the tail distributions may be helpful to anticipate whether extreme observations represent a true deterioration, an artefactual measurement, or a reasonably-expected extreme value arising from long-periods of continuous monitoring. The value of EVT-based methods are typically championed when it is desirable to have a probabilistic novelty threshold that can be adjusted automatically, with reference to the amount of available data under consideration.

An early example of EVT used for a variety of biomedical signals can be found in [92], which uses Gaussian mixtures as a model of normality, and EVT to model anomalous data with respect to that model. Although [92] did not apply EVT to vital-sign data as described in this thesis, it provided a basis for direct applications to vital-sign monitoring, including a doctoral thesis [93], which applied EVT to the same SDU data set as used in this thesis.

Since the CEWS systems described above were designed to mimic univariate-oriented EWS, they fail to account for tail behaviour due to covariance. In contrast,

[3] Fisher and Tippett's EVT states that the extremum (minimum or maximum) of any sample from any distribution must follow one of three distributions. It may be thought to provide the limiting distribution of sample extrema, in the way that the Central Limit Theorem provides the limiting distribution to sample means (which is the Gaussian distribution).

EVT systems have been proposed which account for the multivariate and multi-modal distributions of correlated vital-signs [94–96].

Outside of deterioration detection in critical care, EVT-based methods were used to study patients with Crohn's disease in the doctoral thesis by Niehaus [14]. EVT was used both to model patients's lab-value time-series, as well as to derive features for subsequent machine learning classification.

2.3.4 Cluster-based Methods

Cluster-based methods comprise the third and final broad category of probabilistic methods.

Cluster-based patient models examine how patient groups differ in the distribution of their predictive features. The understanding is that patients with different outcomes will generally diverge according to salient physiological metrics.

Cluster-based methods, when applied to classification, make use of data availability for each patient class of interest. Unlike the logistic regression-based EWS classification, these clustering models typically build a probabilistic model of the predictive feature, conditional on membership to a class. Clustering may also be used as a pre-processing step for novelty detection methods, to create a more succinct set of "normal" data points, as will be seen in the practical implementation of KDEs in Hann [45] (described in Chapter 7) and the vital-sign trajectory clustering of Pimentel [97].

In Yamamoto et al [98], the time-series of dialysis patients were hierarchically clustered. Pulse rate, respiratory rate, and body movement time-series within the same dialysis session were compared via (non-probabilistic) multi-dimensional dynamic time warping, to assess intra-session agreement of the vital-signs. The distance between intra-session vital-signs were then hierarchically clustered via the (probabilistic) Ward method, which is based on Cophenic correlation (a measure of how reliably a dendrogram preserves correlation). Although the various clusters lacked interpretable medical features, it was noted that session clusters followed

seasonal effects. This was interpreted to mean that biological patterns could be influenced by environmental factors.

Schmidt et al [99] used k-means clustering to identify patient subgroups according to their emergency department vital-signs. Intergroup similarities were then compared for clinically significant differences and probabilistically significant differences (via entropy measures and the Kolmogorov-Smirnov test).

Lee and Yetisgen used unsupervised random forest [100] (which trains a discriminator via adversarial generation of synthetic data) to generate patient features from the MIMIC III data set. Summary statistics such as mean, minimum, maximum, and standard deviation, for each vital-sign (heart rate, respiratory rate, mean arterial pressure, and body temperature) were calculated for each patient-day. After unsupervised random forest, k-medioid clustering was used to create interpretable patient clusters to stratify patients by mortality rate.

In the doctoral thesis by Bose [101], statistical features were extracted from heart rate, respiratory rate, and SpO_2 time-series (e.g., mean, median, minimum, maximum, variance, and range). The derived features were clustered via k-means (optimising squared Euclidean difference) and the clusters were linked hierarchically via Ward's linkage procedure. After an initial clustering, further statistical analysis was performed to identify and remove less-important features, and to identify an optimal number of k-means clusters. Having reached a satisfactory number of clusters comprised of the most important features, the driving factors of cardiorespiratory instability were identified. For example, respiratory rate was identified as the driver of 50% of all cardiorespiratory instability events, followed by SpO_2 (33%), heart rate (14%), and blood pressure (4%). SpO_2 was identified as the most frequent driver of first events, as well as the most frequent cause of artefactual alarms. This study is of particular interested because the SDU data under analysis was acquired from the same ward as the data used for this thesis.

2.4 Non-Probabilistic Empirical Approaches to Patient Monitoring

Non-probabilistic machine learning methods have demonstrated strong predictive capacity in many applications, including health care. Well-known healthcare applications of non-probabilistic machine learning include signal-processing, computer vision, bioinformatics, and mortality prediction. Many non-probabilistic methods have a probabilistic interpretation (e.g., neural networks, random forests), but these probabilistic aspects typically receive a minority of the focus when designing and evaluating these methods. Unlike many of the probabilistic empirical systems (e.g., APACHE, Visensia, prognostic regression, CEWS), it is difficult to find many instances of non-probabilistic approaches being implemented in the ward for patient vital-sign monitoring. This "plague of pilots", in which many solutions are proposed but few are used, may be due to the perception of such methods as being a "black box", and therefore not amenable to the transparency required by clinicians and regulators.

Decision trees are an intuitive choice for a method to replace heuristic thresholds with empirical thresholds in early warning systems. The value proposition of decision trees includes (i) validation of current heuristic EWS thresholds by comparison to empirically-derived thresholds, and (ii) automated generation of manually-calculable EWS for bespoke clinical environments [102]. Like the EWS decision-process, decision trees naturally subdivide the feature space according to outcome. In [102], the decision thresholds of NEWS were compared to the decision thresholds that a decision tree would have selected, if presented with the same set of predictive features. Only a single decision tree was used, instead of an ensemble of decision trees. Unlike NEWS's heuristic assignment of scores to high and low vital-sign values, the Decision Tree EWS (DTEWS) assigned a score of 1, 2, or 3, according to the risk of an adverse event at each node (as a proportion of the baseline risk of deterioration). Three clinical outcomes were examined: cardiac arrest, mortality, and ICU transfer. The DTEWS algorithm selected decision thresholds that were very similar to those selected by NEWS. A similar DTEWS approach [103], from

the same group, has been used for laboratory tests. Further examples of decision tree-derived EWS can be found in [104] and [105].

Ensembles of decision trees have also been popular to derive empirical thresholds for early warning systems. For example [106] (using data from the same SDU as used in this thesis) used random forest to differentiate vital-sign artefact from true alarms. The random forest classified artefactual values using (undisclosed) features derived from the time-series.

Kernel-based learning methods, such as support vector machines (SVMs) are also a popular comparator for the GP-based methods described in the next section. SVMs have demonstrated comparable accuracy at forecasting, interpolation, and imputation tasks, but lack the probabilistic descriptions of the uncertainty in those forecasts, which GPs provide [107]. Gultepe [108] used SVMs (and probabilistic Naive Bayes classifiers) to predict mortality in sepsis patients using only 3 vital-signs, along with white blood cell count, and lactate level. The classifiers were advantageous over previous machine learning approaches, and current practice, due to the reduced number of predictive features required to achieve good predictive performance.

2.5 Gaussian Processes for Vital-Sign Modelling and Deterioration Detection

GPs are a flexible and principled way to model a variety of functions, including for regression, classification, time-series, and spatio-temporal modelling tasks. It is, therefore, unsurprising that there exists an ever-growing body of literature describing the use of GPs to model and predict vital-signs. The probabilistic framework of GPs assists in the modelling of vital-signs, which are typically recorded using wearable sensors, which induce various noise components, such as sensor noise, quantisation, and artefact arising from patient movement or signal processing errors. The GP's flexibility allows it to handle a variety of modelling tasks. For example, as a regressor, the GP can perform forecasting, interpolation, and missing-value-imputation tasks with measures of uncertainty in its prediction. This allows GPs to serve both as a modelling approach in its own right, as well as serve as a pre-processing step to

subsequent analytical steps. As a classifier, GPs can circumvent unnecessary and inaccurate assumptions of linear relationships between predictors and outcomes (which concerned the authors of the later APACHE systems [32, 59, 60] as well).

Durichen et al [109] used multi-task GPs to model the correlation between nurse-recorded observations in heart rate and respiratory rate. When compared to single-task Gaussian processes (which do not account for inter-vital-sign correlation) a multi-task approach demonstrated improved estimation of vital-sign values at time points where (i) heart rate, or (ii) both heart rate and respiratory rate were missing. This implies that GPs can improve the imputation of missing vital-sign values. A similar example is given in [110], with the multi-task modelling extended to include systolic blood pressure, as well as heart rate and respiratory rate.

A more concrete description of the value of empirical imputation is provided in [107], which compared (i) univariate GPR-based imputation and (ii) support vector regressor-based (SVR) imputation to more common heuristic approaches of vital-sign imputation that used the population-mean or patient-mean. The empirical methods demonstrated marked improvement in accurate imputation of the missing values over the heuristic methods. The GP and SVR demonstrated comparable performance, while the GPR imputation also provided an estimate of uncertainty in the imputed value. Clifton et al concluded with an illustration of how mean-imputation could falsely reduce a deterioration alarm using a system such as the EVT-based method in [96], whereas the GPR method would (retrospectively) infer the time of deterioration during the period of missing values. An extension of this work is described in [48], which further illustrates the same methods, using nurse observations to validate the accuracy of GP imputations of bed-side monitor data.

Wong et al [111] used Gaussian processes to impute missing vital-sign values as well. By providing a complete set of vital-signs, the values could be fed into a patient status index, such as in the KDE-based method from Hann [45], without resort to a heuristic imputation at the population mean. Furthermore, the posterior distribution of the Gaussian process allowed for probabilistic reasoning over both the missing

vital-sign value and the patient status index which results from those vital-sign values. (This second aspect was less-directly discussed in the imputation work of [107].)

In [112], GPs were used to model the periodic components of ECG in order to estimate a patient's respiratory rate coupled with an estimate of the uncertainty in the respiratory rate.

In Pimentel et al [113], multi-task GPs are used to provide an estimate of heart rate and respiratory rate trajectories, which were subsequently clustered via a metric of local likelihood into four template trajectories. These templates were then used to distinguish between deteriorating and non-deteriorating trajectories in a held-out set of test patients. The value of GPR for modelling the trajectories was several fold, including (i) imputation of vital-sign values at arbitrary and constituent time intervals, (ii) the principled estimation of those values, and (iii) a representation of uncertainty at any point (thereby accommodating greater weight for trajectories with greater certainty). The GPR-based method out-performed deterioration detection using clustering via multidimensional dynamic time warping. (As seen previously in the dialysis example of Yamamoto [98], dynamic time-warping is a popular method to cluster multivariate time-series.) An extension of this work in [97] first generated patient clusters (as described above) using only the time-series for temperature and systolic blood pressure. The patient clusters in 2D were then compared to the clustering when heart rate, respiratory rate, and SpO_2 were also included. Approximately 70% of patients remained in the same cluster for the 5D clustering, which suggests that a smaller number of vital-signs drive the differentiation between patient groups.

In Lasko et al [114], GPs were used to pre-process time-series of uric acid measurements. GPs were used to transform "noisy, irregular, and sparse observations to a longitudinal probability distribution". The time stamps of these posterior GPs were then heuristically time warped, and fed as inputs into an autoencoder with the aim to distinguish between the uric acid measurement features of patients with gout and leukaemia. The auto-encoder-learned features were compared to features designed by clinical experts.

Stegle et. al. [115] examined the use of GPs for free-living HR monitoring with 40 adult subjects. Model-fitting and forecasting were improved by the use of clustering based on "auxiliary" ECG-waveform summary data (for example, the variability of inter-beat intervals, the extrema within these intervals, and the fraction of time these extrema fell outside a credible range). The latter were used to identify noisy periods in the data. The auxiliary variables were used to generate clusters of variables with different levels of noise. The noise model was then a mixture of the different classes with varying noise. After clustering, the kernel accounted for two additive components of HR variability: a discontinuous short-term variability component, and a periodic component for diurnal patterns. Model parameters were found using expectation maximisation to approximate the posterior distribution and then choosing values that maximised log-likelihood.

2.6 Potential Applications of Gaussian Process Modelling State-of-the-Art to Patient Monitoring

Despite the many applications describe above, GPs are yet to have been used to their full potential in handling the varied and complex data presented by patients on the ward or at home. The following is a brief survey of the variations on common GPR models which may be useful in critical care monitoring.

GPR holds several advantages over more traditional regression methods. Most broadly, GPR does not impute a functional form relating the dependent variables to the independent variables. Pre-specified functional forms can include those meant to handle non-linearity and other salient features (e.g. polynomial, fractional-polynomial [82], or Poisson regression) or to specify the extent to which previous observations affect future observations (e.g. models with autoregressive or moving-average components, such as [116]). Such pre-specified functional forms lead to inherent challenges in the modelling process. Most obviously, the pre-specified model may be misspecified and deviate significantly from the actual form of the generative process, especially where data are absent or sparse. The kernel-based modelling of

GPs circumvents these issues by avoiding the need of undesirable assumptions of the functional form. Furthermore, GPs are amenable to probabilistic assessment of a kernel's performance, so that the introduction of superior-performing kernels may be automated in the presence of new data [117, 118]. This is certainly desirable in continuous patient monitoring systems, which may wish to change the model in light of further hours or days of information.

Similarly, GPs are amenable to on-line change-point detection [119, 120], for example, to identify when there has been a change in the underlying generative process, so that the previous model fails to describe the data subsequent to the change-point. Plausible causes of change-points include changes due to deteriorating physiology; or probe detachment leading drift, or bias, in the observations (due to signal processing failures of the waveform. It is plausible that in the course of deterioration, the underlying dynamics of a patient's time-series will deviate from the previous dynamics. These applications might be very useful in those instances in which a patient does not have forecastable trajectory of impending deterioration, but which instead can only be identified after a dynamic change has occurred. As a concrete example, the step-change GP work in Chapter 8, which identifies instances of rapid volatility could be supplemented by the change-point models described in [119, 120]. In the presence of a suspected step-change, a change-point model could evaluate the evidence for two separate GPs, occurring at the time-point of the step-change.

A further flexibility is that GPs models are not constrained to Gaussian likelihoods. This is useful since many important vital-signs measurements are certainly not the product of Gaussian random variation, for example SpO_2 which is (i) constrained between 0%-100% and (ii) fixed at 100% during oxygen therapy, and heart rate, in which noise is typically right-tailed. To avoid heuristic transformations (such as log-transformation for positive-valued vital-signs) warped GPs can learn transformations for both the outputs [121] or inputs [122] of the data to improve the Gaussianity or stationarity of the data, implicitly creating a more accurate probabilistic description of the untransformed data. GPs can also form the basis to

segue into copula process modelling [123], which for multi-vital-sign modelling allows for more bespoke modelling of (i) marginal distributions (e.g. a [0,100]-constrained marginal distribution for SpO_2), and (ii) the correlations between the vital-signs.

Other GP developments may be useful in niche clinical environments. McHutcheon et al [124] examine uncertainty in a GPR model and the importance of precise time-stamps depending on how quickly the functions are changing with respect to time. In a patient-monitoring framework, poorly-synchronised monitoring devices will likely contain errors in the times of the recorded measurements, creating temporal uncertainty. This is especially true in mobile or home patient-monitoring where information might be relayed between multiple devices. A simple example could be the Bluetooth connection of a monitoring device becoming disconnected or desynchronised from other devices being used to monitor the same patient. GPs could be designed to accommodate this uncertainty if it were determined to be significant.

Finally, advancements have been made to ensure that GP analysis can scale to the size of the data. Approximation methods allow the GP to model a large number of data points via a smaller number of (empirically selected) inducing points [125] or pseudo inputs [126]. Approaches such as these may be preferable to more heuristic approaches to "sparsify" or down-sample the data without independent evaluation of its affect on the model.

2.7 Conclusion

This chapter has described how early warning systems for patient deterioration detection may be usefully divided into two categories, (i) those systems with heuristic decision criteria, and (ii) those systems with empirical decision criteria. While heuristic approaches are most common in patient monitoring they are prone to well-established flaws that undermine their capability to forewarn patient deterioration. Empirical approaches are less widely-adopted but aim to address the short-comings of heuristic systems in several ways, including (i) clarifying the clinical performance metrics by which a monitoring system is evaluated, (ii) data-driven selection of

monitoring parameters, and (iii) avoiding arbitrary thresholds in clinical decision making. This thesis proposes GP-based methods of patient monitoring to build upon these empirical methods.

3
Data Description

The methods in this thesis were developed using data from a step-down unit (SDU) at the University of Pittsburgh Medical Center (UPMC). The data set comprises 333 SDU patients whose vital-sign data were recorded over the duration of their stays on ward. Using criteria from similar studies at this UPMC SDU, each patient's time-series was retrospectively annotated for extreme vital-sign measurements indicative of cardio-respiratory instability.

A descriptive analysis of patient vital-signs and the annotated clinical emergency is provided. Attention is given to inter-and intra-patient variability in (i) vital-sign measurements, (ii) emergency-event causes, (iii) length-of-stay (LOS) on ward, and (iv) data missingness.

The large inter- and intra-patient variability in vital-sign values demonstrates a fundamental challenge in early warning systems that are based on extreme-valued vital-sign measurements. This challenge motivates the development of personalised clinical inference via each patient's time-series, which is the contribution of this thesis.

3.1 UPMC Data Collection

The UPMC data set comprises 333 adult patients in the surgical-trauma SDU at the University of Pittsburgh Medical Center (UPMC) Presbyterian Hospital. The patients were recorded as phase 1 of a 3-phase trial to optimise and validate the efficacy of the kernel density estimate (KDE) based monitoring system described in [45]. The goal of phase 1 was to use the collected data to optimise the value of the novelty threshold for the KDE used to generate clinical alerts. The KDE model of normality had been constructed using data from two studies at the John Radcliffe Hospital.

Phase 1 of the UPMC trial acquired vital-sign data via Philips MP30 bedside monitors. The KDE was not used to alert clinicians, and the nursing staff was blinded to the KDE model and the KDE model's outputted early warning score. Phase 1 started in November of 2006 and lasted eight weeks. Phase 2 and 3 of the UPMC study were used, respectively, to train nursing staff to respond to alerts generated by the system, and evaluate the use of the KDE-based early warning system.

This thesis uses only data from phase 1 of the UPMC study. Accordingly, the data corresponds well to the current clinical practice in the United States and United Kingdom. (In contrast, the data in phases 2 or 3 would correspond to clinical practice subsequent to the introduction of the KDE-based method.)

3.2 Individual Patient Data

Each UPMC patient's data comprises vital-sign time-series, and a set of annotated emergency events (these events are called C"-events, and are described in the next section). Examples of two such vital-sign records are shown in Figure 3.1, for 3.1(a) a non-C"-patient with no annotated emergency events, and for 3.1(b) a patient with three annotated emergency events. (The exact definition of C"-events will be given in the next section.)

The vital-sign data contains unique time-series for each of five vital-signs: heart rate (HR), respiratory rate (RR), SpO_2, systolic blood pressure (SBP), and diastolic

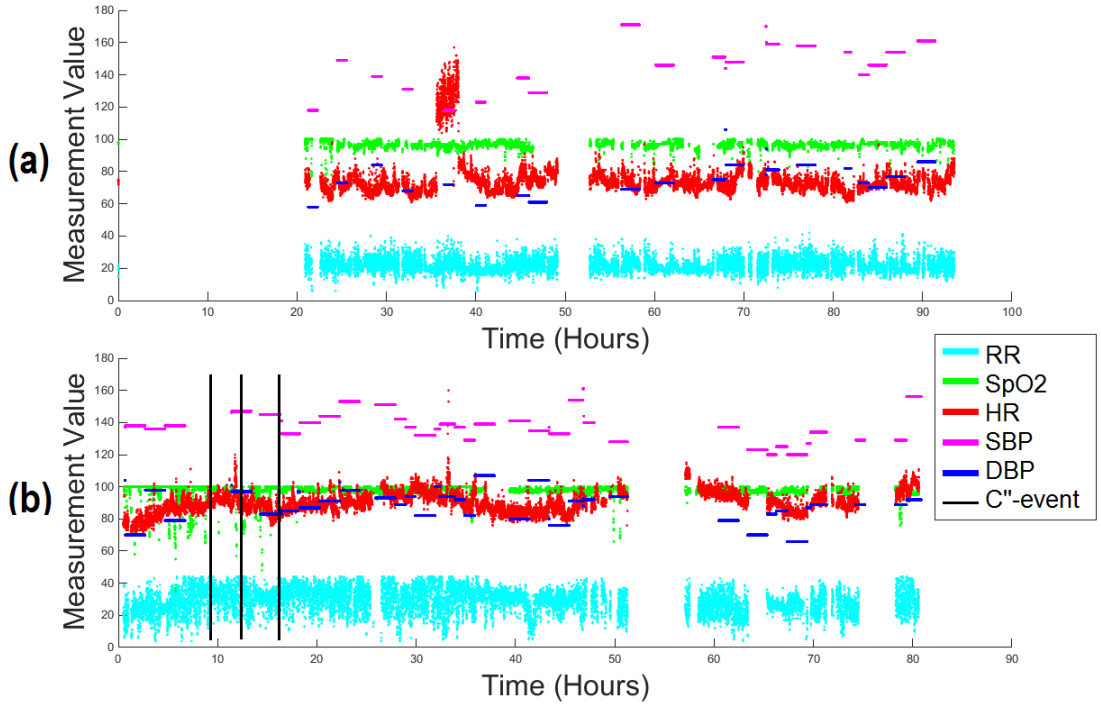


Figure 3.1: A full vital-sign time-series record for (a) a non-C"-patient and (b) a C"-patient. Vital-signs are differentiated by colour. Annotated emergency events are shown by vertical lines. For visual clarity, blood pressure values are held for 30 minutes after the initial measurement. Vital-sign measurements are shown after artefact removal, and down-sampling to $f_s = \frac{1}{60}$ Hz.

blood pressure (DBP), all of which were recorded by the Phillips bedside monitors. Temperature was recorded but not used in any of the models discussed in this thesis. The temperature measurements of the data set were beset with artefacts and were of ambiguous clinical value. For these reasons, temperature is not included in our discussion. The time-series of each vital-sign comprised vita-sign measurements and their associated time-stamps.

Validated clinical emergency events, called C"-events, occurred in some of the patient's vital-sign time-series. Each patient record includes a record of the time-stamp, duration and primary cause of any C"-event. For patients who did not experience any C"-events, this record is empty.

Figure 3.1 illustrates many of the challenges encountered with continuous vital-sign data:

First, each patient has several extended periods of missing measurements. In

3.1(a), the non-C"-patient's time-series begins with only a single initial measurement (at hour 0), followed by 20 hours of missingness. Between approximately 35-38 hours, the non-C"-patient has a step-change with elevated HR. Since this period was not marked as a C'-event (where data exceeded the MET calling limit of 140 bpm), we know that clinical experts considered these measurements to be artefactual. Since sections such as these are neither marked as C'-events nor C"-events, and many of the measurements fall within both the artefactual limits (Table 3.1) and the MET limits (Table 3.2) it is left to the patient monitoring algorithm to identify these questionable measurements and decide how to handle them appropriately. Measurements of this kind were left in the time-series with the understanding that, if taken at face-value, they would represent highly unusual physiological dynamics. Novel approaches to identifying artefacts (using the individual patient's time-series information) are described in Chapter 5.

In 3.1(b), the C"-patient has three C"-events due to high RR (where the MET limit is 36 bpm). The C"-annotation seems appropriate, seeing that RR is visibly elevated over that of the non-C"-patient. On average, the C"-patient's HR is about 20 bpm higher than that of the non-C"-patient. The preponderance of SpO_2 measurements at exactly 100% suggests that the C"-patient received a significant amount of oxygen therapy. (Patients breathing room air typically have SpO_2 values around 95%.) A final observation is SBP and DBP were sampled much more frequently for the C"-patient compared with the non-C"-patient. This suggests that the clinical staff was taking particular interest in this patient following the previous MET-level measurements. This observation supports our choice to focus our predictive methods on only the first C"-event for each C"-patient.

3.3 Annotation of Clinical Emergency Events

3.3.1 Prospective Identification of Emergency Events

Over the course of phase 1 data collection, the medical staff made 7 medical emergency team (MET) calls based on extreme vital-signs (across all 332 Patients). We view these 7 calls as emergency events that were prospectively-identified under

current clinical practice. As described below, these 7 events constituted the minority of emergency events that occurred over phase 1.

3.3.2 Retrospective Identification of Emergency Events

Retrospective annotation of all emergency events was completed by (i) an automated screening of all patient time-series for candidate emergency events, followed by (ii) validation of each candidate event by clinical experts for a final set of validated clinical emergency events. Identical or nearly-identical criteria have been used in numerous other studies at the UPMC SDU to define "cardiorespiratory instability", for example [16, 21, 28, 34].

Automated Screening of Emergency Events

For each patient, the automated screening for potential emergency events was conducted as follows:

1. Artefactual measurements were removed from the recorded vital-sign time-series, according to Table 3.1 (which was taken from [45]).

2. Further artefactual recording errors in respiratory rate were identified and remedied.

3. Vital-signs that would have satisfied the univariate MET calling criteria (for at least 4 of the previous 5 minutes) were identified. The MET calling criteria are shown in Table 3.2 (which was taken from [45]).

4. Alarms occurring close in time were merged into a single alarm event.

This automated process identified 407 candidate emergency events to be validated by clinicians.

	Lower Threshold	Upper Threshold
HR (bpm)	30	300
SDA (mmHg)	20	180
SpO_2 (%)	10	-
Temp (°C)	32	39
RR (bpm)	3	45

Table 3.1: Artefactual thresholds applied to phase 1 patients

	Lower Threshold	Upper Threshold
HR (bpm)	40	140
RR (bpm)	8	36
SpO_2(%)	85	-
SBP (mmHg)	80	200
DBP (mmHg)	-	110

Table 3.2: Univariate MET alarm thresholds

Clinical Validation of Emergency Events

A team of clinical experts were provided with the relevant time-series data to determine whether any of the 407 potential events were indeed non-artefactual. Any events that clinicians confirmed to be non-artefactual exceedances of the MET thresholds in Table 3.2 were labelled as so-called C'-events. There were 237 C'-events between 83 patients.

For each C'-event, the group of clinical experts then decided whether an MET ought to have been called to perform an emergency clinical intervention. C'-events thus confirmed to have warranted an emergency intervention were labelled C"-events. The annotated C"-events each had an associated start-time, stop-time, and primary cause. There were 112 C"-events between 59 patients. (This means that there were 24 patients whose vital-signs exhibited non-artefactual, sustained MET exceedances, but who were not thought to require emergency intervention.)

Patients who experienced a C"-event are labelled C"-patients. Those who did not experience a C"-event are labelled non-C"-patients.

The presence of 112 emergency C"-events when, in practice, only 7 MET calls were made supports the understanding that continuous monitoring can add value to the intermittent observation of nursing staff.

3.4 Characteristics of Annotated Emergency Events

For the purpose of detecting deterioration, we are particularly interested in a patient's first C"-event because it is possible that vital-signs subsequent to that first C"-event may be affected by clinical intervention.[1]

As seen in Table 3.3, SpO_2 comprised almost one-third of first C"-event causes. SBP and HR followed, comprising about one-quarter and one-fifth of first C"-event causes, respectively.

As seen in Figure 3.2, events occurred between 0.1 and 1000 hours after admission. The earliest C"-event occurred within minutes of recording. There is significant heterogeneity, but no obvious relationship between the first C"-event time-stamp and the C"-event's primary cause. Within the same C"-patient's time-series, multiple C"-events tended to share the same primary cause.

	HR	RR	SpO_2	SBP	DBP	Total
Lower Threshold	2	4	20	11	-	37
Upper Threshold	10	6	-	3	3	22
Total	12	10	20	14	3	59

Table 3.3: Count of first C"-event by primary cause

3.5 Missing Vital-Sign Measurement Data

As seen in the example time-series of Figure 3.1, the time-stamps of different vital-signs (e.g. HR, RR, etc.) are not aligned. This is for several reasons, including differing sampling rates, artefact removal, probe detachment, and vital-signs beginning recording at different times. It is helpful to distinguish between two metrics of a patient's available data:

1. Length of stay (LOS): The time from a patient's entry on ward to release from ward.

[1] Similar reasoning is found in the patient exclusion-criteria of APACHE IV [32], to avoid the confounding affects of previous emergency interventions on a patient's physiology.

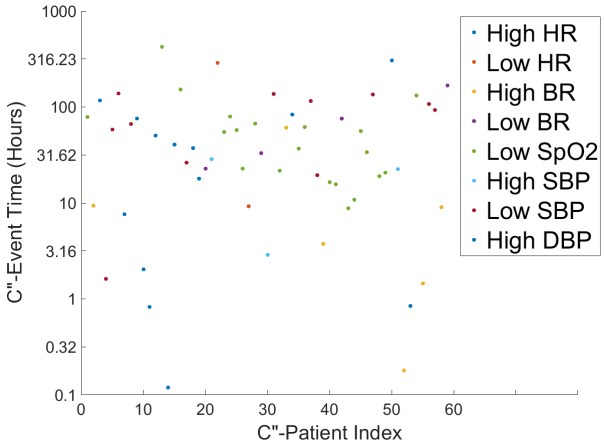


Figure 3.2: First C"-event by primary cause and time-stamp for the 59 C" patients. Event time is presented in logarithmic scale from the time of admission. Each C"-event is coloured according to primary event-cause. The earliest emergency event (patient 14) occurs within minutes of admission, whereas the latest occurs (patient 2) after weeks on ward. No clear pattern is visible to suggest a relationship between the C"-event's cause and time of event.

2. Length of monitoring (LOM): The total duration of (possibly non-contiguous) time that a patient's vital-signs were recorded.

LOS is an informative clinical indication, since a patient will remain on ward until their status progresses (either to a lower- or higher-acuity ward) In contrast LOM indicates data availability, which may be agnostic to clinical condition (unless the patient's condition interferes with vial-sign recording). While LOS is directly available from the patient record, LOM must be calculated with reference to the challenges of missingness and misalignment described above.

LOS and LOM are summarised in several tables and figures. In Figure 3.3, patient LOS is stratified by C"-patient-status. Unsurprisingly, C"-patients were much more likely to have long LOS, presumably because their physiology was identified as abnormal by clinical staff who delayed their discharge from ward. In Figure 3.4 the LOM for each vital-sign is plotted, with patients stratified by C"-patient-status. Since C"-patients were on ward longer, LOM was longer as well. In Table 3.4, the number of patients with LOM < 1 for any vital-sign is tabulated. Missingness was significantly more common among non-C"-patients.

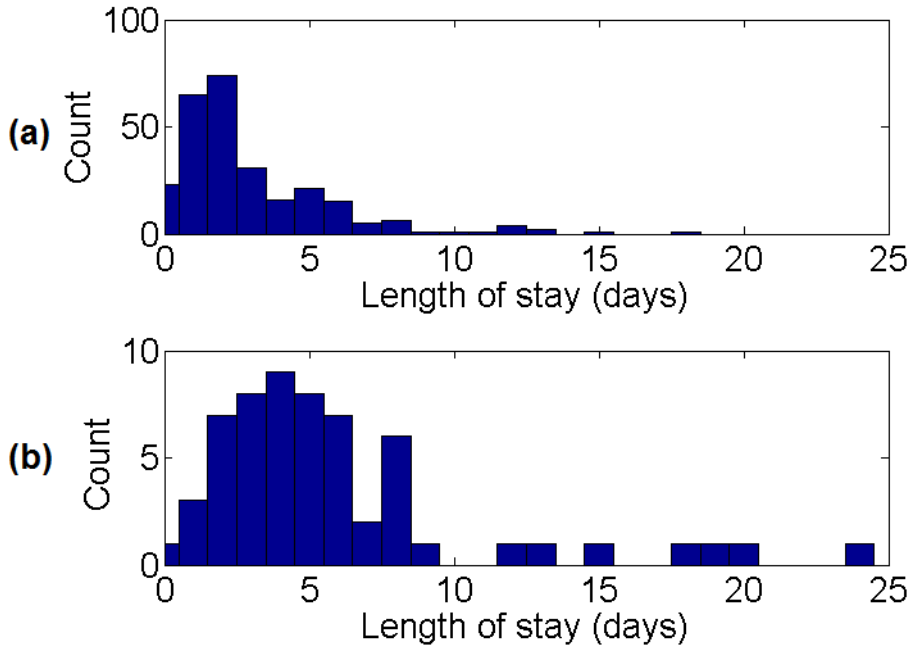


Figure 3.3: Length of stay in (a) non-C"-patients, and (b) C"-patients

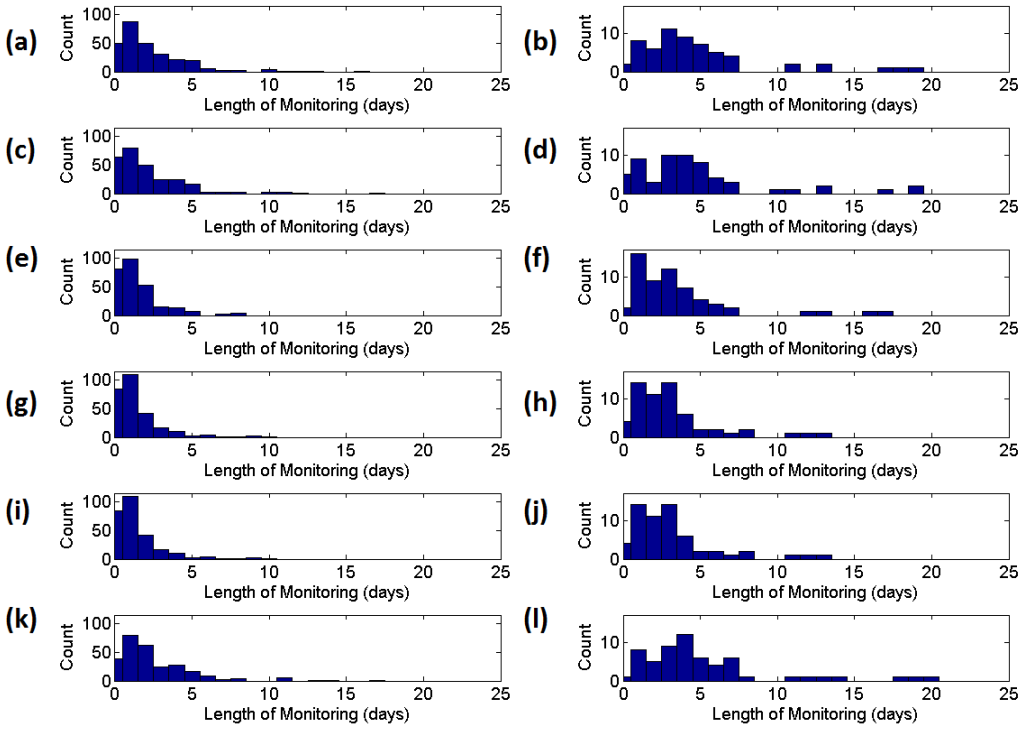


Figure 3.4: Length of monitoring for each vital-sign and across all vital-signs in non-C"-patients (left column) and C"-patients (right column). Vital-signs are ordered as (a-b) HR, (c-d) RR, (e-f) SpO_2, (g-h) SBP, (i-j) DBP, and (k-l) across all vital-signs.

	HR	BR	SpO_2	SBP	DBP	Every Vital
Non-C"-Patients (n=273)	18	33	27	16	16	14
C"-Patients (n=59)	0	3	0	0	0	0

Table 3.4: Counts of patients with less-than 1 hour of data for a particular vital-sign

From Figure 3.3, there were 11 patients whose LOS was less than one hour. From Table 3.4, there were 14 patients whose LOM across all vital-signs was less than 1 hour. In addition to these patients, many remaining patients had an individual vital-sign with less than one hour of recorded data, as seen in Table 3.4. Precise comparison of the rate of missingness is difficult, since there are 59 C"-patients and $333 - 59 - 11 = 263$ non-C"-patients. However, the preponderance of missing vital-sign channels is among the non-C"-patients, which suggests the obvious fact that emergency events can only be observed in the presence of data. It is interesting that, with the exception of three patients (each with less than 1 hour of RR data), all other C"-patients had data available for all five vital-signs.

A further component of missingness is intervals of missing data within a time-series. This is seen in Figure 3.1(a) around hour 50. It is possible, for example, that a non-C"-patient with many hours of data may, in fact, simply be missing data during a period in which abnormal vital-signs occurred. Similarly, if a C"-patient had only a single recorded vital-sign available during the C"-event, then the true "primary cause" vital-sign may have been missed in favour of the available vital-sign. (Table 3.4 does not refute this possibility, since it does not clarify when measurements were available, only that at least an hour of measurements was available at some point.) While, these are all possibilities, these aspects of missingness were not investigated.

As seen in Figure 3.3, C"-patients had longer lengths of stay compared to non-C"-patients. Similarly, from figure 3.4, C"-patients had longer periods of continuous monitoring for each vital-sign 3.4(a-j) and across all vital-signs 3.4(k-l). This could be evidence of the clinician's apprehension over releasing C"-patients.

3.6 Distribution of Continuous Vital-Sign Measurements

As seen in Figure 3.5, a quick overview of univariate vital-sign distributions struggles to usefully differentiate between the average vital-sign measurements of C"-patients and non-C"-patients. This may be partly due to the fact that the preponderance of C"-patients' vital-signs fall within normal ranges, and therefore only differentiate in tail events.

In Figure 3.5, C"-patients appear to have, on aggregate, heavier right-hand tails in HR, BR, and SBP than did non-C"-patients, and a heavier left-hand tail in SpO_2. The left-hand tail of HR, BR, and SBP is (slightly) heavier in HR, BR, and SBP. It is clear that the C"-patients have, on aggregate, more extreme values, as would be expected. It is unclear, however, whether these differences are sufficient to differentiate health status in a timely manner. For example, a C"-patient could have long periods of "normal" physiology prior to rapid onset of the emergency event. Furthermore, extreme valued-measurements within non-C"-patients may occur with sufficient frequency to produce alarms at a similar rate as C" patients.

This view is corroborated by Figure 3.6, which shows the 25th, 50th, and 75th percentile of measurements for each patient. Patients are stratified by C"-patient-status. In both non-C"-patients in Figure 3.6 (top row) and C" patients in Figure 3.6 (bottom row), individual patients can be seen to occupy distinct ranges of measurements, but almost none are near the MET calling range of Table 3.2. Furthermore, the inter-patient range of C" and non-C"-patients are nearly identical. The only notable exception to this is RR, for which C"-patients indexed 50-60 have elevated RR ranges beyond any seen among the non-C"-patient cohort.

3.7 Conclusion

The UPMC data set has been described. Several important aspects of the UPMC data set are (i) the clinical annotation of C"-events, which warranted emergency

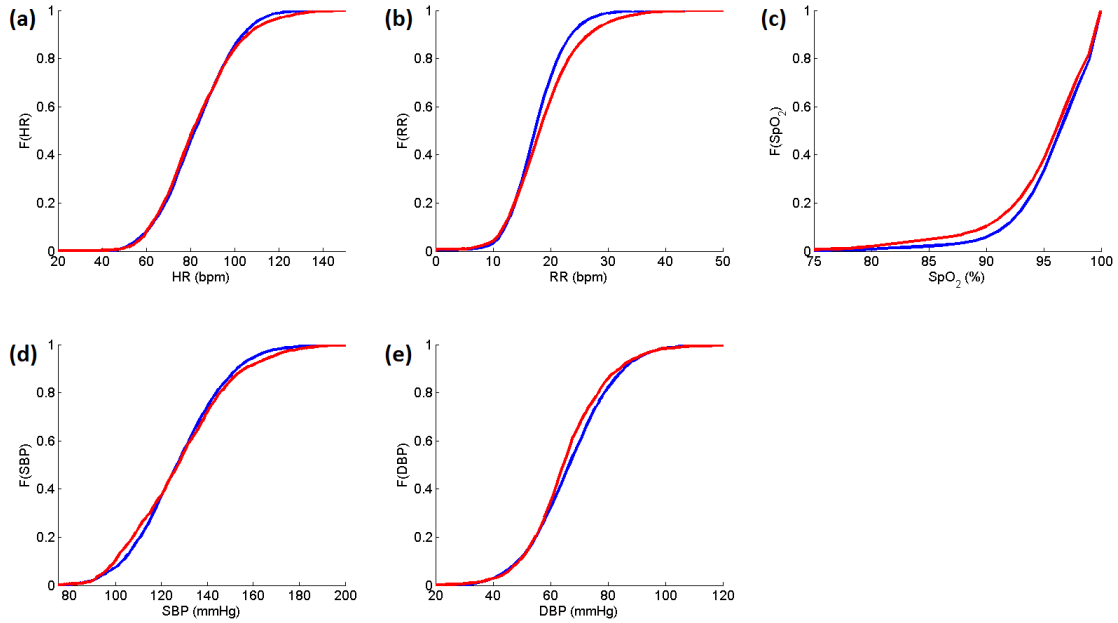


Figure 3.5: Univariate CDF comparison of vital-signs between C" patients (-) and non-C"-patients (-). Differences in tail density can be compared for (a) HR, (b) RR, (c) SpO_2, (d) SBP, and (e) DBP.

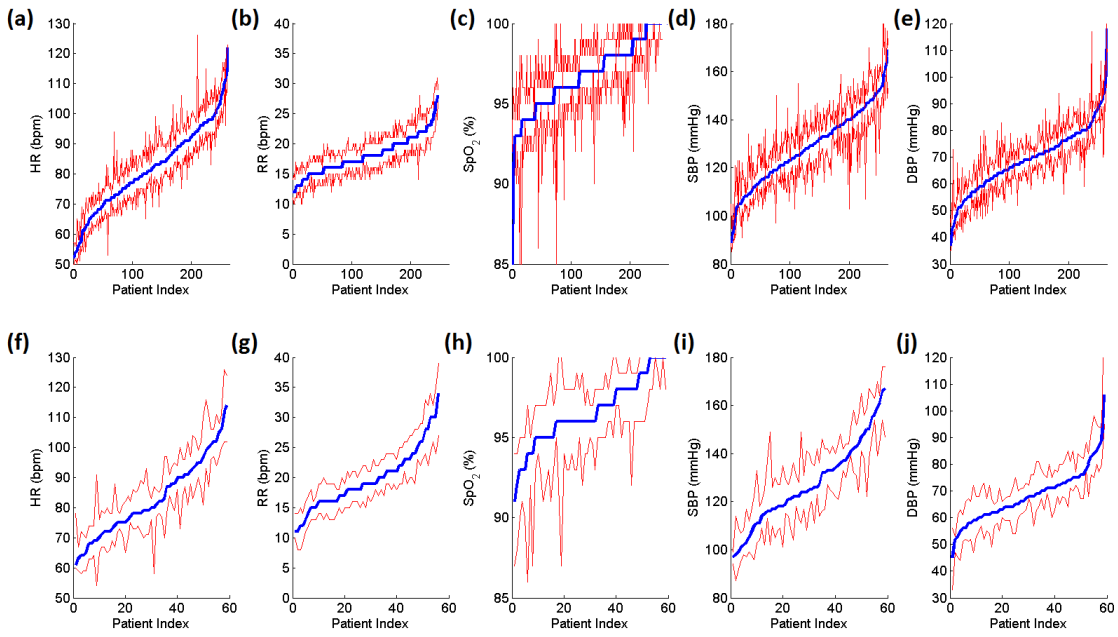


Figure 3.6: Intra-patient vital-sign variability of non-C"-patients (top row) and C" patients (bottom row). The vital-signs are ordered by (a,f) HR, (b,g) RR, (c,h) SpO_2, (d,i) SBP, (e,j) DBP. The blue line shows the median measurement for each patient. Red lines show the 25th and 75th percentile for each patient. Patients in each subplot are indexed by median measurement, for visual clarity.

medical intervention, (ii) the high inter- and intra-patient variability in vital-sign values, and (iii) the confounding caused by vital-sign missingness.

The identification of 112 validated clinical emergency events when only 7 MET calls were made in practice motivates the development of continuous monitoring techniques so that fewer clinical emergencies are missed by clinical staff. The methods developed in this thesis aim to demonstrate (retrospectively) that these emergency events could have been identified in advance without inundating staff with a large number of false-positive alarms.

The high inter- and intra-patient variability in vital-sign values demonstrates the challenge of early warning systems that alarm only when a patient exhibits physiology that is extreme with respect to a patient population, instead of extreme with respect to their own physiology. The methods developed in this thesis will aim to (i) learn, and (ii) exploit personalised physiology information to supplement current monitoring systems.

Finally, there are inherent technical challenges to continuous vital-sign monitoring, particularly those caused by missing vital-sign measurements. Missingness may include complete missingness of a vital-sign record, or a vital-sign not recording within a specific time-period. Ideally, these data would be available for clinical inference. However, the ability to handle missing data is necessary for any automated monitoring system that could be applied realistically to patient monitoring.

4
Methods Review

This chapter describes the fundamental technical elements that will be applied in this thesis. We begin by describing how the univariate Gaussian distribution may be extended to the multivariate Gaussian (MVN) distribution, and then to an infinite-variate Gaussian Process (GP). The GP's probabilistic model is then related to an equivalent state-space model, which is equivalent to a Kalman filter. We then describe Bayesian optimisation, which uses a GP model to select sequential queries to an objective function that we wish to optimise.

Finally, the Parzen kernel density estimate (KDE) is described, which forms the basis of (i) the current state-of-the-art in patient monitoring (discussed in Chapters 7 and 8), and (ii) a candidate method for modelling the marginal distribution of vital-signs, e.g., for artefact detection (used in Chapter 5) and copula modelling (described in future work).

