THE UNIVERSITY *of* EDINBURGH

Dynamical Probabilistic Graphical Models applied to Physiological Condition Monitoring

Konstantinos Georgatzis

Doctor of Philosophy

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh

2017

Abstract

Intensive Care Units (ICUs) host patients in critical condition who are being monitored by sensors which measure their vital signs. These vital signs carry information about a patient's physiology and can have a very rich structure at fine resolution levels. The task of analysing these biosignals for the purposes of monitoring a patient's physiology is referred to as physiological condition monitoring.

Physiological condition monitoring of patients in ICUs is of critical importance as their health is subject to a number of events of interest. For the purposes of this thesis, the overall task of physiological condition monitoring is decomposed into the sub-tasks of modelling a patient's physiology a) under the effect of physiological or artifactual events and b) under the effect of drug administration. The first sub-task is concerned with modelling artifact (such as the taking of blood samples, suction events etc.), and physiological episodes (such as bradycardia), while the second sub-task is focussed on modelling the effect of drug administration on a patient's physiology.

The first contribution of this thesis is the formulation, development and validation of the Discriminative Switching Linear Dynamical System (DSLDS) for the first sub-task. The DSLDS is a discriminative model which identifies the state-of-health of a patient given their observed vital signs using a discriminative probabilistic classifier, and then infers their underlying physiological values conditioned on this status. It is demonstrated on two real-world datasets that the DSLDS is able to outperform an alternative, generative approach in most cases of interest, and that an α-mixture of the two models achieves higher performance than either of the two models separately.

The second contribution of this thesis is the formulation, development and validation of the Input-Output Non-Linear Dynamical System (IO-NLDS) for the second sub-task. The IO-NLDS is a non-linear dynamical system for modelling the effect of drug infusions on the vital signs of patients. More specifically, in this thesis the focus is on modelling the effect of the widely used anaesthetic drug Propofol on a patient's monitored depth of anaesthesia and haemodynamics. A comparison of the IO-NLDS with a model derived from the Pharmacokinetics/Pharmacodynamics (PK/PD) literature on a real-world dataset shows that significant improvements in predictive performance can be provided without requiring the incorporation of expert physiological knowledge.

Acknowledgements

Despite the fact that a PhD is mostly a lonely process, its completion is nonetheless only possible thanks to the contributions of many people apart from the author. Chris Williams has been an excellent primary supervisor, guiding me through this long process with patience and providing me always with insightful advice and invaluable academic guidance. Simon Rogers and Ian Piper have both been very helpful co-supervisors and their fresh ideas have always been a welcome addition to this thesis' content.

Chris Hawthorne's help has been essential for the completion of this thesis. His expert medical advice and provision of relevant data made the applications of the developped methodologies possible. Similarly, Martin Shaw has provided very kindly his data preprocessing code which greatly facilitated me in focussing more onto the modelling tasks at hand.

I also had the benefit of working within a very stimulating environment in the Institute for Adaptive and Neural Computation in the University of Edinburgh. Many of the concepts developped within this thesis have been refined by countless discussions with my colleagues at the University.

My thanks also go to the Scottish Informatics and Computer Science Alliance (SICSA) and the Informatics Graduate School for funding my project.

Last, but certainly not least, my parents and my brother have been of immense help during those four years (and many years before that) and it is thanks to their unfailing emotional support that this thesis was made 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.

(*Konstantinos Georgatzis*)

Table of Contents

Chapter 1

Introduction

This thesis treats the subject of developing dynamical probabilistic graphical models (DPGMs) and how they could be applied to the task of monitoring the physiological condition of patients whose health is in a critical state.

The importance of such a task is hopefully obvious to the reader. An approach that could characterise accurately and in a timely manner the changes in the state of health of a critically ill patient could have highly beneficial consequences. A decision support system could emerge from such an approach with the goal of providing severity-graded alerts to the clinical personnel so they could prioritise their scarce time in a more appropriate manner.

DPGMs are a natural formalism for such a setting. A patient's health evolves over time and thus the dynamics of this temporal evolution need to be explicitly captured. Furthermore, a multitude of complex, interacting factors renders this task into one of high uncertainty, which needs to be incorporated into the modelling process. This leads naturally to the domain of probabilistic calculus and the associated graphical models' framework.

The remaining of this introductory chapter provides a description of the Intensive Care Unit setting in Section 1.1, presents the concept of physiological condition monitoring in Section 1.2 and summarises the contributions of this thesis in Section 1.3. Finally, Section 1.4 provides an overview of the structure of the remainder of the thesis.

1.1 Intensive Care Units

Intensive Care Units (ICUs) constitute departments within hospitals that are especially designed in order to provide advanced care to patients whose condition is critical and their health is deemed to be at very high risk. They are staffed with doctors and nurses who have received special training and they always have a higher staff-to-patient ratio than other clinical departments, which translates into higher levels of patient monitoring and eventually reduced incidences of mortality (Pronovost *et al.*, 2002). ICUs are further divided in sub-categories, depending on their area of specialisation.

The main source of information for this thesis will be extracted from Neuro-ICUs. Neuro-ICUs host patients in critical condition who have been moved there usually after having suffered some serious trauma, like traumatic brain injury (TBI). Brain trauma can occur either after delivery of a focal impact on a specific area of the head, by a sudden acceleration/deceleration within the cranium or potentially by a combination of both movement and sudden impact (Maas *et al.*, 2008). Frequent brain trauma sources include falls, vehicle accidents and violence.

Neonatal ICUs (NICUs) will form the secondary subject of this thesis' models. NICUs specialise in care delivery for ill or prematurely born infants. Often, the lack of development associated with a premature birth is the only complication these infants face. In these cases the goal is simply to provide the necessary physiological support and tackle any additional complications that may arise. In other cases however, prematurity can be associated with serious respiratory or cardiovascular problems or other types of pathology and can result in reduced survival rates (Cooper *et al.*, 1998).

1.2 Physiological condition monitoring

Patients admitted in ICUs need to be, by definition, under constant monitoring. Their vital signs are being monitored on a 24-hour basis by sensors which measure their heart rate, arterial blood pressure, intracranial pressure, respiratory trace, temperature and other biosignals. The monitoring equipment used to capture these biosignals is highly specialised and consists of various devices such as electroencephalography (EEG) and electrocardiography (ECG) monitors which record a patient's brain and heart activity, intracranial pres-

sure monitors that (invasively) measure the pressure that develops inside a patient's skull, a ventilator that provides oxygen and mechanically assists patients' breathing etc. and a bedside monitor which displays all of the aforementioned biosignals. Also, a number of intravenous pumps are used to deliver medication to the patient in a controlled manner. The sampling rate of these monitors can be high (e.g. 125 Hz) which means that the signals have a very rich structure at fine resolution levels. We refer to the task of analysing these biosignals for the purposes of monitoring a patient's physiology as *physiological condition monitoring* (or simply *condition monitoring*).

1.2.1 Clinical need

Physiological condition monitoring of patients in ICUs is of critical importance as their health is subject to a number of potentially fatal physiological episodes which include but are not limited to events such as hypotension, bradycardia, atrial fibrillation, sepsis etc. Therefore, correctly identifying such events can trigger an alarm so that clinical personnel can tackle the situation appropriately.

Additionally, a variety of artifactual processes are present in such settings where the clinical personnel interact with the patients in various ways (e.g. taking blood samples, performing suctions, recalibrating sensors etc.). This frequent interaction coupled with the simplicity of rules behind most ICU alarm systems, which often utilise simple magnitude thresholds to indicate important changes on the state of a patient's health as mentioned in Chambrin *et al.* (1999), results in a very high false alarm rate which can fluctuate between 72% to 99% according to Sendelbach and Funk (2013). This leads very frequently to the development of *alarm fatigue*. Alarm fatigue is defined in that same work as a "sensory overload when clinicians are exposed to an excessive number of alarms, which can result in desensitization to alarms and missed alarms". More importantly even "patient deaths have been attributed to alarm fatigue". This is recognised as a major issue in ICU operation and measures that can lead to a considerable decrease of false alarms need to be taken.

Furthermore, frequent interactions result in the observed data being "contaminated" with artifacts that can potentially interfere with the task of modelling critical events that are solely due to the physiology of the patient (e.g. hypotension or bradycardia) and not due to some external interference (e.g. blood sample). In addition, precise modelling

of artifactual processes can indirectly increase a model's performance on the inference of physiological patterns. This occurs because a large number of potential training data for physiological patterns need to be discarded due to the presence of artifact. For example, a high percentage of potential hypotension episodes end up being discarded if any of the observed channels' values lie outside the range of acceptable physiological values. However, the latter is a usual case when an artifactual process occurs, thus causing clinicians to discard a potentially valid training instance. More specifically, in Tarbet, 2012 it was shown that out of 3158 potential hypotension events (collected from a cohort of 256 patients), 1174 (27%) had to be rejected due to the fact that the recorded blood pressure was found to be outside the physiologically acceptable values, most likely due to an artifactual process such as the taking of a blood sample. Correctly identifying such artifactual process and at the same time providing an estimate of the underlying physiological process could result in "salvaging" many training instances.

1.3 Contributions

This thesis aims to develop models that could be used to tackle the issues identified in the previous section. In summary, we aim to develop a model that will be able to:

i) Identify artifactual events in order to reduce false alarms.

ii) Track the underlying physiology of a patient in the presence of artifact.

iii) Identify events of physiological importance.

To that end, the following contributions have been made:

a) The Discriminative Switching Linear Dynamical System (DSLDS) addresses all three of the above points and was first published in Georgatzis and Williams (2015). The DSLDS is described in Chapter 4.

b) The Input-Output Non-Linear Dynamical System (IO-NLDS) tackles point iii) above and improves on the modelling task at hand by incorporating the effects of drug infusions, in contrast to the DSLDS which does not take those into account. The IO-NLDS was first published in Georgatzis *et al.* (2016b) and is described in Chapter 5.

c) Finally, previous work on Factorial Switching Linear Dynamical Systems which was conducted by Quinn *et al.* (2009) in the context of NICUs, was extended to adult Neuro-ICUs by Georgatzis *et al.* (2016a).

The ultimate goal of this project is to implement an efficient and effective model which could be used as a reliable decision support system by medical staff. We hope that by accurately modelling various processes that occur in the Neuro-ICU environment we could aid in the reduction of uncertainty and in the increase of predictive capability when medical decisions are being made. This could lead to a decrease in the rate of poor patient outcomes and thus the author believes that it is a project with a deeply beneficial prospect.

1.4 Structure of thesis

The remainder of this thesis is structured according to the following plan: Chapter 2 provides an exposition of the datasets that will be used in this thesis and contains details about the acquisition and preprocessing stages, preceding their usage in subsequent chapters. Chapter 3 provides an overview of the relevant literature and work that formed the basis of this thesis both in terms of methodologies and applications. In Chapter 4, the first research contribution of this thesis is described via the presentation of the DSLDS. Chapter 5 contains the details about the second research contribution of this thesis which is comprised of the formulation, development and validation of the IO-NLDS. Finally, Chapter 6 concludes by summarising this work in light of the presented material and includes suggestions for future work.

Chapter 2

Data

The purpose of this chapter is to provide a description of the data that will be used in subsequent chapters in order to achieve the tasks set out in the introductory chapter. Data constitute the driving force behind any machine learning algorithm and indeed provide the testbed for any proposed model. Most modelling assumptions require a large quantity of data and the models' outputs tend to become more reliable as the number of datapoints increases. In the context of physiological condition monitoring, datasets comprise primarily of patients' vital signs, which reflect the evolution of a patient's physiology across time. Additionally, in some cases, annotations are provided by expert clinicians with respect to clinical events of interest. Thus, in those cases, vital signs are accompanied by *labels* which leads to the task of *supervised learning*. Pertaining to the sensitive nature of medical data, it is not as straightforward to obtain high quality data from a large cohort of patients. However, it is reasonably feasible to obtain a considerable amount of data per patient in most cases.

Vital signs belong to a category of measurements that are indexed by time with successive observations usually exhibiting high correlations. These observations constitute what is known as *sequential data* or *time series*, which describes their temporal character, while the less descriptive term, *signal*, is also frequently used. All these terms will be used interchangeably within this thesis.

