THE UNIVERSITY
of EDINBURGH

Dynamical Models for Neonatal Intensive Care Monitoring

Ioan Stanculescu

Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2014

Abstract

The vital signs monitoring data of an infant receiving intensive care are a rich source of information about its health condition. One major concern about the state of health of such patients is the onset of neonatal sepsis, a life-threatening bloodstream infection. As early signs are subtle and current diagnosis procedures involve slow laboratory testing, sepsis detection based on the monitored physiological dynamics is a clinically significant task. This challenging problem can be thoroughly modelled as real-time inference within a machine learning framework.

In this thesis, we develop probabilistic dynamical models centred around the goal of providing useful predictions about the onset of neonatal sepsis. This research is characterised by the careful incorporation of domain knowledge for the purpose of extracting the infant's true physiology from the monitoring data.

We make two main contributions. The first one is the formulation of sepsis detection as learning and inference in an Auto-Regressive Hidden Markov Model (AR-HMM). The model investigates the extent to which physiological events observed in the patient's monitoring traces could be used for the early detection of neonatal sepsis. In addition, the proposed approach involves exact marginalisation over missing data at inference time. When applying the AR-HMM on a real-world dataset, we found that it can produce effective predictions about the onset of sepsis.

Second, both sepsis and clinical event detection are formulated as learning and inference in a Hierarchical Switching Linear Dynamical System (HSLDS). The HSLDS models dynamical systems where complex interactions between modes of operation can be represented as a two-level hidden discrete hierarchical structure. For neonatal condition monitoring, the lower layer models clinical events and is controlled by upper layer variables with semantics sepsis/non-sepsis. The model parameterisation and estimation procedures are adapted to the specifics of physiological monitoring data. We demonstrate that the performance of the HSLDS for the detection of sepsis is not statistically different from the AR-HMM, despite the fact that the latter model is given "ground truth" annotations of the patient's physiology.

Lay Summary

The vital signs monitoring data of an infant receiving intensive care are rich in information about its health condition, and are now routinely recorded in Neonatal Intensive Care Units (NICUs). One major concern about the state of health of NICU patients is the onset of neonatal sepsis. As early signs are subtle and current diagnosis procedures involve slow laboratory testing, sepsis detection based on the monitored physiological dynamics is a clinically significant task. This challenging problem can be thoroughly modelled as real-time inference within a machine learning framework.

In this thesis, we developed probabilistic dynamical models centred around the goal of providing useful predictions about the onset of neonatal sepsis. This research is characterised by the careful incorporation of domain knowledge for the purpose of extracting the infant's true physiology from the monitoring data.

We first investigated the extent to which low-level physiological events observed in the patient's monitoring traces could be used for the early detection of neonatal sepsis. In close collaboration with clinicians from the NICU at the Royal Infirmary of Edinburgh, we defined and annotated a set of clinical events. The task was then formulated as learning and inference in a generative probabilistic model, called the Autoregressive Hidden Markov Model (AR-HMM). When applying our model on a genuine monitoring dataset, we found that it can produce effective real-time predictions about the onset of sepsis.

A practical limitation of the above model is that it requires expert annotations of physiological events as input. In order to eliminate this bottleneck, we developed a hierarchical probabilistic model, called the Hierarchical Switching Linear Dynamical System (HSLDS), able to both detect neonatal sepsis and infer the physiological events from the raw vital signs data. We empirically demonstrated that the performance of this model is not statistically different from the AR-HMM, despite the fact that the latter model is given "ground truth" annotations of the patient's physiology.

Acknowledgements

I am deeply grateful to my primary supervisor, Prof. Chris Williams, who has tirelessly stimulated my progress. His vast expert insight, passion for scientific truth and openness for discussion have made me feel more than privileged to work together. Dr. Yvonne Freer has actively supported my research, provided excellent clinical information, and also the data that made this work possible. It has been a great pleasure to receive advice from Prof. Neil McIntosh, who has also supervised the data annotation process. I am also thankful to Dr. John Quinn whose research and baby monitoring code have been instrumental for my progress.

David Clifton and Amos Storkey have provided many useful comments while examining this thesis. In addition, I would like to thank Guido Sanguinetti and Rod Murray-Smith, who have contributed with excellent advice during my yearly review meetings.

Being part of the Probabilistic Inference Group has been an outstanding experience, as I have had much to learn for its elite faculty members. Thanks to my fellow PhD students with whom I've engaged in numerous exciting (non-)technical discussions.

I would like to thank the Scottish Informatics and Computer Science Alliance (SICSA), the Engineering and Physical Sciences Research Council (EPSRC) and the Informatics Graduate School for funding my PhD studies.

In the end, the biggest thanks go to my family for making all of the above possible.

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.

(*Ioan Stanculescu*)

Table of Contents

Chapter 1

Introduction

Intensive care is the branch of medicine concerned with the diagnosis and treatment of patients whose life-threatening condition requires continuous monitoring and support via medication and equipment. This thesis is concerned with monitoring in a Neonatal Intensive Care Unit (NICU), where infants born three to four months prematurely are treated for their generally fragile condition. NICU monitoring can be viewed as a platform for doctors to gain the evidence necessary for translating their clinical expertise into a medical diagnosis, so that the most effective course of treatment is pursued.

A very important part of this evidence is provided by continuously monitoring the infant's *vital signs*. These regularly include measurements of the heart rate, amounts of various gases in the blood stream, temperatures and blood pressure, and are recorded on a second-by-second basis. This stream of data is widely known to be rich in information about the patient's state of health. Nevertheless, it is the extraction of this information from the monitoring traces where current approaches are deficient. Difficulties include the need to analyse patient physiology across multiple measurement channels and time scales, and the corruption of data with artifact. Expert real-time analysis of these high-frequency, multi-dimensional data leads to overload, and may also be challenging for junior staff. In addition, naïve monitoring software often results into high false alarm rates. Thus, the use of physiological data for answering high-level questions about the infant's state of health, such as the onset of an infection, cardiovascular and respiratory problems remains largely unsolved.

In this thesis, we combine the representational power of machine learning methods with knowledge engineering into a novel framework for employing the vital signs streams towards answering high-level questions about the health condition of NICU patients. Here, we focus on the important problem of making early predictions about the onset of neonatal sepsis.

1.1 Neonatal sepsis

Late-Onset Neonatal Sepsis (LONS) is a bloodstream infection, usually bacterial, occurring during the first days of life. Its onset is a major cause of high mortality, lifelong neurodisability and increased health care costs [Modi et al., 2009]. Estimates show that 10% of all neonates and 25% of Very Low Birth Weight babies (VLBW, < 1500 grams birth weight) are affected, and this number rises to 50% for extremely preterm infants [Stoll et al., 2002, Beck-Sague et al., 1994, Modi et al., 2009]. The patients we considered for our work are VLBW with an average gestation of 27 weeks and a mean birth below 900 grams (see Table 5.1).

The major challenge in successfully treating septic babies is making the diagnosis of infection in the first place. Early signs are subtle and yet it is at this stage that treatment will be effective. A deterioration of the baby's condition over the course of a few hours is a strong symptom for neonatal sepsis, and prompts clinicians to take a blood sample for laboratory testing. However, laboratory cultures can take up to a day before becoming available. Because of the dangers of delaying treatment, antibiotic therapy is usually started at the same time as taking the blood sample. However, applying low thresholds in suspecting sepsis results in a high number of patients being treated unnecessarily for each true case [Griffin et al., 2003]. Thus, if achievable, the early detection of sepsis based on monitoring data would be of great value.

We will continue the discussion on neonatal sepsis in Section 4.1, where we also provide a review of the previous work on early detection.

1.2 Vital signs monitoring

The types of babies considered in this thesis are nursed inincubators primarily because of their lack of development. Cotside devices continuously display measurements of the patient's vital signs, which allows the NICU staff to monitor whether their physiological systems function correctly.

Most of the time the patients are in a "stable" state, and their vital signs display a healthy variation. In many situations, certain acute physiological conditions give rise to characteristic patterns on the monitoring traces. An example of patterns falling into this category is bradycardia, a spontaneous drop in heart rate measurements. Other stereotypical patterns are related to measurement corruption by sensor fault or by operating the monitoring equipment. The simplest type of pattern in this category is a probe disconnection, which is characterised by the temporary lack of monitoring data. At the same time, there are periods of unusual or novel dynamics, which may be harder to explain. Throughout this thesis, we will refer to the occurrence of any pattern different from the "stable" state dynamics as a *low-level clinical event*, or clinical event in short. These notions will be detailed and illustrated in Chapter 3.

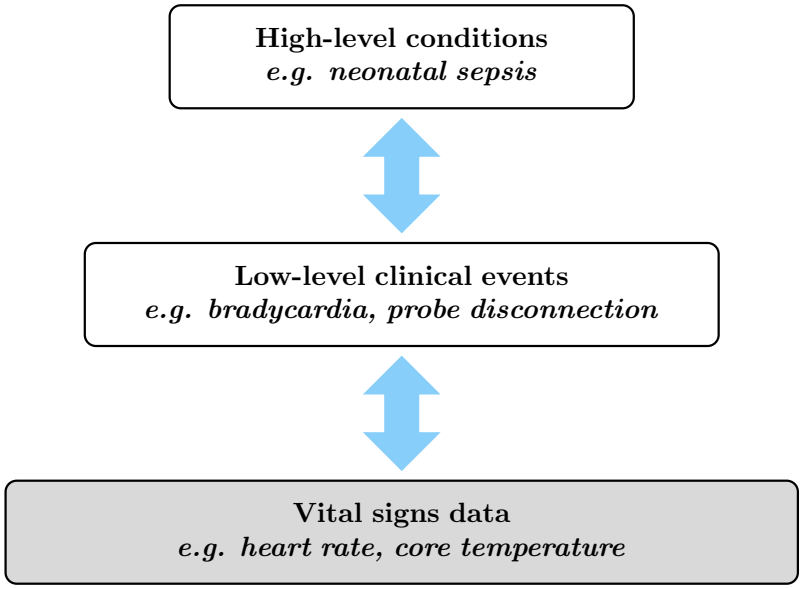


Figure 1.1: Hierarchical framework for neonatal physiological monitoring.

In this thesis, we propose a hierarchical model for neonatal physiological monitoring. Conceptually, the hierarchy has three layers, each incorporating some specific a priori knowledge. The bottom level is the monitoring data layer, and is the only one which can be directly observed. On top of this, we place a layer responsible for describing the vital signs measurements in terms of the clinical events. Generally lasting at most a few minutes, clinical events provide a local low-level explanation of the data. Even though they are a more abstract representation of the patient's condition than the raw monitoring recordings, they cannot be directly used to articulate a clinical diagnosis. Thus, we incorporate further knowledge into a third level of our hierarchy, which is designed to produce high-level explanations of the measurements. In the work presented by this thesis, this layer produces inferences about the onset of neonatal sepsis. The diagram shown in Figure 1.1 illustrates our monitoring framework.

As already mentioned, extracting useful descriptions from the vital signs data is non-trivial. Many of the deployed neonatal monitoring systems suffer from high false positive rates [Tsien and Fackler, 1997, Ahlborn et al., 2000], as they generally oversimplify the complexity of physiological data; e.g an alarm is fired when some pre-set threshold is exceeded. The methods developed in this thesis are most related to the Factorial Switching Linear Dynamical System (FSLDS) described in Quinn, Williams, and McIntosh [2009], and which is also reviewed in Section 4.2.2. Based on the promising results they show on detecting multiple low-level clinical events, we chose to adapt and extend their ideas in order to make high-level predictions about the onset of sepsis.

1.3 Modelling uncertainty in neonatal condition monitoring

Condition monitoring refers to the task of using measurements taken from a dynamical system in order to infer which of several regimes best describes the data. As hinted in the above, the vital signs measurements of a prematurely born baby receiving intensive care can be understood as belonging to several different regimes. Each regime is associated with patterns in the data, but it cannot be directly observed. Given a sequence of vital signs observations, the goal of neonatal condition monitoring is to determine which of the regimes best describes the patient's health condition.

As there are multiple hypotheses competing for explaining the measurements, the theory of probabilities is the rigorous way to handle this uncertainty. The approach we take here is directed graphical modelling (see e.g. Koller and Friedman [2009])[1], where joint probability distributions are factorised into products of local conditional distributions. The conditional independences implied by a directed graphical model are best seen by representing it as a Directed Acyclic Graph (DAG). In this thesis, directed models are used to capture the processes by which the observed data were *generated*. In broad terms, the clinician's prior knowledge about each hypothesis h is encoded in a prior distribution $p(h)$. Further knowledge is used to construct the conditional distribution of the physiological observations y given each hypothesis $p(y|h)$. Thus, we obtain a joint distribution of all the modelled variables, $p(h,y) = p(h)p(y|h)$. In this setting, automated condition monitoring is equivalent to probabilistic inference. The latter consists of applying Bayes rule to determine the posterior belief about the hypothesises in light of the observed data:

$$p(h|y) = \frac{p(h)p(y|h)}{\sum_{h'} p(h')p(y|h')}. \tag{1.1}$$

In generative approaches the modelling task can be separated from the inference task, thus facilitating knowledge integration. For baby monitoring, advantages include ease in adding or removing measurement channels and the principled way in which missing data can be handled (see Sections 4.2.2, 5.2 and 6.2.2).

Vital signs recordings arrive sequentially in time series of the form: $y_1, y_2, y_3, \ldots$, and generative models for this type of data will be discussed in Chapter 2. In general, each data point y_t will be associated with an unobserved variable h_t. NICU monitoring requires real-time inference, and thus the inference goal at any time t is determining the posterior distribution of all past hidden variables given the all the data observed up to time t, $p(h_{1:t}|y_{1:t})$. This posterior is referred to as the *filtering* distribution, but in practice it often suffices to estimate its one-step marginal $p(h_t|y_{1:t})$. Other inference queries discussed in the following chapters are *smoothing* and *prediction*. Smoothing commonly applies to off-line settings, and refines the filtering distribution by also conditioning on all the observations made after time t. Prediction is concerned

[1] Directed graphical models are sometimes referred to as Bayes networks or belief networks.

with estimating the distribution of future latent and observed variables given historical data.

1.4 Thesis Overview

In **Chapter 2** we review several probabilistic dynamical models which could be applied for monitoring the condition of premature babies receiving intensive care. The discussion evolves around models capable to "switch" between several modes of operation, and also around models in which several hidden factors collectively determine the regime followed by the data.

Chapter 3 provides information on how an infant's physiology is translated into the data streams used in this thesis. We discuss the different regimes present in the data, and illustrate their associated patterns. This includes defining physiological and artifactual clinical events.

A review of the previous work done on NICU monitoring is offered in **Chapter 4**. We first discuss neonatal sepsis, and then focus on the condition monitoring FSLDS of Quinn et al. [2009].

Chapter 5 demonstrates an autoregressive hidden Markov model (AR-HMM) for making early predictions about the onset of neonatal sepsis. Model development relies on a priori knowledge about an increase in acute physiological events being a symptom of sepsis. Thus, the proposed AR-HMM is a principled framework for assessing the amount of predictive information about neonatal sepsis offered by the distribution of clinical events. Using an expert-annotated dataset collected from the NICU at the Royal Infirmary of Edinburgh, we perform a thorough empirical analysis showing that our AR-HMM produces effective predictions about the onset of sepsis. The work presented in this chapter is an extension of **Stanculescu, Williams, and Freer [2013]**.

The AR-HMM takes expert event annotations as input, which limits its practical application. In **Chapter 6**, we demonstrate the Hierarchical Switching Linear Dynamical System (HSLDS) for inferring both sepsis and clinical events from the raw monitoring data. The proposed HSLDS is developed as an extension of the FSLDS for neonatal condition monitoring [Quinn, Williams, and McIntosh, 2009]. It adds a higher-level variable with semantics sepsis/non-sepsis, which allows detecting changes in physiological events that signal the presence of sepsis. We empirically show that the HSLDS's performance is not statistically different from the AR-HMM, and is also competitive against discriminative sepsis detectors. The work described in this chapter extends **Stanculescu, Williams, and Freer [2014]**.

Chapter 7 summarises the contributions made by this thesis, and discusses several directions in which the work can be extended.

In **Appendix A**, we provide details of the approximate inference algorithm we use for filtering in a switching linear dynamical system. A complete algorithm for running exact inference in the presence of missing data in an AR-HMM with discrete observations is given in

Appendix B. **Appendix C** shows expectation maximisation derivations for several dynamical models.

Appendix D demonstrates additional novel work on modelling time series with slow linear trends. Based on a particular linear dynamical system parameterisation, we formulate a generative model for such data. We then empirically show that a switching linear dynamical system can successfully discriminate data with linear trends from data without trend.

1.5 Notational conventions

Throughout this thesis scalars will be shown in either lower or upper case italics (e.g y, Y). All vectors are assumed to be column vectors and are written in lower case bold Roman letters (e.g. $\mathbf{v}$). Matrices are always denoted by upper case bold Roman letters (e.g. $\mathbf{M}$). The transposition operation is marked by a T superscript (e.g. $\mathbf{M}^T$).

We also use a shorthand notation for sequences. For instance, the sequence of scalar variables $y_{t_0}, y_{t_0+1}, y_{t_0+2} \ldots, y_{t_1}$ will be denoted as $y_{t_0:t_1}$.

The fact that the random variable $\mathbf{x}$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ will be denoted as: $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The expression $a \perp\!\!\!\perp b | c$ means that random variables a and b are conditionally independent given variable c.

In all the DAGs presented in the following, circles represent continuous variables, squares represent discrete ones, and shaded variables are observed.

Chapter 2

Models for sequences

In this chapter, we review several dynamical models which serve as foundation for the development of our neonatal condition monitoring system. The observed vital signs of a baby receiving intensive care are the result of complex interactions between multiple physiological systems. Thus, a common feature of the modelling techniques discussed is that they attempt to explain the underlying processes that generated the data. For neonatal condition monitoring, such approaches have often facilitated the generally difficult task of a priori knowledge integration.

A probabilistic model for our neonatal condition monitoring task should satisfy several essential requirements.

1. Given a monitoring regime, we need models that accurately capture the evolution of the vital signs channels. These have to account for both the inherent stochastic nature of human physiology and for the noise and artifact processes associated with sensor measurements.

2. The dynamical models should be able to identify the interpretable patterns that intermittently occur on the monitoring traces. The patterns can be associated with various underlying clinical events such as particular states of health, clinical procedures or monitoring equipment operation.

3. We need to account for the fact that the observed data patterns are determined by the interaction of multiple clinical events. Expert knowledge about these interactions is available, and must be incorporated into the model.

4. Once a model has been built, we need to identify some tractable algorithm for inferring the hidden processes from the observed vital signs data and some parameter estimation procedure.

We start with a brief introduction to the Markov chain in Section 2.1. Section 2.2 is concerned with models that employ a hidden discrete structure for explaining time series data. In more detail, we first discuss the Hidden Markov Model (HMM), and then continue with the Autoregressive Hidden Markov Model (AR-HMM) and the Factorial Hidden Markov Model (FHMM). Section 2.3 is dedicated to continuous-variable dynamical models. These can be either fully observed in the shape of autoregressive models or have a hidden continuous state such as the Linear Dynamical System (LDS). Hybrid models (Section 2.4) combine the ideas in the previous sections into models with more representational power such as the Switching Linear Dynamical System (SLDS) and the Factorial Switching Linear Dynamical System (FSLDS).

2.1 Markov chains

The Markov chain is probably the most important concept for modelling *discrete-time* sequential data (see e.g. Grimmett and Stirzaker [2001] or Bishop [2007]). Here, we sketch a probabilistic view. Consider a sequence of data points $y_1, y_2, \ldots, y_T$, which for simplicity we take to be scalars. Their joint probability distribution can be factored as follows using the chain rule:

$$p(y_1, y_2, \ldots, y_T) = p(y_1) \prod_{t=2}^{T} p(y_t | y_1, y_2, \ldots, y_{t-1}). \tag{2.1}$$

Markov chains make the simplifying assumption that the value of the current observation y_t depends only the values of the previous p variables. More formally, we make the conditional independence assumption:

$$y_t \perp\!\!\!\perp y_1, y_2, \ldots, y_{t-p-1} | y_{t-p}, y_{t-p+1}, \ldots, y_{t-1}.$$

This can be interpreted as limiting the "memory" of the process. The positive integer p is commonly referred to as the order of the Markov chain.

Since this still requires operating with a number of conditional probability distributions linear in the length of the sequence, we often add a *time-homogeneity* constraint. That is all the conditional distributions $p(y_t | y_{t-p}, y_{t-p+1}, \ldots, y_{t-1})$ are set to be identical. The joint distribution can then be parsimoniously represented as:

$$p(y_1, y_2, \ldots, y_T) = p(y_1, y_2, \ldots, y_p) \prod_{t=p+1}^{T} p(y_t | y_{t-p}, y_{t-p+1}, \ldots, y_{t-1}) \tag{2.2}$$

The Markov chain cannot be used directly for condition monitoring; however it will be used as a building block by all the models discussed in the following. We will consider both discrete Markov chains (Section 2.2) and continuous ones (Section 2.3). Most often, the chain will not be directly observed, and estimating it will require some type of statistical inference procedure.

2.2 Latent discrete-state models

The following is a discussion of a special class of latent Markov models, where the hidden variables are exclusively discrete. These are referred to as *states* and can be organised either into a single hidden first-order Markov chain as in the HMM (Section 2.2.1) and the AR-HMM (Section 2.2.2) or as a collection of first-order Markov chains as in the Factorial HMM (Section 2.2.3). The space in which the state variables can take values is known as the *state-space*.

The common feature of these discrete-state models is that they share the following two-step data generation process. First, Markov chain states are sampled from some *transition* matrix (i.e. stochastic matrix). This is succeeded by an *emission* process during which observations are drawn from a probability distribution conditioned on the current chain setting. Thus, the value of state variable determines the dynamical regime followed by the data, and the observed sequences can be viewed as a concatenation of regimes.

When applying these models, the usual goal is to infer the states of hidden discrete variables given the observed sequences. Efficient inference routines can be obtained by using the property that all future variables are conditionally independent of the past variables given the present variables.

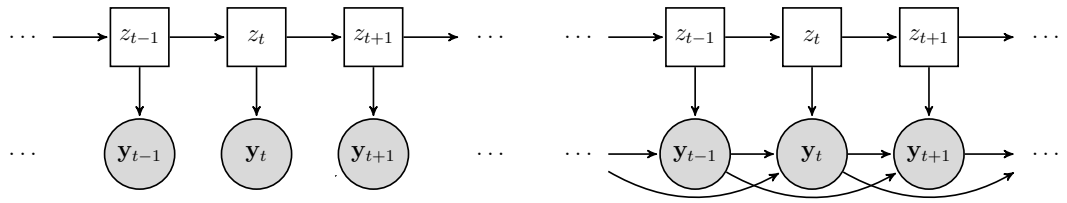
Importantly, the discrete-state models reviewed in this section can be used in both supervised and unsupervised settings. For condition monitoring, we are most interested in supervised modelling as the interpretability of the discrete factors that affect the vital signs data is paramount.

2.2.1 Hidden Markov models

The broad and highly successful applicability of the HMM has been long proven in areas such as speech recognition [Rabiner, 1989], natural language processing [Manning and Schütze, 1999], biological sequence analysis [Krogh et al., 1994], and electrocardiography [Coast et al., 1990, Andreao et al., 2006] to name a few. For a comprehensive review see Rabiner [1989].

Consider a set of d_y-dimensional observations $\mathbf{y}_t \in \mathbb{R}^{d_y}$, $t = 1, 2, \dots T$. If for each data point $\mathbf{y}_t$ we introduce an associated hidden categorical variable z_t, then its distribution $p(\mathbf{y}_t)$ can be modelled as a mixture model: $p(\mathbf{y}_t) = \sum_{z_t} p(z_t) p(\mathbf{y}_t|z_t)$. In addition, if the data points arose as measurements of some sequential process, then the HMM assumes that the discrete z variables compose a first-order Markov chain. The corresponding DAG is given in Figure 2.1a and the joint probability distribution has the following form:

$$p(z_{1:T}, \mathbf{y}_{1:T}) = p(z_1) p(\mathbf{y}_1|z_1) \prod_{t=2}^{T} p(z_t|z_{t-1}) p(\mathbf{y}_t|z_t). \tag{2.3}$$



(a) Standard HMM

(b) AR-HMM with $p = 2$

Figure 2.1: DAGs of several HMM models.

Inference

In the HMM, exact inference can be efficiently performed using the *forward-backward* message passing routine. For our application, the main interest lies in inferring the one-step marginal filtering distribution $p(z_t|\mathbf{y}_{1:t})$. This is obtained by normalising the *forward* messages $\alpha(z_t) \triangleq p(z_t, \mathbf{y}_{1:t})$, which can be computed using the following recursion:

$$\alpha(z_t) = p(\mathbf{y}_t|z_t) \sum_{z_{t-1}} p(z_t|z_{t-1})\alpha(z_{t-1}). \tag{2.4}$$

In an off-line setting, a one-step marginal smoothing distribution can be computed by noting that $p(z_t|\mathbf{y}_{1:T}) \propto \alpha(z_t)\beta(z_t)$, where the *backward* message $\beta(z_t) \triangleq p(\mathbf{y}_{t+1:T}|z_t)$ is determined using the recursion

$$\beta(z_t) = \sum_{z_{t+1}} p(\mathbf{y}_{t+1}|z_{t+1})p(z_{t+1}|z_t)\beta(z_{t+1}). \tag{2.5}$$

Due to the advantage that computing the β messages is independent of computing the α messages, this algorithm is referred to as *parallel* smoothing. Normalised (scaled) versions of the forward and backward recursions are used in practice in order to prevent numerical underflow (see e.g. Bishop [2007, §13.2.4]).

One of the most appealing features of the HMM is that inference is independent of the particular form of the emission process, as long as a normalised probability can be provided.

A different approach to inference is to find the most probable joint sequence of hidden states. This is solved by the Viterbi algorithm, which actually only replaces the summation operations in the forward-backward routine by maximisations.

Learning

In condition monitoring we deal with interpretable discrete factors, and thus can obtain labelled data of the form $\{\mathbf{y}_t, z_t\}$. In this case, the entries of the transition matrix can be estimated as:

$$p(z_t = j|z_{t-1} = i) = \frac{n_{ij} + n_0}{\sum_{j'}(n_{ij'} + n_0)} \tag{2.6}$$

where n_{ij} is the number of transitions from state i to state j counted over all the training data. The constant count n_0 comes from placing a Dirichlet prior on each row of the transition matrix, which prevents probabilities from being too close to zero.

Learning the emission process depends on the modelling choice. If for simplicity we take the conditional distributions to be part of the exponential family, then for estimating $p(\mathbf{y}_t|z_t = j)$ it suffices to compute sufficient statistics from the training samples $\{\mathbf{y}_t, z_t = j\}$.

If labelled data were not available, maximum likelihood parameters can be found via Expectation-Maximisation (EM) [Dempster et al., 1977]. The application of EM to HMM learning is sometimes referred to as the Baum-Welch algorithm.

2.2.2 Autoregressive hidden Markov models

An AR-HMM enhances the HMM architecture by introducing a direct stochastic dependence between observations [Ephraim et al., 1989, Woodland, 1992]. It is designed to explicitly model the (possibly long range) correlations in sequential data. A special class of AR-HMMs, Switching AR (SAR) models, have been widely used in econometrics [Hamilton, 1990] (see Section 2.3.1 for a review of autoregressive models).

In an HMM, the current observation is independent of all the other observations given the current state. Consequently, there is no explicit constraint on time series drawn from an HMM to be smooth. The AR-HMM encourages correlation amongst observations by adding direct dependencies between the current observation and those at the previous p time steps. Technically, in the AR-HMM the emission process is defined by the conditional distribution $p(\mathbf{y}_t|z_t, \mathbf{y}_{t-p:t-1})$. Figure 2.1b shows the DAG of an AR-HMM with $p = 2$.

Samples drawn from an AR-HMM are thus smoother than samples from an HMM, usually making the former a better generative model in many time series problems. We will take advantage of this property in Chapter 5, where we discuss the application of AR-HMM models to neonatal sepsis detection.

Importantly, exact AR-HMM inference only subtly differs from the HMM equivalent (Section 5.2 and Appendix B.1). In addition, learning can be immediately adapted from the HMM routines.

The BP-AR-HMM

The beta process AR-HMM (BP-AR-HMM) [Fox et al., 2010] is an extension of the AR-HMM and was previously applied on intensive care data for unsupervised feature discovery (see Section 4.2.1). The BP-AR-HMM is motivated by the following real world constraints. First, the number of regimes in a time series dataset is often not known a priori. Second, individual sequences might often display only a subset of the regimes. One principled framework to address these constraints is offered by Bayesian nonparametrics (e.g. Teh and Jordan [2010]).

The BP-AR-HMM assumes the following data generation procedure. First, an infinitely large pool of shared regimes is sampled from a prior over regime parameters. Then, for each individual sequence we sample a finite subset of regimes from the pool of shared regimes. Given the subset of regimes, a Markov transition matrix is also sampled, after which the generative process follows the standard AR-HMM. Note that exact inference in the BP-AR-HMM is not possible, and an approximate algorithm has been proposed alongside in Fox et al. [2009].

2.2.3 Factorial hidden Markov models

As previously discussed, a key aspect of neonatal condition monitoring is that the settings of several factors mutually describe the dynamics of the observed data. The Factorial HMM (FHMM) is perhaps the simplest generative model that could be applied to our task.

Originally introduced in Ghahramani and Jordan [1997], the model decomposes the state-space of a standard HMM into a cross-product of K factors $z_t \triangleq f_t^{(1)} \times f_t^{(2)} \times ... \times f_t^{(K)}$. A key assumption is that the factors are a priori independent:

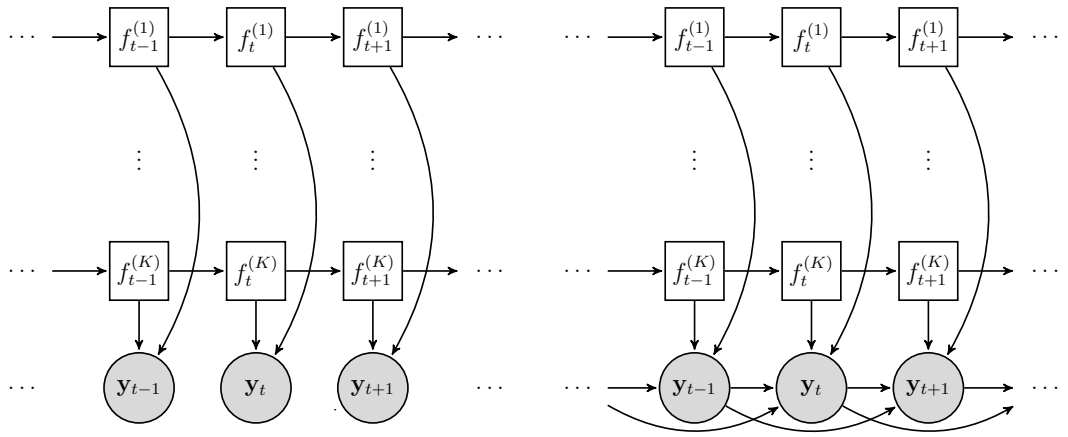
$$p(z_t|z_{t-1}) = \prod_{k=1}^{K} p(f_t^{(k)}|f_{t-1}^{(k)}). \tag{2.7}$$

See the corresponding DAG in Figure 2.2a.

The form of the emission distribution $p(\mathbf{y}_t|f_t^{(1)}, f_t^{(2)}, \ldots, f_t^{(K)})$ is largely unconstrained, but often it is assumed to be Gaussian with mean given by adding individual factor means selected according to the hidden chain settings. This model is sometimes referred to the additive FHMM and has been applied to tasks as such audio separation [Roweis, 2000] and energy disaggregation [Kolter and Jaakkola, 2012].

It is easy to see that the dimension of the state-space is exponential in the number of factors. This means that exact FHMM inference is tractable only in problems with relatively small state-spaces; e.g. 10 to 20 binary factors. For large state-spaces, several approximate inference methods have been applied including the structured mean field approximation [Ghahramani and Jordan, 1997], block Gibbs sampling [Kim et al., 2011] and approximate maximum a posteriori (MAP) inference [Kolter and Jaakkola, 2012].

An interesting candidate model for condition monitoring would be a Factorial AR-HMM (Figure 2.2b), which combines a factored state-space with the correlated observations of the AR-HMM. However, such a model is not the best choice, partly because it is not well-suited for dealing with artifactual measurements. For a more detailed discussion see Quinn [2007, §4.4].



(a) Factorial HMM

(b) Factorial AR-HMM with $p = 2$

Figure 2.2: DAGs of two factorial models.

2.3 Continuous-state models

The models in the previous section attempt to explain sequential data in terms of an underlying hidden discrete structure. In this section we discuss models which focus on explaining time series in terms of transitioning between continuous state variables. The state variables can be either directly observed as in the case of autoregressive models (Section 2.3.1) or hidden as in the case of linear dynamical systems (Section 2.3.2).

2.3.1 Autoregressive models

Autoregressive models are widely used for modelling fully observed time series. Here, we briefly introduce the *AR*, *ARMA* and *ARIMA* models. For comprehensive treatments of this family of models see Brockwell and Davis [1991], Hamilton [1994] or Chatfield [2004].

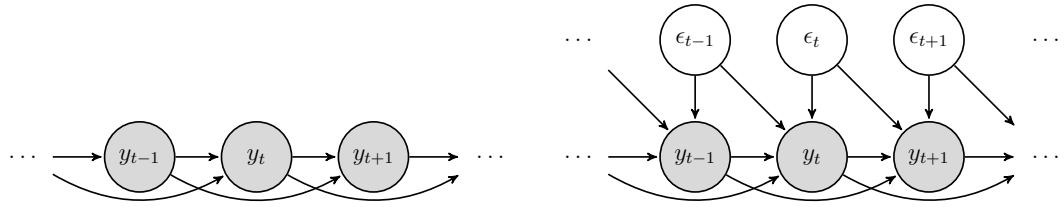
The simplest way to model a vital signs channel is arguably the $AR(p)$ process[1]. It assumes that the value of a stationary[2] time series y_t can be explained by linear regression where the covariates are the previous p observations $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$. Thus, assuming centred data (i.e. zero-mean), an $AR(p)$ model is defined as:

$$y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} = \varepsilon_t, \tag{2.8}$$

where ε_t is Gaussian white noise, $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. The $AR(p)$ model is a special case of the more general family of *ARMA* models for stationary processes.

[1] A (stochastic) process is an ordered collection of random variables. A rigorous treatment is beyond our scope here, but can be found in e.g. Brockwell and Davis [1991, §1].

[2] In this thesis, we restrict ourselves to *weak* stationarity, which means that the first and second moments of a process are time-independent.



(a) AR(2) model (b) ARMA(2,1) model

Figure 2.3: DAGs of autoregressive models

An $ARMA(p,q)$ model describes the data by the following recursion:

$$y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}, \tag{2.9}$$

where $\{\varepsilon_t\}$ is again Gaussian white noise, $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. By setting $p = 0$ we get the special case of moving average (MA) models. In Figure 2.3, we show DAGs corresponding to an $AR(2)$ model and an $ARMA(2,1)$ one. Note that for a generative probabilistic view of $ARMA$ models we defined the conditional distribution $p(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}) = \delta(y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} - \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q})$, where $\delta(\cdot)$ is the Dirac delta distribution.

A useful way to rewrite eq. 2.9 is by using the backward shift operator B, $B^i y_t = y_{t-i}, i \in \mathbb{Z}$. We get:

$$\phi(B) y_t = \theta(B) \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{2.10}$$

where $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ is the autoregressive polynomial and $\theta(B) = 1 + \theta_1 B + \ldots + \theta_p B^q$ is the moving average polynomial.

We will restrict ourselves to the case of *causal ARMA* processes. In a causal dynamical model the output at any time step is independent of all future inputs. This is the usual requirement for any physically realizable system. More formally, we say that an $ARMA(p,q)$ process is causal if it can be written as the following $MA(\infty)$ model

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \tag{2.11}$$

where $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

Then, the following important result holds:

Theorem 2.3.1 *An $ARMA(p,q)$ process where $\phi(z)$ and $\theta(z)$ have no common zeros is causal if only if all the roots of $\phi(z)$ are strictly outside the unit circle [Brockwell and Davis, 1991, Theorem 3.1.1].*

The modelling power of *ARMA* models can be summarised by the following property. Consider a stationary process having some autocorrelation function $\gamma(\cdot)$ converging to zero when $n \to \infty$. Then for any positive integer m there exists an *ARMA* process whose autocorrelation perfectly matches $\gamma(\cdot)$ up to order m [Brockwell and Davis, 1991, §4].

ARIMA models

The *ARIMA* model is a standard tool for modelling non-stationary data. As defined in Section 2.3.1, *ARMA* models only apply to stationary data, which means they cannot capture trends or seasonality. *ARIMA* models assume that *differencing* the data for a certain number of times of times will produce a stationary signal. This can be modelled as an *ARMA* process. The general equation of an $ARIMA(p,d,q)$ model is:

$$\phi(B)(1-B)^d y_t = \theta(B)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,\sigma_\varepsilon^2) \tag{2.12}$$

where again $\phi(B) = 1-\phi_1 B - \ldots - \phi_p B^p$ and $\theta(B) = 1+\theta_1 B + \ldots + \theta_p B^q$.

ARIMA models can be useful for modelling stationary time series as well [Brockwell and Davis, 1991, §9]. A typical scenario is when the autocorrelation function of a stationary time series decays slowly to zero, while the autocorrelation function of the differenced series decays much quicker.

It is important to note that (2.12) determines the second-order statistics of the process $\{(1-B)^d y_t\}$ and not those of $\{y_t\}$. It can be shown [Brockwell and Davis, 1991, §9], that both y_t and $y_t^* \triangleq y_t + C_0 + C_1 t + \ldots + C_{d-1}t^{d-1}, C_k \in \mathbb{R}$ satisfy (2.12).

Vector AR models

Even though we have so far focused on 1-d time series, autoregressive models can be extended to multivariate processes. For instance, a time series $\{\mathbf{y}_t\}$, $\mathbf{y}_t \in \mathbb{R}^{d_y}$ can be explained as a linear regression of the previous p observations:

$$\mathbf{y}_t - \Phi_1 \mathbf{y}_{t-1} - \ldots - \Phi_p \mathbf{y}_{t-p} = \mathbf{v}_t, \tag{2.13}$$

where Φ_i's are $d_y \times d_y$ matrices and $\{\mathbf{v}_t\}$ is a multivariate Gaussian white noise sequence, $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Analogously, the AR-HMM model with continuous observations (i.e the SAR model) introduced in Section 2.2.2 can be extended to multi-dimensional observations. This extension is sometimes referred to as the Switching Vector AR (SVAR) model.

2.3.2 The linear dynamical system

Our choice for modelling the individual regimes present in neonatal monitoring data is the linear dynamical system (LDS) [Kalman, 1960]. Also known as the Kalman filter, the LDS assumes a hidden first-order Markov chain of d_x-dimensional continuous-state variables $\mathbf{x}_t \in \mathbb{R}^{d_x}$ (i.e. a vector $AR(1)$ process) and a noisy observation process connecting this latent process to $d_y$dimensional measurements $\mathbf{y}_t \in \mathbb{R}^{d_y}$ (see Figure 2.4). Both state transition and observation

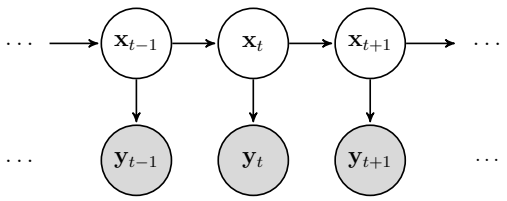


Figure 2.4: DAG of the LDS.

distributions are linear-Gaussian:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \mathbf{A}_t\mathbf{x}_{t-1}, \mathbf{Q}_t), \tag{2.14}$$

$$p(\mathbf{y}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{y}_t; \mathbf{C}_t\mathbf{x}_t, \mathbf{R}_t), \tag{2.15}$$

where $\mathbf{A}_t$ are the square dynamics (system) matrices, $\mathbf{C}_t$ are the observation matrices, and where $\mathbf{Q}_t$ and $\mathbf{R}_t$ are noise covariance matrices. In the following we take the parameters to be time independent, thus discussing the time-homogeneous LDS[3]. Also, to ensure system stability, all the eigenvalues λ_i of the dynamics matrix must lay within the unit circle ($|\lambda_i| \leq 1$). The model definition is completed by introducing an initial conditions distribution for the state-space, $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{x}_1; \mathbf{m}_0, \mathbf{V}_0)$.

Inference

In the LDS, the (marginal) filtering and smoothing distributions can be exactly computed. Both computations are performed recursively.

(Kalman) filtering can be understood as a two step process. First, the *prediction* step estimates the distribution of the hidden state at time t given all historical observations up to time $t-1$:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1}. \tag{2.16}$$

The prediction can be further projected onto the observed variable space, to give the predictive distribution of $\mathbf{y}_t$ given the history:

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})\mathrm{d}\mathbf{x}_t. \tag{2.17}$$

After observing $\mathbf{y}_t$, this result is used to compute the log-likelihood $l_t \triangleq \log p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$.

The subsequent *correction* step incorporates the information brought by observing $\mathbf{y}_t$ into the state estimate:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})}. \tag{2.18}$$

Since eqs. 2.16 to 2.18 exclusively involve Gaussian distributions, Kalman filtering only requires the forward propagation of first- and second-order moments. The resulting recursive

[3]Inference is not simplified by this assumption, but learning will be less involved.

algorithm is shown in Figure 2.5. Note that the matrix recursions are independent of the data, and thus can be pre-computed.