4.1 Introductory Reading

The methods described in this chapter are popular in many engineering applications and therefore are accessible through a rich body of introductory material, which the reader may find helpful. Particular examples of useful introductory material are provided below.

David MacKay's "The Humble Gaussian Distribution" [127] provides the best concise description of the properties of Gaussian distributions that are most pertinent to GPs. Mark Ebden's "Gaussian Processes: A Quick Introduction" [128] provides a concise introduction to GPs for regression and classification (for those unfamiliar with GPs). An overview of Gaussian process covariance functions can be found in David Duvenaud's "Kernel Cookbook" [129], with a more in-depth coverage throughout multiple chapters of his thesis [118]. State-space representations for Gaussian processes (as used in this thesis) are described in [130], more general coverage of the relation between Gaussian Processes and the Kalman filter can be found in [131, 132].

An overview of applications of GPs to optimisation can be found in Shahriari et al [133].

The Parzen kernel density estimate (KDE) implemented by Alistair Hann in [45] is covered extensively in the Chapter 7 of this thesis. The original publications by Rosenblatt [44] and Parzen [43] remain useful introductions.

4.2 Units in probability

Readers familiar with the units inherent in various statistical entities (e.g., mean, variance, probability density functions) may wish to skip this section.

Different vital-signs tend to be measured in different units. The statistical entities that describe these vital-signs are seldom unitless, but instead change according to

- the random variable under consideration (e.g., which vital-sign), and
- the statistic of interest (e.g., mean, variance, etc.).

For brevity, we will stick to the case of continuous variables, as these (i) may be less intuitive to the unfamiliar reader than discrete random variable and (ii) constitute the majority of probabilistic models used in this thesis.

4.2.1 Mean

When random variable Y is measured in units "u", then operations on Y will have units as well. For example, let Y be governed by probability density function (pdf) $p(y)$, then $p(y)dy$, roughly speaking, is the probability that Y falls within the interval of $(y, y + dy)$. Since a probability is unitless, then $p(y)dy$ is unitless, and dy has units of $\frac{1}{u}$. The cumulative density function (cdf), $F(y)$ outputs a probability and therefore is unitless as well.

The expected value, or mean, of Y,

$$\mathrm{E}[Y] = \int_{-\infty}^{\infty} y\, p(y)\, dy \tag{4.1}$$

has units of "$u \times u \times \frac{1}{u}$", which is units of u. For example, given a random sample of heart rates, measured in beats-per-minute (bpm), the mean of that sample would be bpm as well.

4.2.2 Variance

The variance of Y,

$$\mathrm{Var}[Y] = \int_{-\infty}^{\infty} (y - \mathrm{E}[Y])^2\, p(y)\, dy = \mathrm{E}[Y^2] - \mathrm{E}(Y)^2 \tag{4.2}$$

has units of "$u \times u - u \times u$", which is units of u^2. For example, given a random sample of heart rates measurements, measured in beats-per-minute (bpm), the variance of that sample would be $(\mathrm{bpm})^2$, which is beat-squared-per-minute-squared. Note that $\mathrm{Var}[Y] = 0$ when Y is a constant and $\mathrm{Var}[Y] > 0$ otherwise.

4.2.3 Covariance

The covariance between two random variables, Y_i and Y_j, measured in units of u_i and u_j, respectively, is

$$\mathrm{Cov}[Y_i, Y_j] = \mathrm{E}[(Y_i - \mu_i)(Y_j - \mu_j)] = \mathrm{E}[Y_i Y_j] - \mu_i \mu_j, \tag{4.3}$$

which is measured in units of $u_i \times u_j - u_i \times u_j$, which is units of $u_i u_j$. For example, given a random sample of heart rates measurements, measured in beats-per-minute (bpm), and a (paired) random sample of blood pressure measurements, measured in mmHg, the covariance of the heart rates measurements and blood pressure measurements would be measured in units of "bpm × mmHg". Note that $\text{Cov}[Y_i, Y_j]$ may be greater or less than 0, and is equal to 0 only if Y_i and Y_j are independent.

Equations 4.1, 4.2, and 4.3 will form the basis of our inference via Gaussian-distributed random variables, described in the next section.

4.3 Gaussian Processes

In this section we will extend the univariate Gaussian to the multivariate Gaussian, and, finally, to the infinite-variate Gaussian process.

4.3.1 Univariate Gaussian Distribution

The univariate Gaussian random variable, $Y \sim N(\mu, \sigma)$ is parameterised by mean $\mu = \text{E}[Y]$ and variance $\sigma^2 = \text{Var}[Y]$. A common alternative parameterisation of variance is with precision σ^{-2}. The probability density of Y is

$$p(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{4.4}$$

with cumulative density

$$\Phi(y|\mu, \sigma) = Pr(y \leq a|\mu, \sigma) = \int_{-\infty}^{a} p(y|\mu, \sigma)dy = \frac{1}{2}\left[1 + \text{erf}\left(\frac{y-\mu}{\sigma\sqrt{2}}\right)\right] \tag{4.5}$$

where $\text{erf}(x)$ is the error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{4.6}$$

Further useful properties include the survival function

$$S(y|\mu, \sigma) = Pr(y \leq a|\mu, \sigma) = 1 - \Phi(y|\mu, \sigma), \tag{4.7}$$

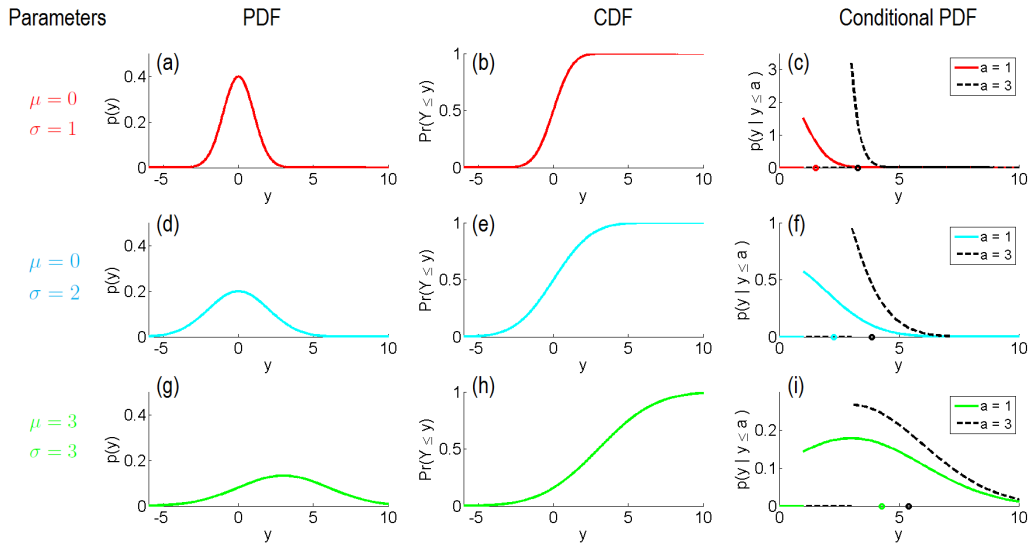


Figure 4.1: Properties of the univariate normal distribution. The PDF, CDF, and conditional distribution are shown for three univariate normal distributions, with different parameterisations. In (c,f,i), showing $p(y|y \geq a)$ plots are shown for $a = 1$ and $a = 3$. Note that the discontinuity in $p(y|y \geq a)$ is induced by $p(y|y \geq a) = 0$ when $y < a$. The conditional mean is marked on the x-axis. These later figures demonstrate how the normal distribution facilitates interesting probabilistic reasoning, which will be applied in the section on Bayesian optimisation.

and the conditional distribution of Y *given* that it exceeds a pre-defined threshold a.

$$p(Y|y \geq a, \mu, \sigma) = \frac{p(y, y \geq a)}{Pr(y \geq a)} = \begin{cases} 0, & \text{if } y < a \\ (1 - \Phi(a|\mu, \sigma))^{-1} \, p(y|\mu, \sigma), & \text{otherwise.} \end{cases} \tag{4.8}$$

This conditional distribution then yields a conditional expectation $\mathrm{E}[Y|y \geq a, \mu, \sigma]$ that is greater than the unconditional expectation, $\mathrm{E}[Y] = \mu$.

These properties are illustrated in Figure 4.1. Roughly speaking, the parameter μ regulates the location of Y, and σ regulates the symmetric spread around μ. Since pdfs must sum to 1, the conditional distribution (in Equation 4.8) is proportional to the unconditional distribution (in Equation 4.4) at all points greater than a. However, the conditional expectation converges towards a as a increases due to the exponentiated quadratic decrease in the tails. Conversely, as a decreases, the conditional expectation converges to μ since the conditional distribution converges towards the unconditional distribution.

4.3.2 Multivariate Gaussian Distribution

The univariate Gaussian, $Y \sim N(\mu, \sigma)$, may be extended to a random vector, $\mathbf{Y} \sim \text{MVN}(\mathbf{m}, \mathbf{K})$. Where

- $\mathbf{Y}$ is an $n \times 1$ random vector, where each element of $\mathbf{Y} = [y_1, ..., y_n]$ is a univariate normal random variable.
- $\mathbf{m}$ is an $n \times 1$ vector of the marginal means. That is, $\mathbf{m}_i = \text{E}[\mathbf{y}_i]$.
- $\mathbf{K}$ is an $n \times n$ positive-definite covariance matrix, with $\mathbf{K}_{i,j} = \text{Cov}(y_i, y_j)$, so that the marginal variances populate the main diagonal and pairwise covariances populate off-diagonal elements.

The pdf of the MVN is

$$p(\mathbf{y}|\mathbf{m}, \mathbf{K}) = (2\pi|\mathbf{K}|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{m})\mathbf{K}^{-1}(\mathbf{y} - \mathbf{m})\right). \tag{4.9}$$

The MVN is the most common model for the joint distribution of univariate Gaussian variables. Linearly correlated Gaussian distributions are jointly MVN. If any linear combination of a set of univariate Gaussian distributions is also Gaussian, then those Gaussian distributions are jointly MVN. However, the MVN is not the only model for correlated variables that are marginally Gaussian. For example, copula models can accommodate marginally normal random variables that are not jointly multivariate normal.

The MVN's cumulative density function, $F(\mathbf{Y}) = Pr(\mathbf{Y}_1 \leq \mathbf{y}_1, ..., \mathbf{Y}_n \leq \mathbf{y}_n)$, has no analytical/closed-form solution. However, numerous numerical methods exist to calculate this probability, along with alternative definitions of multivariate cumulative density (which may or may not have analytical solutions).

As illustrated by the 2-D example in Figure 4.2, the values of $\mathbf{m}$ determine the central coordinates of the multivariate Gaussian, while the diagonal elements of $\mathbf{K}$ determine spread in the direction along the axis in each dimension. The off-diagonal values of K determine the spread along and angle of off-axis variance. In 4.2(a,c) the eigenvectors and eigenvalues of K form the principal directions of variation, and

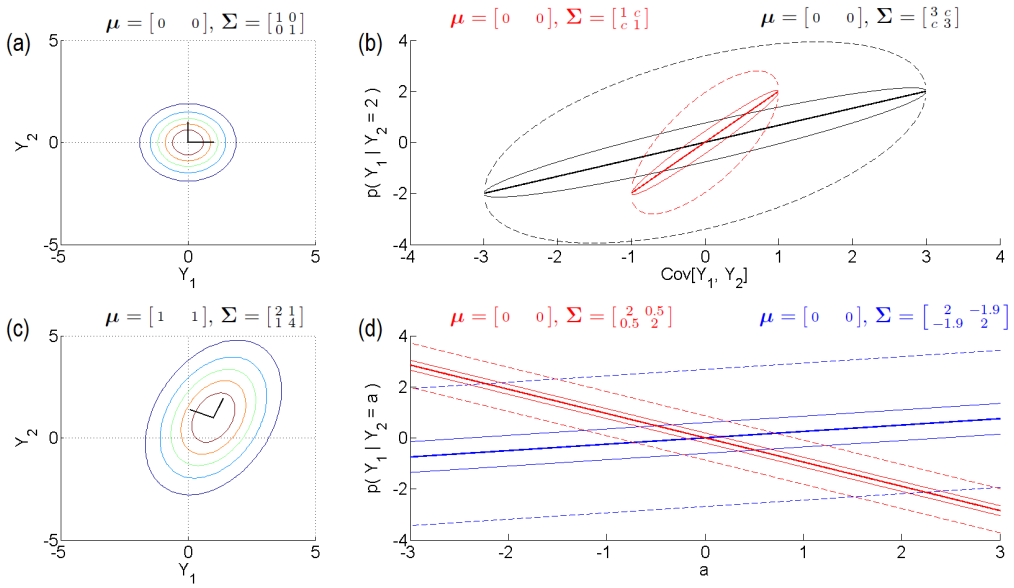


Figure 4.2: Properties of the multivariate normal distribution. MVN parameterisations are shown above each plot. (a) and (c) show contours of the joint PDF of two different MVNs. The eigen vectors of each $\boldsymbol{\Sigma}$ are shown in black, centred at $\boldsymbol{\mu}$. To illustrate the change in conditional expectation, (b) shows the change in $p(Y_1|Y_2 = 2)$ as a function of $c = \text{Cov}[Y_1, Y_2]$, i.e. the off-diagonal element of $\boldsymbol{\Sigma}$ and (d) shows the change in $p(Y_1|Y_2 = a)$ as a function of a for both a negatively (red) and positively (blue) correlated MVN. In (b) and (d) the mean (-), 50% (-), and 95% (- -) quantiles are shown.

the magnitude of the variation, respectively. As pointed out by MacKay [127], the diagonal and off-diagonal elements of the covariance matrix will not necessarily share the same units, and therefore, linear combinations of these different-unit entities may be questionable in interpretation. As per the rules of matrix multiplication, eigen values and eigen vectors are not invariant to unit transformations.

When $\mathbf{Y}_i$ and $\mathbf{Y}_j$ are correlated, our uncertainty in $\mathbf{Y}_i$, conditional on observing $\mathbf{Y}_j$ is now

$$p\left(\mathbf{Y}_i|\mathbf{Y}_j = a\right) \sim N(\,\mathbf{m}_i + \mathbf{K}_{i,j}\mathbf{K}_{j,j}^{-1}(a - \mathbf{m}_j)\,,\, \mathbf{K}_{i,i} - \mathbf{K}_{i,j}\mathbf{K}_{j,j}^{-1}\mathbf{K}_{j,i}\,) \tag{4.10}$$

Intuitively, this is sensible: as our *a priori* uncertainty in $\mathbf{Y}_i$ is equal to $\mathbf{Y}_i$'s marginal variance $\sigma_{i,i}^2 = \mathbf{K}_{i,i}$. From this baseline uncertainty, we may subtract a portion of the uncertainty that is proportional to the known $\mathbf{Y}_j$'s precision, $\mathbf{K}_{j,j}^{-1}$, and the covariance between the known and unknown variables, $\mathbf{K}_{j,i}$. Note that as correlation goes to zero, the uncertainty removed by observing Y_j decreases to zero,

until we are only left with the original marginal variance σ_i^2. This is illustrated for two different parameterisations of the MVN in Figure 4.2(b). At $\text{Cov}[Y_1, Y_2] = 0$, then $p(Y_1) = p(Y_1|Y_2)$. However as $\text{Cov}[Y_1, Y_2]$ approaches perfect linear correlation, we become certain in the values of Y_1, since it will be a linear function of Y_2.

Along similar reasoning, our *a priori* belief in the mean of $\mathbf{Y}_i$ is equal to the marginal mean of $\mathbf{Y}_i$, $\mathbf{m}_i$. Recalling that the covariance of two variables is the expected product of the deviation from their respective means we may update our expectation of $\mathbf{Y}_i$, having seen how much $\mathbf{Y}_j$ has deviated from its mean. The magnitude and directionality from which $\mathbf{Y}_i$ is expected to deviate from $\mathbf{m}_i$, is augmented according to $(a - \mathbf{m}_j)$, which is the magnitude and directionality from which $\mathbf{Y}_j$ *has* deviated from $\mathbf{m}_j$. The magnitude of this augmentation is proportional to $\mathbf{Y}_j$'s precision, $\mathbf{K}_{j,j}^{-1}$, and the linear covariance of $\mathbf{Y}_i$ and $\mathbf{Y}_j$. As seen in Figure 4.2, this is a linear relationship, both as a function of covariance (in 4.2(b)) and as a function of a (in 4.2(d)).

This relation between the conditional distribution of observed and unobserved correlated Gaussian distributions is fundamental to the intuition of the Gaussian process, in which we will model our uncertainty of an infinite number of unobserved points, conditional on a finite number of observed points. Our estimation of the prior mean and covariance will inform the uncertainty in Y at unseen points. At points near our observed values (where covariance is presumably high) we will be more certain as in the poles of the ellipse in Figure 4.2(b). Alternatively, far from the known values, at the centre of Figure 4.2(b), our prior knowledge dominates.

4.3.3 Gaussian Processes

Gaussian Process Model

The GP extends the multivariate Gaussian (of pre-defined dimensionality n) to an infinite-dimensional stochastic process. We define the GP to be a stochastic process for which any finite subset of points, along a domain t, follows an MVN distribution. This MVN will, as before, have both a mean vector, $\mathbf{m}$, and covariance matrix, $\mathbf{K}$ to describe any observed data points.

To populate the elements $\mathbf{m}$ and $\mathbf{K}$ for *any* finite subset (conditional on specifying all t), we replace the mean *vector*, $\mathbf{m}$, with a mean *function* $\mu(t)$. That is, the mean of the Gaussian at point t is $\mu(t)$.

Similarly, the covariance of any two points y_i and y_j, located at points t_i and t_j, is defined by covariance *function* $k(t_i, t_j) = \text{Cov}[y_i, y_j]$. This function $k(t_i, t_j)$, which will be described in detail below, is positive semi-definite and typically decreases as t_i and t_j separate in distance. Note that whereas covariance (as in equation 4.3) takes random variables as arguments, covariance function k takes the *location* of those random variables as arguments.

We return to our initial definition of a GP as "a stochastic process for which any finite subset of points, along a domain t, follow a multivariate Gaussian". Given this specification of $Y(t) \sim \text{GP}(\mu(t), k(t, t'))$, then for any finite vector $\mathbf{t} = [t_1, ..., t_n]$ we have a random vector $\mathbf{Y}(\mathbf{t}) = [y_{t_i}, ..., y_{t_n}]$, with the now-familiar mean vector

$$\mathbf{m} = \text{E}[\mathbf{Y}(\mathbf{t})] = \text{E}[y_{t_i}, ..., y_{t_n}] = [\mu(t_i), ..., \mu(t_n)] = \mu(\mathbf{t}).$$

The covariance matrix is

$$\mathbf{K} = \begin{bmatrix} \text{Cov}[y_1, y_1] & \cdots & \text{Cov}[y_1, y_n] \\ \vdots & \ddots & \vdots \\ \text{Cov}[y_n, y_1] & \cdots & \text{Cov}[y_n, y_n] \end{bmatrix} = \begin{bmatrix} k(t_1, t_1) & \cdots & k(t_1, t_n) \\ \vdots & \ddots & \vdots \\ k(t_n, t_1) & \cdots & k(t_n, t_n) \end{bmatrix}.$$

While GPs do not necessarily specify a functional form over $y(t)$, *a priori*, functional characteristics may be made implicit via the *a priori*-specified mean and covariance functions, which, typically, are parametric. A judicious choice in $\mu(t)$ and $k(t, t')$, and the inference over their respective parameterisations $\theta_{\mu(t)}$ and $\theta_{k(t,t')}$ will be discussed below.

Gaussian Process Mean Functions

The prior mean function $\mu(t)$ specifies the expected value of $y(t)$ in the absence of further information. An absence of information may occur when either no data are observed or when the $y(t)$ under consideration is so far from the observed points (in terms of t) that the covariance between $y(t)$ and the observed points

is (approximately) 0. In these instances, the GP's expectation of $y(t)$ is equal to the prior mean. We relate this back to Figure 4.2(b) where the expectation of Y_1 is the same, whether or not Y_2 is observed. (The variance around this mean will be described in the next section.)

The mean function may may have a parametric or non-parametric functional form. Popular choices of mean function include constant functions (e.g., $\mu(t) = c$, where c is (i) the mean of training data, (ii) a particular value from the observed data, or (iii) a known constant). Mean functions may also be derived from prior data (e.g., from a related data set), and functions derived from physical models (e.g., an exponential function which decays to room temperature when modelling a cooling process). While the prior mean function has little effect on the posterior mean near points at which $y(t)$ has been observed (with some accommodation for high-noise or low length-scale parameterisations of the covariance function, described later), it has a strong effect in the absence of observations (where the posterior returns to the prior mean). When used correctly, the prior mean can provide reasonable descriptions of our uncertainty in sparse-data scenarios.

For many applications, the prior mean function is fixed to be $\mu(t) := 0$, once all observed data have been mean-centered via the transformation $\mathbf{Y} := \mathbf{Y} - \frac{1}{n}\mathbf{Y}^T\mathbf{1}$. This prior mean is equivalent to specifying a constant prior mean function $\mu(t) = c$, where $c := 0$, and therefore not estimated via statistical inference. For ease of explanation, and without loss of generality, $\mu(t) := 0$ will be used unless otherwise noted. However, the parameters of $\mu(t)$ may be estimated (via the likelihood function) simultaneously to the parameters of $k(t, t')$, as described in the inference section.

Gaussian Process Covariance Functions

Selection of the covariance function $k(t, t')$, typically receives the lion's share of attention for GP model selection. This may be sensible, given that, while the prior mean may be "washed out" where data is observed, the implicit effect of the covariance function will always influence the posterior estimate of $y(t)$.

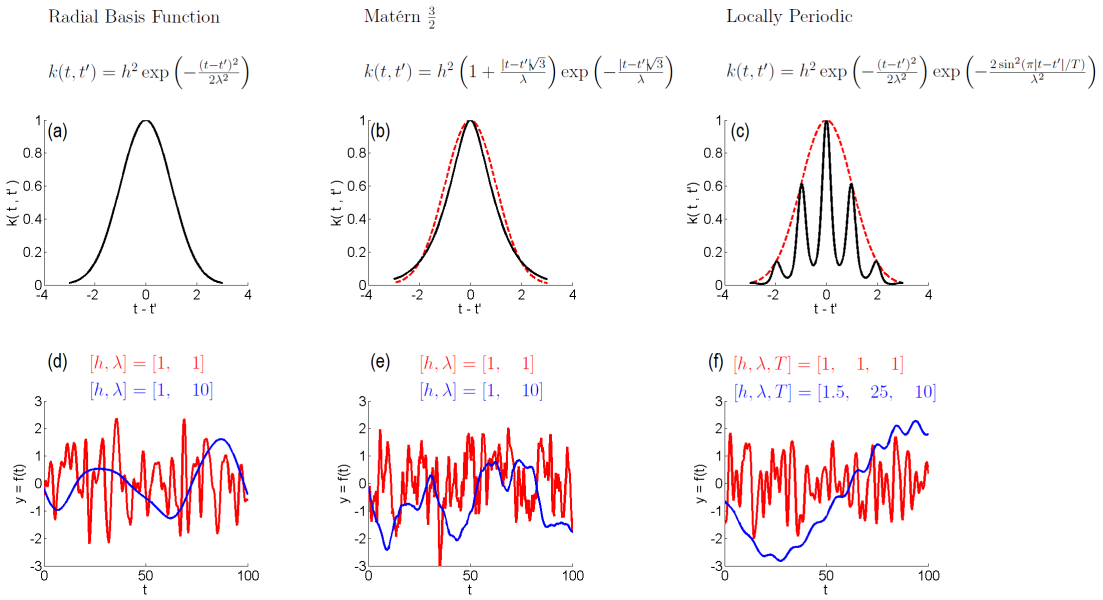


Figure 4.3: Three simple covariance kernels: the RBF, Matérn $\frac{3}{2}$, and local periodic kernel. For each covariance kernel, the mathematical function is given, along with a plot of that function in (a,b,c). In (b,c) the RBF kernel (- -) is plotted for reference. In (d,e,f) a random draw is shown for each kernel under two different parameterisations. The parameterisations are provided above each plot.

For example, in Figure 4.3(d-f), we may see how the length scale hyperparameter, λ, regulates how rapidly the function may change for each of the 3 different covariance functions. The shorter length scales (in red) correspond to more rapidly changing functions. Similarly, the output scale h regulates the magnitude of deviation from the mean. As seen in 4.3(d-e), both draws occupy approximately the same range across the y-axis (since their output scales are identical for both parameterisations). In contrast, increasing the output scale from 1 (red) to 1.5 (blue) in 4.3(f) results in a function that spans a greater range along the y-axis.

The differentiability of the covariance function regulates the smoothness of function $y(t)$. For example, samples from a GP with an RBF covariance function are infinitely differentiable. Although the RBF is a popular modelling choice, such smoothness may or may not be an appropriate representation of less-smooth functions. The Matérn(p) family of covariance functions (where $v = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, ...$) is $p + \frac{1}{2} - 1$ times differentiable. So samples from a Matérn $\frac{3}{2}$ kernel are once-differentiable, and samples from a Matérn $\frac{5}{2}$ kernel are twice-differentiable. The

Matérn family converges to the RBF kernel as p approaches ∞.

The Matérn $\frac{5}{2}$

$$k(t,t') = h^2\left(1 + \frac{|\mathbf{t}-\mathbf{t}'|\sqrt{5}}{\lambda} + \frac{5|\mathbf{t}-\mathbf{t}'|^2}{3\lambda^2}\right)\exp\left(-\frac{|\mathbf{t}-\mathbf{t}'|\sqrt{5}}{\lambda}\right) \tag{4.11}$$

will be used extensively in vital-sign modelling. As a twice-differentiable function, the Matérn $\frac{5}{2}$ provides another option to model un-smooth functions, such as measurements of heart rate.

For GPs with multidimensional inputs, it may be desirable to model all covariance via a single kernel, by allowing the length-scales to vary by input dimension, via an Automatic Relevance Determination (ARD) kernel. The ARD kernel replaces the scalar-valued length scale λ with a vector-valued length scale ν, with a value for every input dimension. The ARD can incorporate a variety of kernels, for example, the Matérn 3/2 ARD kernel

$$\begin{aligned} &k(\boldsymbol{t},\boldsymbol{t}') = h^2\Big(1+\sqrt{3r}\Big)\exp\Big(-\sqrt{3r}\Big), \\ &\text{s.t. } r = \sum_{d=1}^{D}\frac{(\boldsymbol{t}_d-\boldsymbol{t}'_d)^2}{\nu_d}. \end{aligned} \tag{4.12}$$

where the distance between $\boldsymbol{t}$ and $\boldsymbol{t}'$, r, is now regulated by different a length-scale, ν_d in each dimension d of $\boldsymbol{t}$. Variable r is $|\boldsymbol{t}-\boldsymbol{t}'|$. The ARD kernel is of particular interest later in Chapters 4 and in Chapter 6, where it is used for Bayesian optimisation of functions that vary differently according to the input dimension under consideration.

As shown in Figure 4.5, kernels may be combined to incorporate further complexity. The most common additive component is the white noise kernel[1]

$$k(t,t') = \sigma_n^2\delta(t,t') = \begin{cases}\sigma^2, & \text{if } t \text{ is } t' \\ 0, & \text{otherwise.}\end{cases} \tag{4.13}$$

By requiring that t and t' not only be equal in value, but identical indices the GP has codified that the function has some amount of irreducible noise error, and therefore the function has uncertainty, even at points along the domain that have

[1]By convention, we denote the noise-variance parameter to be σ_n where the subscript n denotes "noise", and is unrelated to the typical use of n to denote the total number of data points.

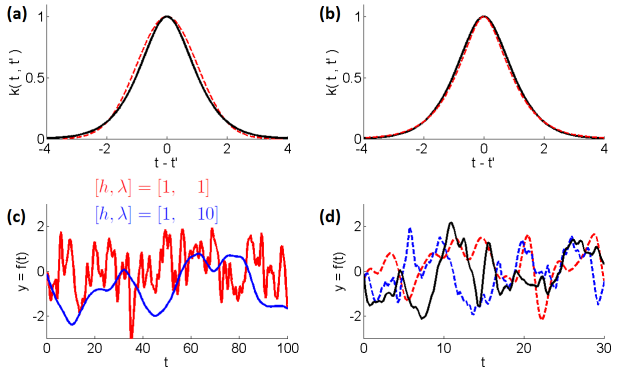


Figure 4.4: In (a,b) the Matérn $\frac{5}{2}$ covariance kernel function (-) is plotted. The (a) RBF (- -) and (b) Matérn $\frac{3}{2}$ (- -) are provided for reference. Compared to the RBF, the Matérn $\frac{5}{2}$ covariance function decreases more rapidly near 0, but maintains a higher covariance in the tails. In (c) a random draw is shown for the Matérn $\frac{5}{2}$ under two different parameterisations. In (d) a draw from the Matérn $\frac{5}{2}$ (-), RBF (- -), and Matérn $\frac{3}{2}$ (- -) is shown, with each draw using identical values for hyperparameters h and λ. Although the functions' hyperparameters are identical, each exhibits a distinct level of smoothness.

already been observed. By excluding a noise kernel, the kernel has codified that any observed point is known with perfect certainty. This latter property is what creates the classic "sausage-link" GP posterior (such as the red GP in Figure 4.5(c)), in which uncertainty is zero at observed points and non-zero else where. This is analogous to the certainty in the value of $p(Y_1|Y_2) = a$ in the poles of Figure 4.2(b). In contrast, by including the white noise kernel, the variance of any point is strictly greater than the covariance of any two points. Even when white noise is not desired in the GP model, the noise kernel with σ_n set to a negligible factor of computational ϵ may be used to avoid matrix singularity during the inference routine.

Beside noise-variance, additive components may help capture multiple sources of signal variation. For example, by adding two RBF kernels with different length-scale parameters, we may capture both long length-scale trend and a short length-scale trend, without to need to trade-off the modelling of one trend for the other. This may allow the GP to adapt to short-term volatility while maintaining its capacity for long-term extrapolation. As seen in Figure 4.5, the GP with two additive components may accommodate the short-term volatility around $t = [50, 60]$, but it also forecasts a further increase in $t > 80$, whereas the simple kernel more quickly converges back to the prior mean (analogous to the central point of Figure 4.2(b)).

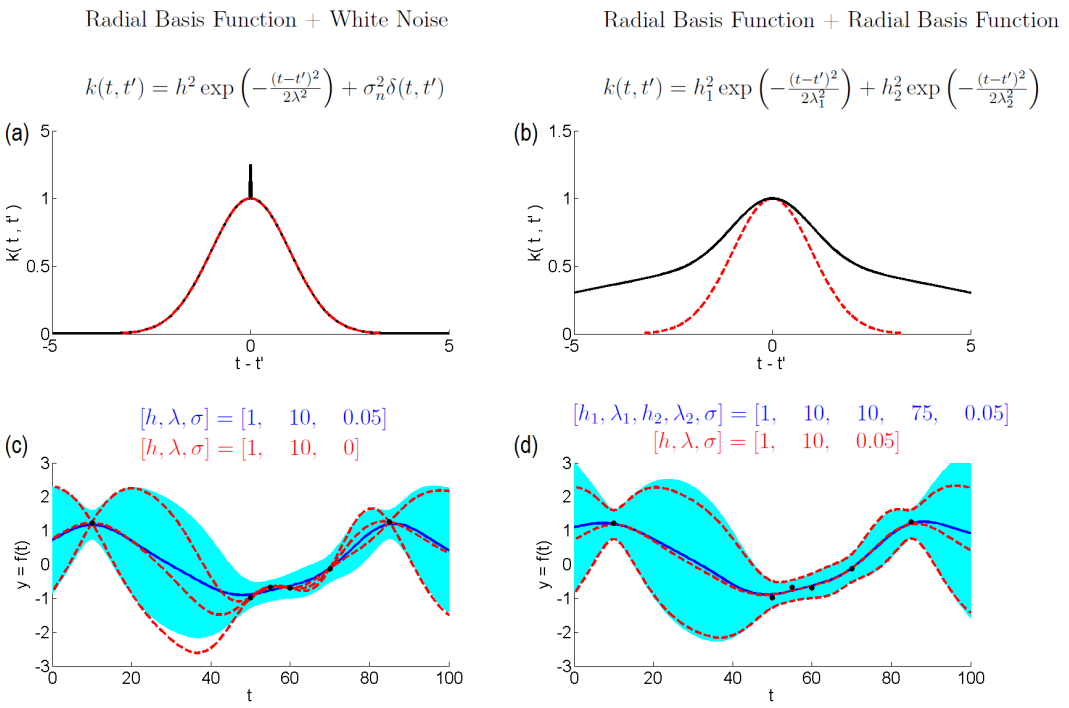


Figure 4.5: Two additive covariance functions: (a) an RBF plus white noise, and (b) an RBF plus another RBF. For reference, a single RBF kernel is in (- -). In (c) and (d) a posterior GP (mean and 95% CI) from each kernel is shown fit to data, along with a reference GP using an RBF kernel in red.

A final property, which is helpful when attempting to reason probabilistically via GPs is the inducement of logical "and"/"or" statements via the kernel. A logical "or" statement may be created via additive kernels, which allows correlation to be high for when it is high for either one *or* the other kernel. This is particularly helpful in multidimensional applications in which only a single dimension has been thoroughly sampled. The logical "and" statement may be induced by multiplication of kernel, requiring correlation to be high only when correlation is high for both of the kernels being multiplied. For example, the locally periodic kernel from Figure 4.3 is a multiplicative kernel. Covariance is only high for Y's that are close in both the RBF component *and* the periodic component in the kernel.

Gaussian Process Hyperparameter Inference

Collating the hyperparameters of the mean and covariance functions into a single vector, $\boldsymbol{\theta}$, we can infer appropriate values of $\boldsymbol{\theta}$ though the posterior log marginal

likelihood (LML)

$$\log p(\mathbf{y}|\theta) = -\frac{1}{2}(\mathbf{y}-\boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{m}) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log(2\pi), \tag{4.14}$$

where $\boldsymbol{\theta}$ influences $\log p(\mathbf{y})$ via the the mean function and covariance function which determine the values of the mean vector, $\boldsymbol{m}$, and the covariance matrix, $\boldsymbol{\Sigma}$.

This crucial inferential step typically either (i) optimises the LML (e.g. via gradient ascent), or (ii) integrates across the LML, e.g. via Markov Chain Monte Carlo (MCMC) [134]. Access to parallel computation can assist MCMC via (i) running multiple parallel MCMC chains, or, alternatively, (ii) parallel processing of the proposals for the likelihood-ratio step between points [135]. Quadrature methods may be used may be used for a more Bayesian integration over $p(\mathbf{y}|\theta)$ to mitigate the short-comings of "fundamentally frequentist" MCMC methods [136]. Approximate methods or GP-based models with non-Gaussian likelihoods typically require recourse to alternative forms of inference.

The primary computational burden of evaluating the LML is the $O(n^3)$ inversion of the covariance matrix, with a memory requirement of $O(n^2)$, where n is the number of observations. This is unfeasible in large data sets or in computationally-contained settings. This is particularly true when performing a large number of evaluations of the LML, as required for the MCMC sampling process used for integration.

State-space representation of Gaussian process regression

There are several ways to reduce the computational burden of GP inference. Computation-saving methods include identifying those covariance matrices which are simpler to invert (e.g. Toeplitz matrices) and Cholesky decomposition of the covariance matrix to update in light of new data [137]. While these methods are extremely useful for particular applications, vital-sign monitoring does not typically lend itself to consistent time-stamps (to take advantage of Toeplitz properties). Vital-sign dynamics themselves may rapidly change (motivating continuous re-estimation), which limits the use of methods requiring identical parameterisations

[137]. Alternative approximate methods to handle large amounts of data may infer the GP via inducing points [125], however these methods require more complex modes of inference [138] that are (i) computationally demanding in their own right and, more importantly, (ii) more difficult to automate for sequential time-series fitting (as is required for the applications in this thesis).

Due to the real-time application of patient monitoring algorithms, it is still desirable to reduce computational burden, where possible. One way to achieve this is by framing the GP as an equivalent state-space model, which requires less-burdensome inference. Due to its Gaussian likelihood, GPs are closely related to, and frequently equivalent to, other least-squares models. (See Sorenson [139] for a intuitive description, written for the casual reader in IEEE Spectrum). This relation may be used advantageously, as equivalent models may require less computational burden to achieve the same model.

Hartikainen et al [130] demonstrate how these requirements can be reduced to $O(m^3 n)$ for covariance inversion and $O(m^2 n)$ for memory, by reformulating $k(\mathbf{x}, \mathbf{x}')$ as an m^{th}-order, scalar, linear time-invariant stochastic differential equation. This computation-saving manipulation does not affect interpretation of the GP models subsequently described.

4.4 Bayesian Optimisation

Bayesian optimisation uses GP inference to identify the global maximum (or minimum) of a black-box function by sequentially selecting queries. Bayesian optimisation requires both a probabilistic inference step, and a decision step, both of which may be imbued with desired levels of complexity. For clarity, Section 4.1 motivates the use of Bayesian optimisation over alternative, more common optimisation procedures. Section 4.2 gives a general overview of the full Bayesian optimisation algorithm. Sections 4.3 and 4.4 provide details about the key components of inference and decision steps, respectively, which enables the algorithm to achieve competitive performance compared to alternative algorithms.

4.4.1 Motivation

A common challenge in optimisation is to identify the global optimum of a black-box (unknown) function that is expensive to evaluate. "Expense" is usually a computational or time expense, however in some applications, the expense could be different, e.g., the monetary expense drilling holes for geographical sampling. Since the function is black-box, direct analytical solutions are not possible (e.g., setting the second derivative to zero, and solving). Convex algorithms are also not possible (e.g., the Simplex methods) due to local optima. Gradient-based methods are furthermore undesirable due added computational burden at each step. By requiring a small number of evaluations to the function, the optimisation protocol must choose between the competing needs to (i) *explore* unknown parts of the function, and (ii) *exploit* further evaluations near known high values.

Bayesian optimisation uses GPs to identify the global optima of functions. Due to the computational expense of fitting a GP, the use of Bayesian Optimisation is typically relegated to applications were the objective function is (i) black-box (thereby hindering the application of analytical optimisation routines), and (ii) expensive to evaluate, which confines the total number of queries to a small budget. Bayesian optimisation takes advantage of many of the benefits of GP modelling, including (i) flexible forms to model the objective function, and (ii) a natural application of probabilistic reasoning to the trade-off between exploration and exploitation. As shown in Figure 4.6, Bayesian optimisation provides a principled conjecture about the global optimum, having only evaluated the function at a small number of points.

4.4.2 Overview of the Bayesian Optimisation Algorithm

We will first present an overview of the Bayesian optimisation algorithm. Pseudo code for the Bayesian optimisation algorithm is presented in Algorithm 1. The text of Algorithm 1 was published in [1] and is © 2017 IEEE. Separate components will be described in greater detail in subsequent sections.

Algorithm 1 Bayesian optimisation algorithm

1: **query** $L(\boldsymbol{s})$ at initial points, $\boldsymbol{s}_{\text{init}}$.
2: $\{L_{\text{prev}}), \boldsymbol{s}_{\text{prev}}\} := \{L(\boldsymbol{s}_{\text{init}}), \boldsymbol{s}_{\text{init}}\}$.
3: **while** Iter $<$ ComputationalBudget

- **estimate** $\boldsymbol{\mu}^*_L$ and $\boldsymbol{\sigma}^*_L$ from data $\{L(\boldsymbol{s}_{\text{prev}}), \boldsymbol{s}_{\text{prev}}\}$.
- **estimate** $p(L^*(s)) \sim N(\boldsymbol{\mu}^*_{L(s)}, \boldsymbol{\sigma}^*_{L(s)})$ from posterior GP over data $\{L(\boldsymbol{s}_{\text{prev}}), \boldsymbol{s}_{\text{prev}}\}$.
- **estimate** $A(\boldsymbol{s})$, from $p(L^*(s))$.
- **query** $G(\boldsymbol{s})$ at $\boldsymbol{s}_{\text{new}} := \arg\max_{\boldsymbol{s}} A(\boldsymbol{s})$.
- $\{L_{\text{prev}}, \boldsymbol{s}_{\text{prev}}\} := \{L_{\text{prev}} \bigcup L(\boldsymbol{s}_{\text{new}}),\ \boldsymbol{s}_{\text{prev}} \bigcup \boldsymbol{s}_{\text{new}}\}$.

4: optimal solution $\boldsymbol{s} := \arg\max_{\boldsymbol{s}_{\text{prev}}} L(\boldsymbol{s}_{\text{prev}})$.

Problem Set Up

We begin with an objective function, $L(s)$, evaluable over the domain s. We would like to identify the global maximum of $L(s)$, while minimising the number of evaluations to $L(s)$. We begin with an initial set of queries to the domain, $\mathbf{s}_{\text{prev}}$, resulting in an initial set of evaluations $l_{\text{prev}} = L(\mathbf{s}_{\text{prev}})$. For all practical purposes, we lack any analytical knowledge that would allow us to optimise $L(s)$ more directly, except through sequential evaluations. We aim to minimise the number of evaluations required, but are hindered by our uncertainty in $L(s)$ at the points that have not been evaluated.

Probabilistic Model

For each new evaluation, we aim to use the information contained in $(l_{\text{prev}}, \mathbf{s}_{\text{prev}})$ to select the next query point $\mathbf{s}_{\text{new}}$ with evaluation $l_{\text{prev}} = L(\mathbf{s}_{\text{new}})$.

We denote s^* to be any point in s that has not been queried, and $L^* = L(s^*)$ to be its respective (true) evaluation in L. Clearly we are uncertain about the value of $L(s)$ at every point that we have not evaluated. A GP will be used to model this uncertainty. As shown in Figure 4.6, the GP presents the possible values of an unqueried s^* as univariate normal distributions, to which we may apply probabilistic reasoning, as described at the beginning of this chapter, particularly Figure 4.1(g,h,i).

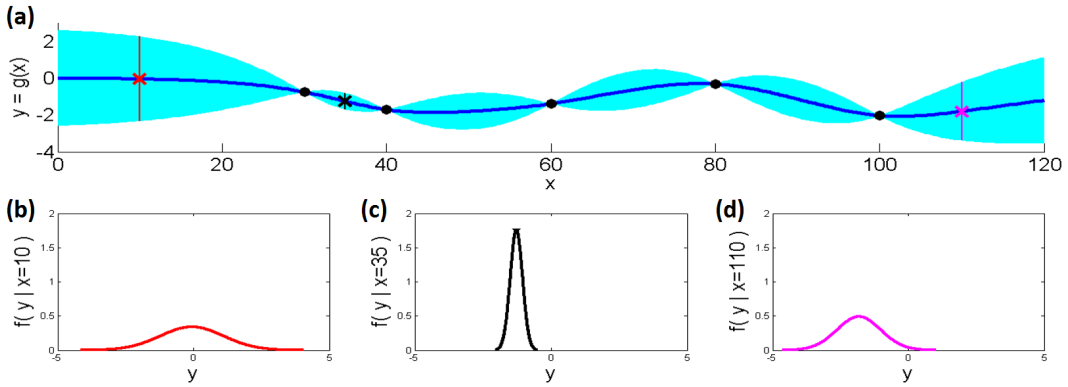


Figure 4.6: Probabilistic reasoning to query a black-box function. In (a) the posterior GP is fit to previously queried points (at $x = 30, 40, 60, 80, 100$) allows a principled form of inference over the un-queried points. Three potential queries at $x = 10$, 35, and 110 are shown in (b), (c), and (d), each corresponding in colour to the vertical line in (a). The desirability of the next query is a function of its marginal univariate Gaussian distribution, which formalises our belief about the potential to identify high values at that location. The best choice of query from among these three options could then be found by maximising the acquisition function, for example EI in Figure 4.7(d,e,f). This would select the point with the highest expected improvement in the objective function.

The approach to deriving these posterior marginals shown in 4.6(b-d) is no different than, say, the GP modelling that uses observed points in a time-series to estimate the uncertainty of values in the future, or for missing values within the time-series. The same creative GP modelling techniques that apply to other applications, apply to GP modelling of the objective function as well. We would like to incorporate a mean function and covariance function that properly reflect the properties of the generative function of the data, in this case, the black-box objective function. Common properties and how to model them will be discussed shortly in the section dedicated to the GP inference step of Bayesian optimisation.

Acquisition Function

The GP probability model alone (and its resultant posterior marginal distributions) is insufficient to select the next query point - we must have a decision rule to select the next query point (from a possibly infinite set of potential queries), given the posterior distribution at each point.

This decision rule is known as an "acquisition function", $A(s)$. Its salient features include that

1. $A(s)$ takes s^* as an argument,
2. $A(s)$ operates on the posterior probability distribution of L at point s^*, and
3. $A(s)$ provides (as an output) a metric to quantify the suitability of querying point s^*.

Since we wish to select the most suitable querying point (i.e., the point at which the acquisition function is largest), it is generally clearer to denote the acquisition function's argument to be s^*, instead of the posterior probability $p(L(s^*))$, although some authors include both as arguments.

There are many possible acquisition functions. For example, we may wish to query the location with the highest expected value. Alternatively, we might select the query that is most likely to improve over the current best value, or with the highest *expected* improvement over the current best value.

Querying and Update

With the next query point selected, the subsequent step is simple: we query the objective function at the new point.

We then update our set of observed points $(l_{\text{prev}}, \mathbf{s}_{\text{prev}})$ to include our newest query $(l_{\text{new}}, \mathbf{s}_{\text{new}})$. That is, $l_{\text{prev}} \coloneqq l_{\text{prev}} \cup l_{\text{new}}$, and $s_{\text{prev}} \coloneqq s_{\text{prev}} \cup s_{\text{new}}$.