The remaining structure of this chapter is as follows: In Section 2.1 a brief description of the various collected vital signs is given. In Section 2.2, details about the various data sources and the data preprocessing stages are given. In Section 2.3, a description of

clinical events of interest is given which are broken down into: i) artifactual events and ii) physiological events. In the same section, we describe the process that was followed in order to annotate the aforementioned events. Finally, this chapter concludes with a brief summary of the presented material in Section 2.4. Parts of this chapter are based on Lal *et al.* (2015).

2.1 Vital signs

Vital signs are defined as signs of life and are used as indicators of a patient's condition. These include various measurements such as blood pressure, heart rate etc. which collectively characterise the patient's observable physiological status. Since the true physiology of a patient is always latent and cannot be observed directly, these vital measurements are the basis upon which all inferences regarding a patient's physiology must be drawn. What follows is a brief description of the vital signs used within this thesis.[1] Since they are collected from various hospitals and can refer to either adults or neonates, normal ranges can vary depending on each case. Also, some vital signs are only collected for adults, whereas others are only collected for neonates (this pertains to the datasets used in this thesis and is not a general remark).

Blood pressure

Blood pressure (BP) is defined as the pressure exerted by the blood against the walls of the blood vessels, especially the arteries (arterial BP) and it depends on the strength of the heartbeat, the elasticity of the arterial walls, the volume and viscosity of the blood, and a person's health, age, and physical condition. Blood pressure is measured in millimeters of mercury (mmHg) and is usually expressed in terms of systolic (maximum) BP (BPsys) and diastolic (minimum) BP (BPdia) which are acquired during the duration of a cardiac cycle. Their (weighted) average is also frequently used and referred to as mean BP (BPmean). BPsys is measured during systole, when the ventricles contract, pumping blood into the arteries and BPdia is measured during diastole, when the ventricles relax and the heart fills with blood. Normal resting BP ranges for adults are 90–119 mmHg for BPsys and 60–79 mmHg for BPdia, according to the American Heart Association, while

[1] Definitions were taken from `http://medical-dictionary.thefreedictionary.com/`.

for neonates the corresponding ranges are 27–55 mmHg and 23–43 mmHg as mentioned in Quinn (2007, table 2.2).

Heart rate

Heart rate (HR) is defined as the number of heartbeats per unit of time, and is most often measured in beats per minute (bpm). The normal resting HR range for adults is 60–100 bpm according to the American Heart Association, while for neonates it is 119–175 bpm according to Quinn (2007, table 2.2).

Intracranial pressure

Intracranial pressure (ICP) is defined as the pressure of the cerebrospinal fluid in the subarachnoid space, which is the space between the skull and the brain. ICP is measured in millimeters of mercury (mmHg) and its normal resting range is at 7–15 mmHg for an adult in the supine position as suggested by Steiner and Andrews (2006). Similarly to the BP signal, ICP can be expressed in terms of systolic ICP (ICPsys) and diastolic ICP (ICPdia) depending on the phase of the cardiac cycle (systole or diastole). ICP is only collected in the case of adults. In this thesis we focus on ICPsys and thus the abbreviations ICP and ICPsys will be used interchangeably.

Respiratory rate

Respiratory rate is defined as the frequency of breathing, usually expressed in breaths per minute. Its normal resting range for adults who are spontaneously breathing is 12–20 breaths per minute according to Ganong and Barrett (1995). Patients admitted in the ICU most often receive mechanically assisted ventilation and thus their respiratory rate is decided by the clinical personnel. The respiratory rate is only collected in the case of adults.

Bispectral index

The bispectral index (BIS) is a statistically derived, dimensionless scale that measures depth of anaesthesia and can take values from 0–100. It is a weighted sum of several features derived from the electroencephalogram (EEG), including time domain, frequency domain, and higher order spectral features. A BIS of 100 means that the patient is fully

conscious, while zero values correspond to EEG silence. A value of 40–60 is considered to be within the appropriate range for general anaesthesia. The exact algorithm used for the calculation of BIS is proprietary but an overview of the methodology is as follows: First, the digitised EEG is filtered to exclude low and high frequency artifacts and is further divided into two-second segments. Next, a set of artifacts, such as electrocardiogram (ECG) and/or pacer spikes, eyeblink events, low-frequency electrode noise etc. are detected and removed. After the EEG signal is deemed artifact-free, two different algorithms, called burst suppression ratio (BSR) and QUAZI, are used to calculate the degree of burst supression[2] observed in the EEG segments. After those two time-domain features, another two, frequency-domain features are calculated, called BetaRatio and SynchFastSlow. BetaRatio is the log ratio of power in two empirically chosen frequency bands: $\log((P_{30-47_{Hz}})/(P_{11-20_{Hz}}))$ and is computed on the Fourier transform[3] of the original EEG signal. The SynchFastSlow feature is the log ratio of the sum of all bispectrum peaks in the band from 0.5–47 Hz over the sum of the bispectrum in the band from 40–47 Hz, where the bispectrum is a higher-order spectrum (compared to the power spectrum one obtains from the standard Fourier transform). The resulting index is then defined as a proprietary combination of those four features. More details about BIS can be found in Rampil (1998). BIS is only collected in the case of adults.

Partial pressure of gases

Various gases are dissolved in a patient's blood, and the partial pressure is used to measure the amount of each. It is defined as the amount of pressure that a particular gas would exert on a container if it was present without the other gases. In the case of neonates a transcutaneous probe, placed on the chest, measures the concentration of oxygen ($TcPO_2$) and carbon dioxide ($TcPCO_2$) in capillary blood underneath the skin by heating the skin to improve perfusion and measuring how much of each gas is emitted through the skin electrochemically. Partial pressures are measured in kiloPascal (kPa) and normal ranges for neonates are 5.5–13.5 kPa and 2.6–7 kPa for oxygen and carbon dioxide respectively. These measurements are only considered in the case of neonates for the purposes of this thesis.

[2]Burst suppression is a distinctive EEG pattern, during which periods of normal to high voltage alternate with periods of low or even zero voltage.

[3]The Fourier transform is a way of transforming a time series from the time domain to the frequency domain. A detailed exposition of frequency-domain (also known as spectral) methods is outside the scope of this thesis but more details can be found in Brigham (1988).

Oxygen saturation

A pulse oximeter, attached to the foot of a neonate, measures the saturation of oxygen (SpO_2) in arterial blood. SpO_2 is defined as what proportion of the blood's capacity to carry oxygen is being utilised and is expressed as a percentage. In the case of neonates, the pulse oximeter shines two wavelengths of light through the neonate's foot into a photodetector which measures the changing absorbance at each of the wavelengths, allowing it to determine the absorption spectrum of oxygenated hemoglobin due to the pulsing arterial blood. These measurements are only considered in the case of neonates for the purposes of this thesis.

Temperature

The core body temperature (CT) and peripheral temperature (PT) are measured by two probes, one of which is placed under the neonate's back and the other is attached to a foot. Temperature is expressed in degrees Celsius (°C) and normal ranges for neonates are 35.7–38.5 °C for CT and 34.6–38.0 °C for PT. These measurements are only considered in the case of neonates for the purposes of this thesis.

Resting values

It should be noted that the provided normal resting value ranges should not be expected to always represent typical values in an ICU environment. In an ICU it is not straightforward to provide absolute normal values for vital signs since a) there is baseline variability in the population (e.g. a given BPsys measurement may be normal for a young woman but severely hypotensive for an older man) and b) patient's physiology is often aggressively manipulated. For example a given BPsys measurement which would be considered very high under some circumstance might be regarded normal if the patient is under the influence of a vasoconstrictor[4]. To some degree, an indication of normal value ranges in an ICU environment can be obtained via the inspection of "early warning scores" definitions, such as the National Early Warning Score (NEWS) as formulated by Smith *et al.* (2013), where a score of 0 indicates normality, while higher scores indicate increasing levels of abnormality/criticality of a patient's physiology.

[4]Vasoconstrictors are medications causing vasoconstriction, which is the narrowing of the blood vessels resulting from contraction of the muscular wall of the vessels.

2.2 Data acquisition and preprocessing

The data have been mainly obtained at two sites: a) the Southern General Hospital (SGH) in Glasgow and the Golden Jubilee National Hospital (GJNH) in Clydebank. We note that later in this thesis, another dataset collected from 15 neonates from the Edinburgh Royal Infirmary (ERI) will be used.

Data acquisition and distribution at a large scale for medical data is considerably restricted by regulations that govern the handling of data of such a sensitive nature. All data used in this thesis have undergone through the appropriate patient consent forms and approval requests, and furthermore they have been deidentified prior to research use. It should also be noted that the collected data are the product of copious and continuous work involving multiple parties for a period of approximately three years.

More specifically, the SGH dataset is part of routinely collected clinical data and none of the administered clinical care was adapted on the basis of this collection proces. Therefore, no formal approval from the Research Ethics Service was required but instead approval was provided by the local Caldicott Guardian[5]

The GJNH dataset was obtained from patients who were formally consented for an independent clinical research project. Patients were approached during their pre-operative clinic attendance and provided with a written and verbal description of the study procedure. They were either consented at this stage or allowed further time to consider their involvement prior to attendance on the day of surgery. The analysis carried out in the context of this thesis falls within the objectives of the original clinical study.

SGH

Data from seven traumatic brain injury (TBI) and two subarachnoid haemorrhage (SAH) adult patients were collected in the Neuro-ICU of the SGH. An average of 33-hour periods were collected from each of these patients. The collected signals were arterial blood pressure (ABP), ECG, pulse oximetry pulse, intracranial pressure (ICP), end tidal CO2 (EtCO2) and the respiratory signal (Resp), from the patients' bedside Philips Intellivue monitors. The ECG signal was sampled at a frequency of 500Hz, while the remaining signals were sampled at the lower frequency of 125 Hz. For seven out of the nine patients, the waveform data were recorded into a laptop computer which was connected to the

[5]More details can be found in https://www.gov.uk/government/groups/uk-caldicott-guardian-council.

bedside monitor. The remaining two patients' data were recorded and transmitted via the waveform capture software ixTrend from ixellence GmbH (2015). This software captures waveform data from all available channels and transmits them on a minute-by-minute basis via a local area network to a SQL database that is hosted on a local server.

The waveform data, which are sampled at a high frequency of 125/500 Hz, are used as source signals in order to produce second-by-second summary values of vital signs. As it has been shown in Quinn (2007), this temporal resolution is more than adequate for the purposes of physiological condition monitoring. The approach of deriving summary measures from high-frequency waveform data is described in detail in Shaw (2013) which in turns makes use of the methodology described in Clifford (2002). The key idea behind the approach is that for each of the aforementioned signals, an *index* channel is selected first. The purpose of the index channel is to facilitate the identification of a physiologically meaningful interval within which the original signal can be then measured. For example, the ECG is used as the index channel for ABP. In order to derive second-by-second ABP measurements, the ECG is processed in order to identify the R–R intervals[6] and then the original signal is normalised, filtered (according to a band-pass filter with a 5–15Hz band) and is subsequently differenced and squared. Finally a running average of the transformed signal is calculated and the results are resampled to 1 Hz. For example, in the case of the ABP signal, the systolic, mean and diastolic channels are obtained per R–R interval and are then resampled.

GJNH

In this case, data from 40 adult patients were collected in the Neuro-ICU of the GJNH. The collected signals were systolic, mean and diastolic BP, HR and Resp. Additionally, the BIS signal was collected for 27 out of the 40 patients. The data were recorded into a laptop computer which was connected to the bedside monitors and they were subsequently transmitted to a remote server hosted by a third party. All derived vital signs are available at a sampling period of $T_s = 15$ seconds, with the exception of BIS which was sampled every 5 seconds and was subsequently downsampled to 15 seconds. Additionally, drug infusion rates were recorded from the bedside drug infusion pumps. The sampling frequency varies, with higher frequency sampling occurring during periods of rapid infusions (with highly variable infusion rates), while lower frequency sampling took

[6] An R–R interval is the interval between two successive ventricle depolarisations in the heart.

place while the infusion rate was more constant. These infusion rates were downsampled to 1 Hz for the rapid-infusion stage and linearly interpolated to 1 Hz for the slow-infusion stage, and were subsequently downsampled to 15 seconds.

