Bayesian Condition Monitoring in Neonatal Intensive Care

John A. Quinn

Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2008

Abstract

The observed physiological dynamics of an infant receiving intensive care contain a great deal of information about factors which cannot be examined directly, including the state of health of the infant and the operation of the monitoring equipment. This type of data tends to contain both common, recognisable patterns (e.g. as caused by certain clinical operations or artifacts) and some which are rare and harder to interpret. The problem of identifying the presence of these patterns using prior knowledge is clinically significant, and one which is naturally described in terms of statistical machine learning.

In this thesis I develop probabilistic dynamical models which are capable of making useful inferences from neonatal intensive care unit monitoring data. The Factorial Switching Kalman Filter (FSKF) in particular is adopted as a suitable framework for monitoring the condition of an infant. The main contributions are as follows: (1) the application of the FSKF for inferring common factors in physiological monitoring data, which includes finding parameterisations of linear dynamical models to represent common physiological and artifactual conditions, and adapting parameter estimation and inference techniques for the purpose; (2) the formulation of a model for novel physiological dynamics, used to infer the times in which something is happening which is not described by any of the known patterns. EM updates are derived for the latter model in order to estimate parameters. Experimental results are given which show the developed methods to be effective on genuine monitoring data.

Acknowledgements

Chris Williams, my principal supervisor, has continually and energetically stimulated my progress, and provided much technical insight. Neil McIntosh has also been a reliable source of encouragement and investigative foresight. I have particularly appreciated both of them conveying the enjoyment of doing research.

Ian Nabney and Simon King provided many helpful comments while examining this thesis. Thanks also to Amos Storkey for assistance on my thesis review panel. I have had the benefit of useful discussions and comments from others including Felix Agakov, Yvonne Freer, Sam Roweis, Charles Fox, David MacKay, Birgit Wefers, François Portet, and many others in both the machine learning group at the Institute for Adaptive and Neural Computation and the neonatal unit at the Royal Infirmary of Edinburgh.

Some of Alex Spengler's baby monitoring code, especially for visualisation, has been useful. Peter Badger provided assistance with NICU monitoring systems. Jim Hunter adapted a version of the Time Series Workbench software for this project, for which I am grateful; Neil McIntosh and Birgit Wefers then spent time poring over it in order to supply the annotations which made experimental evaluation possible.

I am also grateful to the premature baby charity BLISS for research funding.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*John A. Quinn*)

Table of Contents

Notation and abbreviations

Mathematical and statistical notation

Capital letters in bold (e.g. $\mathbf{A}, \mathbf{B}$) denote matrices. Lower case bold letters (e.g. $\mathbf{x}, \mathbf{y}$) denote vectors, and normal lower case letters (e.g. a, b) denote scalar values.

$\sim$	Distributed as
$\propto$	Proportional to
$\approx$	Approximately equal to
$\langle \cdot \rangle_{\mathbf{x}}$	Expectation over the variable $\mathbf{x}$
∇	Difference operator, such that $\nabla y_t = y_t - y_{t-1}$
$\det \mathbf{A}$	Determinant of the matrix $\mathbf{A}$
$\mathbf{A}^\top$	Transpose of the matrix $\mathbf{A}$
$\lvert \cdot \rvert$	Absolute value, magnitude of a complex quantity
$\otimes$	Cross-product
γ_k	Autocovariance function at lag k
ρ_k	Autocorrelation function at lag k
cdf	Cumulative distribution function
pdf	Probability density function
i.i.d.	Independent and identically distributed
ACF	Autocorrelation function
PACF	Partial autocorrelation function
$\mathcal{N}(\mu, \Sigma)$	Normal distribution with mean μ and covariance Σ
$S_x(f)$	Power spectral density of sequence $x_{1:T}$ at frequency f
AR(p)	Autoregressive process of order p

Notation for linear dynamical systems

$\mathbf{A}$	System matrix
$\mathbf{Q}$	System noise covariance
$\mathbf{C}$	Observation matrix
$\mathbf{R}$	Observation noise covariance
$\mathbf{d}$	Drift term
$\mathbf{x}_t$	Hidden state at time t
$\mathbf{y}_t$	Observation at time t
s_t	Switch setting at time t
M	Number of factors in a factorial model
$K^{(m)}$	Number of possible settings for factor m
K	Total number of switch settings
T	Total number of data points in a sequence

For switching models, a parameter with a superscript indicates the value associated with a particular switch setting, e.g. $\mathbf{A}^{(s)}$ is the system matrix for switch setting s. Variables used without subscripts denote an entire sequence, e.g. $\mathbf{x} \equiv \mathbf{x}_{1:T}$.

Kalman filter notation

$\hat{\mathbf{x}}_t$	Estimated mean of latent variable $\mathbf{x}$ at time t
$\mathbf{P}_t$	Covariance of estimate $\hat{\mathbf{x}}_t$
$\tilde{\mathbf{y}}_t$	Innovation at time t
$\mathbf{S}_t$	Innovation covariance at time t
$\mathbf{K}_t$	Kalman gain at time t
μ	Initial state estimate
Σ	Covariance of initial state estimate

Physiological measurements

Core temp.	Core body temperature (°C)
Dia. Bp	Diastolic blood pressure (mmHg)
Sys. Bp	Systolic blood pressure (mmHg)
HR	Heart rate (bpm)
Periph. temp.	Peripheral body temperature (°C)
SpO_2	Saturation of oxygen in pulse (%)
$TcPCO_2$	Transcutaneous partial pressure of CO_2 (kPa)
$TcPO_2$	Transcutaneous partial pressure of O_2 (kPa)

Chapter 1

Introduction

Babies who are born three or four months prematurely are usually in a fragile condition, and one which is surrounded by uncertainty. One important function of a neonatal intensive care unit (NICU) is to try and reduce this uncertainty, so that the right treatment can be given and unwelcome surprises can be avoided. Clinical staff are able to continually articulate their doubts in abstract terms in order to get a handle on the situation. Is that baby stable, or is its condition deteriorating? Should something be done about this baby's decreasing blood pressure? Is that cardiovascular problem getting better or worse? These questions are often difficult to answer.

In general when faced with uncertainty we rely on models, mental or otherwise. Starting with beliefs about what is likely to be happening based on previous experience, we can try to obtain evidence and then use it to update these prior beliefs according to our model of the situation. Some items of evidence reduce uncertainty about particular questions to a greater extent than others. Of the observations gathered about patients in intensive care, *vital signs* provide evidence about physiological function at the most fundamental level. These typically include measurements of the heart rate, blood pressure, temperature, and the quantities of gases dissolved in the bloodstream, up to once per second. Although continuous vital signs monitoring provides a lot of information about a baby's state of health, it is generally agreed that the current technology has important limitations. Problems include the crude way in which alarms are set, and the high false alarm rate; corruption of the data by artifact; the overload of low-level data; and the difficulty of interpretation for inexperienced staff.

The goal of this thesis is to show how statistical models incorporating expert knowledge can be applied to such data in order to overcome some of these problems, producing inferences which help to reduce uncertainty about the condition of the baby.

1.1 Vital signs monitoring in neonatal intensive care

Vital signs monitoring is reviewed in more detail in Chapter 2, but is briefly introduced in this section along with a description of the shortcomings of current technology.

The types of babies considered here are around 26-28 weeks gestation, highly premature compared to the full term gestation of 40 weeks. These infants normally have very low birthweight (VLBW), less than 1500g, and will tend to rely on mechanical ventilation and an environment in which the temperature and humidity are finely controlled. Often these infants are treated for prematurity alone, having no specific problem other than their lack of development. In these cases the object is simply to provide the necessary physiological support and an environment conducive to growth, and remain vigilant for complications. In other cases, infants of this gestation can have serious respiratory or cardiovascular problems, or problems with infection, or several other types of pathology. Less than 80% of infants born at 26 weeks gestation survive till discharge from the NICU (Cooper et al., 1998).

Measurements of the heart rate, blood pressure, temperature and so forth give an indication of the general condition of a baby. In particular, they provide evidence about whether different physiological systems are functioning correctly, e.g. that oxygen is being absorbed into the blood, that the blood is then being perfused around the body correctly, and that the temperature is being regulated appropriately. Much of the time babies can be expected to be in a ‘normal’ state, where a degree of homeostasis is maintained and measurements are stable. In specific situations, characteristic patterns can appear which indicate particular conditions or pathologies. Some patterns are common and can be easily recognised, whereas at other times there might be periods of unusual physiological variation to which it is difficult to attribute a cause.

Alarms in the NICU from the machines providing these measurements are based on limits; e.g. if the heart rate falls below a certain threshold the alarm will sound. As suggested above, false alarms are a problem with current monitoring technology. Lawless (1994) showed that more than 94% of the alarms in a pediatric intensive care unit were ‘false’, i.e. clinically insignificant and not resulting in any action being taken. Tsien and Fackler (1997) found 86% of alarms to be false in a different pediatric ICU, and high false alarms were also reported by Ahlborn et al. (2000) for a comparison involving different monitoring systems. There are also problems related to the different items of monitoring equipment being made by diffeent manufacturers. Chambrin (2001) points out that one problem with alarms in any ICU is that the manufacturer of each piece of equipment assumes that their alarm is the most important. This results in a frequent cacophony of competing alarms from different machines in the NICU, which is disconcerting for parents, unsettling for the babies themselves, and most concerningly causes clinical staff to become desensitised to the presence of the alarms.

One reason for the frequency of false alarms is the fact that the measurements are often corrupted by artifact. Perhaps the most obvious cause of artifactual corruption of the data is when a probe becomes disconnected. For example, if heart rate measurements are being supplied by electrocardiogram (ECG) leads, and one of the leads becomes disconnected, then the measurements will fall to zero. An alarm rings as though the heart has stopped beating. If we are also measuring other channels, and these other measurements appear to be normal, then from the observations we can infer that it is highly unlikely that the change in measurements indicate a problem with the heart. One problem with current technology is therefore that there is no system which is able to take a combination of different measurements and decide whether to produce alarms from this higher level perspective. Besides the ‘dropping out’ of measurements due to probe disconnection, other artifactual patterns are caused during probe recalibrations, handling of the baby or other clinical operations.

It was also suggested above that there can be a problem in having too much monitoring data for staff to be able to pay attention to, and that it can be difficult for staff to interpret the significance of some sequences of measurements. This thesis is particularly concerned with the use of cotside computers to record and display vital signs measurements, a situation in which these issues are potentially more problematic. A general monitor is used alone in most NICUs, which provides only a snapshot of the data covering the last few seconds or minutes. Computerised monitoring provides the ability to see longer term trends in vital signs than would be possible with the general monitor alone. For example, a gradual rise in CO_2 saturation and fall in O_2 may indicate a respiratory problem (see the pneumothorax pattern in Section 2.4.3), despite the values being within a normal range. This type of trend would not be discernable on a standard general monitor displaying only the last few seconds of measurements.

Cunningham et al. (1998) demonstrated that the presence of computerised monitors does not necessarily lead to better decision making in the intensive care unit. Junior staff are particularly likely to be overwhelmed by the quantity of data available. The high frequency of clinical rotations means that junior staff in the NICU may have little practical experience of vital signs monitoring in such a situation, and even common artifactual patterns may be unfamiliar to some. Alberdi et al. (2002) showed the significant extent to which junior clinical staff in a NICU were significantly less able to recognise relevant patterns in physiological monitoring data than were their senior counterparts. Other previous studies have confirmed the desirability of additional processing of the data in order to provide a representation which is more helpful to clinicians (Logie et al., 1998).

To put the state of the technology in context, neonatal intensive care is a relatively young branch of medicine. The development of the speciality began only in the 1960s[1], notably with early work on the use of mechanical ventilation. The field became better established by the 1980s, with more sophisticated monitoring techniques and mortality for premature babies much reduced by new treatments for previously intractable problems—such as surfactant for the avoidance of respiratory distress syndrome. The first computers appeared in neonatal intensive care units for trend monitoring in the early 1990s (Cunningham et al., 1992), having first been used in adult intensive care a few years previously (Sivek et al., 1987). This type of computerised monitoring has provided the opportunity to collect and analyse large quantites of physiological data, of which clinicians are still beginning to take advantage (McIntosh, 2002).

Commercial development of systems which use large quantities of data is also still beginning, notably including the BioSigns system developed by Oxford BioSignals[2]. This system provides an index of a patient's stability based on analysis of several vital signs channels, and is aimed at providing an early warning of any kind of physiological deterioration. The functionality that the BioSignals system provides is related to the work on novelty detection in this thesis (see Chapter 6). In general however we address a somewhat different problem, looking at multiple specific patterns which may have particular known interpretations, as discussed in the next section.

1.2 Condition monitoring

Returning to the issue of uncertainty surrounding the condition of a premature infant, it is clear that the qualities of interest (e.g. whether a particular aspect of the physiology is healthy, or even whether the monitoring equipment is functioning correctly) are usually not directly observable. Different conditions are associated with different patterns of vital signs measurements, but to a clinical observer there may be many competing hypotheses about the cause of a particular pattern. Based on what is already known about the baby, each hypothesis has a prior probability. By collecting observations, or

[1] There is an interesting history of the subject in Christie and Tansey (2001).

[2] `http://www.obsmedical.com`

evidence, it can then be possible to obtain confirmation or disconfirmation given the assumptions made in each hypothesis.

Bayesian methods provide a calculus of uncertainty for determining a rational degree of belief in competing hypotheses in the light of evidence. Assumptions in the hypotheses are formalised in terms of models, which take into account properties of the overall system (in this case, the baby and the monitoring equipment) which produces the observations. An important assumption made in this thesis is that the physiological measurements we receive over time can be viewed as *switching* between different modes of operation. The data can therefore be modelled with an explanatory 'switching' variable which indexes these different modes, where these modes include different states of health and different processes which causes artifactual corruption. A class of generative probabilistic models can be formulated which switch between different regimes, to reflect the fact that regimes change in the data. The temporal evolution of the data in this application is significant, so these models must furthermore switch between different *dynamic* regimes. There are many different choices for how these dynamics might be represented, e.g. using combinations of discrete or continuous variables of different dimensions, described in more detail in Chapter 4.

Given measurements from a dynamical system, the task of trying to infer which of a set of given regimes is being followed (i.e. inferring the value of the switch variable), is sometimes referred to as *condition monitoring*. This is used in other applications, for example to evaluate the state of complex industrial processes from large networks of sensors (Morales-Menedez et al., 2002). In these cases, the measurements exhibit a number of regimes, some corresponding to normal operation and some corresponding to certain failure modes. In the same way, the sequences of measurements produced by a baby in intensive care change between different dynamical regimes, which correspond to different states of health or the operation of the monitoring equipment. Making inferences about the underlying causes in this case is also clearly of value. When identifying patterns in such physiological data, there are two important cases for condition monitoring: *known* regimes, in which a common occurrence (such as the disconnection of a probe) produces a stereotyped pattern which an expert can identify, and *unknown* regimes, in which changes occur which seem to be significant but have unclear cause. This distinction will be elaborated on with examples in Chapter 2.

The advantage of condition monitoring is essentially that it makes it possible to provide a better description of what is happening over time in high level terms. It has already been described how clinical staff reason at a high level, using data at a low

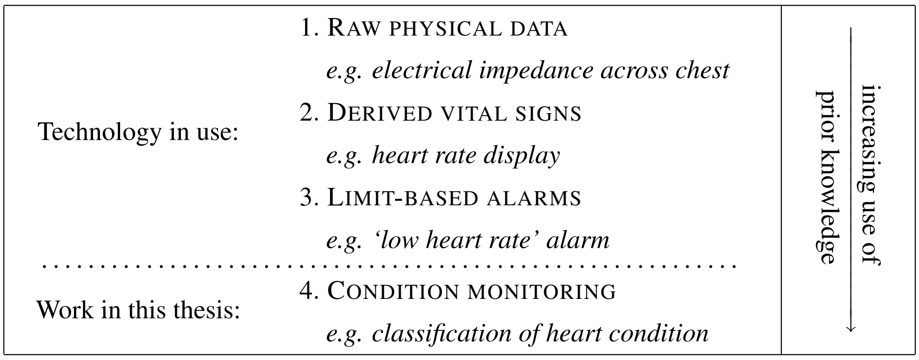

Technology in use:
1. RAW PHYSICAL DATA
e.g. electrical impedance across chest
2. DERIVED VITAL SIGNS
e.g. heart rate display
3. LIMIT-BASED ALARMS
e.g. 'low heart rate' alarm
Work in this thesis:
4. CONDITION MONITORING
e.g. classification of heart condition
increasing use of prior knowledge

Figure 1.1: Increasingly sophisticated approaches to monitoring, incorporating prior knowledge about what the measurements represent. Measurements relating to the heart are used here as examples, but many other sources of data are normally available.

level. Different types of existing monitoring equipment in the NICU can be arranged into a hierarchy in which prior knowledge about what the measurements represent is used to different extents to provide information at different levels of abstraction. Figure 1.1 shows this hierarchy. At the lowest level there are direct representations of raw physical signals, such as the electrocardiogram which shows electrical impedance across the chest. Based on the knowledge that this signal is affected by the heart, and that it is therefore periodic with particular characteristics, the heart rate can be calculated. At the next level, further knowledge can be incorporated about which levels are considered acceptable and which might indicate a problem, giving rise to alarms which are based on limits. Condition monitoring extends this principle to use more knowledge about particular patterns in the data to try and classify what is happening at a higher level.

Taking the simple probe dropout example: at level (1), displaying an electrocardiogram, the readings become excessively noisy and then cease as the probes are disconnected. At level (2), the heart rate drops to zero. At level (3), an alarm is sounded to show that the heart rate is outside the normal range, but without conveying what the cause might be. The idea of condition monitoring at level (4) is to offer more specific and useful information: 'the ECG leads are disconnected'. The same principle applies for more complex patterns, and for patterns involving multiple measurement channels, as introduced in Chapter 2.

A successful condition monitoring system can therefore address the problems raised in the last section as follows:

- False alarms can be reduced, as it is possible to distinguish between artifacts and genuine physiological variation (see e.g. results in Figure 7.4, p114).
- Artifacts can be identified, and displayed accordingly—not as though the data was physiological. More sophisticated models can attempt to estimate the true values of the physiology (see e.g. results in Figure 7.7, p117).
- The problem is also addressed that existing monitoring systems can have false negatives, i.e. that no alarm is sounded when there is an ominous sequence of measurements, for example a sharp downwards trend in blood pressure where the levels are still within the 'normal' range. The formulation of more sophisticated models can help to identify these situations (see e.g. results in Figure 7.13, p125).
- Patterns can be classified at a high level, which provides clinicians with more intuitive representations of the data, and makes interpretation of the vital signs data easier for inexperienced staff.

Notice also that while manufacturers invest a lot of effort in applying signal processing techniques to produce measurements which are as reliable and artifact-free as possible, these systems normally process only a single channel at once. The combination of measurement channels provides a rich source of information, better than the individual channels taken in isolation. A system or model which is able to 'see' different aspects of a baby's physiology has a better chance of providing useful diagnostic information than the current systems which deal with measurements independently.

1.3 Overview of the rest of the thesis

Chapter 2 reviews physiological monitoring of infants in intensive care, and describes the sources of data used in this work. Common monitoring patterns are described, which correspond to different known artifactual and physiological conditions. Examples of other, more rare types of physiological variation are also shown.

Chapter 3 reviews relevant previous work in the field of intensive care unit monitoring and discusses the connections to work in this thesis.

Chapter 4 discusses different probabilistic dynamical models, and the ways in which they are appropriate for the task of monitoring the condition of an infant using physiological data. One model, the Factorial Switching Kalman Filter (FSKF) is developed in particular. The non-factorised version of this model has been used in the past for other applications such as detecting faults in mobile robots (de Freitas et al., 2004), monitoring industrial processes (Morales-Menedez et al., 2002), modelling financial series (Azzouzi and Nabney, 1999), modelling creatinine levels in patients with kidney transplants (Smith and West, 1983), musical transcription (Cemgil et al., 2006), speech recognition (Droppo and Acero, 2004) and synthesising human motion (Pavlović et al., 2000). Application-specific techniques for speeding up these models and handling missing values are proposed.

Chapter 5 shows the ways in which the parameters of the dynamical models in the previous chapter can be estimated for known types of variation in sequences of vital signs measurements. It is shown that by choosing particular parameterisations, it is possible to incorporate a significant amount of domain knowledge in order to make useful inferences. Techniques and software for verifying that the models are well fitted are demonstrated.

Chapter 6 proposes a model for the types of variation in which there is a significant departure from 'normal' dynamics which does not correspond to any known pattern. Parameter estimation is demonstrated by deriving the expectation-maximisation algorithm are this model.

Experimental results are given in **Chapter 7**, which applies the techniques developed in previous chapters to monitoring data from the Edinburgh Royal Infirmary NICU. These show that the models proposed here are able to make accurate classifications about the underlying causes in the cases of known patterns, and to draw attention to other periods of variation that experienced clinicians would say was significant.

Appendix A provides details of software written to perform the experiments. **Appendix B** gives details of two inference routines for switching linear dynamical models which were used to obtain filtering densities in the experiments. **Appendix C** gives derivations of expectation-maximisation algorithms for different forms of linear dynamical models. Some additional novel material is supplied in **Appendix D** concerning the application of autoregressive models to data with different sampling frequencies.

1.4 Clinical and commercial collaboration

The work in this thesis was undertaken as a collaboration between the Neonatology department in the University of Edinburgh and the School of Informatics. The author was regularly present in the neonatal intensive care unit at Edinburgh Royal Infirmary throughout the duration of this work, being responsible for maintaining some aspects of the monitoring systems there. This presence in the NICU has allowed first-hand observation of particular patterns, and familiarity with clinical processes; expert interpretation of data has always been available; and informal input from clinicians about the research has also always been available, as referred to in the acknowledgments.

The Badger Patient Data Management System which operates in this NICU has been developed by the company Clevermed[3]. There has also been close collaboration with this company in the early stages of the research with regards to extracting and storing data, as well as jointly producing interface software for particular items of monitoring equipment.

[3] http://www.clevermed.com.

Chapter 2

Physiological monitoring

Vital signs, literally ‘signs of life’, are used in neonatal intensive care to help diagnose the condition of a baby in a critical state. The state of health of a baby cannot be observed directly, but different states of health are associated with particular patterns of measurements. Given observations of the heart rate, body temperature and so on, inferences can therefore be made about the operation of the underlying physiological systems—e.g. whether they are functioning normally, or whether there seems to be evidence of some pathology.

This inference task is not straightforward, as the complex nature of a baby’s physiology leads to observed patterns which are also complex. A certain set of readings may have been caused by a number of known conditions, and conversely some babies have rare combinations of illnesses and may exhibit patterns of vital signs which are not quite like any that have been seen previously. The difficulties are clear when considering all the things that might happen to a human body, and the fact that we are typically trying to gain information about these possible conditions from a set of around only ten measurements taken once per second. To complicate the problem further, the vital signs which are observed depend not just on the state of a baby’s physiology but also on the operation of the monitoring equipment. There is observation noise due to innacuracies in the probes, and some operations can cause the measurements to become corrupted with artifact. For example, handling a baby may cause a probe to become disconnected (see Section 2.3.1), or a particular probe might need to be recalibrated (see Section 2.3.3).

This chapter describes how data is obtained, and gives examples of common patterns. Section 2.1 opens with a description of the items of equipment used in neonatal intensive care which are relevant to this thesis, in particular the probes used to obtain vital signs measurements and the data channels which each probe provides. The rest of the chapter goes on to provide examples of this data under different circumstances. Much of the time, premature babies in intensive care are in a stable condition, and in Section 2.2 'normal' vital signs observations are described. Section 2.3 goes on to show how artifactual patterns can corrupt the data under particular circumstances, and Section 2.4 gives examples of the manifestations of different physiological occurrences, both common and rare. Section 2.5 summarises.

2.1 Monitoring and other equipment

Equipment is used in the NICU to maintain a stable environment for the infant and to provide treatment, as well as to obtain physiological measurements. A baby with a gestational age of 26 to 28 weeks is normally enclosed in an incubator which controls the temperature and humidity. A general monitor operates at the cotside, which displays short term vital signs information and sounds an alarm when any measurement goes outside a predefined range. The display on this monitor includes the second-by-second data with which this thesis is concerned, such as heart rate, as well as higher frequency waveform data, such as the ECG or respiratory waves. A mechanical ventilator is used in the cases where the baby is not developed enough to breathe unaided. Numerous other items of equipment may be used, for example concerning phototherapy or the controlled administration of drugs, with which this thesis will not be concerned.

2.1.1 Probes and observed channels

This section describes the probes used to take measurements and the observation channels which are obtained from each. When monitoring in intensive care, it is helpful to distinguish between physiological sub-systems which have a certain degree of independency. For example, the cardiovascular system circulates blood around the body, the respiratory system regulates the quantities of different gases in the blood, absorbing oxygen and dispelling carbon dioxide, and the thermoregulatory system is responsible for maintaining homeostasis in body temperature. Each item of equipment we consider here monitors an aspect of the functioning of one of these systems.

Cardiovascular function is monitored by an electrocardiogram (ECG) and a blood pressure sensor. Here the ECG measurements are normally obtained by placing three sensors on the torso which measure electrical impedance corresponding to the beating of the heart, producing a heart rate reading measured in beats per minute (bpm).The

Siemens SC7000 patient monitor used in the Edinburgh Royal Infirmary has a base sampling rate of 225Hz, and calculates a heart rate based on an average of all valid beats in the preceeding 15 seconds. According to the published specifications for this device (Dräger Medical, 2007) there is no special handling for missed or extra heartbeats. A pressure transducer connected to the infant's arterial line monitors the blood pressure, and can also give an indication of the heart rate. This piece of equipment collects the systolic blood pressure (the pressure while the heart is contracting) and diastolic blood pressure (while the heart is at rest). 'Mean blood pressure' is often used by clinicians and is the average of the systolic and diastolic. All blood pressures are measured in mmHg. Samples of the blood pressure in the artery are taken internally by the SC7000 monitor at a rate of up to 32Hz. When ECG leads are not being used, this waveform is used to calculate the heart rate; here there is a smooth periodic evolution in the signal, with a peak each time the heart muscle contracts. As with the ECG-based heart rate estimate, the specifications (Dräger Medical, 2007) do not indicate any special treatment for missed or ectopic beats.

Respiratory function is measured here in two ways—with a transcutaneous probe and with a pulse oximeter. A transcutaneous probe, sited on the chest, measures the partial pressure of O_2 and CO_2 in the blood[1]. This device measures in particular the concentration of these gases in capillary blood underneath the skin. It heats the skin to improve perfusion (blood is drawn into the capillaries from the nearest artery), and measures how much of each gas rises through the skin electrochemically. Partial pressures are measured in kPa.

A pulse oximeter, attached to the foot, measures the saturation of oxygen in arterial blood—a related but different quantity to transcutaneous O_2[2]. Rather than measuring in absolute terms the concentration of each gas in the blood, the principle here is to measure what proportion of the blood's capacity to carry oxygen is being utilised. This is done by shining an infrared light through the foot. The oxygen saturation can be estimated by analysing the light emerging through the tissue, taking into account the absorbtion spectrum of oxygenated hemoglobin. These measurements can sometimes have limited accuracy, however (Stoddart et al., 1997). The oxygen saturation is measured in percent.

[1] Various gases are dissolved in the bloodstream, and the partial pressure is used to quantify the amount of each. It is the amount of pressure that a particular gas would exert on a container if it was present without the other gases.

[2] For a particular baby there is a nonlinear, monotonic relationship between arterial oxygen saturation and transcutaneous oxygen which is known as the 'oxygen dissociation curve'. See Poets (2003) for further comparison of these two quantities.

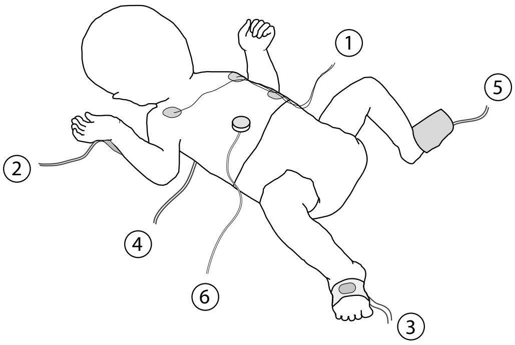


Figure 2.1: Probes used to collect vital signs data from an infant in intensive care. 1) Three-lead ECG, 2) arterial line (connected to blood pressure transducer), 3) pulse oximeter, 4) core temperature probe (underneath shoulder blades), 5) peripheral temperature probe, 6) transcutaneous probe.

The core body temperature and peripheral temperature are measured by two probes, one of which is placed under the baby's back (or under the chest if the baby is prone) and the other attached to a foot.

In addition, environmental measurements (temperature and humidity) are collected directly from the incubator. These measurement channels are summarised in Table 2.1, and the items of monitoring equipment are depicted in Figure 2.1. Normal ranges for each of the measurement channels, for an infant representative of those observed in this thesis, are summarised in Table 2.2. All the measurements described here are taken at one second intervals.

2.2 Normal variation

Having described the sources of monitoring data, we now go on to describe its characteristics under different conditions. The most common condition is physiological stability. Infants in intensive care can be treated for prematurity without having any specific pathology, and they may also have a chronic condition, such as respiratory distress, but remain physiologically stable. We can therefore make the distinction between a 'stable' condition, where long-term problems may or may not be present, and other conditions during which something is happening; an acute pathology or some kind of physiological change. Where the infant is stable and the measurements are free from monitoring artifact (see Section 2.3 for a summary of common artifacts), we refer to the measurements as having 'normal' variation.

System	Measurement	Abbreviation	Units
Cardiovascular	Heart rate (ECG, BP)	HR	bpm
	Systolic blood pressure (BP	Sys BP	kPa
	Diastolic blood pressure (BP)	Dia BP	kPa
Respiratory	Partial pressure of O_2 (TCP)	$TcPO_2$	mmHg
	Partial pressure of CO_2 (TCP)	$TcPCO_2$	mmHg
	O_2 pulse saturation (PO)	SpO_2	%
Thermoregulatory	Core temperature (CT)	Core temp.	°C
	Peripheral temperature (PT)	Periph. temp.	°C
(Environmental)	Ambient humidity (I)	Incu. humidity	%
	Ambient temperature (I)	Incu. temp	°C

Table 2.1: Measurement channels available at the cotside and the abbreviations used to refer to them in the rest of this thesis. (ECG) denotes a measurement derived from the electrocardiogram, (BP) the arterial blood pressure sensor, (TCP) the transcutaneous probe, (PO) the pulse oximeter, (CT) and (PT) the core and peripheral temperature sensors respectively, and (I) the incubator.

Measurement	Normal range
HR	119-175bpm
Sys BP	27-55mmHg
Dia BP	23-43mmHg
$TcPO_2$	5.5-13.5kPa
$TcPCO_2$	2.6-7kPa
SpO_2	86-100%
Core temp.	35.7-38.5°C
Periph. temp.	34.6-38.0°C

Table 2.2: Normal ranges for infants of 26-28 weeks gestation on the first day of life, for each measurement channel. Low and high values are the 3rd and 97th centiles calculated from a large corpus, apart from the values for SpO_2 which is the range quoted in McIntosh et al. (2003).

Because infants with a low gestational age are usually asleep and motionless, there tends to be low variability in their vital signs when in a stable condition. There may be healthy variation in heart rate (and, to an extent, blood pressure) in a stable infant, but other vital signs would be expected to exhibit more homeostasis and stay roughly constant. An example of vital signs measurements during a normal period is shown in Figure 2.2.

2.3 Artifactual events

The measurements received by the cotside PC are affected not only by the physiological condition of the infant, but also by the state of the recording equipment. Measurements are frequently corrupted with artifactual patterns, caused by particular clinical procedures or changes to probes or equipment.

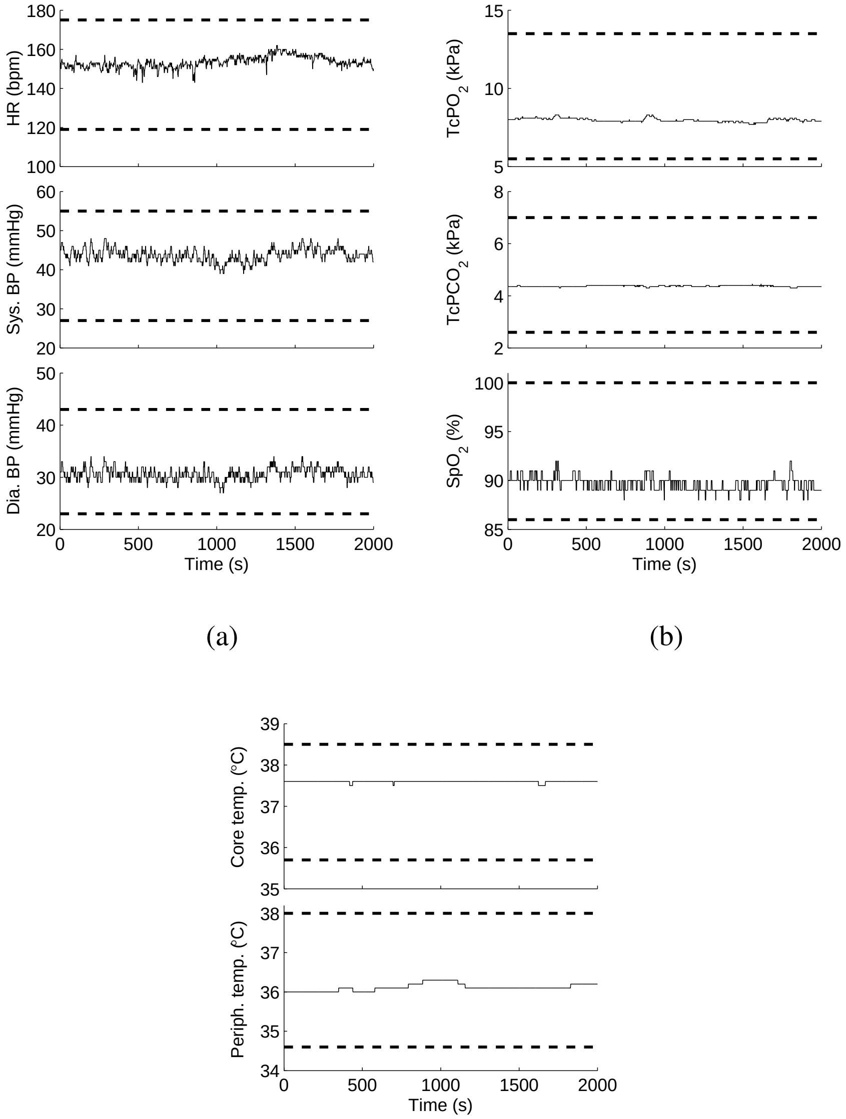
2.3.1 Probe disconnection

The most obvious cause of artifactual readings is a probe disconnection. When probes become physically disconnected from the baby, readings will tend to fall to zero. Temperature probes are an exception, as they decay to the ambient temperature in the incubator when disconnected.

The core temperature probe is placed underneath the baby and detachment is common when the baby is moved, e.g. for an exam or a change of linen. Figure 2.3 shows examples of temperature measurements decaying to the ambient level after the probe has been disconnected.

2.3.2 Blood sampling

A blood sample might be taken every few hours from each baby. This involves diverting blood from the arterial line containing the pressure sensor. The arterial line normally has a pump which passes a small amount of saline solution through the line to keep it clean, and while the sample is taking place, the pump is sealed off and is left acting against the pressure transducer. This causes an artifactual ramp in the blood pressure measurements, shown in Figure 2.4. If the heart rate is being calculated from the pressure sensor rather than the ECG, then it will read as zero for the duration of the sampling procedure.



(a)

(b)

(c)

Figure 2.2: Examples of normal variation, from a baby of 28 weeks gestation in a stable condition. (a) Cardiovascular channels, (b) respiratory channels, (c) thermoregulatory channels. The dashed lines show the 3rd and 97th centiles for each channel for babies of the same gestation, except SpO$_2$ for which the dashed lines show the clinically recommended normal range.

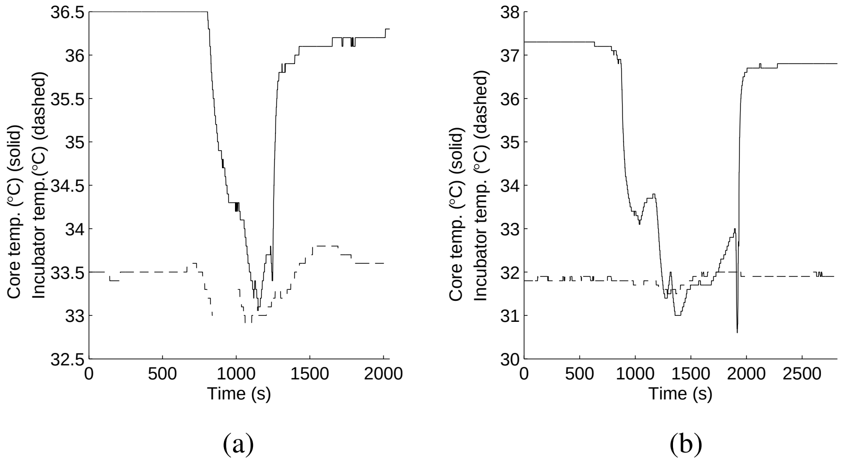


Figure 2.3: Examples of temperature probe disconnection. Note that the temperature probe does not always become completely detached – for example, in panel (b) the probe appears to make temporary partial contact with the baby at around $t = 1000$. In panel (a) the ambient temperature readings from the incubator are themselves interrupted between 800 and 1000 seconds.

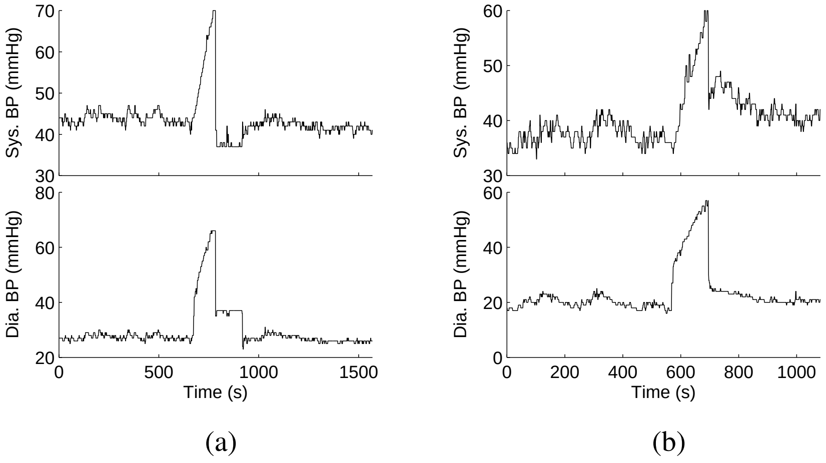


Figure 2.4: Examples of blood sample artifacts. While the sample is being taken the pressure sensor is disconnected from the artery, so that diastolic and systolic pressure measurements are similar. A saline pump acts against the sensor throughout the duration of the sample, causing the artifactual ramp in measurements. Panel (a) represents a special case in which the pump has been stopped at around $t = 800$, in order to prevent damage to the baby's artery.

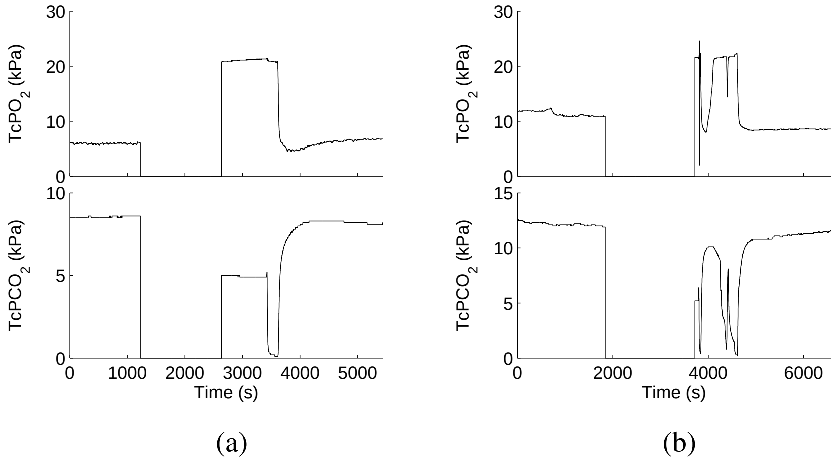


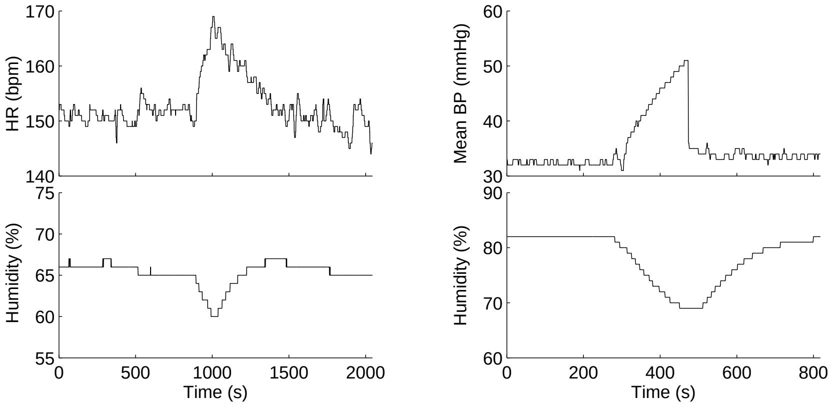
Figure 2.5: Examples of transcutaneous probe recalibrations. Note the different stages of the recalibration, where readings first of all drop to zero, then calibration solutions are applied and the probes reattached. The stages are described in more detail in Section 5.8.4.

2.3.3 Transcutaneous probe recalibration

Transcutaneous probes are prone to drift, and need to be recalibrated every few hours. When the probes have been in continuous operation long enough for potential drift to have become a problem, the readings automatically cease until the probe has been recalibrated. To do this, the probe is subjected first to room air and then a calibration solution. It is then re-attached, causing the measurements to decay from calibration levels to something more accurately representative of the true physiological levels. This artifactual pattern therefore has multiple consecutive stages, shown in Figure 2.5.

2.3.4 Handling and incubator entry

Handling of a baby can affect its stability, and we might expect to see increased variation in vital signs at these times. The pulse oximeter is sensitive to movement, so we also expect to see higher (artifactual) SpO$_2$ variability during handling. Where environmental measurements can be made from the incubator, handling of the baby has more direct visible effect. Each incubator has a closely controlled temperature and humidity, both of which change when a member of clinical staff reaches in through the side portals. Figure 2.6 shows the humidity falling to an ambient level when access is gained to the incubator.