In light of this updated set of information, our probabilistic inference on the unknown points in $L(s)$ may be updated as well. We return to the probabilistic modelling step and repeat again, in sequence, until stopping criteria are met. Stopping criteria typically involve a pre-defined budget of evaluations to $L(s)$, with early termination if further improvement is believed to be unachievable.

In Figure 4.7, we follow the posterior GP and acquisition function over three complete iterations of the optimisation *while* loop in a 1D example. In (a) we see the currently known evaluations of $L(s)$, along with the GP's posterior estimate of the uncertainty in the points that have not been queried. In (d) the corresponding acquisition function identifies the most promising point in s to query next. Having evaluated L at that query point, the set of unknown points is updated in (b), as

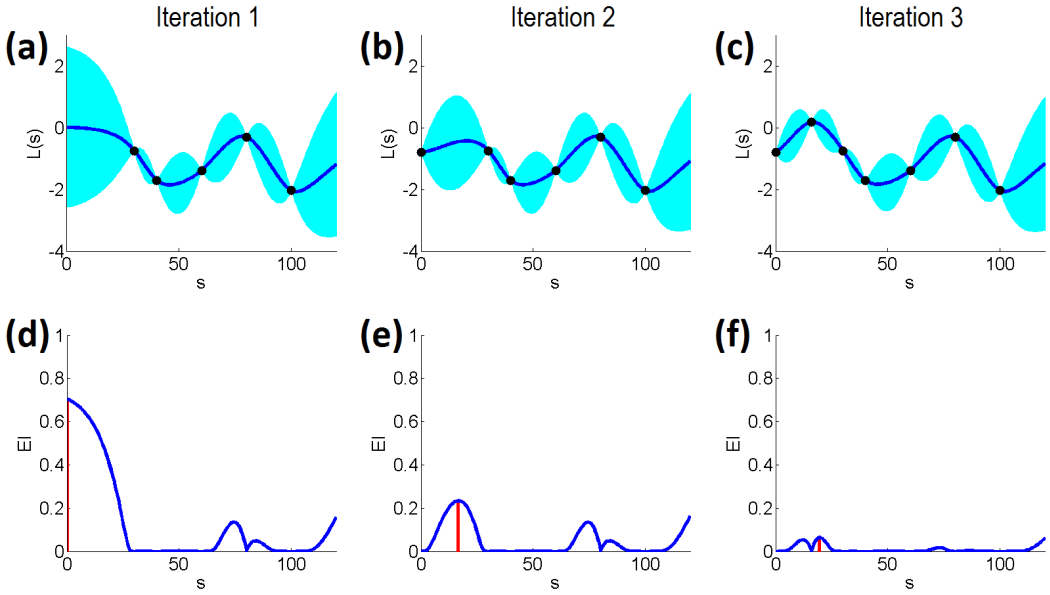


Figure 4.7: Three iterations of Bayesian optimisation. In (a) the posterior GP is shown from the five total queries, followed by (b) six total queries in the second iteration, and (c) seven total queries in the third iteration. The resulting acquisition function (expected improvement, in this case) is shown from the (d) first, (e) second, and (f) third iteration, with the EI-maximising location at each iteration is shown in red.

is the posterior distribution over L. The most promising new query (following this updated information) is shown at the peak of the acquisition function in (e). In this case, the algorithm chooses to exploit the same area instead of exploring under-sampled regions. The estimated EI decreases with each iteration as the algorithm explores more of the space.

This completes our overview of the Bayesian optimisation algorithm. Further details on the GP model and acquisition function are given in the dedicated sections below.

4.4.3 Bayesian Optimisation GP Modelling

As a Bayesian approach to optimisation, we hope to incorporate as much prior information as possible into our GP to reflect the characteristics of the objective function $L(s)$. This may seem counter-intuitive, since the motivation behind Bayesian optimisation is to handle functions that are black-box, non-analytic, and expensive to query. However, the guiding principles of the GP model is largely

the same. Just as before when we attempt to encode characteristics of heart rate as a function of time, we now aim to encode characteristics of objective L as a function of query point s.

The simplest example is measurement noise: if we know that querying the same location s^* will always yields an identical value of $L(s^*)$, then the covariance function will be modelled without a noise kernel (as in the red GP of Figure 4.5(a,c)), since there is no further irreducible error after $L(s)$ has been sampled at that point. Noise-free objective functions are common, for example, for testing parameterisations of computer simulation, or a machine learning algorithm with fixed training/validation data.

Similarly, if we know our function is likely to change rapidly with respect to one dimension, but very slowly with respect to another, then this can be encoded via differing length-scales across dimensions of s. Additive and multiplicative kernels can, respectively, encode logical "and"/"or" statements, allowing appropriate extrapolation far from observed values.

Even estimates of the objective function's gradient can be incorporated as a further predictive dimension of the covariance function k. This may help to better extrapolate in unseen regions and also alleviate the computational risks from querying multiple nearby points [140].

If similar optimisations have been performed in the past, or via expert knowledge, the prior mean function may help encode edge-case behaviour. For example, a machine learning system with guaranteed low performance when parameters are near the edge of the search space may place low priors values at the edges to entice queries away from the edge cases.

Once a prior mean and covariance function have been selected for the GP prior over $L(s)$, the hyperparameters of those functions, collectively denoted as $\boldsymbol{\theta}$, are subject to the same inferential options described earlier, such as maximum a posteriori estimation or slice sampling MCMC of the posterior distribution. In practice, it may be desirable to refit $\boldsymbol{\theta}$ each time a new data point is added to $\{L(\boldsymbol{s}_{\text{prev}}), \boldsymbol{s}_{\text{prev}}\}$, if we anticipate significant further learning of $L(s)$'s dynamics. Alternatively, $\boldsymbol{\theta}$

may be refit every few iterations. This approach would save on computation of fitting the GP model, but at the risk of using a less-appropriate value of $\boldsymbol{\theta}$.

The posterior predictive distribution, $N(\boldsymbol{\mu}_L^*, \boldsymbol{\sigma}_L^*)$, of $L(\boldsymbol{s})$ is calculated identically to that in Equation 4.14, but using the paired queries and evaluations $\{L(\boldsymbol{s}_{\text{prev}}), \boldsymbol{s}_{\text{prev}}\}$. Denoting

- $\boldsymbol{C}$ to be the covariance matrix between values $L(\boldsymbol{s})$ at locations $\boldsymbol{s}$,
- $\boldsymbol{C}^*$ to be the covariance matrix between values $L(\boldsymbol{s})$ and $L(\boldsymbol{s}^*)$ at locations $\boldsymbol{s}$ and $\boldsymbol{s}^*$, and
- $\boldsymbol{C}^{**}$ to be the covariance matrix between values $L(\boldsymbol{s}^*)$ at locations $\boldsymbol{s}^*$,

then $L(\boldsymbol{s}^*) \mid L(\boldsymbol{s}_{\text{prev}})$ is MVN such that:

$$\begin{aligned} \boldsymbol{\mu}_L^* &= \mathrm{E}[L(\boldsymbol{s}^*)] = \boldsymbol{C}^* \boldsymbol{C}^{-1} L(\boldsymbol{s}^*). \\ \boldsymbol{\sigma}_L^* &= \mathrm{Var}[L(\boldsymbol{s}^*)] = \boldsymbol{C}^{**} - \boldsymbol{C}^* \boldsymbol{C}^{-1} \boldsymbol{C}^{*T}. \end{aligned} \tag{4.15}$$

The predictive distribution of equation (4.15) can now be used to describe our posterior uncertainty in $L(s)$ at any point along the domain. In particular, we are interested in the uncertainty at unqueried points, from which we will select the next point to query, as shown in Figure 4.6.

Using this posterior uncertainty to make a decision of where to query next is the task of the acquisition function described below.

4.4.4 Bayesian Optimisation Acquisition Function

With the posterior estimate $L(\boldsymbol{s}^*) \sim N(\mu_L^*, \sigma_L^*)$ from Equation 4.15, our preference between different possible queries in s is a trade-off between the value of L that we expect at the query, $\boldsymbol{\mu}_L^*$, and the uncertainty of L around that expectation, $\boldsymbol{s}_L^*$. There are two plausible locations where the objective function may have high values: (i) query locations where $L(s)$ is known to be large (exploitation), or instead, (ii) query locations far from any previous queries where the objective function *may* be high (exploration).

The acquisition function formalises this trade-off in preferences. Since the posterior distribution of $L(\boldsymbol{s}^*) \sim N(\mu_L^*, \sigma_L^*)$ is identical to that of Equation 4.4, we can reason probabilistically, using the properties of a Gaussian discussed in Equation 4.7 and its conditional expectation from the beginning of the chapter.

Given the current best-found value, L_{best}, popular choices of acquisition functions include probability of improvement (PI) over L_{best}:

$$A(\boldsymbol{s}^*) := \quad \text{PI}(\boldsymbol{s}^*) = 1 - \Phi\Big(L_{\text{best}} \mid \boldsymbol{\mu}_L^*,\ \boldsymbol{\sigma}_L^*\Big), \tag{4.16}$$

or the expected improvement (EI) over L_{best}:

$$\begin{aligned} A(\boldsymbol{s}^*) := \quad \text{EI}(\boldsymbol{s}^*) &= (L_{\text{best}} - \boldsymbol{\mu}_L^*)\Phi\Big(L_{\text{best}} \mid \boldsymbol{\mu}_L^*,\ \boldsymbol{\sigma}_L^*\Big) \\ &\quad + (\boldsymbol{\sigma}_L^*)N\Big(L_{\text{best}} \mid \boldsymbol{\mu}_L^*, \boldsymbol{\sigma}_L^*\Big), \end{aligned} \tag{4.17}$$

where Φ and N are the Gaussian cumulative distribution and probability density, respectively.

The EI acquisition is typically more popular, since EI incorporates the magnitude of improvement over L_{best}, whereas PI gives equal preference to large and small improvements over L_{best}. As seen in equation 4.17, the EI acquisition function yields high values where either (i) the posterior mean is near or greater-than L_{best}, or (ii) the posterior variance is large, allowing opportunity to exceed the current best. A further offset may be added to L_{best} to prefer further exploration and avoid over sampling of a small area, as was done in Figure 4.7.

It is worth noting, that since $A(\boldsymbol{s}^*)$ requires the posterior mean and variance to be calculated for $\boldsymbol{s}^*$, identifying the $\boldsymbol{s}^*$ which maximises $A(\boldsymbol{s}^*)$ is a further optimisation problem in practice. However, since the posterior distribution is inexpensive to evaluate at any point, optimisation of the acquisition function is typically achieved by gradient methods, exhaustive enumeration, random search, or grid search. To avoid computational issues of inverting a singular covariance matrix, $\boldsymbol{C}$, a popular heuristic is to require new queries to be a minimum distance from any current points in s_{prev}.

While PI and EI are two of the simplest acquisition functions, a variety of sensible alternatives exist. For example, EI and PI, as presented, only account for a

single iteration look-ahead, however there are alternatives that account for a large remaining computational budget, or selection when multiple queries can be run in parallel. Other alternatives include upper bound and entropy-based functions.

4.5 Kernel Density Estimation

4.5.1 KDE Model

Kernel density estimation is a non-parametric modelling technique originally introduced by Rosenblatt [44] and Parzen [43]. Instead of relegating the fitted pdf, $p(y)$, to a pre-specified parametric form (e.g., Gaussian, Gamma, etc.), the distribution of the data is modelled by placing areas of probability density around the values where data are observed. This is achieved by placing a kernel, H with a portion of the total density, centred at each data point. The sum of these kernels meets all the properties of a pdf.

For example, using a Gaussian kernel

$$H(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \tag{4.18}$$

on data points $[y_1, \cdots, y_n]$ yields the probability density estimate

$$f(Y) = \frac{1}{n\omega} \sum_{i=1}^{n} H\left(\frac{Y - y_i}{\omega}\right) \tag{4.19}$$

with cumulative density function

$$F(Y) = \frac{1}{n\omega} \sum_{i=1}^{n} \Phi\left(\frac{Y - y_i}{\omega}\right) \tag{4.20}$$

Several kernels, such as the Gaussian kernel in equation 4.18, have non-zero density over the real line, resulting in a KDE with support across the real line. When this is undesirable, e.g., to model vital-signs which are non-negative or subject to upper and lower bounds, common corrective measures include:

- Variable transformation (e.g., log transformation for non-negative variables, or logit/probit transformation for bounded variables),

- Reflection of density outside of boundaries, or
- Truncation and factorisation by a normalisation factor.

Kernels are parametrised by a bandwidth parameter, ω. It is commonly noted that the kernel choice is frequently less important than the choice in bandwidth for many practical applications.

For kernel density estimation over multiple dimensions, the bandwidth parameter, ω, becomes a bandwidth matrix, Ω. This extension is similar, in principle, to estimating a single variance parameter for a univariate Gaussian distribution, but $n + \binom{n}{2} = \frac{n(n+1)}{2}$ unique covariance parameters for the covariance matrix of a multivariate Gaussian.

For clarity, a reasonable outline of KDE modelling, in increasing levels of complexity, may be:

- a univariate KDE
- a multivariate KDE with an isometric kernel
- a multivariate KDE with identical bandwidth parameters in each dimension, i.e., $H = \omega I$
- a multivariate KDE with unique bandwidth parameters in each dimension, i.e., $H = \boldsymbol{\omega}^T I$
- a fully-specified bandwidth matrix H

It is common practice, as seen in Hann [45] (who, in turn, cites Bishop [141]), to use an isometric kernel, which circumvents the need to estimate multiple bandwidth parameters. Similarly, one may use an identical bandwidth parameter for each dimension, that is, $H = \omega \mathbf{I}$. These approaches are typically accompanied with a zero-mean, unit-variance transformation along each dimension, as done in Hann [45].

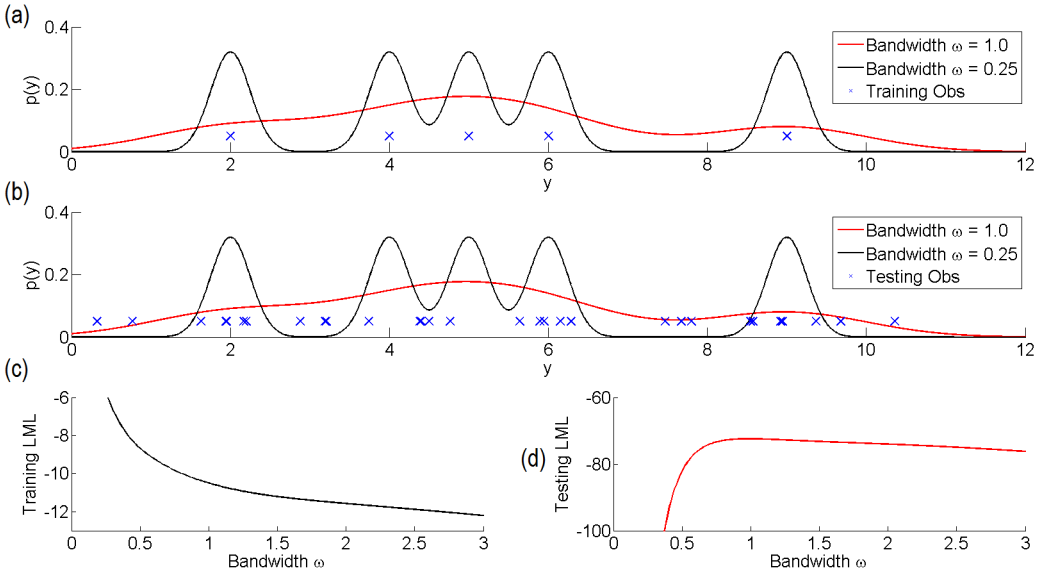


Figure 4.8: KDE bandwidth parametrisation. The locations of training points in (a) and testing points in (b) are shown by (×). A KDE is fit to the five training points in (a), under two possible bandwidths. As seen in (b) shorter bandwidth risk placing lower likelihood where the 30 held-out data points exist. As bandwidth ω decreases towards 0, the log-likelihood of the KDE (c) increases towards infinity for training data, and (d) decreases towards negative infinity for any testing data unseen in the training set.

4.5.2 KDE Inference

Like other probability models, inference on the KDE parameters (in this case the bandwidth parameter or bandwidth matrix) is required. By inspection of equation 4.18, direct maximum likelihood estimation is unsuitable for the bandwidth parameter ω, as the likelihood given ω can be increased *ad infinitum* by decreasing ω further towards zero. This is illustrated in Figure 4.8, where the likelihood of training points may continue to increase, even when the resulting likelihood is nearly 0 for the held-out testing/validation data. Accordingly, bandwidth parameters are typically selected according to rules-of-thumb, prior knowledge, cross validation techniques, or via Bayesian regularisation (e.g., a prior over ω to counteract the likelihood). Particularly popular methods are selection of ω to minimise integrated squared-error (ISE) and mean integrated squared-error (MISE).

The baseline-comparator in Hann [45] uses a rule-of-thumb described in Bishop [141]. This rule-of-thumb, which estimates ω of an isomorphic multivariate KDE to be the average distance between each kernel centre and its 10 nearest neighbours.

The KDEs implemented in the baseline comparators follow Hann's approach, to represent the work done in his thesis.

5 Detection of Artefactual Vital-Sign Measurements

Data cleaning is a first step in many analyses. Robust continuous monitoring of patient vital-signs must rely on vital-sign measurements that are representative of the patient's underlying physiology. Artefactual vital-sign measurements, which are not representative of the patient's physiology, undermine the clinical inference we wish to perform.

In this chapter, we motivate the use of probabilistic artefact detection to move beyond the most common, but insufficient, approaches used in current literature. We present several "archetypes" of vital-sign dynamics, and describe how to detect a subset of these archetypes as measurement artefacts. A computationally light-weight likelihood-based algorithm is proposed and tested on a cohort of patients. An application to patient deterioration detection is then described.

Key elements of this chapter have been published in [2], and the algorithm has been further used for pre-processing vital-signs in [1], [9], and [10].

5.1 Clinical Value of Artefact Detection

As discussed in Chapters 1 and 2, automated monitoring of patient vital-signs may improve the timeliness, accuracy, and transparency of clinical inference. Potential

improvements include (i) reliable monitoring in the absence of staff, (ii) reduced human error, and (iii) complex empirical modelling, estimation, and forecasting of patient health status. The efficacy of these automated systems, however, is undermined by inappropriate handling of artefactual data, which is acquired from the monitoring devices. This, in turn, undermines clinical staff's confidence in the use of automated methods.

There are many potential technical causes for artefactual vital-sign measurements [142]. These causes include (i) partial or complete probe detachment, (ii) algorithmic failure of signal processing (e.g., missed or extra beats in a heartbeat-detector), or (iii) corrupted device signal (e.g., from movement or perspiration). Potential physiological causes abound as well. For example, the outputs of many monitoring devices are validated on healthy patients, who differ significantly from the critically-ill patients whom we wish to monitor.

Regardless of origin, vital-sign artefacts are vexing to automated vital-sign inference because they add a further layer of confounding and complexity into an already complex system. These measurements must be handled to facilitate machine reasoning with regard to the patient's heath condition.

Data-cleaning is a fundamental step in most analyses. However, it is common in many patient-monitoring publications to neither (i) acknowledge nor (ii) handle the presence of artefactual measurements. When artefacts are acknowledged, common solutions include (i) pre-screening to remove physiologically-implausible high or low values, or (ii) introduce further smoothing of the vital-sign measurements under consideration. As an example, already described in Chapter 3, the thesis by Hann [45] removed measurements according to a pre-set threshold (e.g., HR values outside of the 30-300 bpm range). Further median-filtering, and capture-and-hold methods were used to avoid alarms due to transient artefacts.[1]

Both of these approaches have short-comings:

[1]It is understandable, of course, to give attention to these extreme-valued artefacts, since they directly precipitate false-alarms. However, this is no reason to ignore the majority of artefacts (with values within these thresholds), which may induce false alarms more indirectly.

First, smoothing techniques such as mean- or median-imputation obscure the measurement-noise of the monitoring device. This, in turn, biases the statistical descriptions of the time-series, which we would like to use for inference. For example, such smoothing may conflate measurement-noise with short-term variance of the underlying vital-sign.

Second, upper and lower thresholds to remove extremal measurements only address a minority of measurement artefacts. The remaining majority of artefacts (within the upper and lower bounds) would continue to interfere with automated clinical inference.

Principled approaches to identify and handle artefactual dynamics have been published as well [143]. Motivations to use less-involved methods include algorithmic simplicity, and lack of access to the original sensor waveform and clinical annotations.

The method proposed in this chapter is attractive in both its lightweight implementation (making it suitable for general purpose inclusion in wearable sensors) and applicability in the absence of waveform data.

5.2 Annotation of Artefactual Data

Twenty UPMC non-C"-patients with at-least 1 hour of data were selected for annotation.

The first 48 hours of each patient's HR time-series was visually inspected for likely artefacts. The artefacts were manually annotated via a program (designed by the author) using the Matlab graphical user interface.

The visual annotation of artefacts was conducted for each of the 20 patients in sequence. For each patient, HR measurements equal to 0 bpm or greater than 200 bpm were automatically removed and those measurements were recorded as artefacts. Removal of these extreme values reduced the visual distortion in the y-axis (heart rate bpm) caused by a large range of values. Next, each 12-hour segment of available HR data was viewed to (i) annotate large obvious artefacts, such as those in Figure 5.2, and (ii) gain an understanding of the patient's major trends in HR variation over the monitoring period. The artefacts annotated when

viewing the 12-hour segments were recorded and removed from the time-series to further reduce visual distortion. Finally, 2-hour segments of the time-series were viewed for less-obvious artefacts, such as those in Figure 5.2(c). The 2-hour window was advanced in 1-hour increments to maximise the visual context of the time-series. Artefacts annotated in this manner were recorded as artefacts, but not removed from plots of the time-series. Once the entire time-series had been annotated within the 2-hour windows, the entire time-series was replotted with artefacts and non-artefacts in different colours to ensure erroneous annotations were not visible. If erroneous annotations were visible, the patient's time-series was re-annotated for accuracy.

This annotation of 140,556 total HR measurements identified 21,062 artefacts (15%) and 119,494 non-artefacts (85%). Of the artefactual measurements, only 80 (<0.01%) exceeded the 30-300 bpm thresholds used by Hann [45] on the same data set.

The proportion of artefacts within each time-series differed by patient and is shown in Figure 5.1(b). The total number of HR measurements per patient ranged from 1,390 to 13,950, so it is unsurprising that artefacts were identified in each patient's time-series. The large proportion of measurements annotated as artefacts demonstrates how artefacts are more frequent than generally believed, and warrant the attention of those developing automated systems.

Visual inspection of the patients' time-series identified several archetypal HR dynamics that warrant consideration for data-cleaning. Figure 5.2 shows six such archetypes.

Certain archetypes clearly represent undesirable artefacts (which we may not wish to model), such artefacts include 5.2(a) probe attachment, 5.2(b) probe detachment, and 5.2(c) sporadic inter-beat detection failure and ectopic beats. The artefacts in 5.2(a) and 5.2(b) are contextualised by the absence of measurements before/after a non-physiological fall/rise. This suggests that the monitoring probe has just been reattached or detached. In 5.2(c), red measurements are contrasted with the consistent HR around 70 bpm, which are implausible variations over this short time-scale.

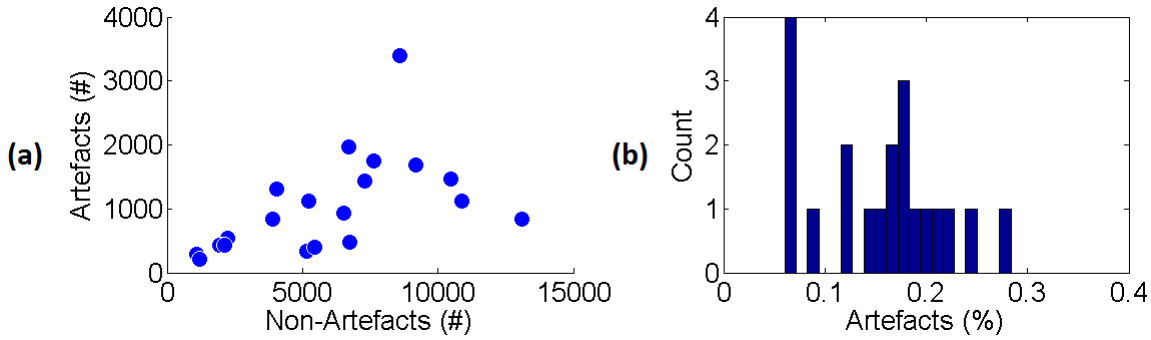


Figure 5.1: The frequency of artefacts in each of the 20 manually-annotated heart rate time series. In (a) the number of annotated artefactual and non-artefactual measurements shows that there is no clear correlation between the length of monitoring and the occurrence of artefacts. In (b) the proportion of artefactual measurements as part of total number of HR measurements shows that the frequency of artefacts per patient ranges from about 1-in-20 to 1-in-3 measurements. Elements of this figure were published in [2], and are © 2017 IEEE.

Other archetypes warrant identification, even if we do not currently wish to model them. These include bigeminy (not shown) and 5.2(d) atrial fibrillation. In the absence of further waveform or signal-quality data, there are many borderline cases as well, such as 5.2(e) which *may* be bigeminy after poor built-in median smoothing, and 5.2(f) which *may* be atrial fibrillation, or simply a high-noise regime due to signal interference coupled with algorithmic failure of beat-detection. Examples such as these illustrate that the annotation process is both subjective, and possibly erroneous for borderline cases.

A final distinction between (i) artefact and (ii) unusual dynamics is that we aim to model each time-series via a homoskedastic Gaussian process.[2] Dynamics such as atrial fibrillation or bigeminy induce heteroskedastic and highly-non-Gaussian noise. To facilitate our automated continuous modelling, we may wish to handle certain "true" dynamics as artefacts for a more reliable and replicable inference in light of our modelling choice. A more principled solution would be (i) to first identify the current time-series dynamic, and then (ii) apply bespoke modelling and inference that is most-appropriate for that physiology. However, in the absence

[2]That is, our model assumes that noise-corruption comes from a single IID Gaussian distribution, which is clearly untrue in light of certain types of artefactual corruption

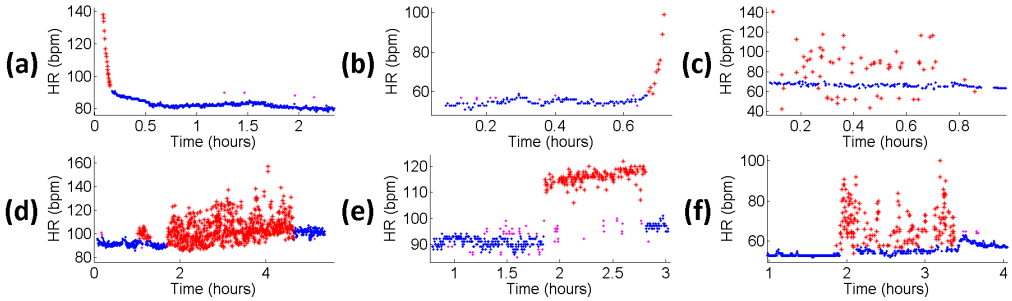


Figure 5.2: Six example HR dynamics. Examples contain non-artefactual measurements (•), example measurements from the archetype (∗), and plausible measurements of a different archetype (•). Archetypes (a)-(c) are transient, whereas (d)-(f) are probably non-artefactual and more persistent. This suggests that they likely differ in the appropriate method to detect these dynamics. Unique dynamics were identified only by time-series, not by waveform data. Elements of this figure were published in [2], and are © 2017 IEEE.

of waveform or further clinical annotation, such a system would be difficult to believe on the UPMC data set.

5.3 IID Modelling of Transient Artefacts

Artefact Model

From Figure 5.2 we see that (uncontroversial) artefactual measurements are outstanding in the context of temporally-proximate measurements (or lack thereof). We aim for an artefact detection algorithm with several attributes:

Computationally, we aim for an algorithm that is (i) lightweight (and may be plausibly embedded within a wearable sensor), (ii) fast (so as not to delay clinical inference), and (iii) transparent/intuitive in its decision process for clinical inspection.

Algorithmically, we aim for an algorithm that (i) may be run on time-series, in the absence of waveform data, (ii) is sensitive to transient measurement artefacts, as in Figure 5.2(a-c), and (iii) is agnostic to persistent noise dynamics, as in Figure 5.2(d-f), which should be left to an alternative form of artefact detection.

Since a distinguishing feature of transient artefacts is that they diverge from temporally-proximate measurements, an intuitive approach is to model all measurements $Y = y_1, ..., y_n$ within a short window of duration τ minutes as i.i.d. draws from a probability distribution $p(y)$. Artefactual measurements will have low likelihood with respect to $p(y)$. Since this requires a further (fallible) estimation

step to infer $p(y)$, we will average the log-likelihood of the measurement, with respect to $p(y)$, over each unique (but over-lapping) τ-length window (indexed $w = 1, ..., W$) that includes the measurement of interest, y

$$z(y) = \frac{1}{W} \sum_{w=1}^{W} \log p_w(y). \tag{5.1}$$

Transient artefacts will have low values of this novelty score $z(y)$, compared to (i) stable (presumably non-artefactual) measurements, and (ii) persistent high-noise regimes.

This approach merely requires the selection of (i) window-length τ, (ii) probability model $p(y)$, and (iii) the inference procedure for $p(y)$. Each is covered briefly below.

Selection of Time Window τ

Across windows of short time-length τ, an outlying measurement becomes well-separated from its neighbours. A short τ comes at the cost of a smaller number of neighbours by which to infer the sample distribution $p(y)$. In contrast, a large τ incorporates more data by which to estimate $p(y)$, but (i) conflates noise variances with the potential that the underlying HR has changed, and (ii) delays the inference as the algorithm must wait until all data is collected.

Values of τ between 5-10 minutes were found to have comparable ability to discriminate artefacts, whereas τ values less than 5 minutes began to compromise discriminative performance. A $\tau = 5$ minutes was selected to minimise lag and computation without compromising performance.

Selection of Inference Method of $\mathbf{p(y)}$

As measurements from physiology, it is desirable to encode physiological knowledge via Bayesian methods. However, the primary consideration for the inference step is that the entire algorithm should remain both computationally light-weight and fast. MCMC-based methods were excluded for this reason, along with the fact that the novelty score of Equation 5.1 already contains an element of model-averaging. For the parametric probability distributions considered to model $p(y)$

(e.g., Gaussian, log-normal, and gamma), all had maximum a posteriori estimation with analytical solutions (for some priors over their respective parameters). However, a consistent and rigorous framework to parametrise these priors, particularly for location parameters, was difficult due to the wide range of possible values across many patients at any time. This challenge was illustrated by the inter- and intra-patient variability in vital-sign values in Chapter 3.

For frequentist inference, the potential parametric models of $p(y)$ each had analytical maximum-likelihood estimates, which makes the inference step predictable in both operation and run-time. Frequentist maximum likelihood estimation was chosen. Non-parametric KDE modelling was also examined. As described in Chapter 4, the KDE requires cross-validated selection of a bandwidth parameter (in place of maximum likelihood selection).

Selection of $\mathbf{p(y)}$

The Gamma distribution over HR demonstrated a marginally superior ability to discriminate between artefactual measurements and non-artefactual measurements compared to normal and log-normal models of $p(y)$. KDE demonstrated poorer discriminative ability than the parametric models, presumably due to placing a minimal level of likelihood on every observed HR measurement.

Final Artefact Scoring Algorithm

Figure 3 illustrates the steps to calculate $z(y)$. At each window τ a gamma distribution

$$p(y) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} y^{\alpha-1} \exp\left(-\frac{y}{\beta}\right), \quad y > 0, \tag{5.2}$$

was fit to the data y within the window. The gamma distribution is parameterised by α, a shape parameter, and β, a scale parameter. The values of α and β are set to their maximum likelihood estimate, that is, the values of α and β that maximise Equation 5.2. These values of α and β are apt to change as the values within the window change. As described in Equation 5.1, the artefact score of each

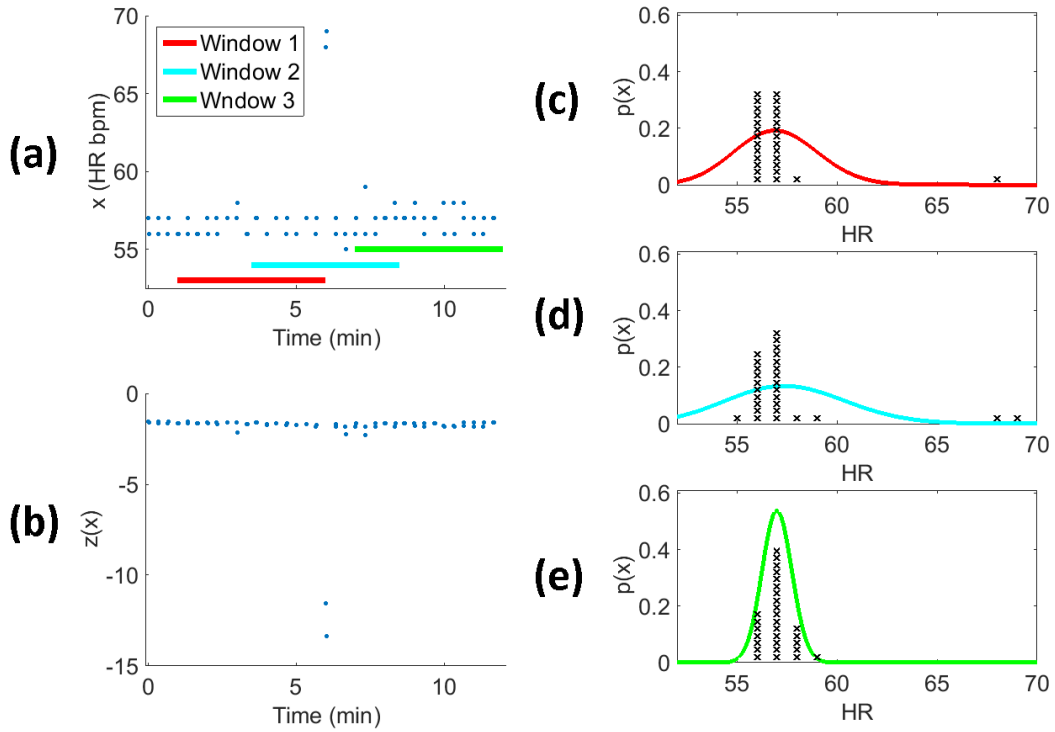


Figure 5.3: In (a), a patient's HR time series is shown with two artefacts around $t = 6$ minutes. The time-series is underlined by three unique windows of $\tau = 5$ minutes. There is substantial overlap of measurements within nearby windows since each new measurement creates a unique window. In (b) the corresponding artefact score is shown for each measurement in (a) with low values at the time of the artefacts. A gamma distribution's PDF (as in equation 5.2) is fit via MLE to the measurements (x) within each window in (c), (d), and (e). While the artefactual measurements in the tails of (c) and (d) visibly flatten the PDF over the window of measurements, their tail-likelihood is significantly lower than the likelihood of the other measurements in the same window. In (e) the artefact-free window is well-described by the fitted Gamma pdf, which further bolsters the average log-likelihood for non-artefacts. When averaged across all windows, the artefact score in (b) is much lower for the transient artefacts, as desired. Elements of this figure were published in [2], and are © 2017 IEEE.

measurement is equal to its average log-likelihood across all W windows in which it is a member. Measurements with low values of z may be considered likely artefacts.

5.4 Discriminative Ability of Artefact Score

The artefact score's ability to discern between artefactual and non-artefactual measurements is examined in the trade-off between the proportion of artefacts and non-artefacts that would be discarded at any given threshold k on the artefact score $z(y)$. Results are shown in Figure 5.4 both for 5.4(a) inter-patient variability

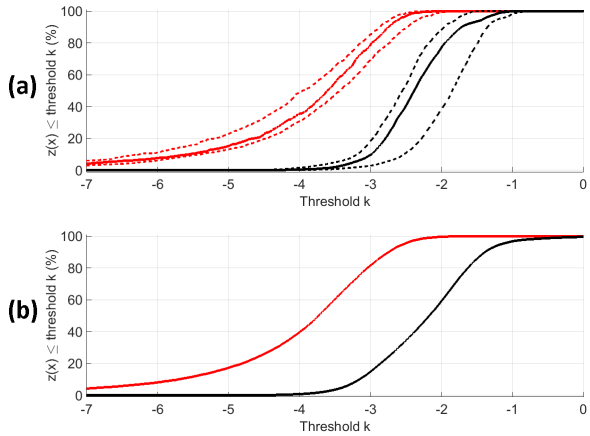


Figure 5.4: Ability to discriminate between artefact (–) and non-artefact (–), assessed on (a) for each patient, and (b) on aggregate. In (a), the CDF of $z(y)$ was calculated for each of the 20 patients under consideration and the the 20%, 50%, and 80% quantiles across all patient's results are shown. In (b), the CDF is shown for all measurements aggregated across all patients. A threshold k to classify artefacts may be determined using such plots, according to what is an acceptable trade-off for a particular application. Elements of this figure were published in [2], and are © 2017 IEEE.

and 5.4(b) on aggregate.

However, since there is an inherent inter-patient variability, a specific threshold k will remove a different proportion of artefactual measurements and non-artefactual measurements, depending on the patient. Even with the uncertainty of inter-patient variability in 5.4(a), it may be seen that the artefact score effectively differentiates between artefacts and non-artefacts. For example, at threshold $k = -4$, half of all patients would have at least 35% of all artefacts removed, and almost no non-artefacts removed. The top 20% of patients at $k = -4$ would have at least 50% of artefacts removed, and the top 80% of patients would have at least 30% of artefacts removed. From the y-axis ranges, we can also see that the artefact scores of artefactual measurements have a much narrower range in artefacts (red) than non-artefacts (black). This suggests that there is only modest inter-patient variability in the proportion of artefacts removed at a given threshold. In particular, the large gap between the 50% and 80% quantiles suggest that most patients have a considerable gap in $z(y)$ values between their artefactual and non-artefactual measurement, up to the point at which nearly all artefacts would be removed.

In Figure 5.4(b), when aggregating all scores, regardless of patient, about 40% of artefacts and <1% of non-artefacts had a score below threshold $k = -4$.

These results suggest that it is possible to remove a large portion of artefactual measurements while removing only a negligible proportion of the non-artefactual measurements. Since the identification of artefacts is only used for machine learning ends (i.e., it does not directly generate clinical alarms) there is little risk from false-positive identifications (beyond that it removes data we would otherwise prefer to have for clinical inference). The results in Figure 5.4 suggest that there is little risk of removing very much data that we would like to retain for clinical inference. When paired with the low inter-patient variability, there is little risk in applying a single threshold across an entire patient cohort. With no further patient-specificity required to remove the most apparent artefacts, such a method may be suitable (for example) to embed within a wearable sensor for general purpose monitoring.

In comparison to current practice (e.g. simple thresholds from Hann [45]), the likelihood-based algorithms adds little to computational demand and run-time. In terms of run-time, the proposed algorithm can compute $z(y)$ for an hour of HR data (collected at $\frac{1}{5}$ Hz to $\frac{1}{3}$ Hz) in less-than 1 second when programmed in Matlab. From this, the time delay in using the proposed algorithm is only as long as the window length τ required to collect the data. In terms of efficacy, the proposed algorithm can remove many artefacts with minimal loss of non-artefactual data. Although measurements exceeding the range of 30-300 bmp are virtually assured to be artefacts, these extreme-valued instances are less than 0.01% of the total annotated artefacts. Further varying of the thresholds proposed by Hann quickly degenerates into removing true measurements since, as seen in Chapter 3, the range of patient values is highly individual. We suspect that the proposed likelihood-based algorithm circumvents the need for patient individualisation by performing inference using only the most pertinent data, that is, the patient's current values at the current time.

5.5 Clinical Applications of Artefact Detection

The ultimate goal of artefact detection is to improve automated clinical inference. Improved data supports such applications as inputs to (i) forecast future vital-signs (as described in Chapter 6 and 8), (ii) identify a change in measurement-noise regimes or patient status, or (iii) infer deterioration (described in Chapters 7 and 8).

To conclude, we demonstrate how simple removal of the most extreme artefacts may improve deterioration detection. Although complete removal is not the only (or best) way to handle measurements deemed artefactual, it illustrates how a simple application may be applied to improve the end goals of this thesis.

As described in Chapter 8, time-series modelling provides an effective tool to identify unusual vital-sign dynamics over time (for example a precipitous rise or fall in HR beyond what would be expected of normal variability). As illustrated in Figure 5.5(a), sequential GP forecasts could identify such step-changes, quantified by forecast likelihood.

This approach is similar to likelihood-based artefact detection method itself, except that we (i) incorporate a Bayesian time-series inference in place of a frequentist IID inference, and (ii) evaluate the likelihood of a window of data instead of individual measurements.

Artefacts hinder this inference by over-stating the noise or volatility of the underlying physiology in the training or forecast window. By removing the most erroneous data we blind our clinical inference to misleading data.

As described in Chapter 3, the UPMC data set has 59 patients with clinically-annotated cardiorespiratory instability events (C"-events), along with the time stamps of those events. By applying the step-change detection algorithm described above, we examine the trade-off between time of early warning (TEW) among the 59 patients with cardiorespiratory instability events (C"-patients) and the false positive alarm rate (FPR) among the 89 patients with no such annotated events (non-C"-patients). (Further details will be given in the dedicated chapter.)

In Figure 5.5, we show the TEW in advance of the first C"-event for each of the 59 C"-patients. The FPR is the rate at which the step-change detector alarmed

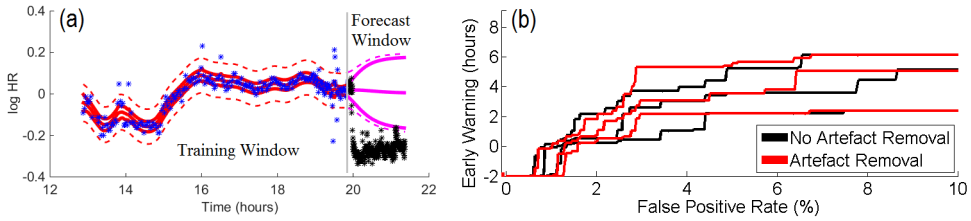


Figure 5.5: Improving patient monitoring with artefact removal. (a) A GP step-change detector may help identity abnormal vital-sign dynamics, however this inference can be hampered by noise in both the training and/or forecast window. (b) GP deterioration detection is performed with (–) and without (–) artefact removal, using threshold of $z(y) \leq k = -4$. Increased sensitivity improves the hours of early warning in advance of the annotated clinical emergency, but also increases the rate of false positive alarms in patients without clinical emergencies. Lines show the 15%, 20%, and 25% quantiles of early warning among the 59 patients with clinical emergencies. Elements of this figure were published in [2], and are © 2017 IEEE.

in the 89 patient without annotated alarms. At a higher level of sensitivity to step-changes, we may detect deterioration earlier in the C" patients. Due to the individual patient dynamics, the TEW in C" patients varies on an individual basis, creating 59 unique TEWs (one TEW per C"-patient). We summarise this TEW distribution by quantiles. We would like to compare the TEW vs FPR trade-off from before (black) and after (red) artefact removal.

For patients with the largest TEW, artefact removal showed nearly identical performance (not shown). That is, the trade-off between TEW vs FPR had negligible improvement for those patients who already benefited from very early warning. In contrast artefact removal showed useful improvement for patients for the lower quantiles of TEW (the patients currently receiving the *least* early warning via step-change detection). This is a positive result, since those patients are most vulnerable to (i) missed warning signs, or (ii) warning signs that are too late for effective clinical intervention. For example, at an FPR of 3%, the patients with the lowest 15% of TEWs (shown by the lowest red and black lines of Figure 5.5) improved from an early warning of 0 hours (no advanced warning, only a contemporaneous warning) to 2 hours of early warning. The improvement of advanced warning from 0 to 2 hours could be invaluable for clinical staff, and potentially life-saving for a patient.

While this example is performed on a small number $(59 + 89 = 148)$ of patients, and therefore is far from conclusive, it suggests that simple methods to improve data quality, may help remove more subtle sources of error.

5.6 Time-Series-based Artefact Detection

The IID model for artefact scoring was attractive for its simple and transparent (and therefore, robust) probabilistic inference. However, it should be noted that a GP could also provide a such a score by replacing the $p(y)$ from Equation 5.2 with the GP's posterior predictive distribution at $y(t)$.

A simplistic way to do this would be to assess the log-likelihood of each measurement in the time-series with respect to the posterior GP. Measurements with low log-likelihood would then be candidates for removal as artefacts. Such a method was used to remove artefacts in [9] and [10]. A further addition could apply EVT to calculate the probability of a measurement at least as extreme as the measurement under consideration. However, when tested, the EVT-based methods were highly sensitive to model misspecification (e.g., the estimation of noise variance) when calculating these probabilities. Alternative scores derived from the GP model could involve metrics drawn from Q-Q plots or Cramér-von Mises test statistics to simultaneously (i) assess individual measurements while (ii) generating useful model-checking metrics for the GP fit to the data.

Moving away from the traditional homoskedastic GP, alternative models could include (i) "fault bucket" GP methods, (ii) data cherry picking regression methods, (iii) mixtures of GPs and (iv) Student-t processes.

Modelling via (i-iii) involves the assumption that the vital-sign data comes from multiple data generative sources or latent functions: one data generative process that produces the non-artefactual vital-sign measurements, and another (or multiple other) data generative processes that produce artefacts. For example, the fault bucket could measure non-artefactual measurements by a GP with a covariance function with a white noise term, such as those illustrated in 4.5(c,d). This GP covariance would then be supplemented with an additional high-variance white

noise kernel, as in Equation 4.13, to model high-noise volatility. This may be an appropriate model for the artefacts in Figure 5.2(c,d), which likely represent signal processing noise and atrial fibrillation, respectively.