Only a subset of the vital signs that were collected from those two sites will be used in subsequent chapters. A list of those is given in Table 2.1 for reference purposes.

Table 2.1: List of captured/derived channels from SGH and GJNH that will be used later in this thesis.

Vital sign	Abbreviation
Blood pressure	BP
Heart rate	HR
Intracranial pressure	ICP
Bispectral index	BIS

ERI

Data comprising 24-hour periods from 15 neonates of the NICU of the ERI have been collected for the purposes of the analysis conducted in Quinn (2007) and will be used for the purposes of this thesis as well. In the case of ERI, all signals are collected by the Siemens SC7000 patient monitor (Dräger Medical, 2007). ECG measurements are recorded at a sampling rate of 225Hz and heart rate measurements are obtained based on an average of identified beats in the preceding 15 seconds. ABP is measured internally by the monitor at a maximum frequency of 32Hz. When ECG measurements are not available, the ABP waveform is used to calculate the heart rate as a smooth periodic signal with peaks produced by the contraction of the heart muscle. A number of additional vital signs, namely $TcPO_2$, $TcPCO_2$, SpO_2, PT and CT are collected as described in Section 2.1.

2.3 Events of interest

The collection of vital signs serves the purpose of identifying various events of clinical interest. These are decided after close consultation with clinical experts and are chosen so

as to include events that are considered to be of major concern in the quest for effective physiological condition monitoring.

2.3.1 Artifactual events

Artifactual events refer to changes in the observed vital signs of a patient that are not associated with a change in their physiology but are rather the product of an artifactual process, such as e.g. the taking of a blood sample. Since, by definition, in ICUs there is a high degree of interaction with the patient by the clinical personnel, artifactual events are a very common occurrence. As such, they contribute considerably to the high false alarm rate observed in ICUs and to the corresponding alarm fatigue as described in Section 1.2. What follows is a description of the most commonly occurring artifacts in the Neuro-ICU of the SGH.

Blood samples

For the accurate modelling of blood sample events, BPsys and BPdia are a sufficient set of channels. Blood pressure in adult patients is measured with the help of an arterial line (A-line) which is usually inserted into the radial artery of the patient. The A-line is connected with a transducer which converts the arterial blood pressure at one end to the actual readings on the monitoring device at the other end. A three-way tap is connected in the A-line which can be set in one of three modes: a) open to the arterial line (for measuring the arterial blood pressure) b) open for taking blood samples and c) open to the air for calibration purposes.

Blood samples are taken from patients in a Neuro-ICU on a regular basis as a way of keeping track of the physiological indices (i.e. quantities measured through blood tests such as the concentration of C-reactive protein etc.) of the patient. Thus, a long period of channel readings can always be expected to contain a significant amount of blood sample events. There are four steps that are followed when a blood sample is being taken: 1) the three-way tap is turned so that the blood can be diverted to a syringe for the needs of blood sampling. During this stage the patient is disconnected from the transducer and a saline pump acts against it, causing an artifactual ramp to the BP measurements. 2) The transducer is exposed to the air for calibration purposes, causing measurements to drop to zero. 3) The tap is turned to its original configuration and BP measurements continue

as normal. 4) The tap is flushed with a pulse of saline solution to remove blood residues and avoid infection, causing a sharp increase in measurements. A state transition diagram describing this process is shown in Figure 2.1, while two representative examples are shown in Figure 2.2.

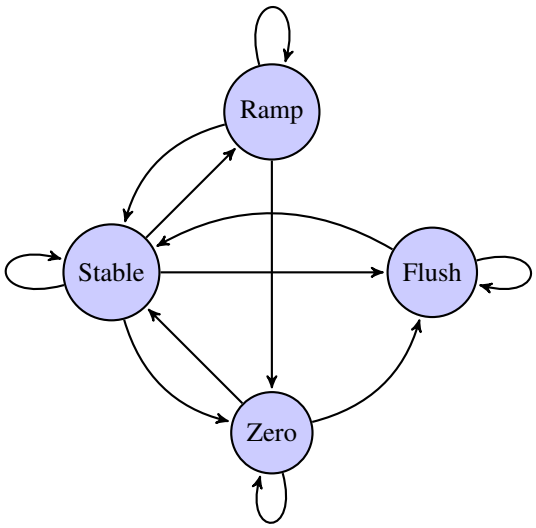

Figure 2.1: State transition diagram of a blood sample event's stages. In practice, all transitions are allowed.

Suctions

Another very frequently occurring event during the hospitalisation of a patient in a Neuro-ICU is that of endotracheal suctioning. During that event an endotracheal tube is inserted in the airway of the patient with the aim of removing pulmonary secretions that have accumulated over time in the patients pulmonary system. Its high frequency is due to the high importance of keeping the pulmonary system of a patient free of secretions. Suctioning is usually combined with a sharp increase in the values of (almost) all observed channels. It can also induce a coughing episode in the patient if the endotracheal tube comes in contact with the carina (the lowest part of the trachea), which causes a reflexive reaction that leads to coughing. A coughing episode is reflected in the observed channels in a very similar way to suctioning and the two events are not easily distinguishable. Two representative examples of suction events are shown in Figure 2.3.

Damped traces

Another frequently occurring artifactual event in the Neuro-ICU is a damped trace event.

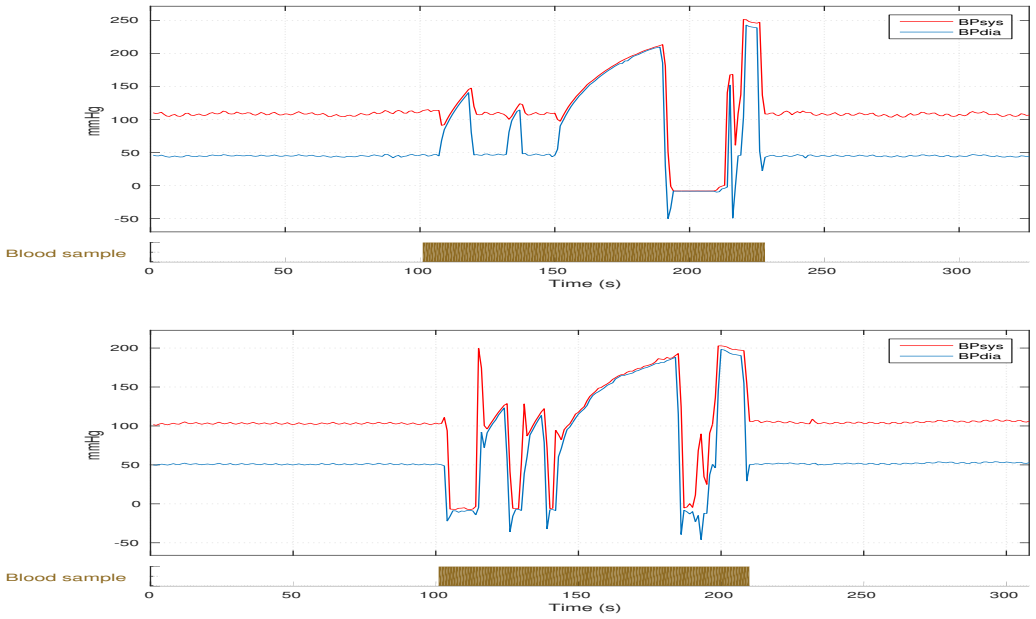

Figure 2.2: Two examples of blood sample events with the annotated periods in gold colour. Blood samples affect only the BP channels.

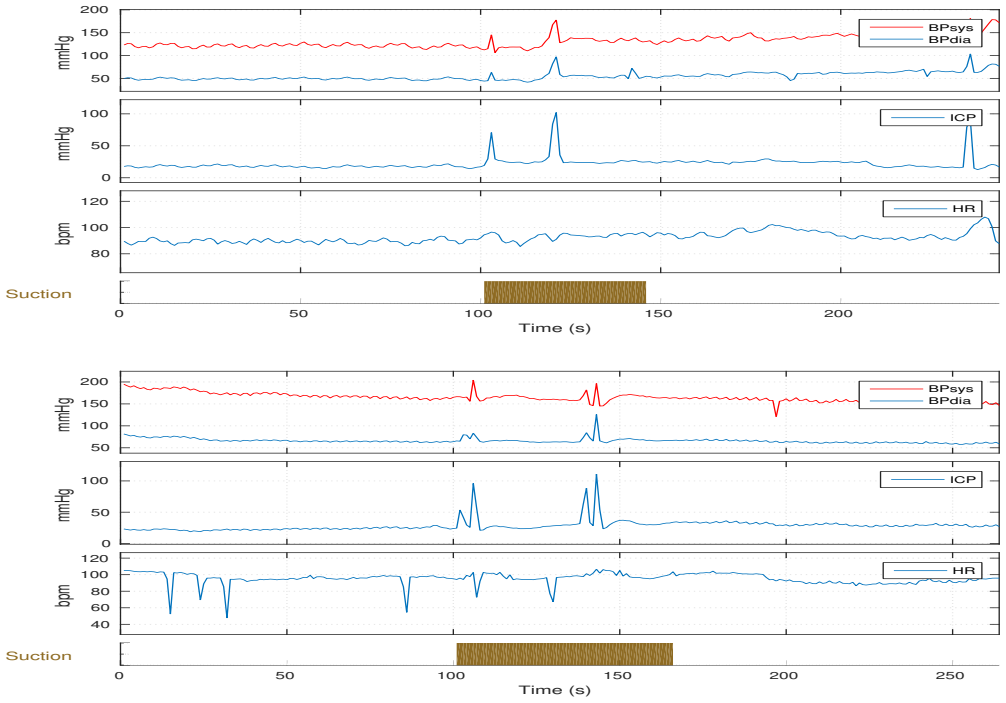

Figure 2.3: Two examples of suction events with the annotated periods in gold colour.

A damped trace occurs when blood residues are accumulated over time within the arterial line causing its occlusion, which in turn leads to a damping of the observed BP signals. The BPsys and BPdia gradually converge to similar values, thus obscuring the BP corresponding to the patient's stable condition. This damping is usually resolved when noticed by a nurse via a flushing of the arterial line which clears it from the blood residue buildup. Since a damped trace can remain unnoticed for several hours at a time (e.g. during a night shift), it can have a considerable contribution to the artifact contamination of the recorded values. Two representative examples of damped trace events are shown in Figure 2.4.

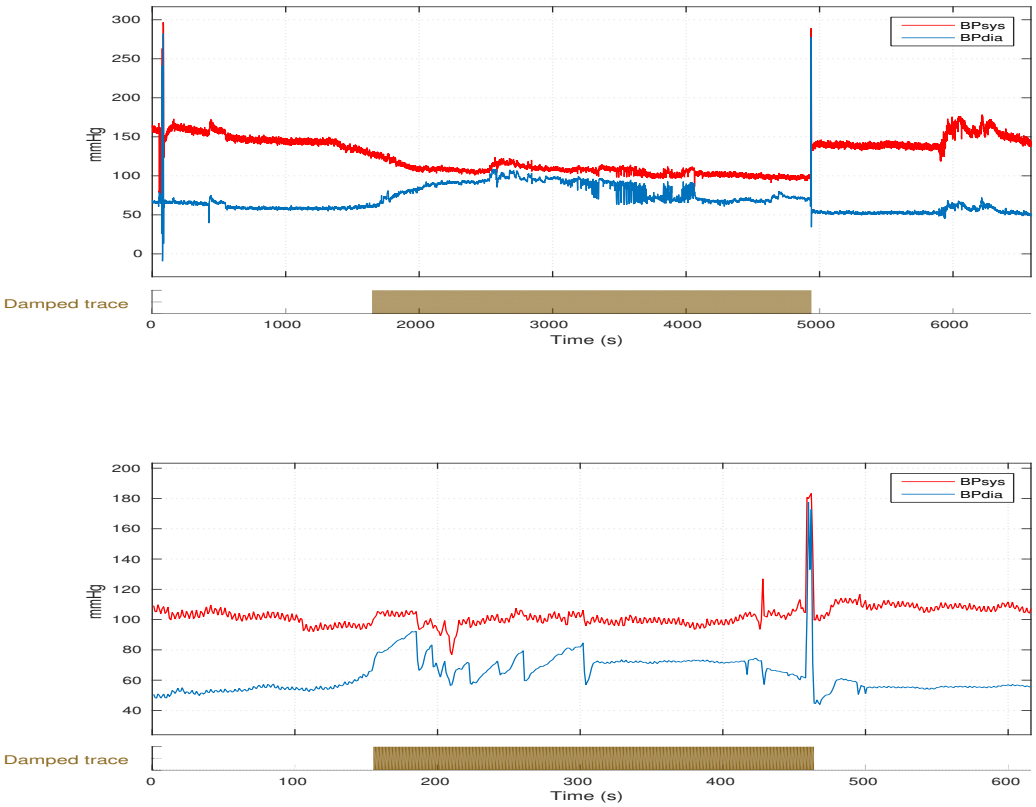

Figure 2.4: Two examples of damped trace events with the annotated periods in gold colour. Damped traces affect only the BP channels.