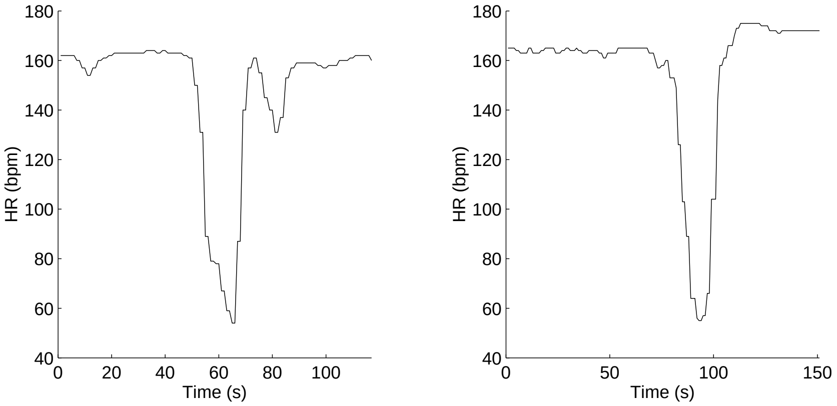
(a) (b)

Figure 2.6: Examples of falling humidity while the incubator portals are open. In panel (a), handling of the baby has caused an increase in heart rate. In panel (b) the incubator has been opened in order to obtain a blood sample. The ramp in mean blood pressure is an example of the blood sampling artifact described in section 2.3.2.

Unlike the artifactual patterns described in the previous sections, the variation in physiological and environmental measurements caused in this situation can be genuine. We class this as an artifact in the sense that an external agent produces changes in the measurements which would not have occurred without the intervention. An important element of our monitoring problem is to identify iatrogenic data—that is, variations in a baby's physiology which have been caused by a clinical intervention rather than occurring spontaneously.

2.4 Physiological events

This section gives examples of specific physiological conditions and the patterns which are associated with them. Some of these patterns, such as bradycardia or desaturation are stereotyped and normally have a similar manifestation. Patterns such as the pneumothorax (section 2.4.3) are less specific, while there are a large class of other patterns of variability which it is difficult to identify a cause for, or which are unlikely to repeat in the same way twice, such as those examples in Section 2.4.4.

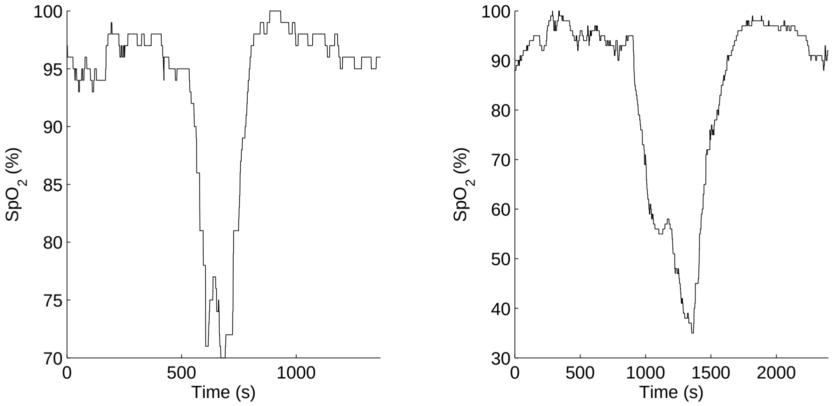


(a) (b)

Figure 2.7: Episodes of bradycardia. These can occur in isolation as in the examples above, or in quick succession in a bradycardia 'storm'. It is rare for the heart rate to make a transient drop almost to zero for more than a few seconds.

2.4.1 Bradycardia

Bradycardia is a slowing of the heart rate, and brief episodes are common for premature infants. It can have many causes, some benign and some serious. They can broadly be classified either as abnormalities in the generation of pacemaker impulses (atrial bradycardia), or a problem with conducting these impulses to the heart muscles (atrioventricular block). The first type of bradycardia can be triggered by many things, including a disturbance at the back of the throat, which can happen routinely—for example with the insertion of a nasogastric feeding tube or preparation for ventilation. Miller et al. (2000) has a further description of the causes and mechanisms involved in neonatal bradycardia. Episodes of bradycardia for different infants are shown in Figure 2.7. Note that in general the heart rate measurements can only be used to establish whether a bradycardia has occurred or not. Diagnosis of the particular cause of the bradycardia (and in particular whether there is a routine explanation or if it indicates a serious problem in cardiovascular function) generally requires expert analysis of the underlying ECG waveform. The normal prevalence of bradycardia for a premature baby of 26-28 weeks gestation is indicated by the data collected in this research, shown in Table 7.1.



(a) (b)

Figure 2.8: Episodes of oxygen desaturation, each lasting a few minutes.

2.4.2 Desaturation

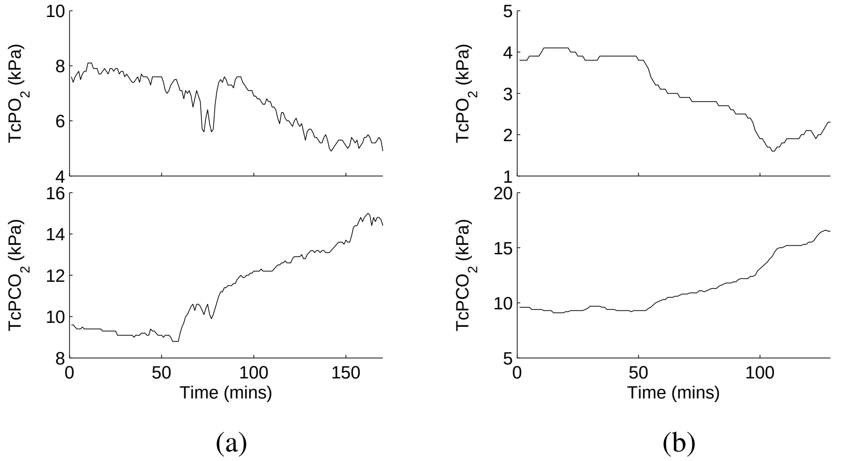
A sudden drop in oxygen saturation is common in neonates. Examples are shown in Figure 2.8. Often, periods of desaturation are associated with apnoea (cessation of breathing). It can also be caused by prolonged bradycardia, as a slower heart rate results in decreased oxygen delivery. See Martin and Faranoff (1998) for a brief discussion of the mechanisms commonly involved. These desaturations are often transient occurrences, as shown in the Figure, where stability is restored either spontaneously or by increasing the concentration of oxygen supplied to the baby.

2.4.3 Pneumothorax

Pneumothorax is a rare, life threatening condition, involving a buildup of air outside the lung. This can cause the lung to collapse, affecting oxygenation (SpO_2 and $TcPO_2$ may fall) and ventilation ($TcPCO_2$ may rise). The patterns in vital signs associated with pneumothorax have been studied by McIntosh et al. (2000). Two examples are shown in Figure 2.9. The prevalence of babies who have at least one pneumothorax is around 1% of those receiving mechanical ventilation (McIntosh et al., 2000).

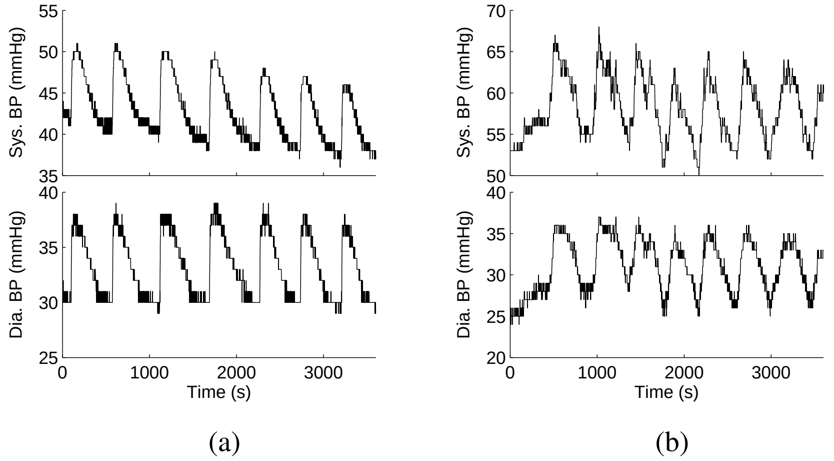
2.4.4 Other types of physiological variation

In addition to the commonly occurring dynamics described above, many other patterns may be observed in certain situations. A few are given here to provide context. Blood pressure waves are one example of a rare (but clear) pattern, where there is periodic



(a) (b)

Figure 2.9: Two examples of the effects of pneumothoraces. Times on these charts are in minutes, rather than seconds. The onset of pneumothorax for case (a) is at 60 minutes, and for case (b) 50 minutes. Note that these examples are relatively unusual in that they are free from artifact. Often the transcutaneous probes supplying these measurements are recalibrated by nursing staff (see Section 2.3.3) in response to changing values.



(a) (b)

Figure 2.10: Two examples of blood pressure waves.

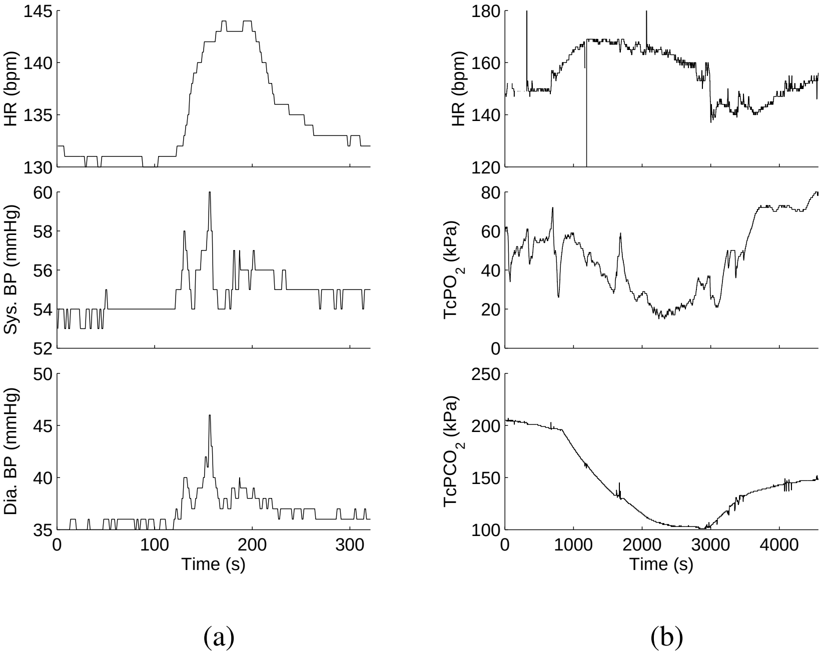


Figure 2.11: Two examples of atypical dynamics. In (a) the baby reacts to the flash of a camera at 120 and 150 seconds, with peaks of blood pressure, and elevated heart rate. (b) An unusual case of physiological variation associated with respiratory distress. The causes are known in both these cases because clinical observers were present, but would be difficult to infer retrospectively.

structure in the blood pressure measurements thought to be caused by an endocrine mechanism. Examples are shown in Figure 2.10, and more information about the phenomenon can be found in Cunningham et al. (1993).

Much of the variation in vital signs measurements does not conform to a particular stereotype, and is peculiar to the baby or the situation, or is a combination of factors. There are also rare conditions about which it has not been possible to gather enough data to establish whether there is an associated monitoring pattern, such as the more unusual monitoring periods in Figure 2.11, one caused by respiratory trouble, and another caused by the baby's reaction to flashes of a camera.

Other types of variation, with more or less stereotyped manifestations, could be caused for example by the administration of drugs, feeding or the carrying out of other procedures, sepsis, or more serious occurrences such as an internal haemorrhage. The variety of potential causes and patterns is large enough that it is clear that modelling all of them explicitly would not be practical. In the data collected for this research, the prevalence of such abnormal variation not fitting any common pattern is indicated by the 'Abnormal' row in Table 7.1.

2.5 Summary

Having described the significance of an infant's vital signs measurements and the processes involved in obtaining them, this chapter has shown the ways in which the dynamics of of these measurements are affected by different underlying factors. These factors include different changes in physiological functioning, clinical procedures which are being carried out, and the operation of the monitoring equipment. Though we do not measure these things directly, there are certain common patterns which it is possible to recognise. However, the number of ways an infant's condition might change, and the non-specific nature of vital signs patterns caused by some of these conditions means that it is not practical to try and model everything explicitly.

Also note that the events described in this chapter can happen independently. A bradycardia could occur while a blood sample is being taken, for example. Each event affects only particular measurements, so in this case the bradycardia would affect the heart rate observations, and the blood sample would add artifactual values to the blood pressure observations.

Further details on physiological monitoring procedures in the NICU are available, for example, in Rennie and Roberton (2001).

Chapter 3

Previous work

Ths chapter reviews relevant previous work on monitoring in intensive care, applied to the types of data and patterns introduced in chapter 2. Work on monitoring for both neonatal and adult intensive care is considered, being similar from a technical perspective.

Previous work is divided here into three main categories. At the most specific level, there has been work on identification of particular individual patterns in monitoring data, for example to diagnose infection or particular respiratory problems. This work is reviewed in section 3.1. The remainder of the chapter considers previous work which has a broader scope, dealing with the fact that this type of data typically contains many different patterns. This work is comprised of two distinct approaches. The type of approach considered first (in section 3.2) is based on using domain knowledge to formulate high-level representations of particular patterns or situations, then to find suitable abstractions of the data in order to apply some matching rules. In this type of work, the goal is to describe what is happening, and sometimes to suggest what to do next; an *interpretation* is put on the data. By contrast, another body of work (discussed in section 3.3) is based on making inferences of a statistical nature from monitoring data using time series analysis techniques. The goal in this case to use the methodology of time series analysis to obtain informative *descriptions* of the data, which offer insight into the underlying processes. As discussed in the summary in section 3.4, these approaches are not mutually exclusive—the work in this thesis is motivated by the idea that it is possible to perform extensive knowledge engineering within a principled (probabilistic) time series analysis framework.

Recall the problems with monitoring described in the introduction, particularly that alarms are too frequent and not informative enough. Attempts to work with monitoring data tend to address either the problem of false alarms (particularly in the form of *artifact detection*), or to identify when an alarm would really be appropriate (detection of physiological patterns).

The work of Spengler (2003) is directly relevant to this thesis, in which a Gaussian mixture model was implemented to represent different artifact processes on a similar, publically available dataset (Hunter et al., 2003) from Edinburgh Royal Infirmary NICU. The Gaussian mixture model is conceptually related to the models proposed later as shown in Table 4.1. The parts of work in this thesis based on hidden Markov models essentially extends this to various dynamic cases. Another directly relevant aspect of Spengler's work was reasoning about the way that different phenomena 'overwrite' observations (related to work in section 5.10.1).

3.1 Monitoring of specific patterns

Some relevant previous work has aimed to identify individual, interesting patterns in monitoring data by exploiting knowledge about the pattern and its underlying physiological processes.

Griffin and Moorman (2001) worked towards classification of neonatal sepsis by looking at moments of deviation of the heart rate. They found that infection was associated with changed variability (specifically, lower skewness) of heart rate observations. Using respiratory information, an algorithm for the detection of pneumothorax (section 2.4.3) was developed by McIntosh et al. (2000), based on the observation that the condition caused increased CO_2 and decreased O2 in the blood as shown in Figure 2.9.

Detection of seizure by analysing electroencephalogram (EEG) signals has been an active area of research in both neonates and adults. It is of only peripheral relevance here though, as we do not deal with this type of data; an example of this work is Roessgen et al. (1998), who create a model of neuronal feedback using linear time-invariant filters driven by an impulse train. This model was used to give a probabilistic classification of seizure on novel data.

Other work on specific pathologies has used analysis of waveform data. For example, detection of sleep apnoea (absence of breathing) has been shown to be possible by looking at ECG data sampled up to several hundred times a second. Different detection algorithms based on analysis of heart rate are compared by Penzel et al. (2002), who found the most effective approaches to be based on features derived from frequency-domain representations of the waveform. Studies of heart rate variability (HRV) in general using high frequency data are abundant and reviewed in Malik (1998). These studies look at ECG signal characteristics as a function of specific physiological systems including certain aspects of the central nervous system, and as such have a more fine-grained approach than the work taken in this thesis. Here the focus is rather on monitoring the overall function of a baby in terms of the high-level physiological systems described in section 2.1.1.

3.2 Work based on abstraction of data

Several pieces of work have also been undertaken to find a framework in which to identify multiple patterns. One approach to this is characterised by taking quantitative, raw monitoring data and extracting a more qualitative representation in order find an interpretation or explanation for different types of variation. In general this type of work aims to represent the descriptions of patterns given by clinical experts and apply them to new data. Because these descriptions tend to be in more abstract or qualitiatve terms than the raw data (they might be in terms of approximate rates of change, or with respect to reference ranges, for example) the idea is to derive a representation of the data at the same level as the expert description and then apply matching heuristics.

Tsien (2000) used decision tree learning to detect artifact corruption in NICU data. The work fits in to the idea of abstracted data in that it was based on an expert's interpretation of data as 'artifactual' or 'significant', and produced high-level, human-interpretable rules for classification. As an application of a statistical technique, this is also related to the work in section 3.3. A large number of derived features were used, including the mean, slope and standard deviation over 3, 5 and 10 samples on each channel (samples were taken each second as in our data). The work resulted in simple rules to determine whether a data point is artifactual not, and whether an alarm should have been sounded or not. These rules treated the feature set at each time point independently; there were no dynamics in the system.

Cao et al. (1999) produced an NICU artifact detection system based on limits and slopes across multiple channels, which used contextual information to improve results. The system included a set of heuristics to look for sections of data which change faster than biologically possible, and to classify data points as true or corrupted depending on whether a point in the vicinity had been definitely categorised as corrupted. This system, the 'Artitector', also produced estimates for the actual values of data points which were classified as artifactual, using a weighted sum of the mean of the last correct data points and the corrupted values. Using a different method, the models in this thesis also produce estimates of 'true' physiology underlying artifactual data, see for example Figure 5.9. Note that in this work by Cao and colleagues, patterns were manually described by an expert rather than learnt.

More general frameworks for expressing expert knowledge and applying it to data have been suggested. Hunter and McIntosh (1999) proposed a system for matching descriptions of patterns given by clinical experts to sequences of monitoring data. This involved finding a segmentation of a sequence of data, e.g. by looking for changes in level, and then applying matching rules specifying a particular level or slope. As an example, the framework was applied to the detection of transcutaneous probe recalibrations (for which the pattern was described in section 2.3.3). A different approach to comparing the temporal evolution of signals with experts' descriptions of patterns was given by Haimowitz et al. (1995). In this work, 'trend templates' (representations of the expected trajectory of a signal) were used to characterise different patterns. In an intensive care monitoring context, the work looked at the evolution of respiratory signals during a clinical operation involving changes in ventilation. Different templates described the possible outcomes of this operation, so that the authors could attempt to automatically classify patterns by looking at the distance of an observed trend from the idealised description.

A more extreme example of the idea of abstracting data and the descriptions of clinical experts is given by Miksch et al. (1996). The goal in this work was not only to detect significant patterns in intensive care unit monitoring data, but also to suggest therapeutic actions. This was done by deriving abstracted representations of the data (e.g. '$TcPO_2$ is substantially below the normal range'), and recording similarly abstracted representations of rules representing best practices (e.g '$TcPO_2$ should reach normal range [after some operation] after 10-20 minutes; if it doesn't then [perform other operation]'). A rule-base was developed describing best practices in ventilation, which was clinically tested with real patients. In the sense that this work aimed to represent very high level knowledge operating on abstracted data it is related in spirit to medical expert systems work, reviewed by Lavrač et al. (2000).

3.3 Work based on statistical time series analysis

The field of time series analysis has a large body of techniques which can be used to work with patterns in this type of data. There are two main motivations to approach the problem from a statistical perspective. First, a range of statistical tools exist which provide a rich description of the data; for example, performing analysis in the frequency domain or in terms of parametric models can provide insights which would be unavailable when dealing only with levels, slopes, standard deviations and so on. Second, the techniques are well-studied and consistent. For example, a probabilistic model provides inferences which are clearly interpretable—some 'meaning' can be attributed to an inferred probability distribution. A useful review of some of the applications of statistical time series analysis to intensive care monitoring is given by Fried et al. (2003).

Rather than providing a high level interpretation, the applications of these methods tend to have been designed to highlight statistical changes in observations. For example, changes in the slope of a sequence of observations, or level changes, or transient spikes, might all be flagged up as being a change which warrants the attention of a clinician.

Relevant work is discussed in increasing order of the sophistication of time-series models used. 'Sophistication' in this context means how much of the signal's phenomenology is incorporated into the analysis. Previous relevant approaches might be divided into three main categories:

- 'simple', involving transforms of the data such as median filtering, or statistics such as the cumulative sum;
- 'black boxes', in which parametric models such as autoregressive processes are used to model data and make predictions;
- 'gray boxes', or 'low-level physical models', in which models such as state spaces (e.g. structural models) are used to give increased physical interpretability.

Candy (2005) uses these distinctions to describe a hierarchy of techniques in a general signal processing context, for which increasingly informative inferences are obtained at the expense of higher model complexity.

The cumulative sum technique for identifying changes in sequences of data is used for intensive care monitoring data by Charbonnier et al. (2004), who use it to find segmentations of monitoring data. In this technique, a straight line is fitted to the data, and the differences between new data points and the predictions under this linear model are analysed. The principle of operation is that for a certain stable linear progression, the cumulative sum of the residuals remains roughly constant: chance positive deviations should be cancelled out by chance negative deviations. If the cumulative sum deviates outside a certain limit, a segmentation occurs, and a new linear model is fitted. Kennedy (1995) uses another simple statistic to tell if the slope of measurements (in this case blood pressures) is changing significantly: the difference between the next observed measurement and the weighted moving average of previous measurements.

The use of parametric models to make inferences about monitoring data has been investigated by Imhoff et al. (1998) and Hoare and Beatty (2000). Imhoff and coworkers learnt low order autoregressive models (see section 4.1) from monitoring data, and used them to infer the periods in which deviations from the learnt model were occurring. At each time step, predictions were made about the next observation under the model. If the observed value fell outside a predefined range of the prediction then this was taken to be a sign of clinically significant change. Hoare and Beatty tried learning different combinations of Autoregressive Integrated Moving Average (ARIMA) models and Kalman filters (the Kalman filters used random walk dynamics) in a similar way, though this time measurements which were far away from predictions were classified as artifact.

Some work closer to that in this thesis was carried out by Smith and West (1983), to analyse sequences of creatinine measurements taken at eight hour intervals from adults having recently had kidney transplants. The work is based on the knowledge that under a particular transformation, these measurements should appear as a straight line if the transplant has gone well. In practice periods of stable function tend to alternate with episodes of rejection, so given a new reading there is motivation to try and assess the probability that a stable regime is being followed or if there is a new trend. The authors use a switching Kalman filter (described later in section 4.6), which they refer to as a ‘multiprocess Kalman filter’. A structural parameterisation (general details in Harvey (1990)) was used to model the evolution of the creatinine measurements, with components representing the level and rate of change of the measurements. The model had four switch settings: stable evolution, changes in level, changes in slope and transient outlying measurements.

Multiple Kalman filters were also used by Sittig and Factor (1990) to a similar end in the analysis of blood pressure and heart rate measurements in a model which the authors refer to as a ‘multistate Kalman filter’. Here four Kalman filters are run independently. The Kalman filters have the same types of dynamics as used by Smith and West—stable evolution, level changes, slope changes and transients—and predictions from each model were compared with observations to give four likelihood terms. The authors try to obtain a posterior probability from these likelihood terms, but do not take into account that there should be a joint distribution over the latent variables in all four Kalman filters. The results therefore do give an indication of slope changes and so on, but have limited interpretability as probabilities.

A common aspect of much of this work involving statistical methods is that little or no interpretation is put on the way in which the data varies. Deviations from some sort of known regime are taken to be cues to alert a clinician that ‘something’ is happening, or as evidence of artifact, or as the output of a preprocessing step. Also note that in general the methods described in this section are closely related to previous work on novelty detection, reviewed in section 6.4.

3.4 Summary

Previous work based on the analysis of particular patterns in intensive care monitoring has been reviewed, along with general approaches to drawing inferences from monitoring data. Some of the general approaches are based on implementing the experience of experts and some are based on taking a statistical perspective. Despite the latter bodies of work appearing to form two camps, the motivation for the work in this thesis is that it is possible to model monitoring data using the state of the art in statistical time series methods which incorporate expert knowledge as a priori information. Some of the previous work reviewed here has successfully incorporated elements of both approaches, e.g. Smith and West. This thesis also builds on different types of previous work by constructing models which can both identify patterns with known interpretations (chapter 5) and which can also represent unusual and unexplained changes in dynamics (chapter 6).

There is much previous work from other domains relevant to the methods used in this thesis, which are discussed in the appropriate chapters. In particular, references for probabilistic dynamical models are given at the beginning of chapter 4, and previous work in novelty detection is described in section 6.4.

Chapter 4

Dynamical models

This chapter reviews different dynamical models which can be applied to the problem of neonatal condition monitoring. Chapter 2 has shown the ways in which the dynamics of a baby's vital signs measurements change depending on different underlying factors, where these factors include the operation of the monitoring equipment, clinical procedures which are being undertaken and the state of health of the baby. The first goal is therefore to find a model that effectively represents the regimes which are known to be associated with particular factors, and the behaviour of the factors themselves. Having obtained a suitable model, it is then our goal to be able to infer the underlying factors giving rise to novel sequences of measurements.

This modelling task can be split into three parts. First, it is necessary to find a way of representing a single dynamic regime—that is, a model for the observations conditional on a particular clinical procedure taking place, a particular state of health, and so on. Second, the model needs to be extended to be able to switch between different regimes. This typically involves the addition of a discrete switch variable which controls the regime. Finally, the factorial nature of the system needs to be taken into account, so that the switch variable is represented as a cross product of a priori independent switch variables for each factor.

The Gaussian distribution and the autoregressive (AR) process are two fully observable models for single regimes. Adding a discrete hidden switch variable with a first-order Markovian dependency to these models yields the Hidden Markov Model (HMM) and the Autoregressive Hidden Markov Model (AR-HMM) respectively. The switch variable can then be factorised to give the Factorial HMM and the Factorial AR-HMM. The Kalman filter is a partially observable model for a single regime, having a set of continous latent variables which we refer to as the *state*. Adding a hidden switch

Single regime		Switching dynamics		Factorised
Gaussian	⟶	HMM (4.2)	⟶	Factorial HMM (4.3)
AR process (4.1)	⟶	AR-HMM (4.4)	⟶	Factorial AR-HMM (4.4)
Kalman filter (4.5)	⟶	SKF (4.6)	⟶	Factorial SKF (4.7)

Table 4.1: Relationships between different dynamical models introduced in this chapter. The first column shows a type of model characterising a single regime. The second column contains the model given by generalising to switching dynamics, and the final column shows the model given by a further generalisation in which the dynamical regime is affected by multiple independent factors. The number of the section in which each model appears in the chapter is given in brackets.

variable in the same manner yields the Switching Kalman Filter (SKF), and factorising the switch variable gives the Factorial Switching Kalman Filter (FSKF). Table 4.1 shows these relationships between the models, and the sections in which each model appears in the chapter.

Methods are outlined for setting parameters in each of these models, though in general we apply domain knowledge wherever possible, decribed in detail in the next chapter. For details on AR models, see Chatfield (1975) or Shumway and Stoffer (2000). A comprehensive introduction to the HMM is given by Rabiner and Juang (1989), and details specific to the Factorial HMM are discussed by Ghahramani and Jordan (1997). The AR-HMM is discussed by Woodland (1992). Various authors have discussed the SKF, see for example Shumway and Stoffer (1991); Murphy (1998); Ghahramani and Hinton (1998); Kim (1994). A useful unifying perspective on many of these models is given by Roweis and Ghahramani (1999).

Parts of this chapter are adapted from Williams, Quinn, and McIntosh (2006).

4.1 Autoregressive processes

An autoregressive (AR) model is a way of characterising a single, stationary dynamical regime, for which much of the methodology was set out by Box and Jenkins (1976). It is based on the idea that the value y_t of a series can be explained in terms of the preceeding values $\{y_{t-p}, \ldots, y_{t-1}\}$. It is therefore a regression based on the variable's own history rather than on any external variables. Where the current value of a series is dependent on the previous p values, we refer to the resulting series as an AR(p) process. The scalar AR(p) model is described by the equation

$$y_t - \mu = \sum_{k=1}^{p} \alpha_k (y_{t-k} - \mu) + \nu_t \ , \tag{4.1}$$

in which ν_t is white noise with variance σ_ν^2, and μ is the mean of the process. In the rest of this chapter, AR models will be assumed to have a mean of zero for convenience. The values of the α_k's and the variance σ_ν^2 therefore characterise the behaviour of the process $y_{1:t}$. For certain parameter settings it might be expected to decay, or oscillate or so on.

An AR model can be viewed in a generative sense in which the process $y_{1:t}$ is *driven* by the noise sequence $\nu_{1:t}$. For a sampling interval of Δ, a white noise sequence contains frequency components with uniform power in the range $-\frac{1}{2\Delta}$ to $\frac{1}{2\Delta}$. The effect of the autoregression is to filter out some of these frequency components. By analysing the properties of the filter it is possible to calculate the spectral characteristics of the resulting sequence.

A linear filter is completely described by its response to a single impulse, in this case easily calculated from the coefficients $\alpha_1, \ldots, \alpha_p$ and known as the *impulse response function*. Taking the Fourier transform of the impulse response function gives the *frequency response function*,

$$H(f) = \left(1 + \sum_{k=1}^{p} \alpha_k \exp -2\pi i f k\right)^{-1} \ , \tag{4.2}$$

which represents the result of driving the filter with a sinusoid of a particular frequency. See Shumway and Stoffer (2000) for a derivation, and Jenkins and Watts (1998) for more detailed background. The output spectrum of the filter when driven by white noise as in Equation 4.1 is given by

$$S(f) = |H(f)|^2 \sigma_\nu^2 \ . \tag{4.3}$$

A useful property of autoregressive processes is that they can approximate any power spectrum to within an arbitrary accuracy, given a high enough autoregression order p.

Maximum likelihood parameters

This model has no hidden variables, so that in general if the order p is known then the parameters with maximum likelihood can be found from training data (see section

5.1 for a discussion of the significance of maximum likelihood parameters). The Yule-Walker equations relate the maximum likelihood autoregressive coefficients $\alpha_1, \ldots, \alpha_p$ to the sample autocorrelation coefficients $\rho_1 \ldots \rho_p$ of a sequence of training data as follows:

$$\rho(k) = \alpha_1 \rho(k-1) + \cdots + \alpha_p \rho(k-p) \ , \tag{4.4}$$

for all $k > 0$. The autocorrelation coefficients are normalised so that $\rho_0 = 1$, and are given by

$$\rho_k = \frac{\sum_{t=1}^{N-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{N}(y_t - \bar{y})^2} \ , \tag{4.5}$$

where $\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t$. With these values, Equation 4.4 can be solved for $\{\alpha_k\}$ as a linear regression.

The noise variance σ_ν^2 can then be estimated by looking at the mean square discrepancy between the observed values y_t and the predicted values $\sum_{k=1}^{p} \alpha_k y_{t-k}$ in the training data.

Differenced AR processes

It is assumed above that the process is stationary, that is that its parameters do not change with time. Real data is often nonstationary, particularly in the types of physiological sequences arising in this application, though stationarity can sometimes be much improved by differencing the raw data. First order differencing cancels the nonstationarity caused by a linear drift, second order differencing cancels a quadratic trend and so on. Applying an AR model to differenced data gives the following (shown here as zero-mean for convenience):

$$y_t - y_{t-1} = \sum_{k=1}^{p} \beta_k (y_{t-k} - y_{t-k-1}) + \nu_t \ , \tag{4.6}$$

and it is straightforward to apply this model to the original, undifferenced data by converting the coefficients as follows:

$$y_t = \sum_{k=1}^{p+1} \alpha_k y_{t-k} + \nu_t \ , \tag{4.7}$$

where

$$\begin{aligned} \alpha_1 &= 1 + \beta_1 \\ \alpha_{k,1<k\leq p} &= \beta_k - \beta_{k-1} \\ \alpha_{p+1} &= -\beta_p \ . \end{aligned}$$

Not all sets of coefficients give rise to a stationary process. It is useful to analyse an AR process in terms of its *auxiliary equation*:

$$a^p - \alpha_1 a^{p-1} - \cdots - \alpha_p = 0 \,, \tag{4.8}$$

having roots $\{\pi_i\}$. The condition $|\pi_i| < 1,\ \forall i = 1, \ldots, p$ is necessary and sufficient for the process to be stationary. In this case where the coefficients $\{\alpha_k\}$ have been determined by converting the coefficients $\{\beta_k\}$ of a differenced AR process, some of the roots of the auxiliary equation will have magnitude greater than or equal to one.

ARIMA processes

The first-differencing in (4.6) can be generalised so that we difference d times. Writing

$$z_t = \nabla^d y_t = \sum_{k=1}^{p} \beta_k z_{t-k} + \nu_t \tag{4.9}$$

we obtain an example of an autoregressive integrated moving average (ARIMA) process on y. In this case we refer to z as an ARIMA(p,d,0) process, meaning that we have an autoregressive process of order p on data which has been differenced d times. In general there might also be a moving average component in such processes, represented by the third parameter. We are not concerned with moving averages here, though—see e.g. Chatfield (1975) for more details.

Vector AR processes

In some situations, such as within the AR-HMM in Section 4.4, it is desirable to use an AR process to represent multivariate data. Equation 4.1 can be simply expanded to the multivariate case,

$$p(\mathbf{y}_t | \mathbf{y}_{t-p:t-1}) \sim \mathcal{N}\left(\mathbf{y}_t; \sum_{k=1}^{p} \Theta_k \mathbf{y}_{t-k}, \Sigma\right) \,, \tag{4.10}$$

where $\mathbf{y}_t$ is an observation of dimension d, and Θ_k is a $d \times d$ matrix of AR coefficients for lag k. Learning of maximum likelihood proceeds analogously to the univariate model, where the Yule-Walker equations are extended to include the cross-correlations between each observation dimension at each lag, giving a system of linear equations with pd^2 unknowns.

4.2 HMMs

Under the assumption that a particular regime gives rise to observations with a Gaussian distribution, a Hidden Markov Model can be used to model the dynamics of the overall system[1]. First, consider a single time slice of the HMM. This is referred to as the conditional Gaussian model and is shown graphically in Figure 4.1(a). There is a hidden discrete variable s which can take K different settings. The distribution of the observations is determined by the value of s such that

$$p(\mathbf{y}|s=i) \sim \mathcal{N}(\mathbf{y};\mu^{\{i\}},\Sigma^{\{i\}}) \; . \tag{4.11}$$

Given a novel data point $\mathbf{y}^*$, the probability of the corresponding setting of s which generated that measurement is given by Bayes' rule,

$$p(s=i|\mathbf{y}) = \frac{p(y|s=i)p(s=i)}{\sum_{j=1}^{K} p(y|s=j)p(s=j)} \; . \tag{4.12}$$

The conditional Gaussian model treats each observation independently, regardless of its time of observation. To add dynamics, we make the state s_t dependent on s_{t-1}, as shown in Figure 4.1(b). There is therefore a Markovian dependency on the hidden variables; a hidden Markov model. The setting of the hidden variable s_t evolves according to a transition matrix Z of dimension $K \times K$, where Z_{ij} is the probability of the hidden variable moving from state i to state j in the course of a single time step. Clearly the rows of Z must sum to unity.

The observation equation for each time step is equivalent to the static case in Equation 4.11:

$$p(\mathbf{y}_t|s_t=i) \sim \mathcal{N}(\mathbf{y}_t;\mu^{\{i\}},\Sigma^{\{i\}}) \; . \tag{4.13}$$

Learning

In general for this application, it is possible to obtain labelled training data $\{\mathbf{y}_t, s_t\}$ for $t = 1, \ldots, N$. Given this it is easy to calculate the maximum likelihood parameters, $\mu^{\{i\}}$ and $\Sigma^{\{i\}}$ for each regime $i = 1, \ldots, K$, taking the sample means and covariances of the observations $\mathbf{y}_t$ associated with each regime. Estimates of the transition probabilities are given by

$$P(s_t=j|s_{t-1}=i) = \frac{n_{ij}+\zeta}{\sum_{k=1}^{M} n_{ik}+\zeta} \; , \tag{4.14}$$

[1] Note that this is a specific case, and that in general any reasonable distribution can be used to model observations in the HMM.

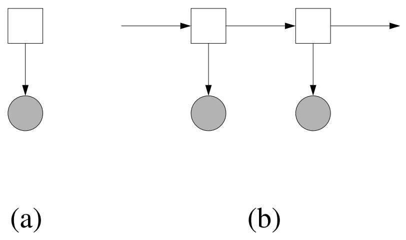


Figure 4.1: Graphical representations of (a) the conditional Gaussian model, (b) the hidden Markov model. Squares are discrete values, circles are continuous and shaded nodes are observed.

where n_{ij} is the number of transitions from state i to state j in the training data. The constant terms ζ (set to $\zeta = 1$ in the experiments described later in this thesis) are added to stop any of the transition probabilities being zero or very small. While a zero probability could be useful for a sequence of states that we know are impossible, in general we want to avoid it. This method is theoretically justified as a maximum a posteriori estimate where the prior is given by a Dirichlet distribution.

If labelled training data was unavailable, then unsupervised learning is possible with this model using the Baum-Welch algorithm.

Inference

For the application of real time clinical monitoring, we are interested in filtering, inferring s_t from the observations $\mathbf{y}_{1:t}$. This involves considering the possibility of having been in each state at every time point. Given estimates of the initial conditions, $p(s_0 = i)$, $i = 1, \ldots, K$, inference proceeds recursively:

$$p(s_t = i|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|s_t = i) \sum_{j=1}^{M} p(s_t = i|s_{t-1} = j) p(s_{t-1} = j|\mathbf{y}_{1:t-1}) \;, \tag{4.15}$$

normalised so that $\sum_{i=1}^{M} p(s_t = i|\mathbf{y}_{1:t}) = 1$. All the historical information relevant to the calculation of $p(s_t|\mathbf{y}_t)$ is contained in the discrete distribution $p(s_{t-1}|\mathbf{y}_{1:t-1})$. Exact inference is therefore possible, where the computation time scales as $O(NK)$. If an offline application was being considered, smoothing could also be done in a backwards pass, incorporating the information about s_t provided by the measurements $\mathbf{y}_{t+1:N}$.

4.3 Factorial HMMs

In the previous section it has been assumed that the discrete hidden state represents a number of mutually exclusive regimes that fully characterise the operation of the system generating the observations. In applications such as neonatal monitoring, there may in fact be many independent factors affecting the observations. In these cases it is possible to factorise the state variable, representing it in terms of all contributing factors. This has advantages in terms of interpreting the output, setting parameters, and space requirements of the model. The Factorial Hidden Markov Model is shown in Figure 4.2(a). In this model, M factors $f_t^{(1)}, \ldots, f_t^{(M)}$ affect the observations $\mathbf{y}_t$. The factor $f^{(m)}$ can take on $L^{(m)}$ different values. The state space is the cross product of the factor variables,

$$s_t = f_t^{(1)} \otimes \cdots \otimes f_t^{(M)} \ . \tag{4.16}$$

with $K = \prod_{m=1}^{M} L^{(m)}$ different values. The value of $f_t^{(m)}$ depends on $f_{t-1}^{(m)}$. The factors are a priori independent, so that

$$p(s_t|s_{t-1}) = \prod_{m=1}^{M} P(f_t^{(m)}|f_{t-1}^{(m)}) \ . \tag{4.17}$$

Notice that the factors are not, in general, a posteriori independent. Specific inference routines for this model are given by Ghahramani and Jordan (1997), who also provide approximate inference algorithms which can be applied if the state space is very high dimensional.

4.3.1 Factorial HMMs for condition monitoring

On the assumption that the observations under each regime have a Gaussian distribution, the FHMM can be used to infer the combination of factor settings underlying a novel sequence of physiological measurements. In many cases this assumption seems unrealistic, however, and a visual inspection of typical vital signs monitoring data such as that shown in Chapter 2 suggests that more complex dynamical structure can be present. A pattern such as a steady rise in heart rate could be represented by chaining together a sequence of regimes with successively higher means, though modelling more complex dynamics such as oscillations would be difficult. One approach to using an HMM-based model for more complex dynamics is to find some features which can be derived from the raw data that provide a stationary representation of a certain dynamical regime. Taking the example of modelling oscillations, in some circumstances

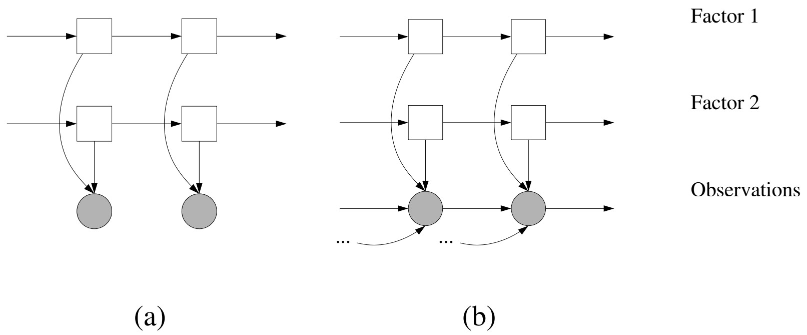


Figure 4.2: Graphical representations of different factorial models, with $M = 2$ factors. Squares are discrete values, circles are continuous and shaded nodes are observed. Examples of factors include probe dropouts, recalibrations, the presence or absence of respiratory or cardiovascular problems. (a) the Factorial HMM, (b) the Factorial AR-HMM.

it might be possible to extract frequency domain features which give such a representation. This can be a compelling approach when there is a high volume of data, as inference within the HMM or FHMM is relatively fast.

It might be thought that training data would be needed for every possible combination of factors, but it can usually be simulated as we know that some of the phenomena we want to model only affect a subset of the channels, or override other phenomena. As an example, one of the artifact processes we would like to model is a probe dropout, where one or more channels go to zero. Clearly this artifact in conjunction with any other event gives the same sequence of observations. This will be elaborated on in Chapter 5.

4.4 Autoregressive HMMs

Observation dynamics can be modelled more generally using autoregressive (AR) processes. It is possible to extend the HMM so that each observation value $\mathbf{y}_t$ depends on the preceeding p observations, giving the autoregressive HMM (AR-HMM) pictured in Figure 4.2(b). This model is referred to by some authors as a switching AR process. Practically, the only change required to an HMM is to alter the observation equation to the following:

$$p(\mathbf{y}_t|s_t = i) \sim \mathcal{N}\left(\mathbf{y}_t; \sum_{k=1}^{p} \Theta_k^{\{i\}} \mathbf{y}_{t-k}, \Sigma^{\{i\}}\right) , \tag{4.18}$$

so that the ith regime is a vector AR(p) process, with autoregressive matrices $\Theta_1^{\{i\}}, \ldots, \Theta_p^{\{i\}}$, driven by noise with covariance $\Sigma^{\{i\}}$. Using the above equation in place of (4.13), inference proceeds as for the HMM.

Where multiple independent factors affect the observations, the state variable in the AR-HMM can be factorised according to Equation 4.16, producing the Factorial AR-HMM, illustrated in Figure 4.2(b).