The matrix $\mathbf{K}_t$, known as the Kalman gain matrix, plays an essential role in understanding LDS filtering. For ease of explanation, assume that both states and observations are 1-dimensional ($d_x = d_y = 1$). By examining eq. 2.24, we can see that the filtering mean is the sum of the predicted mean and the difference between the observed and predicted measurements, weighted by the Kalman gain. In addition, the Kalman gain is inversely proportional to the observation noise variance (eq. 2.23). A larger observation noise results in a smaller Kalman gain, and thus the filter will trust the predicted mean more than the current observation. Conversely, if the observation noise is small compared to the predicted noise, the current observation will have a stronger influence on the filtering mean. Also note that conditioning on the current observation reduces the uncertainty of the estimate (eq. 2.25).

In neonatal monitoring, we use LDS smoothing solely during parameter estimation. A *sequential* smoothing procedure can be run after completing the filtering recursions. The backwards recursion is:

$$p(\mathbf{x}_t|\mathbf{y}_{1:T}) = \int p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t}) p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T}) \mathrm{d}\mathbf{x}_{t+1} \tag{2.26}$$

The key term here is $p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t})$, which is suggestively referred to as the *dynamics reversal* term. It can be obtained using:

$$p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t}) \propto p(\mathbf{x}_t|\mathbf{y}_{1:t}) p(\mathbf{x}_{t+1}|\mathbf{x}_t). \tag{2.27}$$

Again, all the distributions involved are Gaussian, and the smoothed mean $\tilde{\mathbf{x}}_t$ and covariance $\tilde{\mathbf{V}}_t$ are given by:

$$\overleftarrow{\mathbf{A}}_t = \hat{\mathbf{V}}_t \mathbf{A} \left(\mathbf{A} \hat{\mathbf{V}}_t \mathbf{A}^T + \mathbf{Q} \right)^{-1} \tag{2.28}$$

$$\tilde{\mathbf{x}}_t = \hat{\mathbf{x}}_t + \overleftarrow{\mathbf{A}}_t \left(\tilde{\mathbf{x}}_{t+1} - \mathbf{A} \hat{\mathbf{x}}_t \right) \tag{2.29}$$

$$\tilde{\mathbf{V}}_t = \hat{\mathbf{V}}_t + \overleftarrow{\mathbf{A}}_t \tilde{\mathbf{V}}_{t+1} \overleftarrow{\mathbf{A}}_t^T - \overleftarrow{\mathbf{A}}_t \mathbf{A} \hat{\mathbf{V}}_t^T \tag{2.30}$$

This recursion was originally proposed in Rauch et al. [1965]. A more recent description is available in Barber [2012, §24].

Learning

Most often it is not possible to obtain access to ground truth for the state of an LDS. Fortunately, there are several widely applied approaches to unsupervised learning of LDS parameters. Also note that the parameters can be identified only up to a similarity transform (see e.g. [Barber, 2012]).

Ghahramani and Hinton [1996] have proposed using EM for maximizing the observed data log-likelihood. This uses the inference algorithm shown above in the E-step. An interesting

Denote:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \mathcal{N}(\mathbf{x}_t; \hat{\mathbf{x}}_t^-, \hat{\mathbf{V}}_t^-)$$
$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \mathcal{N}(\mathbf{y}_t; \hat{\mathbf{y}}_t, \mathbf{S}_t)$$
$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \mathcal{N}(\mathbf{x}_t; \hat{\mathbf{x}}_t, \hat{\mathbf{V}}_t)$$

Initialise the recursions:

$$\mathbf{K}_1 = \mathbf{V}_0\mathbf{C}^T\left(\mathbf{C}\mathbf{V}_0\mathbf{C}^T + \mathbf{R}\right)^{-1}$$
$$\hat{\mathbf{x}}_1 = \mathbf{m}_0 + \mathbf{K}_1\left(\mathbf{y}_1 - \mathbf{C}\mathbf{m}_0\right)$$
$$\hat{\mathbf{V}}_1 = \left(\mathbf{I} - \mathbf{K}_1\mathbf{C}\right)\mathbf{V}_0$$

For $t = 2$ to T do:

1. **Prediction**

$$\hat{\mathbf{x}}_t^- = \mathbf{A}\hat{\mathbf{x}}_{t-1} \tag{2.19}$$
$$\hat{\mathbf{V}}_t^- = \mathbf{A}\hat{\mathbf{V}}_t\mathbf{A}^T + \mathbf{Q} \tag{2.20}$$
$$\hat{\mathbf{y}}_t = \mathbf{C}\hat{\mathbf{x}}_t^- \tag{2.21}$$
$$\mathbf{S}_t = \mathbf{C}\hat{\mathbf{V}}_t^-\mathbf{C}^T + \mathbf{R} \tag{2.22}$$

2. **Correction**

$$\mathbf{K}_t = \hat{\mathbf{V}}_t^-\mathbf{C}^T\left(\mathbf{C}\hat{\mathbf{V}}_t^-\mathbf{C}^T + \mathbf{R}\right)^{-1} \tag{2.23}$$
$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_t^- + \mathbf{K}_t\left(\mathbf{y}_t - \mathbf{C}\hat{\mathbf{x}}_t^-\right) \tag{2.24}$$
$$\hat{\mathbf{V}}_t = \left(\mathbf{I} - \mathbf{K}_t\mathbf{C}\right)\hat{\mathbf{V}}_t^- \tag{2.25}$$

Figure 2.5: Kalman filtering

observation is that while the LDS is not identifiable, EM always finds the same solution given the same initialisation.

A different approach to LDS parameter estimation is taken by subspace methods [Overschee and Moor, 1996, Boots, 2012]. Several versions of these algorithms have been proposed, but here we sketch the general framework following Gibson and Ninness [2000]. The central idea is that all the "future" observations $\mathcal{Y}^+$ can be regressed on the complete "history" $\mathcal{Y}^-$. The regression consists of two successive projections ($\mathcal{Y}^+ \approx O\mathcal{K}\mathcal{Y}^-$), where $\mathcal{K}\mathcal{Y}^-$ is a projection on the *predictor* space, the space of the hidden states $\mathbf{x}$. Subsequently, the matrix O is used to project the hidden states onto the space spanned by the future observations. The regression is then represented in terms of second-order moments which can be estimated from data. LDS parameters are subsequently fitted by applying a singular value decomposition to the learnt regression coefficient. Subspace methods can be used to initialise EM, and there has been work on understanding the relationship between these two estimation algorithms [Gibson and Ninness, 2000].

Relation to HMMs

Even though the HMM and the LDS have historically developed in different communities, they are intimately related as probabilistic graphical models in the field of machine learning. The HMM and the LDS are both special cases of *state-space models*, which share the same belief network but place no assumptions on the type of the underlying distributions. In addition, the HMM is the dynamical version of mixture models, while the LDS can be understood as a sequential extension of factor analysis. A unified view on these models can be found in Roweis and Ghahramani [1999]. In addition, both HMM and LDS inference algorithms are special cases of belief propagation [Pearl, 1988], an efficient method for running inference on tree-structured models by message passing.

ARMA models in state-space form

The general approach we took for modelling vital signs channels is to assume they can be explained as noisy observations of hidden *ARMA* processes. In the following, we show how such models can be represented as an LDS with a particular parameterisation. The immediate advantage is that exact inference can be run with the standard Kalman recursions.

Let us assume that we cannot directly observe an $ARMA(p,q)$ process $\{Z_t\}$, but can obtain realisations of a noisy version of it, $\{Y_t\}$, and that the added noise is Gaussian. Using the definitions of Section 2.3.1 this process can be formalised as:

$$Y_t = Z_t + \omega_t, \qquad \omega_t \sim \mathcal{N}(0, \sigma_\omega^2) \tag{2.31}$$

$$\phi(B)Z_t = \theta(B)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{2.32}$$

It is equivalent to the following LDS:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}) \tag{2.33}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \tag{2.34}$$

where

$$\mathbf{x}_t = \begin{bmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-r+2} \\ X_{t-r+1} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_{r-1} & \phi_r \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix},$$

$$\mathbf{y}_t = \begin{bmatrix} Y_t \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 1 & \theta_1 & \dots & \theta_{r-2} & \theta_{r-1} \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \sigma_\omega^2 \end{bmatrix},$$

$r = \max(p, q+1)$, $\phi_k = 0$ for $k > p$ and $\theta_k = 0$ for $k > q$. This example of parameterisation has been previously used in Brockwell and Davis [1991, §12.1]. The initial conditions' parameters can be chosen as $\mathbf{m}_0 = [\mu_0, \mu_0, \dots, \mu_0]^T$ and $\mathbf{V}_0 = diag([\sigma_0^2, \sigma_0^2, \dots, \sigma_0^2])$. It will also be useful to note that $Z_t = \mathbf{C}\mathbf{x}_t$.

In Appendix C.1, we show how the state-space representation can be used for EM learning of an $ARMA(p,q)$ process from noisy observations.

2.4 Hybrid models

Many dynamical systems produce complex time series that cannot be satisfactorily modelled by a single LDS (Section 2.3.2). In Section 2.2.1 we have discussed several sequential models which explain time series by dividing them into multiple segments. Conditioned on the segment, these models become fully observed. However, if we model each of these segments as an LDS we get a richer family of models, sometimes referred to as hybrid models. These are characterised by a *discrete-continuous* hybrid hidden state, where the discrete variables determine the current segment, and conditioned on their settings, the continuous variables capture the hidden dynamics of the observed process.

2.4.1 Switching linear dynamical systems

The most popular type of hybrid model is the Switching Linear Dynamical system (SLDS)[4]. It is a generative model for sequential data which switches between S different modes of operation (i.e. regimes). Each mode of operation is modelled as a LDS (Kalman filter), and thus the SLDS can be thought of as a dynamical mixture of LDS models. As the switch settings are hidden,

[4]The model has been also referred to as the Switching Kalman Filter (SKF), Switching State-Space Model (SSSM) or Jump Markov linear System (JMLS).

often the main task is to recover them given the observations. Formally, at time t the SLDS has a discrete-continuous hybrid hidden state consisting of a hidden switch variable s_t and a hidden continuous state $\mathbf{x}_t \in \mathbb{R}^{d_x}$. This hybrid state attempts to explain how measurements $\mathbf{y}_t \in \mathbb{R}^{d_y}$ are generated. More precisely, at any time step t a switch variable s_t sampled from a Markov transition matrix determines the set of LDS parameters:

$$s_t \sim Categorical(\mathbf{\Pi}_{s_{t-1}}) \tag{2.35}$$

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; \mathbf{A}(s_t)\mathbf{x}_{t-1}, \mathbf{Q}(s_t)), \tag{2.36}$$

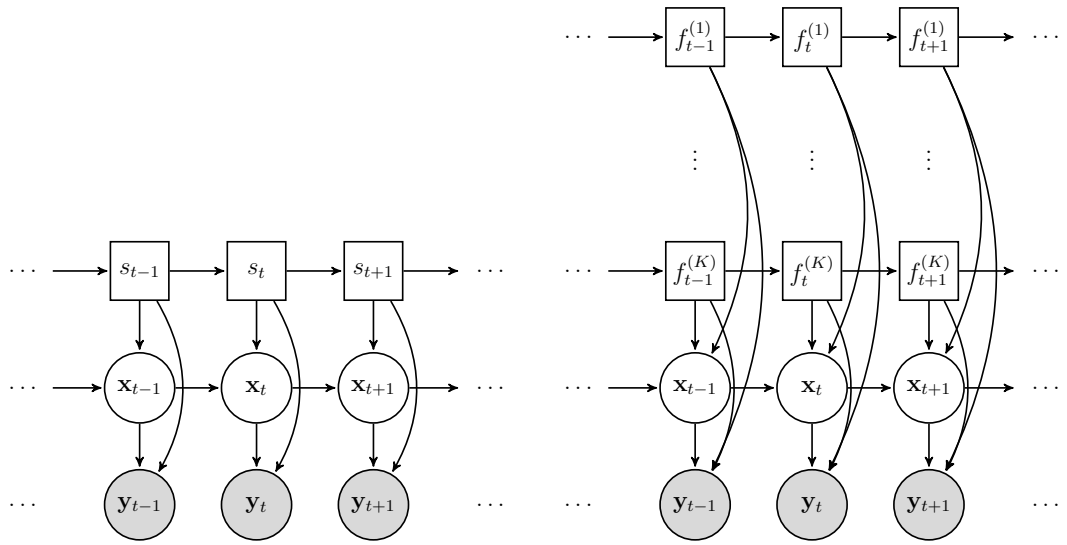
$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{y}_t; \mathbf{C}(s_t)\mathbf{x}_t, \mathbf{R}(s_t)), \tag{2.37}$$

where $\mathbf{\Pi}_{s_{t-1}}$ is the s_{t-1}-th row of the stochastic matrix $\mathbf{\Pi}$. $\mathbf{A}(s_t)$ and $\mathbf{Q}(s_t)$ are the dynamics and dynamics noise covariance matrices, respectively. $\mathbf{C}(s_t)$ and $\mathbf{R}(s_t)$ are the observation and observation noise covariance matrices, respectively. The corresponding DAG is shown if Figure 2.6a. Note that special cases such as switching the parameters of the dynamics or observation processes only, and SLDS models with subtly modified DAGs have been discussed in the literature.

SLDS models have been used in a wide variety of domains, and here we briefly enumerate a few. Navigation and multiple target tracking are treated in Shumway and Stoffer [1991] and Bar-Shalom et al. [2001]. SLDS models for speech recognition have been developed in e.g. Droppo and Acero [2004] and Mesot and Barber [2007]. Modelling financial data is discussed in Kim [1994] and Azzouzi and Nabney [1999]. Medical applications include modelling creatinine levels in patients with kidney transplants [Smith and West, 1983] and modelling the respiration force of a patient with sleep apnea [Ghahramani and Hinton, 2000]. In addition, de Freitas et al. [2004] use an SLDS for automated fault diagnosis in mobile robots, while Morales-Menéndez et al. [2002] and Lerner et al. [2000] employ the same approach for industrial process monitoring. The problem of modelling human motion is tackled in Pavlovic et al. [2000].

Inference

As in the previous sections, the two main objectives in SLDS inference are computing the filtering distribution $p(s_{1:t}, \mathbf{x}_{1:t}|\mathbf{y}_{1:t})$ and the smoothing distribution $p(s_{1:T}, \mathbf{x}_{1:T}|\mathbf{y}_{1:T})$. The main difficulty is that exact inference in SLDS models is computationally intractable. The intuition behind this issue is that the posterior switching variable probabilities at time t need to account for all the possible combinations of switch settings between times t and $t-1$. This results into exact posterior distributions having a number of (Gaussian) components exponential in the length of the sequence, and thus being intractable to compute. A formal treatment of the problem can be found in Lerner and Parr [2001].



(a) Switching LDS (b) Factorial Switching LDS

Figure 2.6: DAGs of multiple regime dynamical mnodels.

Approximate SLDS inference algorithms have been widely studied. In the following, we review some of these approaches focusing on estimating the marginal filtering distribution $p(s_t, \mathbf{x}_t|\mathbf{y}_{1:t})$, as this is more relevant to our application.

SLDS filtering algorithms fall into two classes of approximations: deterministic methods and non-deterministic (sampling) methods.

Deterministic methods approximate the true intractable distribution p by a simpler tractable distribution q. For SLDS filtering, the true marginal continuous state posterior is a mixture of of S^t Gaussians:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \sum_{s_{1:t}} p(\mathbf{x}_t|s_{1:t}, \mathbf{y}_{1:t}) p(s_{1:t}|\mathbf{y}_{1:t}). \tag{2.38}$$

This will be approximated by $q(\mathbf{x}_t|\mathbf{y}_{1:t})$, a mixture of Gaussians with a much smaller number of components. Importantly, the number of components in q is set to be time independent. Similar ideas, but applied to non-linear dynamical systems, date back to the work of Alspach and Sorenson [1972]. These methods are often referred to as Gaussian Sum approximations.

For neonatal conditioning monitoring we use the algorithm described in Murphy [1998], and also referred to as Generalised Pseudo-Bayes 2 (GPB2) [Bar-Shalom et al., 2001]. At each time step, it keeps the continuous state marginal posterior $q(\mathbf{x}_t|\mathbf{y}_{1:t})$ as a mixture of S Gaussians with one component for each possible switch setting:

$$q(\mathbf{x}_t|\mathbf{y}_{1:t}) = \sum_{s_t} q(\mathbf{x}_t|s_t, \mathbf{y}_{1:t}) q(s_t|\mathbf{y}_{1:t}). \tag{2.39}$$

Running the Kalman updates for each possible setting of s_{t+1} produces a posterior distribution $q^*(\mathbf{x}_{t+1}|\mathbf{y}_{1:t+1})$, which is a mixture of S^2 components of the type $q(\mathbf{x}_{t+1}|s_{t+1}, s_t, \mathbf{y}_{1:t+1})$. This

is then collapsed onto a mixture of S components $q(\mathbf{x}_{t+1}|\mathbf{y}_{1:t+1})$ using e.g. moment matching. The full algorithm is given in Appendix A.

Several variations of these ideas are possible. Generalised Pseudo-Bayes 1 (GPB1) [Bar-Shalom et al., 2001] is a computationally cheaper but less accurate alternative, where $q(\mathbf{x}_t|\mathbf{y}_{1:t})$ is collapsed onto a single Gaussian before applying the Kalman updates. The Interacting Multiple Model (IMM) [Bar-Shalom et al., 2001] approximates GPB2 at the cost of GPB1. This is achieved via a different application of collapsing before the Kalman updates. We first compute $q^*(\mathbf{x}_t|s_{t+1},\mathbf{y}_{1:t}) = \sum_{s_t} q(\mathbf{x}_t|s_t,\mathbf{y}_{1:t})q(s_t|s_{t+1},\mathbf{y}_{1:t})$ and then we use moment matching to approximate this mixture by a single Gaussian $q(\mathbf{x}_t|s_{t+1},\mathbf{y}_{1:t})$. The S approximations thus obtained are independently employed as priors for the S filters. A more accurate but more expensive version is discussed in Barber and Mesot [2006], where each "component" $q(\mathbf{x}_t|s_t,\mathbf{y}_{1:t})$ is itself a mixture of I Gaussians, and thus $q(\mathbf{x}_t|\mathbf{y}_{1:t})$ becomes a mixture of $I \times S$ Gaussian components.

Gaussian Sum approximations can be viewed as the application of Assumed Density Filtering (ADF) [Lauritzen, 1992] to SLDS inference. Furthermore, ADF is a special case of deterministic inference methods that minimize the Kullback-Leibler divergence between a true and approximate distribution ($\min_q \mathbb{KL}(p \parallel q)$[5]). However, a rigorous treatment of these connections must take into account the multiple ways in which the collapsing operation can be performed, and is beyond our scope.

Particle filters (PF) are sampling methods often employed for inference in sequential models (see Kantas et al. [2009] for a review). Noting that given the switch settings SLDS inference becomes equivalent to LDS inference, the model is suitable for the application of the Rao-Blackwellised Particle Filter (RBPF) algorithm [Murphy and Russell, 2001]. RPBF reduces the size of the sampling space, and thus improves efficiency. Furthermore, it approximates the continuous-state filtering distribution as a mixture of Gaussians, instead of a sum of Dirac delta distributions as in the standard PF. The algorithm relies on the forward propagation of N particles labelled by $n = 1,\ldots,N$, each consisting of a switch state s_t^n and associated estimates of the continuous-state mean $\mathbf{x}_t^n$ and variance $\mathbf{V}_t^n$. The simplest RBPF would then sample N times from $p(s_{t+1}|s_t^n)$ to get N values of the switch state $\hat{s}_{t+1}^n$. Conditioned on the sampled switch setting pairs $\{s_t^n, \hat{s}_{t+1}^n\}$, we then run Kalman filter updates. This also involves computing the conditional likelihoods $p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t}, s_t^n, \hat{s}_{t+1}^n)$, which serve as particle weights. These weights are then used in a re-sampling step, during which particles are either multiplied or discarded. More accurate sampling distributions can be obtained with the *look-ahead* RBPF [de Freitas et al., 2004], which samples switch states from the optimal sampling distribution. The greater the number of particles used, the bigger the trade-off of speed in favour of accuracy, but in general some increase in speed over deterministic methods is possible as here Kalman updates are usually not computed for all switch setting combinations.

[5] $\mathbb{KL}(p \parallel q) \triangleq \mathbb{E}_p\left[\ln \frac{p}{q}\right]$.

As in the filtering case, several SLDS smoothing algorithms have been studied. Ghahramani and Hinton [2000] have proposed a structured mean field approximation, while Zoeter and Heskes [2003] have adapted expectation propagation [Minka, 2001] for SLDS inference. Barber and Mesot [2006] have developed the state-of-the-art Gaussian Sum approximation called Expectation Correction (EC).

Learning

If the SLDS regimes are interpretable, then some labelled data of the form $\{\mathbf{y}_t, s_t\}$ can be obtained. In this case, learning becomes equivalent to learning one LDS per switch setting. These LDS models can be independently fit as explained in Section 2.3.2, while the transition matrix can be estimated via data counts as discussed in Section 2.2.1.

In the fully unsupervised case, an SLDS can be learnt via EM, where the E-step can use any of the smoothing algorithms enumerated above. Note that variational methods such as the one proposed by Ghahramani and Hinton [2000] have the advantage that each EM step is guaranteed to improve a lower bound on the observed data likelihood [Neal and Hinton, 1998].

2.4.2 The Factorial SLDS

When several discrete factors determine the switch setting of an SLDS, it makes sense to marry it to the FHMM model (Section 2.2.3). The resulting Factorial SLDS (FSLDS) is obtained by representing the switch variable of the SLDS as the cross-product $f_t^{(1)} \times f_t^{(2)} \times ... \times f_t^{(K)}$ [Williams, Quinn, and McIntosh, 2006]. An important assumption made by the FSLDS is that the factors are a priori independent:

$$p(s_t|s_{t-1}) = \prod_{k=1}^{K} p(f_t^{(k)}|f_{t-1}^{(k)}).$$

See the DAG in Figure 2.6b.

The number of possible values the switch variable s_t can take on grows exponentially with the number of factors. As a reminder, approximate inference methods such as GPB2 require S^2 Kalman updates at each time step. In order to reduce computational expense, one may assume that at any time step at most one factor $f_t^{(k)}$ can change its setting [Quinn, 2007, Kolter and Jaakkola, 2012]. For GPB2, this reduces the number of Kalman updates per time step from order S^2 to order $S \log S$.

Quinn et al. [2009] have applied the FSLDS for the detection of physiological and artifactual patterns in neonatal monitoring data. This research is intimately related to our work on sepsis detection (Chapter 6), and we separately provide a detailed description of their application in Section 4.2.2. Cemgil et al. [2006] use the FLSDS as a generative model for polyphonic music transcription. Binary factors store the states of a set of sound generators, while the

continuous state variables model the dynamics of damped oscillators, one for each generator. The observed audio signal is the superposition of the outputs from all the sound generators. FSLDS-like models had also been used to model speech. In the framework set up in Deng [2006], the discrete factors represent different tiers of a phonological model. The continuous state variable are of two types. First, there are states corresponding to phonetic targets which are directly affected by the phonological model. The phonetic targets determine the second type of continuous states, which correspond to the articulatory dynamics. Finally, the visible variables are the observed acoustic signal.

2.5 Summary

This chapter has introduced several models employable for the task of monitoring the physiology of a prematurely born infant receiving intensive care. The methodology described here simplifies understanding the neonatal monitoring FSLDS reviewed in Section 4.2.2, and supported the development of the discrete state sepsis detection framework presented in Chapter 5 and of the hierarchical condition monitoring approach discussed in Chapter 6.

Chapter 3

Neonatal physiological monitoring

It it beyond doubt that the success of a machine learning application heavily depends on the judicious consideration of domain specifics. This type of analysis places practitioners in an advantageous position when incorporating a priori knowledge into their models. In this chapter, we provide a data-focused introduction to neonatal physiological monitoring.

The state of health of an infant receiving intensive care is not an observable quantity, so clinicians must perform some type of inference to articulate their beliefs. These take the shape of a clinical diagnosis, further used to decide on the best course of treatment. A central purpose for continuously monitoring the vital signs of an ICU patient is to provide additional evidence for clinicians to refine their diagnosis. This practise is strongly supported by the fact that patterns in the monitoring data can be associated with different states of health.

Our approach is to apply machine learning to the vital signs data in order to produce useful inferences about the patient's condition. This task is not straightforward. A first set of reasons is related to the intrinsic complexity of the human body. Physiological measurements are the result of intricate interactions between several regulatory systems. While certain stereotypical patterns are caused by known conditions, infants display individual physiological dynamics. Furthermore, certain combinations of illnesses may results into entirely novel monitoring patterns. A second set of reasons is related to clinical procedures and to the operation of the monitoring equipment. In these cases, measurements do not reflect the patient's true state of health, and should not be considered as evidence for diagnosis.

In the following sections, we expand these ideas and support our explanations with illustrative examples of the common patterns. Section 3.1 explains how neonatal physiology is translated into numerical readings via dedicated monitoring devices. We then turn our attention to the patterns present in this data. Section 3.2 introduces the most common state of health which is stability. We then discuss patterns associated with interpretable physiological events in Section 3.3. Section 3.4 presents a set of situations in which data are corrupted by known mechanisms of artifact. In Section 3.5 we treat other generally less common types of

physiological measurement variation.

3.1 Monitored physiological channels

Very Low Birth Weight (VLBW) babies are usually nursed in the controlled environment of an incubator, which allows a fine adjustment of temperature and humidity levels. A large number of items of equipment collecting physiological measurements surround the incubator, and these are generally connected to cotside monitors. The latter allow real-time visualisation of short-term channel dynamics, and can sometimes fire alarms when certain signal thresholds are passed. Importantly, the signals collected by some of these equipment are stored on internal hospital servers for subsequent analysis.

One way to understand the set of clinical probes used in this thesis is by classifying them according to the physiological system they monitor. We discuss three such systems: the respiratory system, the cardiovascular system and the thermoregulatory system.

The respiratory system plays the role of regulating the quantities of various gases present in the blood. The regulatory mechanism is largely understood as oxygen absorption and carbon dioxide dispersion. Here, we had access to the readings of a pulse oximeter attached to one of the infant's feet. The oximeter shines (red and infrared) light through the tissue, and measures the absorbed spectrum [DeMeulenaere, 2007, Salyer, 2003]. The mesurement varies with the ratio of oxygenated and deoxygenated hemoglobin, and is used to derive the proportion to which the patient is able to utilise its capacity of carrying oxygen in the blood. This proportion will be referred to as oxygen saturation (SO)[1], and is measured in percent. Note that pulse oximetry is known to have several limitations including motion artifact, nurse care, ambient lighting, and poor peripheral perfusion [Salyer, 2003, Gerstmann et al., 2003].

The circulation of blood is regulated by the cardiovascular system. This function is monitored by measuring the rate of the heart and the pressure of the blood. Heart rate readings are available from two sources. The primary source is electrocardiography (ECG), which continuously measures the heart's impedance by passing current through several leads attached to the patient's body. For each heart beat, the resulting waveform displays a noticeable sharp peak, the R wave. The time between two R waves is the called the RR interval. Such data are sometimes directly available, but for our work we had access to a processed summary of the waveform, namely the heart rate (HR) measured in beats per minute. Our second source of heart rate readings was the pulse oximeter. This is the PR trace, and its relationship with the HR will be discussed in Section 5.4.2. Measurements of two blood pressure channels are recorded by a pressure transducer connected to the arterial lines. When the heart is contracting, a first channel (BS) captures the systolic blood pressure; when the heart is at rest, the second

[1] Oxygen saturation is often referred to as SpO_2 in the biomedical literature.

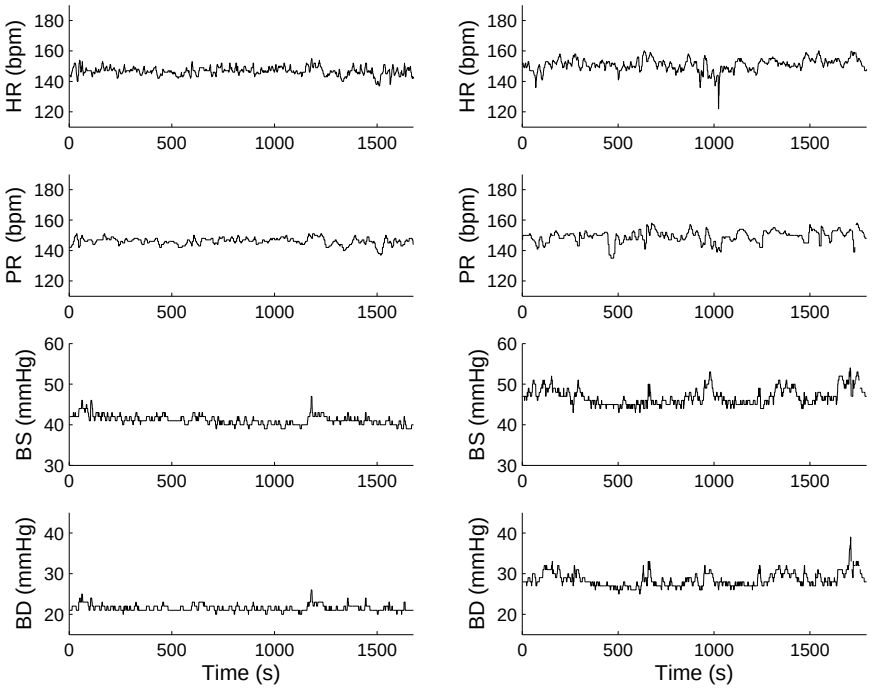


Figure 3.1: Examples of stable dynamics on the cardiovascular system monitoring channels. The first column corresponds a baby of 26 weeks gestation, while the second one to a baby of 25 weeks gestation. Note that the second patient exhibits more physiological variation and higher blood pressure levels.

channel (BD) captures the diastolic blood pressure. Both channels are measured in mmHg.

The thermoregulatory system is responsible for maintaining the body's temperature. In our work, it is monitored by two probes: a core temperature probe (TC) placed under the baby's back and a peripheral temperature probe (TP) attached to one foot. All temperatures are measured in degrees Celsius (°C).

3.2 Stable physiological dynamics

The physiological regime most frequently explaining the monitoring data is *stability*. NICU patients are VLBW babies, which are often treated for prematurity alone. Most of time they are asleep and motionless, and the vital signs traces do not display any type of acute physiology or monitoring device artifact. We define such intervals as periods of *stable* physiological dynamics. Note that this definition does not exclude the possibility that the patient is being treated for some specific pathology or that they suffer from some chronic condition.

We show examples of *stable* physiological evolution on the cardiovascular system channels in Figure 3.1 and on the respiratory and thermoregulatory system channels in Figure 3.2. All traces are characterised by low variability, but note that different babies have different stable

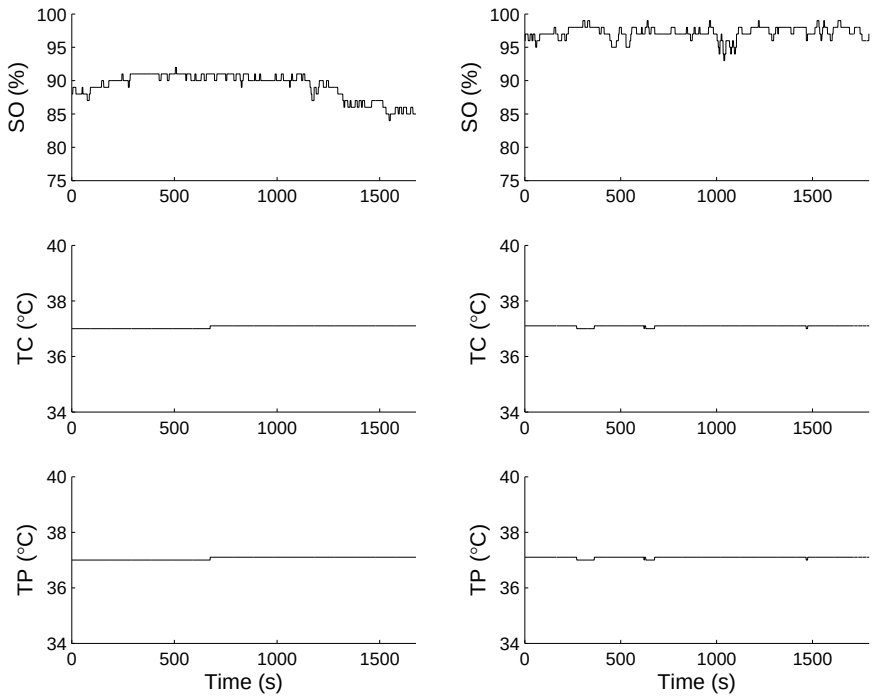


Figure 3.2: Examples of stable dynamics on monitoring channel corresponding to the respiratory and cardiovascular systems. The two columns show data from the same two patients as in Figure 3.1.

dynamics. This is particularly obvious in the case of the higher blood pressure and oxygen saturation levels displayed by the second patient. We will return to this issue in the next chapter, when we discuss the calibration of the condition monitoring FSLDS.

3.3 Physiological events

We now turn our attention to stereotypical patterns appearing in the monitoring data. When such patterns reflect the true values of the patient's vital signs, we refer to them as physiological events. In this section we discuss two types of physiological events: bradycardias and oxygen desaturations. These will play a central role in the sepsis detection frameworks discussed in Chapters 5 and 6. In Appendix D, we will briefly introduce another physiological event, pneumothorax.

Bradycardia

Neonatal bradycardia is a drop in heart rate measurements caused by a slowing of the heart. The possible causes are many, and only some of them are of serious clinical concern. An explanation of the mechanisms involved in neonatal bradycardias can be found in Miller et al. [2000]. Several examples of bradycardia episodes are shown in Figure 3.3.

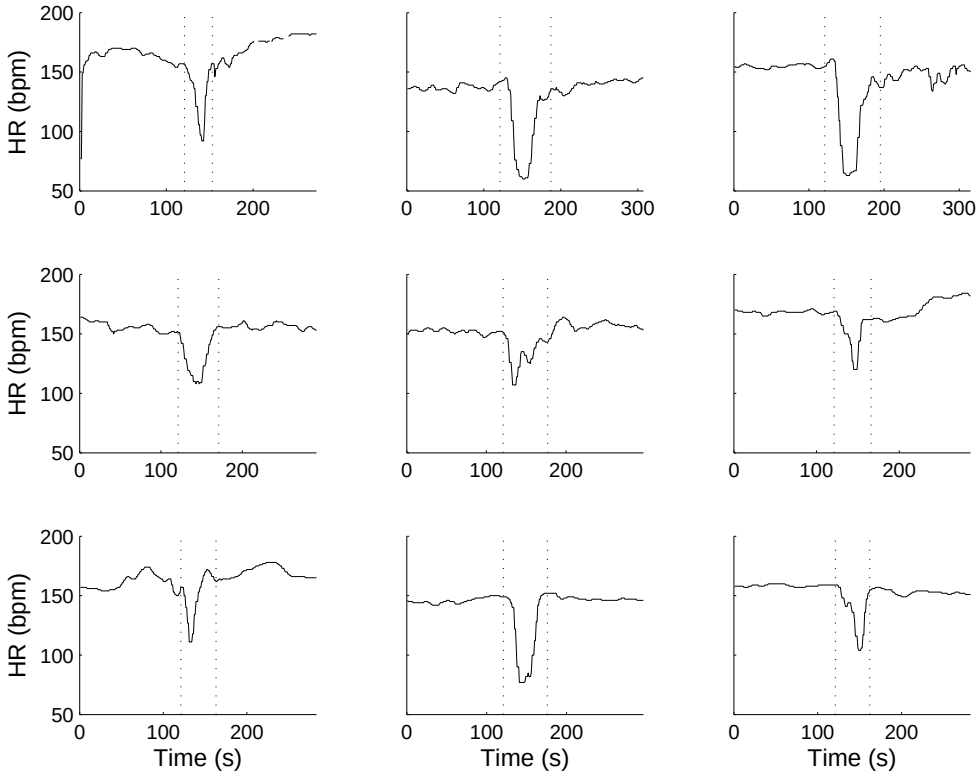


Figure 3.3: Examples of bradycardia selected from several different infants. The vertical dotted lines mark the start and end of a bradycardia episode, and have been provided by an expert annotator.

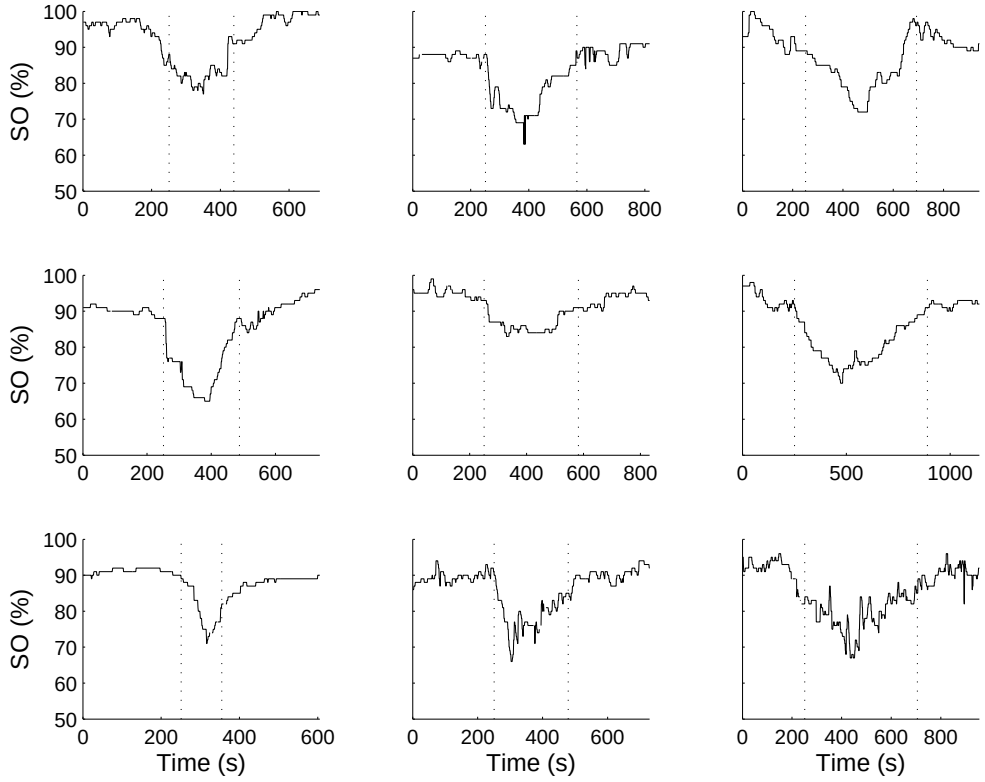


Figure 3.4: Examples of oxygen desaturation selected from several different infants. The vertical dotted lines mark the start and end of a desaturation episode, and have been provided by an expert annotator.

Desaturation

Another stereotypical physiological pattern seen in the monitoring data is oxygen desaturation. The event is characterised by a drop in the saturation of oxygen in arterial blood. Its occurrence is often connected to sleep apnea, a condition when the patient intermittently stops breathing during sleep (See Martin and Fanaroff [1998] for a more detailed discussion of desaturations). SO channel stability is sometimes restored by increasing the concentration of oxygen supplied to the baby [Quinn, 2007, §2.42.].

It is important to mention that 95% confidence limits for pulse oximeters are ± 4 for the interval $70 - 100\%$ [Salyer, 2003]. The same paper also argues that even if SO readings below 70% are less accurate, this is of less importance, as patients who desaturate to such levels need aggressive treatment regardless of measurement precision. Gerstmann et al. [2003] report that pulse oximeters tend to overestimate oxygen saturation below approximately 90%, and that the bias worsens for lower SO values.

Figure 3.4 shows several examples of desaturations selected from different patients. In comparison to bradycardias, these have a longer duration.

3.4 Artifactual events

The physiological channels' recordings do not always reflect the true values of a patient's vital signs. Many of these cases can be explained by known mechanisms of arifact which produce recognisable patterns on the monitoring traces. In the following, we illustrate the main artifactual events occurring in neonatal condition monitoring.

Probe disconnection

Probe disconnections are characterised by a lack of monitoring data due to the temporary removal or malfunctioning of the monitoring devices. In general, when a probe is disconnected, the zero value is stored. The exception are the physiological temperature channels which may decay to the incubator's temperature level.

Blood Sampling

Blood samples are regularly taken for laboratory testing. This involves collecting blood from the arterial line containing the pressure transducer. In order to keep it clean, the line is connected to a pump which releases a saline solution. When taking blood, the pump is blocked by changing the setting of a three-way tap, and ends up acting against the blood pressure sensor. The effect is an approximately linear build-up of pressure, which can be seen as the artifactual ramps in Figure 3.5.

Oximeter error

As already mentioned in Section 3.1, oxygen saturation readings provided by the oximeter are not always accurate. Fortunately, we can test the reliability of the SO trace by using the heart rate measurements also output by the oximeter. If these readings (PR trace) disagree with the ECG heart rate recordings (HR trace), then we have an instance of oximeter error. Similar methods to assess the quality of oximeter readings have been previously applied. For instance, Hay et al. [2002] test the ability of the PR trace to detect bradycardias in neonatal ICU patients by using the HR trace as the gold standard for annotating these events.

Oximeter errors and our approach for identifying them will be treated in detail in Section 5.4.2.