Other

A plethora of other artifactual events have also been annotated. However their frequency and duration compared to the already mentioned events were not deemed high enough

to be included separately into the modelling process within this thesis. These artifacts include events such as patient movement/turning, hygiene (oral, eye care), physiotherapy, clinical examination, change of ventilator circuit, generic signal deterioration and other minor events. These were collectively assigned to a special category and will be modelled accordingly in later chapters. These "other" artifactual events are collectively called the "X-factor" category and two representative examples are shown in Figure 2.5.

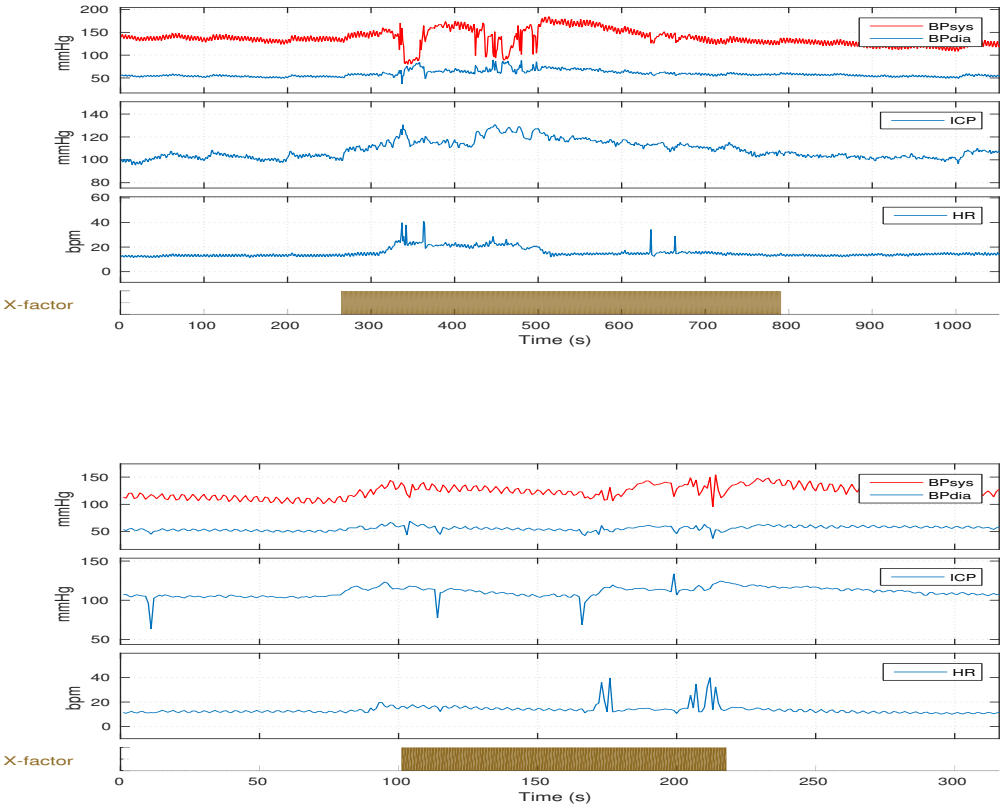

Figure 2.5: Two examples of abnormal events (X-factor) with the annotated periods in gold colour.

As mentioned already, data from the ERI were also used. These included annotated artifactual events which included: a) blood sample events, b) disconnections of the temperature probe, c) openings of the neonates' incubator and d) other artifactual events (X-factor). These are described in detail in Quinn (2007, sec. 2.3).

2.3.2 Physiological events

Physiological events refer to changes in the observed vital signs of a patient that reflect a change in their underlying physiology and can be associated with potentially critical situations. A patient admitted in an ICU is by definition in a critical condition, but shortly after admission a stabilisation of the patient's condition is usually achieved. This stable condition is usually desirable as it allows for the gradual healing of the patient in the absence of other complications.

Stability

Stability characterises the physiology of a patient when the patient is stable, and no artifactual/physiological factors are in effect. These periods are important for our application because they provide the baseline with respect to which the models for the artifactual/-physiological factors will be fitted. An example of stable physiology is given in Figure 2.6.

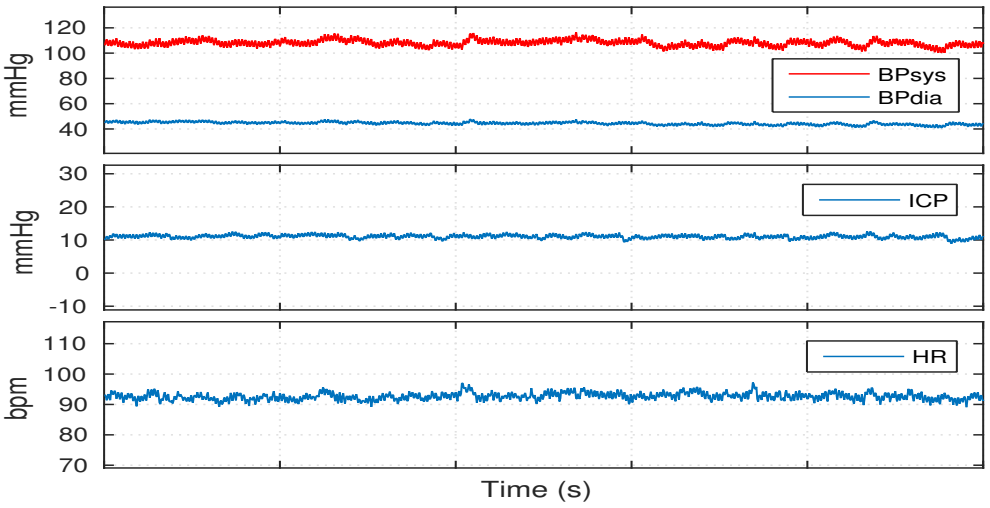

Figure 2.6: Example of a stable patient's vital signs.

Drug effects

The physiology of patients admitted in ICUs is expected to be affected by the drugs which are administered to them. The collected drug infusions from the GJNH correspond to drug dosages of the anaesthetic drug Propofol. Propofol is a sedative and hypnotic drug with rapid onset and short duration of action, which is used intravenously for induction and maintenance of general anesthesia. It thus has the direct physiological effect of increasing a patient's depth of anaesthesia (translating into a decrease of BIS) and the side effect of also decreasing a patient's blood pressure. This latter side effect can sometimes lead to episodes of hypotension (see next paragraph). A representative example of observed vital signs under the effect of a Propofol infusion are shown in Figure 2.7.

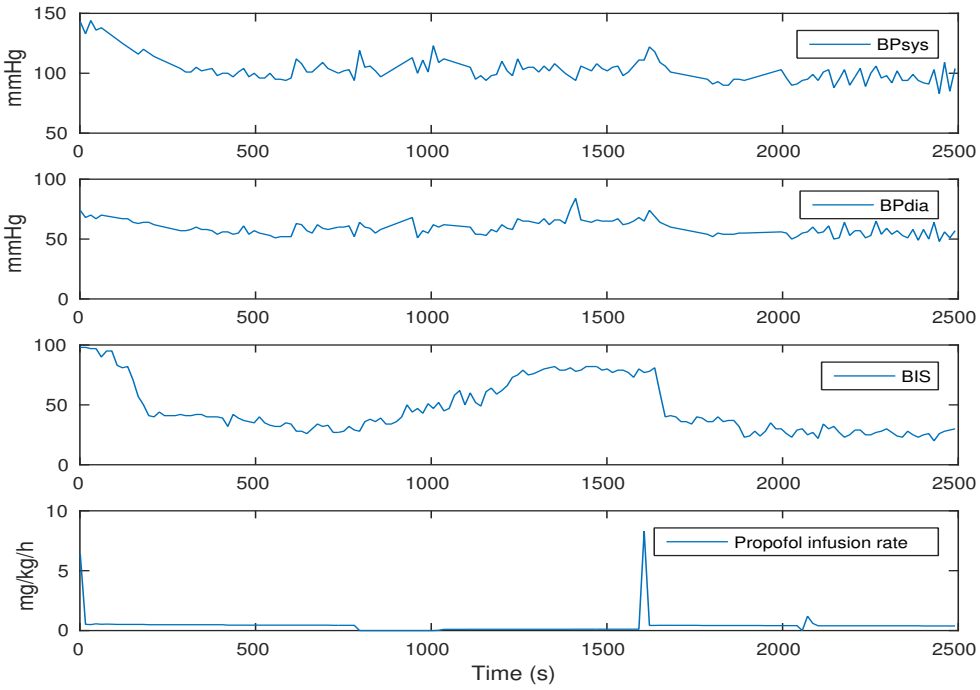

Figure 2.7: Example of vital signs under the effect of drug (Propofol) administration.

Hypotension

Hypotension events will not be directly modelled in the context of this thesis but can be accounted for indirectly by modelling drug effects. According to the EUSIG definition (Jones *et al.*, 1994), a hypotension event refers to severely low BP (BPsys < 90 mmHg or BPmean < 70 mmHg) for at least five minutes and is associated with adverse patient outcome and increased mortality. Thus, averting such episodes could be highly beneficial

to a patient's physiology. Direct modelling of hypotension events is not of interest as they are algorithmically defined and thus identification of such events is trivial. However, predicting their onset is important (see e.g. Donald, 2014) but not within the scope of this thesis.

Bradycardia

Bradycardia is the only known physiological factor which will be explicitly modelled within this thesis (since it was the only one that annotations were available for) and is encountered in the case of the ERI NICU. It refers to a considerable decrease in the heart rate and short incidences can be frequent in neonates. Its severity can vary from non-concerning to critical. More details are provided in Quinn (2007, sec. 2.4.1).

Other

Finally, as in the case of artifactual events, there is a variety of physiological events that are annotated by expert clinicians but are not explicitly modelled due to their rarity[7]. These are treated as a single category (X-factor) and include physiological events such as tachycardia (or tachyarrhythmia), atrial fibrillation, painful stimuli etc.

2.3.3 Data annotation

Data were annotated in two stages in the case of the SGH. First, start times of artifactual events were noted by the nursing staff and/or clinical experts at bedside. Then, a single expert clinician would retrospectively annotate the high frequency data using the bedside notes of start times of events to pinpoint exactly the intervals of artifactual patterns that were spotted in the patients' measurements. During this second stage, the clinician also provides annotations of physiological processes that manifest themselves as abnormal traces in the high frequency data. In the case of the ERI dataset, fifteen (one for each neonate) 24-hour periods were annotated by two clinical experts. Near the start of each period, a 30 minute time segment was used as an example of stability. Each of a set of known physiological and artifactual patterns of interest were then also annotated. Initial estimates as to the durations of these events were made by the author and were subsequently audited and amended as necessary by the clinical experts. The annotation process

[7] These events are deemed rare in the context of the datasets used in this thesis.

is described in more detail in Quinn (2007, sec. 7.1). No expert clinician annotations were provided for the GJNH dataset.

2.4 Summary

In this chapter a description of the various datasets that will be modelled in subsequent chapters has been given. Brief definitions of the various collected vital signs have been given and a description of the mechanisms that were used to collect, preprocess and annotate the available datasets was subsequently presented. Furthermore, descriptions and examples of clinical events of interest that will be the focus of the following chapters were given. The focus was centred around a limited number of clinical events that were deemed important by clinician experts. Also, the exposition was primarily based on the data collected from the SGH and GJNH as this thesis analyses these datasets for the first time. A more concise description was presented in the case of the ERI dataset as this has been described in detail in Quinn (2007).