Mixtures of GPs may be attractive to model time-series that appear to have distinct simultaneous trajectories or for instances were the distribution $p(y(t))$ appears to be multimodal. Physiologically this would include modelling for patients with bigeminy, trigeminy, etc. Time series of this nature could include those in Figure 5.2(e) or Figure 3.1(a), both of which show two distinct HR trajectories. Such methods would be a principled way to handle persistent artefacts, in contrast to the IID method of this chapter, which handles transient artefacts.

The final alternative, Student-t processes, would make use of the heavy-tailed Student-t distribution to model large outliers, such as those in Figure 5.3(a) at $t = 6$. Whereas the Gamma distribution flattens significantly in the presence of large outliers, as seen in the gamma PDFs of 5.3(c) and 5.3(d), the heavily-tailed Student-t distribution may be less-flattened, thereby better distinguishing between artefact and non-artefact. Furthermore, by robustly modelling these transient artefacts, the initial step of removing artefacts may be rendered unnecessary since the time-series model is less affected by such measurements.

Computationally, although GP-based and other time-series-based methods require a larger calculation to fit the model (compared to the IID gamma model), any artefact score would not require recalculation over multiple windows because the model has estimated the mean and noise variance at a given point. Furthermore, the time-series methods may completely circumvent the need for an additional pre-processing step to remove artefacts, if artefacts may be robustly incorporated into the probabilistic model.

5.7 Conclusion

In this chapter we demonstrated that artefactual vital-sign measurements are more common than generally thought and that naive approaches to handling

artefacts are insufficient to either (i) identify most artefacts, or (ii) prepare data for subsequent statistical analysis.

A simple likelihood-based artefact score is introduced to identify transient artefacts while remaining agnostic to more persistent archetypal dynamics. Due to direct inference on each patient's vital-sign measurements, the artefact score has low inter-patient variability. This suggest that the score may be appropriate for use on a patient cohort without need for further patient-specific refinement. The absence of waveforms relegates the current algorithm to general purpose monitoring (that is, monitoring without the goal to identify disease-specific physiology). However, if waveform data were available, there is are various ways in which such information could be included.

Finally, we demonstrated how artefact removal may assist with deterioration detection by improving early warning performance on patients who are may be otherwise missed. Although far from conclusive, the improvement in the low-TEW patients underscores how difficult-to-identify patients may benefit from more involved monitoring.

6

GP Kernel Construction for Patient Monitoring

The GP's Bayesian nonparametric framework accommodates significant modelling flexibility. This flexibility can assist in representing the underlying physiology that generates vital-signs. In this chapter we first demonstrate how a kernel selection procedure may flexibly model a wide range of dynamics from a cohort of patients. Next, we propose to further personalise kernel complexity and parameterisations via the patient's previous monitoring data. These personalised models may be learned using simple or sophisticated optimisation procedures. We demonstrate the improvement in such models by improving worst-case scenario forecasting.

Key elements of this chapter have been published in [1], [3], and [4].

6.1 Clinical Value of GP Kernel Construction

Personalised inference of patient health has vast potential to improve patient monitoring when compared to a single population-based model applied across a large heterogeneous clinical population. Large inter-patient variability requires population-based models to accommodate large ranges of "normal" data that are highly abnormal in a patient-specific or time-specific context. GP modelling of a

patient's time-series implicitly avoids such limitations by performing all inference via the patient's own data.

The use of personalised models has inherent challenges: Reliance on only the patient's data requires that we effectively begin with no hard data. The learning process, therefore, must develop from a robust reasoning pathway by which to balance (i) physiological knowledge and (ii) the data of previous patients with (iii) continuously acquired data from the current patient. We contend that Bayesian methods provide such a reasoning pathway. Furthermore, personalised models must be learned in a timely fashion to assist timely clinical inference.

The challenge of personalised modelling, then, is that we must consider a model (or family of models) that is sufficiently flexible to accommodate a diverse range of patients, but we must learn patient-specific parameterisations and regularisation for the individual patient in a timely manner.

This chapter will illustrate of two main hypotheses: (i) population-based GP priors may adequately model the wide range of vital-sign time-series dynamics exhibited in the SDU, and (ii) further patient-specific refinement of these models can out-perform population-based GP priors for useful clinical tasks. Furthermore, the identification/specification of these personalised models are best determined by sophisticated methods of optimisation, due to highly interrelated GP hyper-parameters used to describe physiology.

Figure 6.1 illustrates the clinical value of good parametrisation or regularisation of a GP model: The GP's flexibility is helpful to model a range of physiology, however this flexibility increases the risk of model misspecification. For example, when we are certain that our model has correctly parametrised a patient model, we may confidently discern between between 6.1(a) predictable and 6.1(c) unpredictable volatility. However, in the presence of 6.1(b,d) poor model specification, it is unclear whether volatility was truly unpredictable, or whether we simply failed to learn from predicable dynamics. Since we may also aim to use unpredictable volatility to infer deterioration (as described in Chapter 8), we must, in turn, aim for a model that minimises (avoidable) poor forecasts.

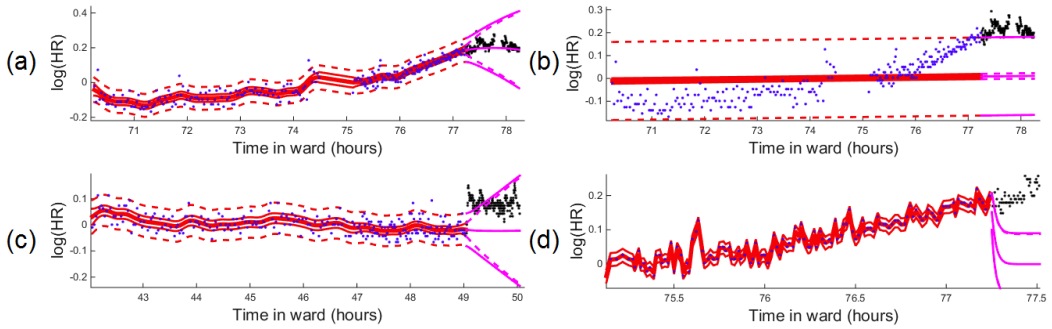


Figure 6.1: Model specification and misspecification. A GP is fit to the training window (•) of each of four time series. A forecast is then made for future measurements (•). Solid lines show the latent mean and its 95% CI. Dashed lines show the 95% CI of observations. GP forecasts may inform us of (a) future values, or (c) unforeseen volatility. In both cases, the GP model require correct parameterisation for robust clinical inference. Some poor forecasts, which may be mistaken for deterioration, could be foreseen, such as when the noise parameter, σ_n^2, is (b) over-estimated, or (d) under-estimated. Elements of this figure were published in [1] and are © 2017 IEEE.

6.2 Data Set

6.2.1 Training, Validation, and Testing Set

The UPMC data set's 169 non-C" training set patients (with at least 1 hour of data both before and after 24 hours on ward) were subdivided into training, validation, and testing sets to learn (i) a patient-cohort model, and (ii) the procedure to learn patient-specific GP models.[1] A testing set of $169 - 43 = 126$ patients will demonstrate the inductive value of these approaches model-learning.

Without patient-specific learning, the GP model (or modelling procedure) will be invariant to the the amount of patient data that is collected. In contrast, patient-specific modelling will, naturally, be a function of the available data. As described in Figure 6.2, we reserve the first 24 hours of each patient's data for any patient-specific learning. the comparison of patient-cohort versus patient-specific model will then be assessed in hours 24-72 for each patient. Since clinical systems are typically assessed against their worst-case performance, we will prefer a modelling procedure that minimises the frequency or magnitude of "very poor" forecasts.

[1] We use the term "validation" as it is used in the computer science literature: to estimate how our modelling choices will perform when using previously-unseen data. In contrast, medical literature uses the term "validation set" to refer to held-out data, used for an unbiased estimate of performance in the final selected model.

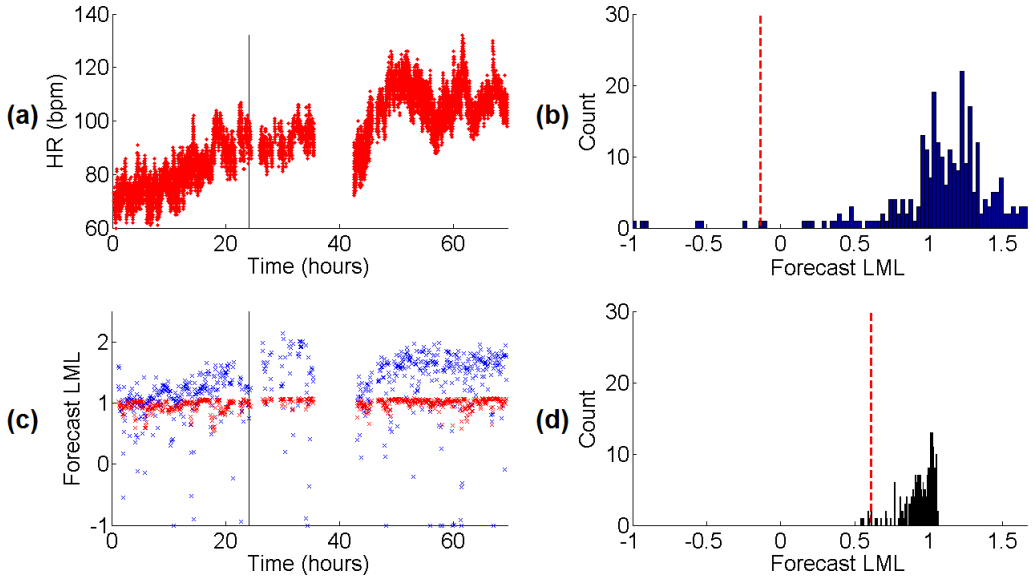


Figure 6.2: (a) Three days of a patient's continuously-acquired HR data. (c) The sequential forecast performance using a cohort-based GP model (•) versus a regularised GP model (•). Note that although the use of an cohort-based GP model (•) tends to result in higher values of forecast LML for this patient, the regularised method (•) avoids the large number of poor estimates and is more robust overall. The distribution of the forecast LMLs from (c) is shown in (b) for cohort-based GP model and (d) for patient-specific regularisation to optimise the 2.5-percentile of forecasts (the vertical dashed line), which is objective function G_1, defined later. Although use of the regulariser certainly avoids worst-case scenario performance, we believe that the regulariser could be learned without such detriment to upper-end performance. This motivates the use of multi-objective optimisation, described later. Elements of this figure were published in [1] and are © 2017 IEEE.

While, for demonstration purposes, the first 24 hours of each patient's data is used for learning, in practice we would not wait 24 hours to begin patient-specific inference. Instead, the patient-specific GP model could be learned and updated continuously, beginning immediately upon admission to the ward. For simplicity, however, we will examine the performance of these models after 24 hours of observation. This 24-hour cut-off is a trade-off between a large enough data set on which to learn a model, and a large enough data set left over on which to assess performance. However, this arbitrary cut-off could be any time $t > 0$, with updates as frequent as desired or as computationally possible.

For a rough estimate of data availability, the number of HR measurements for each patient is shown in Fig. 6.3. The available data is roughly the same

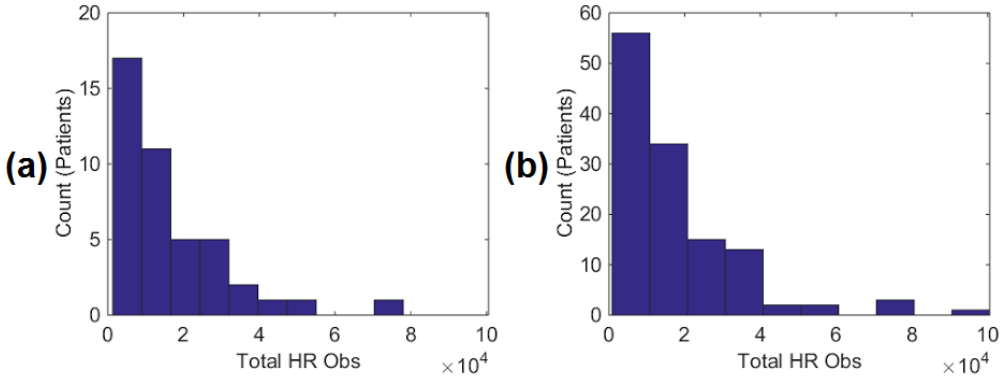


Figure 6.3: The total number of HR measurements (a) for each of the 43 patients in the training/validation set, and (b) for each of the 126 patients in the test set. Patients with fewer measurements (which may be dispersed in time) will have fewer data from which to learn a personalised model (i.e., they may have fewer data in their first 24h), and fewer data with which to test the performance of a personalised model (i.e., they may have fewer data after their first 24h). Elements of this figure were published in [1] and are © 2017 IEEE.

between 6.3(a) the 43 training/validation patients and 6.3(b) the 126 (held-out, previously-unseen) test patients. With HR acquired between $\frac{1}{3}$ Hz to $\frac{1}{5}$ Hz, most patients had more than one day ($\sim 1.7 \times 10^4$ measurements), but less than two days of total usable data.

6.2.2 Data Preprocessing

As described in Chapter 5 (and partnered publication [2]), continuous HR data are typically beset with artefactual corruption due to the measurement process, including partial attachment of the measurement probe (often an ECG electrode or finger-mounted pulse oximeter), or failures to identify the pulsatile complex of wave-forms (e.g., the QRS complex in the ECG) that subsequently confounds HR estimation.

Just as we advocate the modelling of a patient's time-series data using patient-specific (and not population-based) approaches, a principled approach to artefact removal evaluates potential artefacts in light of the patient's current measurements, and not using population-based rules-of-thumb. The likelihood-based artefact score was used to identify measurements that deviated significantly from nearby measurements. Measurements with an artefact score less than -4 were removed as artefacts. Among the annotated patients, this corresponded to roughly 40% of

artefactual measurements and less than 1% of non-artefactual measurements. Due to inter-patient variability, these values may vary between 30% to 50% of artefacts, and 0% to 1% of non-artefacts, for most patients.

6.3 Clinical Performance Objective

As described above, we will illustrate kernel building for the task of sequential vital-sign forecasting. That is, we aim to continuously update our forecast of future vital-signs as we continuously acquire new vital-sign measurements. We simultaneously quantify the accuracy and precision of the forecast via the LML of the future data with respect to the GP's posterior predictive distribution. We aim to optimise forecast LML performance by a personalised-selection of $\boldsymbol{\theta}$ (the vector containing all GP hyperparameters) for each patient.

Figure 6.4 shows the effect of hyperparameter values on forecasting with a 30-minute look-ahead. A forecast is made at hours $[2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6]$. In practice, however, forecasts can be made for any length look-ahead (e.g., 1 minute, or 1 hour), and at any frequency (e.g., continuously with each newly acquired vital-sign measurement, or minutely, or at times requested by clinical staff).

For a GP with a single Matérn 5/2 covariance kernel, a shorter length-scale $\lambda = 5$ better fits the contours of the training data in 6.4(a) compared to the longer length-scale $\lambda = 15$ in 6.4(d), which tends to over-smooth. This results in higher forecast LML values in 6.4(b) and 6.4(c) compared to 6.4(e) and 6.4(f). Furthermore, as shown in 6.4(g) and 6.4(h), the effect of altering the value of one hyperparameter is highly sensitive to the values of the other hyperparameters. This effect is particularly pronounced in worst-case forecasts, which are those forecasts which fall into the lowest quantiles of forecast performance. Examples of particularly poor forecasts include the forecast in (i) 6.4(a) from hours 4 to 4.5, (ii) 6.4(c) from hours 4 to 4.5, and (iii) 6.4(c) from hours 5.5 to 6. In each of these three forecasts, the majority of HR measurements are far outside the 95% predictive bound, and, accordingly, have low forecast LML.

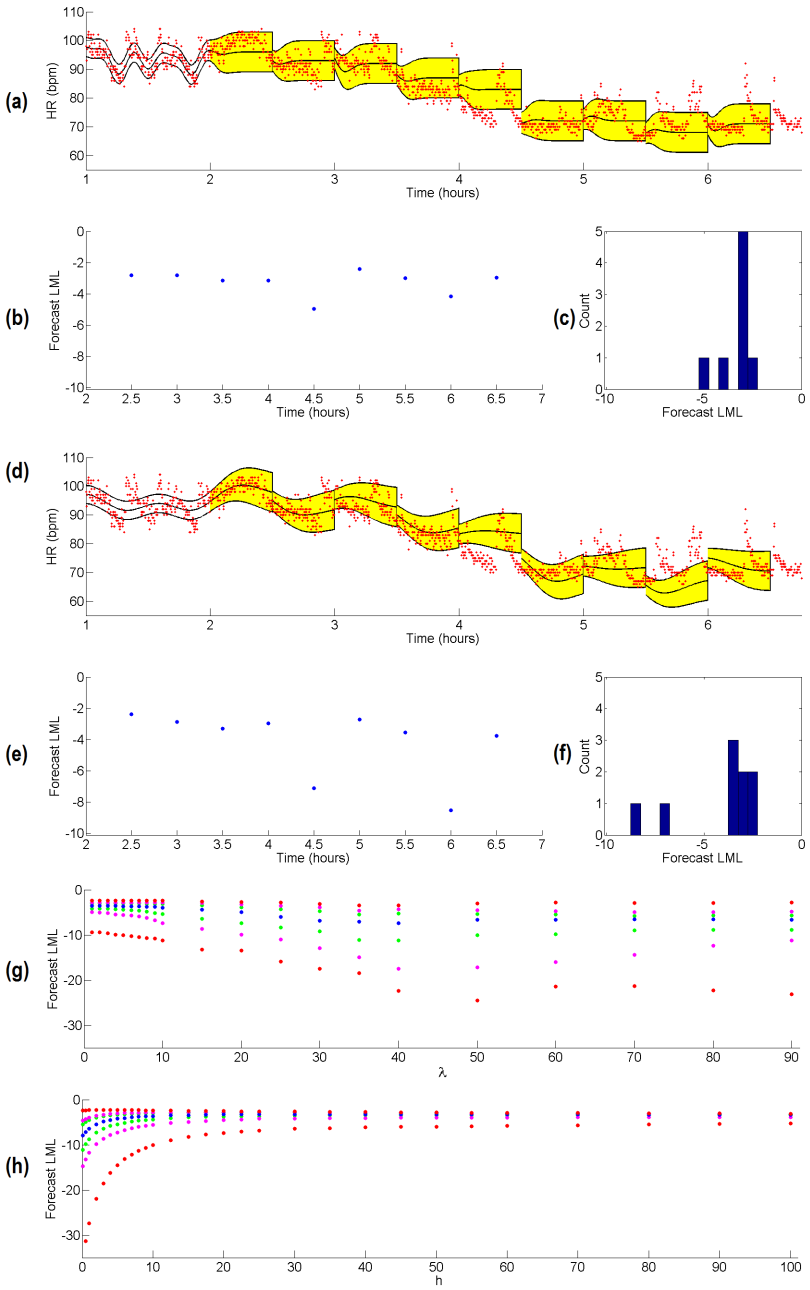


Figure 6.4: Sequential prediction on a time-series. A GP is sequentially fit via a fixed (a) length-scale of $\lambda = 5$ minutes, and (d) length-scale of $\lambda = 15$ minutes. The forecast LMLs at each time point are shown in (b) and (e), respectively. This time-series of forecast LMLs form a distribution, in (c) and (f), which may be summarised by a further statistic, such as mean or a quantile, to describe the overall forecast performance of the choice in hyperparameters. The percentile of these distributions will change by varying the parameters, as shown in (g) by fixing $[h, \sigma_n] = [10, 2.5]$ and varying λ, in (h) by fixing $[\lambda, \sigma_n] = [5, 2.5]$ and varying h. Percentiles are calculated from $4.5 \times 60 = 270$ minutely-forecasts between hours 2 to 6.5, and show the 50 (●), 25 and 75 (●), 10 and 90 (●), and 2.5 and 97.5 (●) percentiles.

Our principle aim is to minimise worst-case forecast performance, which we formalise to be the 2.5-percentile of all forecast LMLs on an individual patient's vital-sign. We will examine forecast performance on each patient separately (instead of on aggregate) (i) to avoid Simpson's paradox in which a "best model" on aggregate is sub-optimal for most or all patients, and (ii) because some patients are simply harder to predict than others, so the maximum-attainable forecast LML will vary by patient-to-patient. From this inter-patient variability, absolute thresholds may confound our understanding of optimal performance for a particular patient.

Defining $\mathcal{L}$ to be the set of all forecast likelihoods of a patient, and $\mathcal{L}_\rho$ to be the ρ^{th} percentile of $\mathcal{L}$ our first performance metric is the 2.5-percentile of $\mathcal{L}$:

$$G_1 = \mathcal{L}_{2.5} \tag{6.1}$$

Noting, from Figure 6.2, that such an objective may prefer models that are far less accurate on average, a second metric is proposed to reward higher general forecast performance, but with greater weight on small quantiles:

$$G_2 = \sum_{\rho=2}^{50} (51 - \rho)\mathcal{L}_\rho \tag{6.2}$$

We seek (i) a cohort-wide GP model, and (ii) personalised GP models that will optimise these metrics for sequential forecasting on individual patient's time-series.

6.4 Gaussian Processes for Patient Cohort Modelling

We will first aim to develop a single GP model which can be applied on any of the UPMC patients at any time. This does not require us to pre-specify every aspect of the GP model (from covariance function to hyperparameter values) in advance. However, we do desire a single, time-invariant model-selection or fitting procedure. Several options exists, ranging from full specification of the covariance function *and* hyperparameter values in advance, to specification of the procedure by which to select the kernel and hyperparameters. Complete *a priori* specification

of the GP model would relegate the model to practically no learning, which is an unrealistic baseline comparator. Accordingly, we will aim to learn a single set of (patient-independent) regularisers to help guide the model fitting. We will aim to learn a set of priors with superior forecasting performance.

To this end we select a finite number of covariance functions with a finite number of priors over the hyperparameters. To give the baseline comparator the best chance of performance, we performed an exhaustive combinatorial search of kernel-prior combinations.

6.4.1 Covariance Function for a Cohort-Wide GP Model

The cohort-wide GP model aims to specify a single covariance structure with which to fit (via MAP estimation) any patient's time-series at any point of time on ward.[2]

For this discussion we will restrain our search to covariance functions $k(t, t')$ composed of 1-, 2-, and 3-kernel additive combinations, where each kernel, $k_a(t, t')$, may be either (i) a radial basis function kernel, (ii) a Matérn 3/2 kernel, or (iii) a Matérn 5/2 kernel. That is,

$$k(t, t') = \sum_{a=1}^{A} k_a(t, t' \,|\, \boldsymbol{\theta}_a) + \sigma_n^2 \delta(t, t') \tag{6.3}$$

where $A \in \{1, 2, 3\}$ and each $k_a(t, t')$ may take one of the following three forms:

$$k(t, t') = h_a^2 \exp\left(-\frac{(t - t')^2}{2\lambda_a^2}\right), \tag{6.4}$$

$$k(t, t') = h_a^2 \left(1 + \frac{|t - t'|\sqrt{3}}{\lambda_a}\right) \exp\left(-\frac{|t - t'|\sqrt{3}}{\lambda_a}\right), \tag{6.5}$$

or

$$k(t, t') = h_a^2 \left(1 + \frac{|\mathbf{t} - \mathbf{t}'|\sqrt{5}}{\lambda_a} + \frac{5|\mathbf{t} - \mathbf{t}'|^2}{3\lambda_a^2}\right) \exp\left(-\frac{|\mathbf{t} - \mathbf{t}'|\sqrt{5}}{\lambda_a}\right). \tag{6.6}$$

Regardless of the selected kernel, the respective hyperparameter sets, $\boldsymbol{\theta}$, of a 1-, 2-, or 3-kernel covariance function are:

[2] As described earlier, the GP prior mean will be set to the mean of the training data. Alternative specifications for the mean function were compared for forecasting tasks, but did not yield improvements. We will use MAP inference to accommodate high-granularity sequential forecasting.

$$\begin{aligned} &\text{For } A = 1, \quad \boldsymbol{\theta} = \{h_1, \lambda_1, \sigma_n\}\,. \\ &\text{For } A = 2, \quad \boldsymbol{\theta} = \{h_1, \lambda_1, h_2, \lambda_2, \sigma_n\}\,. \\ &\text{For } A = 3, \quad \boldsymbol{\theta} = \{h_1, \lambda_1, h_2, \lambda_2, h_3, \lambda_3, \sigma_n\}\,. \end{aligned} \tag{6.7}$$

Note that a 1- or 2-kernel covariance function can also be created from a 3-kernel covariance function by fixing the appropriate output-scales h to 0.

The purpose of this is to allow the GP to model up to three different time scales over which a patient's vital-signs may vary. For example, a GP with a single kernel would aim to capture the single most prominent time scale (i.e., length scale) over which vital-signs tend to vary. A GP with more kernels would aim to capture the most important time scale, while also learning shorter-term time scales (to incorporate greater detail) or longer-term time scale (to incorporate longer trends).

As described, this allows for 3 possible combinations for a 1-kernel covariance function, $3 \times 3 = 9$ combinations for a 2-kernel covariance function, and $3 \times 3 \times 3 = 27$ combinations for a 3-kernel covariance function. We wish to further consider multiple combinations of uninformative priors over the hyperparameters of each kernel. We therefore selected a subset of the possible kernel combinations.

The priors over the hyperparameters of these kernels is described below.

6.4.2 Uninformative Priors for a Cohort-Wide GP Model

A common approach to modelling the heterogeneous physiology of patients is to assign an uninformative prior over the hyperparameters of the covariance function. Table 6.1 lists several common uninformative priors, which are available in the GPStuff Toolbox [144].[3] As seen in Table 6.1, the uninformative priors regularise the fit of the hyperparameters by reducing probability of high values. For example, the log-uniform distribution will regularise against large values more stringently than the square-root uniform, but less stringently than the log-log-uniform distribution. From this, we can codify our preference for certain hyperparameters to have higher

[3] Bishop [141] provides further information on uninformative priors in Section 2.4.3 of his book.

Table 6.1: Uninformative priors for hyperparameter regularisation

Prior	Parameterisation
Uniform*	$p(x) \propto 1$
Square root uniform	$p(x^{1/2}) \propto 1$
Log-uniform	$p(\log x) \propto 1$
Log-log-uniform	$p(\log \log x) \propto 1$
* = Unused in experiments	

Table 6.2: Uninformative prior rules

Rule 1	Rule 2	Rule 3
$h > \sigma_n$	$h_i \propto h_j, \quad \forall i, j$	$\lambda_i < \lambda_j, \quad \forall i < j$

value than other hyperparameters without an explicitly specifying of what those values might be.

For example, we may wish to encode the belief that the time-series' total variance contains more signal than noise. We may achieve this by placing a uniform prior on output-scale, h, and a log-uniform prior on noise variance, σ_n. This results in a greater penalty for a large σ_n than an equally large h. When the model is fit h will be larger than σ_n, in the absence of stronger evidence to the contrary.

Following this intuition, combinations of uninformative priors over the hyperparameters can be selected. For the three kernels described above, we consider the rules in Table 6.2 to generate a plausible set of priors. Table 6.3 counts the total number of prior-kernel combinations for covariance functions of 1, 2, or 3 kernels. Prior combinations compliant with rule 2 are a subset of those compliant with rule

Table 6.3: Prior-covariance function combinations

Covariance Function	Rule 1	Rule 2	Rule 3	Total Combos
1 Kernel	3	3 (1)	3	$3 \times 1 \times 3 = 9$
2 Kernels	2	2 (1)	2	$2 \times 1 \times 2 = 4$
3 Kernels	2	2 (1)	1	$2 \times 1 \times 1 = 2$

Table 6.4: Optimal kernel-prior combinations

Kernel Count	Prior
1	$p(\log h) \propto 1$, $p(\log \lambda_1) \propto 1$, $p(\log \log \sigma_n) \propto 1$
2	$p(\log h) \propto 1$, $p(\log \lambda_1) \propto p(\lambda_2^{1/2}) \propto 1$, $p(\log \log \sigma_n) \propto 1$
3	$p(\log h) \propto 1$, $p(\log \log \lambda_1) \propto p(\log \lambda_2) \propto p(\lambda_3^{1/2}) \propto 1$, $p(\log \log \sigma_n) \propto 1$

1 and, therefore, do not add to the combination count.

Due to high inter-patient variability in both values and (more importantly) time-series dynamics of vital-sign time-series, we do not consider parametric priors, since they imply a level of prior-belief that does not exist. It seems unlikely that parametric priors would yield significant performance gains and therefore, such priors are left untested. However, if such a search were desired, the computational expense to test any particular parameterisation across many patients would be very high. We would recommend a search method such as the Bayesian optimisation search described below to efficiently search through the infinite number of possible options.

6.4.3 Selection of Cohort-Wide GP Model

Any set of priors that satisfies all three rules of Table 6.2 for a covariance function of one to three kernels is considered as a kernel-prior pair and was tested. Such an exhaustive search over kernel-prior combinations was deemed reasonable for several reasons since (i) its offline calculation is a realistic representation of its clinical application, and (ii) it is impossible to deduce the effect of the combinatorial interplay between kernels and priors on performance, and therefore their performance must be learned. Finally, (iii) there is no reason to test only a subset of all possible combinations given that, in clinical practice, any cohort-wide model would be developed off-line.

The performance of each kernel-prior combination is assessed by G_1 (in Equation 6.1) and G_2 (in Equation 6.2).

To simplify comparisons, we took a single candidate from each of the 1, 2, or 3-kernel candidate for final comparison. Accordingly, a single optimal set of

uninformative priors was selected for each kernel of size 1, 2, or 3, shown in Table 6.4. The 2-kernel covariance function was the best among the three final candidates.

Extensions of the Cohort-Wide GP Model

These "general" kernels with light regularisation provide a useful foundation for further personalisation. For example, with a smaller number of pre-determined covariance functions, it is possible to simultaneously compare multiple GP models for performance on a patient in real time. This could be considered an intermediate step between a single cohort-wide model, and patient-specific modelling.

Two such examples include:

1. Identifying the covariance function with best performance in the first 24 hours to use for monitoring in the subsequent hours 24-72.

2. Identifying the covariance function with best performance over the last m-minutes to use for the next forecast.

These approaches are described in [1] and [4], respectively, however they did not result in a significant improvement over the more simple rule of modelling all patients via the best-found cohort-wide 2-kernel covariance function. This suggests that uninformative priors have difficulty making use of full complexity of multiple kernels. Further inspection of the MAP hyperparameter estimates of these models suggested that the uninformative priors were highly susceptible to yielding extremely large length scales across all kernel. This suggests that patient-specific modelling may benefit from improved parameter estimation.

6.5 Personalised Parametrisation of GP Models

Solutions for improved personalised inference may include a combination of (i) MCMC integration [134], which offsets the risk of a single poor choice in hyperparameters; and (ii) regularisation via (cohort-wide or personalised) priors over the hyperparameters.

Instead of (i) or (ii), we aim to learn the combination of kernels and hyperparameters $\boldsymbol{\theta}_a$ that best reflect each individual patient's physiology, as measured by Equations 6.1 and 6.2. A further advantage of pre-selected hyperparameters is that the computations of modelling become very predictable, thereby removing the possibility of computational error or fault during the online inference step.

6.5.1 Optimisation of Patient-Specific models

We aim to learn a single covariance function and set of hyperparameters (for each patient), which optimises G_1 and G_2 from Equations 6.1 and 6.2.

Since G_1 and G_2 performance is a function of the GP kernel and its hyperparameters, $\boldsymbol{\theta}$, we denote our objective functions as $G_1(\boldsymbol{\theta})$ and $G_2(\boldsymbol{\theta})$. For simplicity, we omit the kernel argument, and present only the above 1, 2, and 3-kernel additive Matérn covariance functions. That is, we aim to learn A (the number of kernels) and θ for covariance function

$$k(t,t') = \sum_{a=1}^{A} k_a(t,t' \mid \boldsymbol{\theta}_a) + \sigma_n^2 \delta(t,t') \tag{6.8}$$

where the form of each $k_a(t,t')$ is

$$k_a(t,t') = h_a^2 \left(1 + \frac{|t-t'|\sqrt{3}}{\lambda_a}\right) \exp\left(-\frac{|t-t'|\sqrt{3}}{\lambda_a}\right), \tag{6.9}$$

We formalise our learning goal as optimisation problems:

$$\max\ G_1(\boldsymbol{\theta}), \qquad \text{s.t.} \qquad l_d \leq \boldsymbol{\theta}_d \leq u_d \tag{6.10}$$

and

$$\max\ G_2(\boldsymbol{\theta}), \qquad \text{s.t.} \qquad l_d \leq \boldsymbol{\theta}_d \leq u_d. \tag{6.11}$$

The optimisation parameters u_d and l_d represent, respectively, the upper and lower bounds placed on the d^{th} element of $\boldsymbol{\theta}$. For example, for the 2-kernel covariance function with hyperparameters $\boldsymbol{\theta} = [h_1, \lambda_1, h_2, \lambda_2, \sigma_n]$, then selecting the bounds $[l_2, u_2, l_4, u_4] = [2.5, 45, 60, 600]$ would require that the length-scale of the first kernel fall between 2.5 and 45 minutes and that of the second kernel fall between 60 and 600 minutes.

Although bounding the solution space to fall within l and u necessarily relegates us to an optimum no-greater-than that of the unbounded problem, the constraints l and u are valuable to (i) reduce the search space to the most plausible locations of an optimal solution; (ii) prevent overlapping length scales, which reduces kernel complexity; and (iii) prevent querying of $\boldsymbol{\theta}$ that risk computational singularity in $\mathbf{K}$.[4]

These objective functions are non-analytic and must be sampled to locate global and local optima. As illustrated in Figure 6.4(a) and 6.4(d), each query requires sequential fitting and forecasting of the patient's data via the specified GP covariance function and hyperparameters. Since each sequential query produces a distribution of forecast LMLs values (1 value for each time at which a forecast was made), as in Figure 6.4(c) and 6.4(f). The distributions are summarised into a single metric by objective function $G_1(\boldsymbol{\theta})$ or $G_2(\boldsymbol{\theta})$.

Due to the computational expense of each query to $G(\boldsymbol{\theta})$, methods based on gradient descent or line search are undesirable due to their extremely expensive use of multiple function evaluations to decide the next query. We propose Bayesian optimisation to identify the global optimum. We propose random search [145], which is popular for hyperparameter search as a baseline comparator method.

A common critique of publications comparing the performance of two optimisation algorithms is that only the author's "preferred" method is tuned for best performance on the problem at hand. We will describe how both random search and Bayesian optimisation were tuned (among the 43 training-set patients) to maximise their respective performances.

6.5.2 Random Search

Random search (RS) is a popular global optimisation technique because it (i) is trivial to program, and (ii) exploits the low effective dimensionality of many optimisation problems [145]. RS is most commonly implemented as a random uniform sampling of a hyper-rectangle. Common variants of RS are discussed in [146].

[4] For example, a parameterisation with near-zero noise-variance, σ_n, or extremely-long length-scales, λ.

RS may be tuned to our parameter search by sampling more densely where a patient's optimal parameters, $\boldsymbol{\theta}$, are likely to occur. This tuning may be achieved *a priori*, e.g., by changing the uniform sampling distribution to a distribution with modal values. Alternatively, this may be achieved adaptively, by altering the sampling distribution after each query, as in simulated annealing. Using the 43 patients of the training set, several refinements of RS were attempted, including (i) tuning of the random sampling distributions, (ii) iterative sampling of high-performance regions of the sample space, and (iii) simulated annealing approaches. However, these tunings of RS tended to be no better than a uniform random search within a pre-defined hyper-rectangle. Tuning of the hyper-rectangle bounds l_d and u_d involved running several long (1000 samples) random searches with large sampling bounds. The sampling bounds were tightened to include only regions in which patient-specific optima occurred.

The results of random search over 1, 2, and 3-kernel $\boldsymbol{\theta}$ will be described in the training set results.

6.5.3 Bayesian Optimisation

As described in Chapter 4, popular approaches to tune Bayesian optimisation are via (i) the GP prior over $G(\boldsymbol{\theta})$, and (ii) the acquisition function.[5] We tune the GP model to more-appropriately represent our uncertainty about the unexplored areas the search space, given our observations. We tune the acquisition function to more-appropriately represent the wisest choice of the next query, given our uncertainty in the unexplored areas the search space. Both (i) and (ii) allow us to incorporate useful knowledge for a more efficient sampling of the search domain.

GP Prior

All GP modelling for Bayesian optimisation was implemented in GPML [147]. The GP prior over $G(\boldsymbol{\theta})$ is modelled by a Matérn 3/2 Automatic Relevance Determination (ARD) kernel:

[5] The GP inferential step is tunable as well, but this aspect was not explored in favour of MAP estimation of the GP.

$$
\begin{aligned}
&c(\boldsymbol{\theta}, \boldsymbol{\theta}') = \eta^2 \left(1 + \sqrt{3r}\right) \exp\left(-\sqrt{3r}\right), \\
&\text{s.t. } r = \sum_{d=1}^{D} \frac{(\boldsymbol{\theta}_d - \boldsymbol{\theta}'_d)^2}{\nu_d}.
\end{aligned} \tag{6.12}
$$

where η is the output-scale (equivalent to h for the HR monitoring GPs). The distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, r, is now regulated by different a length-scale, ν_d in each dimension of d of $\boldsymbol{\theta}$. Variable r is $|\boldsymbol{\theta} - \boldsymbol{\theta}'|$.

This covariance function has several desirable features:

First, the ARD component allows $G(\boldsymbol{\theta})$ to change over different length-scales, depending on the hyperparameter of the HR-monitoring kernel k_a being varied. For example, vital-sign GP hyperparameters h and σ_n are measured in log-HR bpm, whereas length-scales λ are measured in minutes. Accordingly, a change in σ_n of 0.01 log-HR bpm may induce a substantial change in $G(\boldsymbol{\theta})$ since this would indicate a large difference in the noisiness of a patient's HR. In contrast, a change in λ of 1 minute would induce no change in $G(\boldsymbol{\theta})$, much less a small change of 0.01. The parameter ν_d of Equation 6.12 allows $c(\boldsymbol{\theta}, \boldsymbol{\theta}')$ to vary more quickly in one dimension (i.e. hyperparameter of $\boldsymbol{\theta}$) than the other.

Second, by encoding the Matérn 3/2 form, we are allowing $G(\boldsymbol{\theta})$ to vary more sharply than, say, the infinitely-smooth RBF kernel.

Third, Equation 6.12 does not contain a white-noise component. This means that we expect $G(\boldsymbol{\theta})$ to produce an identical value if given an identical query $\boldsymbol{\theta}$. This is typical in computer simulations.

Acquisition Function

The acquisition function was selected to be Expected Improvement (EI). Small additive offsets to the best-found value, which prompt greater exploration over exploitation were found to improve performance in a subset of the 43 training patients, but resulted in no improvement (but also no loss) for the remainder of the 43 training patients.

Practical Implementation of Bayesian Optimisation

Several further steps were taken to assist in automated Bayesian optimisation search over a large, physiologically-diffuse patient group. These include:

1. Left-censoring of $G_1(\boldsymbol{\theta}) < -3$, for improved data stationarity.
2. Searching over a pre-specified (sufficiently spread-out) grid to avoid a computationally singular Bayesian optimisation GP covariance matrix.
3. Fixing the values of length scales, ν_d, to avoid a singular Bayesian optimisation covariance matrix.
4. Searching over a subset of the dimensions of $\boldsymbol{\theta}$ at any iteration.

These computational considerations are elaborated and justified in Appendix A.

6.6 Training Set Results

The 43-patient training set was used to tune and then compare the respective ability of RS and Bayesian optimisation to identify high-performing values of $\boldsymbol{\theta}$. Both methods were given 250 total queries to $G_1(\boldsymbol{\theta})$ or $G_2(\boldsymbol{\theta})$ to identify an optimal value. An example of these results is shown in Figure 6.5.

As seen in Figure 6.5, RS successfully optimised hyperparameters for a single kernel. However, RS was unable to take advantage of increased modelling complexity as the covariance function progressed from one (×) to two (×) to three (×) kernels. As more kernel hyperparameters were considered, the effective dimensionality of the search space increased from a 3-dimensional search space to a 7-dimensional search space. While 250 queries was sufficient to thoroughly query a 3-dimensional search space by random chance, it was insufficient to sample a larger search space, given the correlation between dimensions.

The shortcomings of RS cannot be over come with computational brute force: When increasing RS to 1000 queries, RS' performance over 1 kernel remained the same or improved slightly over its performance with 250 queries. RS over 2 and 3

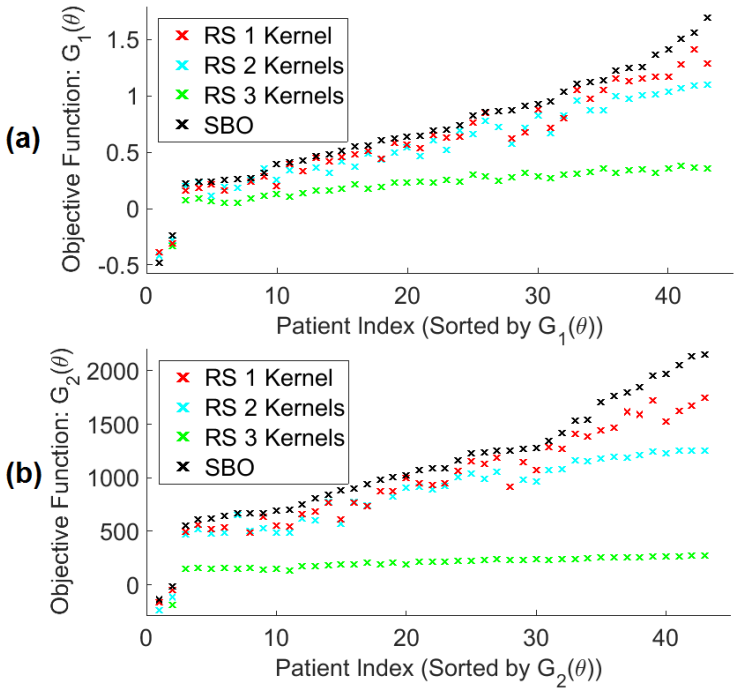


Figure 6.5: Best-found values to optimise (a) G_1 and (b) G_2 for each patient within the training set. A Bayesian optimisation search over G_2 not only identified $\boldsymbol{\theta}$ to maximise (b) G_2 for 30-minute forecasts, but also for tasks it was not optimising, such as (a) G_1 for 60 minute look-ahead. Elements of this figure were published in [1] and are © 2017 IEEE.

kernels improved marginally over its performance with 250 queries, but remained significantly inferior to 1-kernel RS with either 250 or 1000 queries.

Since a single-kernel GP can, effectively, be achieved in 2-kernels and 3-kernels by setting $h_2 = h_3 = 0$, an optimisation algorithm ought to learn (i) to remove the additional kernels by setting the appropriate hyperparameters to 0, and (ii) identify good parameterisations for a single kernel. As the performance of Bayesian optimisation demonstrates, the additional kernels could be used to further optimise the objective function, so long as the hyperparameter space is properly explored.

Bayesian optimisation improved performance as the complexity of the search space increased. Optimising over objective G_1 was sufficient to identify good $\boldsymbol{\theta}$ for both G_1 and G_2.[6] The results proved to be applicable to a wide range of forecast look-ahead windows from 5-60 minutes, suggesting that the solutions were not myopic to a single physiological time-scale.

[6] This suggests that the weighting scheme for G_2 successfully emphasised performance at lower quantiles of forecast likelihood.

For optimisation performance over both G_1 and G_2, for various forecast windows from 1-60 minutes, the Bayesian optimisation algorithm typically out-performed RS in over 90% of patients. A notable exception to this was for forecast windows of 1-5 minutes, in which RS and Bayesian Optimisation performed nearly identically for most patients.[7]

Training set evidence suggests that Bayesian optimisation is superior to RS for the task of identifying patient-specific $\boldsymbol{\theta}$ values. Therefore, the Bayesian optimisation algorithm was selected as the preferred optimiser of patient-specific models. (For reference, the superiority of Bayesian optimisation over RS, was also confirmed in a post hoc analysis in the testing set patients as well.)

6.7 Testing Set Results: Cohort-wide vs Personalised GPs

The performance of cohort-wide GP modelling was compared to GP models with personalised-parameterisations in the 126 test set patients. As described earlier, we simulated the learning process by allowing the personalised GP models access to the first 24 hours of ward data to learn a personalised parametrisation. Forecast performance, as measured by G_1 and G_2, is then assessed in the subsequent hours 24-72, as data is available.[8]

For each patient in the test set, Figure 6.6 shows 6.6(a) G_1 and 6.6(b) G_2 performance of the Bayesian optimisation-found personalised $\boldsymbol{\theta}$, subsequent to the first 24 hours, compared to the population-based uninformative priors. On a patient-by-patient basis, 120 of 126 patients' G_1 performance benefited from personalised modelling. The 6 patients without improvement (2 of which are negligibly different) tended to have less than 2 hours of HR data either before 24 hours (on which to train the personalised model) or after 24 hours (on which to test the personalised $\boldsymbol{\theta}$'s

[7] It is possible that these short windows are comparatively easier to predict and therefore do not require sophisticated methods to identify strong-performing $\boldsymbol{\theta}$.