Patient handling

The incubator's doors are regularly open by the NICU staff, in order for certain clinical procedures to be performed (e.g. changing nappies). Amongst the monitoring channels available

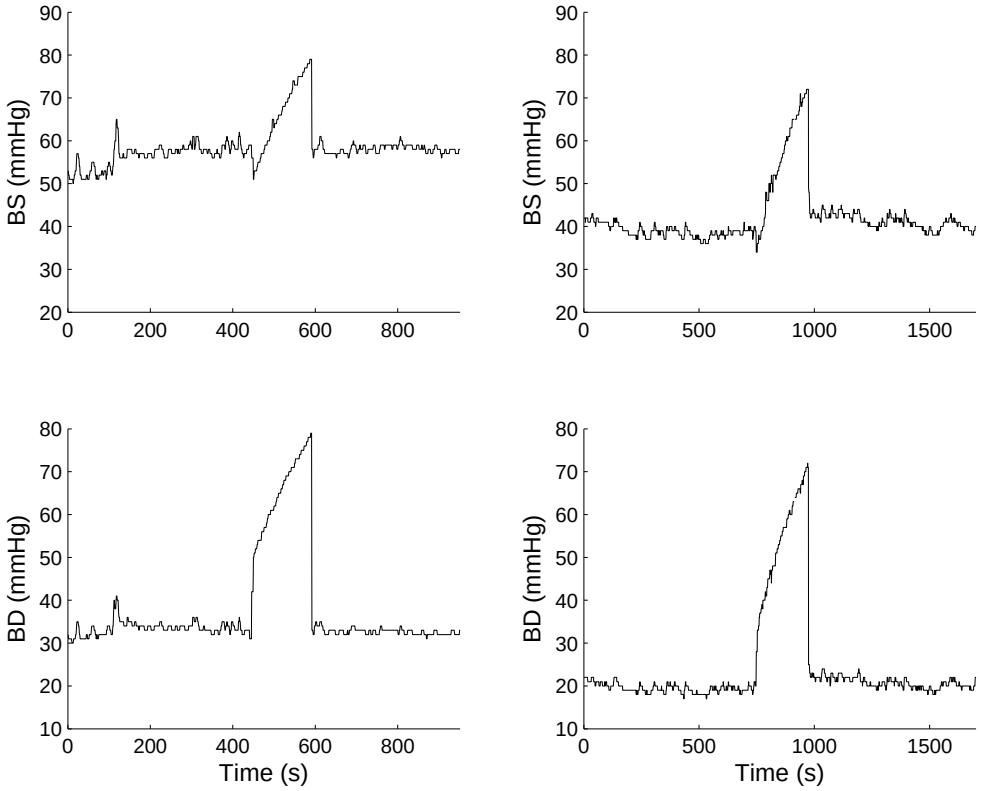


Figure 3.5: Two examples of blood sampling events (left and right columns). When a sample is being taken a saline pump acts against the pressure transducer causing the aritfactual ramps starting around $t = 450$ (left column) and $t = 750$ (right column).

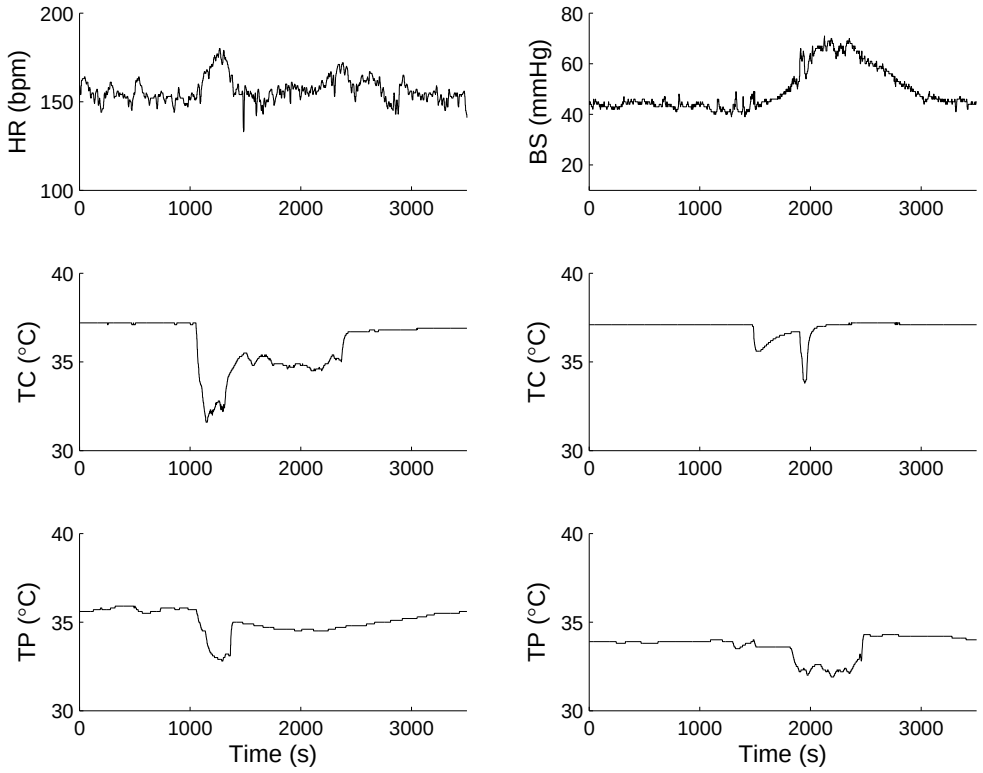


Figure 3.6: Two examples of handling episodes (left and right columns). In the first example, we can notice a sharp increase in heart rate immediately after the doors have been opened. In the second instance, the handling episode is associated with a period of increased blood pressure which lasts for approximately 20 minutes. Note that both heart rate and blood pressure increases are caused by an external agent.

for this research, temperature probes are most sensitive to this procedure[2]. During handling episodes, these generally become detached, and the readings decay to ambient level. In addition, there is increased variability on all the other physiological channels. Two illustrative examples are provided in Figure 3.6.

Note that handling differs from the other artifactual events discussed in this section in the sense that some of the monitoring channels do reflect the true value of the vital signs. This is the case of the heart rate and blood pressure examples in Figure 3.6. Genuine physiological data affected by clinical intervention are referred to as iatrogenic data, and identifying it is an important part of our work described in the following chapters.

[2]Environmental channels such as the incubator's humidity and temperature are arguably better indicators of handling episodes, but were not available here. Quinn [2007] has demonstrated how these signals can be successfully used to detect handling episodes in NICU patients. Nevertheless, more recent incubators are designed to be less sensitive to door opening (Prof. Neil McIntosh, personal communication), and thus alternative handing detection approaches are needed.

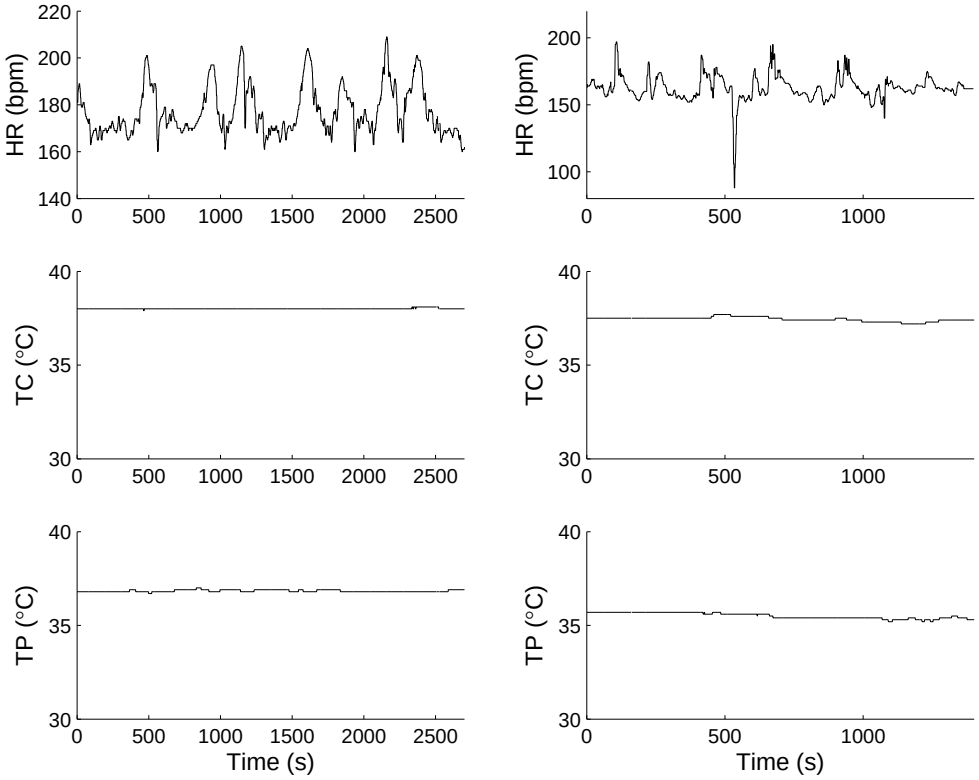


Figure 3.7: Two examples of abnormal dynamics (left and right columns). In both cases, there are several sudden increases in heart rate levels. These occur despite temperature traces being relatively flat, which suggests that the incubator's doors were closed. Note that a bradycardia event (Section 3.3) occurs around time $t = 550$ in the second example.

3.5 Abnormal events

In the previous sections we have discussed patterns of stable variation, and several physiological and artifactual events. In addition to these, there are various other patterns appearing one the monitoring traces. Many of these are rare patterns of true physiology, for which individual modelling would be impractical. Others are entirely novel patterns which may be particular to a certain patient; e.g. caused by the administration of drugs or due to various combinations of clinical conditions. All of these patterns are reunited in the class of abnormal events. We provide two examples in Figure 3.7.

As part of our work we obtained annotations for abnormal events, and subsequently sought to extract meaningful stereotypical patterns. This will be further discussed in Section 5.4.2.

3.6 Summary

This chapter has provided a brief introduction into how the physiology of an infant receiving intensive care is being monitored. We then described the patterns which most frequently appear

in the vital signs recordings. In our framework, these patterns are associated with either patient-specific stable physiological variation, stereotypical physiological events, artifactual events or abnormal events.

The work presented in the following chapters is largely concerned with how the knowledge presented in this chapter can be incorporated in a statistical model for neonatal condition monitoring. Table 5.2 contains the complete list of events used in this thesis, together with the definitions used for annotating them. Information on the prevalence of clinical events is presented Table 5.3.

Chapter 4

Previous work on NICU monitoring

In this chapter we review some of the previous work done on vital signs monitoring. The focus will be on neonatal intensive care unit data, but other relevant work on clinical data will also be presented. We divide the discussion between work on the diagnosis and the early detection of neonatal sepsis, and work on uncovering interesting physiological events and patterns of artifact underlying the vital signs traces.

In Section 4.1 we first discuss some of the challenges currently faced by clinicians when making a diagnosis of neonatal sepsis. We then introduce approaches towards the early detection of neonatal sepsis. Section 4.2 begins by placing this thesis in the wider context of automated vital sign monitoring. We then focus on previous work on representing the complex physiological signals as a concatenation of simpler dynamical regimes. The FSLDS model of Quinn et al. [2009] will be treated in most detail, as their ideas will be integrated into the condition monitoring framework we propose in Chapter 6.

4.1 Work on neonatal sepsis

We separately discuss two types of work on neonatal sepsis. A first body of work tackles several issues surrounding the current diagnosis standard and consists of research directed towards improving this standard. The second body of work is closer to our research goals and focuses on monitoring the infant's vital signs in order to make early predictions about the onset of neonatal sepsis. Most of this monitoring work has concentrated on building features informative for detecting the onset of the infection.

Diagnosis

As already mentioned in Section 1.1, the result of the blood culture is currently regarded as the "gold standard" for diagnosing neonatal sepsis. Nevertheless, it is acknowledged that the test can have poor accuracy [Modi et al., 2009, Griffin et al., 2003, Griffin and Moorman, 2001].

First, small sample volumes and antibiotic therapy can give false-negative results [Modi et al., 2009]. Estimates show that 30% to 40% of sepsis cases have negative blood tests [Griffin and Moorman, 2001]. Second, positive blood cultures do not always imply infection. The reason is that blood samples often contain contaminants [Modi et al., 2009].

To clarify matters, Modi et al. [2009] classify positive cultures into recognised pathogens in pure culture, mixed growth, and skin commensal categories [1]. For cultures in the first category, clinicians are certain that the patient is infected. The latter two categories cannot distinguish true infection from sample contamination. Importantly, statistics show that around two-thirds of the positive cultures are skin commensal or mixed growth. This blood culture classification is the approach followed in the Neonatal Intensive Care Unit (NICU) at the Royal Infirmary of Edinburgh, and has also been adopted for the study design in this thesis (Section 5.3).

Motivated by these problems and the resulting lack of an agreed case definition, Modi et al. [2009] propose a novel case definition for neonatal bloodstream infection. First, they identify 10 binary clinical signs predictive of a positive blood culture. These signs are computed daily and examples include the acute onset of hypotension, increase in bradycardias/apnea and glucose intolerance. The number of present clinical signs is then used to predict a positive blood culture. Based on the classification results, a new case definition is proposed: a baby is infected if either a recognised pathogen is found, or if the test yields mixed growth or skin commensal and ≥ 3 clinical signs are present. Note that this work still relies on laboratory results and thus is not being directed towards an earlier detection of the infection.

Early detection

The majority of the previous work on the early detection of neonatal sepsis has focused on using measures of heart rate variability as sepsis predictors. Unlike the work in this thesis, the methods presented in the following are purely discriminative.

Griffin and Moorman [2001] and Moorman et al. [2006] have previously proposed using heart rate data to discriminate between sepsis and sepsis-like babies pooled together, from a control group. Babies in the sepsis and sepsis-like group have had a blood sample taken for laboratory testing, and the test was positive for sepsis cases and negative for the sepsis-like cases. For the patients in the control group no sample was taken. The authors observed a positive skew in the RR interval (see Section 3.1) histograms in the hours before the clinical suspicion of sepsis, and an absence of skew during normal periods. This finding was quantified by a set of summary statistics referred to as the heart rate characteristics (HRC). This feature set ranges from simple statistics such as mean, quantiles or standard deviations to more

[1] A pure culture contains a single species of organism, whereas a mixed growth culture refers to the presence of several species. Skin commensals are bacteria living on the skin and could be related to either bloodstream infection or sample contamination

complex measures such as sample asymmetry [Moorman et al., 2006]. A notable feature exploiting the sequential nature of the data was the sample entropy [Lake, 2006]. The HRC are subsequently fed to a logistic regression classifier. The model was trained labelling the 24-hour period leading to the collection of a blood sample as positive, and the rest of the data as negative. Moreover, a single label is given to each 6 hour period. A larger dataset was employed for demonstrating that the HRC add predictive information to a classifier using only demographic features to discriminate sepsis and sepsis-like illness patients from controls [Griffin et al., 2003]. More precisely, they showed an increase in AUC (area under ROC curve) from 0.72 to 0.77 on a test set. In recent work, Moorman et al. [2011] they conducted a clinical trail which showed that HRC monitoring can decrease mortality. However, this does not explore the use of other physiological channels for sepsis detection and also assumes access to the high frequency RR data. In addition, unlike the model we develop in Chapter 6, the HRC framework is limited to the detection of sepsis.

The Artemis system [Blount et al., 2010], a stream computing project for neonatal intensive care, sets the detection of sepsis as one of its primary objectives. Their method introduces patient agents (PAs) able to perform multi-dimensional temporal abstraction on monitoring data [Stacey et al., 2007]. This type of work fits into a more general family of methods, in which clinicians apply domain knowledge to build abstract and/or qualitative descriptions of the patterns present in the data. These are subsequently structured in a rule-based model (see Quinn [2007, §3.2] for a brief review). In McGregor et al. [2012], they propose the use of both heart rate and respiratory rate variabilities for real-time sepsis detection. The latter is intended to help discriminate sepsis from confounding factors such as surgery or narcotics. A performance evaluation of this approach has yet to be published.

4.2 Work on modelling physiological monitoring data

A large part of the previous work on vital sign monitoring data (both neonatal and adult) has focused on methods to extract abstract representations predictive of clinical outcomes. Such representations are generally low dimensional descriptions of vital signs patterns or trends. A complementary body of work is aimed at discriminating patterns of true physiology from patterns of artifact.

Research modelling physiological monitoring data can be broadly classified into knowledge-based methods and methods based on statistical time series analysis. In the first category, abstract and qualitative descriptions are extracted from monitoring data by exploiting clinical expertise, and then fed to some rule-based decision system. For instance, Tsien [2000] detects artifact in NICU data by learning decision trees, where the features are knowledge-driven summary statistics of the monitoring traces. Miksch et al. [1996] build a rule matching system for

the ICU, and their model also suggests therapeutic actions based on clinical best practices.

In contrast, our work fits into the second category, where the approach is to model the processes underlying the observed physiology by using statistical methods. The following sections are concerned with methods representing the physiological data from intensive care patients as a concatenation of simpler dynamical regimes. The Factorial SLDS model for neonatal ICU monitoring of Quinn et al. [2009], to which our work is most related, will be discussed in most detail. In the reminder of this section we briefly introduce several examples of applying statistical tools on vital signs data.

As already mentioned, statistical models of vital signs data are often employed for predicting clinical outcomes. A recent example is the work of Wiens et al. [2012], where they measure the risk of an adult patient begin infected with Clostridium *difficile* based on data sources including vital sign measurements, lab results and medication. Each analysed day is labelled as negative or positive, depending on whether the patient eventually became infected during her hospital stay. First, a daily risk score is computed as the output of an SVM. Noting that the risk scores themselves form a time series, the authors report an increase in predictive performance by applying several sequential classification models, including the HMM, to these data.

The goal is sometimes defined as detecting deviations from normal physiology in ICU patients. Such problems are addressed in the novelty detection literature (see Pimentel et al. [2014] for a comprehensive review). An example is the work of Pimentel et al. [2013], where they are interested in monitoring patient recovery after upper-gastrointestinal surgery. A distribution of vital signs in "normal" patients is learnt via kernel density estimation, and is subsequently used to detect "abnormal" recovery data, which are associated with small likelihoods under the learnt distribution. Other work is directed to detecting artifact in the monitoring traces. For instance, Hoare and Beatty [2000] apply sequential models to detect artifact in heart rate data recorded while patients were under anaesthesia. They compute predictive distributions given by *ARIMA* models and LDS models, and classify data points with small likelihoods as artifact.

It is worth noting that our application cuts across several clinical goals as within the proposed model we detect the sepsis infection, infer novel dynamics, and also handle artifact (Chapters 5 and 6).

4.2.1 Switching models for intensive care data

There is a significant body of literature on the application of switching dynamical models to physiological monitoring data. The oldest reference we are aware of is the work of Smith and West [1983], where such models were applied to detect changes in the creatinine levels of patients shortly after they had kidney transplants.

More recently, Lehman et al. [2012] apply switching models to blood pressure (BP) mea-

surements recorded during first 24 hours of a patient's ICU stay. Their main goal is obtaining representations useful for predicting mortality. Similarly to us, they assume the observed dynamical BP patterns are driven by the regulatory systems responding to internal (e.g. disease) or external perturbations. The dynamical patterns are possibly recurrent within the same time series, and crucially shared across several patients. Their approach is to discover the hidden patterns using unsupervised learning. To this end, they employ the BP-AR-HMM model reviewed in Section 2.2.2. For each patient they build a ten-dimensional feature vector corresponding to the mode proportions of the top ten most frequent dynamical regimes inferred by the BP-AR-HMM. The usefulness of this representation is assessed by feeding it into a logistic regression classifier for predicting mortality. The individual predictive performance of mode proportions matches the commonly used acuity score SAPS. Importantly, when the acuity score is combined with the learnt representation performance increases from an AUC of 0.65 to 0.73.

In follow-up work sharing the same goal [Lehman et al., 2013, 2014], they employ the methodologically less intricate Switching VAR, which was briefly introduced in Section 2.3.1. As in the BP-AR-HMM work, the Switching VAR models is used to extract the most frequent dynamical models for each patient. Applying logistic regression on mode proportions, the authors found high risk modes as modes with odds ratios larger than unity, and low risk modes, modes with odds ratios smaller than unity. They found that high risk modes appear to be characterised by a smaller variability compared to the low risk ones. A logistic regressor where the Switching VAR-computed BP features are combined with the APACHE IV acuity score delivered a performance gain not statistically significantly better than APACHE IV on its own. Our application is quite different as we detect and not predict an outcome, and we exploit expert labels for the physiological regimes.

Saria et al. [2010a] also use switching autoregressive models for modelling heart rate data collected from NICU patients. Their model marries hierarchical Bayesian approaches widely adopted for document modelling (see e.g. Blei et al. [2003]) and the BP-AR-HMM. In order to build the analogy with document modelling, they refer to each dynamical regime as a word. Words are chosen to be AR(1) processes. In order to obtain a higher level of abstraction, the authors also borrow the concept of topic. Topics are probability distributions over the space of words, and are manually assigned clinically meaningful semantics such as *healthy* and *lung* (for lung complications). The central inference goal is obtaining posterior distributions over the space of topics. These posteriors are then employed as features in a supervised model for predicting disease grade. The authors report better performance than using either spectral features or AR-HMM features. In addition, they claim that words with higher variance occur more frequently in infants without complications. In contrast, the data patterns used in this thesis are associated with well-defined clinical events (see Chapter 3), and are thus amenable to (partly) supervised learning.

In related work, the same authors developed a NICU morbidity predictor called PhysiScore [Saria et al., 2010b]. They use physiological signs collected during the first 3 hours of life (heart rate, respiratory rate and oxygen saturation), gestational age and birth weight as input to a logistic classifier discriminating high-morbidity (HM) from low-morbidity (LM) neonates. The HM group was defined as patients with major complications (e.g. culture proven-sepsis, intraventricular hemorrage). The authors motivate the inclusion of physiological signal variance as a predictive feature based on their findings in Saria et al. [2010a], reviewed in the previous paragraph. The reported performance of PhysiScore is better than the extensively validated SNAPPE-II score. Interestingly, they also report an excellent leave-one-(patient)-out AUC of 0.97 in infection prediction, where neonatal sepsis was one of the considered infections. Note that this thesis is interested in detecting the onset of LONS, and not in assessing the risk of infection.

4.2.2 The FSLDS for condition monitoring

Some of the work in this thesis extending the FSLDS for neonatal condition monitoring discussed in Williams et al. [2006], Quinn [2007] and Quinn et al. [2009] (see Section 2.4.2 for a review of FSLDS models). Their work splits the physiological monitoring data into segments corresponding to different clinical conditions, such as those described in Chapter 3. More precisely, they identify periods of physiological stability (Section 3.2) and periods explained by certain types of clinical events. In general, clinical events can be of either physiological (Section 3.3) or artifactual nature (Section 3.4). In addition, they introduced an event dedicated to explaining abnormal dynamics (Section 3.5), and which will be further discussed below.

The factors in the physiological monitoring FSLDS correspond to clinical events. Thus, we can discuss physiological and artifactual factors. All factors can take on several discrete settings, one of which corresponds to the clinical event being inactive. When all factors are in the inactive setting we have a period of stable physiological variation.

The neonatal monitoring FSLDS departs from other applications of factorisation to time series modelling in the way factors interact. In previous work this interaction is generally assumed to be an addition of factor contributions, where the individual contributions are determined by factor settings (see Section 2.2.3 and Section 2.4.2). For the NICU monitoring task, domain knowledge is used to define factor interaction in terms of "overwriting" [Spengler, 2003, Quinn, 2007]. For instance, if a bradycardia occurs while the ECG heart rate probe is disconnected, then it cannot be observed. Thus, the disconnection of the heart rate probe has overwritten the bradycardia. Moreover, the activation of any factor overwrites the stability regime.

There can be several continuous state dimensions associated with each of the observed dimensions. This is achieved by modelling each monitoring channel as a univariate LDS (i.e.

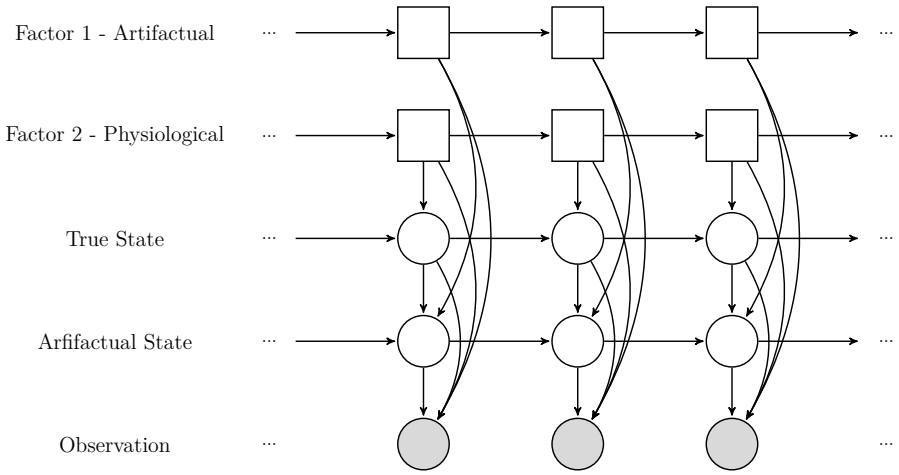


Figure 4.1: An FSLDS for neonatal condition monitoring. This example models two factors, one physiological and one artifactual. Note that an artifactual factor cannot affect true state dimensions.

LDS with 1-d observations). The hidden dynamics of the univariate LDS models are parameterised as low-order $AR(p)$ and $ARIMA(p,d,0)$ processes.

The FSLDS makes the simplifying assumption that the dynamics matrices $\mathbf{A}$ and dynamics noise covariance matrices $\mathbf{Q}$ have a block diagonal structure. A practical avantage is that it is easy to add and remove channels. This is useful as the set of observed channels differs across babies, depending on the clinicians' beliefs about the patient's condition. In addition, the block diagonal structure facilitates the incorporation of factor-channel dependencies.

The continuous states can have dimensions allocated to tracking the "true" dynamics of the vital signs and dimensions allocated to tracking artifactual processes. This is useful in situations such as taking a blood sample (Section 3.4), when the artifact dimensions model the recorded ramps, while the "true" state dimensions can estimate the real blood pressure values.

Certain monitoring patterns occur rarely, and modelling them explicitly would be impractical (see Section 3.5 for examples). At the same time, entirely novel patterns may be caused by certain combinations of clinical conditions. In Quinn et al. [2009], all these "known unknowns" are explained by a single dedicated factor, referred to as the X-factor. The X-factor is intended to flag monitoring data which are not stable and also cannot be explained by any of the known factors. It shares the same parameters as stability, but it has an inflated dynamics noise covariance matrix. The inflation coefficient is always greater than unity, and could be interpreted as how far outside the stable variation a recording should be before it is considered to be not normal.

For training the FSLDS it was possible to obtain expert annotation of the factor labels. Thus, learning largely became equivalent to learning a collection of LDS models, one for each

possible setting of the factors. Because the number of LDS models is exponential in the number of factors, it might seem that training requires large amounts of annotated data and would be computationally demanding. However, in the neonatal application factor models can be learnt independently, and then combined using the factor overwriting rules.

The first stage of FSLDS learning consists of learning the parameters for the stability regime. As a reminder, this corresponds to the setting when all factors are inactive. For learning the LDS corresponding to each observed dimension, Quinn [2007] note that the variance of the observed data is the sum of the variance of the hidden autoregressive process and the observation noise variance. They estimate the latter by inspecting the power spectrum of the observed data. Autoregressive coefficients are learnt via the Yule-Walker equations [Quinn, 2007, 4.5]. We will discuss an alternative estimation procedure in Section 6.2.1.2.

Because stable physiological dynamics are baby-specific, learning this regime is done separately for each monitored baby. This procedure is referred to as calibration, and requires a priori identifying a period of stability on the patient's traces. In the original work of Quinn [2007] this identification was performed *manually*. More recently, we have developed a framework to *automate* FSLDS calibration [Williams and Stanculescu, 2011]. Our approach was to build a classifier able to extract intervals of stability from the monitoring traces, and thus eliminated the FSLDS deployment bottleneck caused by manual calibration.

After learning stability, the other known regimes are separately fit. Then, overwriting rules are applied to learn the full factorial model.

Quinn et al. [2009] have explored running FSLDS filtering with both the GPB2 method described in Murphy [1998] and the RBPF proposed in Murphy and Russell [2001] (also see Section 2.4.1). Allowing both methods the same time budget, GPB2 consistently produced better filtering results, and thus we chose to adopt it for the work in Chapter 6.

4.3 Summary

The previous work on detecting neonatal sepsis has focused on the application of knowledge engineering for building features descriptive of the patterns found in the observed data. A criticism of this type of work is that does not address the problem of building a statistical model for these data. However, we have reviewed in this chapter several bodies of work showing that generative probabilistic models of the intensive care data can be successfully developed. Most relevant to this thesis is the FSLDS of Quinn [2007], which stands out for the extensive use of expert knowledge. The work we show in this thesis extends this modelling framework in a hierarchical fashion in order to incorporate knowledge about sepsis, and thus allow its early detection.

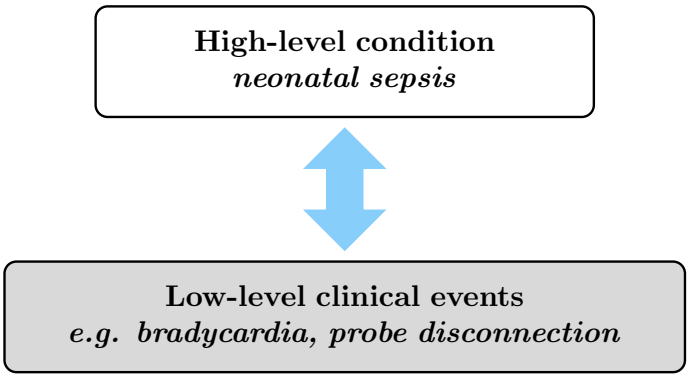


Figure 5.1: Inferring sepsis from low-level clinical events.

Chapter 5

A discrete state model for sepsis detection

In this chapter, we propose a neonatal sepsis detection framework which relies on probabilistic models for monitoring the evolution of baby-generated clinical events observed on the patient's vital signs traces (Figure 5.1). It has been previously asserted that an increased incidence of such events is a symptom of sepsis [Modi et al., 2009]. Starting from this hypothesis, we study the amount of predictive information about neonatal sepsis that can be extracted from the distribution of clinical events. First, using domain knowledge, we define and annotate a set of (low-level) clinical events. Our main contribution is the formulation of sepsis detection as inference and learning in an AR-HMM. In addition, we show how exact inference can be obtained in the presence of missing data. The effectiveness of the method is tested both on prediction of sepsis/normality on a second-by-second basis, and in terms of detected sepsis episodes. We also study the relevance of individual clinical event streams and compare our approach against several discriminative models. In the following chapter, we will extend the work presented here to automatically infer the clinical events from the monitoring data.

The times series model we use for the early detection of neonatal sepsis is described in Section 5.1. Then, inference in the presence of missing data is discussed in Section 5.2. Section 5.3 is concerned with the study design and introduces our neonatal sepsis dataset. We then define the set of clinical events that affect these data, explain how labels for these events have been obtained, and show the results of the annotation process (Section 5.4). Section 5.5 explains sepsis labelling and model fitting. Empirical results on applying the AR-HMM model for sepsis detection are presented and analysed in Section 5.6. A summary of the chapter is provided in Section 5.7.

Parts of this chapter have been adapted from Stanculescu, Williams, and Freer [2013].

5.1 AR-HMMs for sepsis detection

In the following, we explain how the distribution of clinical events can be modelled by an AR-HMM for the purpose of neonatal sepsis detection. A more general discussion about the AR-HMM has been provided in Section 2.2.2.

For the neonatal sepsis detection task, the hidden state variables of the AR-HMM are modelling the state of the infection. They can take one of J values and are organised as a first-order Markov chain with parameters $\pi_{j|i} = p(z_t = j|z_{t-1} = i)$ and $\pi^1_j = p(z_1 = j)$. Observations f_t in the general AR-HMM can be continuous, but for our purposes we restrict the discussion to the discrete case. Furthermore, we introduce direct dependencies only between consecutive observations. Conditioned on the state z_t, the emission process is again a first-order Markov chain parametrised by $\psi_{l|mj} = p(f_t = l|f_{t-1} = m, z_t = j)$ and $\psi^1_{l|j} = p(f_1 = l|z_1 = j)$[1,2]. The joint probability distribution for a sequence of length T is:

$$p(z_{1:T}, f_{1:T}) = \pi^1_{z_1} \psi^1_{f_1|z_1} \prod_{t=2}^{T} \pi_{z_t|z_{t-1}} \psi_{f_t|f_{t-1}z_t}. \tag{5.1}$$

At each time step t, we are observing a set of K clinical events $f^{(1)}_t, f^{(2)}_t, \ldots, f^{(K)}_t$. For each of them, $f^{(k)}_t$ denotes which of its possible $L^{(k)}$ settings clinical event k takes on at time t. Thus, AR-HMM observations are given by the cross product:

$$f_t = f^{(1)}_t \otimes f^{(2)}_t \otimes \ldots \otimes f^{(K)}_t$$

and can take one of $L = \prod_{k=1}^{K} L^{(k)}$ settings. The events are assumed to be conditionally independent:

$$p(f_t \mid f_{t-1}, z_t) = \prod_{k=1}^{K} p(f^{(k)}_t \mid f^{(k)}_{t-1}, z_t)$$

[1] The choice of a first-order Markov process does not imply a loss of generality, as higher-order dependencies have equivalent first-order representations albeit with exponentially higher number of settings.

[2] Note that for the neonatal monitoring application we denote AR-HMM observations by f_t, as opposed to the standard y_t notation used in Chapter 2.

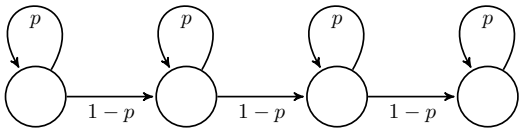


Figure 5.2: Topology giving the transition matrix for the new hidden state variables in a explicit duration model with $\tau_0 = 4$.

Each of them is modelled as a Markov chain with parameters $\{\psi^{(k)}_{l|mj}, \psi^{1^{(k)}}_{l|j}\}$. Notice that, in general, clinical events are not marginally independent. In addition, they tend to have long runs in the same discrete state. This also motivates our preference for an AR-HMM over a standard HMM, where observations are correlated only through the hidden variables.

5.1.1 Explicit duration modelling

HMM-like models make the implicit assumption that the time spent in each hidden state follows a geometric distribution. For sepsis monitoring, we expect episodes of infection to last for at least a few hours. Thus, assuming a geometric distribution for their duration is likely to be a performance limiting factor. Starting form the initial ideas of Ferguson [1980], a large body of work has been dedicated to methods that explicitly model the time spent in each regime (see e.g. Rabiner [1989], Murphy [2002] and Johnson [2005]).

The approach we take here has been discussed in Johnson [2005] and Murphy [2012, §17.6]. The main idea is to replace each state variable with τ_0 copies of itself. Each copy shares the same emission distribution as the original variable. Transitioning between the new states is given by the topology exemplified in Figure 5.2. The distribution of staying times becomes:

$$p(\tau|p,\tau_0) = \binom{\tau-1}{\tau_0-1} p^{\tau-\tau_0}(1-p)^{\tau_0}. \tag{5.2}$$

It is defined for $\tau \geq \tau_0$ and is equivalent to the negative binomial distribution [Murphy, 2012, §17.6]. Its mean and variance are $\mathbb{E}[\tau] = \tau_0/1-p$ and $\mathrm{Var}[\tau] = p\tau_0/(1-p)^2$ respectively. While some alternative solutions offer more flexibility in modelling the distribution of staying times [Murphy, 2012, §17.6], the method chosen in this thesis has the advantages of simplicity and faster inference.

5.2 Inference

Our main interest lies in real time sepsis detection, where we want to infer the onset of the infection from the patient's historical monitoring data up to a query time. Technically, this corresponds to computing the filtering distribution $p(z_t|f_{1:t})$. It is also useful to study if observing

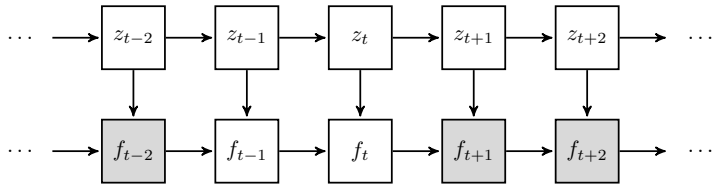


Figure 5.3: DAG of an AR-HMM with missing data.

future data improves the filtering prediction. This can be addressed by computing the smoothing distribution $p(z_t|f_{1:T})$. The latter is also useful for unsupervised parameter estimation. We first show how the forward-backward algorithm [Rabiner, 1989] is applied for AR-HMM inference. Then we explain how we extend it to address the problem of missing data.

The forward-backward algorithm is a message passing routine which exploits conditional independence relationships for doing exact inference in HMMs. In the AR-HMM, past observations are independent of future observations given both the current state and the current observation: $f_{t_0} \perp\!\!\!\perp f_{t_1}|z_t, f_t, \quad \forall t_0, t_1 \quad t_0 < t < t_1$. Using this we can write:

$$\begin{aligned} p(z_t, f_{1:T}) &= p(z_t, f_{1:t})p(f_{t+1:T}|z_t, f_{1:t}) \\ &= p(z_t, f_{1:t})p(f_{t+1:T}|z_t, f_t) \\ &\triangleq \alpha(z_t)\beta(z_t), \end{aligned} \tag{5.3}$$

where we have introduced the forward messages $\alpha(z_t) \triangleq p(z_t, f_{1:t})$ and the backward messages $\beta(z_t) \triangleq p(f_{t+1:T}|z_t, f_t)$. These messages can be computed recursively in a forward pass for α and in a backward pass for β [Ephraim et al., 1989, Woodland, 1992]. See Appendix B.1 for a derivation. When the likelihoods are precomputed, the total computational cost is $O(TJ^2)$.

If is often the case that we do not have access to observations at all time steps. The DAG in Figure 5.3 illustrates this situation. Here, we make a Missing at Random (MAR) assumption [Little and Rubin, 1987], which means there is no need to explicitly model the missing data mechanism. For sepsis modelling, we deal with missing data mainly when the patient is being handled by clinical staff. The sources of missing data will be detailed in Section 5.4.

One advantage of generative probabilistic models is that they can handle missing data in a principled way by marginalisation. For a sequence of length T, let $\mathcal{V}$ be the set of time steps for which we have observations. We define $f^v_{t_0:t_1} = \{f_t | t_0 \le t \le t_1, t \in \mathcal{V}\}$ as the set of observed variables between t_0 and t_1. Using this notation, $f^v_{1:T}$ is the set of observed variables for the given sequence. Similarly let $\mathcal{M} = \{1:T\} \setminus \mathcal{V}$ and $f^m_{1:T}$ be the set of missing observations. The goal of filtering becomes computing

$$p(z_t|f^v_{1:t}) = \sum_{f^m_{1:t}} p(z_t, f^m_{1:t}|f^v_{1:t}),$$

while for smoothing we want

$$p(z_t|f^v_{1:T}) = \sum_{f^m_{1:T}} p(z_t, f^m_{1:T}|f^v_{1:T}).$$

In the AR-HMM such marginalisations need to consider the direct dependencies between consecutive observations. For instance, if $t-1 \in \mathcal{M}$ then the forward message at time t must take into account the uncertainty about the unobserved quantity f^m_{t-1}. Our solution is a simple extension of AR-HMM inference. For $t \in \mathcal{M}$ only, we now compute $\alpha(z_t, f^m_t) \triangleq p(z_t, f^m_t, f^v_{1:t})$ and $\beta(z_t, f^m_t) = p(f^v_{t+1:T}|z_t, f^m_t)$. A full explanation is given in Appendix B.2. After recursively obtaining these messages the desired inference results for $t \in \mathcal{M}$ are obtained by marginalisation (e.g. $p(z_t, f^v_{1:t}) = \sum_{f^m_t} \alpha(z_t, f^m_t)$). If $|\mathcal{V}| = T_v$ and $|\mathcal{M}| = T_m$, then the computational expense increases to $O(T_v J^2 + T_m J^2 L^3)$. Since we expect the amount of missing data to be relatively small compared to the size of the dataset, the increase will be modest.

For neonatal condition monitoring the observations are a cross-product of discrete variables (Section 5.1). Missing data can independently occur for each of the monitored events. This means that at certain time steps only some of the composing factors of f_t are observed. We only need to marginalise over the remaining ones. Extending the missing data inference routine for this case was straightforward.

In practice, forward and backward messages defined as above exponentially decay to zero. In order to prevent such underflow issues, we have derived a scaled version of the recursions (see Appendix B.3). It follows the same reasoning as shown in [Bishop, 2007, §13.2.4] for the HMM.

Finally, note that inference in the explicit duration AR-HMM shares the same routines with a standard AR-HMM. Taking advantage of the state topology constraints explained above, the cost of the forward backward algorithm becomes $O(TJ(J+2(\tau_0-1)))$. The cost was derived by noting that the explicit duration AR-HMM models $J\tau_0$ states, and each of them has on average $(J+2(\tau_0-1))/\tau_0$ predecessor states.

5.3 The neonatal sepsis dataset

We have collected anonymised data from VLBW babies admitted at the NICU in the Royal Infirmary of Edinburgh between 2008 and 2011. All the analysed patients were intrinsically unstable, and thus nursed in incubators. The data consists exclusively of physiological monitoring channels sampled once per second. These are: heart rate, core and peripheral temperatures (TC and TP) and oxygen saturation (SO). Heart rate measurements are available from two sources: ECG leads (HR) and pulse oximeter (PR). Our samples are monitoring windows with a duration of 30 hours and fall into one of the following two categories: the sepsis group or the control group. Sepsis samples have been selected such that the time the positive blood sam-

Table 5.1: Population Demographics: Gestation, Birth Weight (BW) and Post Partum Age

Group	Statistic	Gestation	BW	Age
Sepsis	mean	27.2 weeks	873 gr	14.5 days
	std.dev.	1.5 weeks	256 gr	8.5 days
Control	mean	26.7 weeks	837 gr	15.2 days
	std.dev.	1.7 weeks	139 gr	14 days

ple was taken occurs precisely 24 hours after the start of the window, and control samples are extracted to align by the time of day.