Chapter 3

Background & related work

This chapter contains an overview of the work which the material presented in Chapters 4 and 5 is based and builds upon. Firstly, we introduce some basic models in Section 3.1, which constitute components of more complex ones presented later or serve as simple baselines for comparison purposes. Secondly, in Section 3.2, the focus turns on models that are developed explicitly for modelling sequential data and thus comprise a more natural approach to this thesis' goals. Furthermore, in Section 3.3 a review is conducted on models that have been developed in relevant sub-fields for the purpose of modelling the behaviour of drugs after their administration into a patient's body. Section 3.4 reviews prior work which has been concerned with the application of machine learning methods for the task of physiological condition monitoring and finally Section 3.5 concludes with a summary of this chapter.

3.1 Basic models

In this section we present a brief overview of a series of models for regresssion and classification which will be used as components of the more complex models presented in subsequent chapters. These models do not explicitly model the temporal continuity underpinning the processes that give rise to sequential data. Rather they rely on the assumption that the modelled data are drawn independently from the same distribution, which are also known as independent and identically distributed (i.i.d.) data. Such an assumption might be restrictive when sequential data are the object of modelling, but nonetheless

these models can still be of considerable use, especially when coupled with time-series models and/or when the temporal aspect can be captured implicitly.

3.1.1 Linear regression

Perhaps the simplest model that can be used for regression is linear regression. According to this model, each observation $y \in \mathbb{R}$ can be modelled as a linear combination of a number of input variables $\mathbf{x} \in \mathbb{R}^{d_x}$ such that a prediction made by the model is of the form:

$$\hat{y} = w_0 + \sum_{j=1}^{d_x} w_j x_j \ , \tag{3.1}$$

where $\mathbf{w} = [w_0 \ w_1 \ \dots \ w_{d_x}]^\top$ denote weights to be estimated. The above equation is usually presented as $\hat{y} = \mathbf{w}^\top \mathbf{x}$ for compactness where an extra element equal to 1 is prepended to the input vector so that $\mathbf{x} \in \mathbb{R}^{d_x+1}$. Assuming we have a set of training data $\{(\mathbf{x}_i, y_i), ..., (\mathbf{x}_N, y_N)\}$, the parameters of the model are usually fitted via the *least squares* method which minimises the residual sum of squares (RSS):

$$RSS(w) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \ . \tag{3.2}$$

Differentiating eq. 3.2 with respect to $\mathbf{w}$ and setting to zero yields the formula for estimating the weights $\mathbf{w}$. It can be shown that minimisation of the RSS is equivalent to maximising the likelihood function under the assumption that the residuals follow a Gaussian distribution, as shown e.g. in Bishop (2006, sec. 3.1.1). Linear regression is a very well studied model and one can find many relevant examples in the literature apart from the already cited reference; thus, further exposition of this model within this thesis is deemed unnecessary.

3.1.2 Logistic regression

The linear regression model can be adapted to classification tasks, giving rise to what is known as logistic regression. The same modelling idea as in linear regression forms the basis of logistic regression with the addition that the output of the model is non-linearly transformed so as to always reside within the $[0, 1]$ interval, which is more appropriate for

classification purposes. This is done as follows:

$$\hat{y} = \sigma(\mathbf{w}^\top \mathbf{x}) \,, \tag{3.3}$$

where σ is the logistic sigmoid function:

$$\sigma(x) = \frac{1}{1+\exp(-x)} \,. \tag{3.4}$$

The logistic sigmoid has the convenient property that it maps the whole real line to the $[0,1]$ interval and thus its output can be interpretted as the probability of an input belonging to a class or not (after an appropriate choice of a threshold). Parameter estimation cannot be conducted in closed form as in the case of linear regression but rather an iterative algorithm like *iteratively reweighted least squares* needs to be used. Further details can be found in Bishop (2006, sec. 4.3).

3.1.3 Decision trees

So far, the presented models could only model linear relationships between inputs $\mathbf{x}$ and output y. One way of relaxing this rigid assumption is via a decision tree. Decision trees recursively partition the input space into rectangular, axis-aligned regions and then fit a single model (e.g. a constant for regression or a majority vote for classification) within each region. As such, decision trees can model more complex input-output relationships. Further analysis is outside the scope of this thesis but more details can be found in e.g. Hastie *et al.* (2009, sec. 9.2).

3.1.3.1 Random forests

Decision trees can produce very noisy predictions and are also prone to overfitting when grown in an unbounded fashion. One way of combating both of those limitations is via the use of random forests, which were introduced by Breiman (2001). A random forest is an ensemble of decision trees, where each tree is trained on a bootstrapped sample of size N and for each node in the tree, a random sample of d_r ($< d_x$) variables is selected and the optimal variable/splitting point is selected amongst those d_r variables. After all trees of the forest are fully grown, a prediction is made either as the average of the tree predictions (for regression) or as their majority vote (for classification). More details are given in Hastie

et al. (2009, chap. 15). This bootstrap aggregation (called bagging) and random sampling of variables has the effect of reducing the variance of the ensemble's output and can lead to significantly better performance. Indeed, extensive benchmarks conducted in Caruana and Niculescu-Mizil (2006) and Fernández-Delgado *et al.* (2014) place random forests amongst the best performing models, while being almost hyperparameter-free[1].

3.2 Dynamical models

The models presented so far do not attempt to explicitly model the temporal aspect of a time-series. Our attention is now turned on dynamical models which attempt to capture explicitly the temporal behaviour of the observed data. These models also have a rich literature.

Graphical models

As this thesis is centred around probabilistic modelling, *graphical models* will be often used as a way of representing the probabilistic relationships within the presented models. A graphical model is a diagrammatic representation that merges elements from graph theory and probability theory into a single, compact form. A graph is comprised of nodes and edges connecting the nodes. In the probabilistic graphical model framework, nodes represent random variables and edges represent probabilistic dependences between these variables. Edges can be directed or undirected but we will only focus on directed edges in this thesis. The overall graph then describes the way in which the joint distribution over all random variables decomposes according to the dependences specified by the edges. In this thesis, shaded nodes will be used to denote observed random variables, while unshaded (white) nodes will be used to denote unobserved (latent) random variables. Furthermore, square nodes will denote discrete random variables, while circular nodes will be used for continuous ones. An excellent tutorial on graphical models is given by Heckerman (1998), while a very detailed textbook is provided by Koller and Friedman (2009).

[1] Increasing the number of trees does not result in overfitting and "using fully grown trees seldom costs much" as reported in Hastie *et al.* (2009, 15.3.4).

3.2.1 Autoregressive processes

We start with the presentation of what is considered a classical approach to time-series analysis. This approach was introduced by the seminal work of Box and Jenkins (1970). Other, introductory textbooks, include Chatfield (2003) and Diggle (1990). A more advanced exposition is given by Brockwell and Davis (2009).

One of the simplest time-series models is an autoregressive (AR) process. An AR process models an observation y at time t as a linear combination of a number of p previous observations. As such, the previous observations of the time-series regress on the current one and hence the term "autoregressive". The simplest AR process is an AR(1) process, such that $y_t = \alpha y_{t-1} + w_t$, where w_t's at different times are assumed independent with $w_t \sim \mathcal{N}(0, \sigma_w^2)$, and in general an AR($p$) process is given by:

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + w_t \ . \tag{3.5}$$

The graphical model of an AR(2) process is shown in Figure 3.1.

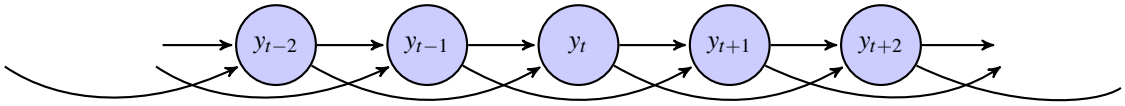

Figure 3.1: Graphical model of an AR(2) process.

3.2.2 Moving average processes

Another simple time-series model is a moving average (MA) process. An MA process models an observation y at time t as a linear combination of a number of $q+1$ previous and current white noise terms. In general an MA(q) process is given by:

$$y_t = \sum_{j=0}^{q} \beta_j w_{t-j} \ . \tag{3.6}$$

AR(p) and MA(q) processes can be combined into ARMA(p,q) processes.

3.2.3 ARIMA processes

An AR process assumes that the modelled time series is (at least weakly) stationary[2]. Real data rarely exhibit stationarity but a non-stationary process can be rendered (near) stationary via differencing. Applying then an AR process on differenced data yields:

$$y_t - y_{t-1} = \sum_{i=1}^{p} \gamma_i (y_{t-i} - y_{t-i-1}) + w_t \ . \tag{3.7}$$

As shown in Quinn (2007), this model can be applied to the initial, undifferenced data such that:

$$y_t = \sum_{i}^{p+1} \alpha_i y_{t-i} + w_t \ , \tag{3.8}$$

where $\alpha_1 = 1 + \gamma_1$, $\alpha_{i,1<i\leq p} = \gamma_i - \gamma_{i-1}$ and $\alpha_{p+1} = -\gamma_p$. The first-order differencing of eq. (3.7) can be generalised to d-order differencing, such that:

$$z_t = \nabla^d y_t = \sum_{i=1}^{p} \gamma_i z_{t-1} + w_t \ , \tag{3.9}$$

which provides an example of an autoregressive integrated moving average ARIMA process. Eq. (3.9) is an example of an ARIMA(p,d,0) process which denotes an AR(p) process on data that have been differenced d times (also denoted as ARI(p,d)). In general, a moving average (MA) component can be included in the construction of an ARIMA process but MA process are not investigated within the framework of this thesis.

3.2.4 Vector AR processes

So far, we have been restricted to y_t being a scalar. However, a vector $\mathbf{y}_t \in \mathbb{R}^{d_y}$ can be modelled as a vector AR (VAR) process such that:

$$\mathbf{y}_t = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \mathbf{G}\mathbf{w}_t \ , \tag{3.10}$$

[2] A stochastic process is *weakly stationary* if its mean is constant and its autocovariance depends only on the lag with respect to which it is computed.

where $\mathbf{A}_i, \mathbf{G} \in \mathbb{R}^{d_y \times d_y}$ and $\mathbf{w}_t \in \mathbb{R}^{d_y}$ is a white noise vector (i.e. $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$). It should be noted that an AR(p) process can be expressed as a VAR(1) process. For example, an AR(2) process can be written as a VAR(1) process in the following way:

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_t \\ w_{t-1} \end{bmatrix} \tag{3.11}$$

3.2.5 Linear dynamical systems

The models that have been described so far in this chapter assume that the processes are fully observed. Alternatively, if we assume that the real underlying temporal process is unobserved (latent) and we have only access to observations that are generated by those unobserved quantities, then we enter the realm of latent variable models (also known as state space models) where the latent space is considered to be a space of (unobserved) states. One of the most well-known latent variable models is the linear dynamical system (LDS), which assumes that the associated random variables are continuous, follow a Gaussian distribution and the temporal process is governed by linear relationships. The LDS was introduced by the seminal work of Kalman (1960) and has been since disseminated in several textbooks including the work of Harvey (1990), Harrison and West (1999), Grewal and Andrews (2001) and Durbin and Koopman (2012), while a very interesting exposition unifying several concepts under the linear and Gaussian assumptions (including LDSs) is given in Roweis and Ghahramani (1999). The LDS, a graphical model representation of which is shown in Figure 3.2, is defined by the following joint distribution:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}_1) p(\mathbf{y}_1 | \mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t | \mathbf{x}_{t-1}) p(\mathbf{y}_t | \mathbf{x}_t) \quad , \tag{3.12}$$

where $\mathbf{x}_t \in \mathbb{R}^{d_x}$ is the latent state vector at time t, which represents the underlying latent process and $\mathbf{y}_t \in \mathbb{R}^{d_y}$ is the observation vector at time t. The involved random variables are distributed according to:

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q}) \ , \tag{3.13}$$

$$\mathbf{y}_t \sim \mathcal{N}(\mathbf{C}\mathbf{x}_t, \mathbf{R}) \ , \tag{3.14}$$

where $\mathbf{A} \in \mathbb{R}^{d_x \times d_x}$ is the dynamics matrix governing the temporal evolution of the latent states which is driven by Gaussian noise with covariance $\mathbf{Q} \in \mathbb{R}^{d_x \times d_x}$ and $\mathbf{C} \in \mathbb{R}^{d_y \times d_x}$ is the observation matrix which projects the latent vector $\mathbf{x}_t$ into observation space corrupted by Gaussian noise with covariance $\mathbf{R} \in \mathbb{R}^{d_y \times d_y}$.