4.4.1 Factorial AR-HMMs for condition monitoring

The Factorial AR-HMM can represent stationary regimes with any power spectrum to an arbitrary accuracy, given a sufficiently high autoregressive order p. It is therefore more likely to be able to be able to represent physiological dynamics adequately compared to the Factorial HMM. However, in practice there are drawbacks to this model. During inference, the model can only confidently switch into a regime if the last p observations have been generated under that regime; there will be a loss of accuracy if any of the measurement channels have dropped out in that period, for example, or another artifactual process has affected any of the readings. Also, observations in this application are caused by both the physiology of the baby and by artifactual means, and the Factorial AR-HMM has no way of distinguishing between these. Finally, there are different sources of uncertainty in the problem of physiological condition monitoring, which the Factorial AR-HMM cannot specifically represent. The physiological dynamics themselves are subject to uncertainty, as for example manifestations of bradycardia will differ slightly between episodes and for different babies. There is also measurement error from the monitoring equipment. The Factorial AR-HMM cannot explicitly model these sources of uncertainty, having only one error term per regime.

4.5 Kalman filters

The Kalman filter is another type of model for a single regime, more general than the AR process, with a continuous hidden variable $\mathbf{x}_t$ (which we refer to as the 'state'). Observations are generated by a linear Gaussian state space, where the state evolves according to

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) \sim \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q}) \,, \tag{4.19}$$

while the observations are generated from the state as follows:

$$p(\mathbf{y}_t|\mathbf{x}_t) \sim \mathcal{N}(\mathbf{C}\mathbf{x}_t, \mathbf{R})\ . \tag{4.20}$$

Here **A** is a square system matrix, **C** is known as the observation matrix, while **Q** and **R** are noise covariances.

AR processes in state space form

A univariate AR process $\tilde{x}_t = \sum_{k=1}^{p} \alpha_k \tilde{x}_{t-k} + \nu_t$, where $\mathrm{var}(\nu_t) = \sigma_\nu^2$, can be modelled in state space form by storing lagged versions of the state, as for example in the following:

$$\mathbf{x}_t = \begin{bmatrix} \tilde{x}_t \\ \tilde{x}_{t-1} \\ \vdots \\ \tilde{x}_{t-p} \end{bmatrix} \tag{4.21}$$

$$\mathbf{A} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \tag{4.22}$$

$$\mathbf{Q} = \begin{bmatrix} \sigma_\nu^2 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \tag{4.23}$$

The observation matrix $\mathbf{C} = [1, 0, \dots, 0]$ picks out the most recent value $\tilde{x}_t$. If the AR process itself is thought to be subject to observation noise then **R** can be set to a positive scalar value. If not then it can be set to zero. Other constructions are possible for storing lagged versions of the state (de Jong and Penzer, 2004; Press et al., 1992, §13.6).

Another class of dynamics which can be represented are *structural models*, for example in which the data follows a random walk or a local linear trend model. Further details of these kinds of models can be found in Harvey (1990).

Learning

Setting of parameters is application specific, and described in detail in the next chapter. One general training method is worth mentioning here, where we assume that each observed dimension in the data is modelled by a hidden autoregressive process of order p. That is, each observation channel is independent of the others, and the dynamics of each channel are governed by a latent autoregressive process $x_t \sim \mathcal{N}\left(\sum_{k=1}^{p} \alpha_k x_{t-k}, \sigma_q^2\right)$. The observations are given by the latent process plus noise, $y_t \sim \mathcal{N}\left(x_t, \sigma_r^2\right)$.

In this case, white observation noise adds a pedestal to the power spectrum of the observed data, the value of which can be estimated by inspection. The variance of the hidden AR process γ_0 is then the variance of the observations minus this pedestal, while the autocovariances of the hidden process $\gamma_{1:p}$ are the same as the empirical autocovariances of the observations. The hidden AR coefficients $\alpha_{1:p}$ can then be obtained with the Yule-Walker equations (Section 4.1), and the system noise variance is given by $\sigma_q^2 = \gamma_0 - \sum_{k=1}^{p} \alpha_k \gamma_k$.

Inference

Given the parameters $\mathbf{A}, \mathbf{Q}, \mathbf{C}, \mathbf{R}$ and observed data $\mathbf{y}_{1:t}$, the Kalman filter equations can be used to find estimators of the value of the hidden state $\hat{\mathbf{x}}_t$ and the estimated error covariance $\hat{\mathbf{P}}_t$. Estimates are built up recursively for successive time steps, and are calculated at each time step in two stages: prediction and correction. Before the data point $\mathbf{y}_t$ has been considered, the prediction stage calculates what the most likely estimators are for time t based only on the estimators for the previous time step and the model parameters $\mathbf{A}$ and $\mathbf{Q}$:

$$\hat{\mathbf{x}}_t^- = \mathbf{A}\hat{\mathbf{x}}_{t-1} \tag{4.24}$$

$$\hat{\mathbf{P}}_t^- = \mathbf{A}\hat{\mathbf{P}}_{t-1}\mathbf{A}^\top + \mathbf{Q} \tag{4.25}$$

where $\hat{\mathbf{x}}_t^-$ and $\hat{\mathbf{P}}_t^-$ denote the estimates at time t made without taking into account the observation $\mathbf{y}_t$. The most recent observation is then used to refine the estimates as follows:

$$\tilde{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{C}\hat{\mathbf{x}}_t^- \tag{4.26}$$

$$\mathbf{S}_t = \mathbf{C}\hat{\mathbf{P}}_t^-\mathbf{C}^\top + \mathbf{R} \tag{4.27}$$

$$\mathbf{K}_t = \hat{\mathbf{P}}_t^-\mathbf{C}^\top\mathbf{S}_t^{-1} \tag{4.28}$$

$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_t^- + \mathbf{K}_t\tilde{\mathbf{y}}_t \tag{4.29}$$

$$\hat{\mathbf{P}}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{C})\hat{\mathbf{P}}_t^- \tag{4.30}$$

in which $\tilde{\mathbf{y}}_t$ denotes the innovation (the difference between the predicted and actual observations), $\mathbf{S}_t$ is the covariance of the innovation, and $\mathbf{K}_t$ is known is the Kalman gain and is the optimal reweighting of the estimate $\hat{\mathbf{x}}_t^-$ given the new observation $\mathbf{y}_t$. Note that $\mathbf{K}$,$\mathbf{P}$ and $\mathbf{S}$ are both independent of the data. Also notice that an effect of the prediction step is always to increase the uncertainty of the estimate. The act of observing a new measurement always decreases the variance of the estimate, as one would expect.

As in previous models, smoothing is possible in an offline application but is not considered here. It is possible to get an idea of how well the parameters have been fitted by analysing the sequence of innovations. If the model is well fitted, the innovation term $\tilde{\mathbf{y}}_t$ should be drawn from a Gaussian distribution with zero mean and covariance $\mathbf{S}_t$. More will be said about model validation in section 5.4.

4.6 Switching Kalman Filters

The Switching Kalman Filter is shown in Figure 4.3(a). In this model, the hidden switch setting s_t affects the hidden continuous state $\mathbf{x}_t$ and the observations $\mathbf{y}_t$. Conditional on a particular switch setting, the model is equivalent to a linear Gaussian state space as in the previous section. The switch setting evolves according to the transition probabilities $p(s_t|s_{t-1})$, and for a given setting of s_t, the hidden continuous state and the observations are related by:

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{A}^{\{s_t\}}\mathbf{x}_{t-1} + \mathbf{d}^{\{s_t\}}, \mathbf{Q}^{\{s_t\}}) \tag{4.31}$$

$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{C}^{\{s_t\}}\mathbf{x}_t, \mathbf{R}^{\{s_t\}}) \tag{4.32}$$

Here $\mathbf{A}^{\{s_t\}}$ is a square system matrix, $\mathbf{d}^{\{s_t\}}$ is a drift vector, $\mathbf{C}^{\{s_t\}}$ is the state-observations matrix, and $\mathbf{Q}^{\{s_t\}}$ and $\mathbf{R}^{\{s_t\}}$ are noise covariance matrices. Note that in this formulation, all dynamical parameters can be switched between regimes. SKFs referred to in the literature sometimes switch only the state dynamics $\{\mathbf{A}, \mathbf{Q}\}$, or the observation dynamics $\{\mathbf{C}, \mathbf{R}\}$.

Inference

The time taken to calculate the exact filtering distribution $p(s_{1:t}, \mathbf{x}_{1:t}|\mathbf{y}_{1:t})$ in the SKF scales exponentially with t, making it intractable. This is because the probabilities of having moved between every possible combination of switch settings in times $t-1$ and t are needed to calculate the posterior at time t. Hence the number of Gaussians

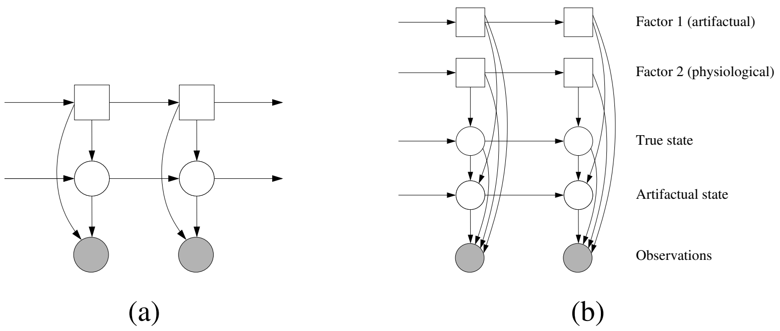


Figure 4.3: Graphical representations of (a) the Switching Kalman Filter, (b) a Factorial Switching Kalman Filter with $M = 2$ factors.

needed to represent the posterior exactly at each time step increases by a factor of K, the number of cross-product switch settings.

The Gaussian Sum approximation (Alspach and Sorenson, 1972) can be used to reduce the time required for inference. At each time step we maintain an approximation of $p(\mathbf{x}_t|s_t,\mathbf{y}_{1:t})$ as a mixture of K Gaussians. Calculating the Kalman updates and likelihoods for every possible setting of s_{t+1} will result in the posterior $p(\mathbf{x}_{t+1}|s_{t+1},\mathbf{y}_{1:t+1})$ having K^2 mixture components, which can be collapsed back into K components by matching means and variances of the distribution, as described in (Murphy, 1998).

Rao-Blackwellised particle filtering (RBPF) (Murphy and Russell, 2001) is another technique for approximate inference, which exploits the conditionally linear dynamical structure of the model to try and select particles close to the modes of the true filtering distribution. A number of particles are propagated through each time step, each with a switch state s_t and an estimate of the mean and variance of $\mathbf{x}_t$. A value for the switch state s_{t+1} is obtained for each particle by sampling from the transition probabilities, after which Kalman updates are performed and a likelihood value can be calculated. Based on this likelihood, particles can be either discarded or multiplied. Because Kalman updates are not calculated for every possible setting of s_{t+1}, this method can give a significant increase in speed when there are many factors. The fewer particles are used, the greater the tradeoff of speed against accuracy, as it becomes less likely that the particles can collectively track all modes of the true posterior distribution. RBPF has been shown to be successful in condition monitoring problems with switching linear dynamics, in the context of industrial processes (Morales-Menedez et al., 2003) and fault detection in mobile robots (de Freitas et al., 2004).

4.7 Factorial Switching Kalman Filters

Factorising the switch variable s_t according to Equation 4.16 yields the Factorial Switching Kalman Filter, shown in Figure 4.3(b). The order of the execution time here is higher than previous models in the chapter, as the Kalman filter equations need to be run at every time point for every combination of factors. When new factors are added, the number of Kalman filter iterations required grows exponentially.

Where there are K possible settings of the switch variable s_t, the Gaussian sum approximation requires K^2 Kalman filter recursions at each step, to calculate the possibility of having moved between every combination of factor settings. With the factorial model, it is possible to constrain the transitions so that only one factor can change its setting at each time step. In the unconstrained version, the value of the i^{th} factor setting $f_t^{(i)}$ can take on any value from $1, \ldots, K^{(i)}$. In the proposed approximation we introduce the constraint that if for any factor setting $f_t^{(i)} \neq f_{t-1}^{(i)}$, then the other factor settings are fixed such that $f_t^{(j)} = f_{t-1}^{(j)}$ where $j \neq i$.

This speeds up inference from order $O(K^2)$ per time step to $O(K \log K)$. Restricting the number of possible transitions in this way works best where the factor settings can be assumed to change slowly relative to the frequency of the observations.

4.7.1 Factorial SKFs for condition monitoring

The FSKF allows us to construct a more sophisticated representation of the causes underlying physiological observations than would be possible in the Factorial HMM or Factorial AR-HMM. For instance, the state can be split into two groups of continuous latent variables, those representing the ‘true’ physiology and those representing the levels of artifactual processes. In turn, factors can be physiological or artifactual processes. Physiological factors can affect any state variable, whereas artifactual processes affect only artifactual state. If the ECG lead drops out and the heart rate measurements go to zero, we do not really believe that the baby’s actual heart rate is zero; in the FSKF, the ‘ECG dropout’ regime can be parameterised with a zero in the observation matrix $\mathbf{C}$, in such a way as to represent that the heart rate is entirely unobserved. The Kalman filter recursions will continue in this regime so that uncertainty is added to the heart rate estimate at each time step ($\mathbf{P}_t^-$ always increases in Equation 4.25). The zero in the observation matrix ensures that the observations cannot affect the estimate—the relevant component of the Kalman gain in Equation 4.28 will also be zero. This formulation of the FSKF for physiological condition monitoring is

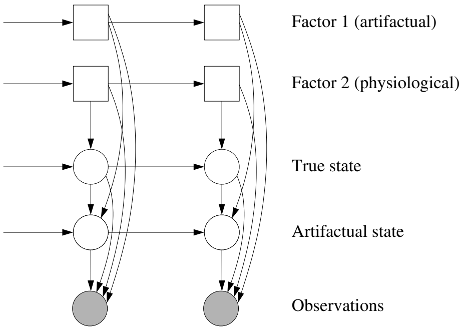


Figure 4.4: Factorial Switching Kalman Filtering for physiological condition monitoring. The state is split up into two sets of variables, containing estimates of the 'true' physiology and of the levels of artifactual processes.

illustrated in Figure 4.4.

The 'artifactual state' variables in Figure 4.4 are useful in practise to keep track of the levels associated with different artifactual processes. For example, when the core temperature probe is disconnected as described in section 2.3.1, the temperature measurements follow an exponential decay. The artifactual state variables can be used to track this decay, while at the same time the estimates of the baby's true temperature can be propagated in the true state variables. The true state can affect the artifactual state, as for example the artifactual temperature decay following a probe disconnection starts at the same level as the true temperature. Note that this dependency is one-way, as there is never any way in which artifactual measurements can affect the true physiology.

Maintaining hidden state in this way helps to overcome the problem with the Factorial AR-HMM in which the last p measurements needed to be from the same regime in order to make confident inferences. Also note that the uncertainty arising from measurement error is also more appropriately modelled than in the HMM-based models.

One cost of this expressive power is the extra computational requirements of inference in the model. A more complex model has other drawbacks, including more parameters to set. A higher number of latent variables to infer from the observations means that we are at risk of creating a 'tower of jelly'. Model verification is therefore more of a consideration with the FSKF, discussed further in Chapter 6.

Factorial HMM		$O(NK^2C_1)$
Factorial AR-HMM		$O(NK^2C_2)$
Factorial SKF	EXACT	$O(K^NC_3)$
	GS	$O(NK^2C_3)$
	CONSTR	$O(NK\log KC_3)$

Table 4.2: Comparison of the complexity of inference of the the three dynamical models introduced in this chapter, in terms of the total number of switch settings K and the number of data points N. To compare the computational requirements of these models, constant terms also need to be considered: C_1 corresponds to the evaluation of a Gaussian density, while C_2 is similar but includes the calculation of a weighted average based on the order of autoregression. C_3 corresponds to one pass of the Kalman filter equations and is typically considerably larger than C_1 or C_2. EXACT denotes inference of the exact posterior, GS denotes the Gaussian sum approximation, and CONSTR denotes the Gaussian sum approximation with factor transitions constrained as described in Section 4.7.

4.8 Summary

This chapter has introduced three factorised switching dynamical models, applicable to the problem of physiological condition monitoring. The Factorial HMM has factor dynamics but no observation dynamics. The Factorial AR-HMM can represent any stationary dynamical regimes with certain drawbacks in operation for condition monitoring, and the FSKF can explicitly model more aspects underlying the physiological monitoring data but with a commensurate increase in the time required for inference. Table 4.2 provides a comparison of the complexity of inference in these three models.

Chapter 5

Parameter estimation

This chapter describes in general how parameters can be set in the models introduced in Chapter 4, and in particular how parameters were set to model some of the known artifactual and physiological regimes described in Chapter 2. Training the FSKF model is the main concern, because of its advantages for our application summarised in section 4.7.1. For experimental comparison in Chapter 7, however, the training procedure for the FHMM is also introduced.

For this application it is possible to obtain sets of labelled training data $D = \{\mathbf{y}_t, s_t\}$, where the labels s_t have been provided by experts. Taking this into account, sections 5.2 and 5.3 describe general methods with which to train these models. We exploit the fact that labelled data allows both the FHMM and the FSKF to be trained by conditioning on the switch setting and dealing with each regime one at a time. Section 5.4 explains how to check that these models are a good fit to the training data, and section 5.4.1 describes software written for this task, the 'Kalman Filter Diagnostic Dashboard'.

Sections 5.5 to 5.8 show how parameters are set for the known regimes in our application, and how these parameters are verified. The details of training for each regime are presented in the same order that they appear in Chapter 2.

5.1 The goal of training

We are interested in finding a model which corresponds to the true distribution of the data as closely as possible. The closer the correspondance, the more accurate the predictions and classifications inferred by the model will be. The true distribution is unknown in general, but we can use the empirical distribution of training data D

as an approximation. Bayes' rule provides a way of calculating a distribution over parameters in the light of training data, by treating the model parameters θ as random variables with a prior $p(\theta)$ over the space Θ:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_\Theta p(D|\theta)p(\theta)\,\mathrm{d}\theta} \;. \tag{5.1}$$

This distribution is often difficult to calculate in a closed form, so it is common practice to calculate a point estimate — the mode of $p(\theta|D)$ — and use this value as a fixed parameter setting. This is known as the maximum a posteriori (MAP) estimate:

$$\hat{\theta}_{\text{MAP}}(D) \quad \leftarrow \quad \arg\max_\theta \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)\,\mathrm{d}\theta} = \arg\max_\theta p(D|\theta)p(\theta) \;. \tag{5.2}$$

When a uniform prior $p(\theta)$ is used, the maximum likelihood (ML) estimate is obtained:

$$\hat{\theta}_{\text{ML}}(D) \quad \leftarrow \quad \arg\max_\theta p(D|\theta) \;, \tag{5.3}$$

reflecting the intuitive principle that the higher the probability of a model with a particular parameter setting having randomly generated the training data, the more plausible it is.

It is clear there are different potential sources of errors in this process. There may be *generalisation error*, since the training data is limited in quantity and so its distribution will tend to differ from the true distribution. *Modelling error* may also arise when the parametric family of distributions in the space Θ does not contain the true distribution. This is likely to be the case to some extent for example where linear-Gaussian dynamics are used to model a complex physiological system. There is also error due to the use of a point estimate of the parameters, rather than the full a posteriori parameter distribution. It is also not often practical to find a global optimum for either $\hat{\theta}_{\text{MAP}}$ or $\hat{\theta}_{\text{ML}}$ when the model contains latent variables.

Likelihood-based training using $\hat{\theta}_{\text{ML}}$ for the FHMM and FSKF is discussed in sections 5.2 and 5.3. Section 5.3.2 describes why it is necessary to calculate $\hat{\theta}_{\text{MAP}}$ in some cases for the FSKF, and how priors might be obtained. Throughout this chapter, potential sources of the above types of errors specific to the application are discussed. Due to the different sources of error it is also necessary to verify that trained models are well fitted to the training data, discussed in section 5.4.

5.2 FHMM training

Conditional on the hidden state s, the HMM or FHMM reduces to a Gaussian model. Given training data which is labelled with the regime being followed at each time step, the maximum likelihood parameters, $\mu^{(i)}$ and $\Sigma^{(i)}$ can be calculated for each regime $i = 1, \ldots, K$ by taking the sample means and covariances of the observations $\mathbf{y}_t$ in the training data which are labelled as being from that regime. Where $\mathbf{y}_{1:N}$ is a sequence of training data which follows regime i, we therefore use the estimators

$$\hat{\mu}^{(i)} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{y}_t \ , \tag{5.4}$$

$$\hat{\Sigma}^{(i)} = \frac{1}{N-1}\sum_{t=1}^{N} (\mathbf{y}_t - \hat{\mu}^{(i)})(\mathbf{y}_t - \hat{\mu}^{(i)})^\top \ . \tag{5.5}$$

We also need to learn the transition probabilities of moving between one state and another. Estimates of the transition probabilities are given by:

$$p(s_t = j | s_{t-1} = i) = \frac{n_{ij} + \zeta}{\sum_{k=1}^{M} n_{ik} + \zeta} \ , \tag{5.6}$$

where n_{ij} is the number of transitions from state i to state j in the training data. The constant terms ζ (set to $\zeta = 1$ in the experiments described in Chapter 7) are added to stop any of the transition probabilities being zero or very small. While a zero probability could be useful for a sequence of states that we know are impossible, in general we want to avoid it. This method is theoretically justified as a maximum a posteriori estimate where the prior is given by a Dirichlet distribution.

If labelled training data was unavailable, then unsupervised learning would be possible with this model using the Baum-Welch algorithm (Rabiner and Juang, 1989). The factorised case is explicity dealt with by Ghahramani and Jordan (1997). Labelled training data means that the model is fully observed in training, making the process simpler.

This concludes the training procedure for the FHMM. We now turn our attention to learning parameters in the FSKF, which is more complicated since this model has hidden variables even when conditioned on a switch setting.

5.3 FSKF training

Conditioned on the switch variable s_t, the SKF or FSKF is equivalent to a linear Gaussian state-space (Kalman filter). As the switch setting is labelled in our training data, this section concentrates on learning the parameters of a linear Gaussian state-space given multiple training examples. Training is different in character to the maximum likelihood estimation in the previous section, because the model to be learnt has a hidden variable, $\mathbf{x}_t$. In addition, the dimension of $\mathbf{x}_t$ is not known and also has to be determined. If this labelling was not available, EM would be possible for the full switching model; see Appendix C. EM training in the full model could also be applied if there was a limited supply of labelled data, but further unlabelled data to hand. To do this, training would first proceed by using the labelled data as described below. Having found an initial parameter setting in this way, EM would then be used to increase the likelihood of the model on the unlabelled data. However, a drawback of using EM on the full FSKF is that we lose the interpretability of the model; the switch settings would no longer necessarily have the same 'meanings'.

We assume that adequate labelled training data is available, and given this the following sections show how linear dynamical parameters are learnt for each regime. The transition probabilities of moving between switch settings are calculated as for the FHMM in (5.6).

5.3.1 Likelihood-based training methods

Under a linear Gaussian state-space model with parameters θ and initial settings $\mathbf{x}_0 \sim \mathcal{N}(\hat{\mathbf{x}}_0, \mathbf{P}_0)$, it is straightforward to calculate the likelihood of a training sequence, $p(y_{1:t}|\theta)$. One way of training is therefore to try and pick values of θ which maximise this likelihood. A model which has a high probability of randomly generating the observed data is intuitively more compelling than a rival model with low likelihood. However, although the likelihood function of a linear Gaussian state-space is straightfoward to evaluate at a point, is is difficult to optimise. A global optimum cannot normally be computed due to the complexity of the hidden variables which would need to be integrated out. There are many local optima, so the procedure is sensitive to initial settings. The next sections discuss methods which can be used to find parameters with high likelihood in the face of these difficulties.

5.3.1.1 EM for linear Gaussian state-spaces

Given initial parameter settings, the Expectation-Maximisation (EM) algorithm can be used to increase in the likelihood of the model with each iteration (it is at least guaranteed that the likelihood will not be decreased).

Calculating a complete likelihood or log likelihood involves integrating out the values of the unobserved variables, such that $L = \int_{\mathbf{x}_{1:t}} \log p(\mathbf{y}_{1:t}, \mathbf{x}_{1:t}|\theta, \mathbf{x}_0)\, d\mathbf{x}_{1:t}$ (referred to below as the 'true likelihood'). This is not tractable to optimise directly as there are many unobserved variables to integrate over. The EM algorithm is based on the assumption that if the hidden variables $\mathbf{x}_{1:t}$ were observed, then maximising the likelihood would be easier. The principle is to use an estimate of the parameters to calculate a distribution for the hidden variables (the 'E-step'), then use this distribution over the unobserved variables to maximise the likelihood with respect to the model parameters (the 'M-step'). The latter step updates the parameters in such a way as to maximise the expected complete data log likelihood:

$$\hat{\theta}_k \leftarrow \arg\max_{\theta} \int \log p(\mathbf{y}_{1:t}, \mathbf{x}_{1:t} | \theta) p(\mathbf{x}_{1:t} | \mathbf{y}_{1:t}, \hat{\theta}_{k-1}) \, \mathrm{d}\mathbf{x}_{1:t} \; . \tag{5.7}$$

Given some initial settings, successive applications of these updates are guaranteed to increase the true likelihood of the model, and will cause the parameters to converge to a local optimum. Note that when the true likelihood has many modes (which is normally the case), the value of this optimum is sensitive to the initial settings. A derivation of this algorithm for a linear Gaussian state-space is given in Appendix C.

5.3.1.2 Training as a hidden AR process

Another method of training a linear Gaussian state-space makes the assumption that each observed dimension in the data is a hidden autoregressive process of order p. In this case, white observation noise adds a pedestal to the power spectrum of the observed data in (4.3), the value of which can be estimated by inspection. The variance of the hidden AR process γ_0 is then the variance of the observations minus this pedestal, while the autocovariances of the hidden process $\gamma_{1:p}$ are the same as the empirical autocovariances of the observations. The hidden AR coefficients $\alpha_{1:p}$ with maximum likelihood can then be obtained with the Yule-Walker equations (Section 4.1), and the system noise variance is given by $\sigma_q^2 = \gamma_0 - \sum_{k=1}^{p} \alpha_k \gamma_k$. This method can also be used for training under the assumption that each observed channel is a hidden ARIMA process, using the conversion described in section 4.1.

5.3.2 Shortcomings of likelihood-based training

One shortcoming of the training schemes described so far is that they do not provide any facility for determining the dimensions of the hidden variables $\mathbf{x}$. In general the likelihood can always be increased by adding to the number of parameters in the model (increasing the dimensionality of $\mathbf{x}$), though as the number of parameters increases relative to the size of the training data the generalisation error can also be expected to increase. Therefore some other model-learning principle needs to applied in conjunction with increasing the likelihood of training data.

For a linear Gaussian state-space there are also practical difficulties with the size of the parameter space and the number of local optima in the likelihood function. Unlike the FHMM, where a Gaussian regime is straightforward to train with maximum likelihood, there is a large set of potential parameterisations for a linear Gaussian state-space. Because the dimension of the hidden state $\mathbf{x}_t$ is unbounded, the search space is theoretically limited only by computational capacity. The task of setting parameters in our dynamical model can be thought of as a search through this large space of potential parameterisations and settings. Because of the size of the search space, the number of local optima, and the number of possible initial settings with which to begin the search, it is likely to be difficult to carry out. In the following section, we discuss how model priors $p(\theta)$ can be used to address these problems.

5.3.3 Model priors to reduce the search space

Prior information about which models are likely to be a good fit[1] makes the search more tractable, by using the information to concentrate on particular parts of the search space and rule out others. Priors are incorporated by using Bayes' rule to calulate the MAP estimate in (5.2). A priori information might be obtained for our purposes in the following ways:

1. Through the use of domain knowledge. In some cases, for example, the physics underlying the measurements are known which makes a particular parameterisation likely.

2. Analysing particular statistics can give an indication of the model structure. For instance, the ACF and PACF might indicate a hidden AR model. In this case the PACF should provide an indication of the model order p.

3. A low prior probability can be put on models with large numbers of parameters. This might be done with an information criteria such as Akaike's Information Criterion (AIC) (Akaike, 1974) or the Bayesian Information Criterion (BIC) (Schwarz, 1978). In general we need to penalise models with too many parameters, as the generalisation error is likely to be higher.

Priors might be formally quantified, allowing the explict maximum a posteriori calculation in (5.2), or we might informally use a priori reasoning to rule out all but a

[1] The term 'model' is overloaded here; earlier it was used to refer the overall statistical construction analysing the data (e.g. the FSKF), whereas here we take it to mean a particular parameterisation and setting of parameters for a certain regime.

particular class of models. For example, when a probe cools to an ambient temperature, the measurements will be governed by Newton's law of cooling and will therefore take the form of an exponential decay. Knowing this, we can restrict the search space to models of this type, as in section 5.8.2. The same type of reasoning is used when selecting a model for a fall in incubator humidity associated with handling (section 5.8.5), and again when modelling the effects of multiple factors being active (section 5.10.1). The latter principle, that models with fewer parameters are more compelling a priori, is used informally when selecting classes of dynamical models for different observation channels in section 5.7.2. Here the idea is to find a class of dynamical models for each channel which verification tests show to be adequate, using the lowest number of parameters (e.g. the lowest order of autoregression). If a low-order autoregressive model fits an observation channel well, then it is not necessary to continue searching through the space of higher-order models.

Though it is clear that the use of informative prior information will improve training, considering the sources of error described in section (5.1) it is still desirable to find ways of verifying the fit of a model. By this we mean checking whether our modelling assumptions seem consistent with training data, in order to try to obtain some assurance that generalisation error and modelling error will be limited when the model is applied to novel data. The discussion will therefore now turn to methods of verification for linear Gaussian state-space models.

5.4 Verification methods

Having arrived at a parameter setting for a particular regime, we need to verify that it is appropriate. Although we consider there to be a 'true', stationary distribution for each regime, the training data is limited in quantity and so may not be completely representative of this distribution. As well as estimating whether generalisation error will be a problem, it is also natural to look for some indication that the assumptions made during modelling are justified.

The most obvious test of model fit is performance on unseen data, as carried out in Chapter 7. However this uses up potential training data, and several other diagnostic procedures are possible which do not require labelled data. For example, a simple but effective test for any generative model is to sample data from the model and check by eye whether it has the desired characteristics. Modelling assumptions can also be checked by analysing statistics regarding the model's operation on the training data.

Note that the likelihood of a model on training data is not informative for verification, as the information in this statistic has been ‘used up’ in the training phase; we have already tried to maximise it as far as possible. However, there are other statistics which indicate the goodness of fit which were not used directly to train the model.

Candy (1986) describes tests for verification of a linear Gaussian state-space, where the model is applied as a Kalman filter to training data and the state estimates and innovation sequence[2] are considered. There are certain properties of the innovation sequence $\tilde{\mathbf{y}}_t$ which are particularly useful for validating a particular model. Recall that the Kalman filter equations provide an estimate of the covariance of the innovations $\mathbf{S}_t$ in (4.27), which is the covariance of the state prediction ($\mathbf{P}$), projected into the data space ($\mathbf{CPC}^\top$) and added to the covariance of the measurement noise ($\mathbf{CPC}^\top + \mathbf{R}$). Also recall that this estimate is independent of the observed data; it describes the uncertainty under ideal conditions, when the observations are drawn from a distribution that exactly corresponds to the model. Our first test is therefore to examine how closely this estimator corresponds to the actual innovation sequence when applied to training data, knowing that ideally $\tilde{\mathbf{y}}_t \sim \mathcal{N}(0, \mathbf{S}_t)$:

Test 1 *The innovations $\tilde{\mathbf{y}}_t$ should come from a Gaussian distribution with zero mean and covariance $\mathbf{S}_t$.*

It can also be shown (Candy, 1986) that each component of the innovations sequence should be i.i.d. white noise, giving the second test criterion:

Test 2 *The innovation sequence should be uncorrelated in time.*

When both of these criteria are satisified for some training data, the model can be considered ‘tuned’. These tests can be practically carried out in different ways. First we can plot the ith component of the innovation, $\tilde{\mathbf{y}}_{i,t}$, with $\pm\sqrt{2\mathbf{S}_{ii,t}}$, where $\mathbf{S}_{ii}$ is the ith component of the diagonal of $\mathbf{S}$. If the model is wrong then it is often easy to see whether there is any signal in the innovations this way, and whether the mean or variance do not match. The innovation covariance $\mathbf{S}_t$ normally converges quickly. After convergence, it is therefore possible to plot the cumulative distribution function (cdf) for each component of the innovations against the theoretical cdf of the distribution $\mathcal{N}(0, \mathbf{S}_i^*)$, where $\mathbf{S}_i^*$ denotes the value which $\mathbf{S}_{ii,t}$ converges to as t becomes large. This can again be checked by eye, or the Kolmogorov-Smirnov test can be used to heuristically check the hypothesis that the distributions are matched. Discrepancies between

[2] To recap section 4.5: at each time step, the innovation is the difference between the predicted and actual observations, $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{C}\hat{\mathbf{x}}_t^-$.

the expected and actual distributions of the innovations are also well shown on a Q-Q plot, described in section 5.4.1. A plot of the autocorrelation of each component of the innovation sequence can also be used to assess whether there is any structure left in the innovations.

Another diagnostic which may be useful is to plot components of the Kalman gain **K** and the prediction covariance **P**. Similarly to the estimated innovation covariance, both these values are unrelated to the data. It can be assessed whether the value that each component converges to is reasonable or not. If **K** and **P** have components with a high magnitude, the model may not be specific enough. If all the components are low, it may need to be questioned whether the projected accuracy is feasible given the application. This latter case can be a problem when the prior **Q** on the model is too strong, which in practice causes observations to be ignored. Our next test is therefore as follows:

Test 3 *The Kalman gain and prediction covariance should converge to appropriate levels.*

Note that convergence is not always desired, however. See for example the dropout model in section 5.8.1, a regime for which the uncertainty is always expected to increase over time. Also note that each of the tests which have been descibed so far are applied to a single component of $\tilde{\mathbf{y}}_t$, $\mathbf{K}_t$ or $\mathbf{P}_t$ at a time. This might be time consuming to assess when the observations are high-dimensional. It is possible to formulate summary statistics for all dimensions in the innovation at once, but these are not considered here.

A further test, as mentioned above, is to look at the characteristics of data sampled from the model. This is an effective way to assess whether the model has captured the relevant characteristics of the regime it is supposed to be modelling:

Test 4 *Samples from the trained state-space should look like the training data.*

Other checks are possible too, for example by training another Kalman filter using a general parameterisation which will not be expected to do particularly well. For example, random walk dynamics might be used to model a decay in temperature. Verifying that this model has an inferior fit can serve as a sanity check.

The following section describes software which was produced to automate parts of this verification process.

5.4.1 The Kalman Filter Diagnostic Dashboard

Software was written to help evaluate the tests above, referred to as the 'Kalman Filter Diagnostic Dashboard' (KFDD). Particularly relevant is the ability to draw the innovation sequence for a particular Kalman filter applied to training data, the innovation autocorrelation, and distribution of the innovations compared to what would be expected under ideal conditions[3]. The KFDD therefore takes the parameters of a Kalman filter and a sequence of data as arguments, and displays a number of useful diagnostic charts.

Using this software, the ideas in the previous section can be illustrated with a simple example. The KFDD can be given a set of parameters and a sequence of data generated by a model with those parameters. This provides an example of an ideally fitted model. If we then provide the KFDD with the same parameters but data generated from a different model, the effects of the mismatch should be visible.

Consider two dynamical models, $\mathcal{M}_1$ and $\mathcal{M}_2$. $\mathcal{M}_1$ is a hidden AR(3) process, $y_t^{(1)} \sim \mathcal{N}(x_t, \eta_1)$ where $x_t \sim \mathcal{N}(\sum_{k=1}^{3} \alpha_k x_{t-k}, \eta_2)$. $\mathcal{M}_2$ is the same hidden AR(3) process with a sinusoidal component added, such that $y_t^{(2)} = y_t^{(1)} + 0.5\sin(\frac{t}{20})$. Samples $y_{1:T}^{(1)}$ and $y_{1:T}^{(2)}$ can be taken from these two models respectively, shown in Figure 5.1. This example is chosen to make explicit the idea that $\mathcal{M}_2$, and therefore $y_{1:T}^{(2)}$, contains structure which is not present in $\mathcal{M}_1$.

Figure 5.2 shows diagnostic information for the correctly fitted case, where $\mathcal{M}_1$ is applied as a Kalman filter to data $y_{1:T}^{(1)}$. The distribution of the innovations is first assessed by plotting it on a quantile-quantile (Q-Q) plot against the expected normal distribution. In this type of plot, if the data is drawn from a Gaussian distribution with the same variance (and in this case zero mean) it will appear as a straight line across the diagonal. It can therefore be seen that in this case the distributions are a close match. There is also clearly insignificant autocorrelation in the innovations, and from the plot of the innovations they appear to be iid white noise with the expected variance.

In Figure 5.3, model $\mathcal{M}_1$ is applied to data $y_{1:T}^{(2)}$, which we know does not have the same structure. Here it is clear to see that the extra periodic structure in the sample is present in the residuals, and that the model mismatch is also manifested as significant correlation in time. The distribution of the innovations shown in the Q-Q plot in the top left panel is no longer a close match to the expected normal distribution.

[3] It is not possible to plot the expected cdf for the entire innovation sequence, as the innovation covariance changes over time. However, the covariance normally converges rapidly (see Figure 5.4) so the part of the sequence after convergence can be considered in isolation.

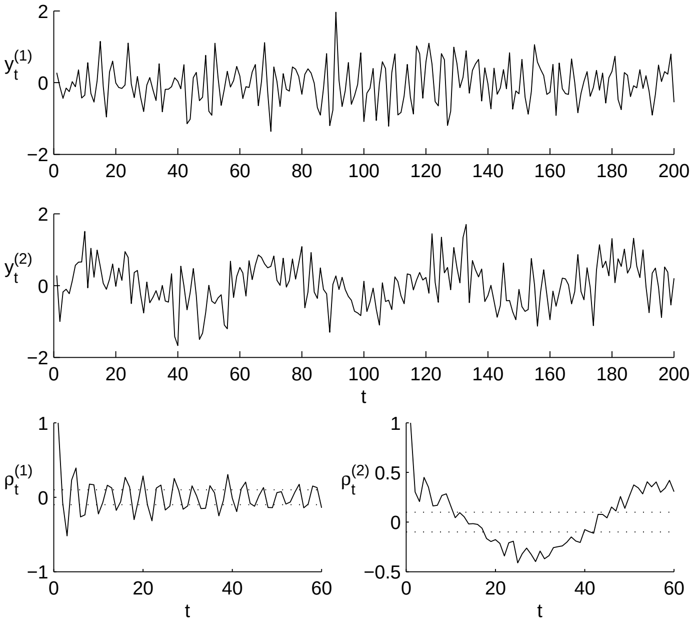


Figure 5.1: Sampled sequences from models $\mathcal{M}_1$ and $\mathcal{M}_2$. Top: Sample from model $\mathcal{M}_1$, a hidden AR(3) process with $\alpha = [.1, -.6, .3]$, and $\mathbf{Q}_{11} = \mathbf{R}_{11} = 0.1$. Middle: A sample from model $\mathcal{M}_2$, the same sequence with a sinusoidal component added such that $y_t^{(2)} = y_t^{(1)} + 0.5\sin(\frac{t}{20})$. Bottom: autocorrelations of the two samples.

The KFDD can also be used to plot the relevant components of the Kalman gain and prediction covariance, as shown in Figure 5.4. This gives an indication of the model's confidence under ideal conditions. In addition it is possible to generate further samples from the given model, to plot the model's predictions with error bars, to return the likelihood of the data given the model, the covariance of the innovations and the distance between the expected and actual innovation cumulative distribution functions (for Kolmogorov-Smirnov hypothesis testing). The software is publically available online with a demonstration at `http://www.homepages.inf.ed.ac.uk/s0348608`.

This software was used to verify the fit of models constructed for each regime in the application. The way that these regimes were trained is now presented in the remainder of this chapter.

5.5 Summary of regimes

Training data was obtained for a number of infants, with 24 hours of second-by-second data for each infant. Labels of which regimes were active at each time step were

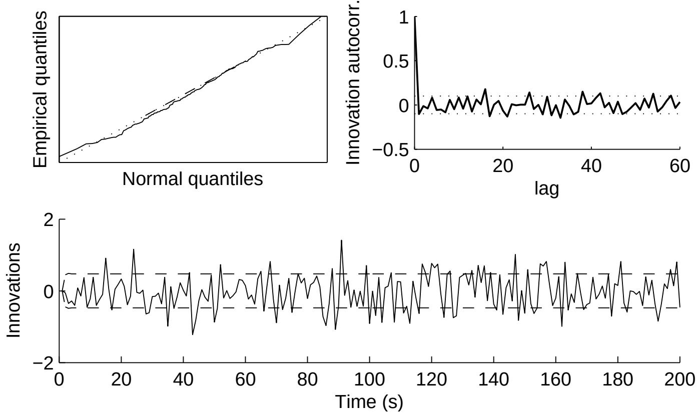


Figure 5.2: KFDD demonstrating a correctly fitted model. Top left: the empirical and expected cumulative distribution functions are similar. Top right: the innovations are not significantly correlated in time. Bottom: there does not appear to be structure in the innovations sequence.

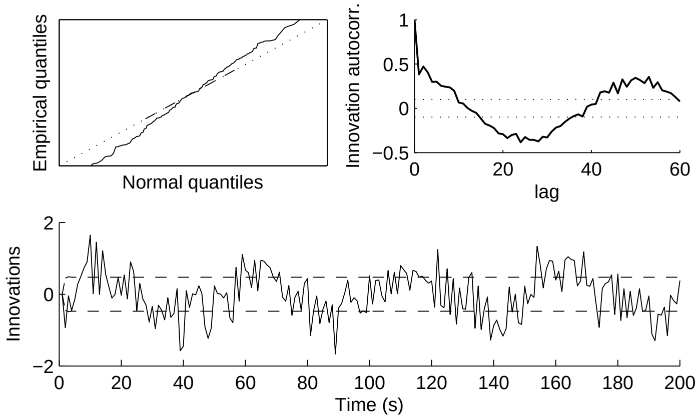


Figure 5.3: KFDD demonstrating a badly fitted model. A sequence from model $\mathcal{M}_2$ was used, containing a sinusoidal component, and model $\mathcal{M}_1$ was applied as a Kalman filter. The sinusoidal component can be seen in the innovation sequence (bottom). and this extra component causes the innovations to become correlated in time (top right). The distribution of the innovations is no longer normal with the expected variance (top left).

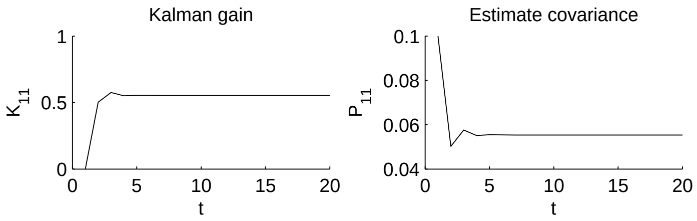


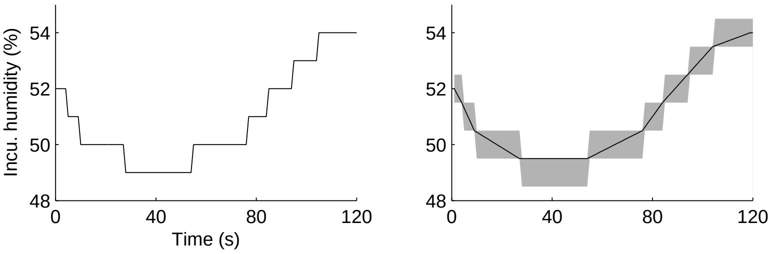
Figure 5.4: Components of the Kalman gain and estimate covariance for the model. K goes to 0.5, which is to be expected since Q and R are equal.

provided by clinical experts. Therefore each set of monitoring data was accompanied by a binary channel for each factor, indicating whether the factor was active at that time point. The annotation process is described in section 7.1.