For the sepsis group, we firstly considered monitoring all babies who had at least one blood sample taken for culture analysis. The group was refined to include only samples where the culture grew organisms ordinarily considered as pathogenic, leading to a diagnosis of "proven sepsis". This was 10% of the original group, as 65% of the samples were negative, and the remaining 25% were allocated to either the mixed growth or skin commensal categories. For the control group, there was no suspicion of sepsis in a consecutive three day period around the selected intervals and no blood sample had been analysed. In addition, there was no recorded evidence of any clinical condition other than severe prematurity.

In order to investigate the utility of multi-channel data for sepsis detection, we selected babies for which all the channels above were present. These are needed for defining the events given in Table I. Since there was no systematic reason for the absence of any of these five channels, this is an unbiased selection criterion. During this step, 20% of the sepsis samples were removed. Finally, in some cases, the bedside devices consistently failed to record measurements (or probes were displaced) for extended periods of time. We placed a data availability threshold of 50% for all channels. This resulted in a reduction from 26 to 18 sepsis samples.

Under the same data availability criteria, we have selected sufficient control samples to provide an equal amount of data to the sepsis group. The main reason for this choice was the time-expensive data annotation process required for model fitting (see Section 5.4.2). However, we chose control neonates such that the demographics of the two patient group are matched (Table 5.1).

In summary, we are studying 36 samples divided as follows:

- the sepsis group: 18 samples obtained from 18 different patients,
- the control group: 18 samples obtained by taking 2 samples from each of 9 different patients[3].

[3]One control sample initially selected has been discarded due to an atypical oxygen saturation trace. This was most likely caused by a fault with the monitoring equipment. The sample was readily classified as an outlier.

As it is arguably important to discriminate sepsis/non-sepsis periods for the same neonate, three patients have samples in both sepsis and control groups. Summing up, we are analysing a total of 24 different neonates.

5.4 Sepsis detection by monitoring clinical events

Our approach for neonatal sepsis detection is centred around the idea that the onset of the infection is associated with an increase in low-level clinically significant events. In the following, we first summarise the knowledge about NICU monitoring data used for defining and annotating these events (Section 5.4.1). We then explain how we efficiently combined expert and automated event annotation together (Section 5.4.2). We conclude the section by providing further interpretation on clinical event occurrences, some useful visualisations and connect our findings to previous work (Section 5.4.3).

5.4.1 Clinical event definitions for sepsis detection

We found it useful to follow the ideas in Quinn et al. [2009] and classify clinical events into physiological events and artifactual events. During physiological events the monitoring traces reflect the true values of the baby's vital signs. Artifactual events occur when the traces are corrupted by faults with the monitoring equipment and do not reflect the true state of the patient. We have provided further discussion on this clinical event classification in Chapters 3 and 4.

Table 5.2 gives the list of clinical events we use, together with their brief descriptions. Note that the present list is adapted from the one proposed in [Quinn et al., 2009] for the purposes of this work. Here, we chose not to monitor blood sampling episodes, because of the small amount of blood pressure data available. While several definitions of bradycardia are possible, this thesis employs the one used in the NICU at the Royal Infirmary of Edinburgh, and also in Hazinski et al. [2010]. Note that the inclusion of the X-factor [Quinn et al., 2009] makes the list exhaustively cover all the patterns appearing in the monitoring data. Since the X-factor can be either physiological or artifactual, it cannot be directly used for inferring the patient's state of health.

5.4.2 Clinical event annotations

In order to efficiently obtain labels for the events defined above, we chose to combine expert and automated annotation.

In collaboration with clinicians from the Royal Infirmary of Edinburgh, annotations were initially obtained for bradycardia, desaturation, handling and for the X-factor. In subsequent data exploration, the X-factor annotations were inspected for any recurring patterns potentially predictive of sepsis. We found low amplitude bradycardia-like patterns to display a higher

Table 5.2: Exhaustive list of clinical events monitored for detecting neonatal sepsis.

Event	Type	Brief Description
Probe dropout	artifactual	lack of monitoring data due to temporary removal or malfunctioning of the monitoring devices
Handling	artifactual	some clinical procedure is performed (e.g changing nappies); the incubator's doors are thus open; the indication is a decay in TC and/or TP together with increased variability or dropouts on the other physiological channels
Bradycardia	physiological	sharp fall in HR (PR) of at least 30 beats per minute (bpm) from a reference level followed by a sharp recovery
Oximeter error	artifactual	disagreement between the oximeter (PR) and EEG (HR) heart rates; the disagreement is associated with temporary malfunctioning of the oximeter; this translates into the unreliability of the SO trace
Desaturation	physiological	sudden fall in SO followed by recovery; desaturations are commonly associated with SO falling below 85%
X-factor	any	non-normal pattern occurring on at least one physiological channel that cannot be explained by ANY of the events above

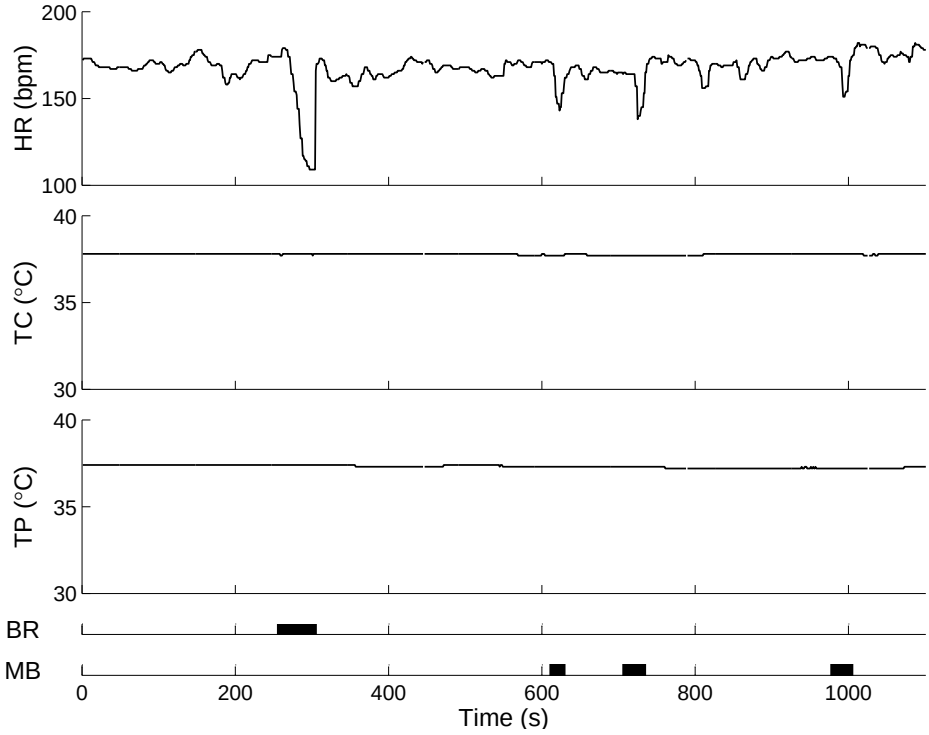


Figure 5.4: Examples of heart rate channel events. An instance of bradycardia ('BR') occurs around time $t = 275$. This is followed by three low amplitude bradycardia-like patterns occurring around times $t = 625$, $t = 700$ and $t = 975$. We defined such events as mini-bradycardia ('MB') instances. Note that the temperature traces are relatively flat, suggesting that no clinical procedure was performed.

incidence in the hours before the positive test (see Figure 5.7c). These patterns would often appear in clusters and close to drops in heart rate significant enough to be classified as bradycardias. As these events they did not fall into our standard working definition (Table 5.2), they were not initially annotated. We chose to separately introduce them as mini-bradycardias: "bradycardias" with a heart rate drop of 15 to 30 bpm (see Figure 5.4). Thus, annotations for mini-bradycardias were a later addition.

Less clinical expertise is required for annotating the remaining two events, probe dropouts and oximeter errors. Thus, both of these artifactual events were handled automatically. The monitoring equipment already marks probe dropouts by recording the value 0. Dropout statistics depend on the channels affected by each clinical event, but on average we lack monitoring data for 2% of the time.

Oximeter errors are characterised by a disagreement between the two heart rate channels (Table 5.2). However, apart from the oximeter error, we found that another source for this disagreement is that the channels are not temporally aligned. Thus, we first aligned HR traces with respect to the PR ones by maximizing their cross-correlation, and refer to the aligned HR

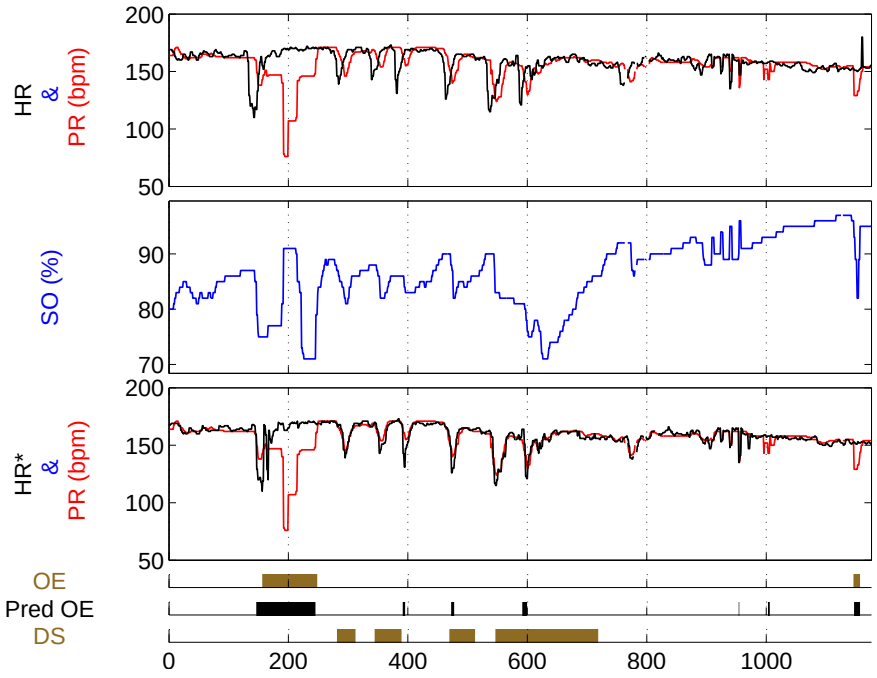


Figure 5.5: True and inferred oximeter errors ('OE's). In the top panel, we see examples of the two sources of disagreement between the HR and PR traces. Around time $t = 200$ there is a clear instance of oximeter error. The problem of HR and PR alignment becomes obvious during the 4 desaturation ('DS') instances, but it is noticeably alleviated in the third panel where we plot HR*, the temporally aligned ECG heart rate. At the bottom of the plot, we show that the two oximeter error instances have been correctly identified ('Pred OE').

trace as HR*. An illustrative example is provided in Figure 5.5.

Our proposed oximeter error detector is a standard HMM with Gaussian emissions applied to the difference between the aligned heart rate trace, HR*, and the PR trace. We tested this procedure by comparing it against expert oximeter error annotations obtained for three 24-hour monitoring windows corresponding to three randomly selected neonates from our dataset. The proposed HMM has two regimes with meaning oximeter error and normal (i.e reliable SO readings), and the labels were used for supervised learning of maximum likelihood parameters. An Area under the ROC curve (AUC) of 0.96 obtained by a three-fold leave-one-patient-out cross-validation encouraged us to apply this simple method over the whole dataset. Inferred oximeter error events such as shown in Figure 5.5 have been produced by binarising the filtering distribution at the threshold corresponding to the equal error rate (EER). For further detail on the ROC curve and its summary statistics see Section 5.6.

5.4.3 Interpretation and visualisation

We now turn our attention to analysing and interpreting the outcome of the annotation process. Key to our approach is the fact that not all physiological event instances should be used for sepsis detection, but only the *baby-generated* ones. Also, labels for physiological events cannot be provided for all the data. The following paragraphs will explain these statements in detail. After that we provide a visualisation of the time evolution of the number of baby generated physiological events, and conclude the section by relating our findings to previous work on sepsis detection.

We begin by defining *baby-generated* physiological events. Our main observation is that instances of the other physiological events can be frequently seen during handling episodes (see Section 3.4 and Table 5.2). A typical example is shown in Figure 5.6. In such cases we cannot distinguish whether the events are caused by the baby's true state of health or merely because an extremely fragile patient is being handled by the clinical staff. Our solution was to not use these instances for sepsis detection. Consequently, we rely exclusively on physiological events happening outside handling episodes. Only such instances can be confidently classified as being *baby-generated*.

Another difficulty was that we cannot label clinical events at all time steps. First, during probe dropouts there is no access to the true values of the baby's vital signs and consequently it is impossible to provide annotations. Second, during oximeter error events one cannot annotate desaturations.

Table 5.3 summarizes the output of the data annotation process. Baby-generated physiological events display a higher incidence in the sepsis group. Also, the amount of patient handling does not differ much between the two groups. The same conclusion can be drawn about the numbers of both X episodes and oximeter errors.

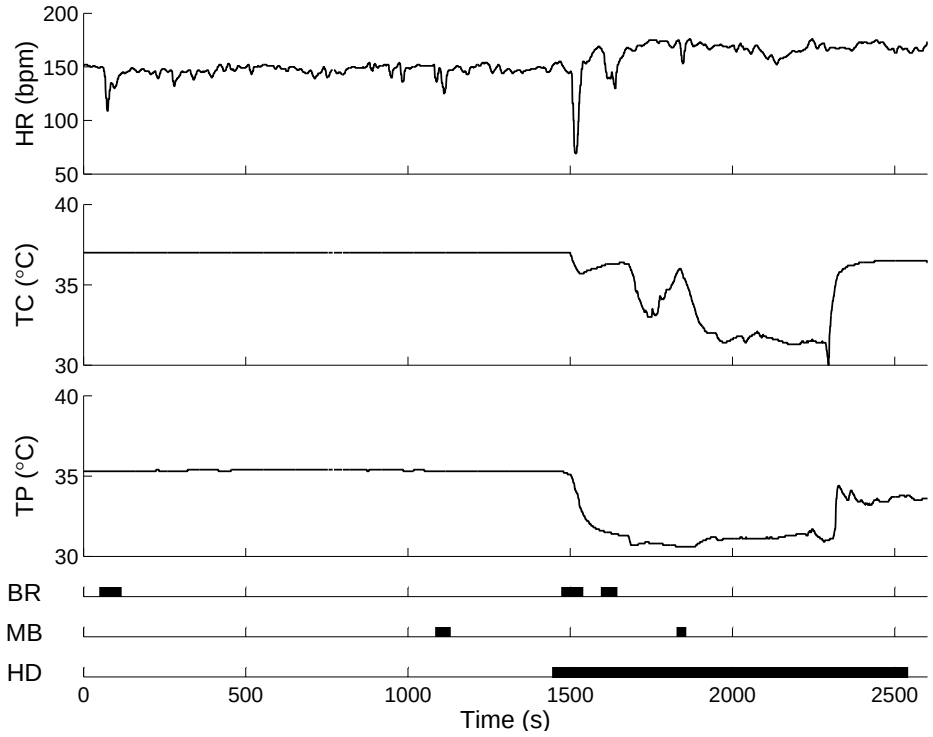
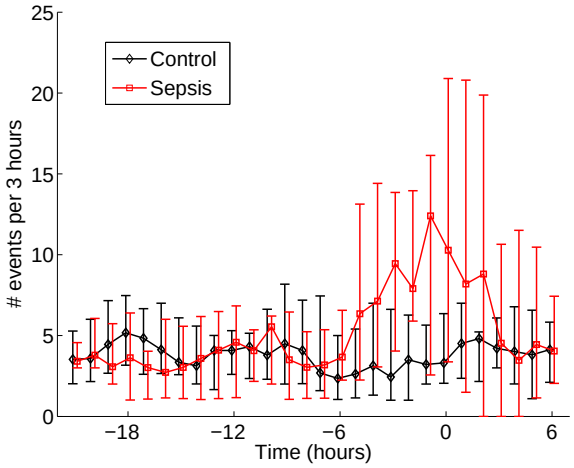
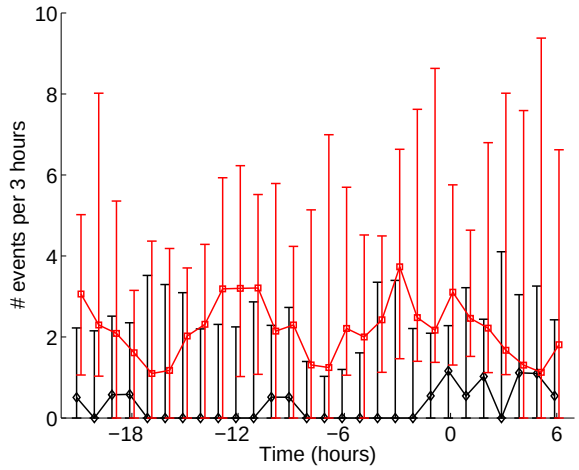


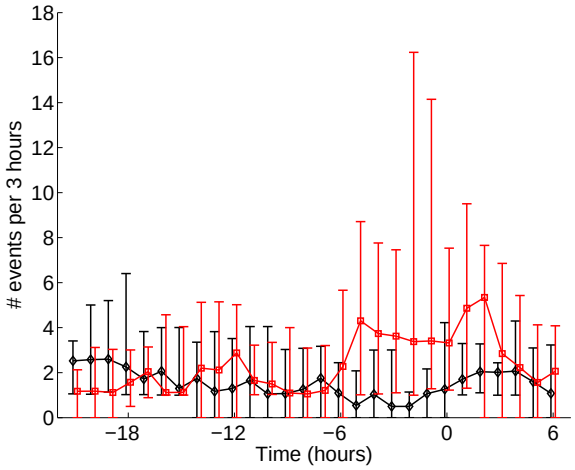
Figure 5.6: Examples of clinical events affecting neonatal monitoring data. Instances of bradycardia ('BR') and mini-bradycardia ('MB') can be observed on the ECG heart rate trace (HR). Around time $t = 1450$, a sudden fall in both the core and peripheral temperatures (TC and TP) signals the start of a handling ('HD') event. Note that physiological events occurring during handling episodes are not used for sepsis detection.



(a) Bradycardia



(b) Desaturation



(c) Mini-Bradycardia

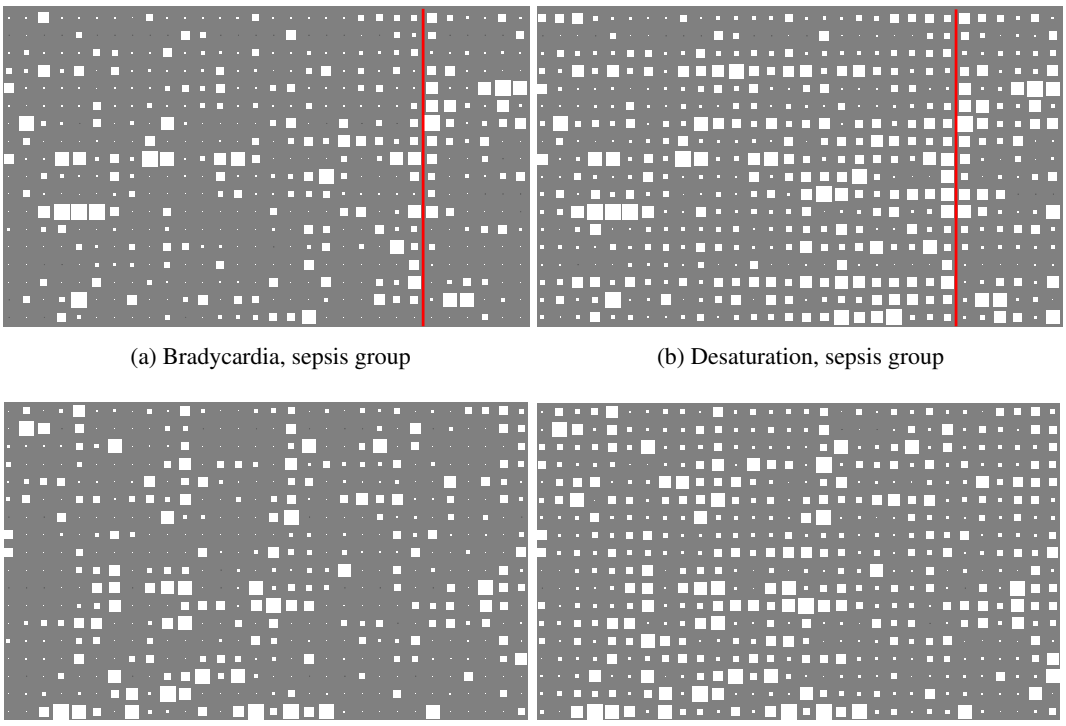
Figure 5.7: Time evolution of the median weighted number of baby generated physiological events for both sepsis and control groups. The data has been aligned such that for babies in the sepsis group 0 denotes the time the positive blood sample was taken. The counts are computed hourly and summarize the preceding 3 hour period. The error bars mark the first and third quartiles. We have used a small offset between the two patient groups to improve readability.

Table 5.3: Clinical event incidence (number of events), total and median durations for the sepsis/control groups. Only *baby-generated* physiological events have been considered. The total amount of data for each group is $18 \times 30 = 540$ hours.

Event	Group	Incidence	Total (hrs)	Median (sec)
Bradycardia	Sepsis	1128	13.9	38.5
	Control	773	8.4	37
Desaturation	Sepsis	742	32.3	101
	Control	231	10.5	124
Mini-Bradycardia	Sepsis	598	10.7	42
	Control	374	4.1	34
Handling	Sepsis	201	41.7	510
	Control	205	53.7	592
X	Sepsis	227	10.3	94
	Control	175	7.0	114
Oximeter error	Sepsis	4051	44.6	16
	Control	3395	36.4	18

Figure 5.7 shows *baby-generated* physiological event counts evolving through time for both patient groups. The samples in the sepsis group are naturally aligned using the time of the positive blood test. Event counts have been weighted according to the label availability constraints discussed above (i.e. patient handling, probe dropout and oximeter error). More precisely, if $p\%$ of a monitoring interval could be annotated for *baby-generated* physiological events, the count for that interval was multiplied by $100/p$. For the sepsis group, there is an increase in baby generated bradycardias and mini-bradycardias in the 9 hours before the positive test. Also, desaturations are generally more present in the sepsis group, but seem to be less informative about the onset of the infection.

The periods of time for which annotations of *baby-generated* events could not be provided will be treated as missing data. The sources of missing data identified in this section are handling episodes, probe dropouts and oximeter errors. In Section 5.2 we mentioned that missing data will be treated under the MAR assumption, and now we justify this choice. Firstly, summary statistics for the missing data sources do not differ much between the sepsis and control group (see Table 5.3), which suggests missing data is independent of the presence of sepsis. In addition, we also investigated whether the amount missing data increases near the time of physiological deterioration for patients in the sepsis group. Figure 5.8 shows the amount of time annotations for bradycardias and desaturations could not be provided. Note that we do not present missing data results for mini-bradycardias, as these are identical to the results for



(a) Bradycardia, sepsis group

(b) Desaturation, sepsis group

(c) Bradycardia, control group

(d) Desaturation, control group

Figure 5.8: Hinton diagrams showing the amount of missing data for baby-generated bradycardias and desaturations. Each row is a sample, and each column corresponds to one hour of monitoring data. The size of the white squares is proportional to the fraction of the hour marked as missing data, and the red vertical line in the sepsis group marks the time the positive blood culture has been collected. Despite certain sepsis samples display an increase in the amount of missing data in the hours before the clinical suspicion of sepsis, there appears to be little difference between the two patient groups.

bradycardias. Largely due to oximeter errors, missing data affects the desaturation annotations more than the bradycardia ones. Importantly, missing data constantly appears in our dataset, but a higher incidence near the clinical suspicion of sepsis is not apparent. The latter point further supports our MAR assumption.

Our findings about the incidence of heart rate events prior to clinical suspicion of sepsis can be associated with the work in Griffin and Moorman [2001], Moorman et al. [2006], Flower et al. [2010]. In Flower et al. [2010], the authors analyse inter-beat (RR) data and report an increase in the frequency of heart rate decelerations near the time of the clinical diagnosis of sepsis. We did not have access to the RR data, but we found it worthwhile to check whether a positive skew in the RR histograms translates into a negative skew of HR data, due to the inverse relationship between intervals and frequencies. Clearly, the HR trace is derived from the RR data, and thus part of the RR frequency spectrum is lost during processing and cannot be

observed in the HR frequency spectrum. However, by computing the sample skewness of the HR channel, we found that indeed lower values of skewness often characterise the hours before the positive blood test. Furthermore, by removing the bradycardias and mini-bradycardias from the analysis most of the skewness is eliminated. Thus, we report that the distribution of heart rate events can be used to at least partly explain the positive skew displayed by the RR histograms.

5.5 Parameter estimation

We fit an AR-HMM model to observations of $K = 3$ baby generated physiological event channels: bradycardias, desaturations and mini-bradycardias. The hidden state is chosen to be a binary variable which can take on values $z_t = normal$ or $z_t = sepsis$. In the following, we explain how we label the presence of sepsis in the training data. These labels are then used for supervised learning of the AR-HMM parameters.

Labelling the sepsis variable is different for the two patient groups. For the sepsis group, we know the exact time of the positive blood test. Following consultation with clinicians, it was agreed that labelling the period of 6 hours before this moment as sepsis would be reasonable. The onset of the infection cannot be assumed to be an instantaneous event. Thus, we define a transition period in which the patient progresses from being in the *normal* state to being in the *sepsis* state. We take this to be the 12 hours between between 18 and 6 hours before the positive test. This period is left unlabelled and will not be used for either training or testing. All monitoring data before the transition period (i.e. the first 6 hours of a sample in the sepsis group) is labelled as *normal*. The later choice is based on the assumption that it is unlikely it would take more than 18 hours between the onset of the infection and the time clinicians become suspicious of sepsis. More precisely, a physiological deterioration would have become apparent, and thus a blood sample would have been collected for laboratory testing. Finally, we do not assign a label to the data after the positive test, as this is likely to be affected by the patient's response to treatment and has less relevance for the task of real-time sepsis detection. A visualisation of these sepsis labelling definitions is provided in Figure 5.9. All the data in the control group is labelled as *normal*.

Using annotations simplifies parameter estimation. Our optimization goal is maximizing the joint probability of the labelled hidden states and the corresponding observations. Maximum a posteriori (MAP) estimates of the state transition probabilities are given by:

$$\hat{\pi}_{j|i} = \frac{n_{j|i} + n_0}{\sum_{j'}(n_{j'|i} + n_0)}, \tag{5.4}$$

where $n_{j|i}$ is the number of times we transition from hidden state i to hidden state j. Here we use a symmetric Dirichlet prior with parameter $n_0 = 1$, in order to prevent estimates from being

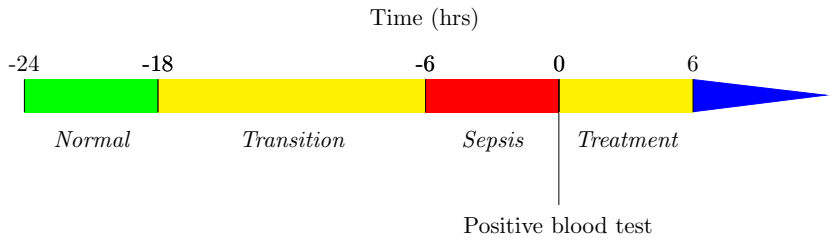


Figure 5.9: Sepsis variable labels corresponding to a patient for whom the outcome of the blood test was positive.

too small when data counts are low. Similarly, we learn the emission probability parameters using:

$$\hat{\psi}^{(k)}_{l|mj} = \frac{n^{(k)}_{l|mj} + n_0}{\sum_{l'}(n^{(k)}_{l'|mj} + n_0)}, \tag{5.5}$$

where $n^{(k)}_{l|mj}$ is the number of times event k transitions from setting m to setting l when the hidden state takes on value j.

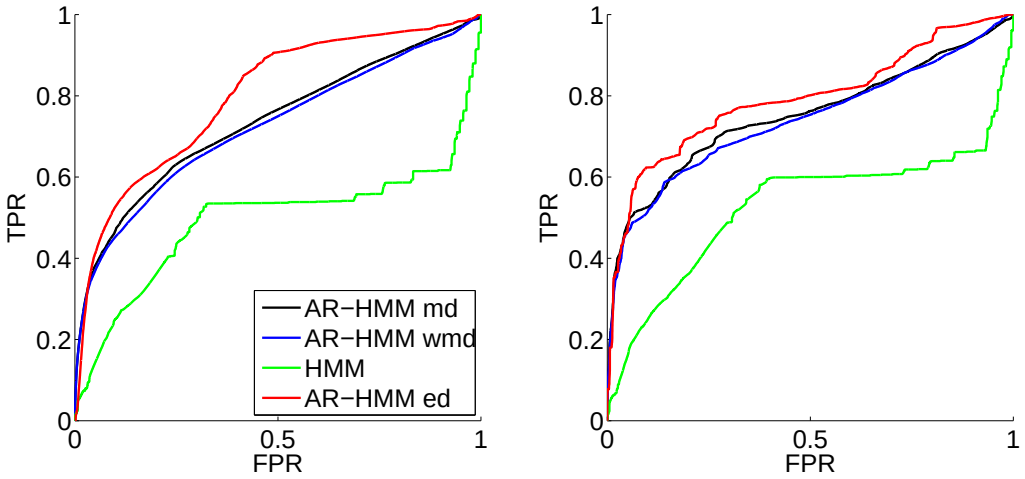
In Section 5.4 we explained why it was not always possible to annotate baby-generated events. In theory, we could use an expectation-maximization (EM) procedure to account for missing data when estimating parameters. However, the total amount of annotated data is much larger than the amount of missing data. Thus, we would expect the benefits to be minimal.

Note that in the absence of any sepsis labels the AR-HMM can be trained in an unsupervised manner. Maximum likelihood (ML) estimates of the values of the parameters are usually determined by optimizing the probability of the observations with EM. The inference procedure described in Section 5.2 can be used in the expectation step.

5.6 Experiments

In this section we describe experimental results for detecting neonatal sepsis on the data introduced in Section 5.3. We show the models' performance and discuss learnt parameters in Section 5.6.1. The relevance of individual physiological event streams is examined in Section 5.6.2. An alternative *episode-based* analysis is presented in Section 5.6.3. We finish in Section 5.6.4, where we present a comparison between our proposed AR-HMM and some discriminative models for sepsis detection.

The quality of the *second-by-second* inferences is measured against the sepsis labelling defined in section 5.5. This labelling translates into 6 hours of *normal* data and 6 hours of *sepsis* data for each sepsis patient, and 30 hours of *normal* data for each control patient. Consequently, we have six times more *normal* data than *sepsis* data in our dataset. In order to account for this class imbalance when reporting the second-by-second results, we chose to draw ROC curves (e.g. Fawcett [2004]). ROC curves show the dependence between the false positive rate (FPR)



(a) Filtering (b) Smoothing

Figure 5.10: ROC curves corresponding to different models for sepsis detection.

and the true positive rate (TPR), and are thus insensitive to class imbalance. We report the area under the ROC curve (AUC) and the equal error rate (EER)[4]. If misclassification costs had been available, we could have visualised the expected cost in ROC space as explained in Provost and Fawcett [2001]. Note that the actual choice of an operating point on the ROC curve will likely take into account the relationship between the absolute numbers of false positives and true positives, which directly depends on the class imbalance.

For the *episode-based* analysis, we use precision-recall (PR) curves[5] [Raghavan et al., 1989]. We report the average precision (AP), which can be understood as the area under the PR curve [Everingham and Winn, 2012, §3.4], and the maximum F-score[6] value over the PR curve. Note that both the second-by-second and the episode-based analyses are projections of the inferences onto different metrics, and thus can reveal different performance aspects.

All the results in this section have been obtained using cross-validation. For each of our 36 30 -hour monitoring samples, we separately test models trained on the remaining $35 = 36 - 1$ monitoring windows. The performance curves have been drawn by merging the predictions obtained for each sample [Fawcett, 2004].

In terms of visualising the results, posterior distributions are given as gray-scale horizontal bars, with white meaning 0 sepsis probability and black corresponding to probability 1.

[4] EER is the error rate computed at the threshold for which the FPR equals the false negative rate (FNR). Note that FNR = 1 - TPR.

[5] Precision is defined as TP/(TP+FP) and recall equals the TPR.

[6] The F-score is the harmonic mean of precision and recall.

Table 5.4: Summary statistics obtained by cross-validation in a second-by-second analysis.

	Filtering		Smoothing	
	AUC	EER	AUC	EER
AR-HMM md	0.74	0.33	0.75	0.29
AR-HMM wmd	0.72	0.34	0.73	0.32
HMM	0.50	0.46	0.53	0.40
AR-HMM ed	0.80	0.30	0.79	0.27

Table 5.5: Monte Carlo estimates of the expected number of baby generated events over a $T = 3$ hour period.

State	Bradycardia	Desaturation	Mini-Bradycardia
Normal	4.29	1.63	2.10
Sepsis	11.01	5.95	6.65

5.6.1 Model evaluation with a second-by-second analysis

In the following we first compare the performance of several sequential models for sepsis detection. We then analyse the fitted emission distributions and provide visualisations of the posteriors distributions produced by the model delivering the best results.

ROC curves for several sepsis models are given in Figure 5.10, and the corresponding summary statistics are presented in Table 5.4. For our sepsis detection task, we are mainly interested in real time prediction. Thus, smoothing results can be interpreted as an upper bound for the predictive power of the selected physiological events.

"AR-HMM md" is the standard AR-HMM model which handles missing data. If we do label the missing data assuming no baby-generated physiological event was happening, we obtain a model without any missing data, "AR-HMM wmd". This approach performs worse, mostly due to long handling events happening during sepsis episodes wrongly classified as normal. The marginalisation performed in the missing data approach helps to correctly classify these periods as sepsis. The benefits of explicitly modelling events as Markov chains with AR-HMMs are clear when compared to an "HMM", whose performance is close to that of a random classifier.

"AR-HMM ed" combines explicit duration modelling with exact marginalisation over missing data at inference time. For each hidden state, the parameters of the corresponding event duration distribution (eq. 5.2) can be learnt from the sepsis labelling. However, due to the lack of diversity in the length of labelled sepsis episodes (Section 5.5), we treat the number of copies of original state variables, τ_0, as a hyper-parameter, and consider values in the set $\{5, 10, 15, 25, 50, 100\}$. To avoid bias in model comparison due to the hyper-parameter τ_0, we

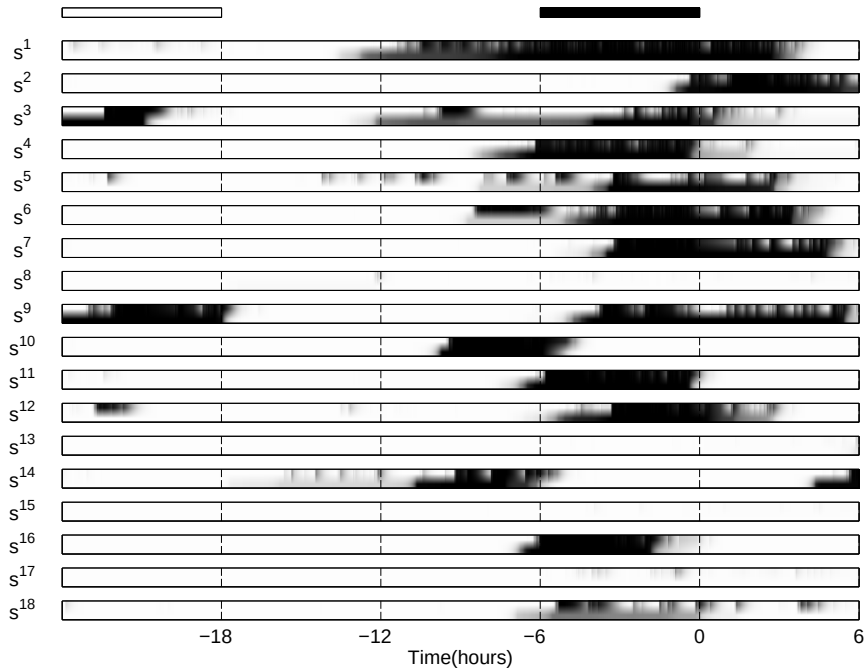
determine the performance of the explicit duration model by applying nested cross-validation (see e.g. Varma and Simon [2006]). Here, nested cross-validation differs from the standard cross-validation in the sense that an additional (inner) cross-validation step is run on the training data of each (outer) cross-validation fold, in order to select the best τ_0. Note that an inner cross-validation is separately run for each outer cross-validation step. Thus, nested cross-validation estimates the generalisation performance, but does not provide a single optimal τ_0. Table 5.4 shows that the explicit duration model delivers the best performance for both filtering and smoothing.

The fitted emission distributions can be used to characterise the *sepsis* and *normal* regimes. Since the learnt ψ parameters are hard to interpret directly, we show an alternative representation which can be easily associated with the information in Figure 5.7. More precisely, we used a Monte Carlo approach to estimate the expected number of physiological events over a $T = 3$ hour period. For each event-regime pair, we separately sampled the corresponding Markov chain. In all cases, $N = 5000$ samples of length $T = 3$ hours have been empirically found to suffice for convergence of the estimated number of physiological events. Table 5.5 shows these estimates. In the *sepsis* state we see an approximately 3-fold increase in the expected incidence of the monitored events compared to periods when the patients are not infected. This is in-line with the event count evolution presented in Figure 5.7.

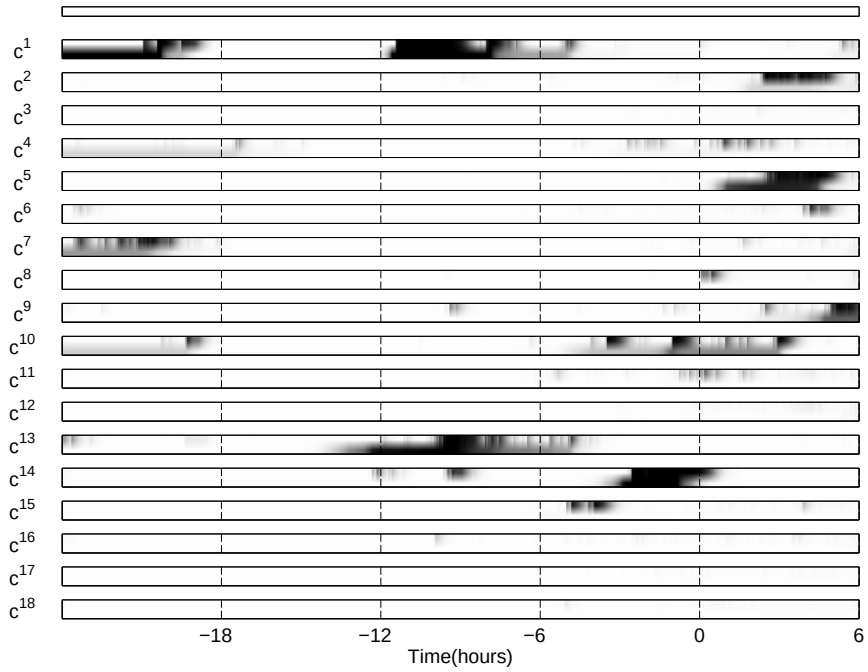
Figure 5.11 shows inference results for the model delivering the best performance, the explicit duration AR-HMM. For 12 samples in the sepsis group ($s^{1-7}, s^9, s^{11}, s^{12}, s^{16}$ and s^{18}) a relatively long sepsis episode is identified during the 6 hours before the positive blood test. In all but 2 samples (s^2 and s^3), the sepsis episode is detected at least 3 hours before the positive test. For 2 cases (s^{10} and s^{14}) sepsis episodes are flagged mostly during the transition period rather than the sepsis one. In the remaining 4 samples (s^8, s^{13}, s^{15} and s^{17}) no clear sepsis episode has been identified. The sepsis periods flagged in the control group are usually short. We believe that many of them can be explained by handling events which do not display corresponding falls in either TC or TP channels.

5.6.2 Physiological event evaluation

It is useful to understand which types of physiological events contribute most for detecting sepsis. Since bradycardias and mini-bradycardias are closely related, we phrase this question as asking whether the monitoring of desaturations brings additional information about sepsis compared to monitoring only the heart rate. In Figure 5.12 and Table 5.6 we compare an explicit duration AR-HMM monitoring all events ("ALL") with one monitoring only heart rate channel events ("BR+MB") and one looking only at desaturations ("DS"). For all event types, ROC curves have been obtained using nested cross-validation. This analysis shows that monitoring desaturations on top of monitoring the heart rate channel does not give better performance.