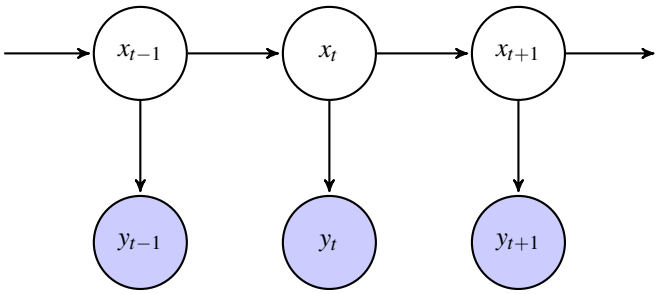

Figure 3.2: Graphical model of the LDS. The continuous latent state at time t is denoted by $\mathbf{x}_t$ (white circular nodes). The shaded circular nodes correspond to the observed variables, $\mathbf{y}_t$.

3.2.5.1 Inference in LDS

Since all relationships within the LDS are linear and governed by Gaussian noise, inference of the latent states, which amounts to computing the filtering distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, can be carried out in an exact manner using the Kalman Filter algorithm. Also, due to the Gaussian assumption, moments up to second order comprise a sufficient set of statistics for a full characterisation of the filtering distribution. Therefore, by representing $p(\mathbf{x}_t|\mathbf{y}_{1:t}) \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, inference translates into calculating $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$. Assuming that the parameters of the model $\{\mathbf{A}, \mathbf{Q}, \mathbf{C}, \mathbf{R}\}$ are known, exact inference can then be broken down into two steps: a) the prediction step and b) the update step, as shown e.g. in Murphy (2012).

Prediction step

The prediction step involves inferring the predictive distribution of the latent state at time t given observations up to time $t-1$ and is given by:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \sim \mathcal{N}(\boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}) \,, \tag{3.15}$$

$$\boldsymbol{\mu}_{t|t-1} = \mathbf{A}\boldsymbol{\mu}_{t-1} \,, \tag{3.16}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \mathbf{A}\boldsymbol{\Sigma}_{t-1}\mathbf{A}^\top + \mathbf{Q} \,. \tag{3.17}$$

Update step

The update step involves inferring the posterior distribution of the latent state at time t given observations up to time t and is given by:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \,, \tag{3.18}$$

which is equal to:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \,, \tag{3.19}$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \hat{\mathbf{y}}_t) \,, \tag{3.20}$$

$$\boldsymbol{\Sigma}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{C})\boldsymbol{\Sigma}_{t|t-1} \,, \tag{3.21}$$

where $\hat{\mathbf{y}}_t = \mathbf{C}\boldsymbol{\mu}_{t|t-1}$ is the predicted observation and $\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^\top(\mathbf{C}\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^\top + \mathbf{R})^{-1}$, is the Kalman gain matrix. It should be noted that for time-series forecasting, we usually need to compute the predictive posterior density for the observations which is given by:

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \sim \mathcal{N}(\mathbf{C}\boldsymbol{\mu}_{t|t-1}, \mathbf{C}\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^\top + \mathbf{R}) \,. \tag{3.22}$$

The predict-update cycle outlined above describes what is known as filtering and is an on-line inference algorithm in the sense that it can be performed in real-time. Since the focus of this thesis is on on-line physiological condition monitoring, inference will be (mainly) referring to filtering in the following chapters.

3.2.5.2 Learning in LDS

The previous section focussed on inference under the assumption that the parameters of the LDS were known. Apart from models that have very well established, underlying theoretical properties underpinning them, in most cases the set of a model's parameters will need to be determined under a parameter estimation scheme.

EM

The most common estimation algorithm for LDSs is the expectation-maximisation (EM) algorithm which provides maximum likelihood estimators (MLEs) for the model's parameters. EM is an iterative algorithm and was developed by Dempster *et al.* (1977) for deriving MLEs in the presence of missing data, but can be readily extended to latent variable models by treating the latent variables as missing data. The EM is an appealing learning method since it is numerically stable and also possesses the property that the likelihood is guaranteed to increase (or stay the same) per each iteration. A derivation of EM is given in Appendix A.

Spectral learning

An alternative method for learning a LDS is spectral learning, also known as subspace system identification in the control theory community (see e.g. Van Overschee and De Moor, 1996). Spectral learning has become increasingly popular in the machine learning community, being used for learning in a variety of models and is a well established learning method in the case of LDSs. The first step in this approach is to arrange the set of observations $\mathbf{y}$ in a specific matrix form, called block Hankel matrix:

$$\mathbf{H} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3 & \cdots & \mathbf{y}_\tau \\ \mathbf{y}_2 & \mathbf{y}_3 & \mathbf{y}_4 & \cdots & \mathbf{y}_{\tau+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_r & \mathbf{y}_{r+1} & \mathbf{y}_{r+2} & \cdots & \mathbf{y}_{r+\tau-1} \end{bmatrix}, \tag{3.23}$$

with $\mathbf{H} \in \mathbb{R}^{(r\times d_y)\times\tau}$. Matrix $\mathbf{H}$ is effectively comprised of stacked observation sequences with a "horizon" τ, where each subsequent sequence in the stack is shifted by one time unit. The key observation is that this matrix, which is comprised exclusively of observable quantities, can be decomposed according to the parameters of the LDS, revealing them in the process. The most commonly used decomposition is the singular value decomposition (SVD, see e.g. Press *et al.*, 1996, sec. 2.6). Decomposing $\mathbf{H}$ according to SVD yields $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{(r\times d_y)\times\ (r\times d_y)}$ is a matrix whose columns are the left-singular vectors of $\mathbf{H}$, $\mathbf{\Sigma} \in \mathbb{R}^{(r\times d_y)\times\tau}$ is a rectangular diagonal matrix whose entries are the singular values (i.e. the square root of the eigenvalues) of $\mathbf{H}$ and $\mathbf{V} \in \mathbb{R}^{\tau\times\tau}$ is a matrix whose columns are the right-singular vectors of $\mathbf{H}$. Estimates for the parameters can then be obtained as follows:

$$\hat{\mathbf{C}} = \mathbf{U} \quad \hat{\mathbf{A}} = \mathbf{X}_{1:\tau-1}\mathbf{X}^{+}_{2:\tau} \tag{3.24}$$

where $\mathbf{X} = \mathbf{\Sigma}\mathbf{V}^\top$, $\mathbf{X}_{1:\tau} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\tau] \in \mathbb{R}^{d_x\tau}$ and $\mathbf{X}^+$ is the Moore-Penrose pseudoinverse (Penrose, 1955) of $\mathbf{X}$. The Moore-Penrose pseudoinverse of a matrix $\mathbf{M}$ is a generalisation of the inverse of a matrix ($\mathbf{M}^{-1}$), is used as a way to approximate $\mathbf{M}^{-1}$ when $\mathbf{M}$ is not invertible and can be calculate via SVD since $\mathbf{M}^+ = \mathbf{V}_M\mathbf{\Sigma}^+_M\mathbf{U}^\top_M$, where $\mathbf{M} = \mathbf{U}_M\mathbf{\Sigma}_M\mathbf{V}^\top_M$ determines the SVD of $\mathbf{M}$ and $\mathbf{\Sigma}^+_M$ can be calculated by taking the reciprocals of its non-zero elements and then transposing it. The noise covariance matrices $\hat{\mathbf{Q}}$ and $\hat{\mathbf{R}}$ can then be directly estimated from residuals. We should also note that the dimensions of matrix $\mathbf{H}$ need to be specified. There is not one simple rule for this but practical advice is given in Van Overschee and De Moor (1996): The number of block rows r needs to be "sufficiently large", i.e. at least as large as the maximum order, d_x, of the system to be identified, which in turn is determined by keeping all singular values of $\mathbf{H}$ above a threshold. The number of columns τ is usually taken to be $\tau \to \infty$ for theoretical guarantees. However, in the case of finite data, a usual practice is to take $\tau = N - r + 1$, where N is the number of available samples.

Spectral learning is an appealing approach because it is non-iterative and therefore is faster than a maximum likelihood based method. However, it leads to sub-optimal results when compared to the EM algorithm. In fact, spectral learning can be considered as one iteration of a sub-optimal version of the EM algorithm as argued in Smith and Robinson (2000). In practice, spectral learning algorithms can provide a consistently good starting point for the EM algorithm with a very low computational cost. The EM algorithm then can improve this starting point by further optimisation, as pointed out in Barber (2012, sec. 24.5.3) or Murphy (2012, sec. 18.4.4).

3.2.5.3 Stability in LDS

In real-world applications it is important for the dynamics of the learned model to be stable. This is usually overlooked in the machine learning community but can be of critical importance especially when modelling patients' vital signs. A system with linear dynamics is stable if the spectral radius of its dynamics matrix $\mathbf{A}$ is at most 1.[3] In Siddiqi *et al.* (2007) it is shown how one can project an unstable matrix back to the space of stable matrices by treating this projection process as a constrained convex optimisation process. In Siddiqi *et al.* (2007), this approach is applied to LDS parameters that are obtained by the spectral learning algorithm but the approach remains virtually unchanged in the case of parameters learned under EM, as shown in Boots (2009).

3.2.5.4 AR process as LDS

Concluding the presentation of the LDS, it is worth stating that there is a connection between AR processes and LDSs, since any AR process can be cast into LDS form. For example, an AR process as defined in eq. (3.5) with added observation noise $v_t \sim \mathcal{N}(0, \sigma_v^2)$ can be cast into state space form as follows:

[3] The spectral radius of a square matrix is defined as the modulus of its largest eigenvalue.

$$\mathbf{x}_t = \begin{bmatrix} x_t & x_{t-1} & \dots & x_{t-p+1} & x_{t-p} \end{bmatrix}^\top , \tag{3.25}$$

$$\mathbf{A} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} , \quad \mathbf{Q} = \sigma_w^2 \mathbf{e}\mathbf{e}^\top , \tag{3.26}$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} , \quad \mathbf{R} = \sigma_v^2 , \tag{3.27}$$

where $\mathbf{e} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^\top$. Once cast into state space form, one can use the already presented LDS methodology for inference and learning.

3.2.6 Discrete-latent-variable models

Retaining the same graphical structure as in an LDS, but with latent random variables being discrete instead of continuous, a hidden Markov model (HMM) is obtained (Rabiner, 1989). HMMs' joint distribution factorises as in the case of LDSs and are thus formally equivalent to LDSs and admit analogous inference and learning routines with the only difference that the latent space is discrete. HMMs (as LDSs) are generative models which means that they can be used to generate sequences of observations. However, often we are only interested in the conditional distribution of the latent variable given an observation at time t. In that case, a graphical model that is especially tailored to this task can be more appropriate, giving rise to the Maximum Entropy Markon Model (MEMM) (McCallum *et al.*, 2000). The joint distribution of a MEMM factorises according to $p(\mathbf{x}, \mathbf{y}) = p(x_1|\mathbf{y}_1) \prod_{t=2}^{T} p(x_t|x_{t-1}) p(x_t|\mathbf{y}_t)$. The graphical models of those two approaches are depicted in Figure 3.3.