It is important to set the parameters for the normal dynamical regime correctly. This is done in section 5.7. Different babies can have different baseline dynamics, and so this training is done on a per-baby basis. Note that we learn all the other dynamical regimes (artifactual patterns and so on) across all babies, making the assumption that the parameters of these non-normal dynamics generalise to every baby.

Sections 5.8.1 to 5.8.5 describe training models for the artifactual dynamics in sections 2.3.1 to 2.3.4: probe dropouts (the simplest pattern to deal with, where the readings on the affected channel go to zero); temperature probe disconnection (where core temperature measurements decay to an ambient level); blood sample artifact; recalibration of the transcutaneous probe, and opening of the incubator. Sections 5.9.1 and 5.9.2 describe setting parameters for two genuine physiological phenomena, bradycardia and blood pressure waves respectively. In each of these cases the methods used to set parameters are described. Where appropriate, each section describes non-standard situations in which the fitted models will not be as accurate and therefore represent sources of modelling error.

Section 5.10 deals with the task of combining these individual dynamical models into a single factorial model with all regimes. In particular, it describes how to form such a model without needing training examples from every possible combination of factors. Section 5.10.1 explains how these parameters are combined into a joint model, and section 5.10.3 shows the physical interpretation which can be applied to each dimension of the hidden state.



(a) (b)

Figure 5.5: Quantisation correction. (a) Original incubator humidity measurements, quantised to steps of 1%. (b) Shaded areas show the range of possible underlying values before quantisation, and the solid line shows points of low uncertainty connected to give a signal with less discontinuity.

5.6 Preprocessing

It is necessary to apply preprocessing to some of the data channels because of measurement quantisation. This is a problem in cases where the resolution of the data is low relative to the amount by which a reading might change over successive time steps. One observation channel for which this is a problem is incubator humidity, for which an example is shown in Figure 5.5(a). This discontinuous, coarse series of measurements is clearly a problem to process using a model with Markovian dynamics which considers only a limited number of past measurements. The same problem occurs with temperature measurements, both for the ambient incubator temperature and the core and peripheral physiological temperatures.

Consider the unquantised, underlying physical values which gave rise to the quantised observations. We want to find a way of reconstructing the underlying values from the measurements as far as possible. For a single time point, the quantised observation is the best we can do; it is the average of all possible underlying values before rounding off occurred. However, for a sequence of readings, it is possible to obtain a new sequence which is closer to the underlying values and less discontinuous. Because the underlying humidity and temperatures can only vary slowly, it is possible to know their values more accurately at the points where the quantised values change. If the underlying value is rounded up at time $t-\delta$ and rounded down at time $t+\delta$, then at time t it must have been close to the rounding threshold. This can be visualised with the aid of Figure 5.5(b), which shows the range of possible underlying values for the quantised signal in panel (a). In Figure 5.5(c) a linear interpolation is made between the transition points, to give a signal with less discontinuity.

Note that we are assuming that there is a certain amount of inertia in the system (e.g. thermal inertia in the case of temperature observations), such that there is unlikely to be any sudden discontinuity in the signal. For the moment we are also disregarding forms of measurement error other than quantisation.

There are some details which need to be observed in implementation. Where the signal is constant for a long period of time, this is regarded as the best estimate. In other words, there needs to be some cutoff in waiting for the next transition point. The method has implications for real time operation, as we rely on future knowledge of the signal to give the current. In practice, this is done by processing data after a fixed lag. See section 8.2 for real-time implementation issues. It is also necessary to treat zeros as special values. In all our physiological measurements, a zero has a special significance, indicating that measurements were not available. Therefore it is inappropriate to change these values.

An alternative preprocessing scheme might be to use a moving average to smooth dicontinuities. A large window would have to be used in cases such as that shown in Figure 5.5. A moving average filter with a large window could result in informative frequency components being removed from the signal.

More sophisticated methods of quantisation error would also be possible. For example, it is clear that the assumption of Gaussian observation noise is not valid when the data is highly quantised, and this might motivate the adoption of a more suitable noise model. Changing the noise model in the FSKF would however affect the practicality of inference, as for instance the Gaussian sum approximation would no longer be appropriate. The simple preprocessing technique described here is fit for purpose because it is fast and adequate for the task of determining whether the incubator is open or shut as in section 5.8.5.

5.7 Normal dynamics

The physiological systems underlying the observation channels are too complicated to model explicitly, being governed by complex interactions between a number of different sub-systems including the central nervous system. Instead, the approach adopted here is to try to find relatively simple models that are compelling according to the principles described in section 5.4.

Model selection is guided here by observations about examples of normal variation, such as those shown in Figure 2.2. A general observation about vital signs sequences is that they tend to contain dynamics at different time scales. Examples of ‘quick’ phenomena (of the order of seconds or minutes) might be bradycardia, or effects relating to handling. ‘Slow’ dynamics are related to changes in the baby’s baseline condition, which might take place over the course of hours or days. This could be caused by long term chronic conditions, or simply the ageing process; for example, a rule of thumb for mean blood pressure in a healthy infant is is that its value in kPa is normally similar to the gestational age in weeks. We are mainly concerned here with phenomena which happen on a rapid time scale.

The approach used for fitting dynamical models to each observation channel is first illustrated with heart rate observations, which are generally the least stable and most difficult to model of the observed channels. Section 5.7.2 then goes on to show how this approach is adapted to model the other observed channels. Conceptually, this approach is comprised of two parts. The first is to find a class of dynamical models which are suitable for each observation channel, for which the verification methods introduced in section 5.4 are used. Having obtained an appropriate model (e.g. with an appropriate order of autoregression) the second step is to learn the parameters for a particular baby given a training sequence. The implication of this for using the system in practice is that ‘normal’ training data must be supplied to the system by a clinician, by annotating a period of data in which a baby was stable. No verification need be done at the cotside though, as we consider such tests to have been completed in the first step.

5.7.1 Heart rate dynamics

Looking at examples such as those in Figure 2.2(a) and the top panels of Figure 5.6, it can be observed first of all that the signal tends to vary around a slowly drifting baseline. This motivates the use of a model with two hidden components: the signal x_t, and the baseline b_t. These components are therefore used to represent the true heart rate, without observation noise.

As an example, there might be a random walk baseline (with low system noise so that it varies slowly) and an AR(p) signal component as follows:

$$x_t - b_t \sim \mathcal{N}\left(\sum_{k=1}^{p} \alpha_k (x_{t-k} - b_{t-k}), \eta_1\right) , \tag{5.8}$$

$$b_t \sim \mathcal{N}(b_{t-1}, \eta_2) . \tag{5.9}$$

This describes the hidden, 'true' dynamics. It is temporarily assumed for the moment that the measurement noise $\mathbf{R}$ is known. There are two other noise terms, η_1 controlling the variance of signal component, and η_2 controling the variance of the baseline process. As a specific example of how this can be represented in state-space form, an AR(1) signal with a random walk baseline can be formulated as follows:

$$\mathbf{x}_t = \begin{bmatrix} x_t \\ b_t \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \alpha_1 & 1-\alpha_1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} \eta_1+\eta_2 & 0 \\ 0 & \eta_2 \end{bmatrix}, \quad \mathbf{C} = [1\ 0]\ . \tag{5.10}$$

This AR(1) model with a random walk baseline was used by Gordon and Smith (1990) to model creatinine levels in adults. Higher AR orders are generalised in a similar form to (4.23). The signal can be trained as an ARIMA process and expressed in AR form using (4.8). The model can be further generalised so that an AR(p_1) signal varies around an AR(p_2) baseline, as given by the following equations:

$$x_t - b_t \sim \mathcal{N}\left(\sum_{k=1}^{p_1} \alpha_k (x_{t-k} - b_{t-k}), \eta_1\right)\ , \tag{5.11}$$

$$b_t \sim \mathcal{N}\left(\sum_{k=1}^{p_2} \beta_k b_{t-k}, \eta_2\right)\ . \tag{5.12}$$

For example, an AR(2) signal with AR(2) baseline has the following state-space representation:

$$\mathbf{x}_t = \begin{bmatrix} x_t \\ x_{t-1} \\ b_t \\ b_{t-1} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \alpha_1 & \alpha_2 & 1-\alpha_1 & -\alpha_2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & \beta_1 & \beta_2 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \tag{5.13}$$

$$\mathbf{Q} = \begin{bmatrix} \eta_1+\eta_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \eta_2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{C} = [1\ 0\ 0\ 0]\ . \tag{5.14}$$

The measurements are therefore generally taken to be made up of a baseline with low frequency components and a signal with high frequency components. Training the heart rate model proceeds by taking sequences of training data and removing high frequency components by applying a symmetric 300-point moving average filter. The resulting signal is taken to be the low frequency baseline. The residual between the original sequences and the moving-averaged sequences are taken to contain both stationary high frequency hemodynamics as well as measurement noise. These two signals can be analysed according to standard methods and modelled as AR or differenced

AR processes of arbitrary order. Heart rate sequences were found to be well modelled by an AR(2) signal varying around an ARIMA(1,1,0) baseline. An ARIMA model is a compelling choice for the baseline, because with a low noise term it produces a smooth drift. The baseline model is expressed in un-differenced form as a non-stationary AR(2) model, and the two AR coefficients are scaled down by multiplying by $(1-10^{-3})$. This slight damping makes the baseline mean-reverting, so that the resulting signal is stationary. This has desirable convergence properties for dropout modelling in section 5.8.1. Having found this initial setting of the model parameters, EM updates are then applied (as given in Appendix C). This has been found to be particularly useful for refining the estimates of the noise terms $\mathbf{Q}$ and $\mathbf{R}$.

Examples of the heart rate model being applied as a Kalman filter to heart rate sequences are shown in Figure 5.6. The top panels show sequences of noisy heart rate observations, and the lower panel shows estimates of the 'true' high frequency and low frequency components of the heart rate. Figure 5.7 shows verification tests from the KFDD. The autocorrelation plot in Figure 5.7 shows that the innovations are not completely white, with significant autocorrelation up to lag 10, though in practice this level of model fit is adequate for distinguishing normality from factors such as bradycardia (as demonstrated by the experiments in Chapter 7). In order to establish how much of the structure of the data is being represented by the model, it can be useful to compare the autocorrelation of the original data (shown in Figure 5.8) with the autocorrelation of the innovations (top right panel in Figure 5.7). In this case it can be seen that the innovations are clearly 'whiter' than the data, which gives us some confidence in the model.

5.7.2 Other channels

The remaining observation channels, with the exception of peripheral temperature, are modelled according to the same principle. The best specific model for the signal and baselines are summarised in Table 5.1. Some of the remaining channels such as core and peripheral temperature are more stable and are well modelled by processes of a lower order.

As well as estimates of underlying physiology, the hidden state $\mathbf{x}_t$ contains estimates of the levels associated with different artifactual processes. These are artifactual core temperature (section 5.8.2), artifactual blood pressures (section 5.8.3), and artifactual transcutaneous channels (section 5.8.4). During normal dynamics these estimates need to be set to appropriate levels even though they are not observed. The artifactual

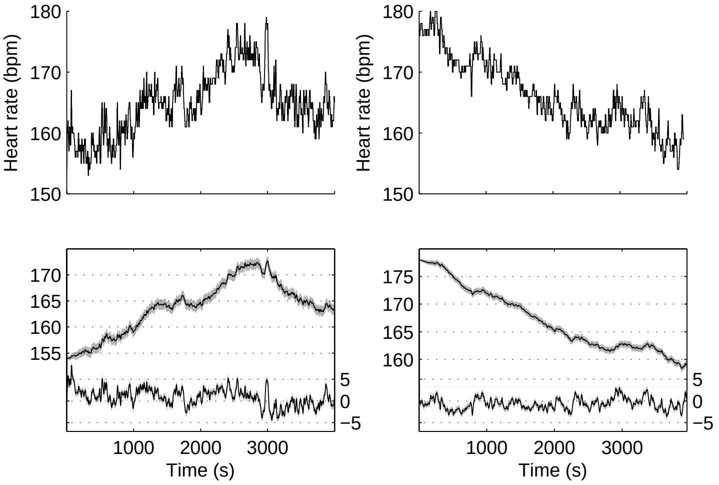


Figure 5.6: In these two examples, HR measurements (in the top left and top right panels) are varying quickly within normal ranges. The estimates of the underlying signal (bottom left and bottom right panels) are split into a smooth baseline process and zero-mean high frequency component. The slight gray shadow around the estimates indicates two standard deviations.

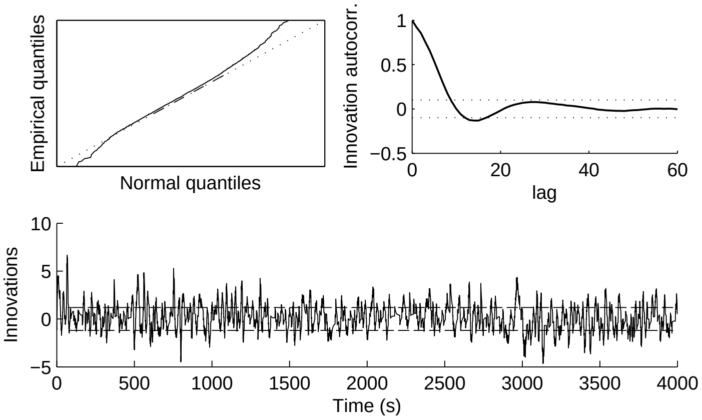


Figure 5.7: Verification statistics for the heart rate model. According to these, particularly to the autocorrelation in the top right panel, there is still some correlated signal left in the innovations, but within acceptable limits.

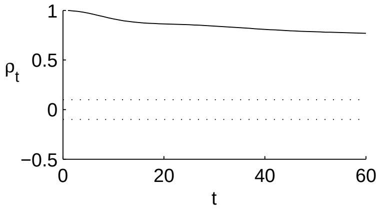


Figure 5.8: Autocorrelation of the original heart rate data. Comparing this with the top right panel in Figure 5.7 gives an indication that much of the structure in the data has been included by the model.

OBSERVATION CHANNEL	MODEL	
	Signal	Baseline
Heart rate	AR(2)	ARIMA(1,1,0)
Systolic BP	AR(2)	ARIMA(1,1,0)
Diastolic BP	AR(2)	ARIMA(1,1,0)
Core temperature	AR(1)	AR(1)
Peripheral temperature	AR(1)	AR(1)
$TcPO_2$	AR(2)	AR(1)
$TcPCO_2$	AR(2)	AR(1)
SpO_2	AR(1)	
Incubator temp.	AR(1)	AR(1)
Incubator humidity	AR(1)	

Table 5.1: Models for each observation component.

patterns all start at physiological levels, so in the normal dynamical model the system matrix $\mathbf{A}$ simply sets the estimate of the artifactual channels to be the same as the estimate of the corresponding true physiological channel. Specifically, during normal dynamics, artifactual blood pressure dimensions in the state are tied to the most recent diastolic blood pressure estimate. During a blood sample, the physiological state is not observed and the two sets of variables evolve independently. Artifactual core temperature dimensions are tied to the most recent physiological core temperature estimate, but evolve independently during a core temperature probe disconnection. Similarly, the artifactual transcutaneous state dimensions are tied to physiological $TcPO_2$ and $TcPCO_2$ during normal dynamics, but can otherwise vary independently according to the dynamics in the different stages of a transcutaneous probe recalibration. At any one time, only one variable is observed for each observation channel—a physiological variable during normal dynamics or an artifactual variable when the corresponding artifactual factor is active.

5.8 Artifactual dynamics

Recall from the description in section 4.7.1 that the hidden state in the FSKF can be constructed so that it has dimensions which fall into two categories. There are quantites representing the true physiology or environment, and quantities representing the levels of different artifactual processes. The regimes described in this section (apart from the ‘incubator open/handling’ regime, section 5.8.5) cannot affect the true physiology, only the estimates of artifactual quantities and the way that the state is projected into the observation space.

5.8.1 Drop-outs

Probe dropouts, which cause the observations to go to zero, are simple to model in this framework by taking normal dynamics and changing the appropriate entry in the observation matrix to zero. This indicates that the relevant underlying physiology is entirely unobserved. In this way, the estimates of the underlying physiology are unaffected. Normal dynamics continue to update the estimates of the true physiology, but without being updated by the observations. The Kalman gain is always zero in (4.28) so that the new observations have no weight upon the estimates. Uncertainty therefore increases, as we would expect, until $\mathbf{P} = \mathbf{Q}(\mathbf{I} - \mathbf{A}^{\top}\mathbf{A})^{-1}$, the stable state for (4.25).

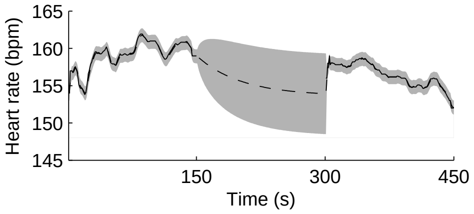


Figure 5.9: Estimates of true heart rate. Readings drop out between times 150 and 450. The solid line indicates a heart rate measurement. The dashed line indicates the mean estimate of the true heart rate, with the shaded area shows two standard deviations of the estimated distribution.

The operation of the filter under these conditions can be assessed with the KFDD. Figure 5.9 shows a section of heart rate data, and the estimates as to the true heart rate. The heart rate drops out in the middle of the sequence. A Kalman filter is used to infer the underlying heart rate, with the appropriate dynamics at each time point. First the normal heart rate regime is used. When the dropout begins, the observation matrix is set to zero and Kalman filter inference is resumed. When the dropout ends, the original observation matrix is reinstated. The dashed line in the Figure indicates the estimated heart rate, and the gray area shows two standard deviations of the estimate. During the dropout the estimate decays to the mean, and the variance of the estimate increases appropriately.

Note that the estimates will not always converge using this technique. If the dynamics are non-stationary then the covariance of the estimate $\mathbf{P}$ can increase without bound. For example, random walk dynamics where the system matrix $\mathbf{A} = \mathbf{I}$ will not converge.

5.8.2 Temperature probe disconnection

The artifactual measurements caused by the disconnection of a core temperature probe reflect the transition of the probe from thermal equilibrium with the baby's body to equilibrium with the air in the incubator. According to Newton's laws of cooling, a cool or hot body with some thermal inertia has a temperature decay to the ambient level which is exponential.

An exponential decay appears as a straight line when plotted on a logarithmic ordinate scaling. This can be seen in Figure 5.10, for both the disconnection and reconnection stages. The decay rate should be the same for each disconnection, since the same type of probe is used which therefore has the same thermal inertia. This gives a way of telling whether the probe is cooling according to Newton's laws of cooling or whether the baby is getting colder, for which there is no reason to assume the same type of dynamics.

An exponential decay model (given by an AR(1) process) is fitted using the Yule-Walker equations. The method described in section 5.3.1.2 is used to estimate the observation noise.

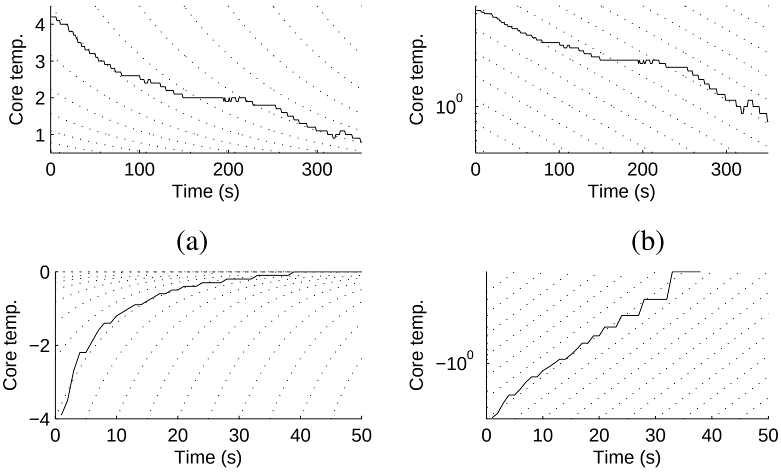
The level to which the core temperature measurements decay to should be the same as the incubator's ambient temperature, which is also recorded. Sometimes this is not the case however, for a number of reasons. The level may be higher due to the core temperature probe being removed and placed near the hot air pump which creates a heated stream of air around the edge of the incubator. The mean level might be higher if the probe becomes swaddled in linen, as the insulating effect will reduce the heat flux between the probe and the air in the incubator. It may also be lower due to the probe being placed near to a portal or removed from the incubator completely, for example in order to pass it back in through a different portal. These special cases are not modelled explicitly and are therefore potential sources of modelling error.

5.8.3 Blood sampling

Artifactual measurements obtained during the collection of an arterial blood sample can be modelled structurally by noting some properties of the process. First, arterial blood pressure is picked up by the measuring equipment as a high-frequency waveform, and the second by second systolic and blood pressure measurements are the points on this wave when the heart is contracted and relaxed. During the blood sample procedure, the pressure transducer is completely isolated from the heart. There is therefore no wave and the systolic and diastolic measurements should be equal.

At the start of the procedure, the pressure acting on the transducer is the same as the baby's systolic blood pressure. As explained in section 2.3.2, this pressure rises due to the action of a saline pump against the closed section of arterial line. This suggests the use of a linear drift term in the model. There is a complication, however, that different amounts of saline are pumped through different lines, so the increase of pressure per time step is not always constant. When the pressure becomes very high, the saline pumps also automatically 'back off' a little, so that the pressure gradient is not constant even during a single blood sample operation.

These uncertainties are straightforward to model in state-space form as a structural



(a)

(b)

(c)

(d)

Figure 5.10: Core temperature measurements. Panels (a) and (b) show a period in which the temperature probe has become disconnected, and panels (c) and (d) show the period after it has been reconnected. Panels (b) and (d) show log plots. Dotted lines show deterministic trajectories for the corresponding fitted AR(1) process. Solid lines show measurements, corrected to have a baseline of zero. During the probe disconnection, there are other dynamics approximately between times 80 and 240, possibly caused by the temperature probe becoming wrapped in linen or being placed near a heating vent.

model. The average gradient can be learnt for all blood samples, and used as a constant linear drift term. A state-space is then formulated which has a random walk on the differences of the data with a small noise term. In this way, the average drift is used as an initial guess, and the differenced random walk term can alter this guess to converge with the data. The small magnitude of the noise term has a smoothing effect on the estimates, so that the hidden state estimation is not affected by jumps in the data caused by measurement error.

Figure 5.11 shows inferences for the underlying pressure in the line caused by the pump, given two sequences of noisy observations on the systolic blood pressure observation channel. During these two example blood sample sequences, different amounts of saline were being pumped down the arterial line, and the Kalman filter estimate of the drift can be seen to converge to the correct value in each case. Having constructed this to work with a single pressure measurement, the observation matrix can then be adjusted to include the diastolic measurement channel as well, generated

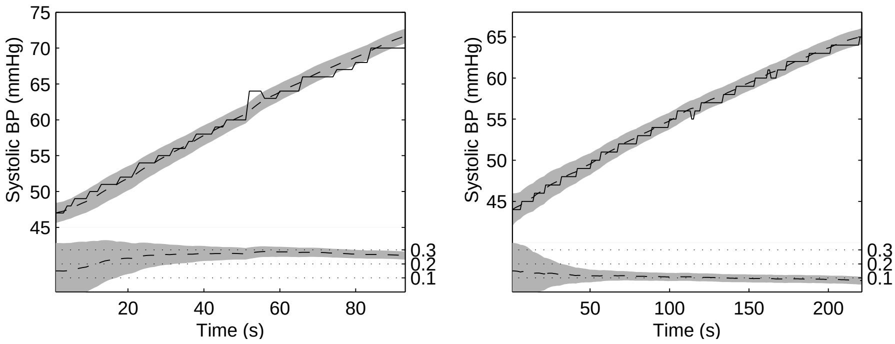


Figure 5.11: Operation of the blood sample model on two example sequences of systolic blood pressure measurements. The top plots show the observations (solid lines), and the estimates of the level (dashed) with two standard deviations. The lower plots show the estimated drifts, with two standard deviations. In each case, the underlying drifts are different, and the estimated drift converges to the correct level. The disruption in measurements in the example on the left at around $t = 55$ is noise from the measuring equipment.

from the same underlying state dimension.

5.8.4 Transcutaneous probe recalibration

A recalibration of the transcutaneous probe is often preceded by a dropout of the TcPO_2 and TcPCO_2 channels, as the probes are subject to drift and automatically shut off after a certain period of time without calibration. The different stages of the calibration process are shown in Figure 5.12. These correspond to the application of a calibration solution to the probe, then the removal of this solution so that the probe gives a reading in room air, then the reapplication of the probe to the baby. After this final step, the levels of the measurements decay to the true physiological levels. Note that there is no single point at which the recalibration can be said to have ended, which means that the discrete switch setting is not strictly appropriate in this case.

The constant levels of the first stage do not require any dynamics to model — just a mean and a variance. The other two stages are modelled as exponential decays. Table 5.2 summarised the modelling of each stage. Unlike the other factors in this chapter, the recalibration has multiple consecutive stages. The transition probabilities between each stage are estimated as usual with (5.6), which sets the expected dwell time in each stage, and the expected sequence of stages.

Period	Recalibration stage	Model type
A→B	Dropout	Model with dropout factor
B→C	Calibration	Constant levels
C→D	Room air	Constant O_2, exponential decay for CO_2
D→end	Probes re-applied	Exponential decay for both channels

Table 5.2: Stages in a transcutaneous recalibration process, and the way in which the dynamics of each stage are modelled.

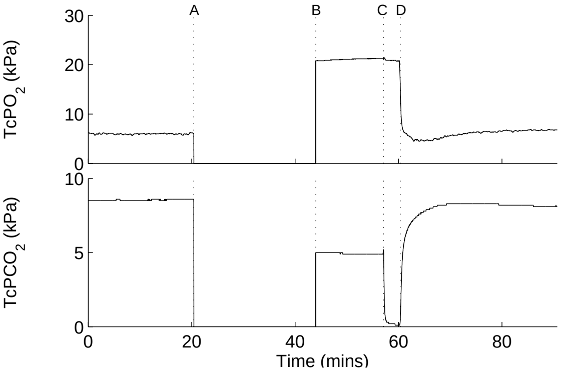


Figure 5.12: Transcutaneous probe recalibration. At point A, readings from the probe automatically stop. At B, a calibration solution is applied to the probe to give reference levels. At C, the probe is exposed to room air, and at D is reapplied to the baby's chest.

5.8.5 Opening of the incubator

Incubator humidity and temperature are closely regulated, so that with all incubator portals shut the ambient humidity and temperature readings normally have low variance. As described in section 2.3.4, when a portal is opened there is a significant drop in these readings. These drops can be modelled as an AR(1) decay, where the level to which these measurements drop is unknown but cannot be lower than the humidity and temperature of the room. These ambient properties of the room are not measured directly, but are subject to guidelines and can therefore be roughly estimated. According to White (2006), the acceptable temperature and humidity ranges for a NICU are 22-26°C and 30-60% respectively. The mean levels of the corresponding ambient measurements for the 'incubator open' process are therefore set to 24°C and 45%. Given this, the remaining hidden AR(1) parameters are obtained according to the technique described in section 5.3.1.2.

Note that while these are reasonable settings to use in general, there are some circumstances in which this model may not be accurate. For example, a clinician or parent may have two arms through the side portals and block enough of the opening that the incubator is able to maintain some heating and humidification.

The opening of the incubator implies that an intervention to the baby is taking place. This can be expected to have some kind of physiological effect, normally an increase of variance on the cardiovascular channels and a slight decrease in peripheral temperature due to the influx of room air in the incubator. Parameters can then be set by repeating the process for training normal dynamics on data which was obtained during handling episodes. In practice, this tends to result in physiological dynamics are similar to the normal dynamics but with a larger system noise term. In Chapter 6 this effect is used to model more general unknown dynamics.

5.9 Dynamics of known physiological events

This section deals with training dynamical models for genuine physiological changes. These models affect the estimate of the true physiological state rather than emitting artifactual measurements, and therefore 'overwrite' the normal dynamics on certain channels. Section 5.9.1 describes training a model for the dynamics relating to periods of bradycardia. We then look at training a model to represent blood pressure waves, a much more rare phenomenon.

5.9.1 Bradycardia

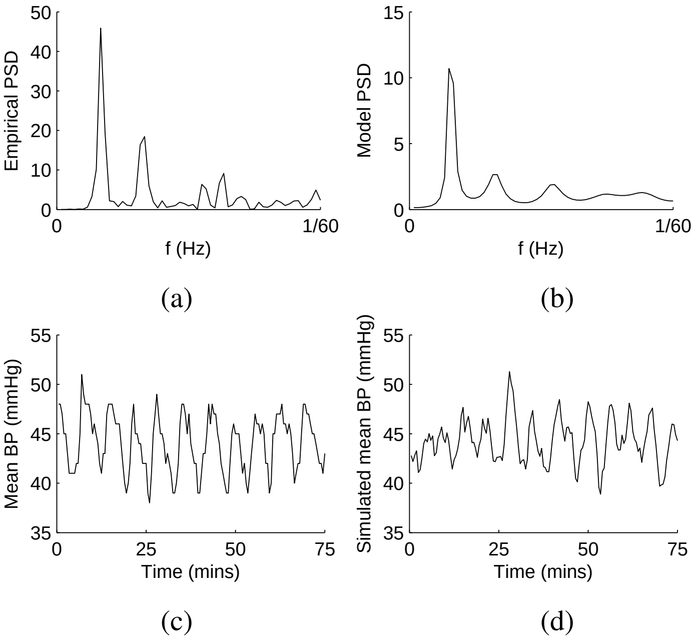
Bradycardic drops and subsequent rises in heart rate as shown in Figure 2.4.1 were found to be adequately modelled by retraining the ARIMA(1,1,0) model for baseline heart rate dynamics. The high frequency heart rate dynamics were kept the same as for the stable heart rate regime. As for the normal regime, this model learnt in terms of hidden ARIMA processes was used as an initial setting and updated with three iterations of EM.

The working clinical definition of bradycardia in neonates is a drop in HR below 100bpm. It may therefore be wondered why this factor is not modelled simply as a constant level. However, a constant level model would not be specific enough. Problems would arise firstly in the case of ECG lead noise. When the ECG pads are moved, there is occasionally high spurious variance on the HR channel. Although this is not modelled explicitly, the bradycardia factor should ideally have a low likelihood for such sequences of observations. Secondly, the heart rate can be volatile during handling, and may drop to low levels but without the characteristic smooth drop which occurs with spontaneous bradycardia. Cardiovascular instability caused by handling should ideally be picked up by the ‘incubator open/handling’ factor.

5.9.2 Blood pressure waves

Blood pressure waves are a rare occurrence, but have a distinctive pattern that is natural to model in state-space form. There were no examples of blood pressure waves in the main body of training data, but example data was taken separately from a different baby for purposes of illustration. Observing the examples of this phenomenon in Figure 2.10 it is firstly clear that the pattern occurs on a slower time scale than the frequency of the observations (one wave every few minutes, compared to the data points taken each second). Any raw training data will therefore contain a lot of high frequency information which is likely to be of low relevance. For the following example, instances of blood pressure waves were taken from one baby and downsampled by a factor of 30.

Figure 5.13 illustrates how the model is constructed. Because the dynamics are highly periodic, it makes sense to view the problem in the frequency domain. The data is first differenced to improve stationarity, and the power spectrum of a sequence of downsampled, differenced training data is shown in panel (a). The main frequency components can be seen, the largest at approximately $\frac{1}{360}$Hz, corresponding to a wave-



(a) (b)

(c) (d)

Figure 5.13: Spectra of differenced blood pressure waves. (a) Spectrum of downsampled training data. (b) Spectrum of ARIMA(10,1,0) model. In panel (c), genuine blood pressure waves similar to those in Figure 2.10 are shown, and panel (d) contains a sample from the model, which can be seen to have similar frequency components.

length of around six minutes. An AR(10) model was trained on the differenced training data, and the spectrum of the model, given by (4.3) is shown in panel (b). The operation of the model can be checked by comparing a section of the training data (in panel (c)) with a sample from the model (in panel (d)). Because we are comparing the undifferenced observations, the sample sequence was generated from the ARIMA(10,1,0) process given by (4.7). Note that because we are dealing here with downsampled data, the dynamics of this regime can be modelled without a separate high-frequency signal component. Recall that such a component was needed for modelling the stable blood pressure dynamics on a 1-second time scale in section 5.7.2.

Appendix D shows how a dynamical model which has been trained at a low time resolution can be applied to the original, high frequency data. See also section 8.2 for a discussion of real-time inference and the way in which this pattern might be classified in a practical setting. Unlike the dynamics described previously in this chapter, the blood pressure wave model is not formally evaluated in Chapter 7, for the reasons given in section 7.1.

	Heart rate	Sys BP	Dia BP	$TcPO_2$	$TcPCO_2$	SpO_2	Core temp.	Periph. temp.
Dropouts	■	■	■	■	■	■	■	■
Blood sample		■	■					
Core temp. probe disconnection							■	
TCP recalibration				■	■			
Handling	■	■	■			■		
Bradycardia	■							
Normal	■	■	■	■	■	■	■	■

Table 5.3: Partial ordering of factors overwriting physiological channels. Factors higher on the list overwrite the channels on lower factors, so for example a dropout (which can occur on any channel) overwrites that channel, regardless of which other factors are active.

5.10 Constructing a factorial model

Having considered the training of each regime independently of the others, this section describes how these models can be combined together to give an overall factorial model for the data.

5.10.1 Setting parameters in factorised models

Recall from section 2.5 that some factors ‘overwrite’ each other. For example, if a bradycardia takes place at the same time as an ECG probe disconnection, the same measurements are obtained as though there was just the disconnection (a sequence of zeros). Every factor overwrites the measurements given by the normal dynamics. A partial ordering of which factors overwrite which others is given in Table 5.3.

The significance of this is that examples of every combination of factors do not need to be found in order to train our factorial model. The factors can be trained independently, and then combined together by reasoning about which channels are overwritten for each combination.

5.10.2 Inference of drop-outs in the FSKF

It is computationally expensive to add extra factors to the FSKF, each extra factor causing the time required for inference to increase exponentially. The addition of five explicit binary dropout factors (one for each probe or data source) would therefore slow down inference by a factor of 2^5. It is also clear that calculating whether a dropout has occurred or not is a very simple operation. In this case the FSKF inference routine can be adapted, regardless of whether particle filtering or an analytical approximation is being used. If an observation is zero, we can set the probability of a dropout on that channel to unity, and zero otherwise. The only change required for inference is to insert a zero in the appropriate place in the observation matrix. It is natural to deal with dropouts in such a way since a zero is a special value encoding the fact that data is unavailable. Its actual numerical value is not significant.

5.10.3 Organisation of hidden state dimensions

Each dimension in the hidden state $\mathbf{x}_t$ has an interpretation as either an estimate of an underlying aspect of the baby's physiology, or an estimate of the level of an artifactual process. As described in section 5.8, artifactual estimates can be affected by the true physiology, but not the other way around. When the FSKF is trained as described in the preceding sections, the interpretation of the hidden state dimensions is as given in Table 5.4.

5.11 Summary

This chapter has shown in general how parameters for dynamical models for condition monitoring can be learnt with labelled training data, and specifically how knowledge about the underlying processes for this application can be used. It has also shown how these models can be verified as being a good fit to the data.

After training, each dimension of the hidden state in the FSKF has a physical interpretation, as listed in Table 5.4. The full model has a hidden state dimensionality of 33, and 48 possible switch settings (the cross product of all factor settings, excluding dropouts). However, it can also be trained with a subset of factors or observation channels if required.

Only common, known regimes have been dealt with in this chapter. As discussed in section 2.5, monitoring data may exhibit many other patterns with uncertain causes. Handling of other types of variation is discussed in Chapter 6.

PHYSIOLOGICAL / ENVIRONMENTAL		ARTIFACTUAL	
1-4	True heart rate		
5-8	True systolic BP	27-28	Artifactual blood pressure
9-12	True diastolic BP	29	Artifactual core temperature
13-14	True core temperature	30	Artifactual TcPO$_2$
15-16	True peripheral temperature	31	Artifactual TcPCO$_2$
17-19	True TcPO$_2$		
20-22	True TcPCO$_2$		
23	True SpO$_2$		
24-25	True incu. temp.		
26	True incu. humidity		

Table 5.4: Interpretation of each dimension of the hidden state $\mathbf{x}_t$. Artifactual state dimensions $\{$23,24$\}$ are affected by blood sample dynamics, 24 by core temperature probe disconection, and $\{$25, 26$\}$ by transcutaneous probe recalibration.

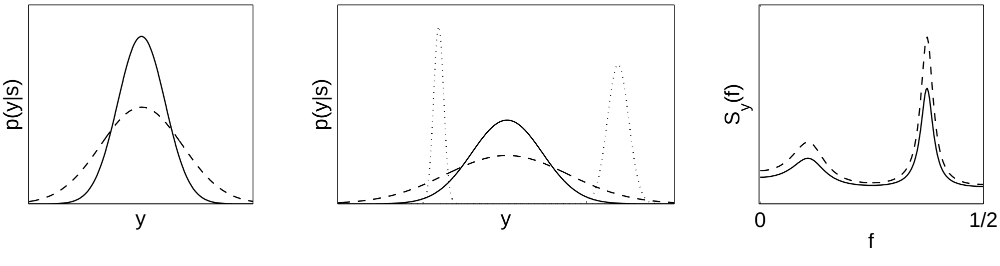
Chapter 6

Novel dynamics

So far we have assumed that physiological monitoring data contains a limited number of regimes, corresponding to different known artifactual or physiological conditions. As described in section 2.4.4 on page 24, however, in practice the data exhibits many unusual patterns. In fact, the number of potential unusual patterns is so great that it would be impractical to explicity include every possibility in a model. Examples include rare dynamical regimes caused by sepsis, neurological problems, or the administration of drugs, even a change of linen or the flash of a camera. Considering these possibilities, it can be useful to include a factor in the condition monitoring model which represents all unusual cases.

This chapter presents a method for modelling previously unseen dynamics as an extra factor in the FSKF, referred to as the ‘X-factor’. This represents all dynamics which are not normal and which also do not correspond to any of the known regimes. A sequence of data can only be said to have novelty relative to some reference, so the model is learnt taking into account the parameters of the normal regime. The inclusion of this factor in the model has two potential benefits. First, it is useful to know when novel regimes are encountered, e.g. in order to raise an alarm. Second, the X-factor provides a measure of confidence for the system. That is, when a regime is confidently classified as ‘none of the above’, we know that there is some structure in the data which is lacking in the model.

Section 6.1 introduces the X-factor model, and section 6.2 describes how to estimate its parameters. This involves both finding an initial parameter setting (section 6.2.1) and also the use of expectation maximisation to update the parameter setting in the light of new – possibly unlabelled – data (section 6.2.2). For the sake of clarity in each of these sections, a simpler case involving static data is described, which is



(a) (b) (c)

Figure 6.1: (a) Class conditional likelihoods in a static 1D model, for the normal class (solid) and the X-factor (dashed). (b) Likelihoods of the normal class and X-factor in conjunction with other known, abnormal regimes (shown dotted). (c) The power spectral density of a latent AR(5) process with white observation noise (solid), and that of a corresponding X-factor process (dashed).

then generalised to switching linear dynamics. Section 6.2.3 provides a demonstration of the way in which X-factor parameters are learnt as part of the FSKF, first for synthetic data and then for genuine physiological data containing both novel dynamics and known patterns. More extensive evaluation is performed in chapter 7. Alternative novelty detection schemes, and their links to the X-factor method, are discussed in section 6.3. Section 6.4 discusses the links with other work in novelty detection, and Section 6.5 summarises. Parts of this chapter were adapted from Quinn and Williams (2007).

6.1 The X-Factor

To begin with, imagine that we have independent, one-dimensional observations which normally follow a Gaussian distribution. If we expect that there will also occasionally be spurious observations which come from a different distribution, then a natural way to model them is by using a wide, flat Gaussian with the same mean. Observations close to the mean retain a high likelihood under the original Gaussian distribution, while outliers are claimed by the new model.

The same principle can be applied when there are a number of known distributions, so that the model is conditionally Gaussian, $\mathbf{y}|s \sim \mathcal{N}\left(\mu^{(s)}, \Sigma^{(s)}\right)$. For condition monitoring we are interested in problems where we assume that the possible settings of s represent a 'normal' mode and a number of known additional modes. We assume here that the normal regime is indexed by $s = 1$, and the additional known modes by

$s = 2, \ldots, K$. In this static case, we can construct a new model for unexpected data points by inflating the covariance of the normal mode, so that

$$\Sigma^{(*)} = \xi \Sigma^{(1)}, \qquad \mu^{(*)} = \mu^{(1)} \, , \tag{6.1}$$

where normally $\xi > 1$. This type of construction for unexpected observations is referred to as an 'X-factor'. As the name suggests, the intention is to apply it later to a model in which s is factorised, the FSKF. The parameter ξ determines how far outside the normal range new data points have to fall before they are considered 'not normal'[1].

The likelihood functions for a normal class and a corresponding X-factor are shown in Figure 6.1(a). Clearly, data points that are far away from the normal range are more likely to be classified as belonging to the X-factor. For condition monitoring this can be used in conjunction with a number of known classes, as shown in 6.1(b). Here, the X-factor has the highest likelihood for regions which are far away from any known modes, as well as far away from normality.

We can generalise this approach to dynamic novelty detection by adding a new factor to a trained FSKF (section 4.7), by inflating the system noise covariance of the normal dynamics

$$\begin{aligned} \mathbf{Q}^{(*)} &= \xi \mathbf{Q}^{(1)} \, , && (6.2) \\ \left\{ \mathbf{A}^{(*)}, \mathbf{C}^{(*)}, \mathbf{R}^{(*)}, \mathbf{d}^{(*)} \right\} &= \left\{ \mathbf{A}^{(1)}, \mathbf{C}^{(1)}, \mathbf{R}^{(1)}, \mathbf{d}^{(1)} \right\} . && (6.3) \end{aligned}$$

This model for changes in dynamics is mentioned by West and Harrison (1999). To see why (6.2) and (6.3) are a dynamic generalisation of (6.1), first consider the specific case of a hidden scalar AR(p) process,

$$x_t \sim \mathcal{N}\left(\sum_{k=1}^{p} \alpha_k x_{t-k}, \sigma_q^2 \right), \qquad y_t \sim \mathcal{N}(x_t, \sigma_r^2) \, . \tag{6.4}$$

The power spectral density for the hidden process x_t at frequency f is given by

$$S_x(f) = \frac{\sigma_q^2}{\left| 1 - \sum_{k=1}^{p} \alpha_k e^{-2\pi i f k} \right|^2} \, , \tag{6.5}$$

where $-\frac{1}{2} \leq f \leq -\frac{1}{2}$, assuming one observed value per unit of time. By inflating σ_q^2 (as specified in (6.2)) we observe that the power is increased at each frequency. The observed process has the spectrum $S_y(f) = S_x(f) + \sigma_r^2$. As the scale of $S_y(f)$ is determined by the magnitudes of the two noise variances, inflating σ_q^2 will have the effect of increasing the power at every frequency, as illustrated in Figure 6.1(c).