(a) Sepsis group

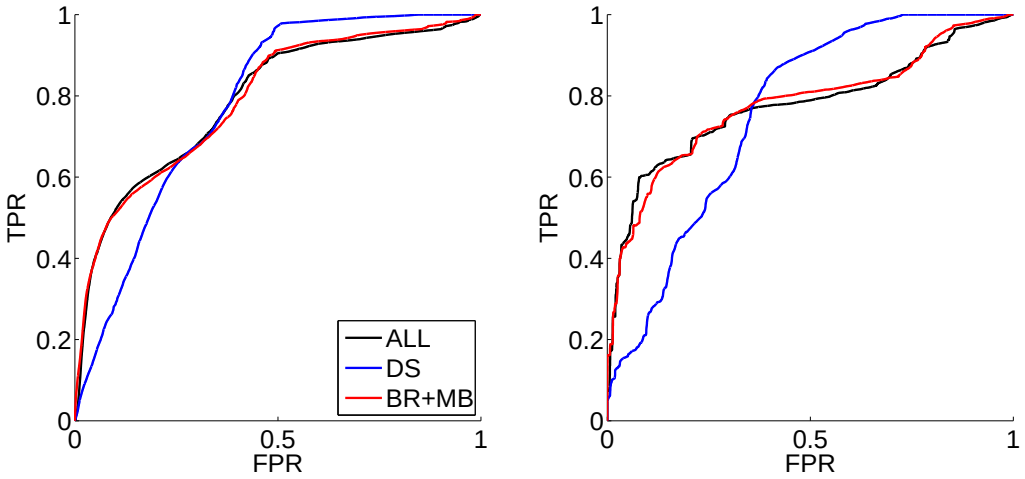


(b) Control group

Figure 5.11: Cross-validation inference for both patient groups using the explicit-duration AR-HMM. The top row of each figure represents the sepsis labelling. Normal periods are white, sepsis periods are black. Transitioning and treatment periods are not assigned any label. For each sepsis sample s^k or control sample c^k the top row of the corresponding image represents the filtering distribution and the bottom row represents the smoothing distribution.

Table 5.6: Summary statistics for the second-by-second analysis of explicit duration AR-HMMs modelling several sets of physiological events.

		ALL	BR+MB	DS
Filtering	AUC	0.80	0.80	0.78
	EER	0.30	0.30	0.30
Smoothing	AUC	0.79	0.79	0.76
	EER	0.27	0.27	0.32



(a) Filtering (b) Smoothing

Figure 5.12: ROC curves corresponding to different sets of physiological events modelled with an explicit duration AR-HMM for sepsis detection.

Also, due to better TPR values at high FPRs, monitoring only desaturations delivers a surprisingly good performance. However, when choosing an operating point from these ROC curves, we are more interested in the performance at low FPRs.

The event-type analysis is continued in the next section when looking at episode-based analysis.

5.6.3 Episode-based analysis

We also propose evaluating our models from the perspective of detecting *episodes* of infection. This analysis is intended to be closer to clinical practice than the second-by-second evaluation. Similar procedures have been used in a variety applications such as object detection [Everingham and Winn, 2012, §4.4], activity monitoring [Fawcett and Provost, 1999], keyword spotting [Young et al., 2006, §17.19] or clinical event detection [Quinn, 2007, §7.2.1]. In the following, we begin by describing our proposed evaluation method and then show the results of applying

Table 5.7: Summary statistics obtained by cross-validation in an episode-based analysis.

	Filtering		Smoothing	
	AP	F-score	AP	F-score
AR-HMM md	0.59	0.61	0.60	0.65
AR-HMM wmd	0.56	0.59	0.57	0.63
HMM	0.10	0.29	0.19	0.32
AR-HMM ed	0.59	0.65	0.63	0.69

it to the sepsis detection task. In the remainder of the section, we compare our approach with related work.

We consider that for the task of sepsis episode detection, evaluation via a PR curve is a better choice than via a ROC one, as true negative (TN) episodes are hard to define, and are not necessary for a PR curve. The algorithm we propose for the *episode-based* analysis is as follows:

1. The posteriors distributions over the binary sepsis variable z_t are converted to binary strings by thresholding.

2. Strings of 1s obtained at the above step correspond to predicted sepsis episodes. Since true episodes last for at least a few hours, we keep only instances longer than 1 hour.

3. A predicted episode can be either:

   - True Positive (TP), if it overlaps with the *sepsis* period but not with any *normal* period.
   - Unlabelled, if it is exclusively contained in either the transition or treatment periods.
   - False Positive (FP), otherwise.

In order to produce the PR curve, we chose as thresholds the quantiles of the marginal posterior distribution.

Importantly, if multiple true positives are detected for a sepsis patient, then only the first is recorded. This means that for any patient, only the first "alarm" is assumed to be significant and shall be used for evaluation.

In order to draw the PR curve, the counts needed to be normalised. Note that we normalise recall by the total number of positive examples (i.e. the number of infected patients) in our dataset. Thus, if for example we fail to detect the sepsis episode in a sepsis group sample, this error was penalised by recall.

Table 5.8: Summary statistics for the episode-based analysis in which explicit duration AR-HMMs are used to model different sets of physiological events.

		ALL	BR+MB	DS
Filtering	AP	0.59	0.53	0.33
	F-score	0.65	0.59	0.46
Smoothing	AP	0.63	0.61	0.25
	F-score	0.69	0.65	0.42

Table 5.7 shows the AP and maximum F-score for the same set of models as discussed in Section 5.6.1. The performance of the explicit duration model is again computed using nested cross-validation, as described in Section 5.6.1. Most of the findings in Table 5.4 are confirmed using PR metrics. Again, the performance of the explicit duration model exceeds the other models, although its filtering AP score equals that of the standard AR-HMM model.

Table 5.8 shows summary statistics from the physiological event evaluation of Section 5.6.2, now using the episode-based analysis. The latter reveals larger performance differences between the models than previously seen in Table 5.6, which showed results of the second-by-second analysis. Monitoring desaturations in addition to monitoring the heart rate channels improves both filtering and smoothing performance. When we assess the predicted infection episodes, monitoring only desaturations performs much worse than in the second-by-second analysis.

Our episode-based analysis approach is probably closest to the evaluation of the object detection task in Everingham and Winn [2012], partly because both draw PR curves. Another important similarity is that both methods normalise the TP count dividing by the total number of positives in the training set. We also use their algorithm for computing AP, where they set the precision for a given recall r to the maximum precision obtained for any recall greater than r. This method implicitly produces a monotonically decreasing PR curve. Moreover, the monotonicity constraint makes practical sense, as deviating from it implies considering dominated classifiers on the PR curve.

Fawcett and Provost [1999] have elaborated the Activity Monitoring Operatic Characteristic (AMOC) for evaluating solutions to the task of inferring event onset in time series. They propose a flexible ROC inspired framework which involves a scoring function and a false alarm function, both of which are application specific. The scoring function has the form $s(\tau, \alpha)$, and measures the value of flagging an alarm at time α, when the true onset of the positive event was at time τ. The false alarm function, $f(\alpha)$, quantifies the penalty for a false alarm at time α. Like us, the AMOC framework only accounts for the first TP. They also acknowledge that TNs are not well defined in this context, and normalise the false alarms by time metrics to obtain a

"false alarm rate". AMOC curves can be applied to our application, but only after deciding on appropriate forms for the s and f functions, and on the normalisation of false alarm counts.

Another related evaluation has been used in Quinn [2007] to assess clinical event detection. Their method draws an ROC curve, but involves thresholding the posteriors as a preprocessing step, which adds a free parameter to the evaluation. Also, they do not pay a penalty for missing an event.

5.6.4 Comparison with discriminative models

In order to study the benefits of modelling the distribution of clinical events using an AR-HMM we compared our framework with results obtained from using several discriminative models. Our approach was to extract features from the event annotations and use these features as input for the discrimonative models. We chose to apply logistic regression, decision trees[7] and a binary Gaussian Process classifier (bGPC). For the latter model see Rasmussen and Williams [2005, §3.3] and Rasmussen and Nickisch [2013][8]. As all of these approaches are designed especially for i.i.d. data; we first describe how training and testing data were obtained. We then show results and compare them to those obtained from using the AR-HMM.

In any of our discriminative models for sepsis detection, a data point consists of a collection of features computed over the preceding 1-hour interval and a label. The data points are extracted every 15 minutes from the monitoring data. For each of the baby-generated physiological events (see Section 5.4.3) we extract the following features: number of instances and total duration. Both of these counts are weighted proportionally to the total duration of missing data in the selected 1 hour interval. Other features we use are the number of handling instances in the 1-hour interval and the total duration of these instances. In order to label each of the data points, we directly applied the labelling scheme introduced in Section 5.5.

Table 5.9 shows summary statistics for both *second-by-second* and *episode-based* analyses. The results have been obtained using the same cross-validation procedure as in the previous sections. Among the discriminative models, logistic regression delivered the best performance for both types of analysis. This can be explained by over-fitting, as both the decision tree and bGPC outperformed logistic regression when training and testing on the whole dataset. However, when compared to the AR-HMM filtering results discussed in the previous sections,

[7] Here, we used Matlab's implementation of decision trees [Breiman et al., 1984], with the default tree splitting condition of at least 10 impure nodes. For pruning, the sub-tree with the best misclassification error was chosen by 10-fold cross-validation.

[8] We chose a bGPC with squared-exponential covariance, and learnt the hyper-parameters by maximising the marginal likelihood. The latter step is also known as Automatic Relevance Determination [Rasmussen and Williams, 2005, §5]. For bGPC inference we used the Laplace approximation [Rasmussen and Williams, 2005, §3.4] and to speed up inverting the covariance matrix we applied the FITC approximation with 100 inducing points [Snelson and Ghahramani, 2006].

Table 5.9: Leave-one-out sepsis inference summaries for several discriminative approaches

	Second-by-second		Episode-based	
Model	AUC	EER	AP	F-score
logistic regression	0.67	0.38	0.59	0.59
decision tree	0.61	0.45	0.43	0.52
bGPC	0.64	0.41	0.53	0.59

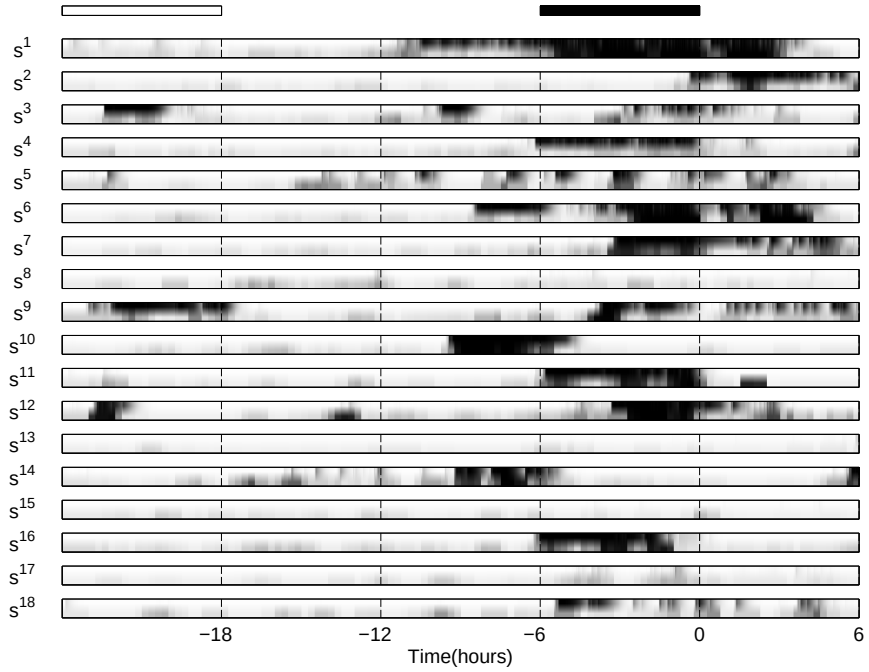
logistic regression does significantly worse in terms of AUC (0.80 for the AR-HMM and 0.67 for logistic regression), and matches them in terms of AP (see Tables 5.4 and 5.7).

Figure 5.13 provides a visual comparison between the filtering distribution of the explicit-duration AR-HMM and logistic regression. The latter model appears to perform worse than the autoregressive model at detecting true episodes of infection (see samples s^4, s^6 and s^{11}). In addition, logistic regression is more uncertain through the entire duration of the monitoring samples.

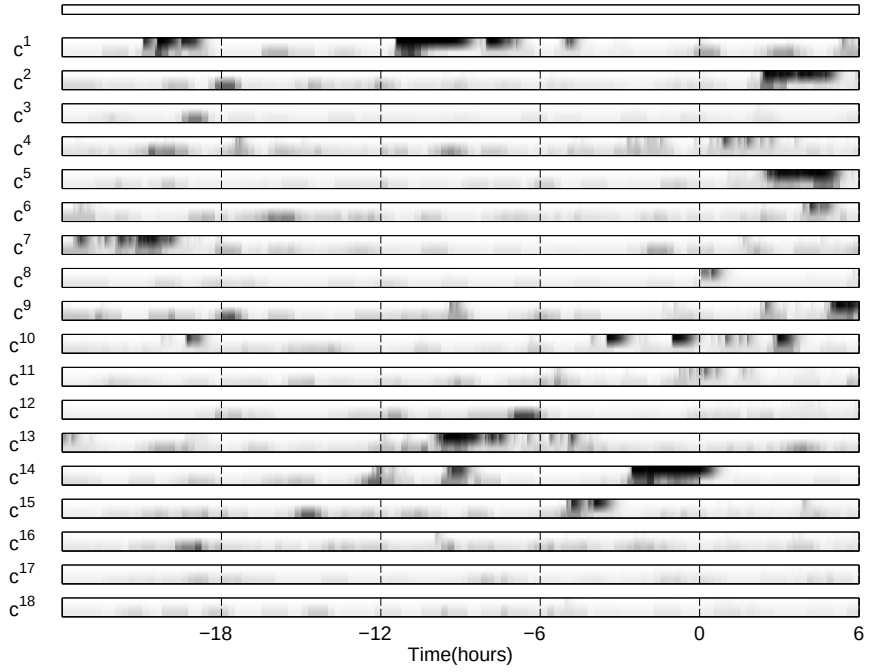
5.7 Summary

This chapter introduced a hidden variable probabilistic model capable of making early predictions about the onset of neonatal sepsis. The key characteristic of our approach was the extensive use of domain knowledge to facilitate both learning and inference. We have discussed the study design and described the annotation process. In addition, we have explained the sources of missing data in this application and provided a solution marginalising over such periods of missing at inference time. Apart from evaluating our models with a standard ROC analysis, we have proposed the more clinically relevant *episode-based* analysis.

The results show that by monitoring the incidence of baby-generated physiological events we can often detect sepsis well in advance of the time a positive blood test was taken. Importantly, marginalising over missing data increased performance. We have provided empirical evidence that monitoring oxygen desaturations in addition to heart rate channel events does bring performance improvements. In addition, we have experimentally shown that the best sepsis detection results within the AR-HMM models considered are achieved by explicit duration modelling. Finally, it has been empirically demonstrated that our generative models perform at least as good as several discriminative benchmark approaches.



(a) Sepsis group



(b) Control group

Figure 5.13: Cross-validation inference for both patient groups. The top row of each image corresponds to the filtering distribution of the explicit-duration AR-HMM. The bottom row corresponds to logistic regression.

Chapter 6

A hierarchical switching linear dynamical system applied to the detection of sepsis in neonatal condition monitoring

In the previous chapter, we have presented a condition monitoring framework which takes expert annotations of low-level clinical events as input, and which produces probabilistic predictions about the onset of sepsis. However, the need for expert annotations is an important limitation for the practical implementation of the model. In this chapter, we discuss a model which takes vital signs monitoring data as input and outputs both sepsis predictions and posterior distributions of clinical events (Figure 6.1).

Section 6.1 introduces the Hierarchical Switching Linear Dynamical System (HSLDS), our proposed model for dynamical systems with complex interactions between modes of operation. The relation to previous work on hierarchical models for sequential data is discussed in Section 6.1.1. We continue by briefly showing how inference in the HSLDS is run (Section 6.1.2). Model training is explained in Section 6.1.3, where we introduce a "deep learning"-inspired algorithm for fitting factor transition matrices. The application of the HSLDS to the task of detecting sepsis in NICU patients is discussed in Section 6.2. In more detail, we explain data preprocessing (Section 6.2.1.1), learning measurement channel models and physiological event distributions (Section 6.2.1.2), learning clinical factor transitions (Section 6.2.1.3) and inference in the presence of missing data (Section 6.2.2). Section 6.3 presents the experiments we have performed for assessing the performance of the HSLDS for neonatal condition monitoring. We discuss sepsis detection results in Section 6.3.1 and physiological event posteriors in Section 6.3.2. A brief summary of the chapter is provided in Section 6.4.

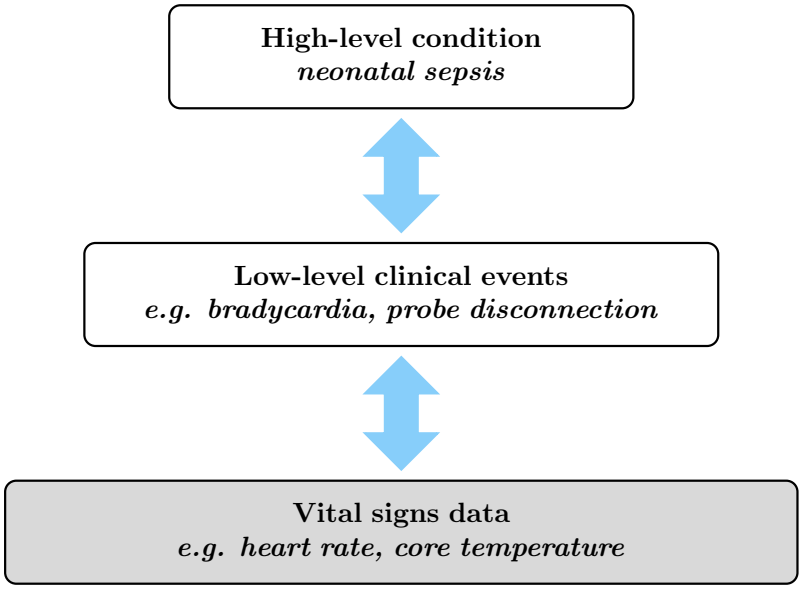


Figure 6.1: Inferring sepsis and low-level events from the raw monitoring data.

Parts of this chapter have been adapted from Stanculescu, Williams, and Freer [2014].

6.1 The HSLDS

The switching linear dynamical system (reviewed in Section 2.4.1) models sequences in which several modes of operation switch in explaining the data. When the modes of operation are collectively determined by the states of K factors, the discrete state variable of the SLDS is factorised and the resulting model is referred to as a Factorial Switching Linear Dynamical System (FSLDS). The FSLDS has been introduced in Section 2.4.2 and its application to neonatal condition monitoring has been discussed in Section 4.2.2. We reiterate that an important assumption made by the FSLDS is that the factors are a priori independent.

In the HSLDS, we propose relaxing this assumption by introducing a hierarchical structure for the discrete hidden variables. The discrete state is now represented by two layers of variables (see Figure 6.2). The top layer variable z_t controls the Markovian dynamics $p(f_t^{(\cdot)}|z_t, f_{t-1}^{(\cdot)})$ used by each factor. Thus, the top layer can capture hidden correlations between the factors. Conditional on the setting of the top layer switch variable z_t, the model becomes equivalent to an FSLDS. Thus, the HSLDS can be thought of as a dynamical mixture of FSLDS models. If we define a full expansion of the discrete hidden state as $s_t \triangleq z_t \otimes f_t^{(1)} \otimes f_t^{(2)} \otimes ... \otimes f_t^{(K)}$, then the joint distribution of the HSLDS can be written as:

$$p(s_{1:T}, \mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = p(s_1)p(\mathbf{x}_1|s_1)p(\mathbf{y}_1|\mathbf{x}_1, s_1)\prod_{t=2}^{T} p(s_t|s_{t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t)p(\mathbf{y}_t|\mathbf{x}_t, s_t), \quad (6.1)$$

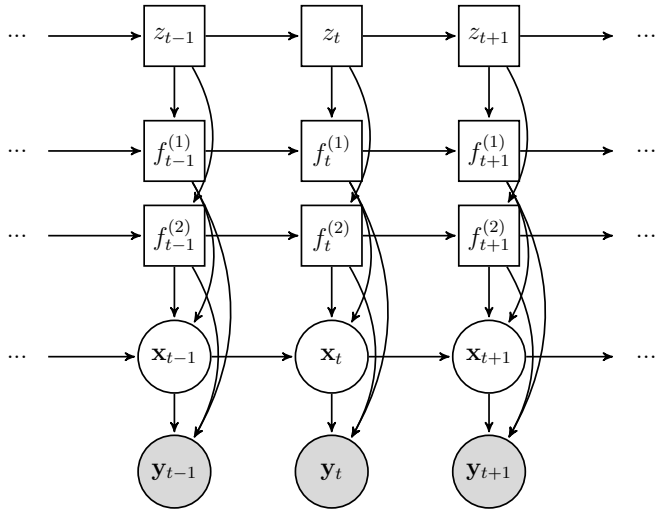


Figure 6.2: DAG of an HSLDS with two factors ($K = 2$).

where

$$p(s_1) = p(z_1) \prod_{k=1}^{K} p(f_1^{(k)}|z_1),$$
$$p(s_t|s_{t-1}) = p(z_t|z_{t-1}) \prod_{k=1}^{K} p(f_t^{(k)}|z_t, f_{t-1}^{(k)}).$$

Note that the top hidden layer is conditionally independent of the continuous variables given the factor settings:

$$\mathbf{x}_{1:T}, \mathbf{y}_{1:T} \perp\!\!\!\perp z_{1:T} | \mathbf{f}_{1:T}, \tag{6.2}$$

where we have defined $\mathbf{f}_t \triangleq [f_t^{(1)}, f_t^{(2)}, \ldots, f_t^{(K)}]$. This simplifies both learning and inference.

6.1.1 Relation to previous work on hierarchical models for sequences

In Chapter 2, we have reviewed probabilistic models suitable for modelling vital signs monitoring data. Here, we compare our proposed HSLDS with several previously proposed hierarchical models for sequential data.

The only previous use of an HSLDS we are aware of in the literature is the work of Zoeter and Heskes [2003]. They use an HSLDS with the goal of producing hierarchical visualisations of sequential data. More precisely, their motivation is to allow a successive refinement of a visualization, starting from projecting onto a single LDS with a two-dimensional (2-d) hidden space. This can be broken down into a SLDS of 2-d LDS models, and then each 2-d LDS can be further independently decomposed into a SLDS. Thus, a set of lower-level states correspond to one higher-level state. Also note that their use case involves interaction from the user to initialise the decomposition.

In contrast, we more naturally think of building our model bottom up, first identifying a set of factors for the FSLDS, and then modelling their correlations with a top-level variable. Notice that in our work the state of the top-level variable affects all of the second-level variables below it.

There are also some similarities between our work and the paper by Taylor et al. [2010]. In their model, the **x** dynamics are modelled by an Implicit Mixture of Conditional Restricted Boltzmann Machines (imCRBM). This is similar to us in that the CRBM part of the model uses a number of discrete latent variables (analogous to our **f**'s) to affect the **x** dynamics. The implicit mixture variable (analogous to our z) switches between different dynamics models. Of course, the details of the model are quite different as it is in part undirected, and that there are no explicit chains of discrete latent variables through time; instead these variables "hang off" the **x** chain.

The HMM model has also been elaborated hierarchically by Fine and Singer [1998] to give the hierarchical hidden Markov model (HHMM). As in our HSLDS, discrete states at different levels of the HHMM hierarchy have the ability to model different stochastic levels present in sequential data. The goal is to capture multi-scale correlations between observations. A notable difference is that the HHMM does not model a hidden continuous state as does the HSLDS.

Another related approach is the hierarchical sequential classification framework of Jordan et al. [1997]. Their Hidden Markov decision tree (HMDT) is a temporal extension of a probabilistic decision tree [Jordan and Jacobs, 1993]. The HMDT adds the constraint that a decision at any level of the tree also depends on the decision made at the previous time step at the same level. Note that unlike the HSLDS, the HMDT is a discriminative approach which models the conditional distribution of the outputs given the inputs.

6.1.2 Inference

Since real-time inference is the major concern in physiological condition monitoring, we are mainly interested in marginal filtering distributions. More precisely, we require sepsis predictions of the form $p(z_t|\mathbf{y}_{1:t})$ and clinical event posteriors $p(f_t^{(\cdot)}|\mathbf{y}_{1:t})$. These marginal posteriors can be immediately obtained from the one-step filtering marginal of the fully expanded state $p(s_t|\mathbf{y}_{1:t})$. Thus, running SLDS inference suffices for HSLDS inference. Note that the more general goal of SLDS filtering is inferring $p(s_{1:t}, \mathbf{x}_{1:t}|\mathbf{y}_{1:t})$.

Exact SLDS inference requires computing Gaussian mixtures with a number of components exponential in the length of the sequence [Lerner and Parr, 2001]. A review of approximate SLDS has been provided in Section 2.4.1. Here, we apply the Gaussian Sum approximation as described in Murphy [1998]. The algorithm is provided in Appendix A.

When the hidden discrete state is a cross-product of variables, we can speed up inference by allowing at most one variable to change its setting at each time step. This procedure has

been previously discussed in Quinn et al. [2009] or Kolter and Jaakkola [2012].

A particular aspect of the baby monitoring application is the presence of several missing data sources. The treatment of this problem will be discussed in detail in Section 6.2.2.

6.1.3 Learning

For performing condition monitoring, HSLDS learning is similar to FSLDS learning to a large extent [Quinn, 2007, §5]. FSLDS learning has also been reviewed in Section 4.2.2. Here, we first emphasize the most significant common aspects between HSLDS and FSLDS learning, and then focus on specifics of the former.

The central assumption for learning is that there are a number of interpretable regimes for which labelled data are available. In the HSLDS, labelled data are of the form $\{\mathbf{y}_t, z_t, \mathbf{f}_t\}$.

As in the FSLDS case, the availability of labelled data makes learning equivalent to learning one LDS model for each switch setting. We parameterise LDS dynamics as autoregressive processes and will further discuss this choice in Section 6.2.1. In general, ML parameters for the LDS can be found using Expectation Maximisation (EM) as proposed by Ghahramani and Hinton [1996].

Learning is performed independently for each factor, and then the fitted parameters are carefully combined for each switch setting. This procedure is greatly simplified by considering the interactions between factors. For instance, the activation of one factor might "overwrite" any effect of another factor on certain observation channels. As already explained in Section 4.2.2, in the neonatal monitoring application domain knowledge is used to define a factor overwriting ordering. This will be further discussed in Section 6.2.1.

For the HSLDS in particular, we use the conditional independence between the continuous variables and the top layer discrete variables (eq. 6.2) to simplify learning further. This means that the parameters of the continuous variable distributions do not depend on the setting of z_t.

A straightforward way of learning the Markov transition matrices for individual factors $p(f_t^{(\cdot)}|z_t, f_{t-1}^{(\cdot)})$ would be to make use of the labelled data and maximize the conditional likelihood $p(\mathbf{f}_{1:T}|z_{1:T})$. Estimates of the factor transition probabilities have the following form:

$$p(f_t^{(\cdot)} = l | z_t = j, f_{t-1}^{(\cdot)} = m) = \frac{n_{l|mj}^{(\cdot)} + n_0}{\sum_{l'}(n_{l'|mj}^{(\cdot)} + n_0)}, \tag{6.3}$$

where $n_{l|mj}^{(\cdot)}$ is the number of transitions from state m to state l for factor $f^{(\cdot)}$ under the z-regime j, counted over all the training data. The constant count n_0 comes from placing a Dirichlet prior which prevents probabilities from being too close to zero.

However, we have found that an alternative "deep learning" style method can give rise to better results (Section 6.3.1). Although the $\mathbf{f}$ data are available at training time, at test time

these labels must be inferred from the $\mathbf{y}$ data. Hence it makes sense to build a model which looks at the actual inferences of the factors, rather than the ground truth labels.

If $\mathbf{Y}$ is the training set of sequences and the corresponding $\mathbf{F}$ are treated as hidden variables, we could use EM and attempt to optimise $p(\mathbf{Y}|\mathbf{Z})$. The M-step is equivalent to maximizing the expected complete data log-likelihood:

$$Q = \mathbb{E}_{p(\mathbf{X},\mathbf{F}|\mathbf{Y},\mathbf{Z})} \log p(\mathbf{Y},\mathbf{X},\mathbf{F}|\mathbf{Z}), \tag{6.4}$$

where $p(\mathbf{X},\mathbf{F}|\mathbf{Y},\mathbf{Z})$ was computed in the preceding E-step using the old parameter settings. In Appendix C.2, we provide a full expansion of the expected complete data log-likelihood for a standard SLDS, which is useful for monitoring the learning process.

Taking partial derivatives of eq. 6.4, we find that factor transition estimates are of the from:

$$p(f_t^{(\cdot)} = l|z_t = j, f_{t-1}^{(\cdot)} = m) = \frac{\tilde{n}_{l|mj} + n_0}{\sum_{l'}(\tilde{n}_{l'|mj} + n_0)}, \tag{6.5}$$

where

$$\tilde{n}_{l|mj} = \sum_t p(f_{t-1}^{(\cdot)} = m, f_t^{(\cdot)} = l|\mathbf{Y},\mathbf{Z}) I(z_t = j),$$

which is commonly referred to as a "soft" data count; I is the indicator function, and the sum is taken over all t in the training data.

Running EM until convergence is likely to be unsatisfactory, as there are no guarantees that the learnt factor transition matrices would produce good factor posteriors. Our solution is to approximate $p(\mathbf{F}|\mathbf{Y},\mathbf{Z})$, by $p_{FSLDS}(\mathbf{F}|\mathbf{Y})$. Here, the FSLDS model is trained using the standard learning routine of Quinn et al. [2009] and the factor models discussed in Section 6.2.1, and is thus unaware of the existence of multiple z-regimes. In practice, we found it sufficient to obtain "soft" counts of pairwise filtering marginals $p_{FSLDS}(f_{t-1}^{(\cdot)}, f_t^{(\cdot)}|\mathbf{y}_{1:t})$ for each training sequence. Since FSLDS posteriors do not depend on the learnt HSLDS parameters, the method is non-iterative.

This procedure follows ideas in the "deep learning" literature [Hinton et al., 2006] where layer-wise training of a model is carried out. Similar ideas can also be found e.g. in Karklin and Lewicki [2005] or Farhadi et al. [2009], although in all these cases the models are not for time series.

Finally, estimates of the Markov transition matrix $p(z_t|z_{t-1})$ are learnt from the z-labels. Also note that in the absence of the labelled data, unsupervised learning for the full model would be possible using EM.

6.2 Application to neonatal condition monitoring

This section is concerned with applying the HSLDS to condition monitoring in Neonatal Intensive Care Units (NICUs). Neonatal sepsis and its diagnosis have been discussed in Section 4.1,

and a description of the dataset collected for this research has been provided in Section 5.3. Here, we explain how the neonatal monitoring problem can be solved by formulating it as learning and inference in an HSLDS.

6.2.1 Learning a sepsis detection model

We first turn our attention to how the baby monitoring HSLDS is trained. We begin by explaining the need for preprocessing on some of the measurement channels. We then discuss parameter fitting for the continuous variable distributions and finish the section with learning the hidden discrete layers of the HSLDS.

6.2.1.1 Preprocessing

In order to alleviate the problem of measurement quantisation, preprocessing was needed on several monitoring channels. The oxygen saturation and temperature channels are most affected. The problem occurs when the resolution of the monitoring device is small relative to the possible changes of the signal over a time step.

In our work, we have applied the preprocessing method discussed in Quinn [2007]. Their solution relies on a system inertia assumption which translates into the smoothness of the affected channels. It essentially performs linear interpolation between time steps at which readings change value.

We have also experimented with the alternative solution of modelling the quantisation noise as part of the observation noise of the Kalman filter. Given a quantisation step Δ, we assumed the quantisation noise follows the uniform distribution $\mathcal{U}(-0.5\Delta, 0.5\Delta)$. Then, the observation noise variance is set by minimizing the KL divergence between a (zero-mean) Gaussian and the assumed uniform distribution (i.e. $\mathbf{R} = \frac{1}{12}\Delta^2$). However, this approach proved inferior to the smoothing solution described above.

Another interesting idea is to assume that instead of observing a single quantised recording y_t at any time step, we observe an interval of the form $\mathrm{Y}_t = [y_t - 0.5\Delta, y_t + 0.5\Delta]$. The goal of the Kalman filter changes from estimating $p(\mathbf{x}_t|y_{1:t})$ to determining $p(\mathbf{x}_t|\mathrm{Y}_{1:t})$. Such a solution is discussed in Duan et al. [2008], alongside an approximate inference routine. In theory, the inference procedure can be incorporated into an EM learning routine, but we found this procedure too laborious for our purposes.

6.2.1.2 Learning continuous variable distributions

A natural classification of the regimes appearing in the NICU monitoring application is: *stability*, *known factors* and *unknown factors* (see Section 4.2.2).

Babies within the NICU are in a stable condition for much of the time, generally being

asleep and motionless. We named this regime *stability* (see Section 3.2) and separately fit an LDS model to each measurement channel. There are many ways to parameterise an LDS, and also several training algorithms can be applied (Section 2.3.2). For the baby monitoring application, we chose to parametrise the hidden dynamics as autoregressive processes. One way to train such LDS models was proposed by Quinn [2007] and has been already discussed in Section 4.2.2. However, we empirically found a different procedure to be more numerically stable. Our solution was to adapt the standard unconstrained EM routine of Ghahramani and Hinton [1996] to the constraints implied by the state-space parameterisation of ARMA models given in Section 2.3.2. In this setting the E-step is equivalent to running the standard Kalman smoother, while M-step updates can be found in Appendix C.1.

As in the FSLDS case, separately fitting an LDS to each measurement channel results in the dynamics and observation matrices to have a block structure. This has some advantages including simplifying the addition or removal of physiological channels or the easy incorporation of factor-channel dependences.

When clinical events associated with stereotypical patterns occur on the monitoring traces, the regimes will be referred to as *known factors*. Several such factors have been introduced in Chapter 3, and some have been modelled in the FSLDS framework [Quinn et al., 2009]. Here, we focus on two physiological events: braydcardias and desaturations (see Section 3.3).

Only bradycardias have been previously monitored in the FSLDS framework [Quinn et al., 2009]. The idea was to model the event using the same parameter set as for normality, except for an inflated system noise covariance matrix on the heart rate channel (HR). The inflation coefficient was learnt via an EM procedure. First, a disadvantage of this model is that sampling from it often produces traces dissimilar to real bradycardias (see the second row of Figure 6.3). Second, other clinical events can be falsely classified as bradycardias. One such event is tachycardia, a physiological event consisting of a sudden raise in HR measurements followed by recovery.

In this work, we started from the observation that both bradycardias and desaturations are characterised by a drop in the monitored signal (a slowing of the heart rate for bradycardias, and a decrease in the saturation of oxygen in arterial blood for desaturations), after which measurements rise back. Thus, it makes sense to model these factors as two-stage events. The first stage corresponds to measurements dropping and can be explained by an exponential decay, the discrete time equivalent of which is an $AR(1)$ process. To set the mean of the decay process, we first compute the empirical distribution F of minimum channel measurements during events. The quantile q^* of F corresponding to $F(q^*) = 0.05$ is chosen to be the decay mean. In the second stage of the event the measurements rise back to approximately the original level. This will be referred to as the recovery stage. Recovery dynamics are also modelled as an $AR(1)$ process, where the mean is now the same as the channel's $stability$ mean. The

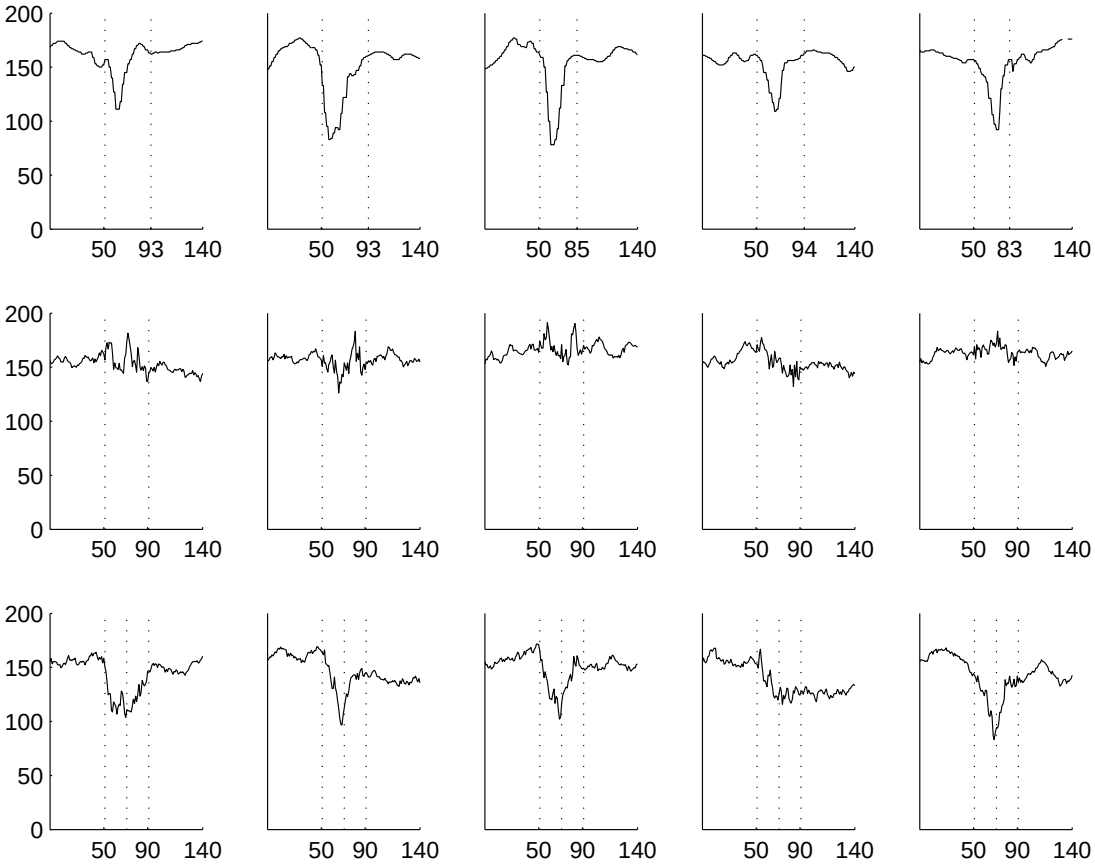


Figure 6.3: Real versus synthetic bradycardias. First row: real bradycardias examples. Second row: bradycardias sampled from the inflated system noise model. Third row: bradycardias sampled from the "decay-recovery" model. The vertical dotted lines represent expert annotation for the real data and switching times of the SKF for the synthetic data. Samples drawn from our proposed model look similar to the real instances. See text for further explanation.

Table 6.1: Overwriting ordering of factors

Channel	Bradycardia	Desaturation	X	Stability
HR	•		•	•
SO		•	•	•

parameters for both decay and recovery models are learnt by running EM, where we chose the dynamics initialisation $\mathbf{A} = \mathbf{0}$.

An advantage of the decay-recovery approach is that samples from the model look more similar to real data when compared to the inflated system noise model. The first row in Figure 6.3 shows five real bradycardias selected from a single randomly chosen patient. The second row shows five samples from the inflated system noise covariance matrix model and the final row has samples from the novel decay-recovery model. The samples are drawn from a SLDS, where for simplicity we have fixed the transition times a priori. We have used stability parameters learnt from the heart rate data of the same patient, and bradycardia models trained on examples from all the other patients. Judging by the quality of the samples obtained, we can conclude that the decay-recovery model is the better generative model.

Finally, certain events cannot be explained by either stability or by any of the known factors. Here, we follow the X-factor approach of Quinn et al. [2009] for modelling these "known *unknowns*". A description of the X-factor can also be found in Section 4.2.2. Note that as the X-factor can claim patterns of both physiology and artifact, we do not use it directly for inferring the presence of sepsis.

Once the factor models have been separately learnt, they are combined using the overwriting order shown in Table 6.1. For each measurement channel, factors placed towards the left of the table overwrite factors placed towards the right.

6.2.1.3 Learning discrete variable distributions

In our model, the top discrete layer of the HSLDS models the state of the sepsis infection. Here, we assume z_t is a binary variable taking on values $z_t = sepsis$ or $z_t = normal$.

Labelling the sepsis indicator variable was non-trivial, and the proposed the labelling scheme has been explained in Section 5.5. As in the AR-HMM discussed in the previous chapter, we employed the labelling scheme to obtain an estimate of $p(z_t|z_{t-1})$ using data counts.

For learning the z-conditioned *known* factors' transition matrices, we apply the procedure explained in Section 6.1.3; see eq. 6.4 and the surrounding text. The X-factor's incidence is assumed to be independent of the state of the infection, and thus the factor transition matrix is copied from the previously learnt FSLDS.

Table 6.2: Missing data sources affecting baby-generated physiological events.

	Bradycardia	Desaturation
Handling	•	•
Oximeter error		•
HR dropout	•	
SO dropout		•

6.2.2 Inference with missing data

We reiterate that this work is centred around the idea of monitoring baby-generated bradycardias and desaturations in order to predict the onset of sepsis. However, there are periods of time during which labels for these events cannot be provided even by an expert annotator. We will treat such periods as missing data. There are three distinct sources of missing data: probe dropouts, oximeter errors and patient handling. As these sources have already been treated in detail throughout Section 5.4, here we only highlight them in HSLDS context. We then explain how inference can be performed in presence of missing data periods.

Patients are regularly handled by clinical staff (e.g. for changing nappies). For sepsis detection we chose to analyse only physiological events happening outside such episodes (see Section 5.4.3). The work presented in this chapter still relies on having expert annotations for handling. Note that Quinn et al. [2009] have shown that these episodes can be inferred by monitoring environmental channels such as the incubator's humidity, but such channels have not been available in this work (also see Section 3.4).

Probe dropouts can be readily recognised by the zero values on the recorded channels.

An oximeter error occurs when there is a disagreement between the HR and PR traces. Here, we adopt the approach in Stanculescu et al. [2013], where an automated oximeter error detection algorithm has been applied as a preprocessing step. For a detailed description of the method see Section 5.4.2.

Table 6.2 shows how physiological events are affected by the presence of each missing data source.