[8] The runtime of the Bayesian optimisation personalised $\boldsymbol{\theta}$ estimation for a patient with a full 24 hours of data is estimated to be less-than an hour running in Matlab, so a fraction of the forecast LML between hours 24-25 would not have access to the personalised $\boldsymbol{\theta}$ after 25 hours, but, could have a different personalised $\boldsymbol{\theta}$ calculated from some earlier time point.

performance). Most encouragingly, the largest gains in G_1 were made for patients with the worst G_1 under population-based regularisation. These patients would be most likely to generate alarms, and plausibly are those patients with the physiology that is most-difficult to quantify. Patient-specific improvement was as equal to or better than what is shown in Figure 6.6 for forecast depths of 1-45 minutes. Results were slightly worse (than shown in Figure 6.6) at a forecast depth of 60 minutes.

In Figure 6.6(c-d), when comparing forecast LML aggregated across all patients, the forecast LML of the cohort-wide GP models had heavier left-tails. The cohort-wide GP model had hundreds of worst-case forecasts (left-censored at -2), whereas the patient-specific models had tens of instances of worst-case forecasts.

6.8 Conclusion

In this chapter we demonstrated that a single loosely-regularised GP model may be sufficient, although not-optimal, to model a wide range of vital-sign time-series physiologies across a patient cohort. These cohort-wide models are sub-optimal, in part, because they struggle to parametrise more-complex GP covariance functions.

Patient-specific parameters improves the robust forecasting of patient vital-signs, e.g., heart rate. While simple optimisation techniques, such as RS may identify optimal parameters for simple GP models, more sophisticated models are required to successfully parametrise more complex GP models, since the effective dimensionality of the search space is much larger.

Multi-objective optimisation via a vectorised objective function, such as G_2 may be a useful technique to succinctly weight multiple forecasting goals with differing priorities.

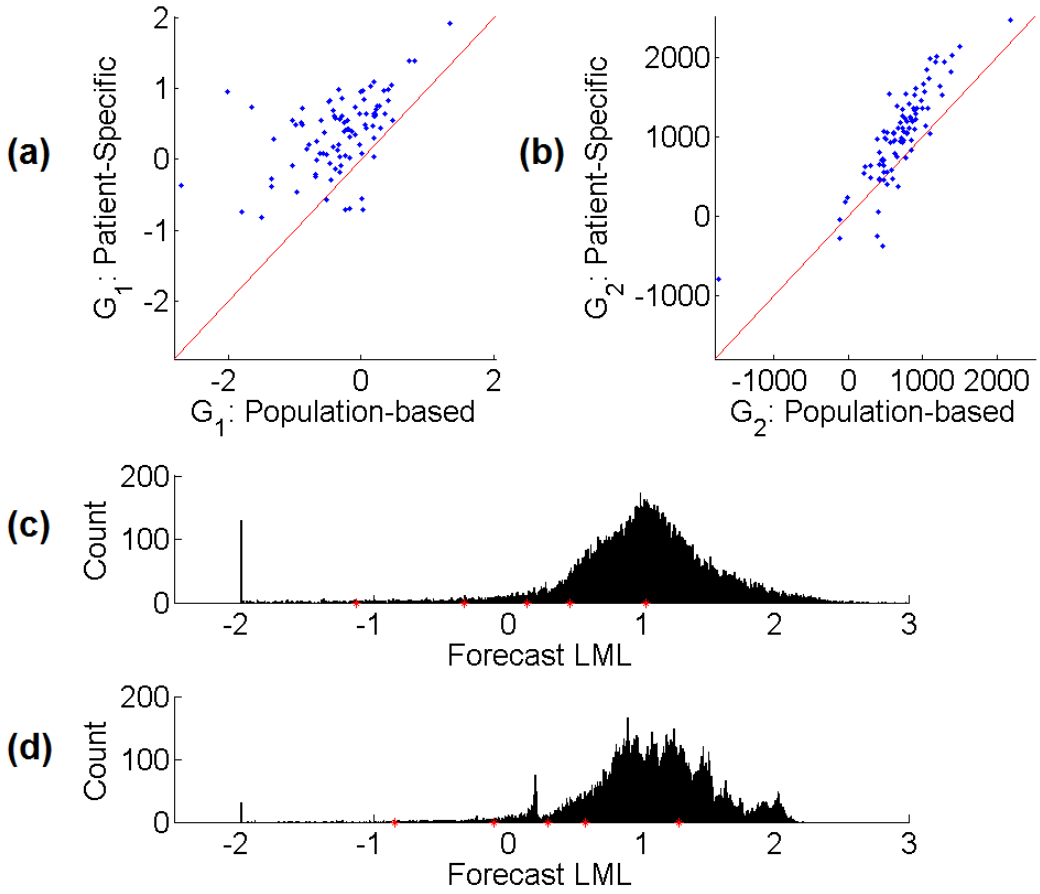


Figure 6.6: Test set performance on cohort-wide and personalised GP modelling. Intra-patient improvement measured by (a) G_1 and (b) G_2 can be seen by personalised modelling. The aggregated forecast performance of (c) uninformative priors can be compared to (d) patient-specific $\boldsymbol{\theta}$'s. Forecast LMLs were left-censored at -2 for visual-clarity. The forecasts LML values in (c) have a significant mode at -2 due to this censoring, where the forecast LML values in (d) do not. For (c) and (d), red points mark the 1, 2.5, 5, 10, and 50-percentiles. Elements of this figure were published in [1] and are © 2017 IEEE.

7

Baseline Comparators for Deterioration Detection

This chapter describes four baseline comparator methods against which Gaussian process (GP) methods will be compared. Three comparator methods are based on heuristic approaches in current clinical practice. The fourth comparator method is based on an FDA-approved empirical approach, which is also in current clinical practice.

Each of the three heuristic baseline comparators are simple thresholding techniques, which represent current clinical practice in early warning score calculations, particularly, those done manually by nursing staff.

The empirical method is a kernel density estimate-based (KDE) approach to novelty detection. As an FDA-approved monitoring algorithm, it represents the current state-of-the-art in continuous multi-parameter vital-sign monitoring.

Each of the 4 comparator methods is evaluated according to its trade-off between two performance metrics: (i) the false positive alarm rate (FPR), and (ii) the time of early warning (TEW). This performance trade-off is assessed for each of the four comparator method on patients from the UPMC data set.

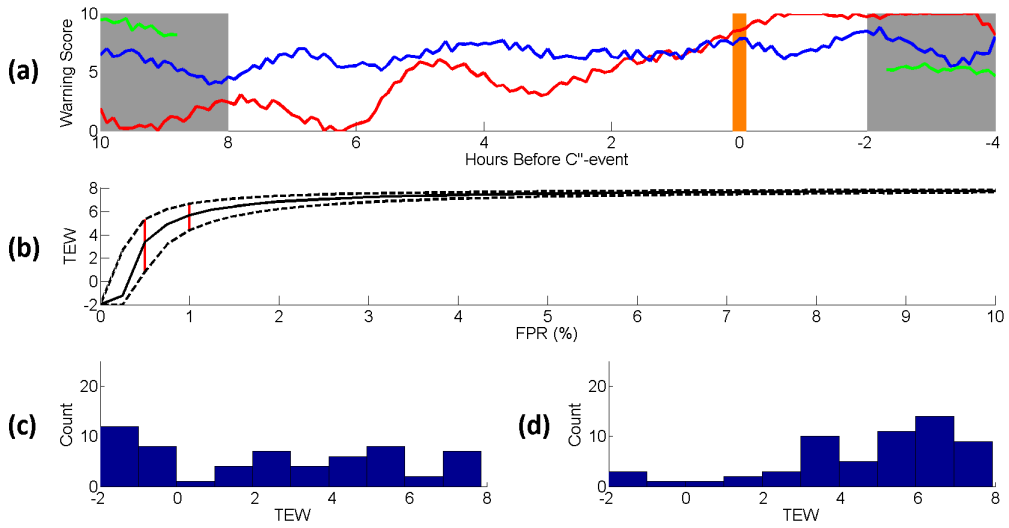


Figure 7.1: The construction of a TEW vs. FPR plot. In (a) the early warning scores of three patients (red, blue, and green) are shown near their respective C”-event, at time 0 (orange). Plausible alarms are considered if they occur within the 8 hours before until 2 hours after the event (white), and not considered if they occur outside of this time period (grey). In (b), for any alarm sensitivity we have a corresponding false-alarm rate in the non-C”-patient set (which we aggregate into a single proportion across all patients) and the TEW across the 59 C”-patient. To visualise the dispersion of TEW values, we plot the 33, 50, and 67 percentiles of the 59 C”-patients. The TEW distributions at two distinct FPR values are shown in (b) in red, and the constituent TEW values are plotted in (c) and (d), illustrating how, for each patient, TEW increases monotonically with FPR.

7.1 Overview of Deterioration Detection

7.1.1 Data

As described in Chapter 3, the UPMC data set held-out 89 non-C”-patients and (all) 59 C”-patients to evaluate deterioration detection. To recap, the first C”-event for each of the 59 C”-patients will be evaluated, since subsequent C”-event data may be influenced by clinical interventions (as a reaction to the first C”-event).

In Figure 7.1(a), we show a generic early warning score (EWS), which are calculated from vital-sign measurements of each of three patients. There is a unique EWS time-series for each of three different patients (red, blue, and green). For each patient, we would anticipate the warning score to (i) be low in the absence of abnormal physiology, and (ii) escalate near to the emergency event at time 0 (orange). We must select a threshold on the EWS to determine when to generate an alarm and identify a patient as deteriorating.

Traditional metrics of classification performance use a confusion matrix, describing the incidence of true positive, false positive, true negative, and false negative predictions. In contrast, the TEW vs. FPR performance metric incorporates both (i) the clinical ambiguity of a patient's time-series prior to deterioration, as well as (ii) the time-value of early alarms. The calculation of a TEW vs. FPR plot is described in Figure 7.1. Alarms on non-C"-patients are false positive alarms. For a generic EWS calculated over time, in 7.1(a), alarms on C"-patients falling in the time period of 8 hours before until 2 hours after the C"-event (in orange) are true positive alarms. A TEW is the time between a C"-patient's first true alarm (within this window) and his first C"-event (in orange). Alarms prior to 8 hours before, or following 2 hours after the C"-event (in the greyed-out region) are not considered due to their ambiguous status. That is, it is less certain that an alarm in this region is specific to the abnormal physiology related to *this* C"-event.

A false negative would be the failure of the EWS to escalate sufficiently to surpass an alarm threshold within the alarmable-window. The TEW of such cases is censored at -2 hours, which is 2 hours after the C"-event at 0 hours. This is the worst possible TEW result. The desired TEW vs. FPR plot of 7.1(b) is achieved by modulating the threshold required to trigger, which, in turn, modulates our alarm sensitivity in the window of 7.1(a), but at the cost of more frequent false positives among non-C"-patients.

At a given alarm threshold, each patient will differ in the TEW due to differing personal physiology in the period surrounding their C"-event. We are therefore interested in the the distribution of TEWs for all 59 C"-patients at any particular FPR. To achieve this, we plot the 33, 50, and 67-percentiles of the TEW distribution at each FPR. For example, at two different FPR values, marked in red in 7.1(b), we can see that the span of TEW quantiles differ since they are drawn from the individualised patient TEWs shown in 7.1(c) and 7.1(d). We are interested in the distribution of TEW because it informs important clinical considerations, such as worst-case performance on the hardest patient cases. Such patients are of special

interest to machine-monitoring applications, since they have the greatest potential to benefit compared to easy-to-identify deteriorating patients.

7.1.2 Evaluating Performance

A challenge posed is that with only 59 C”-patients and 8 C”-event causes, there is insufficient data to have separate training and validation sets for deterioration detection (as we had for the forecast tasks of the previous chapter). It is difficult, then, to both (i) tune a method for optimal TEW vs. FPR performance, and also (ii) produce an unbiased estimate of the method’s performance on a held-out patient set.

In the absence of a large held-out test set, we will underscore several aspects of the methods presented:

First, each early warning method is based on novelty detection [47], and therefore no models are trained with direct reference to TEW in C”-patients. This separates them from many traditional parametric machine learning approaches, in which parameters are directly tuned to perform well on a training set but, necessarily, perform less-well on the held out validation or test set.

Second, these methods have few (or no) tunable parameters to derive an EWS. This means fewer avenues by which the method’s TEW vs. FPR performance may improve with respect to the training set, but also fewer ways in which it may fail to generalise to the test set.

Third, in addition to few tunable parameters, the parameter themselves have fewer “reasonable” values, and very little inter-patient variability in the optimal value of those parameters. This means that any choices for parameter values tend to be the same regardless to whether a subset of patients is held out and that optimal values for one patient tend to have robust performance for other patients.

For the methods discussed in this and the next chapter, we will highlight the aspects above to illustrate how these models more readily generalise, without loss of performance.[1]

[1]Similar results were demonstrated in the Chapters 5 and 6 as well. In Chapter 5, the artefact scoring algorithm was sufficiently personalised that it experienced very little inter-patient variability, which would cause inferior performance on a held-out patient set. Similarly, in Chapter 6, the Bayesian optimisation methods performed better (compared to the alternative methods) in

7.2 Selection of Comparator Methods

There are many methods from which to select a baseline comparator. As discussed in Chapter 2, most early warning systems have a quantitative score but take either an empirical or heuristic approach to derive that quantitative score. To represent heuristic methods (most commonly used in hospitals) we select (i) a single-vital-sign thresholding method (also known as a "triggering system"), and (ii) a multiple-vital-sign thresholding method (also known as a "scoring system"). To represent the technical state-of-the-art in empirical patient monitoring, we select (iii) a KDE-based novelty detection algorithm.

Each method escalates a patient's EWS, with respect to their vital-sign measurements in a different way, as shown in Figure 7.2 for univariate vital-sign changes, and in Figure 7.3 for jointly-varying vital-signs. Each system differs in the granularity of alarmable values. For example, the NEWS-based system scores each vital-sign from a value of 0 to 3. This means (for an individual vital-sign) that there are four alarmable values (including an alarm at 0, which would induce a 100% FPR). The four alarmable values also implies that there are only four distinct points along the TEW vs. FPR curve, from which to select a preferred trade-off. Systems with a greater number of alarmable values would, in turn, accommodate a more granular range of setting along the TEW vs. FPR curve.

The KDE values (in both Figures 7.2 and 7.3) are scaled and right-censored to fall between 0 and 3, for comparability to the trigger and NEWS-based approaches. Escalations of the KDE-based score are continuous-valued and based on data from a set of healthy patients. This is why (i) small fluctuations in warning scores can be within the range of highly-normal values, and (ii) each vital-sign is not given the same maximal warning value (as is done with NEWS-based methods).

The variation of warning score over multiple vital-signs in Figure 7.3 is even more informative. In each plot, the warning score is gradated from low-warning regions (dark blue), which would raise no medical alarm, to high-warning regions (dark red) which would be certain to raise a medical alarm. As can be seen in the NEWS-based

the testing set than in the training set, which is atypical in most machine learning settings.

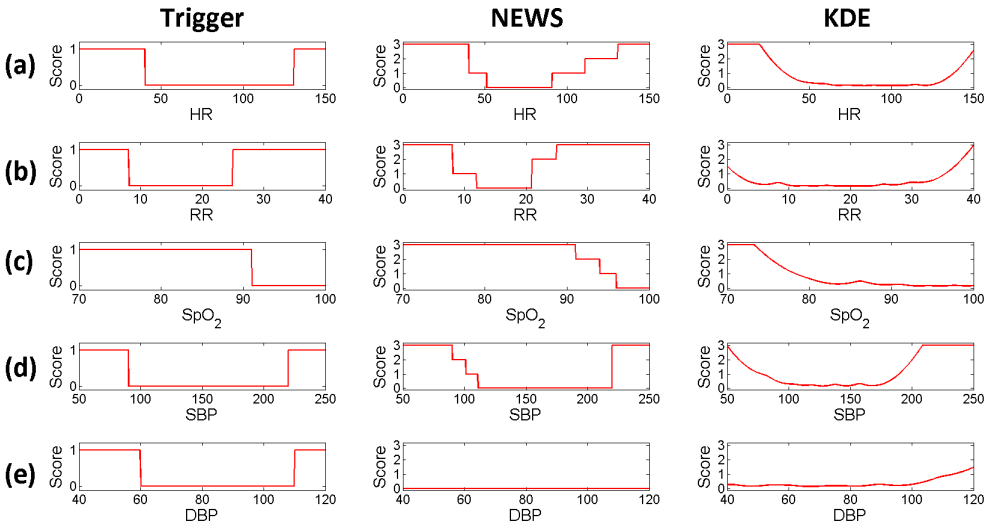


Figure 7.2: Escalation of warning score by vital-sign in the trigger, NEWS-based, and KDE-based early warning system. Vital-signs are ordered by row, as (a) HR, (b) RR, (c) SpO_2, (d) SBP, and (e) DBP. Notably, each method escalates the score of vital-signs with extreme measurements, however, each differs in the granularity of the escalation. NEWS has no escalation with regard to DBP since DBP is not included in the NEWS calculation.

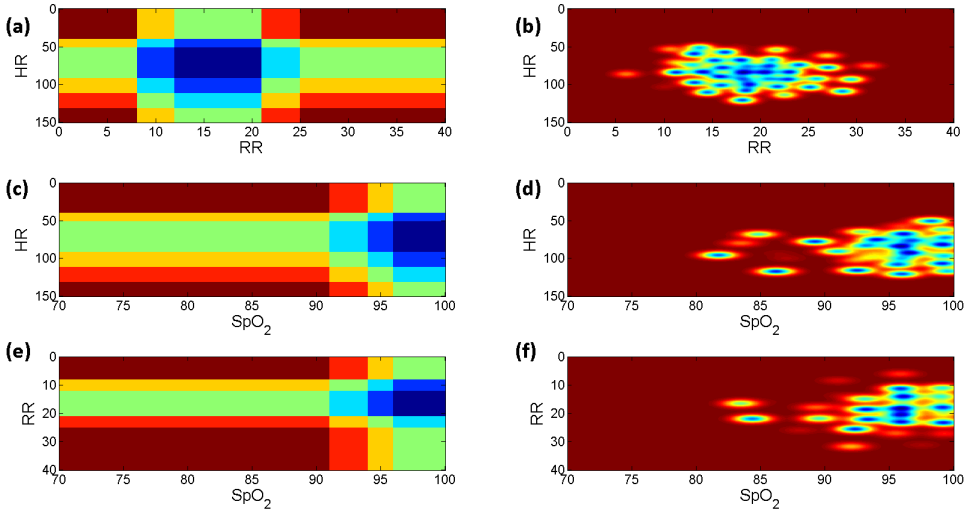


Figure 7.3: Multivariate escalation of NEWS-based (left-column) and KDE-based (right-column) warning scores in HR, RR, and SpO_2. In each plot, the warning score is gradated from low-warning regions (dark blue), which would raise no medical alarm, to high-warning regions (dark red), which would be certain to raise a medical alarm.

escalations of 7.3(a,c,e), the joint abnormality of vital-signs is not incorporated, for example, to acknowledge correlated vital-sign measurements. In contrast, the data-derived KDE method in 7.3(b) consider high HR values to be less alarm-worthy when RR is high as well, since the KDE acknowledges a positive correlation between HR and RR in the patient population. Furthermore, in 7.3(d,f), a highly-alarming score is present in nearly all of the region where $SpO_2 < 90$. In contrast, the NEWS-based system considers, for example, $SpO_2 < 90$ to be less-alarm worthy when HR and RR fall into normal ranges, than when they are extreme as well.

However, we can also see plausible short-comings of such a data-driven method as well: For example, there are approximately 5-6 low-warning regions (blue) apparent in 7.3(d,f) where SpO_2 is less-than 90. While this, in itself, may not be inappropriate given the data, the fact that these regions are separated from other low-warning regions by high-warning (dark red) regions suggests that this is a more likely a facet of the data on which the KDE was trained, instead of an apt description of alarm-worthy regions.

Heuristic Comparators

We consider both the trigger and scoring systems to be heuristic because the selection of decision thresholds for such methods are largely heuristic.

For the trigger system, we consider each of the five available vital-signs in the UPMC data set (HR, BR, SpO_2, SBP, and DBP).

For the scoring system, we use the thresholds provided by NEWS (for the available vital-signs). NEWS is selected over alternatives, such as MEWS, because it incorporates the fewest proportion of parameters that are *not* included in the UPMC data set. Since the aggregate warning score is not identical to that of the NEWS score, we call it a "NEWS-based" system. Scoring system thresholds may be tuned, depending on the patient cohort particular to the ward, and therefore the NEWS-based comparator may not be identical to the system currently preferred by some SDU clinicians. However, as shown in Chapter 2, different scoring systems do not vary greatly in their selected thresholds. We do not anticipate large

differences in performance between the selected NEWS-based comparator and the existing alternatives.

Empirical Comparators

For empirical comparators, it is reasonable to believe that distinct empirical methods may differ substantially in performance. (In fact, that assumption underlies this thesis.) With many alternative methods from which to select, the KDE-based method has several attributes to recommend it:

First, KDE-based models have been studied extensively in relation to the UPMC SDU, where this data set was collected. This creates significant transparency in terms of best-practice, implementation, and model design, compared to other methods, e.g., APACHE scores for which the parameters are propriety and therefore kept secret.

Second, an FDA-approved version of the KDE-based method was developed and is subsequently implemented in the UPMC SDU. This indicates that the KDE method represents a clinician-preferred algorithm from among the many that are available. From this, the KDE is a realistic representative of an empirical system that may be on the ward.

7.3 Heuristic Comparator: Trigger System

Extreme values of a single vital-sign parameter is indicative of patient deterioration. Trigger systems typically designate upper and lower limits on the "healthy" range for each vital-sign. Outside of this range, a clinician should be informed. This means that for a single vital-sign two thresholds must be selected: an upper threshold and a lower threshold.

In the simplest case, we select only the upper threshold or only the lower threshold for each vital-sign. An alarm is sounded when the upper or lower threshold is exceeded.[2] Since, unlike NEWS, a standard trigger threshold is not published, we will allow the trigger threshold to vary for each vital-sign.

[2]Typically, the exceedance must persist for several minutes before an alarm is generated. Such a criterion was tested for each baseline method but did not improve performance. Further discussion can be found in Appendix C: Alarm Hold Criterion.

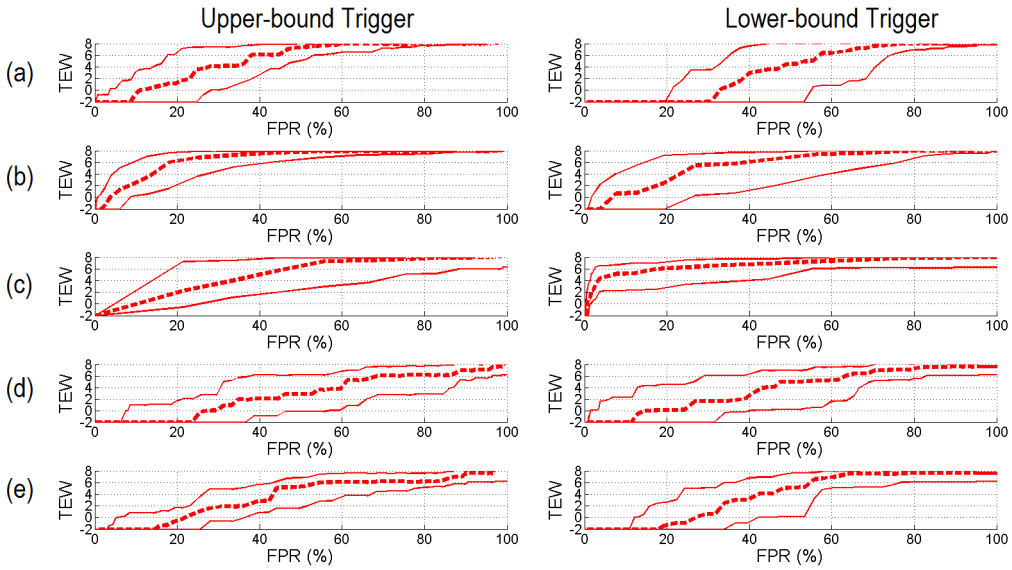


Figure 7.4: TEW vs FPR trade-off for univariate upper-bound and lower-bound trigger systems in (a) HR, (b) RR, (c) SpO_2, (d) SBP, and (e) DBP. The upper-bound trigger on SpO_2 is included for completeness. TEW and FPR are modulated by the upper or lower bound on each vital-sign that triggers an alarm. Each line represents the 33, 50, and 67 percentile of TEW for the respective FPR.

The calculations of TEW and FPR (described earlier in Figure 7.1) are shown for upper-bound and lower-bound triggering systems in Figure 7.4. An upper threshold alarm on SpO_2 is shown for completeness but is not considered to be of interpretable value. Note that the FPR of upper and lower bounds are equivalent to the survival function $1 - F(y)$ and cumulative density function $F(y)$ respectively, of measurements of the 89 non-C"-patients. Visually, the superior trade-off is as close to the upper-left corner as possible (similar to an ROC curve).

Unsurprisingly, the upper threshold on HR and RR, along with the lower thresholds on RR and and SpO_2 seem to have the best trade-off between TEW and FPR. This corroborates our understanding that measurements from a single vital-sign, which (i) act as a bellwether for abnormality across multiple vital-signs, and (ii) are reliably acquired, may provide a sound foundation for monitoring to be as continuous as possible.

As a final note on this simplest of thresholding methods: When viewed in light that (i) the clinical staff identified only 7 of 112 deterioration events, (ii) of those 7 patients, at least 2 were identified too late to save and (iii) alarm fatigue in

Vital-Sign	3	2	1	0	1	2	3
HR (bpm)	≤ 40		41 - 50	51 - 90	91 - 110	111 - 130	≥ 131
RR (bpm)	≤ 8		9 - 11	12 - 20		21 - 24	≥ 25
SpO_2 (%)	≤ 91	92 - 93	94 - 95	≥ 96			
SBP (mmHg)	≤ 90	91 - 100	101 - 110	111 - 219			≥ 220

Figure 7.5: NEWS scoring table used to create (i) the NEWS-based multivariate scoring baseline comparator, and (ii) a univariate trigger system baseline comparator which incorporates both high and low values on which to trigger.

rampant across clinical practice, it is reasonable to assert that a result, such as any subplot in Figure 7.4, may represent an upper-bound performance on current clinical practice. In terms of false alarms, there is little reason to believe that these trigger systems differ significantly from trigger systems in clinical practice. So many missed deterioration events in the original study suggests that the sporadic bedside checks are too infrequent to identify deterioration in a timely manner, if at all. There is value in continuous monitoring algorithms simply because they are continuous, regardless of whether the monitoring algorithm is particularly sophisticated.

Clearly, alarming on a single vital-sign in a single direction may only detect a minority of alarm-worthy vital-sign measurements. Trigger systems may incorporate both an upper bound and a lower bound on each vital-sign. However, this creates two tunable parameters from which to generate the binary alarm score. To create a single score from both high and low values we use the NEWS scoring thresholds for each individual vital-sign, which are shown in Figure 7.5. From Figure 7.5, we can see that each vital-sign is assigned a value from 0 to 3, depending on how high or low the value is. By alarming on this score, instead of the vital-sign value itself we may examine a trigger system which alarms on both high and low values.

Compared to the single-direction trigger system (shown in Figure 7.4) the univariate scoring systems (shown in Figure 7.6) results in a significant reduction of granularity in the TEW vs. FPR plot, as each vital-sign now has (at most) 4 alarmable thresholds. Overall performance in TEW vs. FPR trade-off is reduced as well. The low granularity could be deduced *a priori* from the system having only 4 alarmable values (that is, only 4 achievable settings along the TEW vs. FPR curve), however this effect can also be seen in Figure 7.6, with each curve

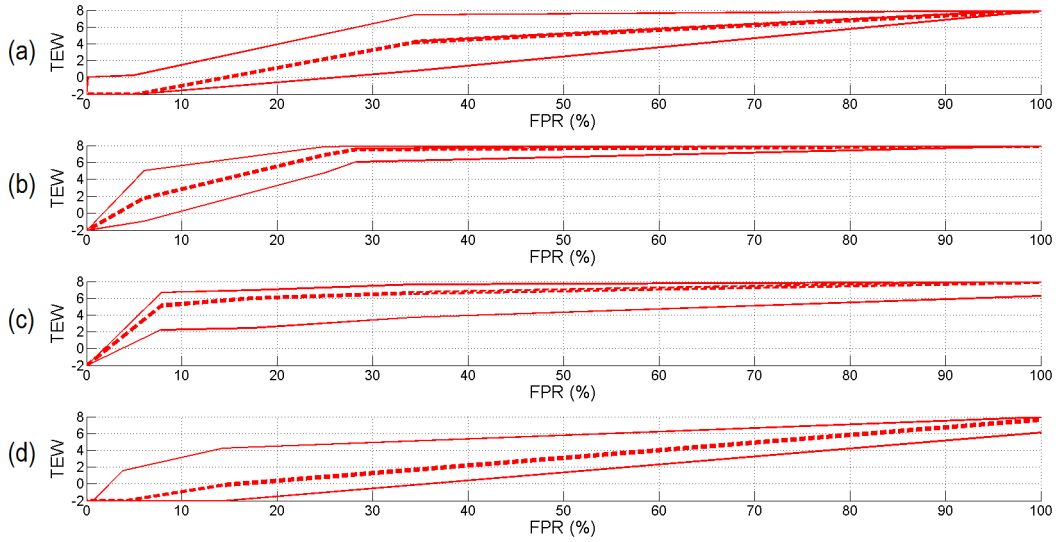


Figure 7.6: TEW vs. FPR trade-off of an univariate trigger system with upper-bound and lower-bound scoring based on NEWS. Vital-signs are (a) HR, (b) RR, (c) SpO_2, and (d) SBP. DBP is not included since it is not included in NEWS scoring. TEW and FPR are modulated by the upper on each vital-sign's NEWS-score, at which an alarm is triggered. Each line represents the 33, 50, and 67 percentile of TEW for the respective FPR.

having only 3 inflection points. The clinical inference on both high and low values is insufficient to overcome the value of an empirically selected threshold in either direction. Worse yet, the low-granularity creates a TEW vs. FPR trade-off that largely falls outside a clinically viable FPR around 0%-5%.

Alternative methods that would retain this granularity could include (i) fixing either the upper or lower alarm bound as a constant, while allowing the other (upper or lower) alarm bound to vary, or (ii) simultaneously modelling the irregularity of high and low values as tail behaviour of some reference distribution. Both approaches are more empirical than both of the presented trigger systems. However, alternative (i) is computationally involved without addressing the trigger system's short comings. Alternative (ii) is, effectively, a univariate version of the KDE method described in the Section 5 of this chapter.

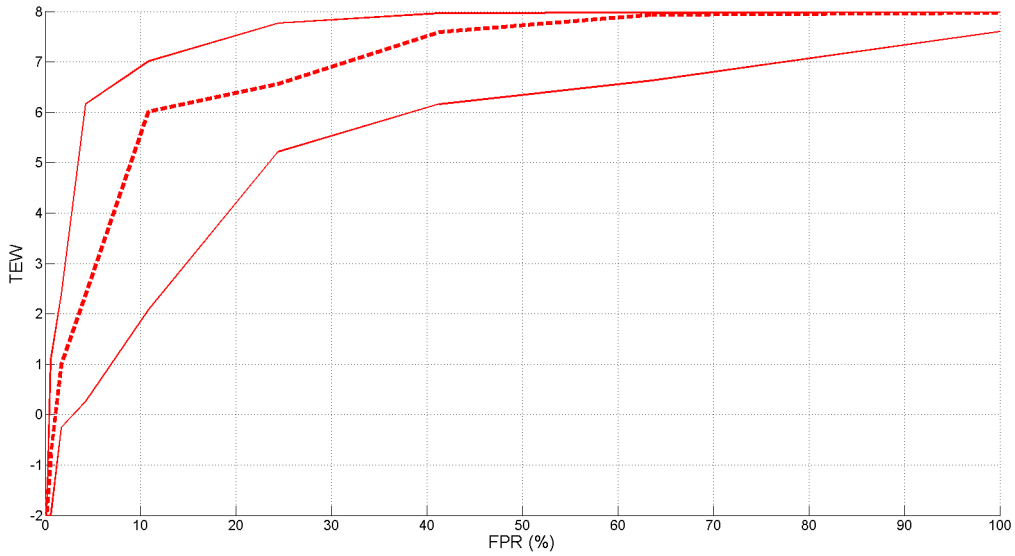


Figure 7.7: TEW vs FPR trade-off of a NEWS-based scoring system. TEW and FPR are modulated by the upper threshold on the total NEWS-score (summed across the individual HR, RR, SpO_2,and SBP score from Figure 7.5), at which an alarm is triggered. Each line represents the 33, 50, and 67 percentile of TEW for the given FPR.

7.4 Heuristic Comparator: Scoring System

Patient deterioration can manifest in different vital-signs at different times. Optimal monitoring should therefore consider multiple vital-signs for signals of deterioration. More concretely, an early warning system should identify (i) large abnormalities in individual vital-signs, as well as (ii) smaller joint abnormalities across multiple vital-signs. As seen in Chapter 2, current EWS attempt to achieve both (i) and (ii) by summing warning scores from individual vital-signs.

The NEWS-based scoring system in Figure 7.5 is implemented to simulate such a system. Since the UPMC data set lacks several decision variables incorporated in NEWS, only the available vital-signs are used for the calculation. Figure 7.7 shows the TEW vs. FPR plot of the NEWS-based comparator. Noticeably, the performance improvement of the five-vital-sign scoring system is marginal compared to the single vital-sign, single-threshold system. As stated before, this representation of NEWS performance is likely an over-estimate of the NEWS system in current practice since no time is lost due to sporadic in-person monitoring.

7.5 Empirical Comparator: KDE Model

A KDE-based model of patient normality represents the current technical state-of-the-art in empirical methods. The UPMC data set was first collected in order to train and evaluate such a model in 2008. As a testament to its success, KDE-based monitoring was retained in the UPMC SDU after the conclusion of the study, and KDE methods continue to generate publications using UPMC SDU data. The KDE-based model, Viscensia, was FDA-approved in 2012.

A review of KDE modelling can be found in Chapter 4. The KDE-based model for patient monitoring finds its motivation in probabilistic novelty detection [47]. As shown in Figure 7.8, the KDE models the joint distribution of the vital-signs from a "healthy" patient group. The novelty of a new measurement is quantified by the inverse-log-likelihood of the new measurement with respect to the KDE.

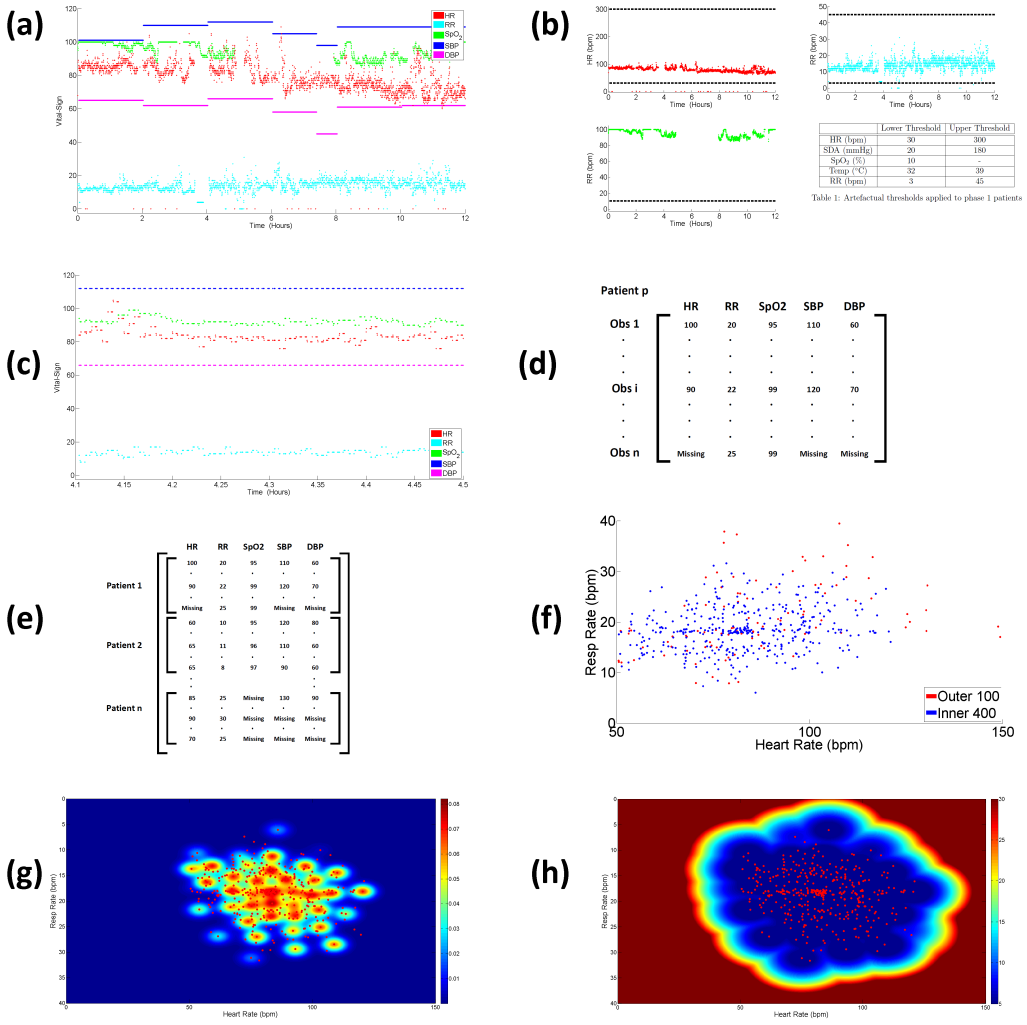
We examine two versions of the KDE model:

1. The **baseline KDE** models the joint distribution of HR, RR, SpO_2, SBP, and DBP using all available data in the 174 training set patients. This KDE model replicates the modelling process in Hann [45], which is the only model described in that thesis.

2. The **KDE with vital-sign volatility features** uses the original five vital-signs but includes a further three features quantifying short-term volatility in HR, RR, and SpO_2. These features are the standard deviation of HR, RR, and SpO_2 over a short time window of the last m minutes. This method attempts to incorporate abnormal vital-sign volatility into the deterioration detection process, which is not captured in the baseline model.

7.5.1 Baseline KDE Model

Construction of the baseline model in Hann [45] is described over the length of a chapter in his thesis, and is outlined in Figure 7.8, with further details in Appendix B. The multi-step process involves 7.8(b) data cleaning, 7.8(c) alignment and transformation of data, imputation and/or removal of missing data, 7.8(f)



	Lower Threshold	Upper Threshold
HR (bpm)	30	300
SDA (mmHg)	20	180
SpO_2 (%)	10	-
Temp (°C)	32	39
RR (bpm)	3	45

Table 1: Artefactual thresholds applied to phase 1 patients

Figure 7.8: Creation of a KDE-based novelty score. Each patient's (a) 5-vital-sign time-series is first (b) cleaned via artefact removal of any measurements that exceed a pre-determined artefact threshold, as described in Table 1 of Chapter 3: Data Description. The clean vital-sign data is then (c) aligned via their time-stamps so that each unique vital-sign measurement may comprise a unique 5-dimensional data point. These aligned data points are then (d) collated into a matrix of IID observations, which, in turn, are (e) collated across all training set patients. In (f), the multitude of IID measurements are clustered into 500 vital-sign centroids, with the outer-most 100 clustered removed as a form of data cleaning. (g) A KDE is fit to the remaining 400 clusters, which, in turn, is (h) converted into novelty score.

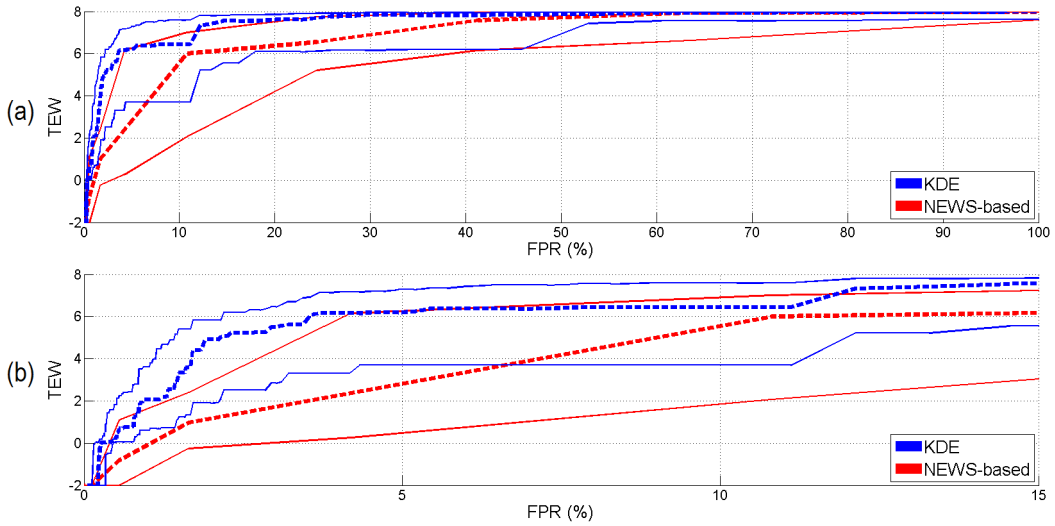


Figure 7.9: TEW vs. FPR trade-off of KDE-based early warning system displaying (a) the entire FPR range from 0% to 100%, (b) within the more clinically-actionable range of 0% to 5% FPR. TEW and FPR are modulated by the upper threshold on the KDE-based novelty score, at which an alarm is triggered. Each line represents the 33, 50, and 67 percentile of TEW for the respective FPR.

generation of representative clusters in the data, and further cleaning of those clusters - before 7.8(g) finally fitting a KDE. A detailed description is in "Appendix B: Construction of Kernel Density Estimate Model of Patient Normality".

In Figure 7.9, the trade-off between TEW and FPR of the first KDE model is shown. Without the need for any expert tuning of thresholds (as required for the expert-tuned NEWS system thresholds), the KDE clearly outperforms the heuristic trigger and scoring systems representative of current practice. In particular, the KDE-based system not only outperforms the heuristic methods in terms of area-under the curve, but also accommodates a much higher granularity, compared to the NEWS-based methods which have only a small number of triggerable settings.

7.5.2 KDE Model with Vital-Sign Volatility Features

The baseline KDE provides an EWS based on the magnitude of patient vital-signs. However, we may also be interested in the clinical implications of erratic vital-sign volatility, even when those vital-signs fall within normal magnitudes. For example, the step-change detection algorithm, developed in Chapter 8, demonstrates high TEW vs. FPR performance without any reference to magnitude, just volatility.

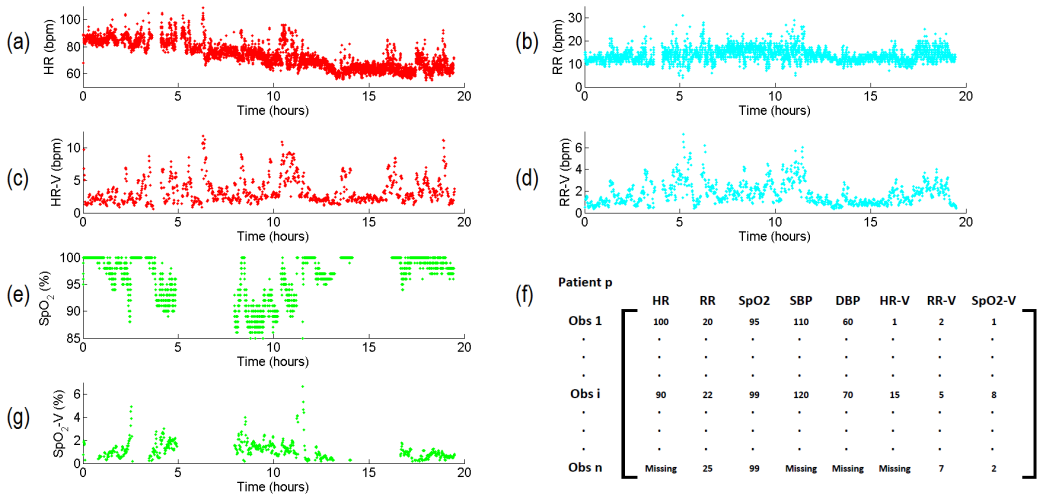


Figure 7.10: Features of the 8-feature KDE model. For (a) HR, (b) RR, and (e) SpO_2 a volatility metric is calculated as the standard deviation of the vital-sign over the last 5-minute. This provides 3 additional features: (c) HR-V, (d) RR-V, and (g) SpO_2-V. The volatility feature of vital-signs SBP and DBP (not shown) are not calculated due to the infrequency of their measurement. In (f) the new aligned vital-sign data includes HR, RR, SpO_2, SBP, DBP, HR-V, RR-V, and SpO_2-V. This matrix may replace the aligned patient data in 7.8(d) to create a KDE over all 8 vital-sign features.

To incorporate this volatility dynamic into our KDE model, we take the original patient vital-sign time-series (post-artefact removal) and create 3 new time-series: HR volatility (HR-V), RR volatility (RR-V), and SpO_2 volatility (SpO_2-V). Figure 7.10 shows this, using the original HR, RR, and SpO_2 time-series in 7.10(a,b,e), and to derive the volatility features HR-V, RR-V, and SpO_2-V in 7.10(c,d,g). The 3 new features are appended to each patient's aligned data set, as in 7.10(f). After replacing the 5-feature matrix in 7.8(d) with the 8-feature matrix in 7.10(f), the remainder of the KDE development was identical to that described in Figure 7.8 for the baseline KDE (i.e., vital-sign alignment, scaling, k-means clustering, etc.). The new KDE describes the joint-distribution over 8 random variables: the five original vital-signs and the three derived volatility features.

The additional volatility features did not improve performance with respect to the TEW vs FPR trade-off. Several windows of varying length m-minutes were tested but with little change to performance. There are several reason why the KDE's performance might be unresponsive to the additional features. This will be left for the discussion at the end of the chapter.