[1] The notation ξ is chosen by association with the Greek word ξενος, or *xenos*, meaning 'strange'.

In the static case, the distributions of the normal regime and the X-factor have the same eigenvectors, with the eigenvalues being increased in the X-factor model by a factor of ξ. Under an AR(p) model driven by Gaussian noise, any sequence of x's (and also the y's) are jointly Gaussian. The eigenfunctions are sinusoids and the eigenvalues are given by the power spectrum. Hence inflating the system noise has created a dynamical analogue of the static construction given above.

Note that the nature of the measurement noise, and hence the value of the parameter $\mathbf{R}^{(s)}$, is assumed to be the same for both the normal regime and for the X-factor. Care needs to be taken that the known factor dynamics do not have a very high variance compared to the normal dynamics. It is clear from Figure 6.1(b) that the X-factor will not be effective if any of the factors are wider than normality. This can be ascertained by examining the spectra of the different model dynamics.

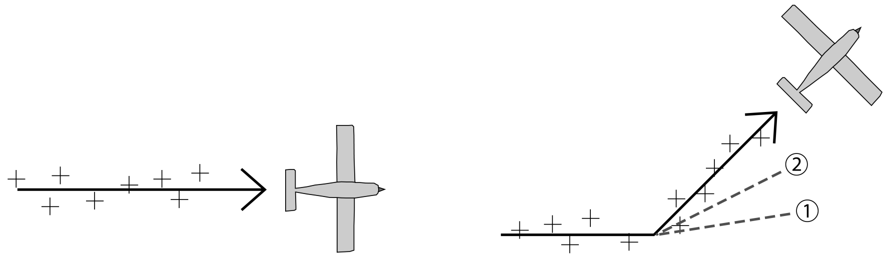
6.1.1 Inference under a novel regime

The operation of the dynamical model given by (6.2) and (6.3) can also be understood by considering how the estimates of $\mathbf{x}_t$ are updated in the light of new observations, as compared to the normal model. Intuitively, the parameters $\mathbf{Q}$ and $\mathbf{R}$ in a linear dynamical system represent prior strengths of belief in the model and the data respectively. If $\mathbf{Q}$ is strong (e.g. has a low determinant) compared to $\mathbf{R}$, then estimates will not change significantly given outlying observations. If $\mathbf{R}$ is strong relative to $\mathbf{Q}$, then in general previous estimates will be ignored, and the new data will be believed. The X-factor puts a weaker prior on the model, and therefore represents the common-sense idea that if we are unsure what is happening then we need to take new observations more seriously; we can no longer be so confident as to whether changes in a sequence are due to the underlying dynamics or to noise. This is true of many situations in which dynamics change unexpectedly, and the concept is illustrated in Figure 6.2 briefly employing a different, intuitive setting — the problem of tracking the position of a plane.

This concept is straightforward to express in terms of the Kalman filter equations. For a single linear dynamical state space, given by $\{\mathbf{A},\mathbf{Q},\mathbf{C},\mathbf{R},\mathbf{d}\}$, the estimate $\hat{\mathbf{x}}_t$ and the covariance of the estimate $\mathbf{P}_t$ are given by the Kalman filtering equations. The new state is predicted as follows:

$$\begin{aligned} \hat{\mathbf{x}}_t^- &= \mathbf{A}\hat{\mathbf{x}}_{t-1} + \mathbf{d} \, , && (6.6) \\ \mathbf{P}_t^- &= \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}^\top + \mathbf{Q} \, , && (6.7) \end{aligned}$$

where $\hat{\mathbf{x}}_t^-$ and $\mathbf{P}_t^-$ denote the estimates at time t made without taking into account the



(a) (b)

Figure 6.2: The effects of making inferences with the X-factor, illustrated with the problem of tracking the position of a plane from noisy radar measurements. The black arrow represents the true previous positions of the plane, and the crosses represent radar measurements. (a) While the plane is following a steady ('normal') course, a Kalman filter with a strong prior on the dynamics can be used to make optimal estimates of the position. (b) If the plane makes an unexpected diversion, the estimates according to the normal regime (the dashed line labelled '1') are suboptimal, as a strong prior on the model prevents the outlying observations from having much effect. If we detect that there has been some change in dynamics and make inferences with an X-factor — weakening the model prior — the estimates (labelled '2') are affected more by the outlying observations, and are therefore preferable. The same reasoning applies directly to unexpected changes in physiological dynamics.

observation $\mathbf{y}_t$. The most recent observation is then used to refine these estimates.

Consider the inference of $\hat{\mathbf{x}}_t$ under the X-factor dynamics if $\mathbf{y}_t$ is drawn from a new, unknown distribution. We initially make the same prediction $\hat{\mathbf{x}}_t^-$ as we would under the normal regime, but the estimate has more uncertainty since $\mathbf{Q}$ (given by Equation (6.2) for the X-factor) is greater. The extra uncertainty in the prior of the dynamics should therefore lead us to expect that the new observations will carry more weight when updating estimates of $\mathbf{x}_t$; this is the desired behaviour under an unknown regime.

Because of this type of effect, it is already known that adding fictitious system noise to a Kalman filter is a practical trick which can improve performance when it seems that the data has not been modelled correctly (Grewal and Andrews, 2001, §7.2.4). The X-factor extends this idea by using the inflated model for 'other' dynamics in combination with a well-fitted model for normal dynamics.

6.1.2 Usage in a factorial model

Section 5.10 described how factors in a factorial model overwrite different dimensions in the hidden state. As the X-factor operates on every state dimension, there are two possibilities for combining it with other known factors: either it can overwrite everything, or it can be overwritten by everything (except normality). For this application the latter approach is more sensible, so that for example if there is a period of unusual dynamics and an ECG probe dropout then the dropout dynamics generate the heart rate observations and the X-factor generates all other channels.

6.2 Parameter estimation

The parameter estimation techniques in Chapter 5 are inappropriate for setting the value of ξ, as we assume that entirely representative training data is unavailable. However, it is necessary for any novelty detection scheme to make some assumptions about the instances which are to be classified as novel. If this description of what is novel does not come from the data, it must be supplied as prior knowledge. Two parameter setting schemes are therefore described here. In the first, the setting of ξ is done entirely a priori, while in the second training data is used.

6.2.1 Initial parameter setting

By making sufficiently strong assumptions about the data, a value for ξ can be arrived at without examining any empirical statistics. Clearly care needs to be taken as to whether these assumptions are warranted. However, it is at least useful to obtain a value with which to initialise more data-dependent training techniques as given in section 6.2.2. The approach here is to again consider the problem in the context of a static outlier detector as in Figure 6.1(a). The same arguments apply to the classification of outliers in the hidden dynamics.

When trying to evaluate whether a point y came from a particular distribution other than a known one, D, the arbitrary likelihood $p(y|D) = 0.05$ is sometimes used as a guide to significance. This is done despite the fact that the problem is strictly not well-posed without an explicit alternative distribution. Fisher (1958) says about the 5% level that it is 'convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.'[2] Based only on this statistic, a type I error (a false rejection, i.e. that we decide a sample did not come from the proposed distribution, when in fact it did) can be expected to occur around once in every twenty trials. Depending on the application this is often considered to be an acceptable margin of error.

If the task is to classify an individual point as being from a particular Gaussian distribution or not, a reasonable starting point is therefore to check whether it falls within two standard deviations of the mean. Referring again to Figure 6.1(a), in this static case we would therefore set the variance of the X-factor mode $\sigma^2_{(*)}$ such that the points of intersection of the likelihood functions occur at two standard deviations of the 'normal' mode (which has variance $\sigma^2_{(1)}$). The ratio of the variances can be calculated directly by writing the expression for the two Gaussian distributions having an equal value at the point $r\sigma_1$,

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}\sigma_{(1)}}\exp\left(-\frac{r^2\sigma^2_{(1)}}{2\sigma^2_{(1)}}\right) &= \frac{1}{\sqrt{2\pi}\sigma_{(*)}}\exp\left(-\frac{r^2\sigma^2_{(1)}}{2\sigma^2_{(*)}}\right), &(6.8)\\
\frac{\sigma_{(*)}}{\sigma_{(1)}}\exp(-\frac{r^2}{2}) &= \exp\left(-\frac{r^2\sigma^2_{(1)}}{2\sigma^2_{(*)}}\right), &(6.9)\\
\log\xi &= r^2\left(1-\frac{1}{\xi}\right), &(6.10)
\end{aligned}$$

where $\xi \equiv \frac{\sigma^2_{(*)}}{\sigma^2_{(1)}}$. For $r = 2$ this has a non-trivial solution at $\xi \approx 50$. For a weaker significance measure, $r = \sqrt{2}$ has a solution at $\xi \approx 5$. For the dynamical X-factor, a weaker significance level is often appropriate since there can be a strong temporal

[2] In the context of a normal (Gaussian) distribution.

prior which also influences whether the system switches out of the normal mode or not (because the normal regime has a long expected dwell time, the prior probability of switching out of it is low). One could try to factor in the prior transition probabilities explicity, or use the lower value $\xi = 5$ as a rule of thumb. Preliminary experiments have been successful with the latter as a starting point, as shown for example in Figure 6.4.

These are therefore reasonable estimates for ξ either as initial settings or where it is impractical or undesirable to look at specific examples of novel dynamics. However, the limitations of this parameter setting scheme should be borne in mind:

- The parameter setting is based only on the expected number of type I errors (X-factor false positives) and says nothing about the number of type II errors (false negatives);
- The parameter setting is somewhat arbitrary, having been made without looking at any examples of novel dynamics.

The assumptions behind this method of choosing the parameter should also be considered:

- Normal dynamics are well modelled, and in particular they are linear-Gaussian;
- The specified deviation in dynamics (of 2 or $\sqrt{2}$ standard deviations from normality) corresponds to a level of clinical significance.

It is unlikely that these assumptions are strictly accurate in practice, particularly the latter. This motivates finding a way to refine the parameter setting in the light of real data.

6.2.2 Updating parameters with training data

This section describes methods which can be used to set the value of ξ in the situation where there is training data available. The training data may be labelled to different extents:

1. Fully annotated. Labels are available for the times in which known and novel patterns are occurring (ground truth for all factors including the X-factor);
2. Partly annotated. Labels are available for the times in which known patterns are occurring, but not novel patterns;

3. Not annotated.

Learning parameters in each of these cases is a supervised, semi-supervised or unsupervised problem respectively. Each case is considered in this section. Once again the problem is first addressed for the static 1D case, shown in Figures 6.1(a) and (b). The dynamic case follows analogously.

6.2.2.1 X-factor updates for a static model

Consider one dimensional i.i.d. data which is fully annotated. We take there to be $K-1$ 'regular' switch settings (denoted by $s_i = 1,\ldots,K-1$) and the X-factor (denoted by $s_i = *$). In this case the value of ξ can be calculated directly by finding the ratio between the variance of points which have been generated by the X-factor and the variance of the normal mode σ^2,

$$\xi = \frac{\sum_{i=1}^{N}(x_i-\mu_{(1)})^2 I(s_i=*)}{\sigma_1^2\sum_{i=1}^{N} I(s_i=*)} . \tag{6.11}$$

If s_i is unknown, then it is uncertain which observations were generated by the X-factor and ξ must be updated with EM. An initial setting of ξ (e.g. $\xi = 50$ from the previous section) is used to calculate the expected distribution $p(\mathbf{s}|\mathbf{y})$. The expected complete data log likelihood is

$$\begin{aligned} Q &= \sum_{i=1}^{N}\sum_{k=1}^{K} p(s_i=k|y_i,\theta_{\text{old}})\log p(s_i=k,y_i|\theta) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K-1} p(s_i=k|y_i,\theta_{\text{old}})\left(-\log(2\pi)^{\frac{1}{2}}-\log\sigma_{(k)}-\frac{(x_i-\mu_{(k)})^2}{2\sigma_{(k)}^2}\right) \\ &\quad+\sum_{i=1}^{N} p(s_i=*|y_i,\theta_{\text{old}})\left(-\log(2\pi)^{\frac{1}{2}}-\log\xi^{\frac{1}{2}}\sigma_{(1)}-\frac{(x_i-\mu_{(1)})^2}{2\xi\sigma_{(1)}^2}\right) , \end{aligned} \tag{6.12}$$

where we are summing over the $K-1$ known switch settings and the X-factor. Differentiating with respect to ξ gives

$$\frac{\partial Q}{\partial\xi} = \sum_{i=1}^{N} p(s_i=*|y_i,\theta_{\text{old}})\left(-\frac{1}{2\xi}+\frac{(x_i-\mu_{(1)})^2}{2\xi^2\sigma_{(1)}^2}\right) . \tag{6.13}$$

Setting this partial derivative to zero yields the M-step update expression,

$$\xi = \frac{\sum_{i=1}^{N}(x_i-\mu_{(1)})^2 p(s_i=*|y_i,\theta_{\text{old}})}{\sigma_{(1)}^2\sum_{i=1}^{N} p(s_i=*|y_i,\theta_{\text{old}})} . \tag{6.14}$$

Note the resemblance to (6.11).

6.2.2.2 X-factor updates for the SKF

The updates in the dynamic case are analogous. Instead of applying them to the levels of the observations, however, we are now interested in deviations in the underlying state over time. It can be shown that the estimate for ξ when labelling is present is

$$\tilde{\xi} = \frac{\sum_{t=2}^{T}(\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1})^\top {\mathbf{Q}^{(1)}}^{-1}(\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1}) I(s_t = *)}{\sum_{t=2}^{T} I(s_t = *)} . \qquad (6.15)$$

The intuition is similar to the static case, in that the update expression calculates a Z-score between the successive state estimates and the normal system noise covariance, including each time frame for which the X-factor is known to have been active. Where the X-factor labelling is unavailable, the reestimation formula is

$$\tilde{\xi} = \frac{\sum_{t=2}^{T}(\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1})^\top {\mathbf{Q}^{(1)}}^{-1}(\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1}) p(s_t = *|\mathbf{y}, \theta_{\text{old}})}{\sum_{t=2}^{T} p(s_t = *|\mathbf{y}, \theta_{\text{old}})} . \qquad (6.16)$$

which calculates expectations over the posterior probability that X-factor was active at each time point. See section C.4 for a derivation.

The update in (6.16) may not be straightforward to apply due to difficulty in incorporating all the observed training data into the estimates $\hat{\mathbf{x}}_t$ and $p(s_t = *|\mathbf{y}_{1:N})$ in the E-step. In (6.15), the SKF is conditioned on a series of switch settings and so the Kalman filtering and smoothing equations can be used to calculate this expectation exactly. In (6.16), where the switch settings are unobserved, the smoothed expectations are difficult to calculate, due to the exponentially growing number of possible histories in both the filtering and smoothing stages of inference. Techniques are available to calculate smoothed estimates (Klaas et al., 2006; Barber and Mesot, 2006; Ghahramani and Hinton, 1998). The approach used here, however, is to replace $\langle \mathbf{x}_t \rangle_{\mathbf{x}_{1:N}, s_{1:N}|\mathbf{y}_{1:N}}$ (denoted by $\hat{\mathbf{x}}_t$ above) with the filtered expectation $\langle \mathbf{x}_t \rangle_{\mathbf{x}_{1:t}, s_{1:t}|\mathbf{y}_{1:t}}$ and to replace $p(s_t|\mathbf{y}_{1:N})$ with the filtered distribution $p(s_t|\mathbf{y}_{1:t})$ to give a 'pseudo-EM' algorithm in the same manner as Shumway and Stoffer (1991).

6.2.2.3 X-factor updates for the Factorial SKF

The factorial case is a little more complicated due to the possibility that different combinations of factors can overwrite different channels. For example, if a bradycardia is occurring in conjunction with some other, unkown regime, then the heart rate dynamics are already well explained and should not be taken into account when reestimating the X-factor parameter ξ.

Some additional notation needs to be introduced to deal with these effects. First, it is assumed that the hidden dynamics have a (square) block structure, where each observation component is represented by a number of hidden state dimensions (see Table 5.4 on page 86 for the application specifics). The term d_c is used to denote the dimensions of the hidden state $\mathbf{x}_t$ which pertain to the observation channel c. The notation $\mathbf{x}_{\{t,d_c\}}$ represents the values of these dimensions of the hidden state at time t, and this variable is assumed not to share any dependency with $\mathbf{x}_{\{t,d_{e\neq c}\}}$ (conditional on a particular switch setting). Similarly, $\mathbf{A}_{d_c}^{(s)}$ and $\mathbf{Q}_{d_c}^{(s)}$ represent the sub-matrices which describe the dynamics on state dimensions d_c for regime s. Recall for the factorial model that s is a cross-product of K factor settings, the known factors denoted by $f_1 \dots f_{K-1}$ and the X-factor denoted by f_*. Finally, the function $o(f_i,s,c)$ encodes the knowledge as to which channels are overwritten by which factors:

$$o(f_i,s,c) = \begin{cases} 1 & \text{if factor } f_i \text{ affects state dimensions } d_c \text{ in switch setting } s \text{ ,} \\ 0 & \text{otherwise .} \end{cases} \tag{6.17}$$

Using this notation, the updates can be given as

$$\xi \quad = \quad \frac{\sum_{t=2}^{T}\sum_s \sum_c o(f_*,s,c)a(t,c)p(s_t|\mathbf{y},\theta_{\text{old}})}{\sum_{t=2}^{T}\sum_s \sum_c o(f_*,s,c)p(s_t|\mathbf{y},\theta_{\text{old}})} \; . \tag{6.18}$$

where

$$a(t,c) = \left(\mathbf{x}_{\{t,d_c\}} - \mathbf{A}_{d_c}^{(1)}\mathbf{x}_{\{t-1,d_c\}}\right)^{\top} \mathbf{Q}_{d_c}^{(1)^{-1}} \left(\mathbf{x}_{\{t,d_c\}} - \mathbf{A}_{d_c}^{(1)}\mathbf{x}_{\{t-1,d_c\}}\right) \; . \tag{6.19}$$

A derivation is given in section C.5. The function $a(t,c)$ is again effectively a Z-score. The block-diagonal conditional independence structure of the dynamics allows the calculation to be factorised with respect to each group of dependent state dimensions. The update in (6.18) is conceptually similar to (6.16), though the new value of ξ is influenced by every switch setting in which the X-factor overwrites any hidden state, and calculated in terms of each set of hidden state dimensions d_c.

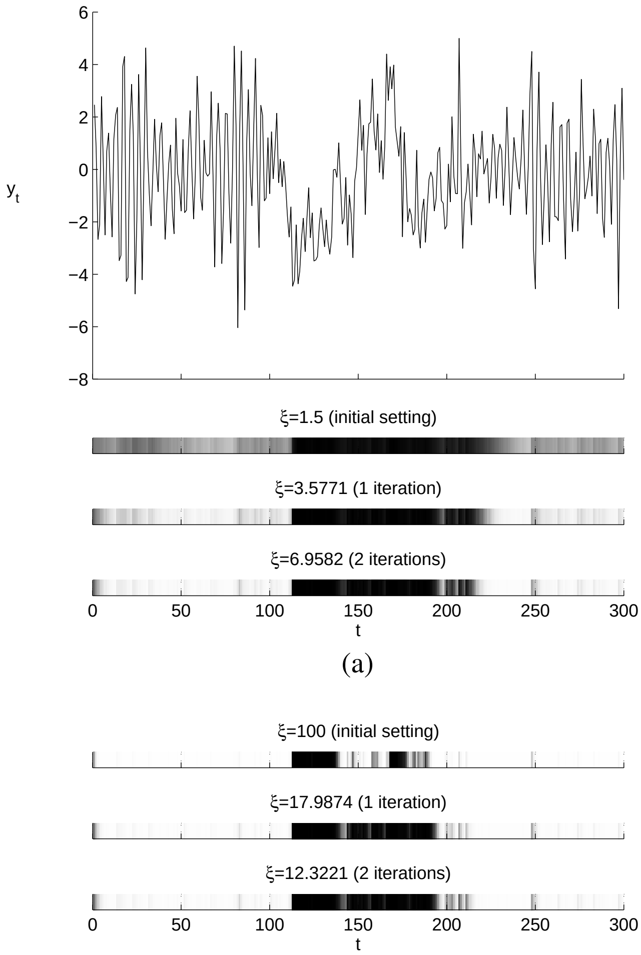
This section on parameter estimation is concluded with two additional comments about setting ξ. First, as with any generative model it is possible to look at samples created with different settings for ξ and see if they create the types of novel dynamics which might be expected. Second, it might be the case that novel dynamics are in general not the same as those seen in the training data. If training examples had only mild deviations from normality, for example, and more extreme excursions were to be expected in general, then the fitted value of ξ could be increased accordingly. Conversely, if the training examples were thought to be extreme cases then ξ could be decreased.

6.2.3 Example X-factor training

The ideas in the previous section are illustrated with examples of parameter estimation for the X-factor. First a synthetic example is used to illustrate operation in a setting in which all properties of the data and the model are known. Data was generated from two arbitrary AR(2) regimes, and concatenated as shown in the top panel in Figure 6.3. The first regime was considered ‘normal’ and the second ‘abnormal’. Gaussian noise was added to the whole sequence to make the dynamics latent. A switching Kalman filter was set up with two switch settings. The first switch setting corresponded to normality, with dynamics set to the same parameters used to generate the first process. The second switch setting was set up according to (6.2) and (6.3) with hand-picked values of ξ. The top probability plots in Figure 6.3(a) and 6.3(b) show the results of inferring the switch settings given low and high initial settings of ξ. In these plots, white corresponds to a posterior probability of 0 and black to a posterior probability of 1. As expected, in the first case a low initial setting of ξ caused false positives, and in the second case the high initial setting caused false negatives. In each case, two iterations of EM were used to reestimate ξ causing the estimates to quickly converge to a level which reduced errors. After 8 iterations the two estimates converge to within three decimal places, at $\xi = 10.934$.

Next, an example involving physiological data is considered, shown in Figure 6.4. As well as being genuine monitoring data, this extends the previous example in that it is multidimensional and contains an instance of a known pattern, the blood sample artifact. The top panel in Figure 6.4 shows the period of observations, in which a blood sample is taken for around 180 seconds, and a period of physiological disturbance follows with a duration of around 780 seconds. The onset and duration of the significant physiological disturbance were determined by a clinical expert (NM).

A factorial switching Kalman filter was set up to model this period of data. Two factors were included: the blood sample factor, trained with other examples of blood sample artifacts from the same baby, and the X-factor with an initial setting of $\xi = 5$ according to section 6.2.1. Normal dynamics were trained using the preceding 1000 seconds of observations, during which time the baby was stable. With this initial configuration, inferences were made, and the posterior probabilities are shown in the top horizontal probability plot in Figure 6.4. The initial setting of ξ appears to be too high, as much of the 13-minute period of physiological instability is not claimed by the X-factor. There are also false positives for the X-factor on either side of the period of the blood sample.



(a)

(b)

Figure 6.3: EM updates for the X-factor, applied to synthetic data. ‘Normal’ data is generated from the same AR(2) regime with $\alpha_1 = 0.1,\ \alpha_2 = -0.9$ for the intervals $\{$1,100$\}$,$\{$201,300$\}$. A second, randomly chosen AR(2) process with $\alpha_1 = 0.8,\ \alpha_2 = 0.1$ was used to generate the data in the period $\{$101,200$\}$. Gaussian noise with variance 0.5 was added to the whole sequence. The shaded horizontal plots show the posterior probabilities that the X-factor is active when run with an SKF which incorporates the true normal dynamics but which has no knowledge about the second process. The topmost of these inference plots show the posterior calculated with a hand-picked value for ξ, and the plots beneath show the results after two iterations of EM. In panel (a) a low initial setting was used for ξ, resulting in type I errors. After two steps of EM, the value of ξ has increased appropriately to reduce the error rate. In panel (b), a high initial setting of ξ causes type II errors, which also improve after two steps of EM.

The false positives illustrate a point about the operation of the X-factor, which is that in some cirumstances a higher setting of ξ can cause a switch into the X-factor more rapidly than had there been a low setting of ξ. If there is a short sequence of outlying observations, the state estimates under the X-factor with a high ξ can quickly (perhaps erroneously) adapt to levels of these outliers. If the outlying observations are correlated for a few time frames then the X-factor, having rapidly adapted, becomes a closer fit than the normal model. A lower level of ξ causes the state estimates to adapt more slowly, which 'softens' the switching of the X-factor.

Starting from this initial inference, four iterations of EM were used to reestimate ξ, the aim being to update it from the data without its value being affected by the values which were thought to be artifactual. The results of this process are shown in the lower probability plot. Here, both false positives and false negatives have been reduced. Further EM iterations did not significantly change the estimate.

6.3 Alternative novelty detection schemes

This section describes other methods which might be used for novelty detection in the context of the FSKF, their relative strengths and weaknesses and relationships to the X-factor method.

6.3.1 Threshold method

An alternative scheme for novelty detection might be to monitor the likelihood that each dynamic model could have generated the observations, and to classify an observation as novel if it falls below a certain threshold. The data sequences for which all models have low likelihood (relative to the likelihoods for data sequences known to be normal). This test has been shown to be practical in a number of settings: it is a common aspect of much of the previous work in novelty detection reviewed in section 6.4.

For i.i.d. static data, the threshold test and the X-factor are equivalent. The setting of ξ in Figure 6.1(a) determines the level of the threshold. In this case, the value of the X-factor is solely to make the alternative model explicit and thefore to enable estimation methods such as EM to be applied.

For the dynamic case, they are not equivalent. The important point is that the estimate $\hat{\mathbf{x}}_t$ is updated in a different way under the X-factor than under the normal regime — it is not simply the same distribution with wider variance, as explained in section

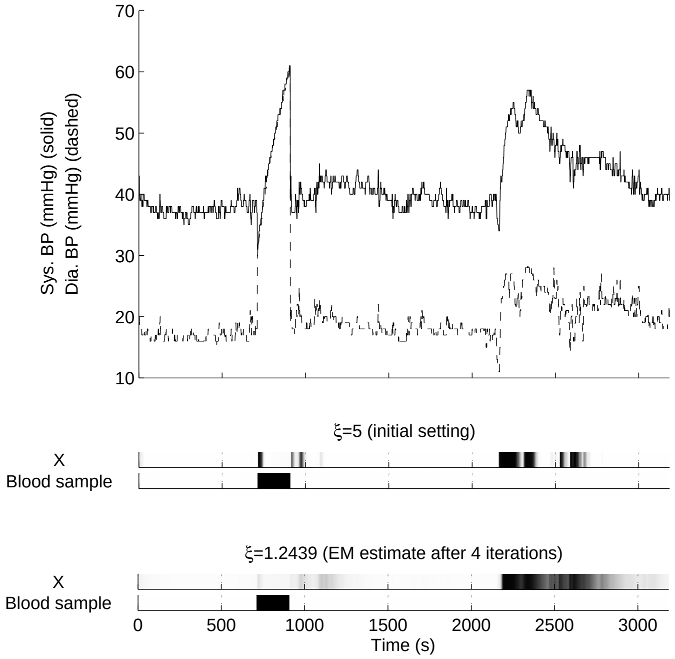


Figure 6.4: EM updates for the X-factor, applied to physiological data. The top panel shows a period of blood pressure measurements, during which there is a blood sample between approximately 700 and 900 seconds, and a disturbance of unknown cause between approximately 2200 and 3000 seconds. Normal dynamics were trained using 1000 seconds of data preceding the period shown above. The horizontal plots show the posterior probabilities that the data is governed by the X-factor or the blood sample factor. In the first pair of horizontal plots, an initial setting of ξ gives sub-optimal X-factor inferences, with both false positives near the blood sample and false negatives around the disturbance (see text for details). After reestimation with four iterations of EM, the inferences are improved.

6.1.1. Because switching into the X-factor dynamics has the effect of weakening the prior on the model, observations have more weight in updating these estimates. Therefore the X-factor has a different purpose to the threshold test, being used to model unusual dynamics rather than just to detect them.

The threshold test is also based on analysing the likelihood of each regime, statistics which have limited interpretability. While the likelihood of the normal mode is related to the idea of novelty, it is not consistent to draw inferences from it alone (an example of a prosecutor's fallacy (MacKay, 2003, §2.1)). This difficulty is caused by the test being made outside the model; the only way to obtain a meaningful probability is to compare specific models. This also means that the process of finding the level at which to set the threshold is more reliant on heuristics than estimating parameters for the X-factor. There is no way to learn the right setting with EM or any related technique. Similarly to the null hypothesis test, this scheme has the convenience of not needing to specify an explicit alternative model, at the expense of obtaining a result which is somewhat arbitrary.

6.3.2 Inflated observation noise

A natural alternative to formulating a new regime by inflating the normal system noise covariance is to inflate the normal observation noise covariance. The parameters of the regime would then be given by $\{\mathbf{A}^{(*)},\mathbf{Q}^{(*)},\mathbf{C}^{(*)}\} = \{\mathbf{A}^{(1)},\mathbf{Q}^{(1)},\mathbf{C}^{(1)}\}$ and $\mathbf{R}^{(*)} = \xi\mathbf{R}^{(1)}$. This model has been put forward by West and Harrison (1999) as an outlier detector.

In general, for physiological monitoring, changes in the dynamics are more relevant than changes in the observation process. Outliers caused by the observation process in this context would correspond to changes in the monitoring equipment (such as noise from an ECG lead), whereas changes in the hidden dynamics are related in the model to the baby's physiology.

Note that these two methods of detecting changes in dynamics are not mutually exclusive. It would be possible to formulate two X-factors, one of which has inflated system noise and the other having inflated observation noise, or one X-factor which has both covariances inflated.

6.3.3 Alternative dynamics

It would have been possible to use a very general dynamical model for the X-factor, for example a white noise process. However, this runs the danger that it will be such a poor fit to the observed dynamics that one of the known regimes will claim data, even when it ‘should’ be picked up by the X-factor. In practice it is very difficult to obtain meaningful inferences when the X-factor dynamics are of a different form to the normal dynamics, as shown in the experiments in section 7.3.

6.4 Novelty detection

This chapter has introduced a model for novel dynamics, which relates to the field of novelty detection. There is a large body of work on statistical approaches to novelty detection, reviewed in Markou and Singh (2003). In general the problem is applied to static data, where the goal is to learn the density of training data and to raise an alarm for new data points which fall in low density areas, based on some threshold. The Gaussian mixture model is commonly used due to its property of universal approximation, see for example Roberts and Tarassenko (1994). This approach can be extended to a time-series context by modelling the next observation $p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t})$ based on the earlier observations, and detecting observations that have low probability. Many different methods have been used in order to obtain a predictive density for new observations, such as converting the sequence into symbolic strings (Keogh et al., 2002), searching tree structures containing wavelet coefficients (Shahabi et al., 2000), or regression with support vectors (Ma and Perkins, 2003). Such approaches define a model of normality, and look for deviations from it, e.g. by setting a threshold.

A somewhat different approach is to define a very broad “outlier” distribution as well as normality, and carry out probabilistic inference to assign patterns to the normal or outlier components. By doing this, outlying observations are explicitly modelled in a generative sense, rather than being detected heuristically. For time-series data this approach was followed by Smyth (1994), who considered the use of an unknown state when using a HMM for condition monitoring.

6.5 Summary

This chapter has shown how novelty detection can be incorporated into the condition monitoring model, suggested sensible initial parameter settings and provided a method for updating parameters with labelled or unlabelled training data. This construction, the X-factor, is useful both as an indicator of novel dynamics which could have clinical significance and as a measure of confidence for the model.

Related work in novelty detection has been reviewed in section 6.4, in which most of the methods are based on formulating a prediction for new data points and classifying that point as novel if it has low likelihood. The X-factor approach is different in that it explicitly represents novel dynamics. Smyth (1994) considered the use of an unknown state when using a HMM for condition monitoring, which uses a similar idea to the X-factor but in a simpler context, as in his work there is no factorial state structure and no explicit temporal model (c.f. the FSKF).

Chapter 7

Experiments

This chapter describes experiments used to evaluate the model for condition monitoring. Section 7.1 gives details of preparatory steps common to all experiments, including the labelling of data by clinical experts. Experiments done to evaluate the classification of known patterns are described in section 7.2, while section 7.3 describes experiments done to evaluate the X-factor.

Some conventions in plotting the results of these experiments are adopted throughout this chapter. Horizontal bars below time-series plots indicate the posterior probability of a particular factor being active, with other factors in the model marginalised out. White and black indicate probabilities of zero and one respectively[1]. On these plots, 'BS' denotes the blood sample factor, 'BR' denotes the bradycardia factor, 'IO' denotes the incubator open (handling of baby) factor, 'TD' denotes the temperature probe disconnection factor, and 'TR' denotes the transcutaneous probe recalibration factor. In general the plots show a subset of the observation channels and posteriors from a particular model—this is indicated in the text.

7.1 Setup

24-hour periods of monitoring data were obtained from fifteen premature infants in the intensive care unit at Edinburgh Royal Infirmary. The babies were between 24 and 29 weeks gestation (around 3-4 months premature), and all in around their first week of life. The specific periods of data were chosen to be the first 24 hours of continuous monitoring data available for each infant. These recordings had been made as routine

[1] A convenient property of the models evaluated here, from the perspective of visualisation, is that the factor posteriors tend be close to zero or one. This is partly due to the fact that the discrete transition prior $p(s_t|s_{t-1})$ is usually heavily weighted towards staying in the same switch setting (long dwell times).

Factor	Incidences	Total duration	Average duration
Incubator open	690	41 hours	3.5 mins
Core temp. disconnection	87	572 mins	6.6 mins
Bradycardia	272	161 mins	35 secs
Blood sample	91	253 mins	2.8 mins
TCP recalibration	11	69 mins	6.3 mins
Abnormal (other)	605	32 hours	3.2 mins

Table 7.1: Number of incidences of different factors, and total time for which each factor was annotated as being active in the training data (total duration of training data $15 \times 24 = 360$ hours).

practice with the Badger Patient Data Management System operating in the Edinburgh Royal Infirmary NICU.

Each of the fifteen 24-hour periods was annotated by two clinical experts. At or near the start of each period, a 30 minute section of normality was marked, indicating an example of that baby's current baseline dynamics. Each of a set of known common physiological and artifactual patterns (bradycardia, opening the incubator, core temperature probe disconnection, blood sampling and transcutaneous probe recalibration) were also marked up[2]. The factors are added in reverse order of total duration in Table 7.1.

Finally, it was noted where there were any periods of data in which there were clinically significant changes from the baseline dynamics not caused by any of the known patterns. The two annotators first looked together at a set of measurement channels from another baby in order to agree on a policy as to what types of changes constituted clinical significance. Having agreed on the best annotation for this sample monitoring data, the two annotators then independently marked up the 'Abormal (other)' category on the fifteen 24-hour sets of measurements. The software package TSNet (Hunter, 2006) was used to record these annotations, and the recorded intervals were then exported into Matlab. The number of intervals for each category, as well as the total and average durations, are shown in Table 7.1. The figures for the 'Abnormal' category were obtained by combining the two annotations, so that the total duration is the number of points which either annotator thought to be in this category, and the number of incidences was calculated by merging overlapping intervals in the two annotations

[2] For expediency, some initial estimates as to the durations of known patterns were made by the author. The clinical experts were then able to check these and amend as necessary.

Inference type		Incu. open	Core temp.	Blood sample	Bradycardia
GS	AUC	0.87	0.77	0.96	0.88
	EER	0.17	0.34	0.14	0.25
RBPF	AUC	0.77	0.74	0.86	0.77
	EER	0.23	0.32	0.15	0.28
FHMM	AUC	0.78	0.74	0.82	0.66
	EER	0.25	0.32	0.20	0.37

Table 7.2: Inference results on three passes of the evaluation data, using a second-by-second evaluation. Inferences were made with a FSKF with Gaussian sum approximation (GS), Rao-Blackwellised particle filtering (RBPF) and with a Factorial Hidden Markov Model (FHMM). AUC denotes area under ROC curve and EER denotes the equal error rate.

(two overlapping intervals are counted as a single incidence).

The rest of this chapter shows the results of performing inference on this data and comparing it to the gold standard annotations provided by the clinical experts. Two of the possible factors discussed earlier are omitted from the formal evaluation. Recalibrations of the transcutaneous probe were not practical to evaluate here due to a shortage of transcutaneous measurements (the appropriate probe was only being used in one of the fifteen recording periods)[3]. Blood pressure waves are a relatively rare occurrence and only minor examples were present in the annotated dataset. Evaluation of these particular factors is therefore left as future work.

7.2 Known factors

In order to maximise the amount of test data and reduce the possibility of bias, evaluation was done with three-fold cross validation. The fifteen 24-hour data periods were split into three groups of five (grouped in order of the date at which each baby first arrived in the NICU). Three tests were therefore done for each model, in each case testing on five babies and training on the remaining ten, and summary statistics were obtained by averaging over the three runs.

The quality of the inferences made were evaluated using area under the receiver operating characteristic curve (AUC) and equal error rates (EER). These statistics are

[3] Note that the diagonal AR parameterisation of the hidden dynamics makes it easy to add or omit factors from the model.

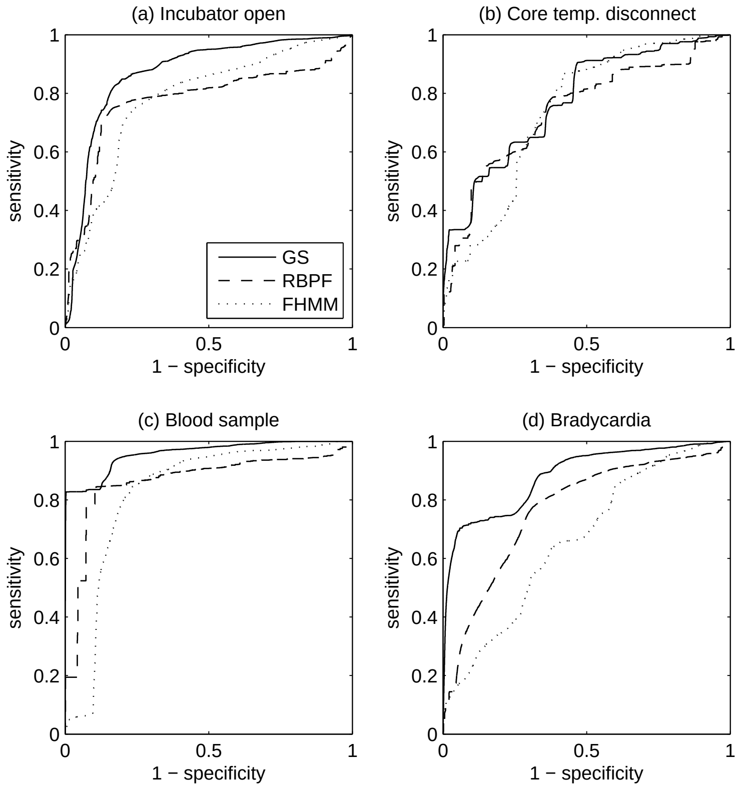


Figure 7.1: ROC curves for classification of known factors, using a second-by-second evaluation. Inferences were made with a FSKF with Gaussian sum approximation (GS), Rao-Blackwellised particle filtering (RBPF) and with a Factorial Hidden Markov Model (FHMM).

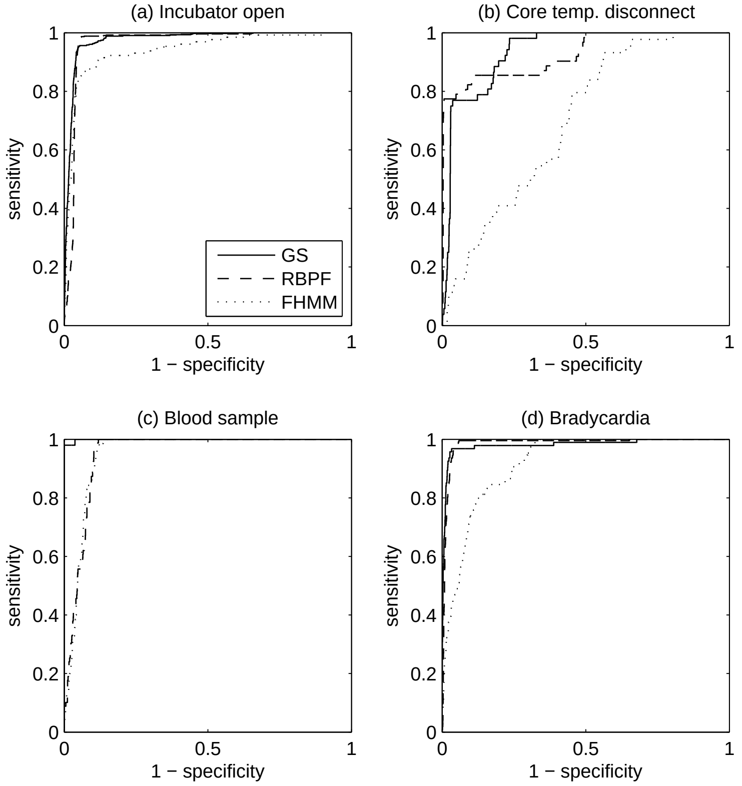


Figure 7.2: ROC curves for classification of known factors, using an event-based evaluation. Inferences were made with a FSKF with Gaussian sum approximation (GS), Rao-Blackwellised particle filtering (RBPF) and with a Factorial Hidden Markov Model (FHMM).

Inference type		Incu. open	Core temp.	Blood sample	Bradycardia
GS	AUC	0.98	0.94	0.99	0.98
	EER	0.05	0.17	0.02	0.03
RBPF	AUC	0.96	0.93	0.95	0.98
	EER	0.05	0.15	0.11	0.04
FHMM	AUC	0.94	0.70	0.95	0.92
	EER	0.11	0.41	0.10	0.16

Table 7.3: Inference results on three passes of the evaluation data, using an event-based evaluation. Inferences were made with a FSKF with Gaussian sum approximation (GS), Rao-Blackwellised particle filtering (RBPF) and with a Factorial Hidden Markov Model (FHMM). AUC denotes area under ROC curve and EER denotes the equal error rate.

a useful summary of performance when there are disparities in the numbers of points of each class. ROC curves are calculated using two different methods; the first looking at model posteriors on a second-by-second basis (explained in the rest of this section), and the second in terms of detected events (section 7.2.1).

If a particular threshold is used to determine the class of each inference then there would be a certain false positive rate (α) and false negative rate (β). If the threshold was 0, for example, then α will be 0 and β will be 1. As the threshold is increased, α decreases and β rises. Some threshold will give equal error rates such that $\alpha = \beta$. This error rate gives a useful insight into the performance of the system as it is constrained by both sensitivity and specificity. EER can also be interpreted as the distance along the diagonal of the ROC plot from the top left corner (the point representing sensitivity and specificity of 100%) to the point of intersection with the curve.