For performing inference with missing data, we extend the ideas in Quinn et al. [2009]. Whenever a missing data source is present, the measurements do not carry information about the true physiology of the patient, and thus should not influence the hidden state estimates. The latter continue to evolve according to the dynamics equations, but without measurement update. Technically, rows of the observation matrix are set to zero whenever there is missing data on the corresponding measurement channel. For these channels the Kalman gain will be zero. Thus, the corresponding hidden continuous state dimensions will be estimated with increasing uncertainty before reaching the stable state of the Kalman filter.

Table 6.3: Sepsis inference summaries using 9-fold cross-validation

	Second-by-second		Episode-based	
Model	AUC	EER	AP	F-score
AR-HMM	0.72	0.34	0.62	0.65
HSLDSdeep	0.69	0.37	0.51	0.54
HSLDSukf	0.66	0.39	0.59	0.65
HSLDSkf	0.62	0.41	0.45	0.47

In general, the stable state covariance $\mathbf{P}$ for a Kalman filter is the solution of the Lyapunov equation $\mathbf{P} = \mathbf{APA}^T + \mathbf{Q}$. This is given by: $\text{vec}(\mathbf{P}) = (\mathbf{I} - \mathbf{A} \otimes \mathbf{A})^{-1} \text{vec}(\mathbf{Q})$, where vec() is the vectorisation operator and $\otimes$ is the Kronecker product [Hamilton, 1994]. Also note that for any stable Kalman filter, the state estimates will exponentially converge towards the process's mean.

6.3 Experiments

This section describes the experiments we have performed to assess the neonatal condition monitoring model introduced in Section 6.2. The detection of sepsis is discussed in Section 6.3.1. Section 6.3.2 is concerned with the quality of physiological event posteriors.

For assessing the HSLDS model we use the same dataset as that used to analyse the AR-HMM model. For details on data collection and population demographics see Section 5.3. The annotation process has been described in Section 5.4. Note that for simplicity in the following we have included mini-bradycardias in the definition of bradycardias.

In order to reduce bias, we tested all our predictions using N-fold cross-validation. Considering the size of our dataset we decided to use $N = 9$ folds. Each fold contains 4 data samples, 2 from each patient group. The 2 control samples are chosen such that they belong to the same patient. Apart from these constraints, the folds have been randomly chosen.

6.3.1 Sepsis detection

To better understand the effectiveness of the HSLDS, we first compare its predictions against filtering results obtained with the AR-HMM model handling missing data, which was discussed in Chapter 5[1]. We then test the HSLDS against a discriminative approach of inspired by the work of Griffin et al. [2003].

While the HSLDS infers the posterior distributions of bradycardias and desaturations, the AR-HMM uses expert annotations of these events as input. Note that in the AR-HMM it

[1] In the previous chapter, this model was referred to as the "AR-HMM md".

was possible to run inference exactly and also to marginalise over the missing data exactly (see Section 5.2). In the HSLDS, we use the approximate inference algorithm discussed in Section 6.1.2 and handle missing data as explained in Section 6.2.2. For the purposes of this chapter, the central question is how well the HSLDS inferences match those of the AR-HMM.

In the following, we discuss several HSLDS models. The models differ in the way factor transitions are learnt as follows:

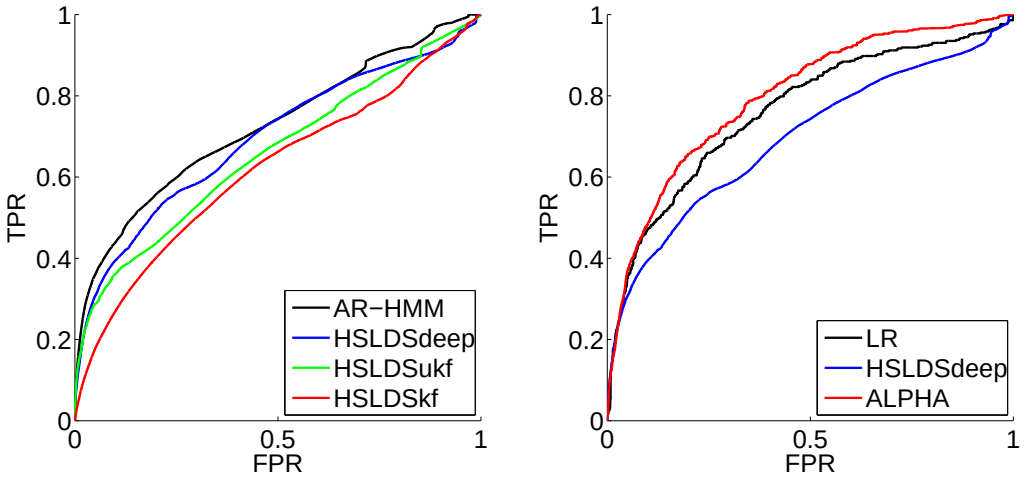
- HSLDSdeep is learnt using the greedy "layer-wise" procedure explained in Section 6.1.3;
- HSLDSukf (i.e. HSLDS with UnKnown Factors) is obtained by using 10 EM iterations to optimise $p(\mathbf{y}_{1:T}|z_{1:T})$, where in the M-step we update the factor transition matrices only using eq. 6.5;
- HSLDSkf (i.e. HSLDS with Known Factors) uses the expert annotations to learn the factor transitions.

In a similar fashion to assessing the discrete state models of the previous chapter, we evaluate the results using two different metrics. This offers the possibility to reveal different aspects of performance, as described previously.

Firstly, our main interest is in the *second-by-second* analysis of the inferences produced by our hierarchical models. For this purpose, we use the z-labels and drew the ROC curves shown in Figure 6.4a. The AUC and EER were computed by aggregating predictions over all folds, and are shown in Table 6.3. Compared to HSLDSukf and HSLDSkf, HSLDSdeep produced results which are closer, albeit still inferior, to the AR-HMM benchmark.

We obtained more insight into how the HSLDS predictions compare against the AR-HMM results via an N-fold cross-validated paired t-test on the AUC. We found the performance difference between the AR-HMM and our proposed HSLDSdeep model not to be statistically significant ($p = 0.552$). This is a good indication that the HSLDSdeep model can be used instead of the AR-HMM, and thus significantly reduce the amount of expert input required. At the same time the performance difference between the AR-HMM and the HSLDSukf model is statistically significant ($p = 0.049$). Even though this p-value is close to the standard significance level $\alpha = 0.05$, the null hypothesis is rejected, and thus HSLDSukf should not be used instead of the AR-HMM. For the latter model, HSLDSkf, we also obtained a statistically significant difference when compared against the AR-HMM ($p = 0.0064$).

We provide the second-by-second sepsis inference results produced by both the AR-HMM and the best performing HSLDS model, HSLDSdeep, in Figure 6.5. In general, there is a strong correlation between the predictions of the two models and we find the inferences of HSLDSdeep to be a good match to those obtained with the AR-HMM. However, in samples s^2, s^7 and s^{11} HSLDSdeep detects sepsis noticeably later than the AR-HMM, and in samples



(a) Comparison against the AR-HMM

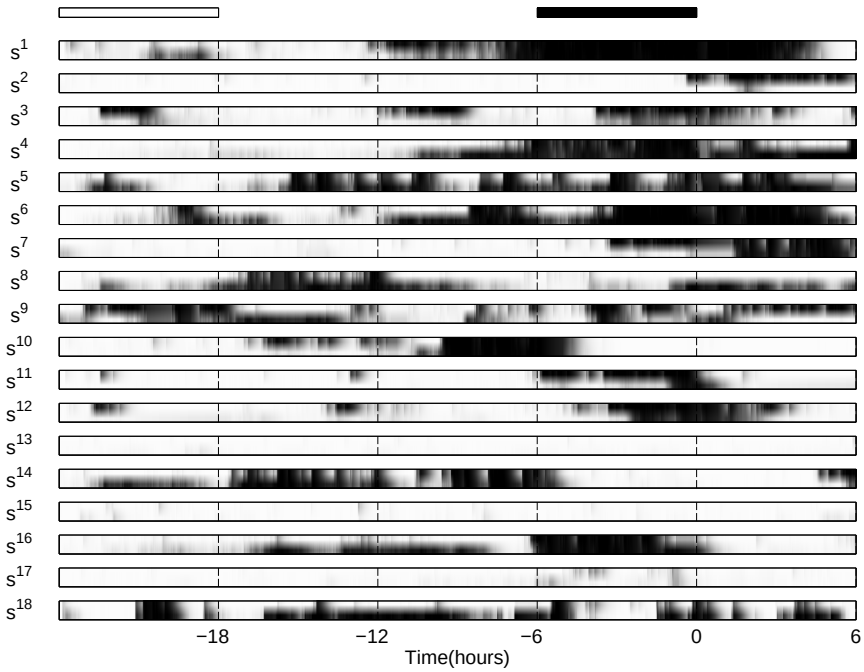
(b) Comparison against LR

Figure 6.4: ROC curves for comparing the performance of several HSLDS models against other sepsis detection models. In panel (a), we notice that HSLDSdeep is closest in performance to the AR-HMM model. Panel (b) shows that although the performance of the logistic regression model (LR) is better than HSLDSdeep, an α-mean combination of their predictions (ALPHA) outperforms both. See text for detailed explanations.

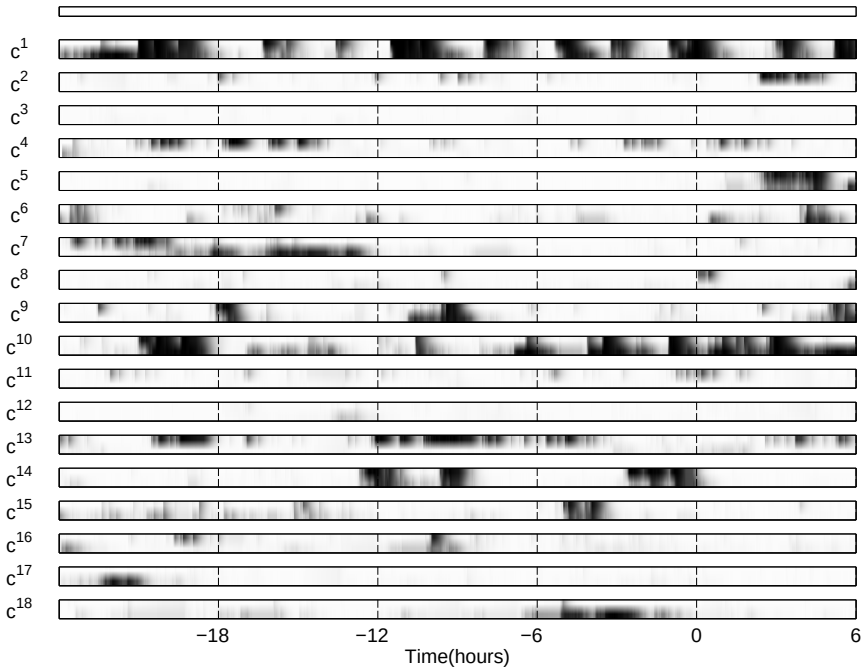
s^4 and s^6 it does so earlier. In the control group, HSLDSdeep does slightly worse on samples c^7 and c^{18}, but outperforms the AR-HMM on samples c^4 and c^{13}.

Secondly, we analyse the inferred *episodes* of infection and draw precision-recall (PR) curves (See Section 5.6.3 for details on this evaluation procedure). Here we report average precision (AP) and the maximum F-score (see Table 6.3). In terms of the former metric, the performance of HSLDSdeep is again closer to the AR-HMM than the HSLDSkf. However, HSLDSukf delivers the best AP in the episode-based analysis. This is partly due to the fact the unsupervised learning resulted in factor transitions with short staying times, which in turn produced less smooth sepsis predictions. At the same time, our episode-based analysis ignores episodes shorter than 1 hour, as these would be too short to be considered real episodes of infection.

We have also compared the HSLDS with a purely discriminative sepsis detector. The latter is a logistic regression (LR) model replicating the work of Griffin et al. [2003]. In a similar fashion to the ideas in Section 5.6.4, we chose to extract features over 1 hour intervals and did so every 15 minutes. The feature set consists of summary statistics including the mean, the median, the 10th and 90th quantiles, variance, skewness, kurtosis and sample asymmetry. This replicates the HRC feature set used in Griffin et al. [2003] (See Section 4.1 for more details). Again, we naturally adapted our sepsis labelling scheme as previously discussed in Section 5.6.4.

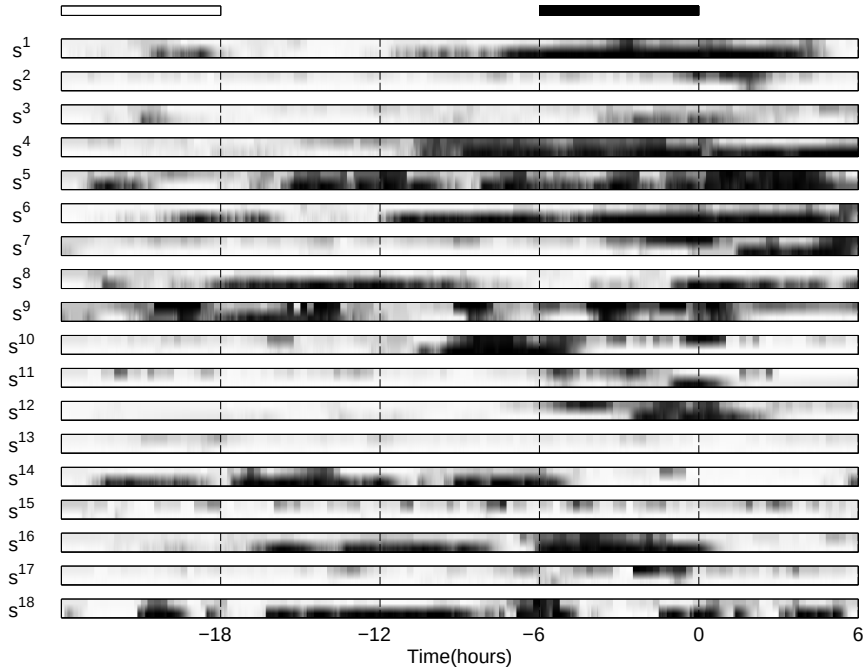


(a) Sepsis group

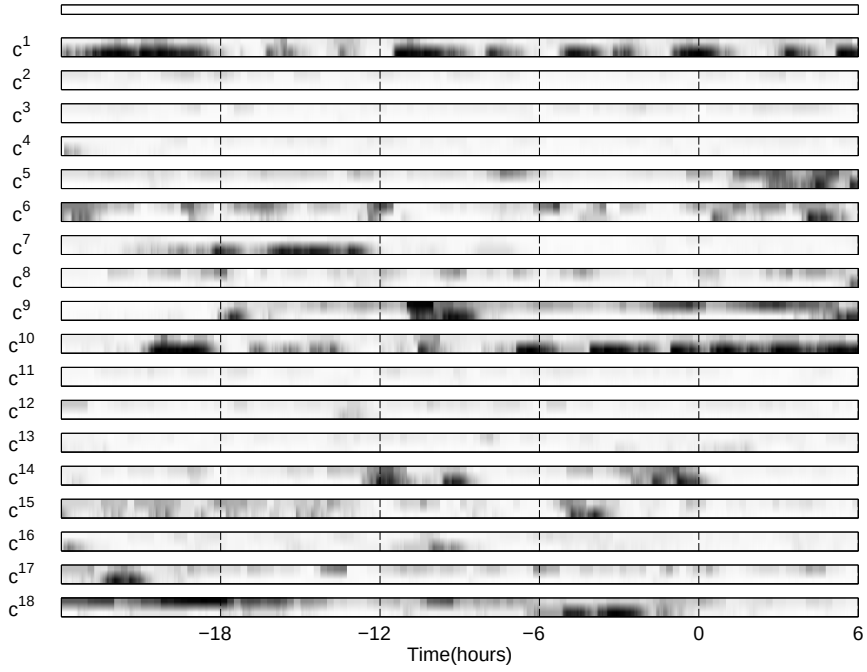


(b) Control group

Figure 6.5: Sepsis filtering distributions obtained using 9-fold Cross-Validation. On the x-axis, 0 denotes the time the positive blood sample was taken. For each group, the top row represents the sepsis labelling: normal periods are white (probability 0), sepsis periods are black (probability 1); transitioning and treatment periods are not assigned labels. For each data sample the top row corresponds to the AR-HMM model, the bottom row corresponds to HSLDSdeep.



(a) Sepsis group



(b) Control group

Figure 6.6: Sepsis filtering distributions obtained using 9-fold Cross-Validation. For each data sample the top row corresponds to logistic regression, the bottom row corresponds to HSLDS-deep.

Table 6.4: HSLDS, logistic regression and thier α-mean aggregation: sepsis inference summaries using 9-fold cross-validation

	Second-by-second		Episode-based	
Model	AUC	EER	AP	F-score
HSLDSdeep	0.69	0.37	0.51	0.54
LR	0.76	0.30	0.55	0.60
ALPHA	0.80	0.28	0.62	0.63

Figure 6.6 represents a comparison between logistic regression (top row for each sample) and HSLDSdeep (bottom row for each sample). HSLDSdeep is better at identifying the sepsis episodes for samples s^1, s^3, s^6, s^8 and s^{16}. At the same time, logistic regression seems to do better on samples s^2, s^7 and s^{12}. In the control group, the HSLDS outperforms logistic regression on samples c^6, c^9 and c^{18}, but does worse on samples c^1, c^7 and c^{10}. Summary ROC scores are given in Table 6.4.

As the predictions of logistic regression and HSLDSdeep are different, it makes sense to analyse the opportunity of combining their predictions. Here, we use α-integration ([Amari, 2007]), which is flexible framework for expert aggregation. Using a single scalar parameter, α-integration covers a broad set of expert aggregation methods: max-pooling ($\alpha = -\infty$), conventional mixture ($\alpha = -1$), product of experts ($\alpha = 1$) or min-pooling ($\alpha = \infty$). The integrated model for sepsis detection, ALPHA, is obtained as:

$$p_{ALPHA}(z_t|\mathbf{y}_{1:t}) = \frac{1}{C} f_\alpha^{-1}\left\{\frac{f_\alpha\left[p_{HSLDSdeep}(z_t|\mathbf{y}_{1:t})\right] + f_\alpha\left[p_{LR}(z_t|\mathbf{y}_{1:t})\right]}{2}\right\}, \tag{6.6}$$

where

$$f_\alpha[p(z)] = \begin{cases} \frac{2}{1-\alpha} p(z)^{(1-\alpha)/2}, & \alpha \neq 1 \\ \log p(z), & \alpha = 1, \end{cases}$$

$$C = \sum_{z_t} f_\alpha^{-1}\left\{\frac{f_\alpha\left[p_{HSLDSdeep}(z_t|\mathbf{y}_{1:t})\right] + f_\alpha\left[p_{LR}(z_t|\mathbf{y}_{1:t})\right]}{2}\right\}$$

As directly optimising summary performance scores with respect to α is non-trivial, here we chose to compare values in the set $\{\pm\infty, \pm 100, \pm 10, \pm 5, \pm 2 \pm 1, \pm 0.5, 0\}$.

When aggregating HSLDSdeep and logistic regression predictions, we found that the best *second-by-second* results are obtained by max-pooling (see Figure 6.4b and Table 6.4). The fact that combining our generative framework with the ideas of Griffin et al. [2003] produces more accurate predictions about sepsis than either of these two approaches separately suggests that the methods are extracting somewhat distinct information about the presence of sepsis from the monitoring traces. Moreover, the aggregated model produced the best results when applying the same aggregation approach to the *episode-based* analysis. In this later experiment, the best performance has been obtained for $\alpha = -2$ (Table 6.4).

Table 6.5: Factor Inference Summaries Using 9-fold Cross-Validation

		Bradycardia	Desaturation	X-factor
FSLDS	AUC	0.85	0.81	0.63
	EER	0.21	0.28	0.40
HSLDSdeep	AUC	0.86	0.82	0.60
	EER	0.21	0.27	0.42
HSLDSukf	AUC	0.84	0.81	0.58
	EER	0.22	0.28	0.44
HSLDSkf	AUC	0.86	0.82	0.60
	EER	0.21	0.27	0.42

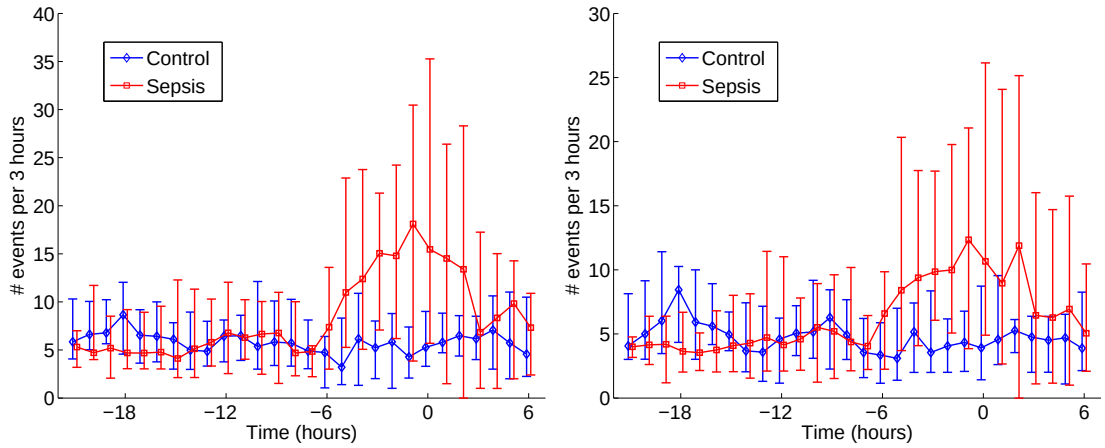
6.3.2 Physiological event posteriors

An important feature of the HSDLS framework is that apart from sepsis predictions the model can provide inferences concerning clinical events. Filtering distributions for these events can be obtained after marginalising the sepsis variable from the HSLDS posteriors. Here we discuss the two physiological events used for sepsis detection, bradycardia and desaturation, and the X-factor. As we have labelled data for the predicted factors, summary results computed by aggregating predictions obtained with 9-fold cross-validation are shown in Table 6.5.

Even though the FSLDS has been trained solely for inferring clinical events, there is little difference between its perfomance and the HSLDS models. The HSDLS models involve a more complex hidden discrete structure over which approximate inference is carried out, but at the same time are arguably better generative models for the data. HSLDSdeep and HSLDSkf delivered equal performance on inferring the hidden clinical factors. Both of them outperform HSLDSukf, mostly due to the unsupervised learning of factor transitions used to train the latter.

Note that bradycardia and X-factor inferences obtained using an FSLDS have been previously assessed in Quinn et al. [2009]. The bradycardia results reported here are very similar to that work, but X-factor predictions are worse. Results on oxygen desaturation have not been previously reported.

We also found it interesting to compare the true incidence of baby-generated physiological events against the inferred one. For this purpose we obtained inferred events by binarising factor posteriors (at threshold 0.5). Figure 6.7 shows a comparative visualisation of the time evolution of annotated and inferred bradycardias. The counts have been weighted in accordance to the amount of missing data in the analysed 3 hour periods. On both plots, there is a clear increase in the incidence of bradycardias in the hours before the sepsis diagnosis.



(a) Labelled events (b) Inferred Events

Figure 6.7: Median weighted number of true and inferred bradycardias separately computed for each patient group. The counts were computed hourly and summarize the preceding 3 hour period. Error bars mark first and third quartiles. The small offset between the two patient groups was used to improve readability.

6.4 Summary

In this chapter, we have proposed a framework for performing condition monitoring in situations when the factors that govern the data can be organised in a hierarchy. The structure of our model allows domain knowledge to be naturally incorporated. In addition, we have described a "deep learning" inspired training method.

The effectiveness of the HSLDS has been demonstrated for the difficult task of detecting the onset of sepsis in NICU patients. When compared against an AR-HMM model which heavily relies on expert annotations, we found the performance difference not to be statistically significant. Furthermore, we found out that by combining the HSLDS results with the method proposed by Griffin et al. [2003], we can outperform both of these approaches. It has also been empirically demonstrated that the HSLDS can produce clinical event posteriors as good as an FSLDS model exclusively trained for this task.

Chapter 7

Conclusions and further work

7.1 Summary of contributions

The main contributions of this thesis can be grouped as follows:

- In Chapter 5 we presented a model for the early detection of neonatal sepsis based on the distribution of clinical events observed in the monitoring traces. This involved:
  - The formulation of sepsis detection as learning and inference in an AR-HMM;
  - The presentation of an exact AR-HMM inference algorithm which marginalises over missing data;
  - Experimental results showing the effectiveness of the model on genuine NICU data, both in terms of second-by-second inferences and episodes of infection.
- The development of a condition monitoring model for inferring both sepsis and clinical events from the raw vital signs data was discussed in Chapter 6. This involved:
  - The formulation of condition monitoring as learning and inference in a Hierarchical Switching Linear Dynamical System (HSDLS);
  - The adaptation of parameterisation, learning and inference to the specifics of the vital signs monitoring task;
  - Experimental results showing the HSLDS performance is not statistically significantly different from the AR-HMM, despite the latter requiring "ground truth" annotations of the physiological factors. In addition, when combined with a discriminative approach, an improvement in performance was observed.
- Additional novel work on modelling data with slow linear trends has been presented in Appendix D. Using synthetic data, we demonstrated an SLDS with switch settings normal/trend, which could be used for the early detection of trends in NICU monitoring traces (e.g. pneumothorax).

7.2 Future work

In the remainder, we discuss several directions in which the work presented in this thesis can be extended. We first discuss methods for improving sepsis detection, continue with extensions on modelling the clinical regimes and then highlight some ways of obtaining better inference results.

The results discussed in Chapters 5 and 6 show that we can often detect sepsis well in advance of the time the positive blood sample is collected. However, summary performance scores could be improved, and there are various ways in which this could be achieved. First, in Chapter 5 we only modelled dependencies between consecutive observations of physiological events, but higher-order correlations can be easily incorporated in our framwork. Second, further knowledge about sepsis could be integrated by exploring the predictive power of patterns different from those used in this thesis. For instance, we could introduce a factor for monitoring the difference between core and peripheral temperatures [Lyon et al., 1997], and one for detecting periods of low measurement variability [McGregor et al., 2012]. Another idea would be to find reoccurring patterns in the X-factor annotations. This could be done either by expert analysis as in the case of mini-bradycardias (Section 5.4.2) or in an unsupervised setting. In the latter case we could either use ML learning via EM as in e.g. Ghahramani and Hinton [2000] or a Bayesian non-parametric approach such as in Fox et al. [2009]. It would also clearly be useful to see how our findings apply to a dataset containing samples from a larger number of babies, and also to extend the work to babies where the blood test result was mixed growth or skin commensal (see Sections 4.1 and 5.3). In addition, exploring data earlier than 24 hours before the positive test is worth considering.

The most prevalent monitoring regime in the patients analysed in thesis was physiological stability. Our models account for the fact that stability is baby-specific, and in previous work we showed how stability periods can be automatically identified [Williams and Stanculescu, 2011]. However, we have so far assumed that stable dynamics are stationary. This assumption is adequate for the 30 hour monitoring windows considered in this thesis, but is unlikely to apply for long term monitoring[1]. One solution would be to periodically re-run the automated method discussed in Williams and Stanculescu [2011]; e.g. every 24 hours. A more elegant approach would be to model parameter dynamics as a hidden stochastic process evolving on coarser time scale. This could be achieved by a binary discrete switch variable whose setting would decide between either copying the parameters at the previous time step or following some type of stochastic process, for instance a random walk. Other novel work could look at correlations between the learnt stability parameters and covariates such as age, gestation, birth weight and actual weight. Such knowledge could be later used to construct priors over the

[1] Note that the NICU stay of a VLBW baby could extend up to a few months.

space of stability parameters. It may also be worthwhile to explore jointly modelling the physiological channels in the FSLDS/HSLDS. Research in this latter direction should account for different sets of channels available for different patients, and should also revisit the modelling of factor interactions.

Future work could also look at developing the slow trend detection approach we described in Appendix D. First, the model is yet to be tested on genuine monitoring data; e.g. for inferring pneumothorax [McIntosh et al., 2000]. Second, the trend detection SLDS demonstrated in Section D.4 does not allow smooth transitions out of the trend regime. The solution to this issue could be related to changepoint modelling ideas (see Eckley et al. [2011] for an introduction).

It is also worth investigating whether inference results obtained with the models developed in this thesis could be improved. Fixed-lag smoothing promises more accurate estimates at the cost of delaying inference at time t in order to condition on the observed data up to time $t+\Delta$, where $\Delta \in \mathbb{N}^*$. In the AR-HMM, efficient fixed lag-smoothing could follow the recursive approach showed in Russell and Norvig [2003, §15.3] for HMMs. For the (H)SLDS, a very simple fixed-lag smoother can be obtained by not re-sampling particle components at times earlier than $t-\Delta$ [Kantas et al., 2009].

Other possible ideas of future work include adapting explicit duration modelling for SLDSes [Murphy, 2002], using spectral methods for learning vital signs dynamics [Overschee and Moor, 1996], and fitting non-linear dynamical models to NICU regimes [Murphy, 2012, §18.5].

Appendix A

SLDS filtering

Our algorithm of choice for SLDS filtering is the GPB2 method [Murphy, 1998, Bar-Shalom et al., 2001], and has been outlined in Section 2.4.1. Here, we first expand the ideas and then provide the full algorithm in Figure A.1.

We assume that at time $t-1$, the approximate one-step marginal continuous state filtering distribution is kept as a mixture of S Gaussians, $q(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$, with one component for each possible switch setting:

$$q(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) = \sum_{s_{t-1}} q(\mathbf{x}_{t-1}|s_{t-1}, \mathbf{y}_{1:t-1}) q(s_{t-1}|\mathbf{y}_{1:t-1}). \tag{A.1}$$

Then, GPB2 runs as follows:

1. **Continuous state update**. The first part of SLDS filtering is concerned with updating the continuous states. For each possible combination of settings $\{s_{t-1}, s_t\}$ Kalman updates are run. In the FLSDS and HSLDS models, during inference we allow at most one term of the cross product to change its setting, and thus at any time step only a limited number of s_{t-1} settings are considered for each value of s_t (also see Sections 2.4.2 and 6.1.2).

   As with the standard LDS, the first part of inference is the *prediction* step. We begin by computing a predicted hidden state:

   $$q(\mathbf{x}_t|s_{t-1}, s_t, \mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|s_t, \mathbf{x}_{t-1}) q(\mathbf{x}_{t-1}|s_{t-1}, \mathbf{y}_{1:t-1}) \mathrm{d}\mathbf{x}_{t-1}, \tag{A.2}$$

   and then determine the conditional likelihood of the current observation given the observation history, and the current and previous switch settings

   $$q(\mathbf{y}_t|s_{t-1}, s_t, \mathbf{y}_{1:t-1}) = \int p(\mathbf{y}_t|s_t, \mathbf{x}_t) q(\mathbf{x}_t|s_{t-1}, s_t, \mathbf{y}_{1:t-1}) \mathrm{d}\mathbf{x}_t. \tag{A.3}$$

   The subsequent *correction* step computes:

   $$q(\mathbf{x}_t|s_{t-1}, s_t, \mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|s_t, \mathbf{x}_t) q(\mathbf{x}_t|s_{t-1}, s_t, \mathbf{y}_{1:t-1}). \tag{A.4}$$

2. **Discrete state update**. It is useful to start by computing the pairwise discrete state filtering marginal:

$$q(s_{t-1}, s_t|\mathbf{y}_{1:t}) \propto q(\mathbf{y}_t|s_{t-1}, s_t, \mathbf{y}_{1:t-1})p(s_t|s_{t-1})q(s_{t-1}|\mathbf{y}_{1:t-1}). \tag{A.5}$$

Then, the one-step filtering marginal can be immediately obtained by marginalisation:

$$q(s_t|\mathbf{y}_{1:t}) = \sum_{s_{t-1}} q(s_{t-1}, s_t|\mathbf{y}_{1:t}) \tag{A.6}$$

3. **Collapsing**. At this point the continuous state filtering distribution is approximated by as a mixture of S^2 Gaussians:

$$q^*(\mathbf{x}_t|\mathbf{y}_{1:t}) = \sum_{s_{t-1}, s_t} q(\mathbf{x}_t|s_{t-1}, s_t, \mathbf{y}_{1:t})q(s_{t-1}, s_t|\mathbf{y}_{1:t}) \tag{A.7}$$

The mixture $q^*(\mathbf{x}_t|\mathbf{y}_{1:t})$ will projected onto a mixture of S Gaussians $q(\mathbf{x}_t|\mathbf{y}_{1:t})$ which shares the same form as eq. A.1:

$$q(\mathbf{x}_t|\mathbf{y}_{1:t}) = \sum_{s_t} q(\mathbf{x}_t|s_t, \mathbf{y}_{1:t})q(s_t|\mathbf{y}_{1:t}). \tag{A.8}$$

In order to understand the collapsing operation it is useful to rewrite eq. A.7 as:

$$q^*(\mathbf{x}_t|\mathbf{y}_{1:t}) = \sum_{s_t} \left[\sum_{s_{t-1}} q(\mathbf{x}_t|s_{t-1}, s_t, \mathbf{y}_{1:t})q(s_{t-1}|s_t, \mathbf{y}_{1:t}) \right] q(s_t|\mathbf{y}_{1:t}). \tag{A.9}$$

It is now easier to see that the components $q(\mathbf{x}_t|s_t, \mathbf{y}_{1:t})$ can be obtained by collapsing Gaussian mixtures of type $q^*(\mathbf{x}_t|s_t, \mathbf{y}_{1:t}) = \sum_{s_{t-1}} q(\mathbf{x}_t|s_{t-1}, s_t, \mathbf{y}_{1:t})q(s_{t-1}|s_t, \mathbf{y}_{1:t})$ onto single Gaussians. Here, we use moment matching for the collapsing operation.

The full algorithm is shown in Figure A.1, and uses the following notation:

$$\begin{aligned}
M_t(j) &= q(s_t = j|\mathbf{y}_{1:t}) \\
M_t(i,j) &= q(s_{t-1} = i, s_t = j|\mathbf{y}_{1:t}) \\
W_t(i,j) &= q(s_{t-1} = i|s_t = j, \mathbf{y}_{1:t}) \\
q(\mathbf{x}_t|s_t = j, \mathbf{y}_{1:t}) &= \mathcal{N}(\mathbf{x}_t; \hat{\mathbf{x}}_t^j, \hat{\mathbf{V}}_t^j) \\
q(\mathbf{x}_t|s_{t-1} = i, s_t = j, \mathbf{y}_{1:t}) &= \mathcal{N}(\mathbf{x}_t; \hat{\mathbf{x}}_t^{i(j)}, \hat{\mathbf{V}}_t^{i(j)}) \\
q(\mathbf{x}_t|s_{t-1} = i, s_t = j, \mathbf{y}_{1:t-1}) &= \mathcal{N}(\mathbf{x}_t; \hat{\mathbf{x}}_{t|t-1}^{i(j)}, \hat{\mathbf{V}}_{t|t-1}^{i(j)}) \\
L_t^{i(j)} = q(\mathbf{y}_t|s_{t-1} = i, s_t = j, \mathbf{y}_{1:t-1}) &= \mathcal{N}(\mathbf{y}_t; \hat{\mathbf{y}}_{t|t-1}^{i(j)}, \mathbf{S}_{t|t-1}^{i(j)})
\end{aligned}$$

$$\left\{M_t(j), \hat{\mathbf{x}}_t^j, \hat{\mathbf{V}}_t^j\right\} = GPB2\_forward\left(\left\{M_{t-1}(j), \hat{\mathbf{x}}_{t-1}^j, \hat{\mathbf{V}}_{t-1}^j\right\}\right)$$

1. **Continuous state update**

   - For each $s_{t-1} = i, s_t = j$, $(i = 1,,2\ldots,S, j = 1,,2\ldots,S)$ do

     (a) **Prediction**

$$\begin{aligned}
\hat{\mathbf{x}}_{t|t-1}^{i(j)} &= \mathbf{A}(j)\hat{\mathbf{x}}_{t-1}^{i} \\
\hat{\mathbf{V}}_{t|t-1}^{i(j)} &= \mathbf{A}(j)\hat{\mathbf{V}}_t^i\mathbf{A}^T(j) + \mathbf{Q}(j) \\
\hat{\mathbf{y}}_{t|t-1}^{i(j)} &= \mathbf{C}(j)\hat{\mathbf{x}}_{t|t-1}^{i(j)} \\
\mathbf{S}_{t|t-1}^{i(j)} &= \mathbf{C}(j)\hat{\mathbf{V}}_{t|t-1}^{i(j)}\mathbf{C}^T(j) + \mathbf{R}(j) \\
L_t^{i(j)} &= \mathcal{N}(\mathbf{y}_t; \hat{\mathbf{y}}_{t|t-1}^{i(j)}, \mathbf{S}_{t|t-1}^{i(j)})
\end{aligned}$$

     (b) **Correction**

$$\begin{aligned}
\mathbf{K}_t^{i(j)} &= \hat{\mathbf{V}}_{t|t-1}^{i(j)}\mathbf{C}^T(j)\left(\mathbf{C}(j)\hat{\mathbf{V}}_{t|t-1}^{i(j)}\mathbf{C}^T(j) + \mathbf{R}(j)\right)^{-1} \\
\hat{\mathbf{x}}_t^{i(j)} &= \hat{\mathbf{x}}_{t|t-1}^{i(j)} + \mathbf{K}_t^{i(j)}\left(\mathbf{y}_t - \mathbf{C}(j)\hat{\mathbf{x}}_{t|t-1}^{i(j)}\right) \\
\hat{\mathbf{V}}_t^{i(j)} &= \left(\mathbf{I} - \mathbf{K}_t^{i(j)}\mathbf{C}(j)\right)\hat{\mathbf{V}}_{t|t-1}^{i(j)}
\end{aligned}$$

2. **Discrete state update**

$$\begin{aligned}
M_t(i,j) &= \frac{L_t^{i(j)} p(s_t = j | s_{t-1} = i) M_{t-1}(i)}{\sum_{i'}\sum_{j'} L_t^{i'(j')} p(s_t = j' | s_{t-1} = i') M_{t-1}(i')} \\
M_t(j) &= \sum_i M_t(i,j)
\end{aligned}$$

3. **Collapsing by moment matching**

   - For each $s_t = j$, $(j = 1,,2\ldots,S)$ do

$$\begin{aligned}
W_t(i,j) &= \frac{M_t(i,j)}{\sum_{i'} M_t(i',j)} \\
\hat{\mathbf{x}}_t^j &= \sum_i W_t(i,j)\hat{\mathbf{x}}_t^{i(j)} \\
\hat{\mathbf{V}}_t^j &= \sum_i W_t(i,j)\left(\hat{\mathbf{V}}_t^{i(j)} + \hat{\mathbf{x}}_t^{i(j)}\left(\hat{\mathbf{x}}_t^{i(j)}\right)^T\right) - \hat{\mathbf{x}}_t^j\left(\hat{\mathbf{x}}_t^j\right)^T
\end{aligned}$$

Figure A.1: One iteration of the GPB2 algorithm for SLDS inference.

Appendix B

AR-HMM inference with missing data

One of the advantages of generative probabilistic models is that they can elegantly handle missing data by marginalisation. In this appendix, we provide a message passing algorithm for running exact inference in an AR-HMM despite the presence of missing data. We begin with the standard forward-backward equations for AR-HMM inference (Appendix B.1). The algorithm is then extended for marginalising over missing data (Appenidx B.2). We also provide a "scaled" version of the recursions, without which any practical application of the algorithm would be computationally difficult (Appendix B.3).

B.1 Inference without missing data

When there is no missing data, the messages in eq. 5.3 can be recursively computed as follows:

$$\begin{aligned}
\alpha(z_t) &= p(z_t, f_{1:t}) \\
&= \sum_{z_{t-1}} p(z_{t-1}, z_t, f_{1:t}) \\
&= p(f_t|z_t, f_{t-1}) \sum_{z_{t-1}} p(z_{t-1}, z_t, f_{1:t-1}) \\
&= p(f_t|z_t, f_{t-1}) \sum_{z_{t-1}} p(z_t|z_{t-1}) p(z_{t-1}, f_{1:t-1}) \\
&= p(f_t|z_t, f_{t-1}) \sum_{z_{t-1}} p(z_t|z_{t-1}) \alpha(z_{t-1}), \qquad \text{(B.1)} \\
\beta(z_t) &= p(f_{t+1:T}|z_t, f_t) \\
&= \sum_{z_{t+1}} p(z_{t+1}, f_{t+1:T}|z_t, f_t) \\
&= \sum_{z_{t+1}} p(z_{t+1}|z_t) p(f_{t+1:T}|z_{t+1}, f_t)
\end{aligned}$$

$$= \sum_{z_{t+1}} p(z_{t+1}|z_t)p(f_{t+1}|z_{t+1},f_t)p(f_{t+2:T}|z_{t+1},f_{t+1})$$
$$= \sum_{z_{t+1}} p(z_{t+1}|z_t)p(f_{t+1}|z_{t+1},f_t)\beta(z_{t+1}). \tag{B.2}$$

B.2 Inference in the presence of missing data

We now extend the recursions above to handle missing data.

Let $\mathcal{V}$ be the set of time steps for which we have observations. We would like to treat both $t \in \mathcal{V}$ and $t \notin \mathcal{V}$ in a unified framework. Thus, we introduce a function $V(f_t) : \{1,\ldots,L\} \rightarrow \{0,1\}$

$$V(f_t) = \begin{cases} \delta_{f_t f_t^v} & \text{if } t \in \mathcal{V} \\ 1 & \text{if } t \notin \mathcal{V}, \end{cases}$$

where δ_{ij} is the Kronecker delta. For any $t_0 < t < t_1$, the following holds:

$$p(z_t, f^v_{t_0:t_1}) = \sum_{f_t} p(z_t, f^v_{t_0:t-1}, f_t, f^v_{t+1:t_1})V(f_t). \tag{B.3}$$

If f_t is not observed, the summation in eq. B.3 represents the marginalization of the hidden variable f_t from $p(z_t, f^v_{t_0:t_1}, f_t)$. If f_t is observed, then the summation only selects the term $p(z_t, f^v_{t_0:t-1}, f^v_t, f^v_{t+1:t_1})$.