3.2.7 Switching linear dynamical systems

An LDS is restricted by the fact that all involved relationships are modelled as linear. Although many real-world processes can be approximated by LDSs in a satisfactory manner,

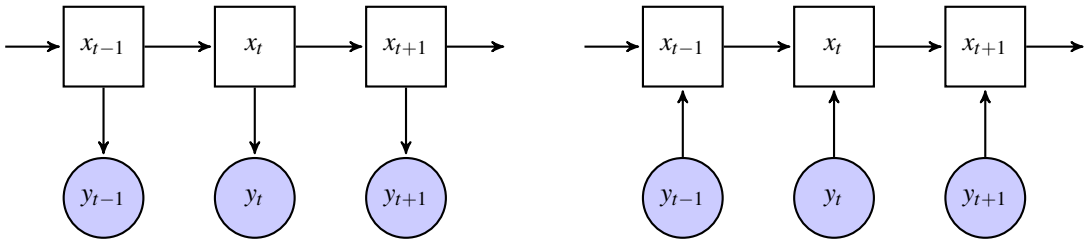

Figure 3.3: Graphical model of the HMM (left) and the MEMM (right). Note that in the case of the MEMM the conditional probability $p(x_t|\mathbf{y}_t)$ is modelled directly.

it is expected that there are cases where such approximations might produce poor results. One could then use non-linear dynamical systems, an example of which we will encounter in Chapter 5. However, there are cases where an observed process could be well approximated by multiple LDSs instead of one. For example, let us assume that a patient's status can be modelled satisfactorily by an LDS when their condition is stable.[4] Let us also assume that their status can be modelled by a different LDS (i.e. with different parameters) when their condition is unstable (assuming an appropriate definition of instability is provided). In that example a simple LDS learned over both conditions might provide poor results, but a collection of the two, where each one is used appropriately depending on the patient's status could provide much better results. This is the idea behind a switching LDS (SLDS) (Ghahramani and Hinton, 2000). At each time t a "switch" variable $s_t \in \{1,...,S\}$ determines which of a set of LDSs should be used. Thus, conditioned on s_t, the SLDS "reverts" to an LDS and in some sense the SLDS provides a piece-wise linear approximation to the observed non-linear relationships. The SLDS's joint distribution is defined as:

$$p(\mathbf{s},\mathbf{x},\mathbf{y}) = p(s_1)p(\mathbf{x}_1|s_1)p(\mathbf{y}_1|\mathbf{x}_1,s_1)\prod_{t=2}^{T} p(s_t|s_{t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1},s_t)p(\mathbf{y}_t|\mathbf{x}_t,s_t)\ , \qquad (3.28)$$

and the graphical model which induces this factorisation is shown in Figure 3.4. The involved parameters remain as in the case of the LDS, where now a different set of LDS parameters is associated with each switch state (i.e. there are S different sets of LDS parameters). Additionally, we now also need to establish the transition relationship $p(s_t = j|s_{t-1} = i) = S_{ij}$ which governs transitions from state i at time $t-1$ to state j at time t.

[4] Here, "stable" refers to a patient's physiological health status as defined in Section 2.3.2 and is not to be confused with the notion of stability defined for LDSs.

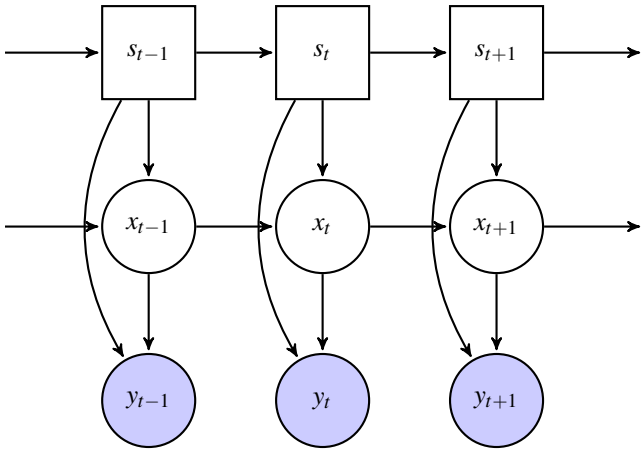

Figure 3.4: Graphical model of the SLDS. The discrete latent state is represented by the white square nodes and is denoted by s_t. The continuous latent and observed variables ($\mathbf{x}_t$ and $\mathbf{y}_t$ respectively) are represented as in the case of the LDS.

This can be parameterised by a stochastic (i.e. each of its rows sums to 1) matrix $\mathbf{T}$, the elements of which can be estimated as:

$$\hat{T}_{ij} = \frac{n_{ij} + \zeta}{\sum_{k=1}^{S} n_{ik} + \zeta} , \tag{3.29}$$

where n_{ij} denotes the number of transitions from discrete state i to discrete state j in the training dataset and ζ is a small value (set to 1 in this thesis as in Quinn, 2007), that ensures no state transition is set to zero.

3.2.7.1 Inference in SLDS

Exact inference in the SLDS is intractable since its computational complexity scales exponentially with time, as shown in Lerner and Parr (2001). For an intuitive explanation, let us assume that at time $t-1$ the joint posterior distribution of the latent states is a set of S Gaussians. Then, at time t, due to the summation over the states s_{t-1}, the joint posterior will be a set of S^2 Gaussians. Under the same reasoning, it then becomes apparent how computational demands scale exponentially with time.

One way of resolving this intractability is by ensuring that the number of Gaussians used to model the joint posterior density remains bounded across time. Similarly to the previous paragraph, if we assume that at time t a set of S^2 Gaussians is obtained, then we can collapse this back to S Gaussians by matching moments (up to second order) for the

distribution obtained for each setting of s_t. This is known as the Gaussian Sum algorithm and was introduced by Alspach and Sorenson (1972).

Another approximation method that can be used to overcome the intractability is Rao-Blackwellised particle filtering (RBPF). The RBPF technique exploits the conditionally linear dynamic structure of the SLDS to avoid sampling both s_t and $\mathbf{x}_t$, or in other words it takes advantage of the fact that if we know s_t, we can compute $\mathbf{x}_t$ exactly using the Kalman filter algorithm. Thus, RBPF proceeds by using a particle filter to estimate the distribution of s_t and exact computations (i.e. one Kalman filter per particle, where the value of the switch state for each particle is obtained by sampling from the transition probabilities) to estimate the mean and variance of $\mathbf{x}_t$.[5] RBPFs have been used before in settings with switching linear dynamics as in De Freitas (2002) for the purposes of fault diagnosis.

In Quinn *et al.* (2009) a practical comparison between the two inference methods (in the context of a variant of an SLDS) showed that the Gaussian Sum approximation outperforms RBPF when the computational cost is the same for both methods. The number of particles used was such that both methods took an (approximately) equal amount of computational time[6]. Based on these findings, we will be using the Gaussian Sum approximation later in this thesis.

3.2.7.2 Learning in SLDS

Under the assumption that training data are available for each switch setting s_t, the learning process decomposes into learning independently a number of S LDSs, one for each configuration of the switch variable. This can be done as already explained in Section 3.2.5.2 for the LDS specific parameters and the transition matrix probabilities are given by eq. (3.29). It should be noted that it is possible to fit an SLDS even if no switch variable labels are provided by using EM, as shown in Quinn (2007, app. C.3).

[5] For a tutorial on particle filters see Doucet and Johansen (2009).

[6] A set of limited, preliminary experiments considering the same trade-off between effectiveness and computational efficiency led to the same conclusions in our case and thus further exploration of the RBPF was not deemed necessary.

3.2.7.3 Factorial SLDS

In some applications, such as in condition monitoring, it could be possible that the observations are affected by the contribution of multiple independent factors. In these cases it is possible to factorise the state variable, representing it in terms of all the involved factors. This could prove advantageous in terms of model interpretability and space requirements. Such an approach is presented in Ghahramani and Jordan (1997) in the context of factorial hidden Markov models. In a similar vein, Quinn *et al.* (2009) present a variant of the SLDS called the factorial SLDS (FSLDS), which can be formulated by assuming that the switch variable factorises according to the cross-product of J factors, so that $s_t = f_t^1 \otimes f_t^2 \otimes \ldots \otimes f_t^J$ and $p(s_t|s_{t-1}) = \prod_{j=1}^{J} p(f_t^j|f_{t-1}^j)$. A graphical model of the FSLDS can be seen in Figure 3.5. Inference and learning takes place as in the case of the SLDS.

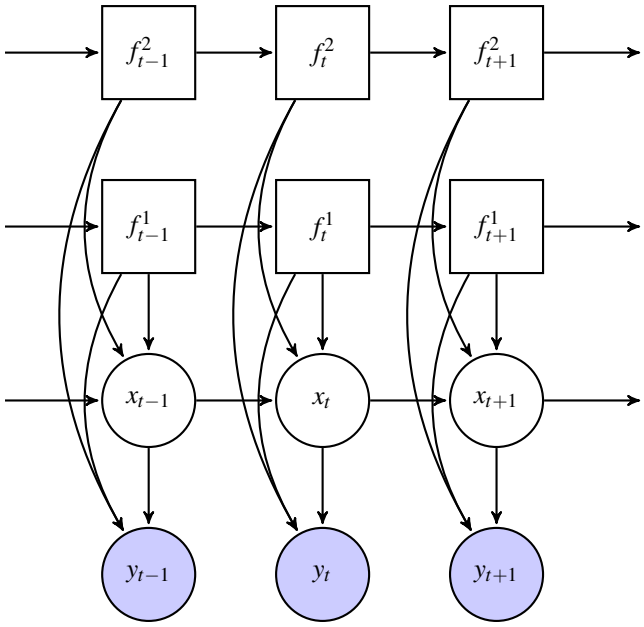

Figure 3.5: Graphical model of the FSLDS with two factors f^1 and f^2. The remaining nodes are as in the case of the SLDS.

3.2.7.4 Hierarchical SLDS

The FSLDS can also admit a hierarchical decomposition of its discrete latent structure in order to model more complex interactions between modes of operation, e.g. when there is evidence (or is assumed) that the transitions between latent discrete variables of lower

levels in the hierarchy are governed by the dynamics of latent discrete variables of higher levels as is the case in Stanculescu *et al.* (2014). This leads to the formulation of a hierarchical SLDS (HSLDS) in which, analogously to the FSLDS, the discrete latent space can be decomposed as $s_t = z_t \otimes f_t^1 \otimes f_t^2 \otimes ... \otimes f_t^J$, where z_t denotes the top level discrete latent variable, and the transition probability of the joint discrete space is factorised as $p(s_t|s_{t-1}) = p(z_t|z_{t-1}) \prod_{j=1}^{J} p(f_t^j|z_t, f_{t-1}^j)$. A graphical model of the HSLDS can be seen in Figure 3.6.

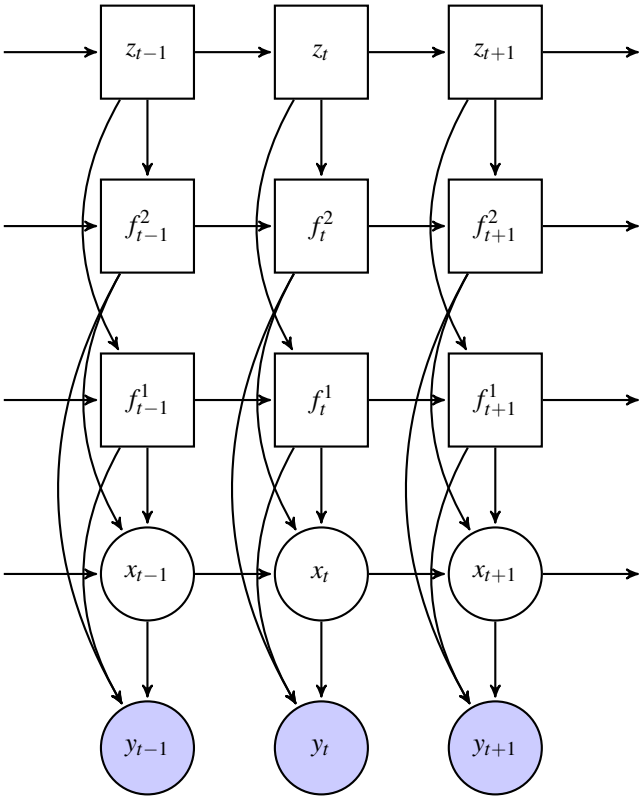

Figure 3.6: Graphical model of the HSLDS with two levels of hierarchy, z and f, and two factors, f^1 and f^2. The remaining nodes are as in the case of the SLDS.