Summary statistics for three types of models are given in Table 7.2, and the corresponding ROC curves are shown in Figure 7.1. Four factors are considered (incubator open, temperature probe disconnection, bradycardia and blood sample), and inferences are made for this set of factors with an FSKF, first with the Gaussian sum approximation, and then with Rao-Blackwellised particle filtering. The number of particles was set so that inference time was the same as as for the Gaussian sum approximate inference, in this case $N = 71$. For comparison, the same set of factors was inferred with the FHMM model, in which training and inference were carried out as described in section 4.2. The performance of the FHMM is a useful comparison because it has similar structure to the FSKF but with no hidden continuous dynamics. For all factors, the effect of adding the continous latent dynamics is to improve performance, as can be seen by comparing the FHMM performance to the two FSKF models. RBPF inferences tend to be less accurate than those made with the Gaussian-sum approximation. This is at least partly due to the inability of the model to sample effectively from all the latent space when there is a high number of switch settings, and in this case the number of possible switch settings (16) is significant relative to the number of particles (71). It can be seen that core temperature probe disconnection is in general the most difficult factor to infer, partly because very long periods of disconnection are eventually misclassified by the model as being normal.

7.2.1 Event-based evaluation

The known factor models were also evaluated from the perspective of detecting events, as we are interested in the pattern of errors not only second by second but also in terms of 'hits' and 'misses' in picking up occurrences of patterns. The methodology used for this type of evaluation is based on a comparable problem in the speech recognition literature, known as *word spotting*. In this type of problem, an audio sequence containing speech is analysed to try to find occurrences of particular key words. Established evaluation frameworks for this task exist (see e.g. HResults in Young et al. (2001, §14.15)), which standardise the process of determining whether events have been detected or not.

The first stage in such an evaluation is to take the posterior probabilities for each factor at every time frame, and convert them into estimates of detected 'events' with start and end times. The posteriors are thresholded to give binary estimates at each time frame, and consecutive sequences of 1's are taken to be instances of detected events. A score is calculated for each event by averaging the posterior probability mass (in the original, un-thresholded posterior) within that period.

Evaluation proceeds one factor at a time. The algorithm which we use to draw an ROC curve in this context is based on HResults and is as follows:

1. The detected events are sorted by decreasing order of score.

2. The true positive rate (TP) and false positive rate (FP) are initialised to zero.

3. It is calculated whether each detected event was a 'hit' or a 'miss'. To do this, the midpoint of each reference event (the doctors' annotation) is calculated. If the

detected event begins before the midpoint and ends after it, then the reference event has been hit (the detected event is a true positive). Otherwise it has been missed (a false positive).

4. For a hit, TP is incremented, otherwise FP is incremented.

The (TP,FP) pairs obtained after considering each event in order can be used to plot a point on the ROC curve. By working through the detected events in decreasing order of confidence we build up the curve from the bottom left-hand corner. Note that the curve must be normalised, so that when the final event has been evlauated $\text{TP} = \text{FP} = 1$. When the posteriors from all 15 test babies are combined, we obtain an overall ROC curve from which summary statistics can be calculated.

Figure 7.2 shows the overall ROC curves for each factor for the three different methods of inference. Note that after normalisation the true positive rates are equal to the sensitivities, and false positive rates are equal to $(1 - \text{sensitivity})$. Table 7.3 shows summary statistics. Under this method of evaluation, the AUC and EER figures are much improved. This suggests that the errors made by the model are often to do with inferring the wrong start and end times for different factors, and that occurrences of factors as a whole tend to be picked up.

AUC and EER statistics were also calculated for individual babies and plotted in Figure 7.3. This plot shows that for detection of incubator opening there were two outlying babies, with low AUC and high EER. In both cases the poor performance was caused by variable humidity measurements where nursing staff made changes to the incubator settings during the course of the monitoring period. Core temperature probe disconnection is somewhat variable between babies, while blood sampling and bradycardia have consistent performance.

7.2.2 Examples of operation

Specific examples of the operation of these models are now given. Figures 7.4-7.7 show inferences of switch settings made with the FSKF with Gaussian sum approximation (denoted ‘GS’ in Table 7.2). In each case the switch settings have been accurately inferred. Figure 7.4 shows inference of temperature probe disconnection and blood sampling in conjunction with handling, and Figure 7.5 shows inference of further bradycardia, blood sampling, and handling of the baby. Note in 7.5(a) that it has been possible to recognise the disturbance of heart rate at $t = 800$ as being caused by handling of the baby, distinguished from the bradycardia earlier where there is no

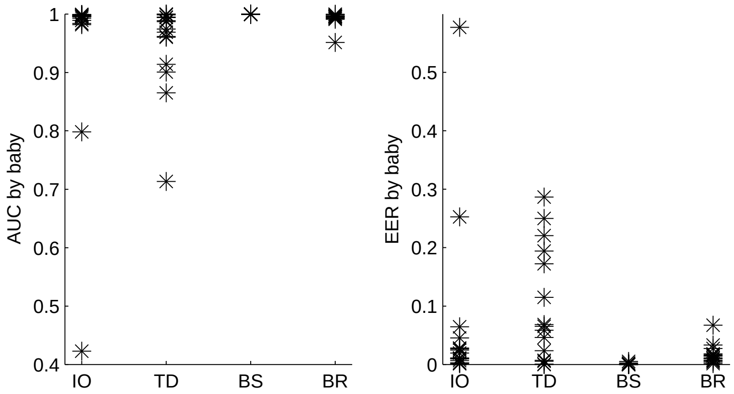
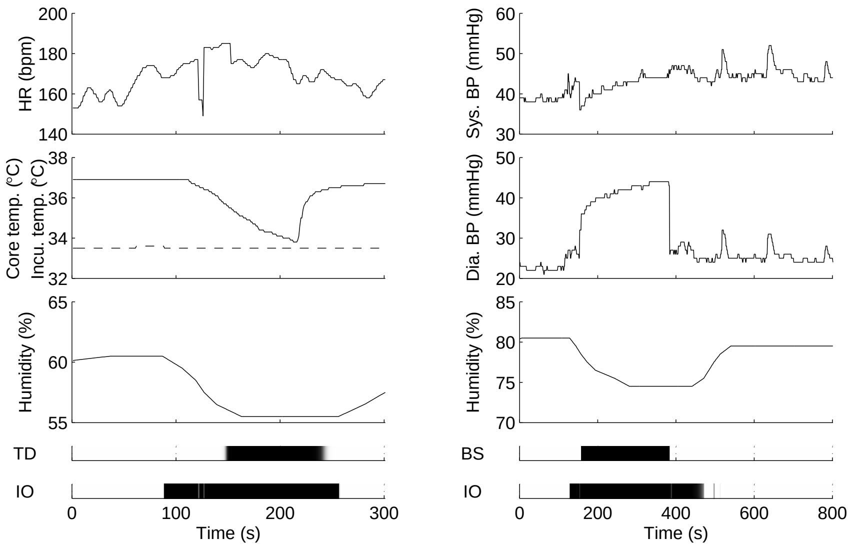


Figure 7.3: AUC and EER for detection of known factors, using event-based evaluation, and plotted by results on each of the 15 babies. IO denotes opening of the incubator, TD denotes core temperature probe disconnection, BS denotes blood sampling, and BR denotes bradycardia.

evidence of the incubator having been entered. Figure 7.6 shows examples of the transcutaneous probe recalibration, correctly classified in conjunction with a blood sample and a core temperature probe disconnection. Note that in 7.6(b) the recalibration and disconnection begin at around the same time, as a nurse has handled the baby in order to access the transcutaneous probe, causing the temperature probe to become detached. More examples of inferred switch settings for known factors are given in the next section, in conjunction with the X-factor.

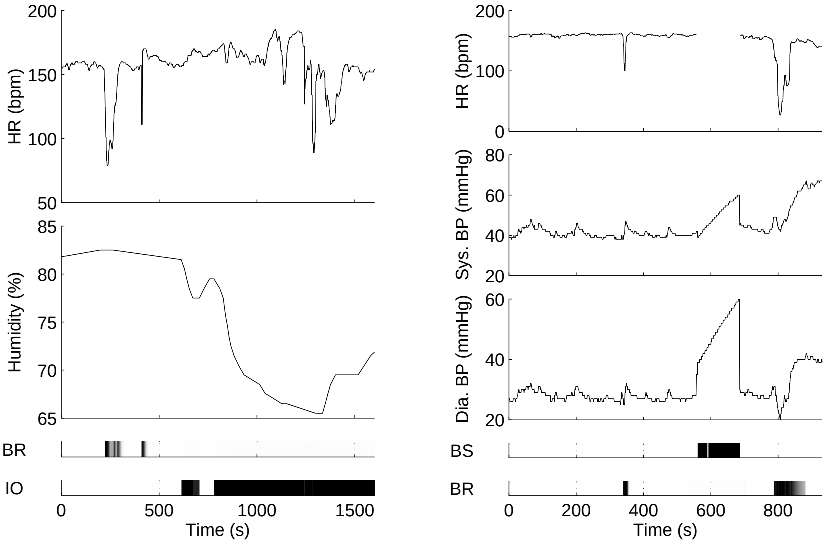
For the blood sample and temperature probe disconnection factors, the measurement data bears no relation to the actual physiology, and the model should update the estimated distribution of the true physiology in these situations accordingly. Figure 7.7 contains examples of the inferred distribution of true physiology in data periods in which these two artifacts occur. In each case, once the artifactual pattern has been detected, the physiological estimates remain constant or decay towards a mean. As time passes since the last reliable observation, the variance of the estimates increases as described in section 5.8.1.

Figure 7.8 shows inferences under the FSKF model and the FHMM model, which help to show why the FSKF has superior classification performance. Recall from section 5.7 that many of the observation channels can exhibit a slowly drifting baseline. During the period shown there is a bradycardia, after which the baseline heart rate is elevated. Both the FSKF and the FHMM pick up the period in which there is a brady-



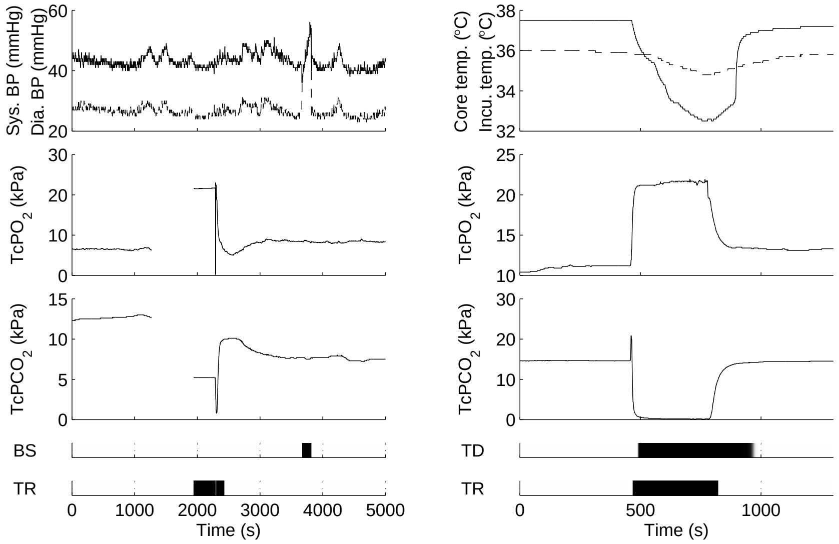
(a) (b)

Figure 7.4: Correctly inferred distributions of switch settings, for two situations involving handling of the baby. TD denotes a core temperature probe disconnection, IO denotes opening of the incubator and BS denotes a blood sample. In panel (a) the incubator is entered and in the course of handling the baby the core temperature probe is disconnected (time 110) and then reconnected (time 210). In panel (b) the incubator is opened in order to take a blood sample. Note the instability in the heart rate caused by handling in panel (a) at time 120.



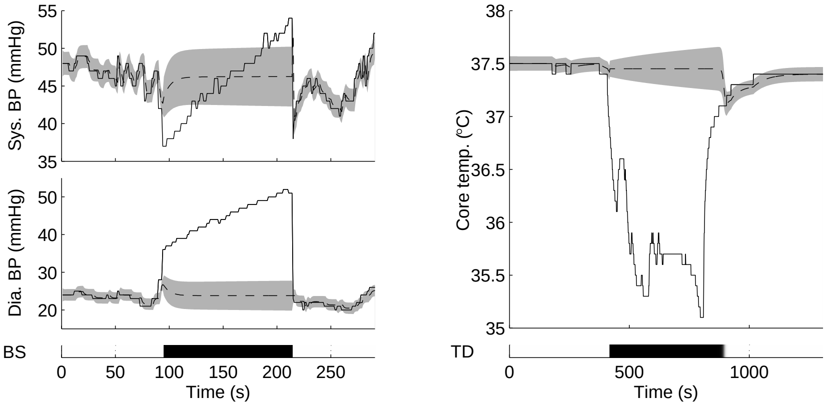
(a) (b)

Figure 7.5: Inferred distributions of switch settings, for two further situations in which there are effects due to multiple known factors. BR denotes bradycardia, IO denotes an opening of the incubator and BS denotes a blood sample. In panel (a) there are incidences of bradycardia, after which the incubator is entered. There is disturbance of heart rate during the period of handling, which is correctly taken to be associated with the handling and not an example of spontaneous bradycardia. In panel (b), bradycardia and blood samples are correctly inferred. During the blood sample, heart rate measurements (supplied by the blood pressure sensor) are interrupted.



(a) (b)

Figure 7.6: Inferred distributions of switch settings, for two situations involving recalibration of the transcutaneous probe. BS denotes a blood sample, TR denotes a recalibration, and TD denotes a core temperature probe disconnection. In panel (a) the recalibration is preceeded by a dropout, followed by a blood sample. Diastolic BP is shown as a dashed line. Transcutaneous readings drop out at around $t = 1200$ before the recalibration. In panel (b), a core temperature probe disconnection is identified correctly, as well as the recalibration.



(a) (b)

Figure 7.7: Inferred distributions of the true physiological state during artifactual corruption of measurements. Panel (a) shows correct inference of the duration of a blood sample (BS), and panel (b) shows correct inference of a temperature probe disconnection (TD). Measurements are plotted as a solid line, and estimates $\hat{\mathbf{x}}_t$ relating to true physiology are plotted as a dashed line with the gray shading indicating two standard deviations. In each case, during the period in which measurements are corrupted the estimates of the true physiology are propagated with appropriately increased uncertainty.

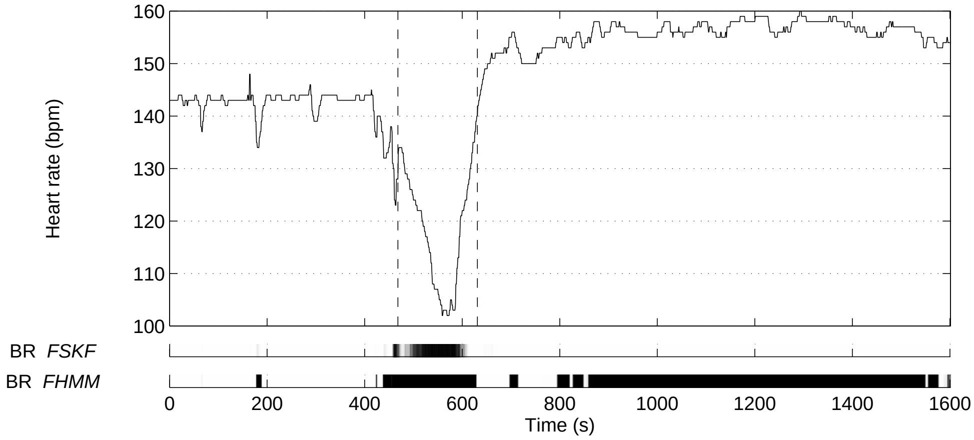


Figure 7.8: Comparison of inference of the bradycardia factor for the FSKF and FHMM models. The dashed vertical lines show a gold standard interval marking the duration of a spontaneous bradycardia. While the FSKF performs an accurate classification, the FHMM is sensitive to drift in the baseline heart rate.

cardia. However, the higher baseline heart rate means that the the normal observations are no longer well modelled in the FHMM, and the bradycardia factor, having high variance, incorrectly claims these measurements. This demonstrated the advantage in developing a parameterisation of normal dynamics which has an explicit changing baseline component, as described in section 5.7.1.

The FSKF experiments both took 14.8 hours to process 360 hours of test data, whereas the FHMM took around 45 minutes. Note that it would not be practical to run the FSKF on this amount of training data without some of the optimisations described earlier. If explicit dropout factors had to be included for six observation channels, rather than using the trick in section 5.8.1, the inference time would be increased by a multiple of 2^6. If the possible switch setting transitions were not constrained as in section 4.7, inference would take four times longer with four factors. With five factors, as in the next section, constraining the transitions makes the inference more than six times faster (see Table 4.2).

A short experiment was also done to compare the quality of inferences with the two FSKF inference schemes (GS and RBPF). The annotated data was separated into a training set of size 10×24 hours and a test set of size 5×24 hours. The four factors in Table 7.2 were evaluated individually (i.e. in a SKF with only one factor at a time), with both Gaussian sum inference and Rao-Blackwellised particle filtering

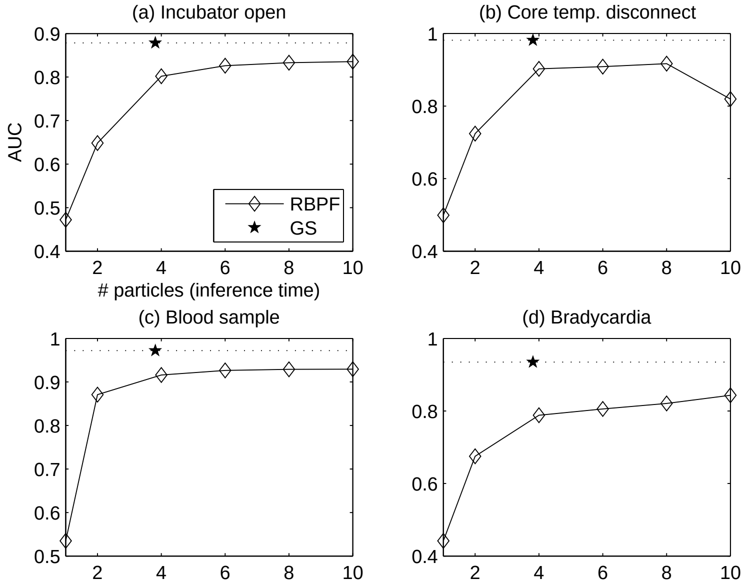


Figure 7.9: AUC for classification of known factors in isolation, using Gaussian sum (GS) and different numbers of particles in RBPF. The alignment of the GS points on the X-axis shows the relative execution time on the test set. Note that as RBPF is stochastic, random effects may decrease the performance even when the number of particles is increased, as at the right of panel (b). Evaluation was done on a second-by-second basis.

with different numbers of particles. The AUCs from these experiments are plotted in Figure 7.9, and the EERs in Figure 7.10. In these plots, the points denoting GS inference are aligned along the X-axis to correspond to the inference time in RBPF (inference time with RBPF increases linearly with the number of particles). In all cases GS outperforms RBPF, even when particles are added such that the inference time for the particle filtering is more than twice that of the GS run.

7.3 Novelty detection

Experiments were done to evaluate the ability of the X-factor to represent novel physiological and artifactual dynamics. Preliminary trials and use of EM estimation showed $\xi = 1.2$ to be a suitable setting (see for example the effects of EM in Figure 6.4). This setting is used for all the experiments described below (except where stated for the alternative X-factor constructions in section 7.3.2).

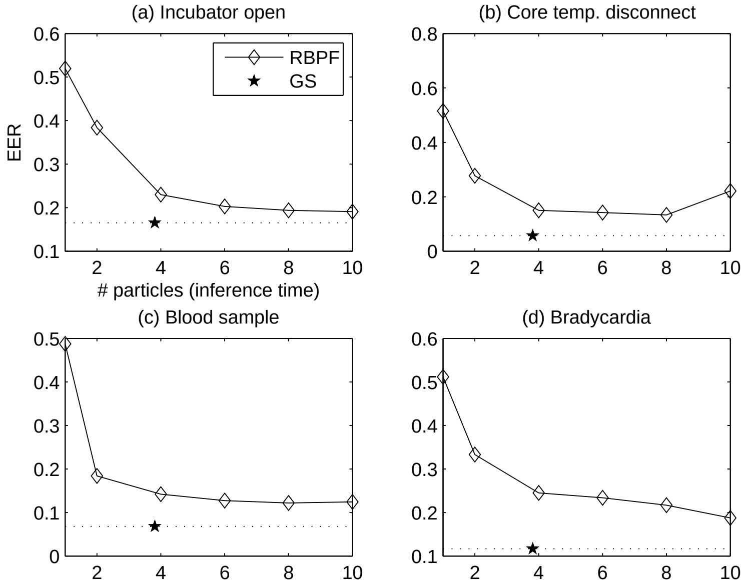


Figure 7.10: EER for classification of known factors in isolation, using GS and different numbers of particles in RBPF as above. Evaluation was done on a second-by-second basis.

7.3.1 Evaluation of the X-factor

An important point to consider when evaluating the X-factor is that there is an element of subjectivity in the ground truth. What counts as significant variation to one clinician may not be viewed the same way by another. Whereas the known factors in the previous section have 'hard' ground truth—there is little ambiguity about the beginning and end of a blood sample, for instance—deciding what counts as abnormal requires more judgement.

Ground truth for the X-factor came from the 'Abnormal (other)' category annotated by the two experts, and combines the two annotations so that the gold standard is the union of the two sets. Figure 7.11 shows a representative example of differences between the two annotations. A period containing some elevated blood pressure at around $t = 500$ is recorded as being significant in both sets of annotations. According to annotation A, the onset and duration of this period are also aligned with changes in SpO_2. In annotation B, an earlier drop in saturation at $t = 290$ is not considered, and the period of significance starts with a slight turning point in the blood pressure measurements and further disturbance of SpO_2. Two further transient drops in saturation are then marked out in annotation A, but not considered significant in annotation B.

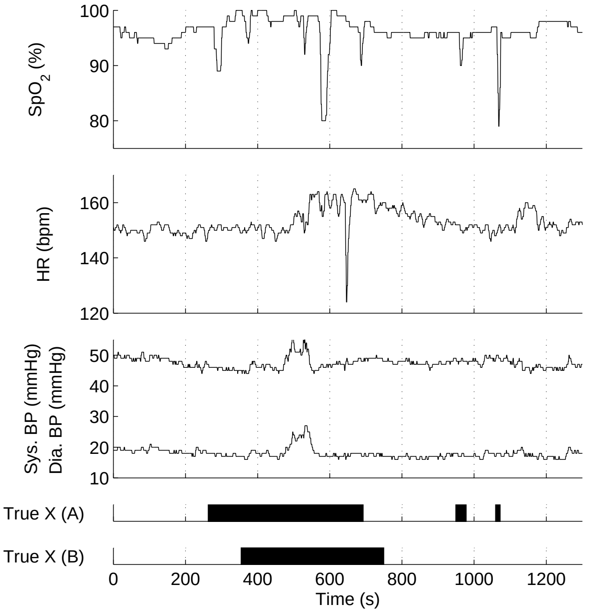


Figure 7.11: Example of discrepancy between the two sets of 'abnormal' notations which make up the ground truth for the X-factor. (A) and (B) denote the two sets of annotations.

An idea of the degree of similarity between the X-factor annotations from the two experts can be obtained by comparing each with the other using ROC analysis. The mean AUC (using annotation set A a gold standard against set B, and vice versa) is 0.85 and the average EER is 0.15. Alternatively, each individual set of annotations can be compared with the union of the two sets (the union of the two sets being taken as the gold standard in the experiments in the rest of this section). In this case the mean AUC is 0.89 and average EER is 0.15[4]. This impression of the amount of subjectivity and the differences between the annotations can be kept in mind when evaluating the model; we would hope for broad agreement between the model and the annotations, but allow that the gold standard is not precisely defined as for the known factors.

Three-fold cross validation was again used to analyse the inferences of different models with different sets of factors. The first model considered contained only the X-factor, the two switch settings therefore being 'normal' or 'abnormal'. The inten-

[4] Because the annotations are binary, in this case there is only one point on the ROC 'curve' between the points (0,0) and (1,1). The AUC and EER can then be calculated directly from this plot, statistics which are somewhat artificial in this context but which allow comparison with the performance statistics given later.

tion with this construction was for it to place probability mass for the X-factor on any period in which anything was happening. As the X-factor here stands in for any known or unknown pattern, the ground truth for this model is the conjunction of all the annotated intervals of every type—known factors and 'abnormal' periods. Another four models are considered, in which the known factors are added to the model one by one. So, for the second model the 'Incubator Open' factor is added and the corresponding intervals are removed from the ground truth for the X-factor. The factors are added in reverse order of total duration in Table 7.1. In the fifth set of factors each known factor has ground truth given by the corresponding annotation, and the X-factor has ground truth given by the 'Abnormal (other)' annotation. Examining the performance of these different models and particular examples of operation gives some insight into the operation of the X-factor, both on its own and in conjunction with the other factors.

Summary statistics are shown in Table 7.4 for a second-by-second evaluation, and again in Table 7.5 for an event-based evaluation, where the models above are numbered (1-5). Only approximate Gaussian sum inference was considered here. The performance in classifying the presence of known factors is almost the same as for when the X-factor was not included (model 'GS' in Table 7.2), only minor variations in AUC and EER being evident. For each of the five models, the X-factor inferences had a rough correlation to the annotations.

It is also useful to calculate the sensitivity and specificity of the X-factor. The posterior probabilities are thresholded at 0.5 to give a binary classification of whether the model believes the X-factor is active or not. The calculation is then made in terms of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN):

$$\begin{aligned}\text{Sensitivity} &= \frac{\text{TP}}{\text{TP}+\text{FN}} = 20.0\% \ , \\ \text{Specificity} &= \frac{\text{TN}}{\text{TN}+\text{FP}} = 95.9\% \ .\end{aligned}$$

This shows that the X-factor is not very sensitive at this threshold (significant variation may not be picked up), but is highly specific (the variation it does pick up is likely to be something that a doctor would find significant). Of course, sensitivity can be exchanged for specificity by adjusting the threshold. For example, a threshold of 0.1 gives a sensitivity of 33.6% and a specificity of 90.5%. The sensitivity-specificity characteristics are shown in Figure 7.12 in the form of an ROC curve.

Examples of the operation of the X-factor are now shown, begining with inferences from model (5) in which the full set of factors is present with the X-factor. Figure 7.13

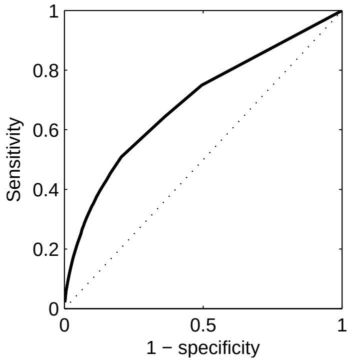


Figure 7.12: ROC curve for the X-factor in model (5), using second-by-second analysis.

shows two examples of inferred switch settings under this model for periods in which there are isolated physiological disturbances. Both the posteriors for the X-factor and the gold standard intervals for the ‘Abnormal (other)’ category are shown. The physiological disturbances in both panels are cardiovascular and have clearly observable effects on the blood pressure and oxygen saturation measurements.

In 7.13(a), the X-factor is triggered by a sudden, prolonged increase in blood pressure and a desaturation, in broad agreement with the ground truth annotation. One characteristic of the X-factor inferences obtained is the ‘trailing off’ effect, in this case slightly beyond the extent of the ground truth intervals. This is to be expected since it is harder to know with filtered inference when the end has been reached, without the benefit of hindsight. See the comments on the possibility of fixed-lag smoothing in section 8.4 for more on this. In 7.13(a) there are two spikes in BP and shifts in saturation which are picked up by the X-factor, also mainly in agreement with the annotation. A minor turning point in the two channels was also picked up at around $t = 2000$, which was not considered significant in the gold standard (a false positive).

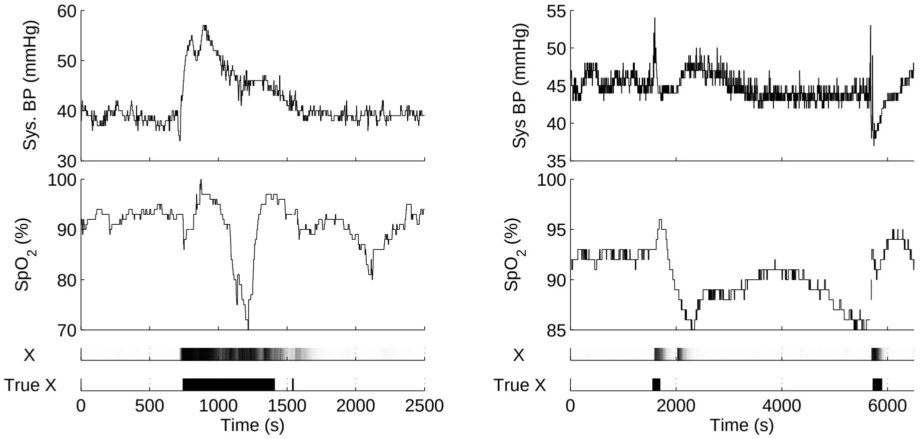
Effects of introducing known factors to model (1) are shown in Figure 7.14. In panel (a), there are two occurences of spontaneous bradycardia, HR making a transient drop to around 100bpm. The X-factor alone in model (1) picks up this variation. Looking at the inferences from model (5) for the same period, it can be seen that the bradycardia factor provides a better match for the variation, and probability mass shifts correctly: the X-factor is now inactive. In panel (b), a similar effect occurs for a period in which a blood sample occurs. The X-factor picks up the change in dynamics when on its own, and when all factors are present in model (5) the probability mass shifts correctly to the blood sample factor. The blood sample factor is a superior description

Model		X-factor	Incu. open	Core temp.	Blood sample	Bradycardia
1	AUC	.72	-	-	-	-
	EER	.33	-	-	-	-
2	AUC	.74	.87	-	-	-
	EER	.32	.17	-	-	-
3	AUC	.71	.87	.78	-	-
	EER	.35	.18	.28	-	-
4	AUC	.70	.87	.78	.96	-
	EER	.36	.18	.28	.14	-
5	AUC	.69	.87	.79	.96	.88
	EER	.36	.18	.28	.14	.25

Table 7.4: Summary statistics for the quality of X-factor inferences, using a second-by-second evaluation. AUC denotes area under ROC curve and EER denotes the equal error rate. In model 1, the X-factor is used on its own to represent all non-normal variation. Ground truth for this model is considered to be the combined annotations for all factors. Known factors are added one by one in the lower rows. The results in the lowest row correspond to the experiment with Gaussian sum inference in Table 7.2 where the X-factor has also been included.

Model		X-factor	Incu. open	Core temp.	Blood sample	Bradycardia
1	AUC	.94	-	-	-	-
	EER	.13	-	-	-	-
2	AUC	.95	.95	-	-	-
	EER	.11	.10	-	-	-
3	AUC	.95	.96	.88	-	-
	EER	.11	.09	.18	-	-
4	AUC	.94	.96	.87	.99	-
	EER	.13	.09	.19	.01	-
5	AUC	.94	.97	.86	.99	.98
	EER	.14	.06	.22	.01	.03

Table 7.5: Summary statistics for the quality of X-factor inferences, using event-based analysis. See caption for Table 7.4.

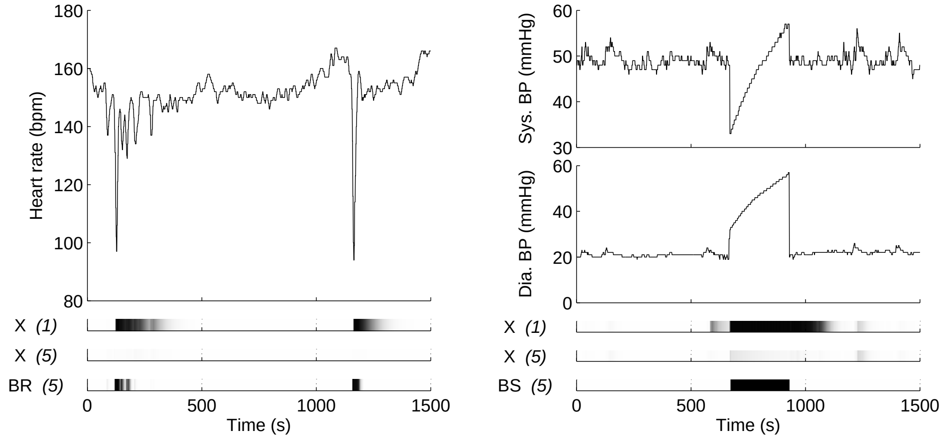


(a) (b)

Figure 7.13: Inferred switch settings for the X-factor, during periods of cardiovascular disturbance, compared to the gold standard annotations.

of the variation, incorporating the knowledge that the true physiology is not being observed, and so able to handle the discontinuity at $t = 900$ effectively.

Figure 7.15 shows examples of inferred switch settings from model (5) in which there are occurrences of both known and unknown types of variation. In 7.15(a) a bradycardia occurs in the middle of a period of elavated blood pressure and a deep drop in saturation. The bradycardia factor is active for a period which corresponds closely to the ground truth. The X-factor picks up the presence of a change dynamics at about the right time, but its onset is delayed when compared to the ground truth interval. This again highlights a difficulty with filtered inference, since at time just over 1000 it is difficult to tell that this is the beginning of a significant change in dynamics without the benefit of hindsight. In panel (b) a blood sample is correctly picked up by the blood sample factor, while a later period of physiological disturbance on the same measurement channels is correctly picked up by the X-factor. Panel (c) shows another example of the bradycardia factor operating with the X-factor, where this time the onset of the first bradycardia is before the onset of the X-factor. The X-factor picks up a desaturation, a common pattern which is already familiar from panel (a). In panel (d), an interaction between the X-factor and the ‘Incubator open’ factor can be seen. From time 270 to 1000 the incubator has been opened, and all variation including the spike in HR at $t = 420$ are attributed to handling of the baby. Once the incubator appears to have been closed, further physiological disturbance is no longer



(a) (b)

Figure 7.14: Inferred switch settings for the X-factor, and for known patterns for models (1) and (5) in Table 7.4. Panel (a) shows two instances of bradycardia (BR), (b) shows a blood sample (BS).

explained as an effect of handling and is picked up by the X-factor.

Having looked at examples of inferences, the operation of the X-factor can be understood further by analysing the pattern of errors which are made under different models. Figure 7.16 shows summed probability masses for the posteriors of different factor settings in two models: that in which only the known factors are present (model ‘GS’ in Table 7.2), and that in which all known factors plus the X-factor are present (model (5) in Table 7.4). For each factor, the posterior probabilities are summed over the periods in which the gold standard showed that factor to be inactive *and* in which there was abnormal variation. That is, the sum of the errors made while abnormal variation was occurring was obtained. Posterior probablities for the normal switch setting were summed over the periods in which the X-factor and any other factor was active. These are summed errors, so smaller is better. It can be seen that the mass of misclassified normality drops significantly with the introduction of the X-factor. The other factors are roughly the same in the two models (mass is slightly smaller for the bradycardia, temperature probe disconnection and blood sample factors), reflecting the similar performance in factor classification in Tables 7.2 and 7.4.

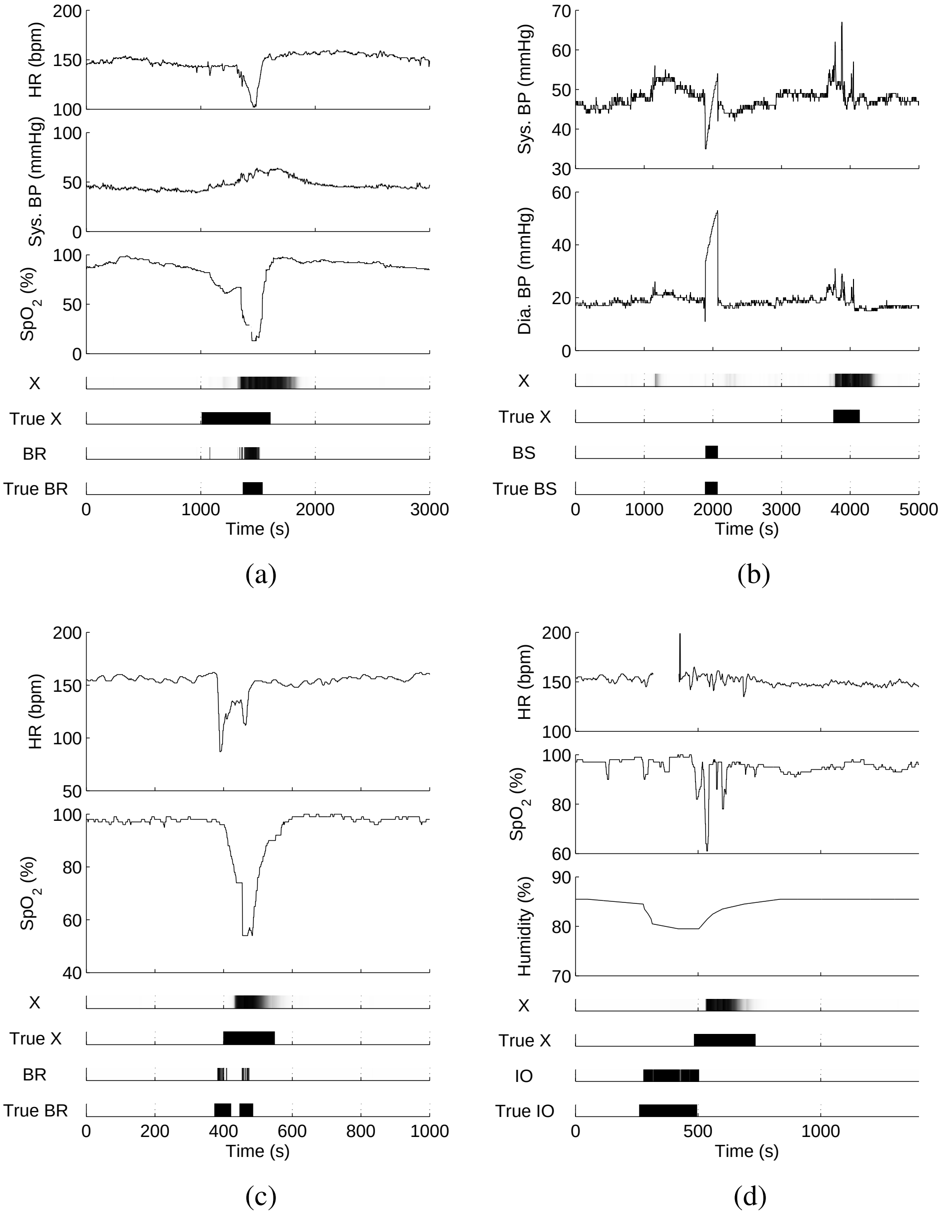


Figure 7.15: Inferred switch settings for the X-factor, in regions where other factors are active. BR denotes bradycardia, BS denotes a blood sample, and IO denotes opening of the incubator. In panel (a) a bradycardia occurs in conjunction with a rise in blood pressure and deep desaturation. The X-factor is triggered around the right region but is late compared to ground truth. In panel (b), unusual BP variation is correctly classified as being due to a blood sample and then of unkown cause. Panel (c) shows bradycardia with a desaturation picked up by the X-factor, and (d) shows the X-factor picking up disturbance after the incubator has been entered.

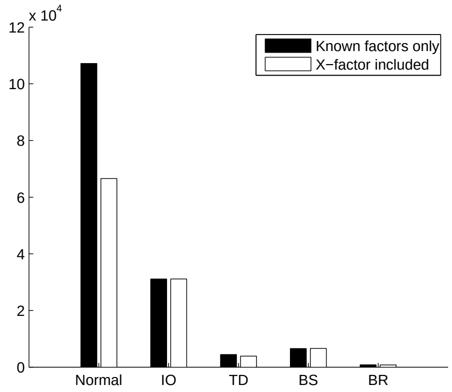


Figure 7.16: Summed probability masses for different factor settings, under models with and without the X-factor. See text for details.

7.3.2 Comparison of alternative X-factor constructions

It was discussed in section 6.3 that other constructions are possible as generative X-factor models. Two particular alternatives were described, firstly in which the observation covariance is inflated instead of the system noise covariance, and secondly in which a white noise process is used to model unusual dynamics.

These alternative models were implemented, and examples of inference are shown in Figure 7.17. In these examples, the setting $\xi = 1.2$ was again used for the standard X-factor. For the other two, different values were tried in order to get the best subjective results from these specific examples. For the inflated observation covariance model, the best value was found to be $\xi = 30$. For the white noise model, the most appropriate variance was found to be given by the variance of the observation channels multiplied by a factor of 50. The models were run with only the X-factor using just the observation channels shown in the figure. In 7.17(a), there is a period of significant disturbance in heart rate and oxygen saturation. This period is chosen as an ‘easy’ example, as the disturbance is highly pronounced. The onset and duration of the period annotated as being abnormal is indicated. The standard X-factor makes an inference which corresponds well to the annotation, though the alternative X-factor constructions are less satisfactory. The inflated observations model tends to respond to sudden discontinuities (e.g. erroneous measurements) rather than more subtle changes in dynamics, and here it can be seen to pick up discontinuities in the heart rate signal. It has been found difficult to produce meaningful inferences with the white noise X-factor, the dynamics being of a different form altogether to normal. In the bottom probability plot in panel (a), inferences from this model are characteristically erratic.

Figure 7.17(b) makes a further comparison between the usual (inflated system noise) X-factor and the inflated observation noise X-factor, for a period in which there is a smooth but marked change in both heart rate and peripheral temperature observations relating to handling of the baby. Again, the inferences from the standard X-factor roughly correlate with the indicated interval in which there are known to be handling effects. The inflated observation model picks up a very short section at around $t = 1600$, where there is another discontinuity in heart rate. Many different values of ξ were tried, though no setting could cause this model to claim any more observations than as shown.

Empirically, the inflated system noise covariance produces inferences which are closer to the desired behaviour than the two alternatives. Note that a factor with inflated observation noise covariance does potentially have value, but this is in detecting outlying measurements in order to ignore them. Consider again the point made in Figure 6.2, that the inflated system noise has the effect of making the model pay more attention to observations in times of uncertainty. Inflating the observation noise has the opposite effect—the signal-to-noise ratio in the model is effectively being lowered, so that there is less confidence in the observations; potentially a useful model in future work for spurious, transient observations which are not part of a longer-term change in dynamics.

7.4 Summary

The FSKF model has been shown to be capable of making useful inferences from noisy, multivariate NICU monitoring data. The experiments support three claims about the methods developed here: (1) the FSKF can be used to identify known artifactual and physiological patterns, and to provude estimates of the true physiology in the cases where artifacts obscure the measurements; (2) the X-factor can operate successfully in conjunction with the known factors, where the classification of the normal factors is preserved and areas in which there are significant deviations from normal dynamics are picked out; (3) the ability of the X-factor to represent many types of deviation from normality is further demonstrated by showing that it can ‘stand in’ for the known factors. When the appropriate known factor is added to the model, probability mass tends to shift to the correct factor.