Applying eq. B.3 together with eq. 5.3 we get:

$$\begin{aligned} p(z_t, f^v_{1:T}) &= \sum_{f_t} p(z_t, f^v_{1:t-1}, f_t, f^v_{t+1:T})V(f_t) \\ &= \sum_{f_t} p(z_t, f^v_{1:t-1}, f_t)p(f^v_{t+1:T}|z_t, f_t, f^v_{1:t-1})V(f_t) \\ &= \sum_{f_t} p(z_t, f^v_{1:t-1}, f_t)p(f^v_{t+1:T}|z_t, f_t)V(f_t) \\ &= \sum_{f_t} \alpha(z_t, f_t)\beta(z_t, f_t), \end{aligned} \tag{B.4}$$

where we have defined the messages:

$$\alpha(z_t, f_t) \triangleq p(z_t, f^v_{1:t-1}, f_t)V(f_t),$$
$$\beta(z_t, f_t) \triangleq p(f^v_{t+1:T}|z_t, f_t)V(f_t),$$

and used the fact that $V^2(f_t) = V(f_t)$. Similarly to eqs. B.1 and B.2, the following recursions

can be written:

$$\begin{aligned}
\alpha(z_t, f_t) &= V(f_t) p(z_t, f^v_{1:t-1}, f_t) \\
&= V(f_t) \sum_{f_{t-1}} p(z_t, f^v_{1:t-2}, f_{t-1}, f_t) V(f_{t-1}) \\
&= V(f_t) \sum_{f_{t-1}} \sum_{z_{t-1}} p(z_{t-1}, z_t, f^v_{1:t-2}, f_{t-1}, f_t) V(f_{t-1}) \\
&= V(f_t) \sum_{f_{t-1}} p(f_t | z_t, f_{t-1}) \sum_{z_{t-1}} p(z_{t-1}, z_t, f^v_{1:t-2}, f_{t-1}) V(f_{t-1}) \\
&= V(f_t) \sum_{f_{t-1}} p(f_t | z_t, f_{t-1}) \sum_{z_{t-1}} p(z_t | z_{t-1}) \alpha(z_{t-1}, f_{t-1}),
\end{aligned} \tag{B.5}$$

$$\begin{aligned}
\beta(z_t, f_t) &= V(f_t) p(f^v_{t+1:T} | z_t, f_t) \\
&= V(f_t) \sum_{f_{t+1}} p(f_{t+1}, f^v_{t+2:T} | z_t, f_t) V(f_{t+1}) \\
&= V(f_t) \sum_{f_{t+1}} \sum_{z_{t+1}} p(z_{t+1}, f_{t+1}, f^v_{t+2:T} | z_t, f_t) V(f_{t+1}) \\
&= V(f_t) \sum_{f_{t+1}} \sum_{z_{t+1}} p(z_{t+1} | z_t) p(f_{t+1}, f^v_{t+2:T} | z_{t+1}, f_t) V(f_{t+1}) \\
&= V(f_t) \sum_{z_{t+1}} p(z_{t+1} | z_t) \sum_{f_{t+1}} p(f_{t+1} | z_{t+1}, f_t) \beta(z_{t+1}, f_{t+1}).
\end{aligned} \tag{B.6}$$

When training an AR-HMM with missing data via EM, the following quantities are needed in the M-step:

$$\begin{aligned}
p(z_t, z_{t-1}, f^v_{1:T}) &= \sum_{f_t, f_{t-1}} p(z_t, z_{t-1}, f^v_{1:t-2}, f_{t-1}, f_t, f^v_{t+1:T}) V(f_t) V(f_{t-1}) \\
&= \sum_{f_t, f_{t-1}} \alpha(z_{t-1}, f_{t-1}) p(z_t | z_{t-1}) p(f_t | z_t, f_{t-1}) \beta(z_t, f_t),
\end{aligned}$$

$$\begin{aligned}
p(z_t, f^v_{1:t-2}, f_{t-1}, f_t, f^v_{t+1:T}) V(f_t) V(f_{t-1}) &= \sum_{z_{t-1}} p(z_t, z_{t-1}, f^v_{1:t-2}, f_{t-1}, f_t, f^v_{t+1:T}) V(f_t) V(f_{t-1}) \\
&= \sum_{z_{t-1}} \alpha(z_{t-1}, f_{t-1}) p(z_t | z_{t-1}) p(f_t | z_t, f_{t-1}) \beta(z_t, f_t).
\end{aligned}$$

Note that the above message passing routine has been designed for discrete observations only. While this suffices for our baby monitoring application, one can imagine marginalising over missing data when observations are continuous. For example, if the emission distribution were Gaussian, the forward messages would become mixtures of Gaussians with a number of components exponential in the length of the missing data sequences. In this case moment matching could be applied in similar way to its application to SLDS inference (see Appendix A).

B.3 Scaling

When applying forward-backward algorithms as described above to longer sequences, numerical underflow could be a serious problem [Bishop, 2007, §13.2.4]. The solution is to swap the α and β messages with their normalised versions and rewrite the forward and backward recursions. This is sometimes referred to as "scaling". In the remainder, we provide a derivation of the "scaled" forward and backward messages for AR-HMM inference in the presence of missing data.

For the forward part we have:

$$\tilde{\alpha}(z_t, f_t) \triangleq \frac{\alpha(z_t, f_t)}{p(f^v_{1:t})} = p(z_t, f_t | f^v_{1:t}). \tag{B.7}$$

Note that in the case of scaled messages conditioning implicitly substitutes the role of $V(f_t)$. We further define:

$$c_t \triangleq \sum_{f_t} p(f_t | f^v_{1:t-1}) V(f_t).$$

If $t \notin \mathcal{V}$, then $c_t = 1$, and if $t \in \mathcal{V}$, then $c_t = p(f^v_t | f^v_{1:t-1})$. In addition:

$$p(f^v_{1:t}) = \prod_{s=1}^{t} c_s.$$

In order to find the recursion for $\tilde{\alpha}(z_t, f_t)$ we substitute the relationship between α and $\tilde{\alpha}$ from eq. B.7 into eq. B.5 to obtain

$$\tilde{\alpha}(z_t, f_t) = \frac{V(f_t)}{p(f^v_{1:t})} \sum_{f_{t-1}} p(f_t | z_t, f_{t-1}) p(f^v_{1:t-1}) \sum_{z_{t-1}} p(z_t | z_{t-1}) \tilde{\alpha}(z_{t-1}, f_{t-1}).$$

However, the term $p(f^v_{1:t-1})$ can be pulled left through the sum over f_{t-1} on the RHS: If $t-1 \notin \mathcal{V}$, then $p(f^v_{1:t-1}) = p(f^v_{1:t-2})$ and it is clear this term can move outside. If $t-1 \in \mathcal{V}$, then the sum over f_{t-1} contains only one non-zero term and again $p(f^v_{1:t-1})$ can be moved left. We can now recognise $c_t = p(f^v_{1:t}) / p(f^v_{1:t-1})$ to write the scaled forward recursion:

$$\tilde{\alpha}(z_t, f_t) = \frac{V(f_t)}{c_t} \sum_{f_{t-1}} p(f_t | z_t, f_{t-1}) \sum_{z_{t-1}} p(z_t | z_{t-1}) \tilde{\alpha}(z_{t-1}, f_{t-1}). \tag{B.8}$$

The scaled version of the backward message is:

$$\tilde{\beta}(z_t, f_t) = \frac{\beta(z_t, f_t)}{p(f^v_{t+1:T} | f^v_{1:t})}. \tag{B.9}$$

Note that:

$$p(f^v_{t+1:T} | f^v_{1:t}) = \prod_{s=t+1}^{T} c_s.$$

Substituting eq. B.9 into eq. B.6 we obtain:

$$\tilde{\beta}(z_t, f_t) = \frac{V(f_t)}{p(f^v_{t+1:T} | f^v_{1:t})} \sum_{z_{t+1}} p(z_{t+1} | z_t) \sum_{f_{t+1}} p(f^v_{t+2:T} | f^v_{1:t+1}) p(f_{t+1} | z_{t+1}, f_t) \tilde{\beta}(z_{t+1}, f_{t+1}).$$

Similar to the $\tilde{\alpha}$ recursion above, the $p(f^v_{t+2:T}|f^v_{1:t+1})$ can be pulled outside both sums on the RHS. First consider $t+1 \notin \mathcal{V}$, then $p(f^v_{t+2:T}|f^v_{1:t+1}) = p(f^v_{t+2:T}|f^v_{1:t})$ and the terms can be pulled through. If $t+1 \in \mathcal{V}$, then the sum over f_{t+1} contains only one non-zero term and again $p(f^v_{t+2:T}|f^v_{1:t+1})$ can be moved left. We recognize $c_{t+1} = \frac{p(f^v_{t+1:T}|f^v_{1:t})}{p(f^v_{t+2:T}|f^v_{1:t+1})}$ and obtain the backward recursion:

$$\tilde{\beta}(z_t, f_t) = \frac{V(f_t)}{c_{t+1}} \sum_{z_{t+1}} p(z_{t+1}|z_t) \sum_{f_{t+1}} p(f_{t+1}|z_{t+1}, f_t)\tilde{\beta}(z_{t+1}, f_{t+1}). \tag{B.10}$$

The smoothing distribution can be expressed in terms of the scaled messages as follows:

$$p(z_t|f^v_{1:T}) = \frac{1}{p(f^v_{1:T})} \sum_{f_t} \alpha(z_t, f_t)\beta(z_t, f_t) = \sum_{f_t} \tilde{\alpha}(z_t, f_t)\tilde{\beta}(z_t, f_t). \tag{B.11}$$

Finally, the following identities are useful for using the scaled messages within the EM optimization routine.

$$p(z_t, z_{t-1}|f^v_{1:T}) = c_t \sum_{f_t, f_{t-1}} \tilde{\alpha}(z_{t-1}, f_{t-1}) p(z_t|z_{t-1}) p(f_t|z_t, f_{t-1}) \tilde{\beta}(z_t, f_t)$$

$$p(z_t, f_t, f_{t-1}|f^v_{1:T}) = c_t \sum_{z_{t-1}} \tilde{\alpha}(z_{t-1}, f_{t-1}) p(z_t|z_{t-1}) p(f_t|z_t, f_{t-1}) \tilde{\beta}(z_t, f_t).$$

Appendix C

EM derivations for dynamical models

In this appendix we provide detail on the application of the EM algorithm for finding ML parameters in several dynamical models discussed in the previous chapters. We first show how hidden $ARMA(p,q)$ dynamics can be learnt from noisy observations. Then, we discuss EM for the unsupervised learning of switching linear dynamical systems.

For simplicity, in the following we consider the training set contains only one sequence, but in practice the results can be easily extended to training sets consisting of several independent sequences.

C.1 ARMA models in SSM from

The application of EM for finding ML parameters for LDSes has been previously discussed in Ghahramani and Hinton [1996]. Here, we adapt their method for the special case when the hidden dynamics are modelled as an $ARMA(p,q)$ process. The state-space representation of *ARMA* processes we employ in the following is the one described in Section 2.3.2.

The E-step can be performed exactly using the standard Kalman smoothing recursions. The subsequent M-step is equivalent to maximizing the expected complete data log-likelihood

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \mathbb{E}_{\mathbf{x}|y,\boldsymbol{\theta}^{old}} \log p(\mathbf{x}, y), \tag{C.1}$$

where $\mathbf{x} = \mathbf{x}_{1:T}$, $y = y_{1:T}$ and $\boldsymbol{\theta} = \{\mathbf{m}_0, \mathbf{V}_0, \mathbf{A}, \mathbf{Q}, \mathbf{C}, \mathbf{R}\}$. Note that the expectation is taken with respect to the posterior computed in the preceding E-step with the old parameter settings.

Making use of the factorisation of the LDS's joint distribution $p(\mathbf{x}, \mathbf{y})$, we see that certain terms in eq. C.1 depend only on a subset of the parameters. Thus, we separately discuss:

- updating the initial conditions $\mathbf{m}_0$ and $\mathbf{V}_0$. The only terms in $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ depending on $\mathbf{A}$ and $\mathbf{Q}$ are:

$$-\frac{1}{2}\ln|\mathbf{V}_0| - \mathbb{E}_{\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}^{old}}\left[\frac{1}{2}(\mathbf{x}_1 - \mathbf{m}_0)^T \mathbf{V}_0^{-1}(\mathbf{x}_1 - \mathbf{m}_0)\right] \tag{C.2}$$

Using the given parameterisation this can be rewritten as:

$$-\frac{r}{2}\ln\sigma_0^2 - \frac{1}{2\sigma_0^2}\mathbb{E}_{\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}^{old}}\left[\sum_{d=1}^{r}(\mathbf{x}_{1,d}-\mu_0)^2\right], \tag{C.3}$$

where $\mathbf{x}_{1,d}$ is the d-th component of $\mathbf{x}_1$. Taking derivatives with respect to μ_0 we get:

$$\tilde{\mu}_0 = \frac{1}{r}\sum_{d=1}^{r}\mathbb{E}[\mathbf{x}_{1,d}] \tag{C.4}$$

Then, we obtain:

$$\tilde{\sigma}_0^2 = \frac{1}{r}\sum_{d=1}^{r}\mathbb{E}[\mathbf{x}_{1,d}^2] - \tilde{\mu}_0^2 \tag{C.5}$$

- updating the dynamics parameters $\mathbf{A}$ and $\mathbf{Q}$. The only terms in $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$ depending on $\mathbf{A}$ and $\mathbf{Q}$ are:

$$-\mathbb{E}_{\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}^{old}}\left[\frac{1}{2}\sum_{t=2}^{T}(\mathbf{x}_t-\mathbf{A}\mathbf{x}_{t-1})^T\mathbf{Q}^{-1}(\mathbf{x}_t-\mathbf{A}\mathbf{x}_{t-1})\right] - \frac{T-1}{2}\ln|\mathbf{Q}|^1. \tag{C.6}$$

Using the given parameterisation this can be rewritten as:

$$-\mathbb{E}_{\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}^{old}}\left[\frac{1}{2\sigma_\varepsilon^2}\sum_{t=2}^{T}(X_t-\mathbf{a}^T\mathbf{x}_{t-1,1:p})^2\right] - \frac{T-1}{2}\ln\sigma_\varepsilon^2, \tag{C.7}$$

where $\mathbf{a}^T=[\phi_1,\phi_2,\ldots,\phi_p]$ and $\mathbf{x}_{t-1,1:p}$ are the first p dimensions of $\mathbf{x}_{t-1}$. Taking derivatives with respect to $\mathbf{a}^T$, we get:

$$\tilde{\mathbf{a}}^T = \left(\sum_{t=2}^{T}\mathbb{E}\left[X_t\mathbf{x}_{t-1,1:p}^T\right]\right)\left(\sum_{t=2}^{T}\mathbb{E}\left[\mathbf{x}_{t-1,1:p}\mathbf{x}_{t-1,1:p}^T\right]\right)^{-1}. \tag{C.8}$$

Then, we obtain:

$$\tilde{\sigma}_\varepsilon^2 = \frac{1}{T-1}\sum_{t=2}^{T}\left\{\mathbb{E}\left[X_t^2\right] - 2\mathbb{E}\left[X_t\mathbf{x}_{t-1,1:p}^T\right]\tilde{\mathbf{a}} + \tilde{\mathbf{a}}^T\mathbb{E}\left[\mathbf{x}_{t-1,1:p}\mathbf{x}_{t-1,1:p}^T\right]\tilde{\mathbf{a}}\right\} \tag{C.9}$$

- updating the observation parameters $\mathbf{C}$ and $\mathbf{R}$. The only terms in $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$ depending on $\mathbf{C}$ and $\mathbf{R}$ are:

$$-\mathbb{E}_{\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}^{old}}\left[\frac{1}{2}\sum_{t=1}^{T}(y_t-\mathbf{C}\mathbf{x}_t)^T\mathbf{R}^{-1}(y_t-\mathbf{C}\mathbf{x}_t)\right] - \frac{T}{2}\ln|\mathbf{R}|. \tag{C.10}$$

Again, using the given parameterisation the latter can be rewritten as:

$$-\mathbb{E}_{\mathbf{X}|\mathbf{Y},\boldsymbol{\theta}^{old}}\left[\frac{1}{2\sigma_\omega^2}\sum_{t=1}^{T}(y_t-x_t-\mathbf{c}^T\mathbf{x}_{t,2:q+1})^2\right] - \frac{T}{2}\ln\sigma_\omega^2, \tag{C.11}$$

[1] To ensure the invertibility of $\mathbf{Q}$ we assume $\mathbf{Q}\leftarrow\mathbf{Q}+\delta\mathbf{I}_r$

where $\mathbf{c}^T = [\theta_1, \theta_2, \ldots, \theta_q]$. Taking derivatives with respect to $\mathbf{c}^T$, we get:

$$\tilde{\mathbf{c}}^T = \left(\sum_{t=1}^{T} \left(y_t \mathbb{E}[\mathbf{x}_{t,2:q+1}^T] - \mathbb{E}\left[x_t \mathbf{x}_{t,2:q+1}^T \right] \right) \right) \left(\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{x}_{t,2:q+1} \mathbf{x}_{t,2:q+1}^T \right] \right)^{-1}. \tag{C.12}$$

Since we did not add any special constraints on it, the update for the (scalar) $\mathbf{R} = \sigma_\omega^2$ is the standard one:

$$\tilde{\sigma}_\omega^2 = \frac{1}{T} \sum_{t=1}^{T} \left\{ y_t^2 - 2\tilde{\mathbf{C}} \mathbb{E}[\mathbf{x}_t] y_t + \tilde{\mathbf{C}} \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^T] \tilde{\mathbf{C}}^T \right\}. \tag{C.13}$$

C.2 Switching Linear Dynamical System

When learning an SLDS with switch settings unknown, the M-step is equivalent to optimizing the expected complete data log likelihood

$$Q(\theta, \theta^{old}) = \mathbb{E}_{s,\mathbf{x}|\mathbf{y},\boldsymbol{\theta}^{old}} \log p(s, \mathbf{x}, \mathbf{y}). \tag{C.14}$$

Using the factorisation of the joint distribution $p(s, \mathbf{x}, \mathbf{y})$, the above can be expanded as follows:

$$\begin{aligned} Q(\theta, \theta^{old}) =& \mathbb{E}_{s,\mathbf{x}|\mathbf{y},\boldsymbol{\theta}^{old}} \log p(s_{1:T}) + \mathbb{E}_{s,\mathbf{x}|\mathbf{y},\boldsymbol{\theta}^{old}} \log p(\mathbf{x}_{1:T}|s_{1:T}) + \mathbb{E}_{s,\mathbf{x}|\mathbf{y},\boldsymbol{\theta}^{old}} \log p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, s_{1:T}) \\ =& \mathbb{E}_{s_1|\mathbf{y},\boldsymbol{\theta}^{old}} \log p(s_1) + \sum_{t=2}^{T} \mathbb{E}_{s_t,s_{t-1}|\mathbf{y},\boldsymbol{\theta}^{old}} \log p(s_t|s_{t-1}) + \\ & \mathbb{E}_{s_1,\mathbf{x}_1|\mathbf{y},\boldsymbol{\theta}^{old}} \log p(\mathbf{x}_1|s_1) + \sum_{t=2}^{T} \mathbb{E}_{\mathbf{x}_t,\mathbf{x}_{t-1},s_t,|\mathbf{y},\boldsymbol{\theta}^{old}} \log p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t) + \\ & \sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_t,s_t|\mathbf{y},\boldsymbol{\theta}^{old}} \log p(\mathbf{y}_t|\mathbf{x}_t, s_t) \end{aligned} \tag{C.15}$$

If we employ the notation $\langle \cdot \rangle_{s,\mathbf{x}|\mathbf{y}} = \mathbb{E}_{s,\mathbf{x}|\mathbf{y},\boldsymbol{\theta}^{old}}[\cdot]$, then we can further refine eq. C.15:

$$
\begin{aligned}
Q(\theta,\theta^{old}) = {} & \langle \log p(s_1)\rangle_{s_1|\mathbf{y}} + \sum_{t=2}^{T}\langle \log p(s_t|s_{t-1})\rangle_{s_t,s_{t-1}|\mathbf{y}} - \frac{1}{2}(d_x+d_y)T\log(2\pi) - \\
& \frac{1}{2}\langle \log\det \mathbf{V}_0(s_1)\rangle_{s_1|\mathbf{y}} - \frac{1}{2}\left\langle \mathrm{tr}\left(\mathbf{V}_0^{-1}(s_t)\langle \mathbf{x}_1\mathbf{x}_1^T\rangle_{\mathbf{x}_1|s_1,\mathbf{y}}\right)\right\rangle_{s_1|\mathbf{y}} + \\
& \left\langle \mathbf{m}_0^T(s_1)\mathbf{V}_0^{-1}(s_t)\langle \mathbf{x}_1\rangle_{\mathbf{x}_1|s_1,\mathbf{y}}\right\rangle_{s_1|\mathbf{y}} - \frac{1}{2}\langle \mathbf{m}_0^T(s_1)\mathbf{V}_0^{-1}(s_t)\mathbf{m}_0(s_1)\rangle_{s_1|\mathbf{y}} - \\
& \frac{1}{2}\sum_{t=2}^{T}\langle \log\det \mathbf{Q}(s_t)\rangle_{s_t|\mathbf{y}} - \frac{1}{2}\sum_{t=2}^{T}\left\langle \mathrm{tr}\left(\mathbf{Q}^{-1}(s_t)\langle \mathbf{x}_t\mathbf{x}_t^T\rangle_{\mathbf{x}_t|s_t,\mathbf{y}}\right)\right\rangle_{s_t|\mathbf{y}} + \\
& \sum_{t=2}^{T}\left\langle \mathrm{tr}\left(\mathbf{Q}^{-1}(s_t)\mathbf{A}(s_t)\langle \mathbf{x}_{t-1}\mathbf{x}_t{}^T\rangle_{\mathbf{x}_t,\mathbf{x}_{t-1}|s_t,\mathbf{y}}\right)\right\rangle_{s_t|\mathbf{y}} - \\
& \frac{1}{2}\sum_{t=2}^{T}\left\langle \mathrm{tr}\left(\mathbf{A}^T(s_t)\mathbf{Q}^{-1}(s_t)\mathbf{A}(s_t)\langle \mathbf{x}_{t-1}\mathbf{x}_{t-1}^T\rangle_{\mathbf{x}_{t-1}|s_t,\mathbf{y}}\right)\right\rangle_{s_t|\mathbf{y}} - \\
& \frac{1}{2}\sum_{t=1}^{T}\langle \log\det \mathbf{R}(s_t)\rangle_{s_t|\mathbf{y}} - \frac{1}{2}\sum_{t=1}^{T}\langle \mathbf{y}_t^T\mathbf{R}^{-1}(s_t)\mathbf{y}_t\rangle_{s_t|\mathbf{y}} + \sum_{t=1}^{T}\left\langle \mathbf{y}_t^T\mathbf{R}^{-1}(s_t)\mathbf{C}(s_t)\langle \mathbf{x}_t\rangle_{\mathbf{x}_t|s_t,\mathbf{y}}\right\rangle_{s_t|\mathbf{y}} - \\
& \frac{1}{2}\sum_{t=1}^{T}\left\langle \mathrm{tr}\left(\mathbf{C}^T(s_t)\mathbf{R}^{-1}(s_t)\mathbf{C}(s_t)\langle \mathbf{x}_t\mathbf{x}_t^T\rangle_{\mathbf{x}_t|s_t,\mathbf{y}}\right)\right\rangle_{s_t|\mathbf{y}}, \qquad \text{(C.16)}
\end{aligned}
$$

where we have also used the identity $\langle \mathbf{a}^T\mathbf{M}\mathbf{b}\rangle_{\mathbf{a},\mathbf{b}} = \mathrm{tr}\left(\mathbf{M}\langle \mathbf{b}\mathbf{a}^T\rangle_{\mathbf{a},\mathbf{b}}\right)$.

In terms of computing the expectations involved eq. C.16 (i.e. the E-step), exact SLDS inference is computationally intractable. One might use approximate SLDS smoothing algorithms such as the Expectation Correction method of Barber and Mesot [2006]. Alternatively, smoothing can be approximated by SLDS filtering. See Appendix A for a Gaussiam sum approximation to filtering.

In general, we can derive M-step updates for any SLDS parameter by setting partial derivatives of eq. C.1 to zero. In the baby monitoring application we have only used the update to the state transition matrix (eq. 6.5). For a complete set of M-step updates see Murphy [1998] or Zoeter and Heskes [2003].

Appendix D

A SLDS for trend detection

In many monitoring scenarios, the presence of a trend in the mean of the data is highly descriptive of the system's condition. Frequently, these trends build up slowly compared to the sampling rate of the data, making their detection hard at high resolution. Moreover, they might be hidden by the variance of the process's dynamics and by noisy observations. We illustrate these problems in Figure D.1a, where we show a sample from our proposed model for time series with linear trend and a detail from this sample.

In the context of neonatal condition monitoring, one useful application for trend detection is predicting pneumothorax. Despite being rare, pneumothorax [McIntosh et al., 2000] is a life threatening event consisting in a build-up of air outside the lung. With respect to the monitoring data, one will see drops in oxygen saturation (SO) and partial oxigen pressure ($TcPO_2$) together with an increase in the partial carbon dioxide pressure ($TcPCO_2$) An example is provided in Figure D.1b. Note that pneumothorax is a slow developing event when compared to the sampling rate of 1 Hz currently used in the NICU.

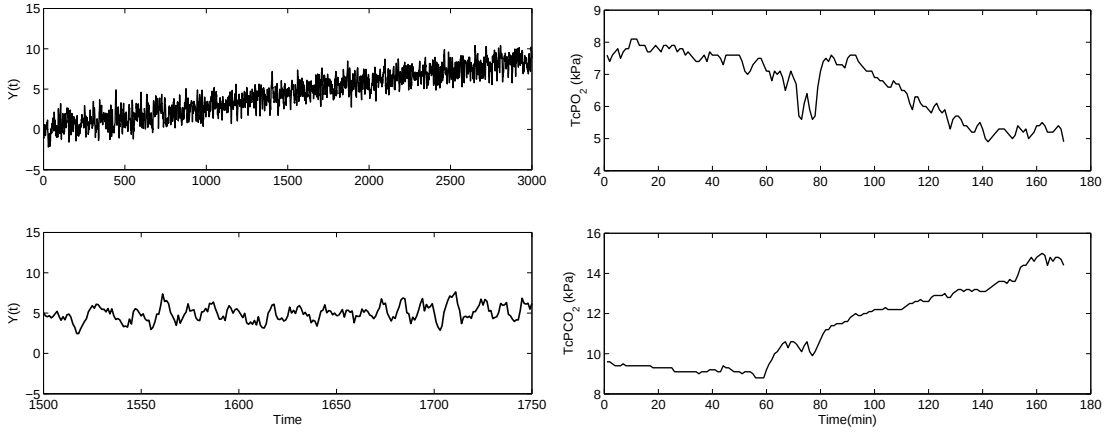
Our approach is to build a generative model for sequences with subtle linear trends in the mean. An important requirement is that the model should be closely related to the one used in the absence of trend. The proposed summation model (see Section D.3) relies on the following two assumptions:

1. In the absence of a trend, the signal is well modelled as an $ARMA(p,q)$ process.

2. The trend is a noisy linear drift on the mean of the process.

The summation model is equivalent to a certain parameterisation of the state-space model discussed in Section 2.3.2. Using this representation we construct a Switching LDS (see Section 2.4.1) able to discriminate periods of trend from periods where the trend is not present.

Starting from a naïve construction which fails to produce the desired linear drift behaviour, Section D.1 highlights some of the difficulties of adding trend to an *ARMA* signal. We then present the local linear trend model in Section D.2. This allows us to introduce the proposed



(a) Synthetic data

(b) Pneumothorax example

Figure D.1: In panel (a) we show a sample from the summation model proposed in Section D.3 (upper plot) and a detail from this sample (lower plot). A pneumothorax example is shown in panel (b). The onset of pneumothorax is at 60 minutes. The example is free from artifact, which is relatively unusual.

summation model in Section D.3, where we also give its state-space representation. Experimental results on synthetic data are presented and discussed in Section D.4. Section D.5 proves that the discussed models are in fact special cases of *ARIMA* models. We conclude and make further work suggestions in Section D.6.

D.1 A naïve construction

ARMA models (reviewed in Section 2.3.1) a are popular tool for modelling stationary time series. In the following, we present the difficulties of incorporating trend into *ARMA* models, by studying a naïve extension which fails to produce the desired behaviour.

Consider the following causal *ARMA* model with drift:

$$\phi(B)Z_t = \theta(B)\varepsilon_t + d_t, \qquad \varepsilon_t \sim WN(0,\sigma_\varepsilon^2) \tag{D.1}$$

where $d_t = d, \forall t \in \mathbb{Z}, d \in \mathbb{R}^*$ and we employed the notation introduced in Section 2.3.1. The intuition is that adding the constant d at each time step would produce a trend with whose slope is precisely d. Also, d_t can be interpreted as a control signal.

Since d_t is constant we can rewrite it as:

$$d_t = \frac{1}{\phi(1)}\phi(B)d_t. \tag{D.2}$$

Dividing by $\phi(1)$ is possible because the fact that 1 is not a root of $\phi(z)$ follows immediately from the causality of the $ARMA(p,q)$ process (Theorem 2.3.1). Replacing this into eq. D.1,

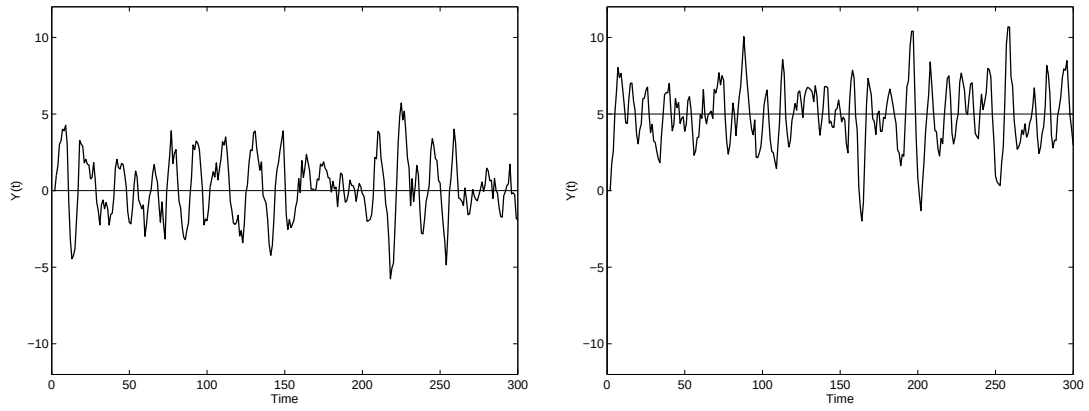


Figure D.2: (a) Sample from the $ARMA(2,1)$ model $(1-1.6B+0.8B^2)Z_t = (1-0.5B)\varepsilon_t$, $\varepsilon_t \sim WN(0,1)$. The flat line represents the mean of the process. (b) Sample from the model described by eq. D.1 using the same underlying $ARMA(2,1)$ as in panel (a) and $d = 1$. The mean of the process, $\mu = d/\phi(1) = 1/(1-1.6+0.8) = 5$, is shown as the flat line on the plot.

rearranging and substituting d for d_t we get:

$$\phi(B)(Z_t - \frac{d}{\phi(1)}) = \theta(B)\varepsilon_t. \tag{D.3}$$

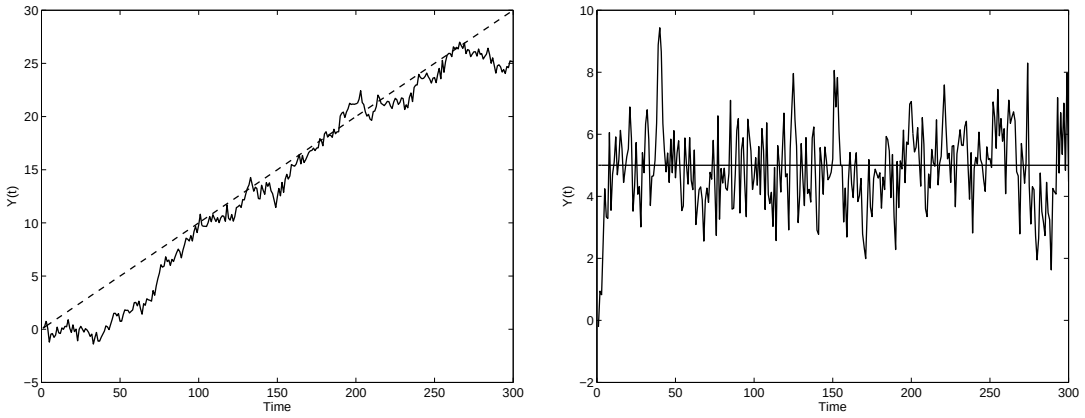
From this new representation, it is easy to see that eq. D.1 is an *ARMA* model with the same dynamics as the original, but with mean $\mu = d/\phi(1)$.

Figure D.2a shows a sample from a $ARMA(2,1)$ model with zero mean. For comparison, Figure D.2b shows a sample from the model described by eq. D.1 and using the same underlying $ARMA(2,1)$ model as in Figure D.2a. This also confirms that our naïve construction cannot generate trends.

D.2 The local linear trend

In this section we first introduce the local linear trend [Harvey, 1991, §2.3], a simple generative model for trends. Then, we prove that a straightforward generalization enabling the model to capture the *ARMA* dynamics of the non-trend data is unsatisfactory.

We observe $\{Y_t\}$, a noisy version of a signal $\{Z_t\}$. Assume that the total increment of the signal, $\{Z_t\}$, is given by the sum of an increment process and some Gaussian white noise, $\{\varepsilon_t\}$. The increment itself is the sum of a constant d and a Gaussian random walk, $\{D_t\}$. The interpretation is that the random walk process allows for small variations in the slope of the signal, while the Gaussian white noise $\{\varepsilon_t\}$ allows signal variability at each time step. This can



(a) (b)

Figure D.3: (a) Sample from the local linear trend with parameters $\sigma^2_\omega = 0.1$, $\sigma^2_\varepsilon = 0.1$, $\sigma^2_\eta = 10^{-6}$ and $d = 0.1$. The dashed line is the mean of the process, $\mathbb{E}[Y_t] = td$. (b) Sample from the local linear trend based generalisation defined by eqs. D.12- D.14. The underlying $ARMA$ is the same model as used in Figure D.2a, and the additional parameters are $\sigma^2_\omega = 0.1$, $\sigma^2_\eta = 10^{-6}$ and $d = 1$. Again, the mean of the process, $\mu = d/\phi(1) = 1/(1-1.6+0.8) = 5$, is shown as the flat line on the plot.

be summarized as:

$$Y_t = Z_t + \omega_t, \qquad \omega_t \sim WN(0, \sigma^2_\omega) \tag{D.4}$$
$$Z_t = Z_{t-1} + D_{t-1} + d + \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma^2_\varepsilon) \tag{D.5}$$
$$D_t = D_{t-1} + \eta_t, \qquad \eta_t \sim WN(0, \sigma^2_\eta) \tag{D.6}$$
$$D_0 = 0 \tag{D.7}$$

Alternatively, $\{Y_t\}$ can be expressed as:

$$Y_t = Z_t + \omega_t, \qquad \omega_t \sim WN(0, \sigma^2_\omega) \tag{D.8}$$
$$Z_t = Z_{t-1} + D_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma^2_\varepsilon) \tag{D.9}$$
$$D_t = D_{t-1} + \eta_t, \qquad \eta_t \sim WN(0, \sigma^2_\eta) \tag{D.10}$$
$$D_0 = d. \tag{D.11}$$

In the second representation we constrain the Gaussian random walk to have mean value $\bar{D}_t = d$ by making the initialization $D_0 = d$. From a fully Bayesian perspective we can have $D_0 \sim \mathcal{N}(D, \sigma^2_d)$. See Figure D.3a for a sample from the local linear trend model.

A special case of the local linear trend is the *integrated random walk* [Harvey, 2006, §2.3], where we set $\sigma^2_\varepsilon = 0$.

Another naïve construction

We now try to generalize the the local linear trend in order to model trends occurring on top of an underlying causal $ARMA(p,q)$ process. Assume that in the absence of trend the signal was well modelled by the system of eqs. 2.31 and 2.32. Then, a model for the drifting version of this process can be obtained by modifying eq. D.5 to inculde an $ARMA$ component

$$Y_t = Z_t + \omega_t, \qquad \omega_t \sim WN(0, \sigma_\omega^2) \tag{D.12}$$

$$\phi(B)Z_t = \theta(B)\varepsilon_t + D_{t-1} + d, \qquad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \tag{D.13}$$

$$D_t = D_{t-1} + \eta_t, \qquad \eta_t \sim WN(0, \sigma_\eta^2) \tag{D.14}$$

We claim that the causality of the original $ARMA(p,q)$ process results into the mean of $\{Y_t\}$ begin a constant. An easy way to see this is by taking the expectation of eq. D.13. Since $\bar{D}_t = 0$ we get:

$$\phi(B)\bar{Z}_t = d \tag{D.15}$$

Noting that 1 is not a root of $\phi(B)$, it immediately follows that $\bar{Z}_t = d/\phi(1)$. Consequently, this means that the system defined by eqs. D.12-D.14 cannot generate a drifting signal. An alternative way of seeing this result is by taking $\sigma_\eta^2 = 0$. Then, eq. D.13 becomes equivalent to the naïve construction in eq. D.1, which we have proven in Section D.1 to be unable of generating trend. The behaviour of this process is also illustrated in Figure D.3b.

D.3 The summation model

In this section we propose the summation model, a generative model for sequences with linear trends. Our solution assumes that in the absence of trend, the times series is well modelled by an $ARMA(p,q)$ process.

The main idea is to have two processes evolving independently in the hidden space, one for the signal and one for the linear trend. We are observing $\{Y_t\}$, a noisy version of their sum. The two hidden processes are:

- The signal process, $\{Z_t\}$, which is precisely the $ARMA(p,q)$ used in the absence of trend (eq. 2.10).
- The trend process, $\{D_t\}$, which is the sum of a linear function with increment d and a random walk.

This can be summarized as:

$$Y_t = Z_t + D_t + \omega_t, \qquad \omega_t \sim WN(0, \sigma_\omega^2) \tag{D.16}$$

$$\phi(B)Z_t = \theta(B)\varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \tag{D.17}$$

$$D_t = D_{t-1} + d + \eta_t, \qquad \eta_t \sim WN(0, \sigma_\eta^2) \tag{D.18}$$

It is easy to see that if $D_0 = 0$ then $\bar{D}_t = \mathbb{E}[D_t] = td$. Consequently, $\mathbb{E}[Y_t] = td$. This means that in the summation model a linear trend can be followed while also preserving the dynamics the signal had in the absence of trend.

State-space representation

In the following we give a state space representation of the summation model. This will allow us to estimate the signal and drift processes from noisy observations by applying the Kalman recursions.

First, we employ the state-space representation of *ARMA* models discussed in Section 2.3.2, and denote the resulting parameters $\{\mathbf{A}^{(s)}, \mathbf{C}^{(s)}, \mathbf{Q}^{(s)}, \mathbf{R}^{(\mathbf{s})}\}$.

We use this notation to write down the summation model as:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{d}_t + \mathbf{w}_t, \qquad \mathbf{w}_t \sim WN(0, \mathbf{Q}) \tag{D.19}$$

$$Y_t = \mathbf{C}\mathbf{x}_t + \omega_t, \qquad \omega_t \sim WN(0, \mathbf{R}), \tag{D.20}$$

where

$$\mathbf{x}_t = \begin{bmatrix} \mathbf{x}_t^{(s)} \\ D_t \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}^{(s)} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix}, \quad \mathbf{d}_t = \begin{bmatrix} \mathbf{0} \\ d \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} \mathbf{Q}^{(s)} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & \sigma_\eta^2 \end{bmatrix}$$
$$\mathbf{C} = \begin{bmatrix} \mathbf{C}^{(s)} & 1 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \sigma_\omega^2 \end{bmatrix}.$$

Not that $\lambda = 1$ is an eigenvalue of $\mathbf{A}$, and thus the summation model lies on the LDS stability boundary. In fact, if Λ is the set of eigenvalues of $\mathbf{A}$, then $\Lambda = \Lambda^{(s)} \cup \{1\}$.

As an alternative, the constant control input $\mathbf{d}_t$ can be dropped and replaced by a suitable initialization of the hidden state (e.g. $D_0 \sim \mathcal{N}(D_0; d, \sigma_{\eta_0}^2)$).