7.6 Discussion

From the TEW vs. FPR figures, it is clear that the empirical approaches to deterioration detection clearly outperform the heuristic threshold-based methods. However, the KDE models have limitations as well: By virtue of being IID models, they struggle to incorporate informative time-series dynamics. Attempts to address this by simply adding more features did not help, in part due to the challenge of estimating density of high dimensions, which hampers model specification in areas with little training data. In 8 dimensions, it's possible that the KDE is over-sensitive to even small deviations from the KDE's centroids (such as those shown in Figure 7.8(f).

7.7 Conclusion

In this chapter we present three baseline comparator methods (five total implementations). Each patient monitoring method represents a method likely to be found on the SDU, such as the UPMC SDU where the dataset was collected. The shortcomings of these baseline methods, from both a technical and clinical perspective are discussed in detail in Chapter 1 and 2.

The heuristic trigger system and the scoring system demonstrate a trade-off between alarm granularity and alarm sensitivity. The KDE method improved significantly over these heuristic methods in terms of TEW vs. FPR trade-off. The addition of further features failed to improve KDE performance. Furthermore, it is clear that empirical systems may outperform heuristic methods, even in the absence of any expert tuning. It is not surprising then, why clinicians may prefer empirical systems if they are (i) transparent, and (ii) available.

These methods will be compared to the performance of GP-based models in Chapter 8.

8

Bayesian Gaussian Processes for Identifying the Deteriorating Patient

In this chapter we demonstrate how Gaussian Process (GP) models can forewarn patient deterioration by identifying informative time-series dynamics. We describe several GP approaches to forewarn deterioration, followed by a more in-depth coverage of GP modelling for step-change detection.

The clinical value of step-change detection for early warning relies on the tendency of vital-sign time-series to exhibit step-changes long before they exhibit extreme values. Since clinical inference is performed over likelihood metrics, it is simple to extend the method, e.g., to include multiple vital-signs or time-scales. Step-change detection monitoring of only a single vital-sign (e.g., HR, RR, or SpO_2) can match or outperform the 5-vital-sign KDE method, in terms of time of early warning (TEW) vs. false positive rate (FPR) trade-off. Further extensions beyond univariate step-change detection build on this high performance of maximising the time of early warning while reducing the rate of false alarms.

Elements of this chapter have been published in [4] and [5], and patented in [6]. Further applications of the work described in this chapter have been used in [8], [9], and [10].

8.1 Overview

GPR Approaches to Deterioration Detection

Significant effort has been dedicated to formalise and automate (i) our belief in the measurements acquired from the bed-side monitors (Chapter 5), and (ii) our belief in inference of the GP fit to data (Chapter 6). The reliability of these tools now allows us to automate the relay of warning signs to clinicians for further inspection.

As illustrated in Figure 8.1, there are many possible warnings that we may wish to bring to a clinician's attention via GP inference on vital-signs.

For example, a GP can forecast the probability of a vital-sign time-series' future values. This allows for the application of probabilistic reasoning to questions such as whether the vital-sign will 8.1(a) exceed the thresholds of of a trigger system, or 8.1(b) achieve a particular EWS, for example those defined by the thresholds of the NEWS scoring system. Compared to sporadic monitoring of vital-signs, the GP-based approach allows for a transparent and principled method to handle the trends and noisiness of vital-sign measurements. Both 8.1(a) and 8.1(b) use GP modelling to infer a patient's deviation from a population-norm.

Alternatively, we may wish to 8.1(d) compare a segment of a patient's time-series to segments from a dictionary of healthy patients. This method would incorporate both (i) the magnitude of values, like current systems, as well as (ii) time-series dynamics, which are currently ignored.

Noting that homeostatic mechanisms should react to personally-abnormal vital-sign values, we may also focus on 8.1(c) unusual dynamics. In particular, if we interpret abnormally rapid increases or decreases of a particular vital-sign as evidence of homeostatic reaction, then we may circumvent the need to learn or incorporate absolute measures of vital-sign abnormality.

Selection of Step-change Detection

Within the UPMC data-set, it is rare for vital-signs to gradually degrade into abnormality. This is, perhaps, unsurprising in light of the large number of C"-events

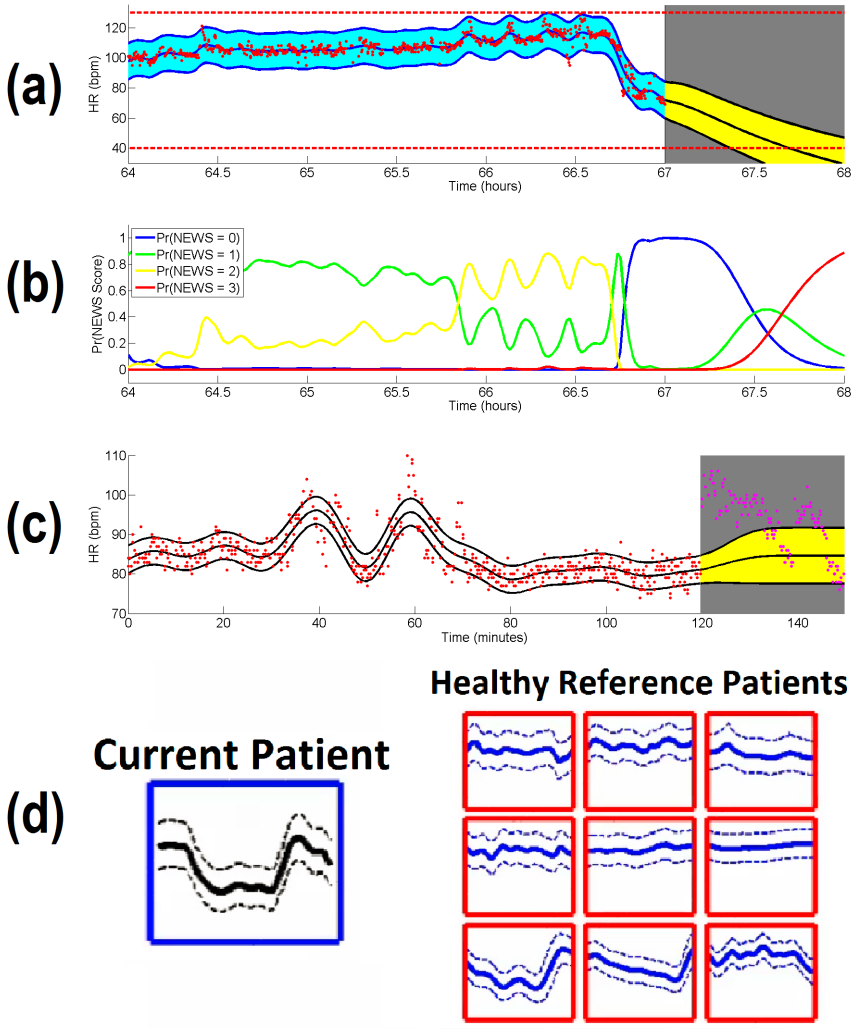


Figure 8.1: Gaussian process inference for deterioration detection. In (a) a GP fits a patient's time-series and identifies the sharp downward trend in HR, and forecasts the the probability that HR will fall below the 40 bpm threshold within the next hour, which would trigger a clinical alarm. (The GP's prior mean is set to the last measurement in the training window.) On the same patient, this same GP fit and forecast can be used (b) to estimate the probability of the HR achieving a particular NEWS early warning score. On a different patient, in (c) the GP forecast may be used as a step-change detector, to quantify the deviation of a patient's vital-sign trajectory from its expected trajectory. As an alternative to comparing vital-signs at only a single point in time, in (d) a segment of the patient's time-series is compared to a dictionary of healthy patients' time-series.

that were undetected in clinical practice. When these predictable degradations do occur, they are typically too close in time to the emergency event to provide actionable early warning. This suggests that GP applications such as in Figure 8.1(a) and 8.1(b) are unlikely to garner significant gains in terms of early warning of deterioration, since there is rarely a prolonged period of evidence for extreme values before they occur.

The comparison of a current time-series to a dictionary of reference patients, as in Figure 8.1(d) can provide early warning gains over currently available methods. Such a method was patented in [6] and may be thought of as an expansion of the KDE methods of Chapter 7 into the personalised probabilistic time-series domain. Furthermore, the method is extensible and transparent in its decision criterion. Although time-series matching can easily be run in real-time, it does require significant memory to hold the reference dictionary.

This chapter will focus on step-change detection methods, as illustrated in Figure 8.1(c). These methods have been published in [5] and [4], and may be extended with minimal computational effort. This makes such methods a realistic contender for clinical implementation across a variety of clinical environments (e.g., both those with and without significant computational resources). The remainder of this chapter will demonstrate that step-change detection is simple, flexible, and extensible from a technical standpoint. More importantly, from a clinical standpoint, such methods are transparent and interpretable for real-time inspection by the clinician.

8.2 Step-Change Detection

Figure 8.2 illustrates how step-change algorithm sequentially fits and forecasts the future distribution of 8.2(a) BR, 8.2(b) SpO2, and 8.2(c,d) HR values. When the future values are consistent with the prediction, as in 8.2(d), the corresponding LML, as shown in 8.2(h) will be high. However if the future values are not consistent with the forecast distribution, as in 8.2(c), then the corresponding LML will be lower, as in 8.2(g). It is reasonable to present the various LML values within a forecast window via a summarising statistic, such as the mean. This also mitigates

the affect of occasional outlying measurements. Since conventional alarm scores are high in the present of abnormal physiology, we will measure step-change warning scores in terms of negative log marginal likelihood (NLML).

Since the step-change detector forecasts over a time-window that contains one NLML value per vital-sign measurement within the window (as seen in Figure 8.2(g) and 8.2(h)), the step-change detector's tunable parameters include (i) the metric to summarise NLML values within the forecast window, and (ii) the time-length of the window. As discussed in Chapter 6, the selection of a GPR model to derive these forecasts is a further tunable element. We will aim to demonstrate that step-change methods are robust to many of these choices or, at the very least, that they out-perform alternatives no matter how these parameters are tuned. To reduce research degrees of freedom, we will relegate our choices to the simplest and most obvious choices for these tunable parameters. We will examine the mean NLML of a 1-minute forecast window. To model HR time-series we will use the two-kernel covariance function (described as the best cohort-wide covariance function) from Chapter 6. Using the same cohort-wide selection procedure described in Chapter 6, we will use a single-kernel Matérn 3/2 covariance function to model the time-series of RR and SpO_2.

In Figure 8.3, we show how step-change metrics may be used as a continuously-monitored warning score in the same manner as the NEWS or the KDE methods described in the previous chapter. It is noteworthy that unlike the NEWS-based or KDE-based warning scores, which tend to be persistent, the step-change detector (by its nature) produces transient warning scores. This can be seen in 8.3(b) where step-change NLML is escalated for only a short period of time, e.g. for HR step-change NLML (red) near hours 71, 73, and 74.5 and for SpO_2 step-change NLML (green) near hours 77.5, 78.5, and 80. This is to be expected since the flexibility of the GP allows it to adjust quickly to the new (volatile) data and resume precise forecasting. In contrast, the KDE-based warning score in 8.3(d) is persistent when elevated, e.g., between hours 79 and 80.

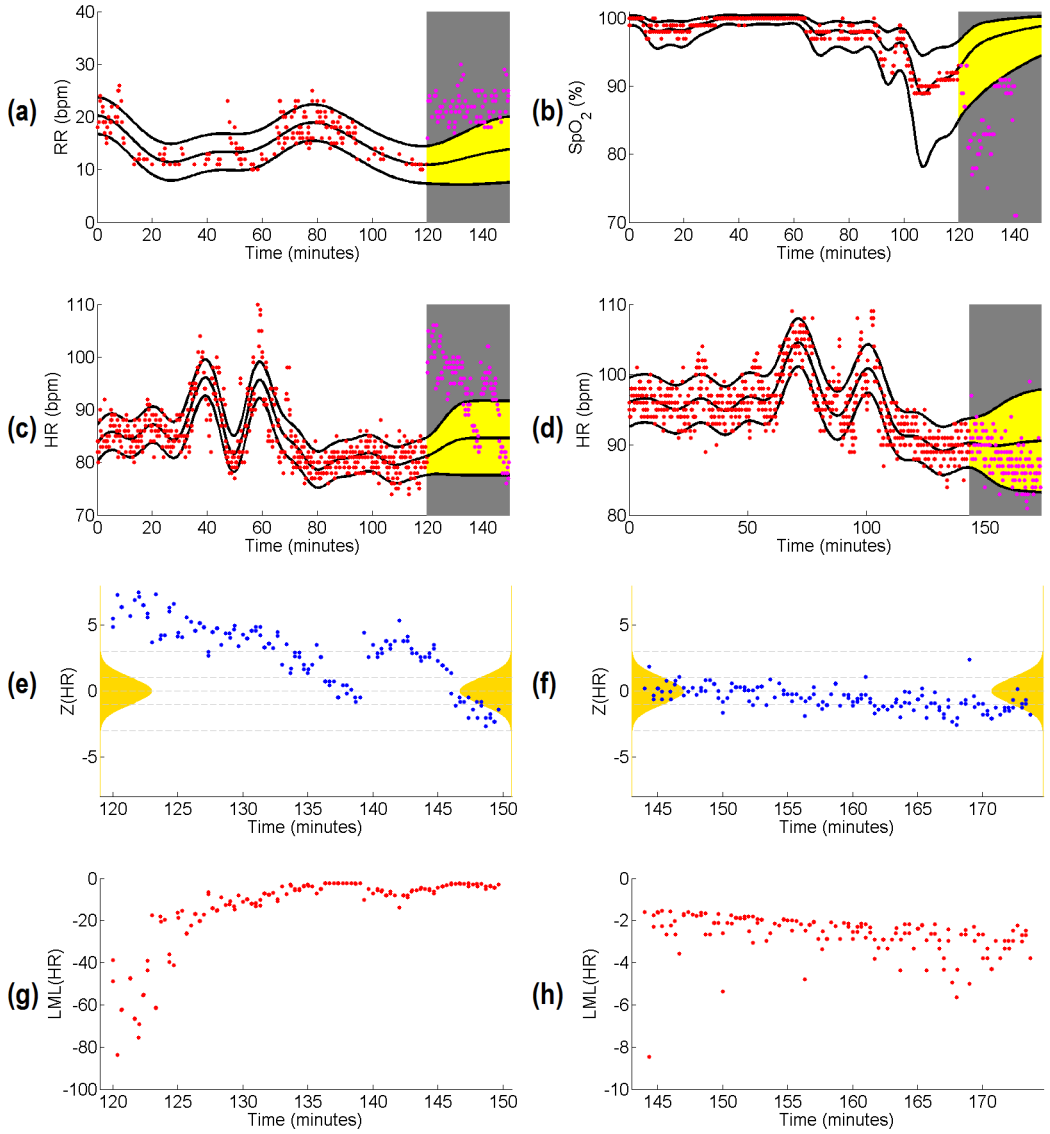


Figure 8.2: Illustration of a GP-identified step-change in (a) RR, (b) SpO_2, and (c) HR. In (d) the HR time-series shows no step-change. GPs are fit to observed data (•) and forecast the distribution of unseen data in the future (•). In (c) the asymmetric GPR confidence bounds on SpO_2 are due to a $\log(101 - SpO_2)$ transformation, to minimise the proportion of the posterior distribution greater than 100%, which is physically impossible. Since the marginal Gaussian distribution changes throughout the forecast window, the z-scores of the forecast-window HR from (c) and (d) are shown in (e) and (f), respectively. In (e) and (f) a $N(0, 1)$ reference distribution (gold) is provided with (- -) denoting mean $\pm$ 0, 1, and 3 standard deviations. The forecast LML of each measurement in the forecast windows of (c) and (d) are shown in (g) and (h), respectively. The LML measurements within a specific time window may be summarised, e.g., by the mean or another statistic. Step-change warning scores are the negative of these LML values, NLML.

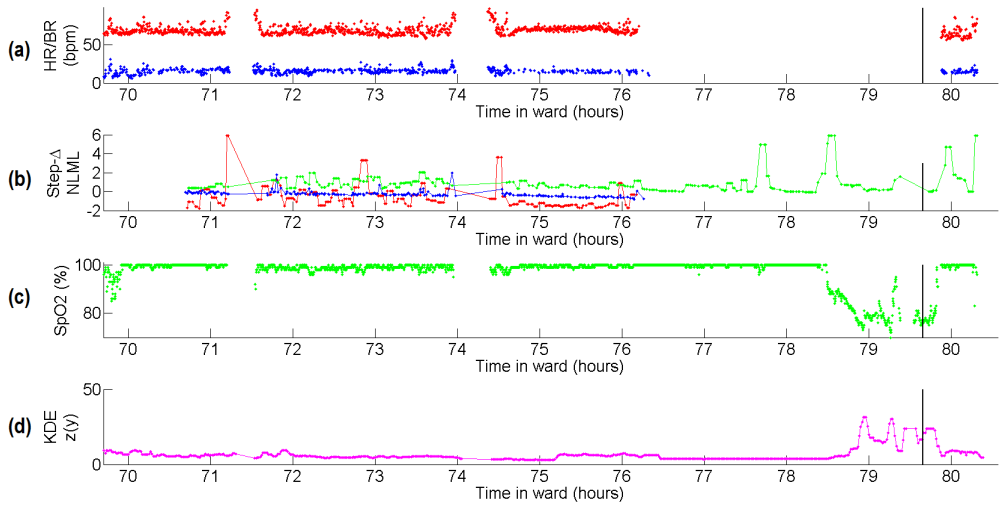


Figure 8.3: Time-series of a patient's vital-signs and early warning scores leading up to an emergency event near hour 80 (black vertical line). The patient's vital-signs in (a) HR (●) and RR (●), and (c) SpO_2 (●) each display various step-change dynamics. SBP and DBP are not shown. In (d), the 5-vital KDE score of the patient vitals has been calculated. It is seen here to escalate at the approach of the emergency event. In (b) the step-change detection novelty score for each individual vital-sign is shown for HR (●), RR (●), and SpO_2 (●). No step-change score is available in the absence of measurements.

Since the warning scores produced by step-change detection are transient instead of persistent, the warnings scores of a step-change detector are apt to be missed if monitored only sporadically by clinical staff. In this, step-change detection would only be appropriate in a continuous computer-assisted monitoring setting since the score indicative of deterioration would need to be recorded and brought to the attention of clinical staff.

8.3 Univariate Step-Change Detection

In Figure 8.4, we show the TEW vs. FPR plots of step-change monitoring on HR, RR, and SpO_2 on the UPMC patients. HR and RR easily out-perform the baseline KDE-model, whereas step-change on SpO_2 performs approximately as well at the KDE method. Since blood pressure is not measured continuously, step-change detection is not applied to SBP and DBP.

Noting that the lines for each method represent the 33, 50, and 67 percentile of TEW, step-change detection on HR in 8.4(a) demonstrates the strongest per-

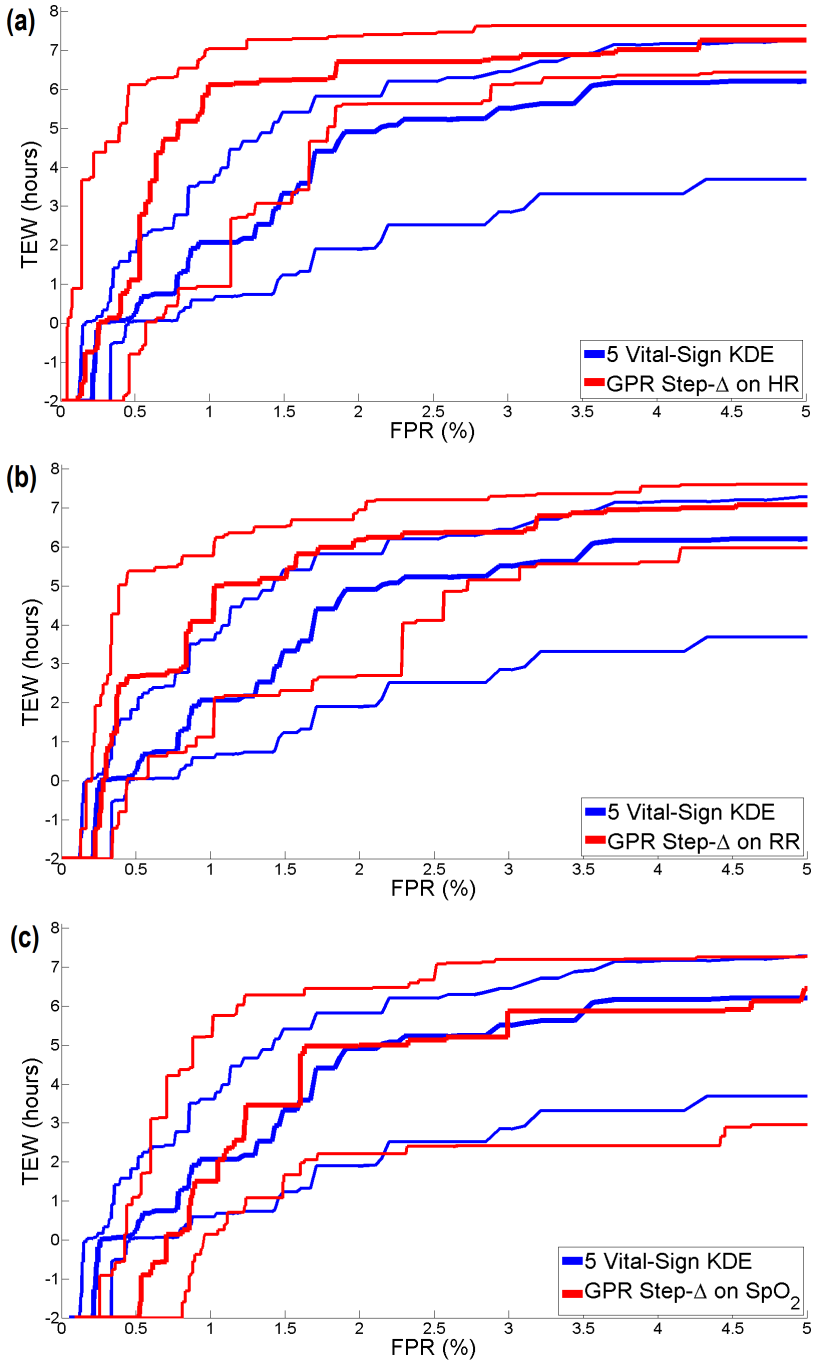


Figure 8.4: TEW vs. FPR plots of a step-change detector on (a) HR, (b) RR, and (c) SpO_2. Each is compared to the baseline KDE model described in Chapter 7. Lines represent the 33, 50 and 67 percentile of TEW at the FPR.

formance, followed by RR, then SpO_2. For example, the 33-percentile TEW of HR step-change detector generally exceeds the median TEW of the KDE system across the FPR range from 0% to 5%. Similarly, the median TEW of the HR step-change detector exceeds the 67-percentile of the KDE method by a significant margin. At an FPR of 1% (for minutely-evaluations for deterioration, this is a false alarm approximately every 100 minutes or 1 hour and 40 minutes) the 33-percentiles are approximately the same for each method at about 0.5 hours to 1 hour. The median TEW improves from 2 hours with KDE to 5 hours with the HR step-change detector. The 67-percentile TEW improves from 3.5 hours with KDE to 7 hours with the HR step-change detector. More importantly, since we prefer to reduce or eliminate worst-case performance, between a 1% and 5% FPR, the 33-percentile of the HR step-change detector escalates rapidly to nearly-match the 67-percentile of the KDE method. This means that (for the same FPR in non-C"-patients) worst-case performance under HR step-change is generally better than average-case performance under the KDE method.

Similar intuition can be applied to 8.4(b-c) to determine that step-change detection on RR also outperforms the KDE. Step-change detection on SpO_2 is superior to the KDE for the 67th percentile, but otherwise nearly identical.

8.4 Bivariate Step-Change Detection

Univariate methods are susceptible to missingness of the single vital-sign under consideration, since missingness removes possibility for clinical inference.[1] The susceptibility of monitoring algorithms to missingness can be seen in Figure 8.3, in which the KDE and SpO_2-based step-change detector continue to produce estimates in the absence of HR and RR, whereas the HR-based and RR-based step-change detectors produce no score because the requisite vital-sign is not available. Therefore,

[1]Vital-sign missingness may arise from anything from unintentional probe-detachment to intentional detachment to facilitate clinical intervention (however such events are not annotated in the UPMC data set). As seen in Chapter 3: Data Description, as well as Figure 8.3, missingness typically does not refer to complete missingness over the patient's stay, but instead, the periods of missingness interspersed between periods of measurement.

we may wish to monitor across multiple vital-signs to (i) maximise vital-sign coverage, in addition to (ii) incorporating a wider variety of informative clinical hypotheses.

As shown in Figure 8.3(b), this has several challenges: First, the forecast NLML of different vitals-signs are not directly comparable in terms of units. As seen in Figure 8.3(a) and 8.3(c) each vital-sign differs in variance (the spread of the vital-sign's distribution), so forecast NLMLs differ in average magnitude. For example, the median NLML of the HR step-change detector is about -2, where for RR and SpO_2 it is about -1 and 0, respectively.

Furthermore, as transient metrics, the step-change evidence may be separated in time across different vital-signs. For example, a step-change in blood pressure may precede a step-change in HR. Furthermore, this lag itself may vary across time, vitals-signs, and patients. While sophisticated approaches to combine these signals are recommended for future work, in the absence of further held-out testing data it would be difficult to determine whether such sophisticated methods were generalisable or over-fit to the data. Instead, we apply a simple approach of adding the NLML value from each vital-sign at each minute. (Addition is a reasonable operation for these joint log-likelihoods, since likelihoods ought to be multiplied if modelled independently.)

Figure 8.5 shows the TEW vs. FPR of each univariate step-change detector, followed by the three unique pairwise combinations of HR, RR, and SpO_2. These bivariate combinations are 8.5(d,e) HR and RR, 8.5(f,g) HR and SpO_2, and 8.5(h,i) RR and SpO_2. All bivariate step-change detectors of the same two vital-sign (e.g "HR and RR" vs. "RR and HR") represent an identical step-change algorithm, and therefore exhibit identical TEW vs. FPR performance.

Each univariate step-change detector benefited significantly from gaining monitoring coverage by including other vital-signs. Most importantly, the 33-percentile of TEW of each bivariate step-change detector increased to at least the median performance of the respective univariate baseline models. This means that worst-case TEW performance is significantly improved, so few patients will have their deterioration signs completely missed or identified too late for clinical intervention.

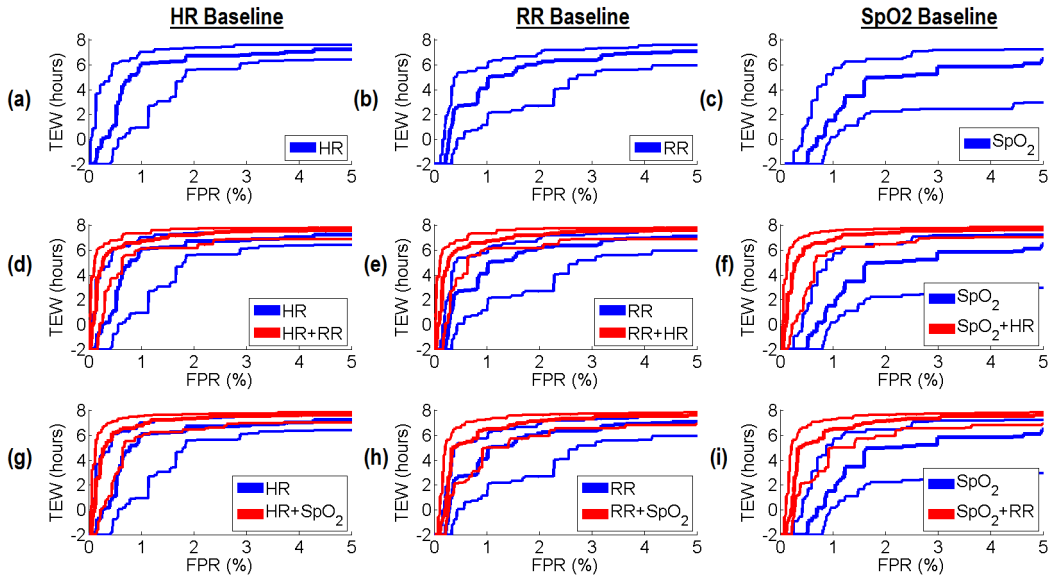


Figure 8.5: TEW vs. FPR plots of bivariate step-change detectors starting with the univariate step-change detectors on (a) HR, (b) RR, and (c) SpO_2. Each of the other vital-signs is then included for $\binom{3}{2} = 3$ unique bivariate combinations: (d,e) HR and RR, (f,g) HR and SpO_2, and (h,i) RR and SpO_2. In each column, the baseline univariate step-change detector is plotted in blue for ease of visual comparison. Lines represent the 33, 50 and 67 percentile of TEW at the respective FPR.

8.5 Trivariate Step-Change Detection

The final step-change detection model includes HR, RR, and SpO_2. As with the bivariate step-change detectors, the NLML of each of the three vital-signs at each minute is added together for a single warning score.

The TEW vs. FPR performance of the trivariate step-change detector is plotted against each of univariate and bivariate step-change detectors in Figure 8.6. As before, the inclusion of further vital-signs increases the coverage of clinical inference in the event of missing vital-sign channels. From 8.6(a), it is clear that most of the TEW improvement has been achieved by an FPR of 0.5%, or a false alarm approximately every three hours of patient monitoring. The trivariate models clearly outperforms the univariate step-change detectors in 8.6(b,c,d). Compared to the bivariate models in 8.6(e,f,g), the trivariate model only improves in the 33-percentile. This is a positive result, in the sense that the inclusion of a further vital-sign improves the consistency of improvement across all patients.

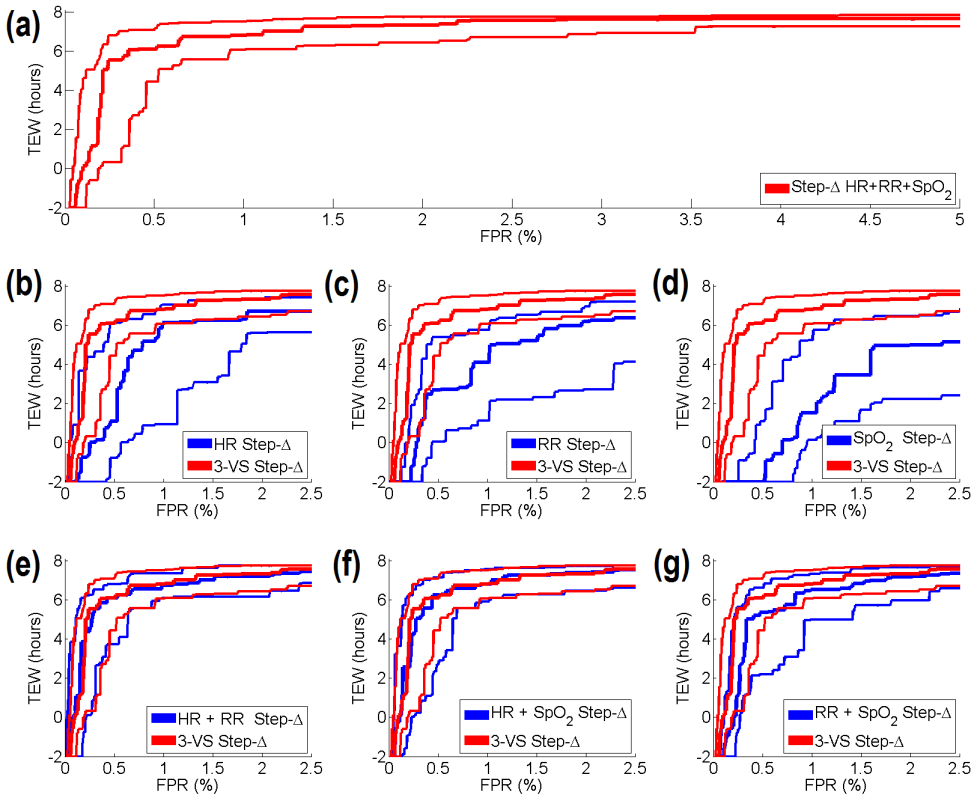


Figure 8.6: TEW vs. FPR plots of (a) the trivariate step-change detector compared to univariate step-change detectors on (a) HR, (b) RR, and (c) SpO_2, and the bivariate step-change detectors on (e) HR and RR, (f) HR and SpO_2, and (g) RR and SpO_2. The trivariate detector is included in each plot (red) for ease of comparison. Lines represent the 33, 50 and 67 percentile of TEW at the respective FPR.

8.6 GP Step-Change vs. Baseline Comparators

8.6.1 TEW vs. FPR Results

Figure 8.7 shows the TEW vs. FPR performance of the 5-vital-sign KDE (the top-performing baseline comparator) against the performance of each of the step-change detection models. We focus on early warning performance in the 0% to 2.5% FPR range due to the desirability to mitigate alarm fatigue. The strong performance of step-change detection methods suggest that step-change detection, or similar methods may provide a useful supplement to the current state of the art in patient monitoring.

Among the univariate step-change detectors, 8.7(a) HR and 8.7(b) RR both outperform the KDE, despite each using only a single vital-sign compared to the

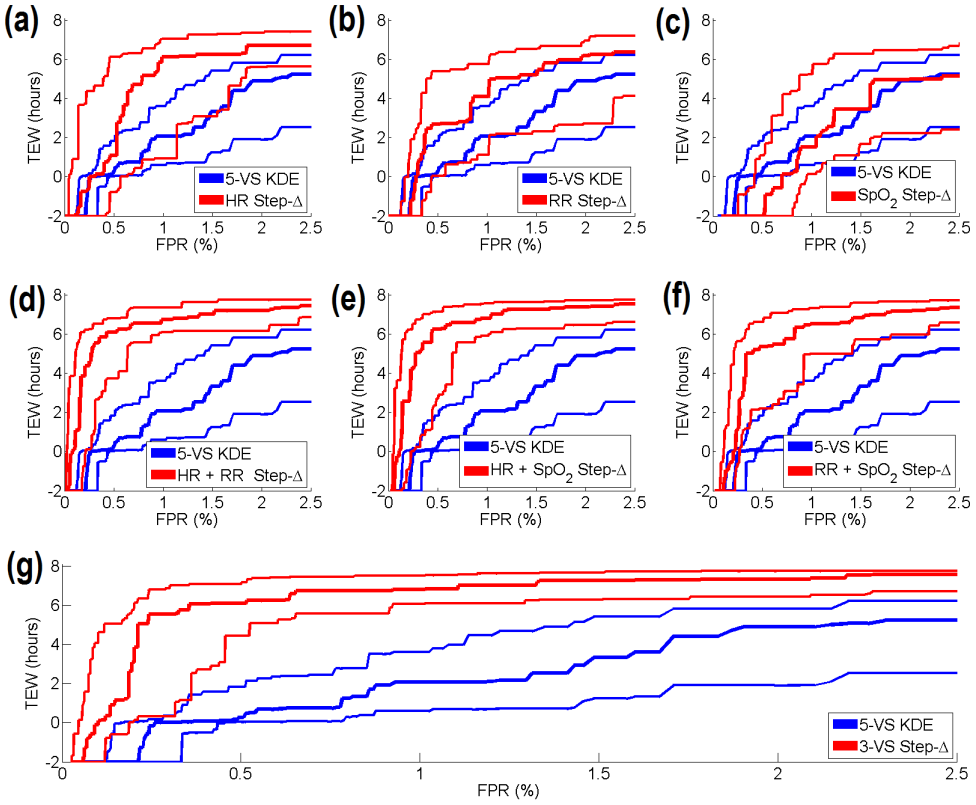


Figure 8.7: TEW vs. FPR plots of KDE baseline comparator (−) next to Univariate step-change detectors (−) on (a) HR, (b) RR and (c) SpO_2; Bivariate step-change detectors (−) on (d)HR+RR, (e) HR+SpO_2, and (f) RR+SpO_2; Trivariate step-change detector (−) on (g) HR+RR+SpO_2 (red). Lines represent 33, 50 and 67 percentiles of TEW at the respective FPR.

KDE's five vital-signs. The step-change on 8.7(c) SpO_2 performs nearly the same as KDE, except for the KDE's superior performance in the 0% to 0.5% FPR range. An explanation for the superior performance of univariate step-change detection over the KDE will be discussed shortly, however this outcome is positive in several respects: First, the step-change method is not dependent on the magnitudes of vital-signs. The information contained within the step-change detection metrics is significantly different from KDE-based or NEWS-based scores. This suggests that such a metric may be a useful supplement to current monitoring. Second, the strong monitoring results on only a single vital-sign suggests that a vast number of clinical variables may not be necessary to achieve optimal or near-optimal monitoring performance. This may be useful, i.e., in resource-constrained settings where fewer monitoring modalities are available. However, both of these benefits are dependent

on the interpretability of the step-change detector when it brings warning scores to the attention to clinical staff.

A final note on the univariate step-change methods is that, within the FPR range of 0% to 0.5%, the KDE frequently outperforms the step-change detector in median and 33-percentile performance. This region roughly corresponds to warnings about 1-hour-or-less prior to the emergency event. Contributing factors to this are several-fold. One factor is the missingness in individual vital-signs near the emergency event, as illustrated for HR and RR in Figure 8.3(a). This puts univariate methods at a disadvantage when the particular vital-sign under consideration is missing. However, this would also put the KDE method at a disadvantage when 3 or more of the five-total vital-signs (that is, any 3 between HR, RR, SpO_2, SBP, and DBP) are missing, since the KDE would no longer produce a score whereas univariate monitoring system on either of the other two vital-signs would continue.

More important than the issue of missingness, is the definition of emergency events. Emergency events, by definition, have highly-abnormal vital-sign values. This means that the KDE-based warning score is nearly-guaranteed to be high in proximity to emergency events because at least one vital-sign will be sufficiently abnormal to contribute to a high warning score. In contrast, the emergency events are not defined according to vital-sign volatility, on which step-change methods are based. Step-changes are, therefore, not guaranteed to occur at the time of event.

When KDE performance is compared to 8.3(d,e,f) bivariate and 8.3(g) trivariate step-change detection, there are fewer caveats to the results, since results are almost uniformly superior. While the improved performance itself may be unsurprising (having already seen that univariate methods themselves produced superior results) the magnitude of difference in performance motivates further inspection to understand why this may be.

The remainder of this chapter discusses possible factors that contribute to the difference in performance between step-change based monitoring and KDE-based monitoring.

The following factors were determined to be particularly important:

1. The physiology at the time of the emergency event differs from the physiology preceding the emergency event.

2. Personalisation of monitoring improves FPR, or, conversely, population-based risk assessment necessitates high FPR.

8.6.2 Physiology Preceding and During the Emergency Event

This section will discuss how the vital-sign measurements tend to differ between (i) the time of the emergency event, and (ii) the time preceding an emergency events. The subsequent sections will then discuss FPR and TEW in particular.

We begin by reasserting that a patient's deterioration is not usually preceded by a long period of gradual decent towards abnormal values. Instead it is more frequently characterised by "shocks" to one or more vital-signs, which are quickly corrected by homeostatic mechanisms. This can be verified by examining plots of patient time-series in the the time period surrounding emergency events. For example, in Figure 8.3, the low SpO_2 which triggered the emergency event near hour 80, was only preceded by about 1 hour of decreasing SpO_2. Otherwise, there was nothing abnormal in the absolute values of HR, RR, or SpO_2. This means that NEWS or KDE-based warning scores would only begin to increase (compared to the general population) in the final hour before deterioration. Therefore, for the KDE to achieve an early warning greater than 1 hour, FPR would need to increase substantially. In contrast, step-change detection identifies at least 3 prominent HR step-changes, 2 prominent RR step-changes, and 2 prominent SpO_2 step-changes, in addition to several smaller step-change episodes. This results in approximately 1 step-change episode per hour. It is unclear whether the KDE warning score was artificially depressed between hours 76 to 80 due to the missingness of HR and RR, however, the KDE does not appear to identify any alarmable episodes between hours 70 to 76, in which all vital-sign are available.

In summary, the KDE method has several advantages and disadvantages.

The KDE method's advantage is that it alarms on vital-sign measurements that (either univariately or jointly) are abnormally high or low. Since the emergency

events of this data set were also annotated as such for their abnormally high or low values, the KDE can alarm reliably at the time of the annotated emergency events.[2]

However the TEW metric places a premium on *advanced* warning, since earlier warning facilitates preventative clinical intervention. If vital-signs are neither abnormally high or low far in advance of the emergency event, then the KDE monitoring method is at a disadvantage.

The KDE is at a further disadvantage with respect to FPR since it attempts to describe the entire patient cohort using only a single model. This is discussed in the next section.

8.6.3 Personalisation Improves FPR

Like the heuristic NEWS approach, the KDE-based EWS suffers from attempting to use a single model to describe each patient. This means that it only matters (i) whether a patient exhibits physiology abnormal to the entire population, not (ii) whether the patient exhibits physiology abnormal to himself. Age- and sex-based early warning scores attempt to adjust for obvious confounding demographic information, however, the intra-group variability is likely to be substantial, given that inter-patient and intra-patient variability is high. As seen in Chapter 3, inter-patient variability is high, even when stratified by C"-status.

To illustrate this, the inter- and intra-patient variability of vital-signs, KDE warning scores, and step-change warning scores are plotted in Figure 8.8 for the 89 non-C"-patients, and Figure 8.9 for the 59 C"-patients.

Inspection of the intra-patient ranges in 8.8(a-c) and 8.9(a-c) show patient vital-signs may operate in completely different dynamic ranges. In extreme cases the upper 2.5 percent of one patient's vital-signs may still be less-than the lower 2.5 percent of another patients vital-signs. This, effectively, removes the possibility of alarming on low values for patients with high-valued vital-signs or alarming on high values for patients with low-valued vital-signs. More importantly with respect to the FPR metric, this means for an "average" patient in the middle of this range

[2]In contrast, the step-change detector has no such guarantee that the requisite physiology (a step-change) will occur at the time of the event.

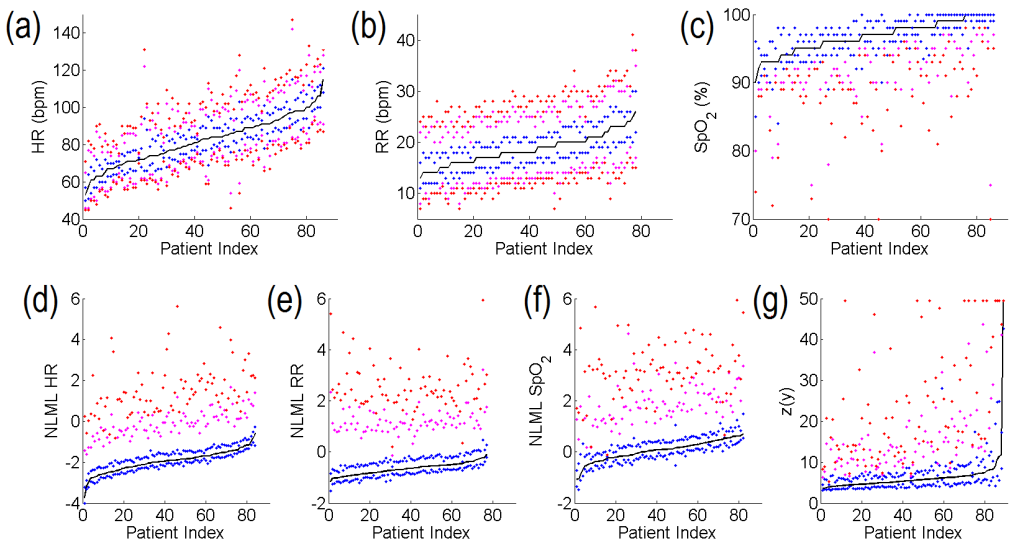


Figure 8.8: Non-C"-patient intra- and inter-patient variability in (a) HR, (b) RR, (c) SpO_2, contrasted to (g) KDE warning scores, and step-change NLML in (d) HR, (e) RR, and (f) SpO_2. For each patient the following percentiles are marked: median (-), 25 and 50 (•), 5 and 95 (•), and 2.5 and 97.5 (•). Patients are indexed by their median value for each metric.

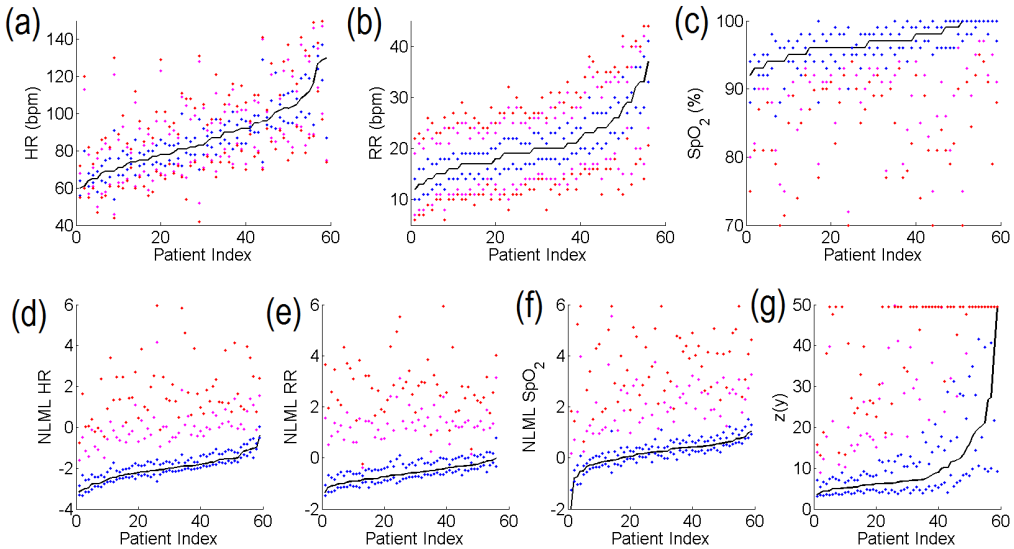


Figure 8.9: C"-patient intra and inter-patient variability in (a) HR, (b) RR, (c) SpO_2, contrasted to (g) KDE warning scores, and step-change NLML in (d) HR, (e) RR, and (f) SpO_2. For each patient the following percentile are marked: median (-), 25 and 50 (•), 5 and 95 (•), and 2.5 and 97.5 (•). Patients are indexed by their median value for each metric.

could not achieve any type of alarm without the monitoring system inducing a nearly-constant state of alarm in other patients.