It has also been demonstrated empirically that the X-factor construction used here

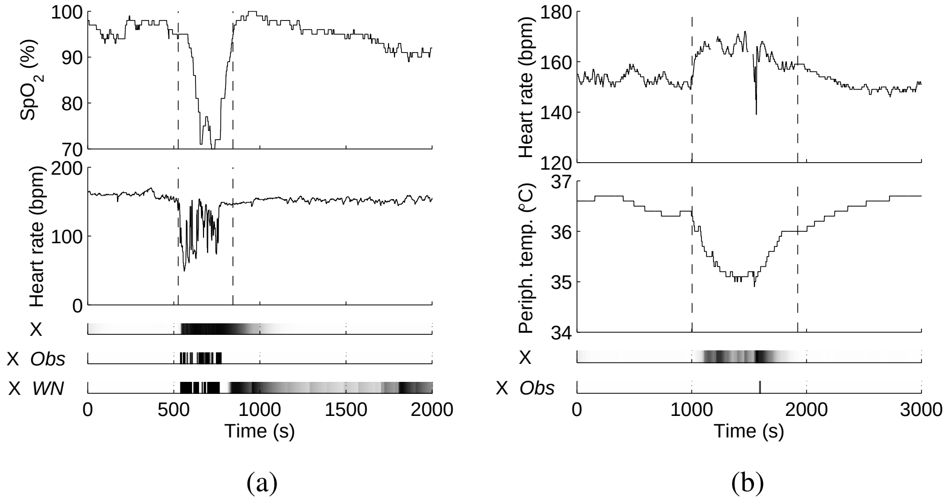


Figure 7.17: Comparisons of alternative X-factor constructions. ‘X’ denotes the usual X-factor construction, in which system noise covariance is inflated. ‘X *Obs*’ denotes the construction in which the observation noise covariance is inflated. ‘X *WN*’ denotes the use of a white-noise process for the X-factor. Dashed vertical lines in panel (a) show the gold standard interval for a physiological disturbance, and in panel (b) show the duration of an episode of handling.

is more appropriate to the task of finding clinically significant deviations in the dynamics of monitoring data then the alternative constructions discussed in section 6.3.

Chapter 8

Conclusions and recommendations for further work

This chapter summarises the main contributions of the thesis, discusses the practicality of real-time application to cotside monitoring, or monitoring with different data collection hardware, and suggests interesting extensions of the work which were beyond the current scope.

8.1 Contributions of this thesis

The main contributions have been as follows:

- The application of the FSKF and FHMM for inferring common factors in physiological monitoring data;
- Finding parameterisations of linear dynamical models to represent common physiological and artifactual conditions, and adapting parameter estimation and inference techniques for the purpose;
- The formulation of a model (the 'X-factor') for novel physiological dynamics, used to infer the times in which something is happening which is not described by any of the known patterns;
- The derivation of parameter estimation expressions for the X-factor model;
- The experimental demonstration that these methods are effective when applied to genuine monitoring data.

Novel material is also presented in Appendix D concerning the application of autoregressive models to data of different sampling frequencies.

8.2 Application to real-time monitoring

The experiments described have been conducted as offline analysis, but very little work would be required in order to operate the system in real time. The preprocessing stage, which lessens the effect of quantised humidity measurements, requires forward information. A real time implementation would therefore have to preprocess some time ahead (say 20 seconds, referring to Figure 5.5), and inference would be done on the delayed, preprocessed measurements. The inference routines at the core of the system perform batch processing (using `for` loops over time), but this is straightforward to recast as a function which takes one value at a time and returns posterior probabilities on the fly.

The Matlab implementation of the FSKF with a set of four known factors and the X-factor operates around 10 times faster than real time on a 3GHz PC.

8.3 Application to alternative monitoring hardware

In principle the system developed here is agnostic to the hardware which collects the data, assuming that the same vital signs can be collected at the same frequency. However, using different monitoring equipment could potentially change the dynamics of the measurements: for example, if a cotside ECG monitor was used which employed a different algorithm to calculate the heart rate. In practice the implementation of hardware interfaces and data collection software is likely to be the bulk of the work required to apply the methods in this thesis to a new setting.

8.4 Further work

There are a number of directions in which the work in this thesis could be continued. One natural extension is the development of additional factors for the model, for example to model pneumothorax (section 2.4.3), the effects of cold stress on temperature dynamics (Lyon et al., 1997), or other serious conditions such as intraventricular haemorrhage (Milligan, 1980).

The experiments with the X-factor have shown that there are a significant number of regimes in the data which have not yet been formally analysed. Future work might therefore also address the problem of learning what different regimes there are in the data, without supervision. One possibility would be to take all the variation flagged by the X-factor and trying to find regularities in it (either with further statistical analysis, or analysis by an expert, or a combination of the two), resulting in further explanatory factors being added to the model. Similar problems have been addressed before, for example in the context of discovering common patterns in genetic sequences (Frey et al., 2005). The latter work incorporated the idea of searching through the data for 'exemplars' of each pattern. This approach is potentially interesting for finding canonical examples of different patterns of physiological variation, but would require some adaptation in order to be used in the context of the SKF or FSKF. A more direct approach to this problem would be to try and learn the full SKF or FSKF model using likelihood-based methods without supervision. This could be done by applying the EM updates as given in Appendix C which do not require labelled training data, or using a method such as variational learning Ghahramani and Hinton (1998). Note that it would be possible to fix the known regimes and try to learn the unmodelled variation by updating parameters in additional factors. The X-factor could prove useful in this context in order to highlight the sequences of measurements which are inadequately represented by the current model, for example by using it as an initial parameter setting for an additional factor.

It would also be interesting to look more closely at the differences in dynamics between babies. It has been assumed in the current work that each baby has different baseline dynamics, but that the known patterns such as bradycardia have the same dynamics for every baby. The first assumption has meant that a section of stable data has to be marked up for each baby, which makes the model less convenient to put into practice. Future work could look at how to relax this assumption, perhaps replacing it with weaker assumptions (such as the monitoring data having to begin at a point where the baby is stable), or by beginning a monitoring session with a loose description of normal dynamics which the model then tries to update 'on the fly'. Note that the idea of 'normality' is somewhat subjective, so there may be problems in determining exactly what constitutes a normal sequence (c.f. the difficulty in obtaining ground truth for the X-factor in section 7.3.1). The second assumption, that physiological dynamics are independent of the baby conditional on a 'known' switch setting (e.g. that all bradycardia have the same dynamics) is not entirely accurate, as there is always a degree of idiosyncracy in different babies. Improving the model in this respect therefore involves analysing the manner in which dynamics can change between babies, and then how to account for this in the model. Previous work on covariate shift, e.g. Sugiyama and Müller (2005) and Storkey and Sugiyama (2007), might be useful in this regard for obtaining estimates of generalisation error.

Another approach to handling the differences between babies might be to see if there are covariates that affect the normal dynamics which could be used to predict model parameters. For example, knowing the gestation of a baby might be useful for getting a better idea of what the normal dynamical regime is likely to be without needing to see any training data from that particular baby. Other covariates might be useful for predicting model parameters, for example the presence or absence of particular pathologies, use of mechanical ventilation and so forth.

Some other suggestions here fall into three broad categories: methods to improve the quality and speed of inference, ways of modelling long-term changes in babies' physiology, and extensions of the X-factor.

Inferences would be likely to improve with fixed-lag smoothing. Some of the discrepancies between inferences in the X-factor and the gold standard annotation were due to variation which can be seen to be the start of a significant pattern in hindsight, but with knowledge only of past observations look like random variation. This also applies to inference of factors such as the opening of the incubator in which the divergence from normal dynamics may be subtle at first. Various authors have examined techniques for smoothing in a switching state-space model (Ghahramani and Hinton, 1998; Barber and Mesot, 2006; Klaas et al., 2006).

Further investigation could be made into speeding up the inference. One scheme which would lessen the time for inference would be to separate the inference of different factors. Recall the trick used for inference described in section 5.10.2, where the next observation is checked to see whether it equals zero or not — if it is, then the dropout dynamics are swapped in automatically. This could be done with the blood sample factor as well (for example). The factor is inferred on its own from blood pressure data, and the posteriors thresholded to give binary estimates of whether there is a blood sample or not at each time step. Inference can then be run for a model with all the other factors, with blood pressure switch settings clamped.

Acute, short term changes in a baby's physiological function have mainly been studied in this thesis, but there can also be changes on a longer time scale. What has been considered the stationary 'normal' dynamical regime is likely to change with age. One of the consequences of this is that in order to monitor a baby over a long time, it may sometimes be appropriate to re-learn the parameters for normal dynamics. The

X-factor could be helpful for this. For example, if a situation occurs in which the X-factor has a high posterior probability for an hour or more, it may be appropriate to consider the current dynamics the ‘new normal’, and re-train. There is a tradeoff in doing this as to how ready we are to reset the model for normal dynamics. If we are too reluctant to retrain, the model is liable to try and make inferences with an inapropriate description of baseline dynamics; too quick to retrain, and there is a danger of any type of variation being seen as normal — an example of a plasticity-stability dilemma. This type of approach to retraining normal dynamics could also be another way to circumvent the need to define training data for normal dynamics for each baby.

Various extensions of the X-factor are possible. First, the single parameter ξ could be extended to the set $\xi_1, \ldots, \xi_D$ in order to model different amounts of uncertainty on each observed channel. It is straightforward to remove the sum over all observed channels from (6.18), and instead calculate the parameter updates with respect to each observed channel independently.

It may also be worth investigating the use of numerical optimisation techniques such as scaled conjugate gradient search for estimation of X-factor parameters. When looking at ways of estimating the observation matrix $\mathbf{R}$ in a switching Kalman filter, Shumway and Stoffer (1991) reported improved results by using a few steps of EM followed by gradient-based search.

Appendix A

Matlab code and data format

The main aspects of the original code used for this research are outlined here. Many other helper functions also exist which are not described. The code and data can be downloaded from the following address:
`http://homepages.inf.ed.ac.uk/s0348608/downloads/thesis_code.tar.gz`.

A.1 Top level scripts

`chooseexperiment`	Select an experiment from list and run it. Includes short demonstrations and long runs, e.g. producing Table 7.2.
`runexperiment`	Rerun the last experiment.
`runxfactorexperiments`	Run the experiments which generate Table 7.4.
`zoomplot`	Show panning, zooming chart with inferences and gold standard annotation.

Experiment settings files are all held in the same directory, with the prefix '`exp_`'. These files can be edited or copied in order to add or remove factors, specify different training/test babies, usage of cross-validation, whether to save inferences to disk, whether to use EM etc.

Ancillary scripts

`factors2statespace`	Full switching state-space model from individual factors.
`factors2meancov`	Full FHMM model from individual factors.
`globalconstants`	IDs of observed channel names, factors and inference types.
`globalsettings`	Global settings for all experiments.
`preprocessing_and_setup`	Preprocess data and set up factorial model structure.
`training_and_inference`	Train the model, make inferences and evaluate.
`train_*`	Fit models for normal and factor dynamics.

A.2 Statistical functions

`ar2statespace`	Convert AR process to state space form (section 4.5).
`arima2ar`	Integrated AR process to AR (section 4.1).
`arspectrum`	Power spectrum of an AR process.
`aryw`	Learn AR coefficients using Yule-Walker equations.
`autocov`	Autocovariance of a univariate sequence.
`chainindex2state`	Cross-product index of a switch setting from factor settings.
`convertdiscontinuities`	Correct quantisation effects (section 5.6).
`factoriseposteriors`	Factorise cross-product inferences.
`fhmmexact`	FHMM inference.
`hiddenarima`	Learn hidden ARIMA model (section 5.3.1.2).
`hiddenrelativear`	Learn state space using method in section 5.7.
`kalmanll`	Likelihood and innovations of a Kalman filter given data.
`kfdd`	Kalman Filter Diagnostic Dashboard (section 5.4.1).
`learnxf`	Update X-factor parameters with EM (section 6.2.2.3).
`mavg`	Moving average with different window functions.
`pacf`	Partial autocorrelation of a univariate sequence.
`rocintervals`	Calculate ROC statistics given ground truth.
`samplestatespace`	Draw a sample from a state space model.
`skf_adf`	SKF inference with Gaussian sum approximation.
`skf_rbpf`	SKF inference with Rao-Blackwellised particle filtering.
`stateindex2chain`	Factor settings from cross-product switch setting index.

A.3 Data format

The data used in the experiments in Chapter 7 is contained in the file `15days.mat`. The struct array `data` has two fields, `raw` and `preprocessed`, each of which is a cell array with elements representing each of the the 15 babies. These elements are also struct arrays, with fields containing the raw physiological data for each baby, and other information such as gestation and anonymised identifiers.

The struct array `intervals` contains annotations provided by the clinical experts. For example, `intervals.BloodSample{3}` contains an array of times for which a blood sample was thought to have occurred for baby 3. This is an $n \times 2$ matrix in which each row represents `[start_index stop_index]` for a particular episode of blood sampling. Indices are relative to the start of the 24 hour monitoring period.

Appendix B

Approximate Inference Methods for the SKF

B.1 Gaussian sum approximation

Figure B.1 shows the algorithm for inference in a switching linear dynamical model with the Gaussian sum approximation (Alspach and Sorenson, 1972). This algorithm makes inference tractable by collapsing K^2 Gaussian mixture components into K mixture components at each time step, considering the likelihood of each mixture component. For further details on the collapsing step, see Murphy (1998).

B.2 Rao-Blackwellised particle filtering

Figure B.2 shows the algorithm for Rao-Blackwellised particle filtering (Doucet et al., 2000). Inference is also made tractable here using a mixture of Gaussians, but in this case the mixture components are chosen stochastically. This method is an improvement over standard particle filtering as it exploits the structure of the SKF. When switch states have been sampled, the density estimates for each particle can be updated using the Kalman filter prediction equations (since these do not depend on the most recent observation). Importance weights are then calculated for each particle, degeneracy is avoided by discarding particles with low importance weights, and the Kalman filter update equations are then used to refine the estimates.

Assumes initial estimates, $\hat{\mathbf{x}}_0$ and $\mathbf{P}_0$.

To calculate $\{\hat{\mathbf{x}}_t,\ \mathbf{P}_t\}$ from $\{\hat{\mathbf{x}}_{t-1},\ \mathbf{P}_{t-1}, \mathbf{y}_{1:t}\}$:

1. **Kalman updates and prediction**

   - For each $s_{t-1} = i,\ s_t = j,\ (\{i, j\} = 1, \ldots, K)$,

$$
\begin{aligned}
\hat{\mathbf{x}}^{-}_{t|i,j} &= \mathbf{A}^{\{j\}}\hat{\mathbf{x}}_{t-1|i} \\
\mathbf{P}^{-}_{t|i,j} &= \mathbf{A}^{\{j\}}\mathbf{P}_{t-1|i}\mathbf{A}^{\{j\}\top} + \mathbf{Q}^{\{j\}} \\
\tilde{\mathbf{y}}_{t|i,j} &= \mathbf{y}_t - \mathbf{C}\hat{\mathbf{x}}^{-}_{t|i,j} \\
\mathbf{S}_{t|i,j} &= \mathbf{C}^{\{j\}}\mathbf{P}^{-}_{t|i,j}\mathbf{C}^{\{j\}\top} + \mathbf{R}^{\{j\}} \\
\mathbf{K}_{t|i,j} &= \mathbf{P}^{-}_{t|i,j}\mathbf{C}^{\{j\}\top}\mathbf{S}^{-1}_{t|i,j} \\
\hat{\mathbf{x}}_{t|i,j} &= \hat{\mathbf{x}}^{-}_{t|i,j} + \mathbf{K}_{t|i,j}\tilde{\mathbf{y}}_t \\
\mathbf{P}_{t|i,j} &= (\mathbf{I} - \mathbf{K}_{t|i,j}\mathbf{C}^{\{j\}})\mathbf{P}^{-}_{t|i,j}
\end{aligned}
$$

2. **Calculate likelihood of each state transition** $\{i, j\}$

$$
p(\mathbf{y}_t | s_{t-1} = i, s_t = j) \quad \sim \quad \mathcal{N}\left(\mathbf{y}_t; \tilde{\mathbf{y}}_{t|i,j}, \mathbf{S}_{t|i,j}\right)
$$

3. **Calculate weights for each state transition**

$$
w_{ij} \quad = \quad \frac{p(\mathbf{y}_t | s_{t-1} = i, s_t = j)}{\sum_{k=1}^{K} p(\mathbf{y}_t | s_{t-1} = k, s_t = j)}
$$

4. **Collapse K^2 mixture components into K**

$$
\begin{aligned}
\hat{\mathbf{x}}_{t|j} &= \sum_{i=1}^{K} w_{ij}\hat{\mathbf{x}}_{t|i,j} \\
\mathbf{P}_{t|j} &= \sum_{i=1}^{K} w_{ij}\left(\mathbf{P}_{t|ij} + (\mathbf{x}_{t|ij} - \hat{\mathbf{x}}_{t|j})\right)
\end{aligned}
$$

Figure B.1: Algorithm for SKF inference with a Gaussian sum approximation.

Assumes initial estimates, $\hat{\mathbf{x}}_0$ and $\mathbf{P}_0$. Inference for N particles proceeds as follows:

1. **Sequential importance sampling**

   - For each particle $i = 1, \dots, N$, sample the switch setting and update predicted densities accordingly:

$$\begin{aligned}
\hat{s}_t^{(i)} &\sim p(s_t | \hat{s}_{t-1}^{(i)}) \\
\hat{\mathbf{x}}_t^{(i)-} &= \mathbf{A}^{(\hat{s}_t^{(i)})} \hat{\mathbf{x}}_{t-1}^{(i)} \\
\mathbf{P}_t^{(i)-} &= \mathbf{A}^{(\hat{s}_t^{(i)})} \mathbf{P}_{t-1}^{(i)} \mathbf{A}^{(\hat{s}_t^{(i)})\top} + \mathbf{Q}^{(\hat{s}_t^{(i)})}
\end{aligned}$$

   - For each particle $i = 1, \dots, N$, calculate importance weight:

$$w_t^{(i)} \propto p(\mathbf{y}_t | \hat{\mathbf{x}}_t^{(i)-}, \mathbf{P}_t^{(i)-}, \hat{s}_t^{(i)})$$

2. **Particle selection**

   - Multiply particles with high importance weights $w_t^{(i)}$ and discard particles with low importance weights.

3. **Kalman updates**

   - For each particle $i = 1, \dots, N$, use the Kalman filter update equations to refine estimates:

$$\begin{aligned}
\tilde{\mathbf{y}}_t^{(i)} &= \mathbf{y}_t - \mathbf{C}^{(\hat{s}_t^{(i)})} \hat{\mathbf{x}}_t^{(i)-} \\
\mathbf{S}_t^{(i)} &= \mathbf{C}^{(\hat{s}_t^{(i)})} \mathbf{P}_t^{(i)-} \mathbf{C}^{(\hat{s}_t^{(i)})\top} + \mathbf{R}^{(\hat{s}_t^{(i)})} \\
\mathbf{K}_t^{(i)} &= \mathbf{P}_t^{(i)-} \mathbf{C}^{(\hat{s}_t^{(i)})^\top} \mathbf{S}_t^{(i)-1} \\
\hat{\mathbf{x}}_t^{(i)} &= \hat{\mathbf{x}}_t^{(i)-} + \mathbf{K}_t^{(i)} \tilde{\mathbf{y}}_t^{(i)} \\
\mathbf{P}_t^{(i)} &= (\mathbf{I} - \mathbf{K}_t^{(i)} \mathbf{C}^{(\hat{s}_t^{(i)})}) \mathbf{P}_t^{(i)-}
\end{aligned}$$

Figure B.2: Algorithm for SKF inference with Rao-Blackwellised particle filtering.

Appendix C

EM derivations for linear dynamical models

This appendix shows how the EM algorithm can be applied to the Kalman filter and the switching Kalman filter under different conditions. EM expressions for the Kalman filter can also be found in Ghahramani and Hinton (1996).

C.1 Kalman filter

The Kalman filter model is factorised as follows:

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}_1)\prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1})\prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t)\ , \tag{C.1}$$

a product of Gaussian densities defined by the parameters $\{\mu,\Sigma,\mathbf{A},\mathbf{C},\mathbf{Q},\mathbf{R}\}$. In EM we try to maximise an auxiliary function, which is set here to the *expected complete data log likelihood*:

$$Q = \int p(\mathbf{x}|\mathbf{y},\theta_{\text{old}})\log p(\mathbf{x},\mathbf{y}|\theta)\,\mathrm{d}\mathbf{x}\ , \tag{C.2}$$

where $\mathbf{x},\mathbf{y}$ with no subscript denote the sequences $\mathbf{x}_{1:T}$,$\mathbf{y}_{1:T}$. The auxiliary function Q can be expanded as follows:

$$\begin{aligned} Q = \int p(\mathbf{x}|\mathbf{y},\theta_{\text{old}})\big[&(d_x+d_y)T\log(2\pi)+\log\det\Sigma+(T-1)\log\det\mathbf{Q}+T\log\det\mathbf{R}] \\ &+(x_1-\mu)^\top\Sigma^{-1}(x_1-\mu)+\sum_{t=2}^{T}(\mathbf{x}_t-\mathbf{A}\mathbf{x}_{t-1})^\top\mathbf{Q}^{-1}(\mathbf{x}_t-\mathbf{A}\mathbf{x}_{t-1}) \\ &+\sum_{t=1}^{T}(\mathbf{y}_t-\mathbf{C}\mathbf{x}_t)^\top\mathbf{R}^{-1}(\mathbf{y}_t-\mathbf{C}\mathbf{x}_t)\big]\,\mathrm{d}\mathbf{x}\ . \end{aligned} \tag{C.3}$$

In this expression, d_x and d_y are the dimensionalities of $\mathbf{x}_t$ and $\mathbf{y}_t$. To maximise Q, the updates to each parameter are found by taking partial derivatives of this function and setting to zero. The notation $\langle\cdot\rangle_{\mathbf{x}|\mathbf{y}} = \int p(\mathbf{x}|\mathbf{y})\cdot \mathrm{d}\mathbf{x}$ and $\hat{\mathbf{x}}_t = \langle \mathbf{x}_t\rangle_{\mathbf{x}|\mathbf{y}}$ is used in the following expressions. Note that $\langle \mathbf{x}_t\mathbf{x}_t^\top\rangle_{\mathbf{x}|\mathbf{y}}$ and $\hat{\mathbf{x}}_t$ are the values calculated by the Kalman filter smoothing equations.

$\langle \mathbf{x}_t\mathbf{x}_{t-1}^\top\rangle_{\mathbf{x}|\mathbf{y}}$ is known as the lag-one covariance smoother. See Shumway and Stoffer (2000) for a derivation.

Updates for μ:

$$\begin{aligned}\frac{\partial Q}{\partial \mu} &= -\frac{1}{2}E\left(\Sigma^{-1}(\mu-\mathbf{x}_1)\right)\\ \tilde{\mu} &= \hat{\mathbf{x}}_1\ .\end{aligned} \tag{C.4}$$

Updates for Σ:

$$\begin{aligned}\frac{\partial Q}{\partial \Sigma} &= -\frac{1}{2}E\left(\Sigma-\mathbf{x}_1\mathbf{x}_1^\top-\mathbf{x}_1\mu^\top-\mu\mathbf{x}_1^\top+\mu\mu^\top\right)\\ &= -\frac{1}{2}\left(\Sigma-\left\langle \mathbf{x}_t\mathbf{x}_t^\top\right\rangle_{\mathbf{x}|\mathbf{y}}+\tilde{\mu}\tilde{\mu}^\top)\right)\\ \tilde{\Sigma} &= \left\langle \mathbf{x}_1\mathbf{x}_1^\top)\right\rangle_{\mathbf{x}|\mathbf{y}}-\tilde{\mu}\tilde{\mu}^\top\ .\end{aligned} \tag{C.5}$$

Updates for $\mathbf{A}$:

$$\begin{aligned}\frac{\partial Q}{\partial \mathbf{A}} &= -\frac{1}{2}E\left(\sum_{t=2}^{T}\mathbf{Q}^{-1}\left(\mathbf{x}_t-\mathbf{A}\mathbf{x}_{t-1}\right)\mathbf{x}_{t-1}\right)\\ \tilde{\mathbf{A}} &= \left[\sum_{t=2}^{T}\left\langle \mathbf{x}_t\mathbf{x}_{t-1}^\top\right\rangle_{\mathbf{x}|\mathbf{y}}\right]\left[\sum_{t=2}^{T}\left\langle \mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top\right\rangle_{\mathbf{x}|\mathbf{y}}\right]^{-1}\ .\end{aligned} \tag{C.6}$$

Updates for $\mathbf{Q}$:

$$\begin{aligned}\frac{\partial Q}{\partial \mathbf{Q}^{-1}} &= -\frac{1}{2}E\left(\mathbf{Q}+\sum_{t=2}^{T}(\mathbf{x}_t-\mathbf{A}\mathbf{x}_{t-1})(\mathbf{x}_t-\mathbf{A}\mathbf{x}_{t-1})^\top\right)\\ &= \frac{T-1}{2}\mathbf{Q}-\frac{1}{2}\sum_{t=2}^{T}\left\langle \mathbf{x}_t\mathbf{x}_t^\top\right\rangle_{\mathbf{x}|\mathbf{y}}-\left\langle \mathbf{x}_t\mathbf{x}_{t-1}^\top\right\rangle_{\mathbf{x}|\mathbf{y}}\mathbf{A}^\top-\mathbf{A}\left\langle \mathbf{x}_{t-1}\mathbf{x}_t^\top\right\rangle_{\mathbf{x}|\mathbf{y}}+\mathbf{A}\left\langle \mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top\right\rangle_{\mathbf{x}|\mathbf{y}}\mathbf{A}^\top\\ \tilde{\mathbf{Q}} &= \frac{1}{T-1}\sum_{t=2}^{T}\left\langle \mathbf{x}_t\mathbf{x}_t^\top\right\rangle_{\mathbf{x}|\mathbf{y}}-\tilde{\mathbf{A}}\left\langle \mathbf{x}_{t-1}\mathbf{x}_t^\top\right\rangle_{\mathbf{x}|\mathbf{y}}\ .\end{aligned} \tag{C.7}$$

Updates for $\mathbf{C}$:

$$\begin{aligned}
\frac{\partial Q}{\partial \mathbf{C}} &= -\frac{1}{2}E\left(\sum_{t=1}^{T}\mathbf{R}^{-1}(\mathbf{y}_t-\mathbf{C}\mathbf{x}_t)\mathbf{x}_t^{\top}\right)\\
&= -\frac{1}{2}\sum_{t=1}^{T}\mathbf{R}^{-1}\left[\mathbf{y}_t\hat{\mathbf{x}}_t^{\top}-\mathbf{C}\left\langle\mathbf{x}_t\mathbf{x}_t^{\top}\right\rangle_{\mathbf{x}|\mathbf{y}}\right]\\
\tilde{\mathbf{C}} &= \left[\sum_{t=1}^{T}\mathbf{y}_t\hat{\mathbf{x}}_t^{\top}\right]\left[\sum_{t=1}^{T}\left\langle\mathbf{x}_t\mathbf{x}_t^{\top}\right\rangle_{\mathbf{x}|\mathbf{y}}\right]^{-1}\;. \qquad \text{(C.8)}
\end{aligned}$$

Updates for $\mathbf{R}$:

$$\begin{aligned}
\frac{\partial Q}{\partial \mathbf{R}^{-1}} &= -\frac{1}{2}E\left(T\mathbf{R}-\sum_{t=2}^{T}\mathbf{y}_t\mathbf{y}_t^{\top}-\mathbf{y}_t\mathbf{x}_t^{\top}\mathbf{C}^{\top}-\mathbf{C}\mathbf{x}_t^{\top}\mathbf{y}_t+\mathbf{C}\mathbf{x}_t\mathbf{x}_t^{\top}\mathbf{C}^{\top}\right)\\
\tilde{\mathbf{R}} &= \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t^{\top}-\tilde{\mathbf{C}}\hat{\mathbf{x}}_t\mathbf{y}_t^{\top}\;. \qquad \text{(C.9)}
\end{aligned}$$

C.2 Switching Kalman filter (switch settings known)

This section describes how the EM algorithm can be applied to switching linear dynamics. The first case considered is that in which the switch settings are available in the training data, and the second case in which they are unobserved.

Where the switch settings are observed, an expected likelihood is taken with regards to $\mathbf{x}$ as before, using the factorisation of the SKF:

$$p(\mathbf{x},\mathbf{y},\mathbf{s}) = p(s_1)p(\mathbf{x}_1|s_1)\prod_{t=2}^{T}p(s_t|s_{t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1},s_t)\prod_{t=1}^{T}p(\mathbf{y}_t|\mathbf{x}_t,s_t)\;, \qquad \text{(C.10)}$$

such that the expected complete data likelihood is given by:

$$\begin{aligned}
Q = -\frac{1}{2}\int p(\mathbf{x}|\mathbf{y},\mathbf{s},\theta_{\text{old}})\big[&(d_x+d_y)T\log(2\pi)+\log\det\Sigma|+\sum_{t=2}^{T}\log\det\mathbf{Q}^{(s_t)}+\sum_{t=1}^{T}\log\det\mathbf{R}^{(s_t)}\\
&+(x_1-\mu)^{\top}\Sigma^{-1}(x_1-\mu)+\sum_{t=2}^{T}(\mathbf{x}_t-\mathbf{A}^{(s_t)}\mathbf{x}_{t-1})^{\top}{\mathbf{Q}^{(s_t)}}^{-1}(\mathbf{x}_t-\mathbf{A}^{(s_t)}\mathbf{x}_{t-1}) \qquad \text{(C.11)}\\
&+\sum_{t=1}^{T}(\mathbf{y}_t-\mathbf{C}^{(s_t)}\mathbf{x}_t)^{\top}{\mathbf{R}^{(s_t)}}^{-1}(\mathbf{y}_t-\mathbf{C}^{(s_t)}\mathbf{x}_t)\big]\,\mathrm{d}\mathbf{x}\;.
\end{aligned}$$

The parameter updates in this case are analogous to (C.4–C.8), but not directly equivalent since we are now taking into account the effects of transitioning from one switch setting into another. Only the update expressions are given here, since the intermediate steps are similar to those in the previous section.

$$\tilde{\mathbf{A}}^{(i)} = \left[\sum_{t=2}^{T} I(s_t=i)\left\langle \mathbf{x}_t\mathbf{x}_{t-1}^\top\right\rangle_{\mathbf{x}|\mathbf{s},\mathbf{y}}\right]\left[\sum_{t=2}^{T} I(s_t=i)\left\langle \mathbf{x}_{t-1}\mathbf{x}_{t-1}^\top\right\rangle_{\mathbf{x}|\mathbf{s},\mathbf{y}}\right]^{-1} \tag{C.12}$$

$$\tilde{\mathbf{Q}}^{(i)} = \frac{1}{T_{i,t\geq 2}}\sum_{t=2}^{T} I(s_t=i)\left[\left\langle \mathbf{x}_t\mathbf{x}_t^\top\right\rangle_{\mathbf{x}|\mathbf{s},\mathbf{y}} - \tilde{\mathbf{A}}^{(s_t)}\left\langle \mathbf{x}_{t-1}\mathbf{x}_t^\top\right\rangle_{\mathbf{x}|\mathbf{s},\mathbf{y}}\right] \tag{C.13}$$

$$\tilde{\mathbf{C}}^{(i)} = \left[\sum_{t=1}^{T} I(s_t=i)\mathbf{y}_t\hat{\mathbf{x}}_t^\top\right]\left[\sum_{t=1}^{T} I(s_t=i)\left\langle \mathbf{x}_t\mathbf{x}_t^\top\right\rangle_{\mathbf{x}|\mathbf{s},\mathbf{y}}\right]^{-1} \tag{C.14}$$

$$\tilde{\mathbf{R}}^{(i)} = \frac{1}{T_i}\sum_{t=1}^{T} I(s_t=i)\left[\mathbf{y}_t\mathbf{y}_t^\top - \tilde{\mathbf{C}}^{(s_t)}\hat{\mathbf{x}}_t\mathbf{y}_t^\top\right] \tag{C.15}$$

In these expressions, I is the indicator function, and T_i denotes the number of time steps in the training data for which $s_t = i$. Since $\mathbf{s}$ is observed, the notation $\langle\cdot\rangle_{\mathbf{x}|\mathbf{s},\mathbf{y}}$ is now used to denote the expectation $\int p(\mathbf{x}|\mathbf{y},\mathbf{s})\cdot \mathrm{d}\mathbf{x}$. Estimates for the initial state and covariance $\{\mu,\Sigma\}$ remain the same as for the Kalman filter.

C.3 Switching Kalman filter (switch settings unknown)

If the switch settings $\mathbf{s} = s_{1:t}$ are not available in the training data, the expected likelihood can be calculated by integrating over $\mathbf{x}$ and summing over each possible setting of s_t. In this case the expected complete data log likelihood will have the following form:

$$Q = \sum_{\mathbf{s}}\int p(\mathbf{x},\mathbf{s}|\mathbf{y},\theta_{\text{old}})\log p(\mathbf{x},\mathbf{y},\mathbf{s})\,\mathrm{d}\mathbf{x}\;. \tag{C.16}$$

This can be expanded to the following:

$$\begin{aligned} Q = -\frac{1}{2}\sum_{\mathbf{s}}\int & p(\mathbf{x},\mathbf{s}|\mathbf{y},\theta_{\text{old}})\big[(d_x+d_y)T\log(2\pi)+\log\det\Sigma \\ &+\sum_{t=2}^{T}\log\det\mathbf{Q}^{(s_t)}+\sum_{t=1}^{T}\log\det\mathbf{R}^{(s_t)}+\log\left(p(s_1)\right)+\sum_{t=2}^{T}\log Z_{s_t,s_{t-1}} \\ &+(x_1-\mu)^\top\Sigma^{-1}(x_1-\mu)+\sum_{t=2}^{T}(\mathbf{x}_t-\mathbf{A}^{(s_t)}\mathbf{x}_{t-1})^\top{\mathbf{Q}^{(s_t)}}^{-1}(\mathbf{x}_t-\mathbf{A}^{(s_t)}\mathbf{x}_{t-1}) \\ &+\sum_{t=1}^{T}(\mathbf{y}_t-\mathbf{C}^{(s_t)}\mathbf{x}_t)^\top{\mathbf{R}^{(s_t)}}^{-1}(\mathbf{y}_t-\mathbf{C}^{(s_t)}\mathbf{x}_t)\big]\,\mathrm{d}\mathbf{x}\;, \end{aligned} \tag{C.17}$$

where Z_{ij} denotes the prior switching probability $p(s_t = i|s_{t-1} = j)$. The updates for this case are in terms of expectations $\langle\cdot\rangle_{\mathbf{x},\mathbf{s}|\mathbf{y}} = \sum_{\mathbf{s}}\int p(\mathbf{x},\mathbf{s}|\mathbf{y})\cdot\mathrm{d}\mathbf{x}$. This is not tractable to compute in general since the number of terms in the sum over $\mathbf{s}$ increase exponentially with the length of the training data sequence. A smoothed approximation could be obtained e.g. by particle smoothing (Klaas et al., 2006), though an easier approach is to calculate the filtered expectations instead. After partial differentiation of Q with respect to each parameter, some straightforward algebra gives the following updates:

$$\tilde{\mathbf{A}}^{(i)} = \left[\sum_{t=2}^{T} p(s_t = i|\mathbf{y})\langle \mathbf{x}_t \mathbf{x}_{t-1}^{\top}\rangle_{\mathbf{x},\mathbf{s}|\mathbf{y}}\right]\left[\sum_{t=2}^{T} p(s_t = i|\mathbf{y})\left\langle \mathbf{x}_{t-1}\mathbf{x}_{t-1}^{\top}\right\rangle_{\mathbf{x},\mathbf{s}|\mathbf{y}}\right]^{-1} \quad \text{(C.18)}$$

$$\tilde{\mathbf{Q}}^{(i)} = \frac{\sum_{t=2}^{T} p(s_t = i|\mathbf{y})\left(\left\langle \mathbf{x}_t \mathbf{x}_t^{\top}\right\rangle_{\mathbf{x},\mathbf{s}|\mathbf{y}} - \tilde{\mathbf{A}}^{(s_t)}\left\langle \mathbf{x}_{t-1}\mathbf{x}_t^{\top}\right\rangle_{\mathbf{x},\mathbf{s}|\mathbf{y}}\right)}{\sum_{t=2}^{T} p(s_t = i|\mathbf{y})} \quad \text{(C.19)}$$

$$\tilde{\mathbf{C}}^{(i)} = \left[\sum_{t=1}^{T} p(s_t = i|\mathbf{y})\mathbf{y}_t\hat{\mathbf{x}}_t^{\top}\right]\left[\sum_{t=1}^{T} p(s_t = i|\mathbf{y})\left\langle \mathbf{x}_t\mathbf{x}_t^{\top}\right\rangle_{\mathbf{x},\mathbf{s}|\mathbf{y}}\right]^{-1} \quad \text{(C.20)}$$

$$\tilde{\mathbf{R}}^{(i)} = \frac{\sum_{t=1}^{T} I(s_t = i)\left(\mathbf{y}_t\mathbf{y}_t^{\top} - \tilde{\mathbf{C}}^{(s_t)}\hat{\mathbf{x}}_t\mathbf{y}_t^{\top}\right)}{\sum_{t=1}^{T} p(s_t = i|\mathbf{y})} \quad \text{(C.21)}$$

The reestimation expression for $\tilde{\mathbf{R}}^{(i)}$ is also given by Shumway and Stoffer (1991). Variational approximations to EM are possible in this case, see Ghahramani and Hinton (1998).

C.4 Switching Kalman filter with X-factor

This section demonstrates how EM can be performed for the case in which the switch settings are observed for known regimes but not for the X-factor. The goal here is to learn the single scalar parameter ξ. The auxiliary function Q is still given by (C.17), with $A^{(*)} = A^{(1)}$ and $Q^{(*)} = \xi Q^{(1)}$. The function is optimised with respect to ξ as follows:

$$\frac{\partial Q}{\partial \xi} = \frac{1}{2}\left\langle \sum_{t=2}^{T}\xi - \sum_{t=2}^{T}(\mathbf{x}_t - \mathbf{A}^{(1)}\mathbf{x}_{t-1})^{\top}\frac{1}{\xi}\mathbf{Q}^{(1)^{-1}}(\mathbf{x}_t - \mathbf{A}^{(1)}\mathbf{x}_{t-1})\right\rangle_{\mathbf{x},\mathbf{s}|\mathbf{y}} \quad \text{(C.22)}$$

Setting this partial derivative to zero implies that

$$\sum_{t=2}^{T}\frac{1}{\xi}p(s_t = *|\mathbf{y}) = \frac{1}{\xi^2}\sum_{t=2}^{T}(\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1})^{\top}\mathbf{Q}^{(1)^{-1}}(\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1})p(s_t = *|\mathbf{y}) \ . \quad \text{(C.23)}$$

The update to ξ is therefore given by

$$\tilde{\xi} = \frac{\sum_{t=2}^{T}(\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1})^{\top}\mathbf{Q}^{(1)^{-1}}(\hat{\mathbf{x}}_t - \mathbf{A}^{(1)}\hat{\mathbf{x}}_{t-1})p(s_t = *|\mathbf{y})}{\sum_{t=2}^{T} p(s_t = *|\mathbf{y})} \ . \quad \text{(C.24)}$$

Intuitively, this update expression calculates a Z-score, considering the covariance of novel points and the covariance of the normal regime. Every point is considered, and is weighted by the probability of having been generated by the X-factor regime.

C.5 Factorial Switching Kalman filter with X-factor

From section 6.2.2.3, we have the notation $o(f_*, s, c)$ which indicates whether the X-factor overwrites state dimensions d_c in switch setting s. The terms in the auxiliary equation Q which include the parameter ξ can now be written down as

$$Q_\xi = \left\langle \sum_{t=2}^{T} \sum_s \sum_c o(f_*, s, c) \left(\log \det \xi \mathbf{Q}_{d_c}^{(1)} - \frac{1}{\xi} a(t,c) \right) \right\rangle_{\mathbf{x},\mathbf{s}|\mathbf{y}} \tag{C.25}$$

where

$$a(t,c) = \left(\mathbf{x}_{\{t,d_c\}} - \mathbf{A}_{d_c}^{(1)} \mathbf{x}_{\{t-1,d_c\}} \right)^\top \mathbf{Q}_{d_c}^{(1)^{-1}} \left(\mathbf{x}_{\{t,d_c\}} - \mathbf{A}_{d_c}^{(1)} \mathbf{x}_{\{t-1,d_c\}} \right) . \tag{C.26}$$

The partial derivative with respect to ξ is

$$\frac{\partial Q}{\partial \xi} = \sum_{t=2}^{T} \sum_s \sum_c o(f_*, s, c) \left(\frac{1}{\xi} - \frac{1}{\xi^2} a(t,c) \right) p(s_t|\mathbf{y}) \tag{C.27}$$

and setting it to zero yields the M-step update

$$\tilde{\xi} = \frac{\sum_{t=2}^{T} \sum_s \sum_c o(f_*, s, c) a(t,c) p(s_t|\mathbf{y})}{\sum_{t=2}^{T} \sum_s \sum_c o(f_*, s, c) p(s_t|\mathbf{y})} . \tag{C.28}$$

The expression $a(t,c)$ is again a Z-score considering the observed dynamical covariance and the covariance of the normal regime. Therefore the effect of this update is to consider, at each time point, every block of hidden state dimensions and all the switch settings in which the X-factor overwrites that block. Each of these possibilities exerts a (possibly conflicting) influence on the new value of ξ, weighted by the inferred probability of the switch setting. As for the SKF, the filtered estimate $p(s_t|\mathbf{y}_{1:t})$ is generally used instead of $p(s_t|\mathbf{y})$ to give a more tractable 'pseudo-EM' algorithm.

Note that (C.25-C.28) would not hold without independent block-diagonal structure in the hidden dynamics.

Appendix D

AR processes on different time scales

Some dynamical regimes in NICU monitoring data take place on a time scale which is very slow relative to the frequency of the measurements. For example, the blood pressure waves described in section 2.4.4 typically have a wavelength of several minutes, whereas observations arrive every second. In order to fit a model to this type of dynamical regime, it can be desirable to downsample the data to remove high frequency components which are not relevant to the pattern. However, this raises the issue of how to then apply the model to new data in conjunction with other models at the original high sampling frequency.

The same problem arises when extra datasets are available in which samples are recorded at a different frequency. For instance, an early version of the monitoring system in Edinburgh Royal Infirmary NICU normally stored physiological data at the rate of once per minute. This data is potentially useful because neonatal intensive care technology of the time was less advanced than at present and serious problems such as pneumothorax occurred more frequently. To take advantage of this data and apply models learnt from it in combination with other models that operate at a finer temporal resolution, it is necessary to find a way of representing them in terms of high-frequency data. Another situation in which this issue might arise would be if training data was thought to contain slower or faster dynamics than might be expected in test data.