In the remainder of this section we provide some insight on the Kalman filtering recursions for summation model. Let hidden state filtered mean be denoted by $\hat{\mathbf{x}}_t$, $\hat{\mathbf{x}}_t \triangleq \mathbb{E}[\mathbf{x}_t | Y_1, Y_2, \ldots, Y_t]$. Also, assume that the Kalman gain matrix, $\mathbf{K}_t$, obeys $\mathbf{K} = \mathbf{K}_t, \forall t \in \mathbb{Z}$[1]. Then, the standard recursion applies:

$$\hat{\mathbf{x}}_t = \mathbf{A}\hat{\mathbf{x}}_{t-1} + \mathbf{d} + \mathbf{K}\big(Y_t - \mathbf{C}(\mathbf{A}\hat{\mathbf{x}}_{t-1} + \mathbf{d})\big). \tag{D.21}$$

In the context of our trend model, we are more interested in filtering estimates of the signal process and of the drift process. We will denote these by $\hat{Z}_t \triangleq \mathbb{E}[Z_t | Y_1, Y_2, \ldots, Y_t]$ and $\hat{D}_t \triangleq \mathbb{E}[D_t | Y_1, Y_2, \ldots, Y_t]$ respectively. Note that $\hat{\mathbf{x}}_t = (\hat{\mathbf{x}}_t^{(s)}, \hat{D}_t)^{\mathrm{T}}$. The recursions follow immediately from eq. D.21 and the definition of the summation model:

$$\hat{\mathbf{x}}_t^{(s)} = \mathbf{A}^{(s)}\hat{\mathbf{x}}_{t-1}^{(s)} + \mathbf{K}^{(s)}\big(Y_t - \mathbf{C}(\mathbf{A}\hat{\mathbf{x}}_{t-1} + \mathbf{d})\big) \tag{D.22}$$

$$\hat{Z}_t = \mathbf{C}^{(s)}\hat{\mathbf{x}}_t^{(s)} \tag{D.23}$$

$$\hat{D}_t = \hat{D}_{t-1} + d + k\big(Y_t - \mathbf{C}(\mathbf{A}\hat{\mathbf{x}}_{t-1} + \mathbf{d})\big) \tag{D.24}$$

[1] This is the steady-state assumption for Kalman filter. It is used often because Kalman recursions tend to converge quickly.

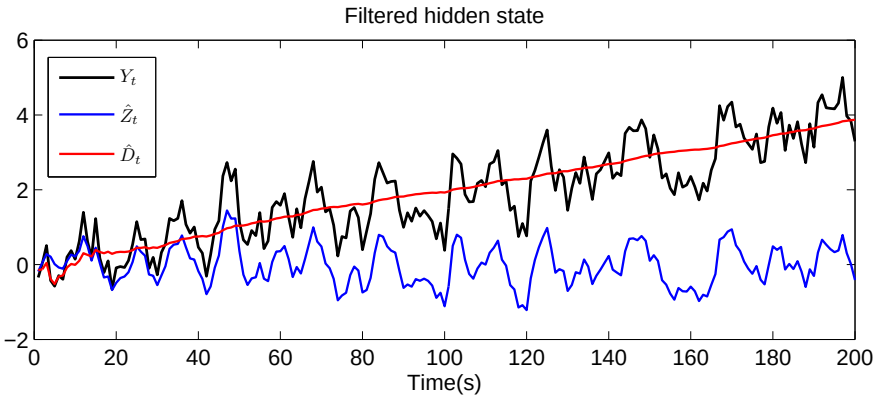


Figure D.4: Y_t is a sample from the summation model with an underlying $ARMA(2,1)$ process. We show that Kalman filtering produces correct results: the filtered signal, $\hat{Z}_t$ (blue) and filtered drift, $\hat{D}_t$ (red).

where we have used the notation $\mathbf{K} = (\mathbf{K}^{(s)}, k)^{\mathrm{T}}$ and took advantage of the block diagonal structure of $\mathbf{A}$. These relations will be used in the next section, where we show some empirical results on applying our trend model to synthetic data.

D.4 Experiments

We now turn our attention to a set of sampling and inference tests on our proposed model. We first sample from the summation model and show some inference results. Since we are primarily interested in discriminating data with a trend from data without trend, we then formulate the problem as inference in a SLDS model (see Section 2.4.1 for a review of SLDS models).

We have already provided a sample form the summation model in Figure D.1a, but now we analyse the model in depth. In Figure D.4 we show another sample from the summation model, $\{Y_t\}$. The underlying stationary process is chosen to be the $ARMA(2,1)$ model:

$$(1 - 1.5B + 0.7B^2)Y_t = (1 - 0.5B)Z_t, \tag{D.25}$$

where $Z_t \sim WN(0, 0.1)$[2]. The parameters of the drift are $d = 0.02$ and $\sigma_\eta^2 = 1e-4$. The observation noise variance is $\sigma_\omega^2 = 0.1$. We see that given the data, the Kalman recursions are correctly inferring the signal component, $\{\hat{Z}_t\}$ (eq. D.23), and the drift component $\{\hat{D}_t\}$ (eq. D.24).

Several sanity checks for the inference results in the summation model are provided in Figure D.5. The focus is on empirically demonstrating that the innovation sequence is Gaussian white noise. The innovation sequence, $\{R_t\}$, is the difference between observed signal and the

[2]The *ARMA* process was chosen such that the roots of the autoregressive polynomial are real and not too close to the unity

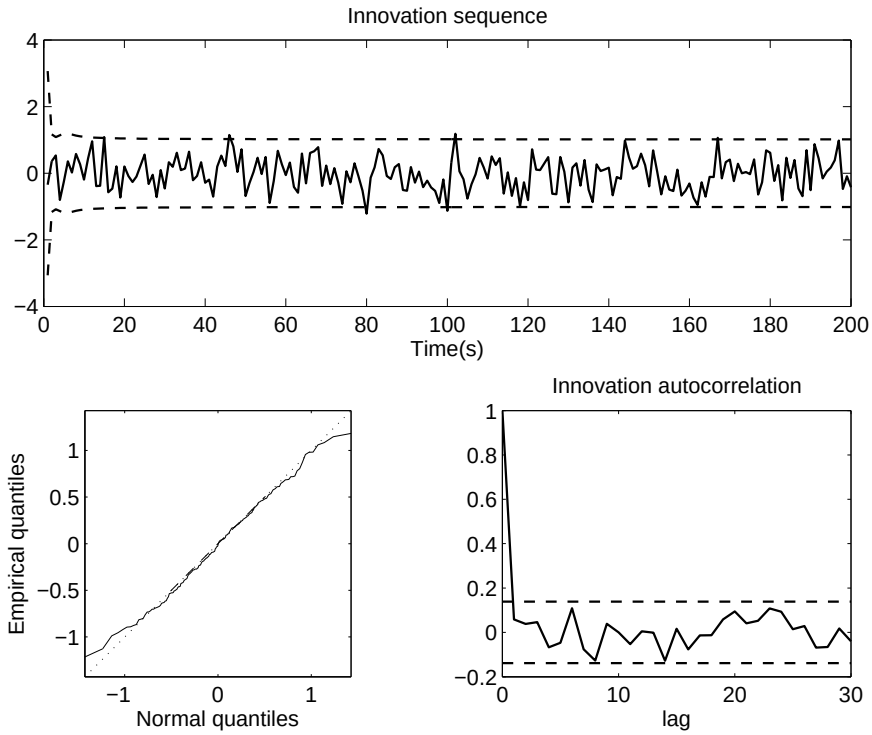


Figure D.5: Top: the innovation sequence (solid line) shown to be between the 2 standard deviations limits (dashed lines) for most of the time. Bottom left: Q-Q plot showing the innovations come from a Gaussian distribution. Bottom right: the autocorrelation of the innovation sequence is insignificant at any non-zero lag.

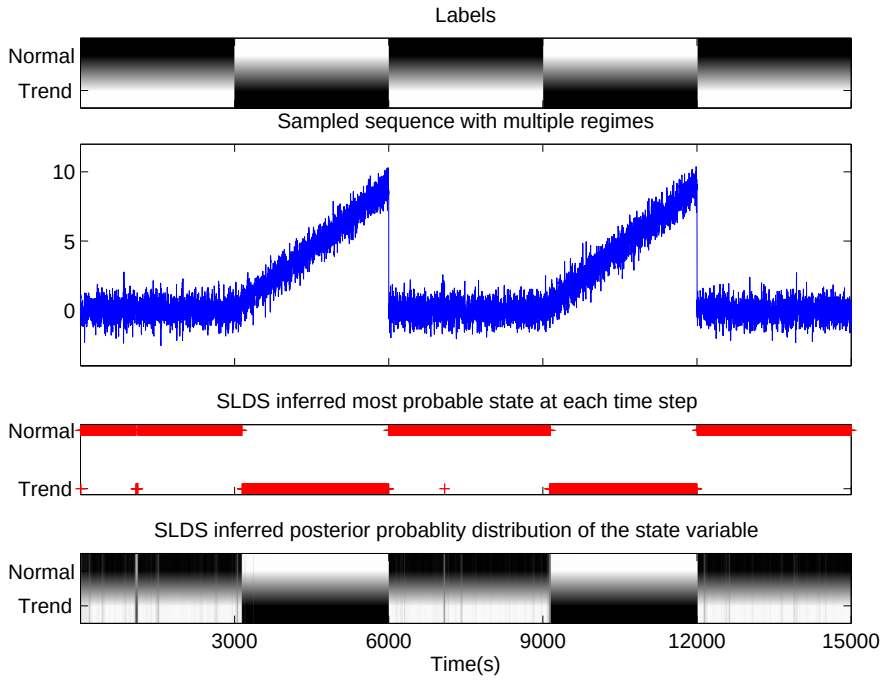


Figure D.6: SKF labels and a sample based on an $ARMA(2,1)$ model (top two plots). Filtered most probable state and filtering distribution (bottom tow plots). Both trend episodes are correctly inferred.

one step ahead Kalman filter prediction ($R_t = Y_t - \mathbf{C}\mathbf{A}\hat{\mathbf{x}}_{t-1}$)). In the top plot, we see that there is no temporal structure in the innovation sequence and that the values fall within ± 2 standard deviations most of the time. The quantile-quantile (Q-Q) plot is almost diagonal, supporting the claim that the innovations come from a Gaussian distribution. In addition, their autocorrelation coefficients at any non-zero lag are insignificant. All these findings mean the innovations sequence is indeed Gaussian white noise.

We now demonstrate that the summation model can be used for the early detection of trends in the data. In order to do this, we build Switching LDS (SLDS) with two hidden regimes. The first regime is described by the stationary $ARMA(2,1)$ process given in eq. D.25. We will call this regime the *Normal* regime. The second regime, *Trend*, will be described by the summation model using the same underlying $ARMA(2,1)$ model. The drift component parameters are set to be $d = 3 \times 10^{-3}$ and $\sigma_\eta^2 = 10^{-6}$. A sample from the SKF is shown in Figure D.6. In this case, the discrete labels are not sampled but a priori fixed, and we learn the discrete state transition probabilities these data. The bottom two plots show the results of running the Gaussian Sum Approximation algorithm (see Appendix A) for approximate SKF filtering. Both trend episodes in the sample are correctly inferred. Also, the filtering probability distribution shows that the model tends to be very confident about its predictions.

An important question is how quickly our proposed model can detect trends. Thus, we plot

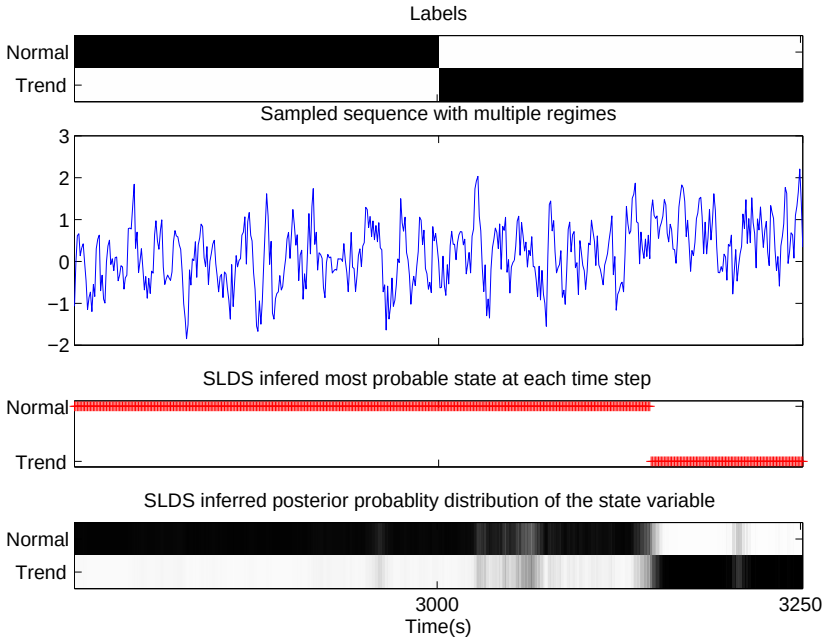


Figure D.7: Detail of Figure D.6. Due to the subtle trend, the model has a delay of approximately 150 seconds in detecting it.

Figure D.7, which is a detailed version of the plots in Figure D.6. Considering that at its early stages the trend is almost impossible to detect, we see that the inference results are reasonably good.

D.5 Relationship with ARIMA models

In this section we show that the trend models discussed above are in fact special cases of *ARIMA* models (Section 2.3.1). We begin by proving that noisy observations of *ARMA* (*ARIMA*) processes are *ARMA* (*ARIMA*) in their own right albeit of higher order. Then, we use these results to show that both the local linear trend and summation models are *ARIMA* processes as well.

D.5.1 Observation noise

In many cases, the true dynamics of a system are hidden by a noisy observation processes. Under linear-Gaussian assumptions, such scenarios can be elegantly tackled in the LDS framework extensively discussed in the previous chapters. In Section 2.3.2, we have analysed the case when the hidden dynamics are given by an $ARMA(p,q)$ model and the observation noise is Gaussian. More precisely, we assume access to the values of a process $\{Y_t\}$, consisting of noisy observations of the hidden process $\{Z_t\}$. The model was defined in eqs. 2.31 and 2.32.

Here, we show that $\{Y_t\}$ is an *ARMA* process as well. First, being the sum of two indepen-

dent stationary processes, $\{Y_t\}$ is stationary as well. If we use eq 2.31 to substitute Z_t for Y_t in eq. 2.32 and rearrange we get:

$$\phi(B)Y_t = \theta(B)\varepsilon_t + \phi(B)\omega_t \tag{D.26}$$

We now introduce $\nu_t \triangleq \varepsilon_t + \omega_t$. Since it is the sum of two Gaussian white noise processes,$\{\nu_t\}$ is also Gaussian white noise. Using this we can re-write eq. D.26 as:

$$\phi(B)Y_t = \theta^*(B)\nu_t, \qquad \nu_t \sim WN(0, \sigma_\varepsilon^2 + \sigma_\omega^2) \tag{D.27}$$

Here, $\theta^*(B)$ is a polynomial of degree $r = \max(p,q)$ whose coefficients can be immediately derived from the coefficients of $\phi(B)$ and $\theta(B)$ using the definition of $\{\nu_t\}$. Notice that $\{Y_t\}$ and $\{Z_t\}$ share the same autoregressive polynomials. We can now conclude that $\{Y_t\}$ is and $ARMA(p,r)$ process.

As in the case of *ARMA* models, for *ARIMA* processes we observe $\{Y_t\}$, a noise corrupted version of the $\{Z_t\}$ process defined in eq. 2.12. If the noise is Gaussian white noise, the noisy process is described by:

$$Y_t = Z_t + \omega_t, \qquad \omega_t \sim WN(0, \sigma_\omega^2) \tag{D.28}$$

$$\phi(B)(1-B)^{(d)}Z_t = \theta(B)\varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \tag{D.29}$$

Using a similar analysis as in the previous section, one can show that $\{Y_t\}$ is an $ARIMA(p,d,r)$ process, where $r = \max(q, p+d)$.

D.5.2 Local linear trend

We now show that the local linear trend is an $ARIMA(0,2,2)$ process.

The following proof is simplified by the results above, where we studied the effect noisy observations processes have on *ARMA* (*ARIMA*) models. The interpretation we give here is that if $Z_t = \mathbb{E}_\omega[Y_t]$ (by eq. D.4) is an *ARIMA* process, then $\{Y_t\}$ is an an *ARIMA* process too.

Thus, our problem becomes equivalent to analysing the process $\{Z_t\}$ given by:

$$Z_t = Z_{t-1} + D_{t-1} + d + \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2) \tag{D.30}$$

$$D_t = D_{t-1} + \eta_t, \qquad \eta_t \sim WN(0, \sigma_\eta^2) \tag{D.31}$$

Let $\Delta Z_t \triangleq Z_t - Z_{t-1} = D_{t-1} + d + \varepsilon_t$. It follows that $\Delta Z_{t-1} = D_{t-2} + d + \varepsilon_{t-1}$. Combining these two with eq. D.31 gives:

$$\Delta Z_t = \Delta Z_{t-1} + \eta_{t-1} + \varepsilon_t - \varepsilon_{t-1} \triangleq \Delta Z_{t-1} + \xi_t + \theta\xi_{t-1}, \tag{D.32}$$

where $\xi_t \triangleq \varepsilon_t + \eta_{t-1}$, $\xi_t \sim \mathcal{N}(0, \sigma_\varepsilon^2 + \sigma_\eta^2)$ and $\theta = -\sqrt{\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + \sigma_\eta^2)}$. We see that $\{\Delta Z_t\}$ is again a non-stationary process. Further differencing produces

$$\Delta\Delta Z_t \triangleq \Delta Z_t - \Delta Z_{t-1} = \xi_t + \theta\xi_{t-1}, \tag{D.33}$$

which is an $ARMA(0,1)$ process (i.e. an $MA(1)$ process). Thus, the mean of the local linear trend, $\{Z_t\}$, is an $ARIMA(0,2,1)$ process. From here, it is straightforward to see that the local linear trend is an $ARIMA(0,2,2)$ process.

D.5.3 The summation model

In similar fashion to proving the local linear trend is an *ARIMA* model we now show that summation model is one too.

We begin by studying the same example as in Section D.4, where we used an $ARMA(2,1)$ model for the hidden signal process. We will first prove that the expectation $\bar{Y}_t \triangleq \mathbb{E}_\omega[Y_t]$ (by eq. D.16) is an $ARIMA(2,1,2)$ process. Using literals instead of numbers, $\{\bar{Y}_t\}$ is defined as:

$$\bar{Y}_t = Z_t + D_t \tag{D.34}$$

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \tag{D.35}$$

$$D_t = D_{t-1} + d + \eta_t, \tag{D.36}$$

where $\varepsilon_t \sim WN(0,\sigma_\varepsilon^2)$ and $\eta_t \sim WN(0,\sigma_\eta^2)$.

We now difference $\bar{Y}_t$ to give

$$\Delta\bar{Y}_t = \Delta Z_t + \Delta D_t = (\phi_1 \Delta Z_{t-1} + \phi_2 \Delta Z_{t-2} + \Delta\varepsilon_t + \theta_1 \Delta\varepsilon_{t-1}) + (d + \eta_t) \tag{D.37}$$

$$\Delta\bar{Y}_{t-1} = \Delta Z_{t-1} + \Delta D_{t-1} = \Delta Z_{t-1} + d + \eta_{t-1} \tag{D.38}$$

$$\Delta\bar{Y}_{t-2} = \Delta Z_{t-2} + \Delta D_{t-2} = \Delta Z_{t-2} + d + \eta_{t-2} \tag{D.39}$$

Substituting ΔZ_{t-1} and ΔZ_{t-2} from eqs. D.38 and D.39 into eq D.37 and after some manipulation we get:

$$(1 - \phi_1 B - \phi_2 B^2)(\Delta\bar{Y}_t - d) = (1 + \theta_1 B)(1 - B)\varepsilon_t + (1 - \phi_1 B - \phi_2 B^2)\eta_t \tag{D.40}$$

We claim that $\Delta\bar{Y}_t$ is an $ARMA(2,2)$ process. To emphasize this we define $\xi_t = \varepsilon_t + \eta_t$, $\xi_t \sim WN(0,\sigma_\varepsilon^2 + \sigma_\eta^2)$. Also, denote $\alpha_s = \sqrt{\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + \sigma_\eta^2)}$ and $\alpha_d = \sqrt{\sigma_\eta^2/(\sigma_\varepsilon^2 + \sigma_\eta^2)}$. Then, eq. D.40 becomes:

$$(1 - \phi_1 B - \phi_2 B^2)(\Delta\bar{Y}_t - d) = (1 + \gamma_1 B + \gamma_2 B^2)\xi_t \tag{D.41}$$

where $\xi_t \sim WN(0,\sigma_\varepsilon^2 + \sigma_\eta^2)$ and

$$\gamma_1 = (\theta_1 - 1)\alpha_s - \phi_1\alpha_d$$

$$\gamma_2 = -\theta_1\alpha_s - \phi_2\alpha_d.$$

We have just proven that the mean of the summation model based on an $ARMA(2,1)$ hidden signal process is an $ARIMA(2,1,2)$ process. Consequently using the findings of Section D.5.1, we get that the summation model is an $ARIMA(2,1,3)$ process. At the same time, $\{\Delta\bar{Y}_t\}$ is an $ARMA(2,2)$ process with mean $\mu = d$.

Table D.1: Equivalence to *ARIMA* processes

Model	*ARIMA* equivalent
$ARMA(p,q)$ + Gaussian white noise	$ARIMA(p,0,r)$, $r=\max(p,q)$
$ARIMA(p,d,q)$ + Gaussian white noise	$ARIMA(p,d,r)$, $r=\max(p+d,q)$
local linear trend	$ARIMA(0,2,2)$
summation model	$ARIMA(p,1,r)$, $r=\max(p+1,q+1)$

We can generalise our reasoning for summation models with hidden signal processes of arbitrary order. It is straightforward to show that eq. D.40 becomes:

$$\phi(B)(\Delta\bar{Y}_t - d) = \theta(B)(1-B)\varepsilon_t + \phi(B)\eta_t \tag{D.42}$$

where where $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \ldots + \theta_p B^q$. Also, $\{\Delta\bar{Y}_t\}$ is an $ARMA(p,r)$ process with mean $\mu = d$, where $r = \max(p, q+1)$. This means that mean of the summation model is an $ARIMA(p,1,r)$ model. Consequently, the summation model is an $ARIMA(p,1,r')$ process, where $r' = \max(p+1, q+1)$.

D.6 Summary

We have proposed the summation model, a generative model for time series displaying linear trends in the mean. The main idea was to have two latent processes, a potentially rich stationary component and a simple non-stationary linear drift component. The observed process is the noisy sum of these two components.

Using synthetic data, we have then demonstrated that the model can accurately infer both hidden signal and drift components. In addition, we have been able to discriminate trend periods from periods without it by formulating the trend detection task as SLDS inference.

As summarised in Table D.1, we have proven that the summation model is a special case of *ARIMA* model.

Bibliography

V Ahlborn, B Bohnhorst, CS Peter, and CF Poets. False alarms in very low birthweight infants: comparison between three intensive care monitoring systems. *Acta Pdiatrica*, 89(5):571–576, 2000. ISSN 1651-2227.

Daniel L. Alspach and Harold W. Sorenson. Nonlinear Bayesian Estimation Using Gaussian Sum Approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972.

Shun-Ichi Amari. Integration of Stochastic Models by Minimizing α-Divergence. *Neural Comput.*, 19(10):2780–2796, October 2007. ISSN 0899-7667.

R.V. Andreao, B. Dorizzi, and J. Boudy. ECG signal analysis through hidden Markov models. *Biomedical Engineering, IEEE Transactions on*, 53(8):1541–1549, 2006. ISSN 0018-9294. doi: 10.1109/TBME.2006.877103.

M. Azzouzi and IT. Nabney. Modelling financial time series with switching state space models. In *Computational Intelligence for Financial Engineering, 1999. (CIFEr) Proceedings of the IEEE/IAFE 1999 Conference on*, pages 240–249, 1999. doi: 10.1109/CIFER.1999.771123.

Yaakov Bar-Shalom, Thiagalingam Kirubarajan, and X.-Rong Li. *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, Inc., New York, NY, USA, 2001. ISBN 0471221279.

D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

David Barber and Bertrand Mesot. A Novel Gaussian Sum Smoother for Approximate Inference in Switching Linear Dynamical Systems. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 89–96. MIT Press, Cambridge, MA, 2006.

C. M. Beck-Sague, P. Azimi, S. N. Fonseca, R. S. Baltimore, D.A. Powell, L. A. Bland, M. J. Arduino, S.K. McAllister, R. S. Huberman, and R. L. Sinkowitz. Bloodstream infections in neonatal intensive care unit patients: results of a multicenter study. *The Pediatric Infectious Disease Journal*, 13(12):1110–1116, 1994.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

David M. Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

M. Blount, M.R. Ebling, J.M. Eklund, A.G. James, C. McGregor, N. Percival, K.P. Smith, and D. Sow. Real-Time Analysis for Intensive Care: Development and Deployment of the Artemis Analytic System. *Engineering in Medicine and Biology Magazine, IEEE*, 29(2):110 –118, march-april 2010. ISSN 0739-5175. doi: 10.1109/MEMB.2010.936454.

Byron Boots. *Spectral Approaches to Learning Predictive Representations*. PhD thesis, Carnegie Mellon University, December 2012.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

P.J. Brockwell and R.A. Davis. *Time series: theory and methods*. Springer series in statistics. Springer-Verlag, 1991. ISBN 9783540974291.

Ali Taylan Cemgil, Hilbert J. Kappen, and D. Barber. A generative model for music transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 14(2):679–694, 2006.

Chris Chatfield. *The Analysis of Time Series – An Introduction*. Chapman & Hall/Crc, 6th edition, 2004.

D.A. Coast, R.M. Stern, G.G. Cano, and S.A. Briller. An approach to cardiac arrhythmia analysis using hidden markov models. *Biomedical Engineering, IEEE Transactions on*, 37 (9):826–836, 1990. ISSN 0018-9294. doi: 10.1109/10.58593.

N. de Freitas, R. Dearden, Frank Hutter, R. Morales-Menendez, J. Mutch, and D. Poole. Diagnosis by a waiter and a Mars explorer. *Proceedings of the IEEE*, 92(3):455–468, Mar 2004. ISSN 0018-9219. doi: 10.1109/JPROC.2003.823157.

Susan DeMeulenaere. Pulse Oximetry: Uses and Limitations. *The Journal for Nurse Practitioners*, 3(5):312–317, 2007.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

Li Deng. *Dynamic Speech Models: Theory, Algorithms, and Applications*. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2006.

Jasha Droppo and Alex Acero. Noise Robust Speech Recognition with a Switching Linear Dynamic Model. In *Proc. ICASSP*, Montreal, Canada, May 2004. IEEE.

Zhansheng Duan, Vesselin P. Jilkov, and X. Rong Li. State estimation with quantized measurements: Approximate MMSE approach. In *FUSION*, pages 1–6. IEEE, 2008. ISBN 978-3-00-024883-2.

I. A. Eckley, P. Fearnhead, and R. Killick. Analysis of Changepoint Models. In *Bayesian Time Series Models*. Cambridge University Press, 2011.

Y. Ephraim, D. Malah, and B.-H. Juang. On the application of hidden Markov models for enhancing noisy speech. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(12):1846 –1856, dec 1989. ISSN 0096-3518. doi: 10.1109/29.45532.

M. Everingham and J. Winn. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit. http://www.pascal-network.org/challenges/VOC/voc2012, 2012.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, 2004.

Tom Fawcett and Foster Provost. Activity Monitoring: Noticing interesting changes in behavior. In *In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 53–62, 1999.

J D Ferguson. Variable duration models for speech. In *Symp. on the Application of HMMs to Text and Speech*, pages 143–179, 1980.

Shai Fine and Yoram Singer. The Hierarchical Hidden Markov Model: Analysis and Applications. In *Machine Learning*, pages 41–62, 1998.

Abigial Flower, Randall Joseph Moorman, Douglas Lake, and John Delos. Periodic heart rate decelerations in premature infants. *Experimental Biology and Medicine*, 235(4):531–538, 2010.

E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Nonparametric Bayesian Learning of Switching Linear Dynamical Systems. In *Neural Information Processing Systems 21*. MIT Press, 2009.

E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Sharing Features among Dynamical Systems with Beta Processes. In *Neural Information Processing Systems 22*. MIT Press, 2010.

Dale Gerstmann, Ryan Berg, Ron Haskell, Cathy Brower, Kari Wood, Brad Yoder, Loren Greenway, Gordon Lassen, Robert Ogden, Ronald Stoddard, and Stephen Minton. Operational Evaluation of Pulse Oximetry in NICU Patients with Arterial Access. *Journal of Perinatology*, 23(5):378–383, 2003.

Zoubin Ghahramani and Geoffrey E. Hinton. Parameter Estimation for Linear Dynamical Systems. Technical report, University of Toronto, 1996.

Zoubin Ghahramani and Geoffrey E. Hinton. Variational Learning for Switching State-Space Models. *Neural Computation*, 12(4):831–864, 2000.

Zoubin Ghahramani and Michael I. Jordan. Factorial Hidden Markov Models. *Machine Learning*, 29:245–273, 1997.

Stuart Gibson and Brett Ninness. On The relationship between State-Space-Subspace-Based and Maximum-Likelihood System Identification methods. In *IEEE International Conference on Decision and Control, Sydney Australia*, pages 2415–2418, dec 2000.

M. Pamela Griffin and J. Randall Moorman. Toward the Early Diagnosis of Neonatal Sepsis and Sepsis-Like Illness Using Novel Heart Rate Analysis. *Pediatrics*, 107:97–104, 2001.

M Pamela Griffin, T Michael O'Shea, Eric A Bissonette, Frank E Harrell, Douglas E Lake, and J Randall Moorman. Abnormal Heart Rate Characteristics Preceding Neonatal Sepsis and Sepsis-Like Illness. *Pediatr Res*, 53(6):920–6, 2003. ISSN 0031-3998.

G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford University Press, third edition, 2001.

James D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1-2):39–70, 1990.

James Douglas Hamilton. *Time series analysis*. Princeton Univ. Press, Princeton, NJ, 1994. ISBN 0691042896.

A.C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1991. ISBN 9780521405737.

Andrew Harvey. *Forecasting with Unobserved Components Time Series Models*, volume 1 of *Handbook of Economic Forecasting*, chapter 7, pages 327–412. Elsevier, January 2006.

William W. Hay, Donna J. Rodden, Shannon M. Collins, Diane L. Melara, Kathy A. Hale, and Lucy M. Fashaw. Reliability of Conventional and New Pulse Oximetry in Neonatal Patients. *Journal of Perinatology*, 22(5):360–366, 2002.

Mary Fran Hazinski, Ricardo Samson, and Steve Schexnayder. *Handbook of Emergency Cardiovascular Care for Healthcare Providers*. American Heart Association, 2010.

G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

S.W. Hoare and P.C.W. Beatty. Automatic artifact identification in anaesthesia patient record keeping: a comparison of techniques. *Medical Engineering & Physics*, 22(8):547 – 553, 2000. ISSN 1350-4533.

M.T. Johnson. Capacity and complexity of HMM duration modeling techniques. *Signal Processing Letters, IEEE*, 12(5):407–410, 2005. ISSN 1070-9908. doi: 10.1109/LSP.2005.845598.

M. I. Jordan and K. A. Jacobs. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, 1993.

Michael I. Jordan, Zoubin Ghahramani, and Lawrence K. Saul. Hidden Markov decision trees. In *Advances in Neural Information Processing Systems*, 1997.

Rudolph Emil Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

N. Kantas, A. Doucet, S. S. Singh, and J. M. Maciejowski. An Overview of Sequential Monte Carlo Methods for Parameter Estimation. In *General State-Space Models, in IFAC System Identification, no. Ml*, 2009.

Y. Karklin and M. S. Lewicki. A hierarchical Bayesian model for learning non-linear statistical regularities in non-stationary natural signals. *Neural Computation*, 17(2):397–423, 2005.

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

Hyungsul Kim, Manish Marwah, Martin F. Arlitt, Geoff Lyon, and Jiawei Han. Unsupervised Disaggregation of Low Frequency Power Measurements. In *Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM*, pages 747–758, 2011.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN 0262013193, 9780262013192.

J. Zico Kolter and Tommi Jaakkola. Approximate Inference in Additive Factorial HMMs with Application to Energy Disaggregation. In Neil D. Lawrence and Mark Girolami, editors, *AISTATS*, volume 22 of *JMLR Proceedings*, pages 1472–1482. JMLR.org, 2012.

Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjlander, and David Haussler. Hidden Markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

Douglas E. Lake. Renyi entropy measures of heart rate Gaussianity. *Biomedical Engineering, IEEE Transactions on*, 53(1):21–27, Jan 2006.

Steffen Lilholt Lauritzen. Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87:1098–1108, 1992. ISSN 0162-1459.

L Lehman, R Adams, L Mayaud, G Moody, A Malhotra, R Mark, and S Nemati. A Physiological Time Series Dynamics-Based Approach to Patient Monitoring and Outcome Prediction. *J Biomedical and Health Informatics*, (99):1–1, 2014.

L.H. Lehman, S. Nemati, R.P. Adams, and R.G. Mark. Discovering shared dynamics in physiological signals: Application to patient monitoring in ICU. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 5939–5942, Aug 2012.

Li Wei Lehman, Shamim Nemati, Ryan P. Adams, George Moody, Atul Malhotra, and Roger Mark. Tracking Progression of Patient State of Health in Critical Care Using Inferred Shared Dynamics in Physiological Time Series. In *Proceedings of the 35th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 07/2013 2013.

Uri Lerner and Ronald Parr. Inference in hybrid networks: Theoretical limits and practical algorithms. In *UAI*, pages 310–318, 2001.

Uri Lerner, Ronald Parr, Daphne Koller, and Gautam Biswas. Bayesian Fault Detection and Diagnosis in Dynamic Systems. In *In Proc. AAAI*, pages 531–537, 2000.

R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. New York, Wiley, 1987.

A J Lyon, M E Pikaar, P Badger, and N McIntosh. Temperature control in very low birthweight infants during first five days of life. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 76(1):F47–F50, 1997.

Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.

Richard J. Martin and Avroy A. Fanaroff. Neonatal apnea, bradycardia, or desaturation: Does it matter? . *The Journal of Pediatrics*, 132(5):758 – 759, 1998.

C. McGregor, C. Catley, and A. James. Variability analysis with analytics applied to physiological data streams from the neonatal intensive care unit. In *Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on*, pages 1–5, 2012. doi: 10.1109/CBMS.2012.6266385.

Neil McIntosh, Julie-Clare Becher, Steven Cunningham, Ben Stenson, Ian A. Laing, Andrew J. Lyon, and Peter 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.

B. Mesot and D. Barber. Switching Linear Dynamical Systems for Noise Robust Speech Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(6):1850–1858, Aug 2007.

Silvia Miksch, Werner Horn, Christian Popow, and Franz Paky. Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants. *Artificial Intelligence in Medicine*, 8:543–576, 1996.

Michelle S Miller, Kevin M Shannon, and Glenn T Wetzel. Neonatal bradycardia. *Progress in Pediatric Cardiology*, 11(1):19 – 24, 2000.

Thomas Minka. *A family of algorithms for approximate Bayesian inference*. Phd thesis, MIT, 2001.

N. Modi, C. J. Doré, A. Saraswatula, M. Richards, K. B. Bamford, R. Coello, and A. Holmes. A case definition for national and international neonatal bloodstream infection surveillance. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 94(1):F8–F12, January 2009.

Joseph Randall Moorman, Waldemar A. Carlo, John Kattwinkel, Robert L. Schelonka, Peter J. Porcelli, Christina T. Navarrete, Eduardo Bancalari, Judy L. Aschner, Marshall Whit Walker, Jose A. Perez, Charles Palmer, George J. Stukenborg, Douglas E. Lake, and Thomas Michael OShea. Mortality Reduction by Heart Rate Characteristic Monitoring in Very Low Birth Weight Neonates: A Randomized Trial. *The Journal of Pediatrics*, 159(6):900 – 906.e1, 2011. ISSN 0022-3476.

J.R. Moorman, D.E. Lake, and M.P. Griffin. Heart rate characteristics monitoring for neonatal sepsis. *Biomedical Engineering, IEEE Transactions on*, 53(1):126–132, 2006. ISSN 0018-9294. doi: 10.1109/TBME.2005.859810.

Rubn Morales-Menéndez, Nando de Freitas, and David Poole. Real-Time Monitoring of Complex Industrial Processes with Particle Filters. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 1433–1440, 2002.

Kevin Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, July 2002.

Kevin Murphy and S Russell. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In N. de Freitas A. Doucet and N. Gordon, editors, *Sequential Monte Carlo in Practice*. Springer-Verlag, 2001.

Kevin P. Murphy. Switching Kalman Filters. Technical report, U.C. Berkeley, 1998.

K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, 2012. ISBN 9780262018029.

Radford Neal and Geoffrey E. Hinton. A View Of The Em Algorithm That Justifies Incremental, Sparse, And Other Variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic, 1996.

Vladimir Pavlovic, James M. Rehg, and John Maccormick. Learning switching linear models of human motion. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 981–987, 2000.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 1558604790.

Marco A. F. Pimentel, David A. Clifton, Lei A. Clifton, Peter Watkinson, and Lionel Tarassenko. Modelling physiological deterioration in post-operative patient vital-sign data. *Med. Biol. Engineering and Computing*, 51(8):869–877, 2013.

Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215 – 249, 2014. ISSN 0165-1684.

Foster Provost and Tom Fawcett. Robust Classification for Imprecise Environments. *Mach. Learn.*, 42(3):203–231, March 2001. ISSN 0885-6125. doi: 10.1023/A:1007601015854.

John Quinn. *Bayesian Condition Monitoring in Neonatal Intensive Care*. PhD thesis, University of Edinburgh, 2007.

John A. Quinn, Christopher K. I. Williams, and Neil McIntosh. Factorial Switching Linear Dynamical Systems Applied to Physiological Condition Monitoring. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(9):1537–1551, 2009.

Lawrence Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM Trans. Inf. Syst.*, 7(3):205–229, July 1989. ISSN 1046-8188. doi: 10.1145/65943.65945.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)*. The MIT Press, 2005. ISBN 026218253X. URL `http://www.worldcat.org/isbn/026218253X`.

Carl Edward Rasmussen and Hannes Nickisch. The GPML Toolbox, 2013. http://www.gaussianprocess.org/gpml/code/matlab/doc/.

H. E. Rauch, C. T. Striebel, and F. Tung. Maximum Likelihood Estimates of Linear Dynamic Systems. *Journal of the American Institute of Aeronautics and Astronautics*, 3(8):1445–1450, August 1965.

Sam Roweis and Zoubin Ghahramani. A Unifying Review of Linear Gaussian Models. *Neural Comput.*, 11(2):305–345, February 1999. ISSN 0899-7667.

Sam T. Roweis. One Microphone Source Separation. In *In Advances in Neural Information Processing Systems 13*, pages 793–799. MIT Press, 2000.

Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2 edition, 2003. ISBN 0137903952.

John Salyer. Neonatal and Pediatric Pulse oximetry. *Respiratory Care*, 48(4):386–398, 2003.

S. Saria, D. Koller, and A. Penn. Discovering shared and individual latent structure in multiple time series. Technical report, 2010a. URL `http://arxiv.org/abs/1008.2028`.

Suchi Saria, Anand K. Rajani, Jeffrey Gould, Daphne Koller, and Anna A. Penn. Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants. *Science Translational Medicine*, 2(48):48–65, 2010b.

R. Shumway and D. Stoffer. Dynamic linear models with switching. *J. of the American Statistical Association*, 86:763–769, 1991.

A. F. Smith and M. West. Monitoring renal transplants: an application of the multiprocess Kalman filter. *Biometrics*, 39(4):867–878, 1983. ISSN 0006-341X.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian Processes using Pseudo-inputs. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 1257–1264. MIT press, 2006.

Alexander Spengler. Neonatal Baby Monitoring. Master's thesis, University of Edinburgh, School of Informatics, 2003.

M. Stacey, C. McGregor, and M. Tracy. An architecture for multi-dimensional temporal abstraction and its application to support neonatal intensive care. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 3752 –3756, aug. 2007. doi: 10.1109/IEMBS.2007.4353148.

Ioan Stanculescu, Christopher K. I. Williams, and Yvonne Freer. Autoregressive Hidden Markov Models for the Early Detection of Neonatal Sepsis. *Biomedical and Health Informatics, IEEE Journal of*, 2013. ISSN 2168-2194. doi: 10.1109/JBHI.2013.2294692. DOI 10.1109/JBHI.2013.2294692.

Ioan Stanculescu, Christopher K. I. Williams, and Yvonne Freer. A hierarchical switching linear dynamical system applied to the detection of sepsis in neonatal condition monitoring. In *Uncertainty in Artificial Intelligence*, pages 752–761, 2014.

Barbara J Stoll, Nellie Hansen, Avroy A Fanaroff, Linda L Wright, Waldemar A Carlo, Richard A Ehrenkranz, James A Lemons, Edward F Donovan, Ann R Stark, Jon E Tyson, and et al. Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network. *Pediatrics*, 110(2 Pt 1):285–291, 2002.

G. W. Taylor, L. Sigal, D. Fleet, and G. E. Hinton. Dynamic binary latent variable models for 3D human pose tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2010*, 2010.

Yee-Whye Teh and Michael I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

Christine L. Tsien. *Trendfinder: automated detection of alarmable trends*. PhD thesis, Massachusetts Institute of Technology, 2000.

C.L. Tsien and J.C. Fackler. Poor prognosis for existing monitors in the intensive care unit. *Critical Care Medicine*, 25(4):614–619, 1997.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91+, feb 2006.

Jenna Wiens, Eric Horvitz, and John V. Guttag. Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 467–475. Curran Associates, Inc., 2012.

Christopher K. I. Williams and Ioan Stanculescu. Automating the Calibration of a Neonatal Condition Monitoring System. In M. Peleg, N. Lavrac, and C. Combi, editors, *Proc AIME 2011*, volume 6747 of *LNAI*, pages 240–249. Springer, 2011.

Christopher K. I. Williams, John A. Quinn, and Neil 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 distributions. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:509–512, 1992.

S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.

Onno Zoeter and Tom Heskes. Hierarchical visualization of time-series data using switching linear dynamical systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1201–1214, 2003.