In contrast, the dynamic range of step-change NLML for HR, RR, and (to a lesser-extent) SpO_2 is highly consistent and compact across all patients. Very few patients have dynamic ranges in step-change NLML that overlap with the extreme (alarm-generating) values of other patients. For example, the 2.5 quantile values of HR NLML in C"-patients in 8.9(d) are more extreme than those of non-C"-patients in 8.8(d). This indicates that C"-patient have more pronounced HR step-changes than non-C"-patients, as we would expect. More important for maintaining a low FPR, though, is that no patient has average values that would fall within the high NLML range. This leads to a two-fold conclusion: (i) a single threshold can delineate between low-NLML and high-NLML step-changes across all patients, and (ii) highest NLML step-changes occur more frequently in C"-patients than they do in non-C"-patients. It is not surprising then, that step-change detection demonstrates a successful trade-off between TEW and FPR.

The KDE method falls in between thresholding on raw vital-signs and the step-change method. Comparing KDE warning scores between the 89 non-C"-patients in 8.8(g) and the 59 C"-patients in 8.9(g), it is immediately apparent that C"-patients experience a much higher rate of high warning scores than non-C"-patients. The difference is much greater, even, than the difference between NLML step-change values of C" and non-C"-patients). This is expected, since the KDE-based novelty score is persistent in the presence of abnormality. However it can be seen that the 95 (•) and 97.5 (•) percentiles of KDE novelty are still highly-intermingled on an inter-patient basis. To avoid the high FPR that make an early warning system infeasible in clinical practice, the high and low percentiles must be clearly delineated. Otherwise the system will generate a near-constant rate of alarms in non-C"-patients.

While Figures 8.8 and 8.9 help illuminate the FPR trade-off, they do not directly address the timeliness of the alarms. The question of timeliness is investigated next.

8.6.4 Early Warning Examples

We complete this chapter with a series of exemplar patient time-series for which both both step-change and KDE-based warning scores are calculated. These examples are helpful illustrations of (i) where step-change inference may usefully supplement current extreme-value-oriented monitoring, (ii) where step-change detection is not helpful, and (iii) how technical interference is a challenge to both step-change and KDE-based monitoring.

The exemplar time-series are grouped in the following way:

Figures 8.10 and 8.11 show two patients who experienced a C"-event within 1.5 and 3 hours of entering the ward. In these situations, the KDE-based approach is particularly advantageous because it is able to assess deterioration immediately, whereas the time-series-based step-change detector (i) must wait to collect a sufficiently long time-series before beginning inference, and (ii) has less information on which to learn the patients time-series dynamics and provide a reliable forecast. Similarly, Figure 8.12 also suffers from little data near the C"-event.

The patients in Figures 8.13 and 8.14 show a gradual deterioration of vital-signs (with regard to the KDE-based warning score) in the 5 hours preceding the C"-event. Although such patients are uncommon, they exhibit the type of time-series physiology for which the KDE-based method was designed. However, both of these patients, still exhibit erratic dynamics that are identifiable by the step-change detector as well.

In contrast to the patients in Figures 8.13 and 8.14, the patients in Figures 8.15, 8.16, 8.17, and 8.18 exhibit little evidence of deterioration (in absolute terms) until shortly prior to the emergency event. In this, the KDE method would be unable to attain additional hours of early warning without a significant increase in FPR.

Finally, the patients in Figures 8.19 and 8.20 exhibit punctuated periods of both abnormal physiology and non-abnormal physiology, both in absolute terms and in step-changes. In this, the patients differ from those shown in Figures 8.15, 8.16, 8.17, and 8.18 since there is ample evidence of deterioration in advance

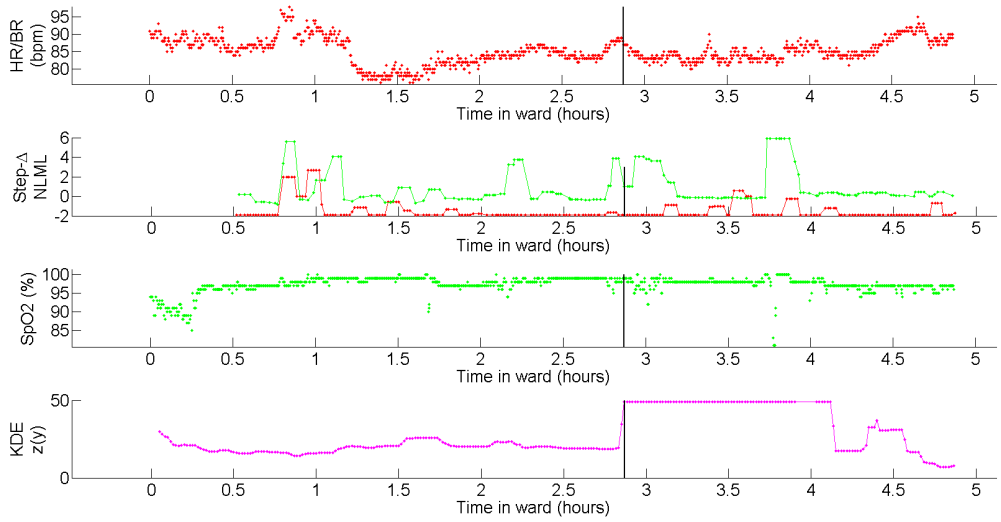


Figure 8.10: A vital-sign and warning score time-series of a patient who deteriorates within 3 hours of entering the ward. The cause of the C"-event is abnormal blood pressure (not shown). Whereas the KDE is able to quantify abnormality from time $t = 0$ the GPR-base step-change detector requires nearly an hour in order to (i) acquire sufficient data, and (ii) identify the next step-change.

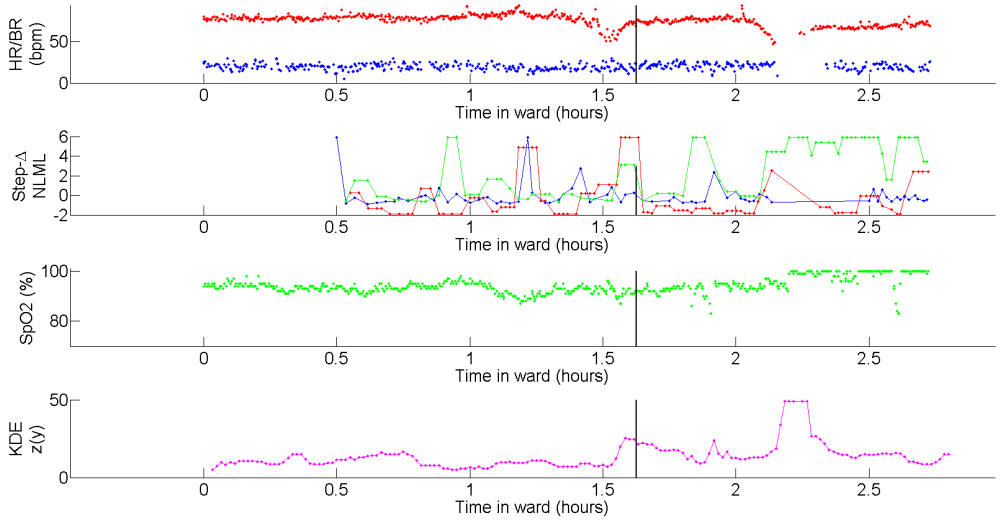


Figure 8.11: A vital-sign and warning score time-series of a patient who deteriorates shortly after 1.5 hours of entering the ward. The KDE and step-change detector exhibit similar early-warning times, since a step-change precipitates the patient's low HR in absolute terms.

of the C"-event, however that evidence is not as gradual or consistent as the patients in Figures 8.13 and 8.14.

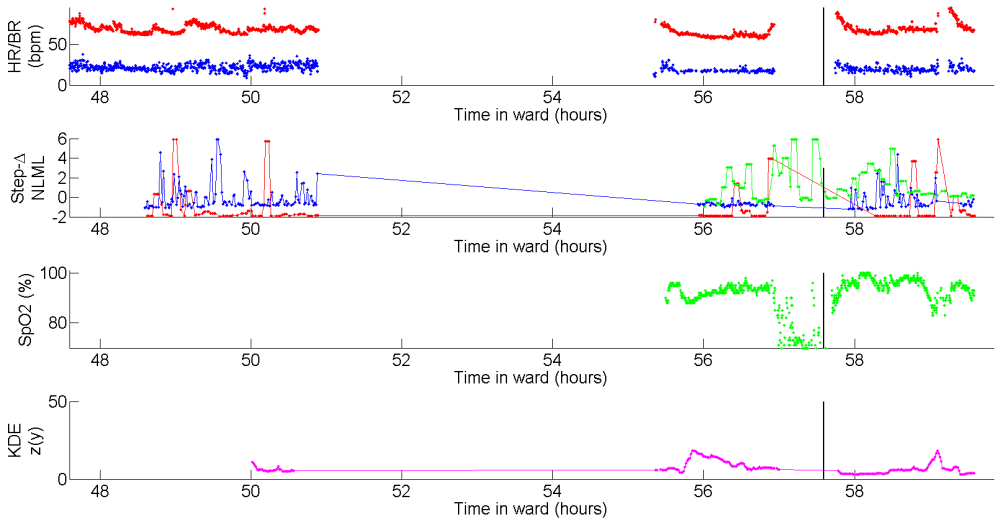


Figure 8.12: A vital-sign and warning score time-series of a patient with significant periods of missing data, both prior-to and during the SpO_2-based C"-event. Since the vital-signs are not abnormal in absolute terms, the KDE fails to identify the deterioration in advance. Furthermore, with only SpO_2 available, the KDE ceases clinical inference. In contrast, any step-change detector that included SpO_2 would continue to perform clinical inference throughout the C"-event.

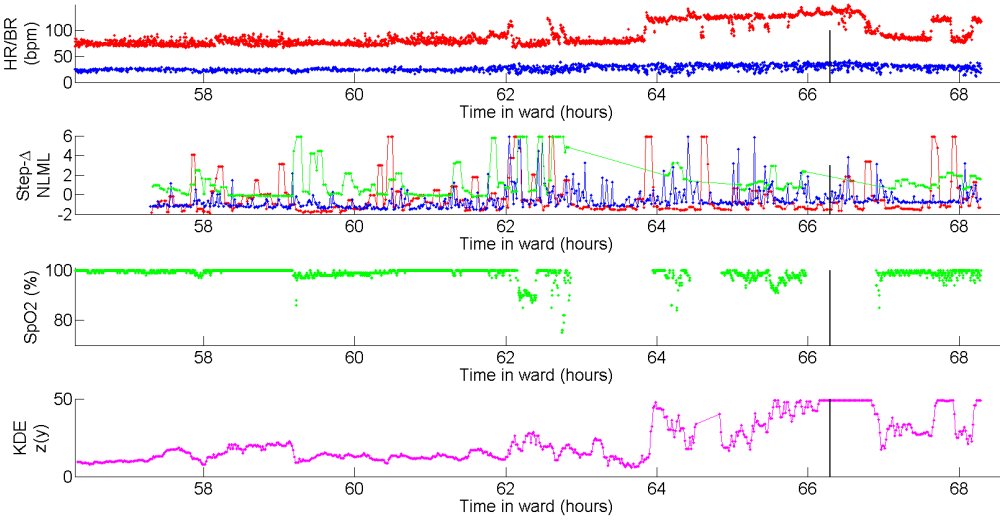


Figure 8.13: A vital-sign and warning score time-series of a patient with long-term escalation in KDE warning score. Although the KDE warning score is punctuated by several large jumps in the hours preceding the C"-event, its general upward trend from hours 60 to 66 is on the account of a the gradual increase in HR and RR over an extended period. HR, RR, and SpO_2 each have a multitude of identifiable step-changes in this period too, which would allow a step-change detector to maintain its advantage in early warning.

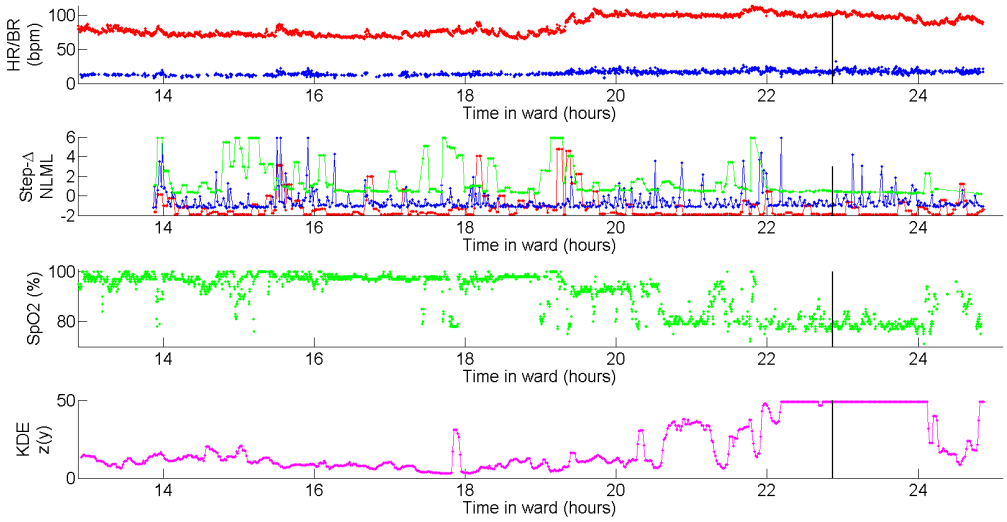


Figure 8.14: A vital-sign and warning score time-series of a patient with long-term escalation in KDE warning score. The KDE warning score steadily increases from hours 18 to 22 as the patient desaturates and HR increases. It is difficult to discern whether the sporadic drops in SpO_2 are artefactual, or instead antecedent to the prolonged desaturation that begins around hour 19. Even without SpO_2, pronounced step-changes in HR and RR occur through the hours preceding the C"-event.

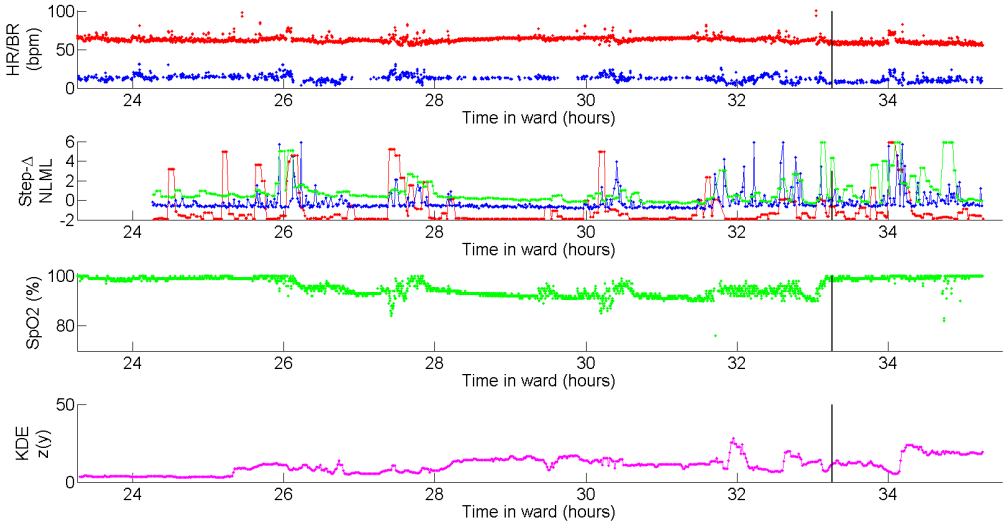


Figure 8.15: A patient with low KDE warning in the hours preceding a C"-event from low RR. With the exception of a small escalation in KDE novelty at hour 32 (caused primarily by low RR as well), there is little evidence of deterioration until nearly an hour after the C"-event. This, in part, reflects that clinical annotation considered only individual vital-signs, even when other vital-signs appear to fall within usual ranges. In this case the abnormally low RR goes unnoticed because of the values of other vitals-signs.

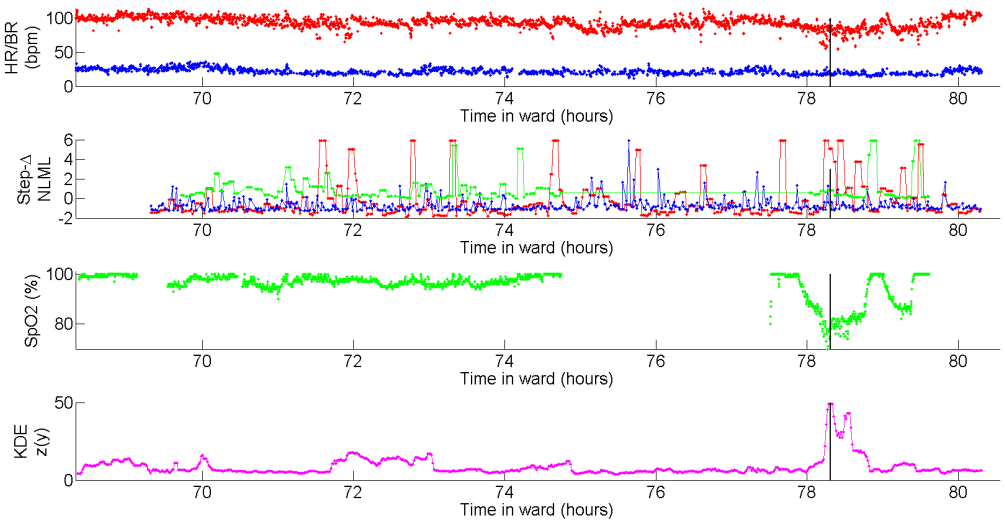


Figure 8.16: A vital-sign and warning score time-series of a patient with low KDE warning but multiple step-changes in the hours preceding a C"-event from low SpO_2. Since the desaturation event at hour 79 is followed by a subsequent desaturation at hour 79, it is unclear whether the C"-event was preceded by several smaller desaturation as well (since this data is missing). It is possible that the KDE would have identified deterioration earlier had SpO_2 been available, however the step-change detector identified at least 1 event per hour in the 8 hours prior.

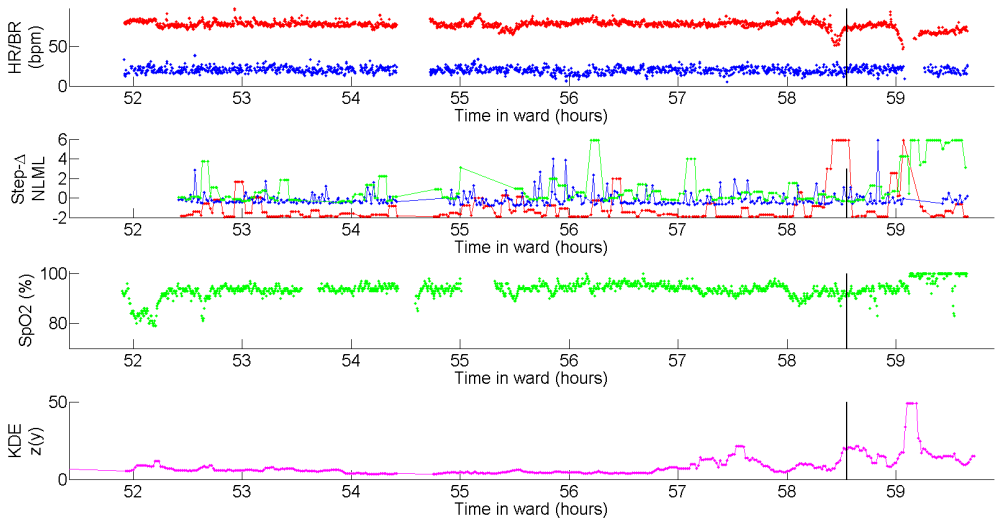


Figure 8.17: A vital-sign and warning score time-series of a patient with little warning from the KDE but multiple step-changes in the hours preceding a C"-event. In this example both the KDE and step-change detector struggled to find evidence of abnormality until the HR drop immediately preceding the C"-event.

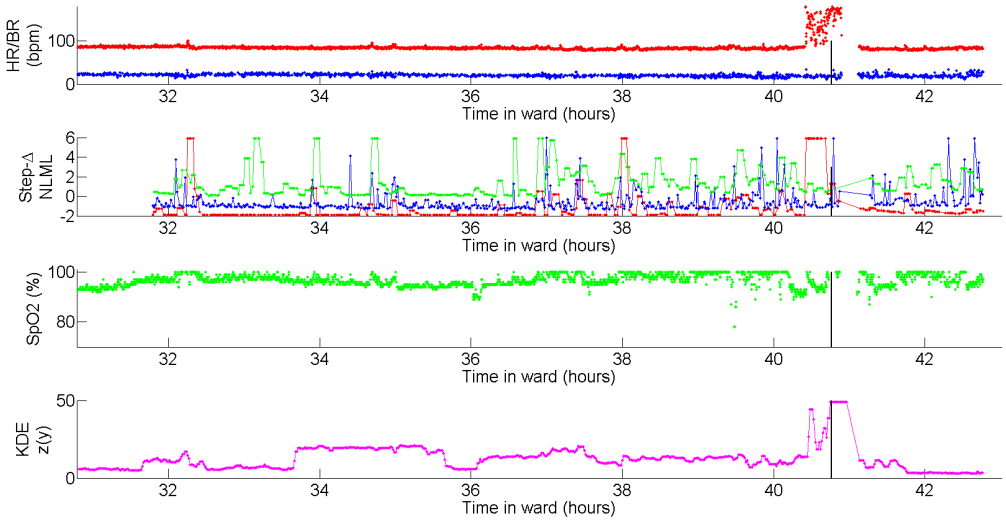


Figure 8.18: A vital-sign and warning score time-series of a patient with little warning from the KDE but multiple step-changes in the hours preceding a C"-event from high HR. This C"-event shows how emergency physiology frequently coincides with noisy signal processing. Since the event is annotated at the end of this noisy period, the clinical annotators have indicated that this is the proportion of the event they believe to be non-artefactual. With the entire HR escalation occurring over the span of 30 minutes, there was little time for the KDE to anticipate the event. However the multiple pronounced step-changes may have for warned of a struggling homeostatic mechanism.

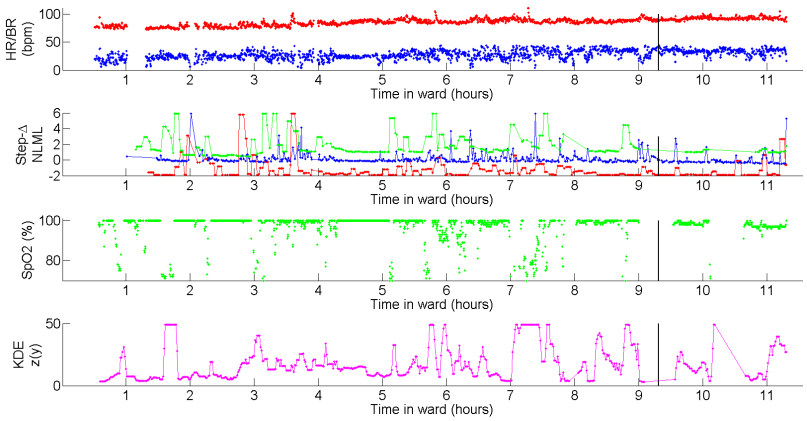


Figure 8.19: A vital-sign and warning score time-series of a patient with multiple escalations in KDE novelty score prior to a C"-event from high RR. The series of highly-escalated KDE novelty scores are largely due to low SpO_2 event which may or may not be artefactual. Since the patient's SpO_2 achieves a consistent value of 100 for much of the time it is possible that the patient was on oxygen therapy. Furthermore, the KDE novelty is not calculated at the time of event because SpO_2, SBP, and DBP are missing at the time of the event.

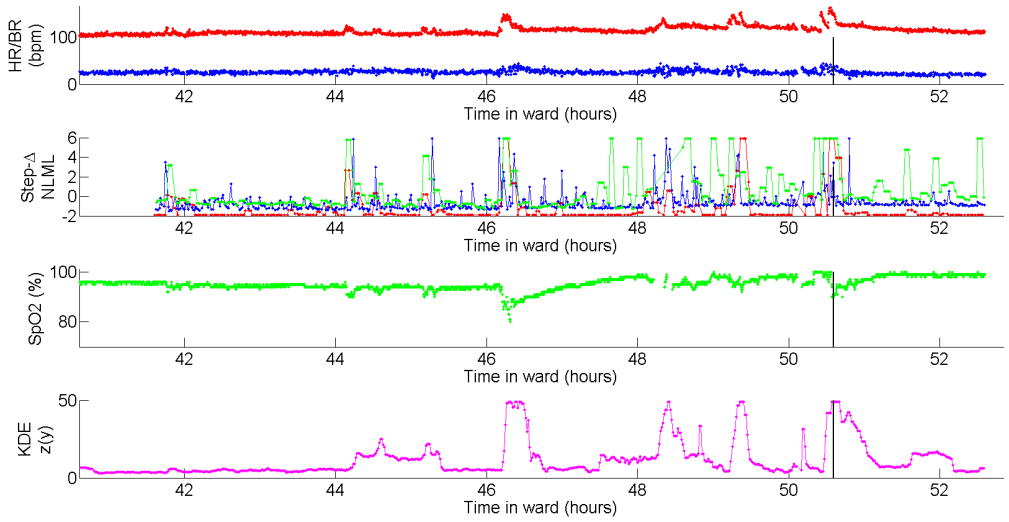


Figure 8.20: A vital-sign and warning score time-series of a patient with multiple escalations in KDE novelty score prior to a C"-event from high HR. Each of the escalations in KDE novelty are due to combination of escalated HR and/or a drop in SpO_2. Due to the sharpness of the HR and SpO_2 volatility, the escalations of KDE warning score are each paired with a pronounced step-change score from at least one vital-sign.

8.7 Conclusion

Current practice in vital-sign early warning scoring focuses on identifying patients with extreme vital-sign measurement values. This is sensible, given that emergency events themselves are typified by vital-signs with extreme values. However, clinical reasoning over time-series offers many ways to identify early physiological indications that current practice tends to ignore.

The probabilistic representation of GP modelling allows us to incorporate a richer range of physiological features beyond magnitude, such as volatility and uncertainty in the patient's current and future vital-sign values. The described methods can provide useful clinical insight beyond or paired with current practice, even when using only a single vital-sign. This is helpful in implementation, since the benefits of a step-change detector may be realised with only a single vital-sign, whereas other empirical monitoring systems may require a diffuse range of vital-signs in order to provide an EWS. Both the baseline KDE method and step-change detection have their advantages, however it is possible the the KDE and related methods are limited due to (i) their population-based approach to modelling patient

abnormality, which increases the rate of false positive alarms, and (ii) the fact that many patients do not exhibit extreme-valued vital-signs until shortly before the emergency event, which decreases the TEW provided by such methods.

The described step-change methods may be applied to a variety of settings including those with constrained computational resources or with a single vital-sign under consideration. Importantly, the described step-change detection models can be run in real-time, even with minimal computational resources, and present salient interpretable physiology to clinical staff to explain the cause of the alarm. By displaying the vital-sign step-change that precipitated the alarm, the GP-based step-change detector is far from a black-box algorithm.

9 Conclusion

9.1 Thesis Contributions

This thesis has described several ways in which probabilistic modelling of patient-specific time-series dynamics may be incorporated into vital-sign data analysis.

These contributions included:

- Development of a computationally lightweight algorithm to identify transient artefactual vital-sign measurements (Chapter 5). This required the identification of a distribution and inference method that could robustly model vital-sign measurement noise to identify low-likelihood measurements as potential artefacts. This resulted in an algorithm with strong inter-patient capability to identify artefactual measurements that were not extreme-valued, which are missed by current methods that place thresholds to remove extreme-valued measurements.

- Identification of a GP covariance function and regularising priors suitable for cohort-wide patient time-series modelling (Chapter 6). The cohort-wide GP model's multiple kernels allowed for modelling of multiple time-series trends, while the regularising priors encoded basic physiological knowledge for reliable automated inference. This resulted in robust automated vital-sign forecasting,

which significantly surpassed the automated forecasting performance of GP models without regularisation via well-chosen priors.

- Development of optimisation methods by which to learn personalised GP models (Chapter 6). This included a simple optimisation method by which to learn a single-kernel personalised GP, as well as more complex optimisation method by which to learn a multi-kernel personalised GP. The latter method proved more effective to navigate a search space in which the input dimensions have a heavily interrelated affect on output. This resulted in the capacity to parametrise more-complex personalised GP models and improved forecasting performance compared to cohort-wide GP models.

- Identification of objective functions quantifying forecast performances with which to (i) reduce worst-case forecasting or (ii) balance the reduction of worst-case forecasting with improved average-case performance (Chapter 6). The outcome was personalised models that did not compromise average performance to achieve robust regularisation against worst-case forecasting performance.

- Development of an early warning score based on personalised time-series modelling to detect erratic volatility (Chapter 8). This involved the development of a 1-, 2-, and 3-vital-sign GP step-change detector. The result was a method of deterioration detection with a superior trade-off between early warning time and false positive rate (when applied to the UPMC SDU data set) compared to a 5-vital-sign KDE-based method, which represents the current technical state of the art.

9.2 Artefact Detection

9.2.1 Result Summary

Vital-sign measurements are beset with artefactual noise-corruption due to a variety of clinical and technical factors including probe-detachment, signal-interference,

and signal-processing error. Common remedies to this noise corruption include (i) removal of implausibly extreme measurement values and (ii) further smoothing of all measurements. The former solution addresses only a minority of vital-sign measurement artefacts, while the latter distorts the measurement noise of the vital-signs.

By performing artefact-inference on a patient's individual vital-sign time-series both of these shortcomings are circumvented by (i) comparing a current measurements to temporally-proximate measurements only, instead of a cohort of patient's measurements, and (ii) identifying specific measurements likely to be artefactual.

This results in an artefact score that can effectively delineate between the majority of artefactual measurements and the majority of non-artefactual measurements. Furthermore, by personalising the artefact inference, the score's performance does not suffer from the high inter-patient variability of vital-sign measurements. Initial evidence was also presented on how removal of artefactual data may improve the false-positive alarm rate in a deterioration detection algorithm.

9.2.2 Future Work

The computationally lightweight calculation of the artefact score allows many possibilities for future work. The most important addition to the artefact analysis would be to validate (via wave-form data) the various archetypal vital-sign dynamics that present in the SDU. This would contribute greatly to the believability of current artefact annotation (in the absence of waveform data).

With believable artefact annotation across multiple vital-sign channels for both transient and persistent artefacts, sensible steps would then include multivariate and time-correlated modelling, since artefacts are likely to correlate across vital-sign channels (e.g. from movement interference, which corrupt PPG measurements for both HR and SpO_2). From a supervised learning perspective, a model-driven approach may aim to classify artefactual dynamics as a function of predictive features, one of which may include the transient artefact score described in Chapter 5.

9.3 Personalised GP Kernel Construction

9.3.1 Result Summary

The monitoring of patient vital-signs via GPs requires not only automated data-cleaning (as described above) but also automated model selection and inference on noisy time-series. Furthermore, there is significant heterogeneity in vital-sign dynamics between patients, and within a single patient's time-series. Addressing this challenge is especially important to avoid worst-case forecast performance, in which a GP produces an extremely inaccurate forecast.

This thesis demonstrated how a GP's covariance function and regularising hyper-priors may be identified in order to apply a single GP model across a patient cohort. Patient-specific GP were then identified, which consistently and significantly improved the robustness of vital-sign forecasting.

A concern with this latter personalised approach is that by optimising a single quantile of forecast performance, we may be ignoring alternative models which are nearly as good, but with superior average performance. This was addressed by the inclusion of a further multi-objective objective function, which demonstrates that a balance may be achieved between these competing objectives.

A final advantage of these personalised model searches is that the patient-specific models were not myopic to the specific forecasting task. Instead, the uncovered parameterisations successfully identified models with high performance for objective functions and for other forecast look-ahead lengths different for which the model was not optimised.

9.3.2 Future Work

Personalised regularisation of GP models can be improved both technically and clinically.

Clinically, personalised models should contain interpretable elements, to be communicated (i.e., in words) to clinical staff. For example, the kernel composition of the personalised model should provide staff with a description of the overall

complexity of a patient time-series. Similarly, parameterisations should be directly translated into terms such as long-term fluctuations, short-term volatility, and noisiness. If extended to multi-task modelling, these same approaches may describe the strength of correlation between vital-signs.

Technically, the identification of personalised models must further overcome the high effective-dimensionality of the hyperparameter search space, which is due to the interrelation of hyperparameter values within and between kernels. This will be especially necessary to build more complex models such as multi-task vital-sign models, and/or identifying priors over the hyperparameters. These goals may be assisted by (i) dimensionality reduction models to learn the modes of variation across dimensions, and (ii) encoding when query points are effectively identical (e.g., different values of λ_i will cease to matter when $h_i = 0$).

Where feasible, heuristic elements should be removed (or replaced with a more reasonable heuristic). For example, left-censoring query values might be remedied via warped-output GPs [148]. If warped GP inference (i) creates an additional computational burden, or (ii) is unreliable in automation, then a practical semi-heuristic alternative may be to learn (one or more) warping functions in an off-line setting, which will be applied to the on-line Bayesian optimisation algorithm.

9.4 GP-based Deterioration Detection

9.4.1 Result Summary

Current vital-sign monitoring aims to identify patients whose vital-sign measurements are persistently extreme-valued. This approach requires a patient's vital-sign to become persistently deranged with respect to a patient cohort before deterioration is identified. However if the goal is to identify the earliest signs of deterioration, then the value of this approach may be limited. Reasons for this may include that (i) extreme values for an individual patient differs from the extreme values of a patient population, and (ii) patients may not exhibit vital-signs that are extreme, long in advance of deterioration. In other words, the physiology at the time of an

emergency events may look very different from the physiology (in the preceding hours) that forewarn the emergency event.

The GP step-change detection algorithm may supplement current extreme-measurement monitoring by identifying instances of erratic volatility in vital-signs. Results from the UPMC data set suggest that step-change methods can provide significantly earlier warning times with a much smaller false positive rate of alarm. This advantage is especially true for those patients who receive the least early warning under current monitoring methods.

The advantage of step-change-based methods for improved early warning time are several-fold:

First, as stated earlier, vital-signs certainly become deranged in the hours preceding cardio-respiratory instability. However, derangement in the form of extremely high or low vital-sign measurements may occur only shortly before cardio-respiratory instability. This means that methods relying on extreme vital-signs (while ideal for identifying patients at the point of cardio-respiratory instability) may struggle to identify the types of vital-sign derangement that occur many hours before cardio-respiratory instability.

Second, one type of vital-sign derangement that *is* present in the hours before cardio-respiratory instability is erratic volatility. This erratic volatility may be described as rapid dynamical changes in a vital-sign time-series, which may or may not result in extreme vital-sign values. This can be seen by visual inspection of vital-sign time-series before or in the absence of a cardio-respiratory instability events. Furthermore, this can be quantified by the reduction in vital-sign forecasting accuracy for patients who experience a cardio-respiratory instability event, compared to the forecast accuracy of patients who did not. The work in previous chapters to (i) remove artefactual data, and (ii) automate robust time-series modelling, allows us to reliably identify these erratic volatility events.

Third, the step-change based method is effective, even when monitoring only a single vital-sign, but improves with the inclusion of multiple vital-signs. From a technical perspective, this allows the step-change detector to continue to monitor

patients, even when vital-sign channels have been dropped (e.g., due to probe detachment). From a physiological perspective, this may relate to the homeostatic mechanism, in that a rapid change in one vital-sign may precipitate a rapid corrective-change in another vital-sign. This means that a step-change in an unmonitored vital-sign is not missed, so much as delayed until its effect is identified in another vital-sign. However, the inclusion of further vital-signs can magnify the deterioration signal by identifying a step-change across multiple vital-sign.

The advantage of step-change-based methods for reduced false-positive alarm rates is also several-fold:

First, the GP-based step-change detection algorithm performs direct, personalised inference on the patient's vital-sign time-series. By learning the patient's typical noise-corruption and rate of change in vital-sign values, the resulting early warning score implicitly controls for uncertainty in vital-sign measurements.

Second, the highly-personalised warning score is less affected by inter-patient variability. This means that a single threshold on the volatility metric will have a similar false alarm rate across all patients. In contrast, inter-patient variability is highly vexing to methods based on extreme values because high values for one deteriorating patient may be low values for another deteriorating patient. This means that a single threshold on a vital-sign value might produce no false positives alarms in one patient, but constant false positive alarms in another patient.

9.4.2 Future Work

The GP-based step-change detector may be in expanded in a variety of ways, ranging from modelling, inference, and clinical development.

As a model of early warning, the step-change early warning score warrants development of a more complex relationship between forecast likelihood and clinical outcome. Specifically, non-linear relationships should be learned to connect early warning scores across vital-signs. Further functional relationships should also be learned to connect multiple warning scores across time. Related to this second point, since the current step-change warning score is transient instead of persistent, there

may be value is summarising cumulatively greater risk as a patient experiences more step-changes over more vital-signs over time.

The GP modelling of vital-sign may also be extended to more realistically represent vital-sign-specific measurement noise and the correlation between vital-signs. More specifically, copula modelling would allow us to diverge from the assumption that each vital-sign is marginally Gaussian and that all vital-signs are jointly multivariate Gaussian. This will have important implications, e.g., for vital-sign measurements such as SpO_2 which frequently achieve their maximum attainable value of 100%, and therefore are clearly non-Gaussian. The obvious advantage would be to provide greater discernment between step-changes and non-step-changes. Furthermore, such methods may provide greater context to the cause of alarm, e.g., volatility in heart rate and respiratory rate vs. volatility in heart rate in the context of a stable respiratory rate.

For development of step-change detection as a useful clinical algorithm, clinical feedback is required on several fronts:

First, clinician feedback would be useful to identify which of the step-change alarms are clinically informative. This would include the identification of (i) alarms on C"-patients that are not clinically informative, and (ii) alarms on non-C"-patient which are clinically informative. This would allow further development of the step-change detector to identify only those dynamics that the clinicians would like to see.

It is equally important to learn how to inform clinicians of step-change-based alarms. As demonstrated earlier, a step-change in vital-signs can be easily visualised by plotting the patient's vital-sign time-series against the forecasted distribution of the GP. This is a first step towards explaining the cause of the alarm without requiring clinical staff to understand the underlying quantitative components. It may be helpful to supplement this visualisation with a written description of the step-change physiology. Such a description could make use of the direction, duration, and magnitude of vital-sign derangement, the GP's hyperparameter values, and the context of other vital-sign values to provide useful context. Such descriptive supplements would require automation as well and therefore require significant

further development for robust performance. The development of automated visualisation and descriptive statements would be require significant collaboration and feedback from the clinicians who would use such alarms.

Developing any of the above requires expanded testing in other data sets with larger numbers of annotated clinical emergencies. Greater data will allow the development of more complex algorithms whose robustness must be verified on held-out data. Furthermore, this would allow us to compare the efficacy of the same algorithm across different hospitals and patient-cohorts. Proving that an algorithm has efficacy across different wards provides the strongest case of the inductive validity of the monitoring method.

Appendices

A

Practical Implementation of Bayesian Optimisation

Bayesian optimisation was used to identify patient-specific Gaussian process (GP) models for robust forecasting. The aim was to identify kernel-complexity and parametrisation with strong forecasting performance, as defined by the objective functions G_1 and G_2, as defined in Chapter 6. The challenge in identifying such kernel-parameter combinations is that the effect of one parameter's values is heavily influenced by the value of the other parameters, creating an optimisation problem with high effective dimensionality. This is why simplistic methods such as random search and simulated annealing struggled to parametrise personalised GPs with more than 1 covariance kernel (corresponding to 3 parameters to optimise).

The patient set under consideration was large and physiologically diffuse. The Bayesian optimisation routine benefited from several practical steps to ensure that the automated optimisation process ran reliably for all patients under consideration.

These practical steps include:

1. Left-censoring of $G_1(\boldsymbol{\theta}) < -3$, for improved data stationarity.

2. Searching over a pre-specified (sufficiently spread-out) grid to avoid a singular covariance matrix of the Bayesian optimisation's GP.

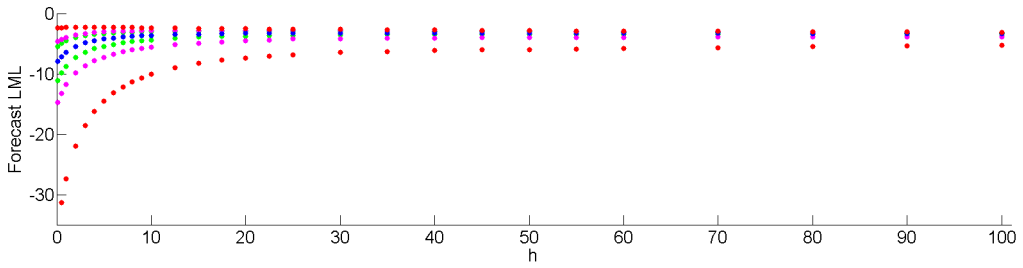


Figure A.1: Forecast LML quantiles at varying values of h. Percentiles are calculated from $4.5 \times 60 = 270$ minutely-forecasts between hours 2 to 6.5, and show the 50 (•), 25 and 75 (•), 10 and 90 (•), and 2.5 and 97.5 (•) percentiles. Objective function G_1 is identical to the bottom red points. Objective function G_2 is roughly equivalent to a linear combination of the bottom half of points. A rapid performance drop off can be seen in $h \in [0, 10]$ compared to $h \in [10, 100]$. Without censoring extremely low values, this would warrant non-stationary GP modelling.

3. Fixing the values of length-scales in the Bayesian optimisation GP to avoid a singular Bayesian optimisation covariance matrix.

4. Searching over a subset of the dimensions of $\boldsymbol{\theta}$ at each iteration of Bayesian Optimisation.

Each of these items is described below.

Left-censoring of G_1 Values

The objective function G_1 is measured in the log marginal likelihood (LML) of a Gaussian distribution. Vital-sign measurements that fall into the extreme tails of a forecast distribution, have LML values that are orders of magnitude lower than measurements near the centre of a forecast distribution. This, in turn, means that personalised models with extremely poor forecasting accuracy have values in G_1 that are orders of magnitude lower than personalised parameterisations with high forecasting accuracy. As shown in Figure A.1, the parameterisations that correspond to the poorest performance are typically edge cases, such as a noise-variance, σ_n, or signal-variance, h, near 0. The modelling of such points requires a non-stationary GP since G_1 changes much more rapidly near these edge cases than away from these edge cases.

To improve data stationarity, values of $G_1(\boldsymbol{\theta}) < -3$ were left-censored to be -3. By censoring these low values, the Bayesian optimisation algorithm remains informed that forecasting performance was low at this point, and is less likely to sample near these points (as it would anyway without censoring).

Since no patient in either the training or test set achieved an optimal $G_1(\boldsymbol{\theta})$ near -3, the censoring does not influence our final estimate of patient-specific $\boldsymbol{\theta}$ performance.

Alternatives to this left-censoring heuristic are to directly model the non-stationarity via a warped-output GP [148]. The challenge to this approach, however, is the same as fitting the Bayesian optimisation GP in general: reliable automated inference without further hand-tuning.

Optimisation via Pre-specified Search Grid

The selection of a query point that is too close to current set of query points runs the risk of creating a GP covariance matrix that is singular. The risk of singularity is especially high for GP's without a white-noise kernel (that is $\sigma_n = 0$). It is typical practice, therefore, to require that new queries be at least a minimum distance from previous queries.

There are many way to achieve this heuristic. In this case, we only accept new Bayesian optimisation queries along a pre-determined grid, in which all points are sufficiently spread out to minimise the risk of singularity.

Fixed Values of Length-scales

The Bayesian optimisation covariance kernel is defined to be

$$\begin{aligned} &c(\boldsymbol{\theta}, \boldsymbol{\theta}') = \eta^2\Big(1 + \sqrt{3r}\Big)\exp\Big(-\sqrt{3r}\Big), \\ &\text{s.t. } r = \sum_{d=1}^{D} \tfrac{(\boldsymbol{\theta}_d - \boldsymbol{\theta}'_d)^2}{\nu_d}. \end{aligned} \tag{A.1}$$

Since each element in $\boldsymbol{\theta}$ is an interpretable GP hyperparameter, describing the dynamics of a patient time-series, we have insight into the values of ν_d. That is, we understand the minimum values over which a change in $\boldsymbol{\theta}_d$ will effect forecast

performance. This directly corresponds to a minimum value in ν_d, the length scale of the Bayesian optimisation GP in that dimension, which may result in a change in G_1.

Computationally, fixing the values of ν_d *a priori* have several advantages: first it removes the possibility of a singular covariance matrix due to low values in ν_d which, in turn, interferes with proper inference on the other parameters of $c(\boldsymbol{\theta}, \boldsymbol{\theta}')$. Secondly, with fewer parameters over which to perform inference, the MAP fitting of the remaining parameter, η, is quicker in addition to being more reliable.