Assuming we have a scalar AR(p) process $\mathbf{X}^{\mathrm{c}}$ to model the dynamics of interest on the 'coarse' time scale (e.g. archived data recorded at one data point per minute),

$$X_t^{\mathrm{c}} = \sum_{i=1}^{p} \alpha_i X_{t-i}^{\mathrm{c}} + Z_t \ , \tag{D.1}$$

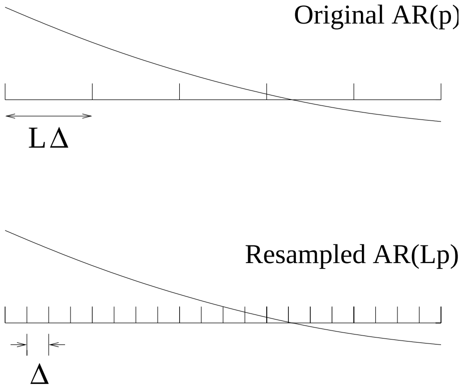


Figure D.1: The original coarse scale process $\mathbf{X}^{\mathrm{c}}$ and the desired resampled process $\mathbf{X}^{\mathrm{f}}$.

we would like to find a process $\mathbf{X}^{\mathrm{f}}$ which represents similar dynamics over a 'fine' time scale (e.g. normal second-by-second data) that covers L times as many sample points. This is illustrated in Figure D.1, where the coarse measurements arrive at intervals of $L\Delta$ and the fine scale measurements arrive at intervals of Δ. The new AR(Lp) process should have the same behaviour as the original, coarse AR(p) process. The coarse scale process is taken to have AR coefficients $\alpha_1^{\mathrm{c}}, \ldots, \alpha_p^{\mathrm{c}}$, while the fine scale process has coefficents $\alpha_1^{\mathrm{f}}, \ldots, \alpha_{Lp}^{\mathrm{f}}$.

An approximation to this can be obtained by simply spreading the coefficients $\alpha_1^{\mathrm{c}}, \ldots, \alpha_p^{\mathrm{c}}$ of the original process $\mathbf{X}^{\mathrm{c}}$ over Lp new coefficients as follows:

$$\alpha_1^{\mathrm{f}}, \ldots, \alpha_{Lp}^{\mathrm{f}} = \underbrace{\frac{\alpha_1^{\mathrm{c}}}{L}, \ldots, \frac{\alpha_1^{\mathrm{c}}}{L}}_{L}, \ldots, \underbrace{\frac{\alpha_p^{\mathrm{c}}}{L}, \ldots, \frac{\alpha_p^{\mathrm{c}}}{L}}_{L}. \tag{D.2}$$

The new $AR(Lp)$ process $\mathbf{X}^{\mathrm{f}}$ that this gives can be seen to be an approximate resampling of $\mathbf{X}^{\mathrm{c}}$. Because they are both stationary processes, we can see how good the approximation is by comparing their spectra. These cannot be compared directly, however, as they cover different frequency ranges. Instead, we want to downsample the fine-scale process and compare the spectrum of that with the coarse-scale process. This is considered in more detail in the next section.

D.1 Spectrum of a resampled process

If the coarse scale process has a power spectral density of $S^{\mathrm{c}}(f)$ at frequency f, and it is resampled to give a fine scale process with power spectral density $S^{\mathrm{f}}(f)$, some way is needed to compare the dynamics of the two to see if it is an accurate resampling. It is not immediately clear how to deal with the high frequency components which may be present in $\mathbf{X}^{\mathrm{f}}$ that cannot be represented by c. A good way to approach the comparison is to subsample the fine process at intervals of $L\Delta$ to give a third process with a spectrum $S^{\mathrm{s}}(f)$ on the original coarse scale. $S^{\mathrm{c}}(f)$ and $S^{\mathrm{s}}(f)$ can then be compared directly.

The Wiener-Khinchin theorem states that the autocovariance of a process is given by the Fourier transform of the power spectrum. Denoting the autocovariance of the fine-scale function at lag k as $\gamma^{\mathrm{f}}(k)$, this relationship is defined as

$$S^{\mathrm{f}}(f) = \Delta \sum_{k=-\infty}^{\infty} \gamma^{\mathrm{f}}(k) e^{-2\pi i k f \Delta}, \quad -\frac{1}{2\Delta} \leq f \leq \frac{1}{2\Delta}, \tag{D.3}$$

$$\gamma^{\mathrm{f}}(k) = \int_{-\frac{1}{2\Delta}}^{\frac{1}{2\Delta}} S^{\mathrm{f}}(f) e^{2\pi i f k \Delta} df\,. \tag{D.4}$$

If we are to subsample this down to the coarse time scale, using sampling steps of ΔL, we can express $S^{\mathrm{s}}(f)$ and $\gamma^{\mathrm{s}}(k)$ in terms of the original process:

$$\gamma^{\mathrm{s}}(k) = \gamma(kL) = \int_{-\frac{1}{2\Delta}}^{\frac{1}{2\Delta}} S^{\mathrm{f}}(f) e^{2\pi i f k L \Delta}, \tag{D.5}$$

$$S^{\mathrm{s}}(f) = \Delta L \sum_{k=-\infty}^{\infty} \gamma^{\mathrm{s}}(k) e^{-2\pi i k f \Delta L}, \quad -\frac{1}{2\Delta L} \leq f \leq \frac{1}{2\Delta L}. \tag{D.6}$$

Expanding (D.6) gives

$$S^{\mathrm{s}}(f) = \Delta L \sum_{k=-\infty}^{\infty} \left[\int_{-12\Delta}^{\frac{1}{2\Delta}} S^{\mathrm{f}}(f') e^{2\pi i f' k L \Delta} df' \right] e^{-2\pi i k f \Delta L}, \tag{D.7}$$

$$= \Delta L \int_{-\frac{1}{2\Delta}}^{\frac{1}{2\Delta}} S^{\mathrm{f}}(f') \sum_{k=-\infty}^{\infty} e^{2\pi i (f'-f) k L \Delta} df'. \tag{D.8}$$

The Poisson sum formula states that $\sum_{n=-\infty}^{\infty} f(n) = \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} f(x') e^{-2\pi i k x'} dx'$. When this is applied to the comb function (a sequence of Dirac-δ functions repeated at interval c), this gives the result

$$C \sum_{k=-\infty}^{\infty} \delta(x + kc) = \sum_{n=-\infty}^{\infty} e^{2\pi i \frac{x}{c}}. \tag{D.9}$$

The Fourier transform of the comb function $\sum_{k=-\infty}^{\infty} \delta(x + \frac{k}{L\Delta})$ has the same form as the sum in (D.8). This is what we would expect when we consider that this is the sampling function for points ΔL apart. Therefore we can write (D.8) as

$$S^{\mathrm{s}}(f) = \frac{\Delta L}{\Delta L} \int_{-\frac{1}{2\Delta}}^{\frac{1}{2\Delta}} df' S^{\mathrm{f}}(f') \sum_{k=-\infty}^{\infty} \delta\left((f - f') + \frac{k}{L\Delta} \right), \tag{D.10}$$

$$= \sum_{k=-\infty}^{\infty} S^{\mathrm{f}}(f + \frac{k}{L\Delta}) \tag{D.11}$$

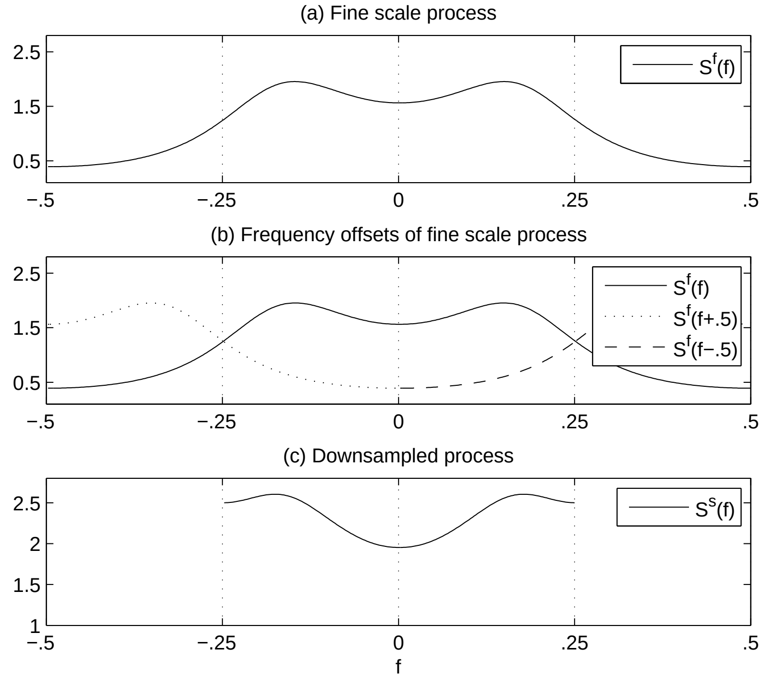


Figure D.2: Illustration of the way in which a fine scale process is downsampled. Panel (a) shows the spectrum of the fine scale process. Panel (b) shows the same spectrum with offset versions overlaid which correspond to the components in (D.11). The components are then summed over frequencies with magnitudes below the Nyquist rate $f_N = \pm 0.25$ to give the subsampled spectrum.

where we assume that the discrete spectrum $S^{\mathrm{f}}(f)$ is 0 for $|f| > \frac{1}{2\Delta}$. Therefore the subsampled spectrum can be expressed as components of the original spectrum, an example of which is shown in Figure D.2. This is a principle known as *aliasing*; when a process with high frequency components is subsampled, any of the original frequency components above the Nyquist frequency $(\frac{1}{2\Delta L})$ are indistinguishable from low frequency components on the new scale[1] (Jenkins, 1968). For high values of L, the subsampled spectrum $S^{\mathrm{s}}(f)$ may have many components.

[1] In the visual domain this is the ‘wagon wheel effect’, explaining why spoked wheels appear to rotate backwards in old film footage with a low frame rate. Notice, for example, that fine scale spectral components at a frequency of $f = 0.3$ in Figure D.2(b) are added to the subsampled spectrum at $f = -0.2$ in D.2(c).

D.2 First order example

We are now in a position to take an AR process, attempt to find an equivalent process at a finer sampling frequency, and compare $S^{\mathrm{c}}(f)$ and $S^{\mathrm{s}}(f)$ to see how well the resampling worked. We first examine the simplest case, an AR(1) process resampled by a factor of 2.

It is possible to obtain an expanded analytical expression for the spectrum of a low order AR process. In general, the spectrum of an AR(p) process can be obtained using

$$\begin{aligned} S(f) &= \Delta\sigma_Z^2 |H(f)|^2 \, , && \text{(D.12)} \\ H(f) &= \left(1 - \alpha_1 e^{2\pi i f \Delta} - \ldots - \alpha_p e^{2\pi i f \Delta p}\right)^{-1} && \text{(D.13)} \end{aligned}$$

where $H(f)$ is the frequency response function of the system, and data points are spaced at intervals of Δ. Taking a coarse scale AR(1) process sampled at steps of $L\Delta$ and setting $L = 2$, this form for the power spectrum can be expanded accordingly to give

$$S^{\mathrm{c}}(f) = \frac{2\Delta\sigma_Z^2}{1 + \alpha_1^2 - 2\alpha_1 \cos 2\pi f 2\Delta} \, , \tag{D.14}$$

where $-\frac{1}{4\Delta} \leq f \leq \frac{1}{4\Delta}$. For a second order AR process, $H(f)$ can be expanded to give the fine scale spectrum

$$S^{\mathrm{f}}(f) = \frac{\Delta\sigma_Z^2}{1 + \alpha_1^2 + \alpha_2^2 - 2\alpha_1(1 - \alpha_2)\cos 2\pi f \Delta - 2\alpha_2 \cos 4\pi f \Delta}, \tag{D.15}$$

where in this case where the sampling interval is Δ and $-\frac{1}{2\Delta} \leq f \leq \frac{1}{2\Delta}$.

We now have (D.14) as the coarse scale process. Using (D.2) to find an equivalent process with half the sampling interval we obtain fine-scale AR coefficients such that

$$X^{\mathrm{f}}(t) = \frac{\alpha_1}{2} X^{\mathrm{f}}_{t-1} + \frac{\alpha_1}{2} X^{\mathrm{f}}_{t-2} + Z^{\mathrm{f}}_t . \tag{D.16}$$

Note that the noise should be scaled since we want to have the same total amount of noise on L steps of $\mathbf{X}^{\mathrm{f}}$ as on a single step of $\mathbf{X}^{\mathrm{c}}$,

$$\begin{aligned} Z^{\mathrm{f}}_t &\sim \mathcal{N}(0, \sigma^2_{Z^{\mathrm{f}}}) \, , \\ \sum_L Z^{\mathrm{f}}_t &\sim \mathcal{N}(0, L\sigma^2_{Z^{\mathrm{f}}}) \;=\; \mathcal{N}(0, \sigma_Z^2) \, , && \text{(D.17)} \end{aligned}$$

so we set $\sigma^2_{Z^{\mathrm{f}}} = \frac{1}{L}\sigma_Z^2$. The spectrum of this new process $\mathbf{X}^{\mathrm{f}}$, according to (D.15), is

$$S^{\mathrm{f}}(f) = \frac{\Delta\sigma_Z^2}{1 + \frac{\alpha_1^2}{2} - \alpha_1(1 - \frac{\alpha_1}{2})\cos 2\pi f \Delta - \alpha_1 \cos 4\pi f \Delta} \, . \tag{D.18}$$

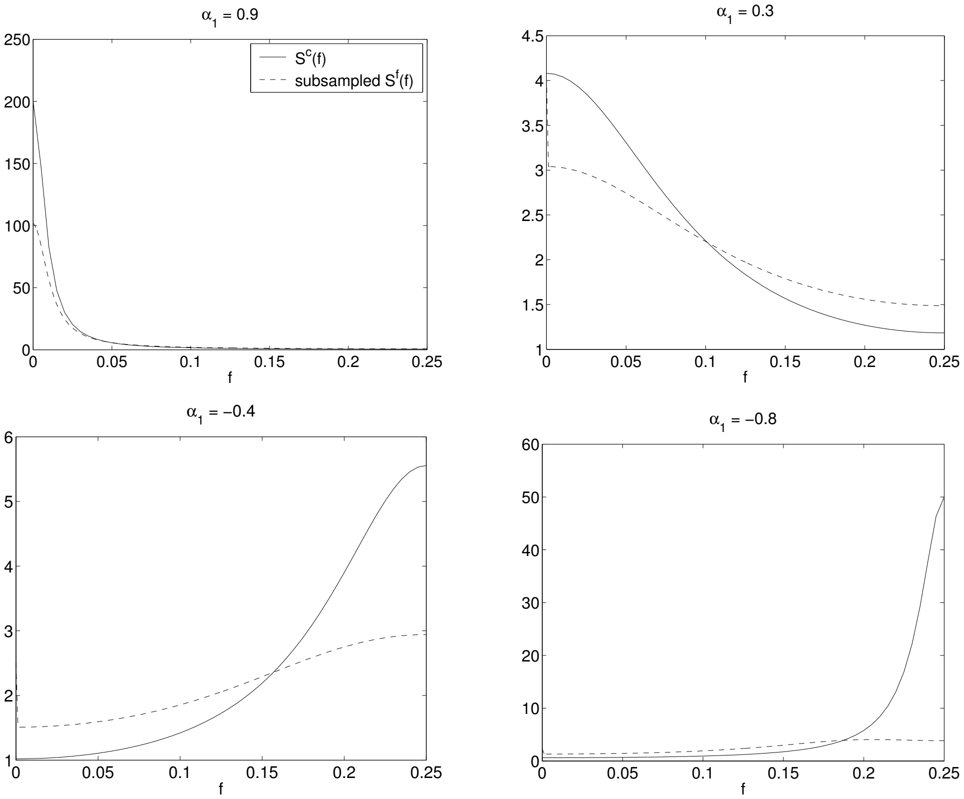


Figure D.3: Power spectral densities of AR(1) processes for varying values of α_1, plotted with subsampled fine-scale processes where $L = 2$. Solid lines show coarse scale processes, and dashed lines show the subsampled fine-scale processes (denoted by $S^{\mathrm{s}}(f)$ in the text).

Now we make a new process $\mathbf{X}^{\mathrm{s}}$ by sampling every second point of $\mathbf{X}^{\mathrm{f}}$. The new spectrum is given by (D.11) and is equal to

$$S^{\mathrm{s}}(f) = \sum_{k=-\infty}^{\infty} \frac{\Delta\sigma_Z^2}{1+\frac{\alpha_1^2}{2}-\alpha_1(1-\frac{\alpha_1}{2})\cos 2\pi(f\Delta+\frac{k}{2})-\alpha_1\cos 4\pi(f\Delta+\frac{k}{2})}. \tag{D.19}$$

Having obtained this form for the spectrum, we can compare it with the original AR(1) spectrum in (D.14) for various values of α_1. Figure D.3 shows examples wth varying α values, for which $\sigma_Z^2 = 1$ and $\Delta = 1$.

The resampling scheme tends provide a reasonable approximation except for α_1 at large negative values. To see the problem, consider the coarse scale process for a negative value of α_1. X_t^{c} changes sign at each time step. An AR(2) process with equal coefficients has no way of realising the equivalent behaviour — two positive values followed by two negative values, and so on.

D.3 Second order example

The first order example is limited in that an AR(1) process can essentially only represent exponential decay. A second or higher order process can oscillate, which is a useful test of the resampling procedure. In this example an AR(2) process is resampled to a fine scale AR(4) process. Using (D.12) and (D.13),

$$S^{\mathrm{c}}(f) = \frac{2\Delta\sigma_Z^2}{1+\alpha_1^2+\alpha_2^2-2\alpha_1(1-\alpha_2)\cos 2\pi f 2\Delta - 2\alpha_2 \cos 4\pi f 2\Delta}, \tag{D.20}$$

while the spectrum of a fine scale AR(4) process is given by

$$\begin{aligned} S^{\mathrm{f}}(f) &= \frac{\Delta\sigma_Z^2}{\left|1-\sum_{k=1}^4 \alpha_k e^{2\pi f k\Delta}\right|^2} \\ &= \frac{\Delta\sigma_Z^2}{\left[1-\sum_{k=1}^4 \alpha_k \cos(2\pi f k\Delta)\right]^2 + \left[\sum_{k=1}^4 \alpha_k \sin(2\pi f k\Delta)\right]^2}. \end{aligned} \tag{D.21}$$

Similarly to before, the fine scale coefficients are estimated with (D.2) and (D.11) is applied to give the subsampled spectrum $S^{\mathrm{s}}(f)$. Results for example AR(2) processes with varying values of α_1, α_2 are shown in Figure D.4, for which $\sigma_Z^2 = 1$ and $\Delta = 1$.

D.4 Optimising scaled coefficients

The previous sections show that sharing out the coefficients into p L-sized blocks give an approximation to the target resampled process. The approximation is not always very good, however, and we turn our attention to using the fine scale process $\mathbf{X}^{\mathrm{f}}$ as a starting point for finding a locally optimal solution.

Given the coefficients of the coarse scale AR process, $\alpha_1 \ldots \alpha_p$, we define a new set of fine scale coefficents which we would like to optimise:

$$\underbrace{\frac{\lambda_1\alpha_1}{L},\ldots,\frac{\lambda_1\alpha_1}{L}}_{L},\ldots,\underbrace{\frac{\lambda_p\alpha_p}{L},\ldots,\frac{\lambda_p\alpha_p}{L}}_{L}.$$

By starting at $\lambda_i = 1, i = 1\ldots p$, and then altering the scaling factors we can search for a better fitting approximation. The optimised fine scale process given by these coeffients is denoted by $S^{f*}(f)$, and $S^{\mathrm{s}*}(f)$ denotes the subsampled version on the coarse scale.

In order to optimise the fine scale process, an objective function is required. One possible error term for a particular setting of λ is given by the sum squared distance

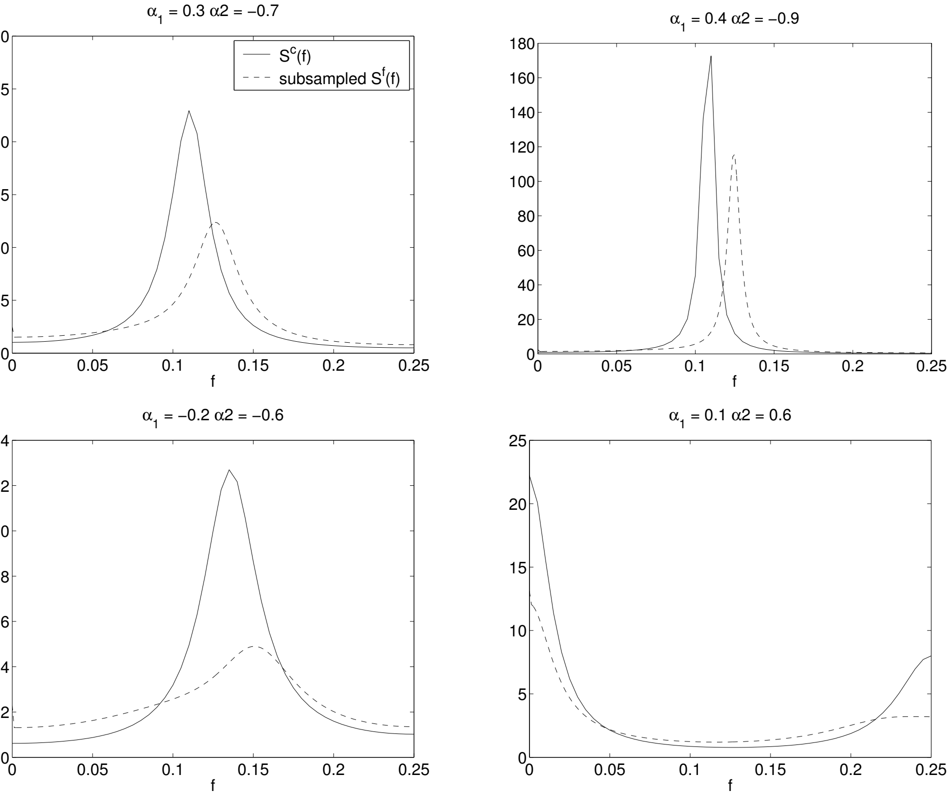


Figure D.4: Power spectral densities of AR(2) processes for varying values of α_1, α_2, plotted with subsampled fine-scale processes where $L = 2$.

between the target coarse scale spectrum and the subsampled optimised spectrum,

$$E(\lambda) = \sum_{f=0}^{\frac{n}{2L\Delta}} \left[S^{c}\left(\frac{f}{n}\right) - S^{s*}\left(\frac{f}{n}\right) \right]^2, \tag{D.22}$$

where n determines the number of points over the frequency range with which to calculate $E(\lambda)$. For all following results, $n = 50$. For a low order process it would be possible to calculate the error across many settings of λ and choose the best one. However, for higher order processes numerical optimisation becomes necessary. The use of scaled conjugate gradient search was investigated to find the local minimum error, with the initial condition $\lambda_i = 1, i = 1 \ldots p$. Results of optimising four AR(2) processes are shown in Figure D.5. The matches to the target spectra are much better after optimisation than with the naïve $\lambda_i = 1$ assumption.

This technique is directly applicable to higher orders of AR processes and larger L. Note that as a by-product of this method, multiple models can be obtained with different values of L as in Figure D.6. Here, mean blood pressure data containing

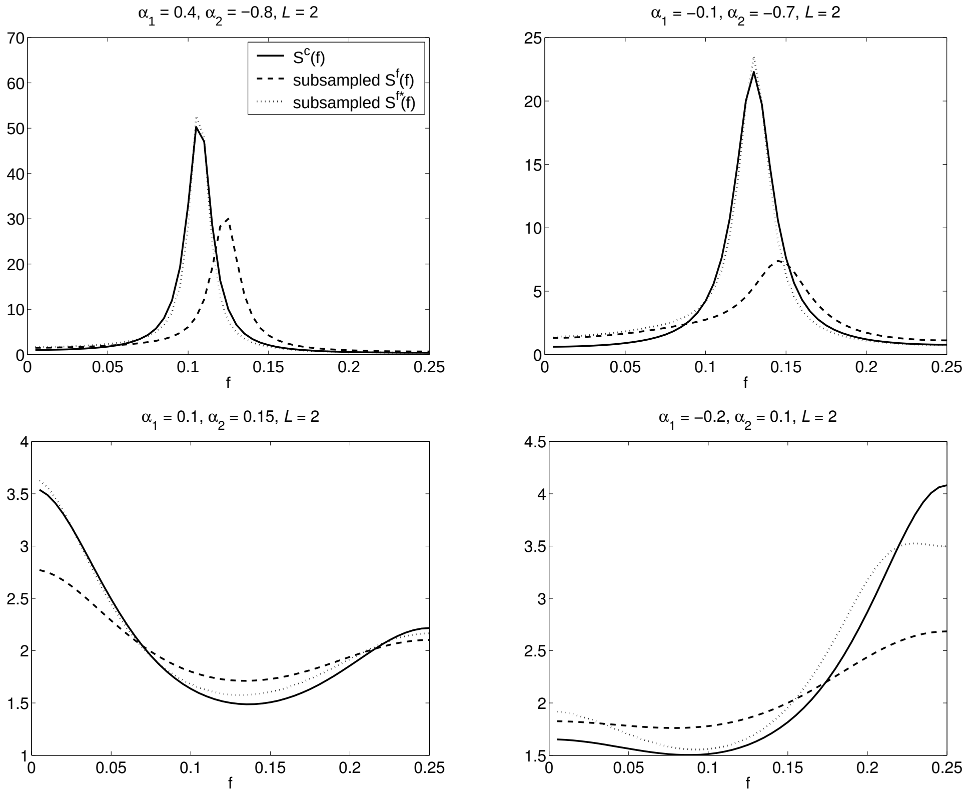


Figure D.5: Spectra of AR(2) processes resampled with a factor $L = 2$ and optimised. The dashed line shows the spectrum of the subsampled initial fine scale process ($S^s(f)$ in the text), and the dotted line shows the spectrum of the optimised subsampled process ($S^{s*}(f)$ in the text).

blood pressure waves (see section 2.4.4) was downsampled by a factor of 20 in order to learn an intrinsic AR(10) model. The plots show the spectrum of the resulting fine scale process when calculated with $L = \{16, 20, 24\}$. The corresponding fine scale processes are therefore AR(160), AR(200), and AR(240) respectively. Alternative models have therefore been obtained which describe blood pressure waves occurring at speeds of 0.8, 1 and 1.2 times that of the examples in the training data.

D.5 Summary

This appendix has set out the problem of applying a dynamical model to new data in which the sampling frequency is higher. A method has been given to reapply a coarse scale process to a fine scale, and for assessing how accurately the dynamics of

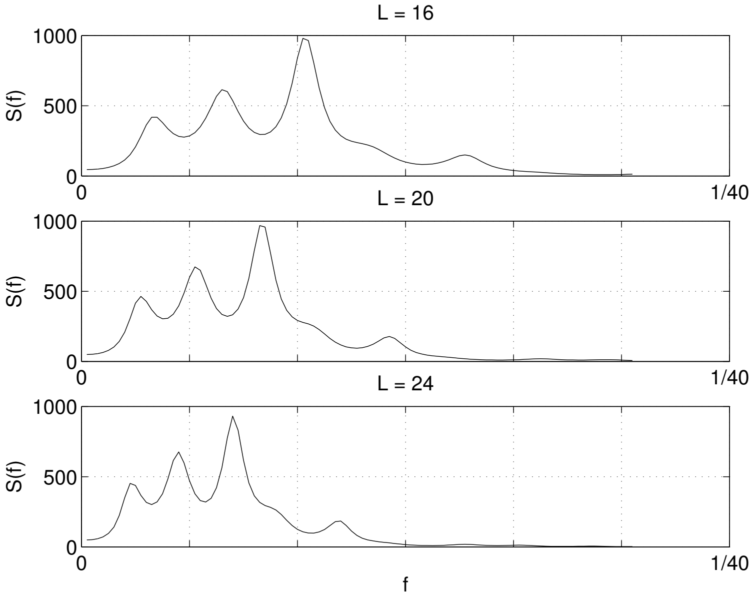


Figure D.6: Spectra of fine scale processes representing blood pressure waves. Example mean blood pressure data containing BP waves was downsampled by a factor of 20, and from this a coarse scale AR(10) model was learnt. The middle plot shows the spectrum of the fine scale process when $L = 20$. Different values of L allow models for waves which are faster or slower.

the new process correspond to the original by subsampling back to the coarse scale. For situations in which the accuracy is not good, it has been shown that results can be improved with numerical optimisation. The process of optimisation is practical for reasonably high order coarse scale processes and high values of L. The process also provides the opportunity to use different values of L in order to speed up or slow down the fine scale process.

Bibliography

V. Ahlborn, B. Bonnhorst, C.S. Peter, and C.F. Poets. False alarms in very low birthweight infants: comparison between three intensive care monitoring systems. *Acta Paediatrica*, 89(5):571–7, 2000.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

E. Alberdi, J.-C. Becher, K.J. Gilhooly, J.R.W. Hunter, R.H. Logie, A. Lyon, N. McIntosh, and J. Reiss. Expertise and the Interpretation of Computerised Physiological Data: Implications for the Design of Medical Decision Support in Neonatal Intensive Care. *International Journal of Human Computer Studies*, 56(3):191–216, 2002.

D.L. Alspach and H.W. Sorenson. Nonlinear Bayesian Estimation using Gaussian Sum Approximation. *IEEE Transactions on Automatic Control*, 17:439–447, 1972.

M. Azzouzi and I.T. Nabney. Modelling Financial Time Series with Switching State Space Models. *Proceedings of the IEEE/IAFE 1999 Conference on Computational Intelligence for Financial Engineering*, pages 240–249, 1999.

D. Barber and B. Mesot. A Novel Gaussian Sum Smoother for Approximate Inference in Switching Linear Dynamical Systems. In *Advances in Neural Information Processing Systems*, volume 20. MIT Press, 2006.

G.E.P. Box and G.M. Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, 1976.

J.V. Candy. *Model-Based Signal Processing*. Wiley-IEEE Press, 2005.

J.V. Candy. *Signal Processing: The Model-based Approach*. McGraw-Hill, 1986.

C. Cao, N. McIntosh, I.S. Kohane, and K. Wang. Artifact Detection in the PO2 and

PCO2 Time Series Monitoring Data from Preterm Infants. *Journal of Clinical Monitoring and Computing*, 15(6):369–378, 1999.

A.T. Cemgil, H.J. Kappen, and D. Barber. A Generative Model for Music Transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 14(2):679–694, 2006.

M. Chambrin. Alarms in the intensive care unit; how can the number of false alarms be reduced? *Crit Care Med*, 5(4):184–188, 2001.

S. Charbonnier, G. Becq, and G. Biot. On-Line Segmentation Algorithm for Continuously Monitored Data in Intensive Care Units. *IEEE Transactions on Biomedical Engineering*, 51(3):484–492, 2004.

C. Chatfield. *The Analysis of Time Series*. Chapman and Hall, 1975.

D. Christie and E. Tansey, editors. *Origins of neonatal intensive care in the UK*, volume 9 of *Wellcome Witnesses to Twentieth Century Medicine*. Wellcome Trust Centre for the History of Medicine at UCL, London, UK, 2001.

T.R. Cooper, C.L. Berserth, J.M. Adams, and L.E. Weisman. Actuarial Survival in the Premature Infant Less Than 30 Weeks' Gestation. *Pediatrics*, 101:975–978, 1998.

S. Cunningham, S. Deere, R. A. Elton, and N. McIntosh. Neonatal physiological trend monitoring by computer. *Journal of Clinical Monitoring and Computing*, 9:221–227, 1992.

S. Cunningham, S. Deere, and N. McIntosh. Cyclical variation of blood pressure and heart rate in neonates. *Archives of Disease in Childhood*, 69:64–67, 1993.

S. Cunningham, S. Deere, A. Symon, R.A. Elton, and N. McIntosh. A Randomized, Controlled Trial of Computerized Physiologic Trend Monitoring in an Intensive Care Unit'. *Crit Care Med*, 26:2053–2060, 1998.

N. de Freitas, R. Dearden, F. Hutter, R. Morales-Menedez, J. Mutch, and D. Poole. Diagnosis by a waiter and a Mars explorer. *Proceedings of the IEEE*, 92(3), 2004.

P. de Jong and J. Penzer. The ARMA model in state space form. *Statistics and Probability*, 70:119–125, 2004.

A. Doucet, N. de Freitas, K.P. Murphy, and S.J. Russell. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In Craig Boutilier and Moisés Goldszmidt, editors, *UAI*, pages 176–183. Morgan Kaufmann, 2000.

Dräger Medical. Sc7000 patient monitor datasheet, 2007.

J. Droppo and A Acero. Noise Robust Speech Recognition with a Switching Linear Dynamic Model. In *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, 2004.

R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 13th edition, 1958.

B.J. Frey, N. Mohammad, Q.D. Morris, Z. Wen, M.D. Robinson, S. Mnaimneh, R. Chang, P. Qun, E. Sat, J. Rossant, B. Bruneau, J.E. Aubin, B.J. Blencowe, T.R. Hughes, and P. Smyth. Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs. *Nature genetics*, 37(9):991–996, 2005.

R. Fried, U. Gather, and M. Imhoff. Pattern Recognition in Intensive Care Online Monitoring. In *Sonderforschungsbereich (SFB) 475*, volume 1, pages 509–512. University of Dortmund, 2003.

Z. Ghahramani and G.E. Hinton. Parameter Estimation for Linear Dynamical Systems. Technical report, Department of Computer Science, University of Toronto, 1996.

Z. Ghahramani and G.E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):963–996, 1998.

Z. Ghahramani and M. Jordan. Factorial Hidden Markov Models. *Machine Learning*, 29:245–273, 1997.

K. Gordon and A.F.M. Smith. Modeling and Monitoring Biomedical Time Series. *J. Amer. Statist. Assoc.*, 85:328–337, 1990.

M.S. Grewal and A.P. Andrews. *Kalman Filtering: Theory and Practice*. John Wiley & Sons, 2001.

M.P. Griffin and J.R. Moorman. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. *Pediatrics*, 107(1):97–104, 2001.

I. J. Haimowitz, P. P. Le, and I. S. Kohane. Clinical monitoring using regression based trend templates. *Artificial Intelligence in Medicine*, 7(6):473–496, 1995.

A.C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1990.

S.W. Hoare and P.C.W. Beatty. Automatic artifact identification in anaesthesia patient record keeping: a comparison of techniques. *Medical Engineering and Physics*, 22: 547–553, 2000.

J. Hunter, G. Ewing, L. Ferguson, Y. Freer, R. Logie, P. McCue, and N. McIntosh. The NEONATE database. In A Abu-Hanna and J Hunter, editors, *Working Notes of the Joint Workshop on Intelligent Data Analysis in Medicine and Pharmacology and Knowledge-Based Information Management in Anaesthesia and Intensive Care, AIME 03*, pages 21–24, 2003.

J.R.W. Hunter. TSNet A Distributed Architecture for Time Series Analysis. In N. Peek and C. Combi, editors, *Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP 2006)*, pages 85–92, 2006.

J.R.W. Hunter and N. McIntosh. Knowledge-Based Event Detection in Complex Time Series Data. In W Horn, Y. Shahar, G. Lindberg, S. Andreassen, and J. Wyatt, editors, *Artificial Intelligence in Medicine: Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (AIMDM'99)*, 1999.

M. Imhoff, M. Bauer, U. Gather, and D. Löhlein. Statistical pattern detection in univariate time series of intensive care in-line monitoring data. *Intensive Care Med*, 24:1305–1314, 1998.

G.M. Jenkins and D. Watts. *Spectral Analysis and Its Applications*. Emerson-Adams, 1998.

R.R. Kennedy. A modified Trigg's Tracking Variable as an 'advisory' alarm during anaesthesia. *Journal of Clinical Monitoring and Computing*, 12:197–204, 1995.

E. Keogh, S. Lonardi, and B. Chiu. Finding surprising patterns in a time series database in linear time and space. *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 550–556, 2002.

C.-J. Kim. Dynamic Linear Models with Markov-Switching. *Journal of Econometrics*, 60:1–22, 1994.

Mike Klaas, Mark Briers, Nando de Freitas, Arnaud Doucet, Simon Maskell, and Dustin Lang. Fast particle smoothing: if I had a million particles. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 481–488. ACM Press, 2006.

N. Lavrač, E. Keravnou, and B. Zupan. *Encyclopedia of Computer Science and Technology*, volume 42, chapter Intelligent Data Analysis in Medicine, pages 113–157. Dekker, New York, 2000.

S. T. Lawless. Crying wolf: false alarms in a pediatric intensive care unit. *Crit Care Med*, 22(6):981–5, 1994.

R. Logie, J.R.W. Hunter, N. McIntosh, K. Gilhooly, E. Alberdi, and J. Reiss. *Medical Cognition and Computer Support in the Intensive Care Unit: A Cognitive Engineering Approach*, pages 167–174. Ashgate, 1998.

A.J. Lyon, M.E. Pikaar, P. Badger, and N. McIntosh. Temperature control in very low birthweight infants during first five days of life. *Arch Dis Child Fetal Neonatal Ed*, 76:47–50, 1997.

J. Ma and S. Perkins. Online Novelty Detection on Temporal Sequences. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618, 2003.

D.J.C MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

M. Malik. Heart rate variability. *Curr Opin Cardiol.*, 13(1):36–44, 1998.

M. Markou and S. Singh. Novelty detection: a review - part 1: statistical approaches. *Signal Processing*, 83:2481–2497, 2003.

R.J. Martin and A.A. Faranoff. Neonatal apnea, bradycardia, or desaturation: Does it matter? *J Pediatrics*, 132(5):783–9, 1998.

N. McIntosh. Intensive care unit monitoring: past, present and future. *Clin Med*, 2(4): 349–55, 2002.

N. McIntosh, J.C. Becher, S. Cunningham, B. Stenson, I.A. Laing, A.J. Lyon, and P. Badger. Clinical diagnosis of pneumothorax is late: use of trend data and decision support might allow preclinical detection. *Pediatric Research*, 48(3):408–415, 2000.

N. McIntosh, P. Helms, and R. Smyth, editors. *Forfar and Arneil's Textbook of Pediatrics*. Churchill Livingstone, 2003.

S. Miksch, Horn, C Popow, and F. Paky. Utilizing Temporal Data Abstraction for Data Validation and Therapy Planning for Artificially Ventilated Newborn Infants. *Artificial Intelligence in Medicine*, 8(6):543–576, 1996.

M.S. Miller, K.M. Shannon, and G.T. Wetzel. Neonatal bradycardia. *Prog Pediatr Cardiol.*, 11(1):19–24, 2000.

D.W. Milligan. Failure of autoregulation and intraventricular haemorrhage in preterm infants. *Lancet*, 1(8174):896–8, 1980.

R. Morales-Menedez, N. de Freitas, and D. Poole. Real-Time Monitoring of Complex Industrial Processes with Particle Filters. In *Advances in Neural Information Processing Systems 15*. MIT Press, 2002.

R. Morales-Menedez, N. de Freitas, and D. Poole. Estimation and Control of Industrial Processes with Particle Filters. *American Control Conference*, 2003.

K. Murphy. Switching Kalman filters. Technical report, U.C. Berkeley, 1998.

K. Murphy and S. Russell. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo in Practice*. Springer-Verlag, 2001.

V. Pavlović, J.M. Rehg, and J. MacCormick. Learning Switching Linear Models of Human Motion. In *Advances in Neural Information Processing Systems 13*. MIT Press, 2000.

T. Penzel, J. McNames, A. Murray, P. de Chazal, G. Moody, and B. Raymond. Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. *Med Biol Eng Comput.*, 40(4):402–7, 2002.

C.F. Poets. Pulse oximetry vs. transcutaneous monitoring in neonates: practical aspects. `http://www.bloodgas.org`, 2003.

W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992.

J.A. Quinn and C.K.I. Williams. Known Unknowns: Novelty Detection in Condition Monitoring. In *Proc 3rd Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2007.

L.R. Rabiner and B.H. Juang. An introduction to hidden Markov models. *IEEE Acoustics, Speech and Signal Processing*, 3:4–16, 1989.

J.M. Rennie and N.R.C. Roberton. *A Manual of Neonatal Intensive Care*. Hodder Arnold, 2001.

S. Roberts and L. Tarassenko. A Probabilistic Resource Allocating Network for Novelty Detection. *Neural Computation*, 6:270–284, 1994.

M. Roessgen, A.M. Zoubir, and B. Boashash. Seizure detection of newborn EEG using a model-based approach. *IEEE Transactions on Biomedical Engineering*, 45(6): 673–685, 1998.

S. Roweis and G. Ghahramani. A Unifying Review of Linear Gaussian Models. *Neural Computation*, 11(2):305–345, 1999.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

C. Shahabi, X. Tian, and W. Zhao. TSA Tree: A Wavelet Based Approach to Improve the Efficiency of Multi-level Surprise and Trend Queries. *Proceedings of the 12th International Conference on Scientific and Statistical Database Management*, 2000.

R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications*. Springer-Verlag, 2000.

R.H. Shumway and D.S. Stoffer. Dynamic Linear Models with Switching. *J. Am. Statistical Assoc.*, 86:763–769, 1991.

D.F. Sittig and M. Factor. Physiologic trend detection and artifact rejection: a parallel implementation of a multi-state Kalman filtering algorithm. *Computational Methods and Programs in Biomedecine*, pages 1–10, 1990.

E.D. Sivek, J.S. Gochberg, R. Fronek, and D. Scott. Lessons to be learned from the design, development and implementation of a computerised patient care management system for the Intensive Care Unit. *Symp Comput Applic Med Care*, 11:614619, 1987.

A.F.M. Smith and M. West. Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter. *Biometrics*, 39:867–878, 1983.

P. Smyth. Markov monitoring with unknown states. *IEEE Journal on Selected Areas in Communications*, 12(9):1600–1612, 1994.

A. Spengler. Neonatal baby monitoring. Master's thesis, School of Informatics, University of Edinburgh, 2003.

S. Stoddart, L. Summers, and M. Ward Platt. Pulse oximetry: What it is and how to use it. *J Neonatal Nursing*, 3(4):12–14, 1997.

A.J. Storkey and M. Sugiyama. Mixture Regression for Covariate Shift. In *Advances in Neural Information Processing Systems*, volume 20. MIT Press, 2007.

M. Sugiyama and K.-R. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4):249–279, 2005.

C. Tsien. *TrendFinder: Automated Detection of Alarmable Trends*. PhD thesis, MIT, 2000.

C.L. Tsien and J.C. Fackler. Poor prognosis for existing monitors in the intensive care unit. *Crit Care Med*, 25(4):614–9, 1997.

M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 1999.

R.D. White. Recommended Standards for Newborn ICU Design. In *Sixth Consensus Conference in NICU Design*, 2006.

C.K.I. Williams, J. Quinn, and N. McIntosh. Factorial Switching Kalman Filters for Condition Monitoring in Neonatal Intensive Care. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*. MIT Press, 2006.

P.C. Woodland. Hidden Markov models using vector linear prediction and discriminative output functions. In *Proceedings of 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 509–512. IEEE, 1992.

S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK Book (version 3.1), University of Cambridge Engineering Department. `http://htk.eng.cam.ac.uk/docs/docs.shtml`, 2001.