Sequential Optimisation over a Subset of dimensions

To encourage greater exploration of the G_1 and G_2 search spaces, the Bayesian optimisation algorithm searches over only a subset of the dimensions of $\boldsymbol{\theta}$ at each iteration. In other words, the sequential approach optimises only a subset, $\mathcal{D}$, of $\boldsymbol{\theta}$'s D total dimensions by holding the remaining values fixed at those of the current best found $\boldsymbol{\theta}$.

Concretely, the sequential optimisation search can be achieved simply by changing line 3iii of Algorithm 1 in Chapter 6 to

$$\begin{aligned} &\textbf{query } G(\boldsymbol{\theta}) \text{ at } \boldsymbol{\theta}_{\text{new}} := \arg\max_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) \\ &\text{s.t. } \boldsymbol{\theta}_d = \boldsymbol{\theta}_d^{\text{best}} \ \forall d \notin \mathcal{D}. \end{aligned} \tag{A.2}$$

The sequence of subsets $\mathcal{D}$ can be selected at random, or by any other selection criterion. However, for simplicity, we opted to tune the dimensions of $\boldsymbol{\theta}$ that are particularly related. For example, the Bayesian optimisation could sequentially tune length-scales ($\lambda_{i=1,..,a}$), followed by variances ($h_{i=1,..,a}$ and σ_n). Alternatively, the algorithm could sequentially tune hyperparameters of a specific kernel, for example, h_i and λ_i of kernel k_i.

This sequential optimisation approach was found to be helpful to increase the exploration of the search space, more so than, e.g., further augmentations of the acquisition function to encourage exploration.

B

Construction of Kernel Density Estimate Model of Patient Normality

This appendix covers the practical design elements to construct a KDE of patient normality. The implementation is the same as that described in Chapter 4 of Hann [45]. A summarising diagram of this process is shown in Figure 8 in Chapter 7, and is reproduced as Figure B.1 of this appendix. The elements of Figure B.1 are described in turn below.

In summary: Each patient's 5-vital-sign time-series (HR, RR, SpO_2, SBP, DBP) is cleaned of any (presumably artefactual) measurements that exceed a pre-determined artefact threshold. Each of the 5 clean vital-sign time-series are aligned to create a unique 5-dimensional data point at any time-stamp. The training data set is created by collating all of these 5-vital-sign data points (irrespective of time-stamp) for all patients in a held-out set of non-C"-patients. This creates an m-by-5 training set, where m is the total number of 5-D data points across all training set patients. These data are scaled to be zero mean and unit variance. A k-means clustering algorithm is applied to identify 500 vital-sign centroids. The 100 outer-most centroids are removed, leaving 400 centroids to describe the joint vital-sign distribution of healthy patients. The joint pdf of this distribution is

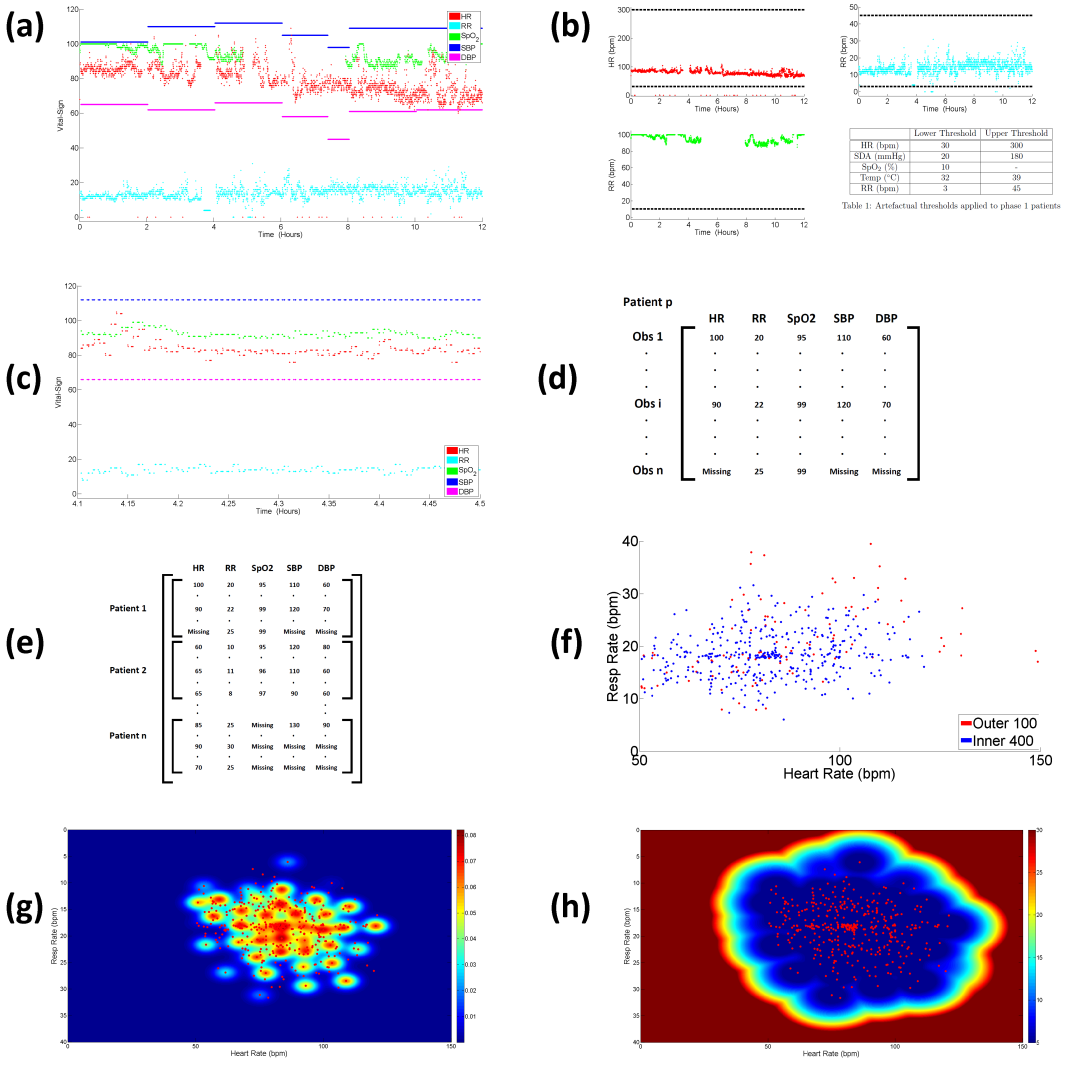


Figure B.1: Creation of a KDE-based novelty score reproduced from Chapter 7. Each patient's (a) vital-sign time-series is (b) cleaned of artefactual measurements, and (c) aligned via their time-stamps. These aligned data points are then (d) collated for each patient, which in turn are (e) collated across all training set patients. To reduce the size of the training set, (f) 500 centroids of the data are identified via k-means clustering. The 400 most-central centroids (•) are kept and the 100 outer-most centroids (•) are removed. In (g) a KDE is fit to the 400 remaining centroids. In (h) the pdf of the KDE is converted into a novelty score, which increases as data moves into the tails of the joint distribution of the vital-signs.

modelled by an isometric kernel density estimate (KDE). The novelty of a new 5-D vital-sign data point is a function of its likelihood with respect to this joint pdf.

B.1 Data Cleaning

In Figure B.1(a) and B.1(b), each patient's time-series was cleaned by removing any artefactual measurements, as determined by Table B.1.

While this approach to artefact removal certainly removes the most extreme artefactual data, as seen in Chapter 5, a threshold approach removes only a minority of all vital-sign measurement artefacts. A particular instance where this may warrant improvement is the SpO_2 threshold of 10%. It is impossible to say *a priori* that an SpO_2 of 10%-70% are artefactual. However, it is highly likely that such measurements are artefactual, especially if transient. (The signal-quality of waveform data at such points would be helpful to make such a decision, but are not available in the UPMC data set.) This leaves many potential artefactual data in the training set. The latter step of removing the 100 farthest centroids attempts to alleviate this issue, as well as remove transient abnormal vital-signs in the training set data.

A further artefactual feature of the UPMC data set involved RR remaining fixed at a single value for an extended period of time. These artefacts were removed as well.

	Lower Threshold	Upper Threshold
HR (bpm)	30	300
SDA (mmHg)	20	180
SpO_2 (%)	10	-
Temp (°C)	32	39
RR (bpm)	3	45

Table B.1: Artefactual thresholds applied to phase 1 patients

B.2 Alignment and Collation of Patient Training Data

In Figure B.1(c), the 5 vital-signs of an individual patient are aligned across time-stamps. These aligned vital-sign measurements are then collated for each patient in

Figure B.1(d) and then collated across all training set patients in Figure B.1(e).

Since the rate of data acquisition varies across each of the 5 vital-signs time-series, the vital-signs must be aligned. This is achieved via a capture-and-hold mechanism, in which the value of a vital-sign is held of a period of time. During this period of time, if a new measurement for that vital-sign is acquired, then the new value is held for the same period of time. If no new measurement is acquired then the vital-sign is considered missing, until a new vital-sign is acquired.

The hold periods are 30 seconds for HR and SpO_2, 1 minute for RR, and 30 minutes for SBP and DBP. The extended hold for SBP and DBP is due to the infrequency of blood pressure measurements.

An aligned patient data set is created as follows: All unique time-stamps in each of the 5 vital-sign time-series were used to create an aligned 5-D vital-sign time-series. For each such time-stamp, if a vital sign did not occur within the capture-and-hold window then it was considered missing at that time-stamp.

To accommodate the very large number of data points for each patient's 5-D time-series, measurements are down-sampled to $\frac{1}{60}$ Hz. Time-stamps were then discarded, since the KDE-based method does not make use of time-series information.

All measurements are collated for all patients in the training set, creating a large, aligned 5-vital-sign training set, shown in Figure B.1(e).

B.3 Data Transformation and Missingness

The mean and standard deviation of each of the 5 vital-signs is calculated. Each vital-sign column (of the 5-vital-sign matrix in Figure B.1(e)) is then transformed to be zero-mean and unit variance. Subsequent to this transformation, any missing vital-sign values were set to the mean (i.e., 0). Individual measurements with three or more missing vital-signs values were removed.

When applied to the test set patients, this convention of (i) scaling, (ii) imputing missing values, and (iii) discarding measurements with excessive missingness, was maintained.

B.4 Representative Centroids of Patient Data

A KDE will be used to estimate the joint density of the 5 vital-signs in the training set. However, it would be infeasible to evaluate such a KDE, with respect to new data, since the training set comprises hundreds of thousands of data points. Such a KDE could not be reasonably employed in a real-time monitoring system.

Therefore, a k-means algorithm is used to find 500 centroids that best represent density of the data comprising hundred of thousands of measurements. (This is the single largest computation of the KDE method.) This is illustrated in Figure B.1(f) with HR and RR measurements transformed back into their original units.

As a least-squares algorithm, many of the k-means centroids only account for outliers and do not represent the points of centrality for most healthy patients. As a step to remove outliers, only the 400 centroids closest to [0,0,0,0,0] are kept. Note that this is not the same as the 400 points nearest to the mean value of the centroids.

An interesting alternative modelling approach would be to perform k-means clustering and artefact removal on a per-patient basis. This might more accurately identify the artefacts by implicitly incorporating the patient-specific context of outlier data.

B.5 Construction of KDE Joint Density Model

From the remaining 400 centroids we now create the KDE, using equation

$$p(\mathbf{x}) = \frac{1}{400\,(2\pi)^{5/2}\,\sigma^5} \sum_{i=1}^{400} \exp\left(-\frac{|\mathbf{x}-\mathbf{x}_i|^2}{2\sigma}\right), \tag{B.1}$$

where $\mathbf{x}$ is the 5-D vector of vital-sign measurements, and $\mathbf{x}_i$ is the i^{th} centroid, from 0 to 400.

The bandwidth parameter σ is calculated according to a rule-of-thumb described in Bishop [141]: For each of the 400 centroids, the 10 closest original data points are identified via K-nearest neighbours. The "variance", σ_i, around centroid $\mathbf{x}_i$ is the average Euclidean distance of $\mathbf{x}_i$ from its 10 nearest neighbours. The bandwidth parameter is then the average of these centroid variances, $\sigma = \frac{1}{400}\sum_{i=1}^{400}\sigma_i$.

By plugging σ into Equation B.1, we now have a joint probability density function over the vital-signs of the healthy patient cohort, as shown in Figure B.1(g). Note that this model has non-zero support for areas corresponding to $SpO_2 > 100\%$. New measurements that are close to the centroids of the healthy group will have high likelihood according to Equation B.1. As a patient deviates from these central points, the likelihood diminishes, indicating a deviation from normality.

B.6 KDE-based Novelty Score

Our formal novelty score will mimic alarm systems, such as NEWS, in which values are high when the patient deviates from physiological normality and low otherwise. This may be achieved by defining the novelty score to be the reciprocal of the log-likelihood of Equation B.1, that is:

$$z(x) = \frac{1}{\log\, p(x)}. \tag{B.2}$$

From Equation B.2, we have a probabilistic metric of patient abnormality, according to how a patient's current vital-signs measurements compare to the current vital-signs of a training set of healthy patients. As shown in Figure B.1(h), the novelty score increases as vital-sign measurements diverge from the centroids.

This score, as described is shown in Figure B.2 for a patient with a C"-event at annotated near hour 23. It can be seen that the patient's vital-signs deviate from normality as they approach the C"-event and the KDE-based novelty score escalates accordingly.

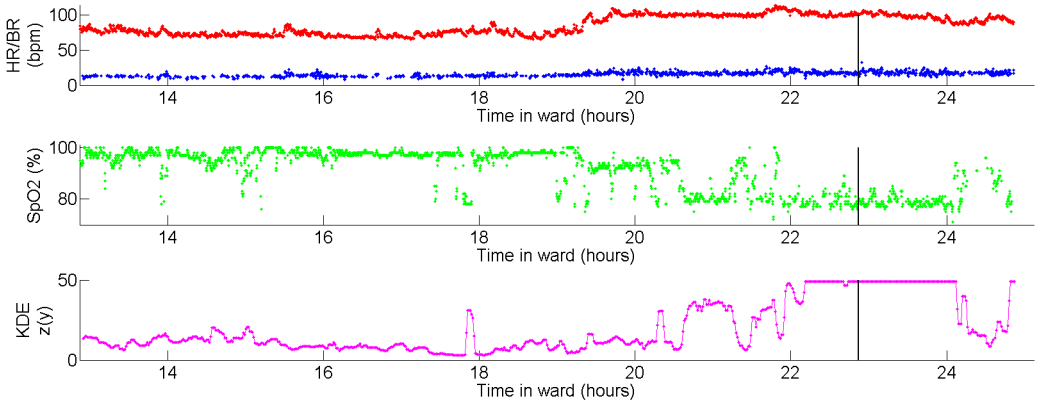


Figure B.2: The KDE-based novelty score $z(x)$ calculated from a patient's time-series. SBP and DBP are not shown. The novelty score $z(x)$ is right-censored at 50 for visual clarity.

C
Alarm Hold Criterion

C.1 Motivation of Alarm Hold Criterion

Each early warning system described produces a single early warning score (EWS) on which to alarm. This EWS can range from the vital-sign itself (for trigger systems) to a composite score across all vitals (for scoring systems) to likelihood-based metrics (for the kernel density and Gaussian process-based systems). An alarm is sounded when the EWS exceeds a predefined alarm threshold (this is distinct from any thresholds used to calculate the score, e.g., the intra-vital-sign thresholds in NEWS).

Typically, the score is required to exceed the alarm threshold for an extended period of time (e.g. exceeding an alarm threshold for 4 of the last 5 minutes) before before presenting an alarm. An alarm status would only cease if the alarm criteria were not met for a period of time as well (e.g. falling below an alarm threshold for 2 of the last 3 minutes). This is in contrast to sounding an alarm at the first instance of a vital-sign measurement or risk score is in exceedance of the threshold.

The choice of m_1 over the last m_2 minutes is a heuristic choice in its own right.

The intuition behind this approach is that vital-sign measurements are typically fraught with artefactual measurements that are unrepresentative of the patientś current physiology. Requiring that a threshold be exceeded for an extended period of time attempts to reduce false alarms caused by these artefacts. This is equivalent

to the alarm being "muted" until the abnormal vital-signs have persisted for a period of time.

Since the implementation of such a system may vary by the acquisition rate of the EWS, it is unclear why this heuristic approach would be preferable over an alternative (and trivial to code) heuristic such as median smoothing of the warning score over a small time window.

C.2 Alarm Hold Criterion for Baseline Comparator Methods

For a fair comparison to clinical practice, all baseline comparator methods were tried both with and without a "four minutes out of five" exceedance requirement.

For both the simple thresholding and the KDE-based novelty score this extra step either (i) did not improve performance or (ii) decreased performance, in terms of the TEW vs. FPR trade-off. To corroborate or contradict this finding, a literature search was conducted but unable to find any published research which compared early warning performance with and without this heuristic.

In the absence of reasons to do otherwise, the results in the baseline comparator chapter only display thresholding results that did not use a holding heuristic, i.e., the best-found results for the baseline methods.

C.3 Alarm Hold Criterion for GP-based Step-Change Methods

Alarm holds were neither used nor tested for GP-based step-change detection since the step-changes are inherently transitory. Therefore the alarm score from these methods do not persist long enough to make alarm holds a viable option.

To improve alignment of step-change scores across vital-signs, it may be helpful to hold the highest NLML value for each vital-sign (i.e. the value most-indicative of step-change in that vital-sign). This was examined, and resulted in an improved

tradeoff between TEW and FPR. However, these results are not presented to reduce the research degrees of freedom in the results.

References

[1] Glen Wright Colopy, Stephen Roberts, and David A. Clifton. "Bayesian Optimisation of Personalised Models for Patient Vital-Sign Monitoring". In: *IEEE Journal of Biomedical and Health Informatics* 22 (2 2018), pp. 301–310.

[2] Glen Wright Colopy et al. "Likelihood-based artefact detection in continuously-acquired patient vital signs". In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2017).

[3] Glen Wright Colopy et al. "Bayesian optimisation of Gaussian processes for identifying the deteriorating patient". In: *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (2017).

[4] Glen Wright Colopy et al. "State-space approximations to Gaussian processes for patient vital-sign monitoring in computationally-constrained clinical environments". In: *MEIBioeng 16* (2016).

[5] Glen Wright Colopy et al. "Bayesian Gaussian processes for identifying the deteriorating patient". In: *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2016), pp. 5311–5314.

[6] David A. Clifton, Marco A.F.P. Pimentel, and Glen Wright Colopy. *System Monitor and Method of System Monitoring*. GB 1613318.3, 2016. URL: http://www.robots.ox.ac.uk/~davidc/pubs/patent_dictionaries.pdf.

[7] Glen Wright Colopy et al. "Vital-Sign Fusion for Novelty-based Detection of Imminent Hypoglycaemia". In: *In review* ().

[8] Glen Wright Colopy and David A. Clifton. "Personalised Volatility Metrics for Temperature Time-Series of Post-Vaccination Infants". In: *In preparation* ().

[9] Tingting Zhu et al. "Modelling Patient-Specific Trajectories Using Hierarchical Bayesian Gaussian Processes". In: *IEEE Journal of Biomedical and Health Informatics* (2018).

[10] Tingting. Zhu et al. "Identifying Patient-Specific Trajectories in Haemodialysis Using Bayesian Hierarchical Gaussian Processes". In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (2018).

[11] David A. Clifton et al. "Health Informatics via Machine Learning for the Clinical Management of Patients". In: *IMIA Yearbook* 10.1 (2015), pp. 38–43.

[12] Theodore R. Rieger et al. "Improving the Generation and Selection of Virtual Populations in Quantitative Systems Pharmacology Models". In: *bioRxiv* (2018). eprint: https://www.biorxiv.org/content/early/2018/01/18/196089.full.pdf. URL: https://www.biorxiv.org/content/early/2018/01/18/196089.

[13] Ruth Clifford et al. “SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage”. In: *Blood* 123.7 (2013), pp. 1021–1031.

[14] Katherine Niehaus. “Phenotypic Modelling of Crohn’s Disease Severity: A Machine Learning Approach”. PhD thesis. University of Oxford, 2016.

[15] Farlex Partner Medical Dictionary. “Step-down unit”. In: (2012). URL: http://medical-dictionary.thefreedictionary.com/step-down+unit (visited on 06/02/2015).

[16] Marilyn Hravnak et al. “Defining the incidence of cardiorespiratory instability in patients in step-down units using an electronic integrated monitoring system”. In: *Archives of Internal Medicine* 168.12 (2008), pp. 1300–1308. eprint: /data/journals/intemed/5682/ioi80028_1300_1308.pdf. URL: +%20http://dx.doi.org/10.1001/archinte.168.12.1300.

[17] Neill KJ Adhikari et al. “Critical care and the global burden of critical illness in adults”. In: *The Lancet* 376.9749 (2010), pp. 1339–1346.

[18] Elizabeth A. Martinez et al. “Identifying Meaningful Outcome Measures for the Intensive Care Unit”. In: *American Journal of Medical Quality* 29.2 (2013), pp. 144–152.

[19] Hannah Wunsch et al. “Variation in critical care services across North America and Western Europe*”. In: *Critical Care Medicine* 36.10 (2008), 2787–e8.

[20] Jack E Zimmerman, Andrew A Kramer, and William A Knaus. “Changes in hospital mortality for United States intensive care unit admissions from 1988 to 2012”. In: *Critical Care* 17.2 (2013), R81.

[21] K. Yousef et al. “Characteristics of Patients With Cardiorespiratory Instability in a Step-down Unit”. In: *American Journal of Critical Care* 21.5 (2012), pp. 344–350.

[22] Andrew D. Harding. “What Can An Intermediate Care Unit Do For You?” In: *JONA: The Journal of Nursing Administration* 39.1 (2009), pp. 4–7.

[23] Neil A. Halpern and Stephen M. Pastores. “Critical care medicine in the United States 2000-2005: An analysis of bed numbers, occupancy rates, payer mix, and costs*”. In: *Critical Care Medicine* 38.1 (2010), pp. 65–71.

[24] Craig M. Coopersmith et al. “A comparison of critical care research funding and the financial burden of critical illness in the United States*”. In: *Critical Care Medicine* 40.4 (2012), pp. 1072–1079.

[25] Gregory S. Cooper et al. “Are Readmissions to the Intensive Care Unit a Useful Measure of Hospital Performance?” In: *Medical Care* 37.4 (1999), pp. 399–408.

[26] A. J. Campbell et al. “Predicting death and readmission after intensive care discharge”. In: *British Journal of Anaesthesia* 100.5 (2008), pp. 656–662.

[27] Annemie Vlayen et al. “Incidence and preventability of adverse events requiring intensive care admission: a systematic review”. In: *Journal of Evaluation in Clinical Practice* 18.2 (2011), pp. 485–497.

[28] Eliezer L. Bose et al. “Cardiorespiratory instability in monitored step-down unit patients: using cluster analysis to identify patterns of change”. In: *Journal of Clinical Monitoring and Computing* 32.1 (Feb. 2018), pp. 117–126. URL: https://doi.org/10.1007/s10877-017-0001-7.

[29] M. D Buist. “Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study”. In: *BMJ* 324.7334 (2002), pp. 387–390.

[30] Helen Hogan et al. “Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study”. In: *BMJ Quality & Safety* (2012). eprint: http://qualitysafety.bmj.com/content/early/2012/07/06/bmjqs-2012-001159.full.pdf. URL: http://qualitysafety.bmj.com/content/early/2012/07/06/bmjqs-2012-001159.

[31] Liam J. Donaldson, Sukhmeet S. Panesar, and Ara Darzi. “Patient-Safety-Related Hospital Deaths in England: Thematic Analysis of Incidents Reported to a National Database, 2010-2012”. In: *PLOS Medicine* 11.6 (June 2014), pp. 1–8.

[32] Jack E. Zimmerman et al. “Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today’s critically ill patients*”. In: *Critical Care Medicine* 34.5 (2006), pp. 1297–1310.

[33] A. Hutchings et al. “Evaluation of modernisation of adult critical care services in England: time series and cost effectiveness analysis”. In: *BMJ* 339.nov11 2 (2009), b4353–b4353.

[34] Marilyn Hravnak et al. “Cardiorespiratory instability before and after implementing an integrated monitoring system*”. In: *Critical Care Medicine* 39.1 (2011), pp. 65–72.

[35] R.C. Bone et al. “Definitions of sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine”. In: 101 (July 1992), pp. 1644–55.

[36] Mitchell M. Levy et al. “2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference”. In: *Intensive Care Medicine* 29.4 (2003), pp. 530–538.

[37] Jean-Louis Vincent et al. “Sepsis definitions: time for change”. In: *The Lancet* 381.9868 (2013), pp. 774–775.

[38] Singer M et al. “The third international consensus definitions for sepsis and septic shock (sepsis-3)”. In: *JAMA* 315.8 (2016), pp. 801–810. eprint: /data/journals/jama/935012/jsc160002.pdf. URL: +%20http://dx.doi.org/10.1001/jama.2016.0287.

[39] Alexandros Pantelopoulos and Nikolaos G. Bourbakis. “A Survey on Wearable Sensor-Based Systems for Health Monitoring and Prognosis”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40.1 (2010), pp. 1–12.

[40] Alistair E. W. Johnson et al. “Machine Learning and Decision Support in Critical Care”. In: *Proceedings of the IEEE* 104.2 (Feb. 2016), pp. 444–466.

[41] *National Early Warning Score (NEWS): Standardising the assessment of acute-illness severity in the NHS*. Tech. rep. London RCP: Royal College of Physicians, July 2012. URL: www.rcplondon.ac.uk.

[42] *National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS*. Tech. rep. London: RCP: Royal College of Physicians, Dec. 2017. URL: www.rcplondon.ac.uk.

[43] Emanuel Parzen. "On Estimation of a Probability Density Function and Mode". In: *The Annals of Mathematical Statistics* 33.3 (1962), pp. 1065–1076.

[44] Murray Rosenblatt. "Remarks on Some Nonparametric Estimates of a Density Function". In: *The Annals of Mathematical Statistics* 27.3 (1956), pp. 832–837.

[45] Alistair Hann. "Multi-parameter monitoring for early warning of patient deterioration". PhD thesis. University of Oxford, 2008.

[46] D. Christopher Bouch and Jonathan P. Thompson. "Severity scoring systems in the critically ill". In: *Continuing Education in Anaesthesia Critical Care & Pain* 8.5 (2008), pp. 181–185.

[47] Marco A. F. Pimentel et al. "Review: A Review of Novelty Detection". In: *Signal Process.* 99 (2014), pp. 215–249.

[48] Lei Clifton et al. "Gaussian Processes for Personalized e-Health Monitoring With Wearable Sensors". In: *IEEE Transactions on Biomedical Engineering* 60.1 (Jan. 2013), pp. 193–197.

[49] R Schoenberg, Daniel Sands, and Charles Safran. "Making ICU alarms meaningful: a comparison of traditional vs. trend-based algorithms". In: (Feb. 1999), pp. 379–83.

[50] Nienke Seiger et al. "Validity of Different Pediatric Early Warning Scores in the Emergency Department". In: *Pediatrics* 132.4 (2013), e841–e850.

[51] Gary B. Smith et al. "A review, and performance evaluation, of single-parameter 'track and trigger' systems". In: *Resuscitation* 79.1 (2008), pp. 11–21.

[52] Gary B. Smith et al. "Review and performance evaluation of aggregate weighted 'track and trigger' systems". In: *Resuscitation* 77.2 (2008), pp. 170–179.

[53] Stuart Jarvis et al. "Aggregate National Early Warning Score (NEWS) values are more important than high scores for a single vital signs parameter for discriminating the risk of adverse outcomes". In: *Resuscitation* 87 (2015), pp. 75–80.

[54] C. Stenhouse et al. "Prospective evaluation of a modified Early Warning Score to aid earlier detection of patients developing critical illness on a general surgical ward". In: *British Journal of Anaesthesia* 84.5 (2000), 663P. URL: https://www.sciencedirect.com/science/article/pii/S0007091217380480.

[55] Alex Psirides, Jennifer Hill, and Sally Hurford. "A review of rapid response team activation parameters in New Zealand hospitals". In: *Resuscitation* 84.8 (2013), pp. 1040–1044.

[56] Alex Psirides and Anne Pedersen. *Proposal for a National New Zealand early warning score & vital signs chart*. Tech. rep. Wellington Regional Hospital, Oct. 2015. URL: wellingtonicu.com.

[57] Alan Monaghan. “Detecting and managing deterioration in children”. In: *Paediatric Care* 17.1 (2005), pp. 32–35.

[58] William A. Knaus et al. “APACHE - acute physiology and chronic health evaluation: a physiologically based classification system”. In: *Critical Care Medicine* 9.8 (1981), pp. 591–597.

[59] W.A. Knaus et al. “APACHE II: a severity of disease classification system”. In: *Crit. Care Med.* 13.10 (1985), pp. 818–829.

[60] William A. Knaus et al. “The APACHE III Prognostic System: Risk Prediction of Hospital Mortality for Critically III Hospitalized Adults”. In: *Chest* 100.6 (1991), pp. 1619–1636.

[61] Jean-Roger Le Gall et al. “A simplified acute physiology score for ICU patients”. In: *Critical Care Medicine* 12.11 (1984), pp. 975–977.

[62] Jean-Roger Le Gall, S Lemeshow, and F Saulnier. “A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study”. In: *JAMA: The Journal of the American Medical Association* 270.24 (1993), p. 2957.

[63] Rui P. Moreno et al. “SAPS3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission”. In: *Intensive Care Medicine* 31.10 (2005), pp. 1345–1355.

[64] Gary B. Smith et al. “Should age be included as a component of track and trigger systems used to identify sick adult patients?” In: *Resuscitation* 78.2 (2008), pp. 109–115.

[65] Mikkel Brabrand and John Kellett. “Mobility measures should be added to the National Early Warning Score (NEWS)”. In: *Resuscitation* 85.9 (2014), e151.

[66] Tom E.F. Abbott et al. “A single-centre cohort study of National Early Warning Score (NEWS) and near patient testing in acute medical admissions”. In: *European Journal of Internal Medicine* 35 (2016), pp. 78–82.

[67] Christian H. Nickel et al. “Combined use of the National Early Warning Score and D-dimer levels to predict 30-day and 365-day mortality in medical patients”. In: *Resuscitation* 106 (2016), pp. 49–52.

[68] John Kellett and Alan Murray. “Should predictive scores based on vital signs be used in the same way as those based on laboratory data? A hypothesis generating retrospective evaluation of in-hospital mortality by four different scoring systems”. In: *Resuscitation* 102 (2016), pp. 94–97.

[69] Jennifer McGaughey et al. “Outreach and Early Warning Systems (EWS) for the prevention of Intensive Care admission and death of critically ill adult patients on general hospital wards”. In: *Cochrane Database of Systematic Reviews* (2007).

[70] Bente Bilben, Linda Grandal, and Signe Søvik. “National Early Warning Score (NEWS) as an emergency department predictor of disease severity and 90-day survival in the acutely dyspneic patient - a prospective observational study”. In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 24.1 (June 2016), p. 80.

[71] C. P. Subbe et al. “Effect of introducing the Modified Early Warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions”. In: *Anaesthesia* 58.8 (2003), pp. 797–802.

[72] Anne L. Solevag et al. “Use of a Modified Pediatric Early Warning Score in a Department of Pediatric and Adolescent Medicine”. In: *PLOS ONE* 8.8 (Aug. 2013), pp. 1–6.

[73] Patricia Benner and Christine Tanner. “Clinical Judgment: How Expert Nurses Use Intuition”. In: *The American Journal of Nursing* 87.1 (1987), p. 23.

[74] David A. Clifton et al. “‘Errors’ and omissions in paper-based early warning scores: the association with changes in vital signs—a database analysis”. In: *BMJ Open* 5.7 (2015).

[75] Gooske Douw et al. “Nurses’ worry or concern and early recognition of deteriorating patients on general wards in acute care hospitals: a systematic review”. In: *Critical Care* 19.1 (2015).

[76] Priscilla K. Gazarian, Elizabeth A. Henneman, and Genevieve E. Chandler. “Nurse Decision Making in the Prearrest Period”. In: *Clinical Nursing Research* 19.1 (2009), pp. 21–37.

[77] Vickie K. Fieler, Thomas Jaglowski, and Karen Richards. “Eliminating Errors in Vital Signs Documentation”. In: *CIN: Computers, Informatics, Nursing* 31.9 (2013), pp. 422–427.

[78] Pauline Gearing et al. “Enhancing patient safety through electronic medical record documentation of vital signs”. In: 20 (Feb. 2006), pp. 40–5.

[79] A. F. Smith and R. J. Oakey. “Incidence and significance of errors in a patient ‘track and trigger’ system during an epidemic of Legionnaires’ disease: retrospective casenote analysis”. In: *Anaesthesia* 61.3 (2006), pp. 222–228.

[80] Douglas G. Altman et al. “Dangers of Using ‘Optimal’ Cutpoints in the Evaluation of Prognostic Factors”. In: *JNCI: Journal of the National Cancer Institute* 86.11 (1994), pp. 829–835.

[81] Patrick Royston, Douglas G. Altman, and Willi Sauerbrei. “Dichotomizing continuous predictors in multiple regression: a bad idea”. In: *Statistics in Medicine* 25.1 (2006), pp. 127–141.

[82] Patrick Royston and Douglas G. Altman. “Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 43.3 (1994), pp. 429–467.

[83] Stanley Lemeshow et al. “A method for predicting survival and mortality of ICU patients using objectively derived weights”. In: *Critical Care Medicine* 13.7 (1985), pp. 519–525.

[84] S. Lemeshow. “Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients”. In: *JAMA: The Journal of the American Medical Association* 270.20 (1993), pp. 2478–2486.

[85] Thomas L. Higgins et al. “Assessing contemporary intensive care unit outcome: An updated Mortality Probability Admission Model (MPM0-III)*”. In: *Critical Care Medicine* 35.3 (2007), pp. 827–835.

[86] Raith E. P. et al. “Prognostic accuracy of th.e SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit”. In: *JAMA* 317.3 (2017), pp. 290–300.

[87] Ferreira F. et al. “Serial evaluation of the SOFA score to predict outcome in critically ill patients”. In: *JAMA* 286.14 (2001), pp. 1754–1758.

[88] Lamontagne F., Harrison D. A., and Rowan K. M. “qSOFA for identifying sepsis among patients with infection”. In: *JAMA* 317.3 (2017), pp. 267–268.

[89] Lionel Tarassenko et al. “BIOSIGN™: multi-parameter monitoring for early warning of patient deterioration”. In: *The 3rd IEE International Seminar on Medical Applications of Signal Processing 2005.* Nov. 2005, pp. 71–76.

[90] Timothy Bonnici et al. “Evaluation of the effects of implementing an electronic early warning score system: protocol for a stepped wedge study”. In: *BMC Medical Informatics and Decision Making* 16.1 (2015).

[91] Lionel Tarassenko et al. “Centile-based early warning scores derived from statistical distributions of vital signs”. In: *Resuscitation* 82.8 (2011), pp. 1013–1018.

[92] S. J. Roberts. “Extreme value statistics for novelty detection in biomedical data processing”. In: *IEE Proceedings - Science, Measurement and Technology* 147.6 (Nov. 2000), pp. 363–367.

[93] Samuel Hugueny. “Novelty Detection with Extreme Value Theory in Vital-Sign Monitoring”. PhD thesis. University of Oxford, 2013.

[94] David A. Clifton, Samuel Hugueny, and Lionel Tarassenko. “Novelty detection with multivariate Extreme Value Theory, part I: A numerical approach to multimodal estimation”. In: *2009 IEEE International Workshop on Machine Learning for Signal Processing.* Sept. 2009, pp. 1–6.

[95] David A. Clifton, Samuel Hugueny, and Lionel Tarassenko. “Pinning the tail on the distribution: A multivariate extension to the generalised Pareto distribution”. In: *2011 IEEE International Workshop on Machine Learning for Signal Processing.* Sept. 2011, pp. 1–6.

[96] David A. Clifton, Samuel Hugueny, and Lionel Tarassenko. “Novelty Detection with Multivariate Extreme Value Statistics”. In: *Journal of Signal Processing Systems* 65.3 (Dec. 2011), pp. 371–389.

[97] Marco A.F. Pimentel et al. “Modelling Patient Time-Series Data from Electronic Health Records using Gaussian Processes”. In: *NIPS 2013.* 2014.

[98] Kazuki Yamamoto, Yutaka Watanobe, and Wenxi Chen. “Clustering Analysis of Vital Signs Measured During Kidney Dialysis”. In: *Trends in Applied Knowledge-Based Systems and Data Science.* Ed. by Hamido Fujita et al. Cham: Springer International Publishing, 2016, pp. 503–513.

[99] T. Schmidt et al. “Clustering Emergency Department patients - an assessment of group normality”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* Aug. 2015, pp. 6824–6829.

[100] Tao Shi and Steve Horvath. "Unsupervised Learning With Random Forest Predictors". In: *Journal of Computational and Graphical Statistics* 15.1 (2006), pp. 118–138. eprint: https://doi.org/10.1198/106186006X94072. URL: https://doi.org/10.1198/106186006X94072.

[101] Eliezer Bose. "Time series analysis and clustering to characterize cardiorespiratory instability patterns in step-down unit patients". PhD thesis. University of Pittsburgh, 2015.

[102] Tessy Badriyah et al. "Decision-tree early warning score (DTEWS) validates the design of the National Early Warning Score (NEWS)". In: *Resuscitation* 85.3 (2014), pp. 418–423.

[103] Stuart W. Jarvis et al. "Development and validation of a decision tree early warning score based on routine laboratory test results for the discrimination of hospital mortality in emergency medical admissions". In: *Resuscitation* 84.11 (2013), pp. 1494–1499.

[104] Michael Xu et al. "A protocol for developing early warning score models from vital signs data in hospitals using ensembles of decision trees". In: *BMJ Open* 5.9 (2015).

[105] M Xu et al. "Validation of an electronic early warning score using decision tree analysis: proposal". In: *Critical Care* 19.Suppl 1 (2015), P503.

[106] Marilyn Hravnak et al. "Machine learning can classify vital sign alerts as real or artifact in online continuous monitoring data". In: *Intensive Care Medicine Experimental* 3.1 (Oct. 2015), A550.

[107] Lei Clifton et al. "Gaussian process regression in vital-sign early warning systems". In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Aug. 2012, pp. 6161–6164.

[108] Eren Gultepe et al. "From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system". In: *Journal of the American Medical Informatics Association* (2013), pp. 315–325.

[109] R. Dürichen et al. "Multi-task Gaussian process models for biomedical applications". In: *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. June 2014, pp. 492–495.

[110] R. Dürichen et al. "Multitask Gaussian Processes for Multivariate Physiological Time-Series Analysis". In: *IEEE Transactions on Biomedical Engineering* 62.1 (Jan. 2015), pp. 314–322.

[111] David Wong, David A. Clifton, and Lionel Tarassenko. "Probabilistic detection of vital sign abnormality with Gaussian process regression". In: *2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE)*. Nov. 2012, pp. 187–192.

[112] M. A. F. Pimentel et al. "Probabilistic estimation of respiratory rate using Gaussian processes". In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. July 2013, pp. 2902–2905.

[113] M. A. F. Pimentel, D. A. Clifton, and L. Tarassenko. "Gaussian process clustering for the functional characterisation of vital-sign trajectories". In: *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. Sept. 2013, pp. 1–6.

[114] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data". In: *PLOS ONE* 8.6 (June 2013), pp. 1–13.

[115] O. Stegle et al. "Gaussian Process Robust Regression for Noisy Heart Rate Data". In: *IEEE Transactions on Biomedical Engineering* 55.9 (Sept. 2008), pp. 2143–2151.

[116] Eliezer Bose, Marilyn Hravnak, and Susan M. Sereika. "Vector Autoregressive Models and Granger Causality in Time Series Analysis in Nursing Research". In: *Nursing Research* 66.1 (2017), pp. 12–19.

[117] David Duvenaud et al. "Structure Discovery in Nonparametric Regression Through Compositional Kernel Search". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13. Atlanta, GA, USA: JMLR.org, 2013, pp. III-1166–III-1174.

[118] David Kristjanson Duvenaud. "Automatic Model Construction with Gaussian Processes". PhD thesis. University of Cambridge, 2014.

[119] Roman Garnett, Michael A. Osborne, and Stephen J. Roberts. "Sequential Bayesian Prediction in the Presence of Changepoints". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML 2009. Montreal, Quebec, Canada: ACM, 2009, pp. 345–352.

[120] Yunus Saatçi, Ryan Turner, and Carl Edward Rasmussen. "Gaussian Process Change Point Models". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML 2010. Haifa, Israel: Omnipress, 2010, pp. 927–934.

[121] Edward Snelson, Zoubin Ghahramani, and Carl E. Rasmussen. "Warped Gaussian Processes". In: *Advances in Neural Information Processing Systems 16*. Ed. by S. Thrun, L. K. Saul, and B. Schölkopf. MIT Press, 2004, pp. 337–344.

[122] Jasper Snoek et al. "Input Warping for Bayesian Optimization of Non-stationary Functions". In: *Proceedings of the 31st International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, 2014, pp. II-1674–II-1682.

[123] Andrew G Wilson and Zoubin Ghahramani. "Copula Processes". In: *Advances in Neural Information Processing Systems 23*. Ed. by J. D. Lafferty et al. Curran Associates, Inc., 2010, pp. 2460–2468.

[124] Andrew McHutchon and Carl Edward Rasmussen. "Gaussian Process Training with Input Noise". In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS'11. Granada, Spain: Curran Associates Inc., 2011, pp. 1341–1349.

[125] James Hensman, Nicolo Fusi, and Neil D. Lawrence. "Gaussian Processes for Big Data". In: (Sept. 2013).

[126] Edward Snelson and Zoubin Ghahramani. "Sparse Gaussian Processes Using Pseudo-inputs". In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. NIPS'05. Vancouver, British Columbia, Canada: MIT Press, 2005, pp. 1257–1264.

[127] David J.C. MacKay. *The Humble Gaussian Distribution*. [Online; posted 11-June-2006]. June 2006. URL: http://www.inference.org.uk/mackay/humble.pdf.

[128] Mark Ebden. "Gaussian Processes: A Quick Introduction". In: *ArXiv e-prints* (May 2015). arXiv: 1505.02965 [math.ST].

[129] David Duvenaud. *The Kernel Cookbook: Advice on Covariance functions*. URL: http://www.cs.toronto.edu/~duvenaud/cookbook/.

[130] Jouni Hartikainen and Simo Särkkä. "Kalman filtering and smoothing solutions to temporal Gaussian process regression models". In: *2010 IEEE International Workshop on Machine Learning for Signal Processing*. Aug. 2010, pp. 379–384.

[131] S. Reece and S. Roberts. "An introduction to Gaussian processes for the Kalman filter expert". In: *2010 13th International Conference on Information Fusion*. July 2010, pp. 1–9.

[132] Ryan Darby Turner. "Gaussian Processes for State Space Models and Change Point Detection". PhD thesis. University of Cambridge, 2011.

[133] B. Shahriari et al. "Taking the Human Out of the Loop: A Review of Bayesian Optimization". In: *Proceedings of the IEEE* 104.1 (Jan. 2016), pp. 148–175.

[134] I. Murray, R.P. Adams, and D.J.C. MacKay. "Elliptical slice sampling". In: *Proc 13 International Conference AISTATS* 9 (2010), pp. 541–548.

[135] Tiangang Cui et al. *Using Parallel MCMC Sampling to Calibrate a Computer Model of a Geothermal Reservoir*. Jan. 2011.

[136] A. O'Hagan. "Monte Carlo is Fundamentally Unsound". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 36.2/3 (1987), pp. 247–249.

[137] M. A. Osborne et al. "Towards Real-Time Information Processing of Sensor Network Data Using Computationally Efficient Multi-output Gaussian Processes". In: *2008 International Conference on Information Processing in Sensor Networks (ISPN 2008)*. Apr. 2008, pp. 109–120.

[138] Michalis K. Titsias. "Variational learning of inducing variables in sparse Gaussian processes". In: *In Artificial Intelligence and Statistics 12*. 2009, pp. 567–574.

[139] H. W. Sorenson. "Least-squares estimation: from Gauss to Kalman". In: *IEEE Spectrum* 7.7 (July 1970), pp. 63–68.

[140] M. A. Osborne, R Garnett, and S. J. Roberts. "Gaussian processes for global optimization". In: *Learning and Intelligent Optimization Conference*. Vol. 3. 2009.

[141] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, 2006.

[142] G.M. Friesen et al. "A comparison of the noise sensitivity of nine QRS detection algorithms". In: *IEEE Transactions on Biomedical Engineering* 37.1 (1990), pp. 85–98.

[143] J.A. Quinn, C.K.I. Williams, and N. McIntosh. "Factorial Switching Linear Dynamical Systems Applied to Physiological Condition Monitoring". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.9 (2009), pp. 1537–1551.

[144] J. Vanhatalo et al. "Bayesian Modeling with Gaussian Processes using the GPstuff Toolbox". In: *ArXiv e-prints* (June 2012). arXiv: 1206.5754 [stat.ML].

[145] J. Bergstra and Y. Benjio. "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research* 13 (2012), pp. 281–305.

[146] Francisco J. Solis and Roger J.-B. Wets. "Minimization by Random Search Techniques". In: *Mathematics of Operations Research* 6.1 (1981), pp. 19–30.

[147] C. E. Rasmussen and H. Nickisch. "Gaussian Processes for Machine Learning (GPML) Toolbox". In: *Journal of Machine Learning Research* 11 (Nov. 2010), pp. 3011–3015.

[148] Edward Snelson, Zoubin Ghahramani, and Carl E. Rasmussen. "Warped Gaussian Processes". In: *Advances in Neural Information Processing Systems*.