3.3 Pharmacokinetics/Pharmacodynamics

In this section we review work that is developed specifically for the purpose of modelling drug behaviour in a patient's body. Establishing a relationship between drug dose and resultant drug concentration and subsequently establishing a relationship between drug concentration and effect (e.g. in observed signals) can lead to better modelling of the observed vital signs and has the potential to lead to better patient treatment overall. The

field of pharmacokinetics (PK) is concerned with the study of the concentration of drugs in tissue as a function of time and dose schedule. The distribution of drugs in the body is affected by complex transfer and metabolic processes, many of which are poorly understood, as mentioned in Bailey and Haddad (2005). On the other hand, the field of pharmacodynamics (PD) is concerned with the biochemical and physiological effects of drugs on a patient's body. We start by providing an introduction to the main approach in PK/PD which is called *compartmental modelling*. Our description consists of simple compartmental models which are represented by systems of linear ordinary and stochastic differential equations (ODEs & SDEs respectively) and then we focus on previous approaches taken in the specific domain of anaesthsesia as described in the medical and control-theoretic literature.

3.3.1 Compartmental models

Compartmental models are an abstraction used in the field of PK to describe the rate of change of a drug's concentration in a subject by accounting for the processes of absorption, distribution, metabolism and excretion of the drug in different parts (compartments) of the human body. An initial approach then is to build a system of ODEs that describe the evolution of drug concentrations at different compartments. The simplest compartmental model one could think of would be a deterministic one-compartment model (1-CM), described by the following ODE:

$$\frac{dC_t}{dt} = -k_{10}C_t \ , \tag{3.30}$$

with initial value $C_0 = D/V$, where D is an injected intravenous bolus dose of the drug, V is the volume of the compartment and k_{10} is the elimination rate of the drug from the compartment. This linear ODE has a solution of the form: $C_t = \frac{D}{V}e^{-k_{10}t}$. Such a deterministic approach does not account for sources of error in our models and most importantly does not take into account uncertainty about the modelled process which is inherent to any such attempt since one cannot possibly expect to capture all sources that contribute to the modelled process either because that would lead to an overly complex model or most frequently because such a complete model is unknown. Therefore, taking a probabilistic perspective leads us to a respective SDE of the form:

$$dC_t = -k_{10}C_t dt + \rho C_t dB_t \ , \tag{3.31}$$

as presented in Ramanathan (1999), with the same initial values and where B_t is a Brownian motion[7] and ρ is a constant. Such a linear SDE also has an explicit solution (Øksendal, 2013, sec. 5.1) with mean and covariance:

$$\begin{aligned} E[C_t] &= C_0 e^{-k_{10}t} \ , \\ Cov[C_t, C_s] &= C_0^2 e^{-k_{10}(t+s)}(e^{\rho^2 s} - 1) \ . \end{aligned} \tag{3.32}$$

Single compartment models can be rather restrictive and have been extended to multi-compartment models which allow for more flexibility in modelling the pharmacokinetic behaviour of a drug. These models are loosely based on physiological arguments with e.g. one compartment modelling the target site of a drug (e.g. brain) and another compartment modelling other components (e.g. tissue) with potentially different transfer/elimination rates but the main motivation is the introduction of additional parameters to allow for extra flexibility in modelling the observed processes. A simple 2-CM is shown in Figure 3.7. This diagram encodes the following deterministic system:

$$\begin{aligned} \frac{dC_{1t}}{dt} &= (-k_{12}C_{1t} + k_{21}C_{2t} - k_{10}C_{1t}) \ , \\ \frac{dC_{2t}}{dt} &= (k_{12}C_{1t} - k_{21}C_{2t}) \ , \end{aligned} \tag{3.33}$$

where C_i is the concentration of drug in compartment i and k_{ij} is the drug's transfer rate from compartment i to j. When $j = 0$ this turns into an elimination rate from compartment i. Thus the diagram in Figure 3.7 describes the drug concentration and rate of transfer in/between C_1 (central compartment) and C_2 (target site). We then might be interested only in the concentration of the drug in the target site and treat the central compartment as a means to describe the transfer in and out of C_2 and out of the subject's body (via the elimination process).

Again, a system of two SDEs might be more appropriate, as shown in Donnet and Samson (2013):

[7] A white noise process can be formally defined as the derivative of a Brownian motion, see e.g. Särkkä and Solin (2014, sec. 3.1).

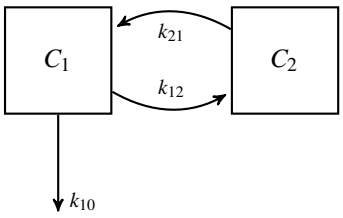

Figure 3.7: A simple 2-compartment model.

$$dC_{1t} = (-k_{12}C_{1t} + k_{21}C_{2t} - k_{10}C_{1t})dt + \rho_1 dB_{1t} \ ,$$
$$dC_{2t} = (k_{12}C_{1t} - k_{21}C_{2t})dt + \rho_2 dB_{2t} \ , \tag{3.34}$$

where B_{1t}, B_{2t} are independent Brownian motions and ρ_1, ρ_2 are the respective diffusion coefficients. All of the above examples admit exact maximum likelihood estimation methods (see e.g. Donnet and Samson, 2013) as they are linear and are not corrupted by observation noise. Non-linear SDEs might be a more appropriate choice in some cases and in Andersen and Højbjerre (2003) an example of a 3-CM non-linear model for modelling glucose and insulin plasma concentration is described but the authors tested this model only on synthetic data.

3.3.1.1 Compartmental models in anaesthesia

One of the most widely used anaesthetics in ICU patients is Propofol. Two well-known models that describe the concentration of Propofol in human subjects are the Marsh model (Marsh *et al.*, 1991) and the Schnider model (Schnider *et al.*, 1998). They are both based on a 3-CM presented in Gepts *et al.* (1987) and shown in Figure 3.8. C_1 denotes the drug concentration in the central compartment, which is the site for drug administration and includes the intravascular blood volume and highly perfused[8] organs, such as the heart, brain, kidney and liver. The highest fraction of the administrated drug is assumed to reside in the central compartment. The remainder is diffused in two peripheral compartments which represent the body's muscle and fat. The drug's concentration in these two compartments is denoted as C_2 and C_3 respectively. The infusion scheme used in Gepts

[8]High ratio of blood flow to weight.

et al. (1987) was a bolus injection, followed by an initial rapid infusion and a slower maintenance infusion, giving rise to a tri-exponentially decreasing infusion rate describing the initial distribution of the drug, its transfer to remote compartments and its elimination from the central compartment. In that work the authors are only interested in the drug concentration in the central compartment (C_1), which is given by:

$$C_1(I+t) = A_1(1-e^{-\pi I})e^{-\pi t} + A_2(1-e^{-\xi I})e^{-\xi t} + A_3(1-e^{-\phi I})e^{-\phi t}\ , \tag{3.35}$$

where I is the infusion duration, t is the postinfusion time and $A_1, A_2, A_3, \pi, \xi, \phi$ are the intercepts and decay parameters of the aforementioned three exponential stages, respectively. The parameters are then learned using an extended least squares estimation method as described in Peck *et al.* (1984).

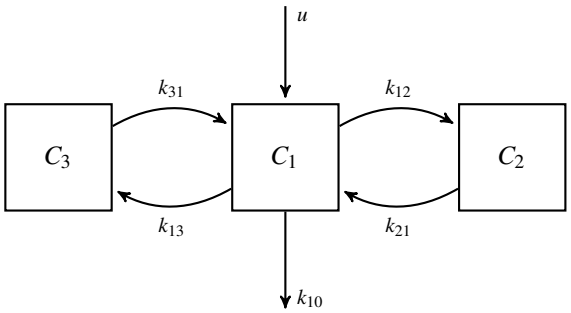

Figure 3.8: A 3-compartment model for Propofol.

So far, this describes the PK component. The PD component has a similarly long history. In Colburn (1981), one, two and three compartmental PK models with the addition of an effect site compartment to model PD effects were investigated. Similarly, in Fuseau and Sheiner (1984), a three compartmental PK model with an additional effect site compartment was studied. An approach which is more motivated by control-theoretic arguments is a PK/PD model presented in Bailey and Haddad (2005), where a three-compartmental PK model with an effect site compartment (which is assumed, for simplicity, to have zero volume and thus equilibrate instantaneously with the central compartment) is cast as a state space model and a sigmoidal function links the drug concentration at the effect site

to the patient's observed level of consciousness, y, as follows:

$$\begin{aligned}
\frac{dC_{1t}}{dt} &= -(k_{10}+k_{12}+k_{13})C_{1t}+k_{21}C_{2t}+k_{31}C_{3t}+u_t \ , \\
\frac{dC_{2t}}{dt} &= -k_{21}C_{2t}+k_{12}C_{1t} \ , \\
\frac{dC_{3t}}{dt} &= -k_{31}C_{3t}+k_{13}C_{1t} \ , \\
\frac{dC_{et}}{dt} &= k_{1e}(C_{1t}-C_{et}) \ , \\
y_t &= h(C_{et}) \ . \qquad (3.36)
\end{aligned}$$

where u_t is the drug dosage at time t, k_{1e} is the drug transfer rate between the central and the effect site compartment, C_{et} is the drug concentration at the effect site at time t and $h(\cdot)$ is a sigmoidal function called the *Hill* function. The Hill function has the following form:

$$h(C_{et}) = E_0 + \frac{E_{max}C_{et}^{\nu}}{EC_{50}^{\nu}+C_{et}^{\nu}} \ , \qquad (3.37)$$

where E_0, E_{max}, EC_{50} and ν are parameters to be estimated. The Hill function is frequently used in the PK/PD literature (see e.g. Goutelle *et al.*, 2008) and a similar sigmoidal function will be used later in this thesis. However, the above model is non-stochastic, and in Bailey and Haddad (2005) it is evaluated on simulated data with parameters fixed on a priori known values. A graphical representation of this approach is shown in Figure 3.9.

3.4 Applications to physiological condition monitoring

In terms of research which is closer to the machine learning community, there has been work which tackles various tasks of interest with respect to physiological monitoring in ICUs. The work which is most similar to this thesis' Chapter 4 is that presented in Quinn *et al.* (2009), where a FSLDS was used to infer artifactual and physiological processes of interest. In that work a generative approach is taken for modelling a number of observed vital signs under the influence of different factors of clinical interest. These factors

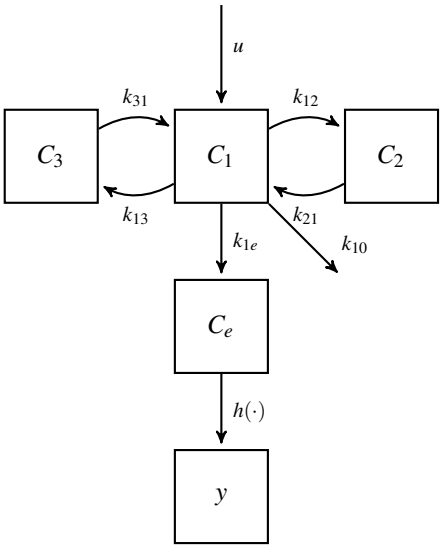

Figure 3.9: A three compartment model with one added effect site compartment.

were divided into two categories: artifactual and physiological ones. The former models artifactual processes such as channel dropouts, while the latter models physiological processes such as bradycardias. A naive construction of the factored space can quickly lead to a high demand for training data, since training instances for each potential combination of factors need to be collected. However, the authors noted that some factors can overwrite others which can mitigate the training data volume requirements. For example, a HR channel dropout will overwrite the HR observations even in the presence of a bradycardia (which would have affected the HR observations in a different way), and therefore this explicit combination of factors does not need to be part of the training set, since it is equivalent to a dropout event in terms of the observed signals. This overwriting mechanism leads to a partial ordering of the factors that allows the FSLDS to be trained with less training data. Overall, this work made use of carefully constructed processes which require a significant level of expert clinical knowledge and can also be time consuming to develop and validate. In the same work, a method for handling events that are classified as abnormal but do not fall into the category of known abnormalities is presented; this tackles in some sense the task of anomaly detection in a patient's physiology (see also Quinn and Williams, 2007).

In a similar vein, work presented by Aleks *et al.* (2009) makes use of a dynamic Bayesian network to detect short artifactual events in a neonatal ICU based solely on minute-by-minute blood pressure measurements. Stanculescu *et al.* (2014) turn their focus on mod-

elling a physiologically critical condition for neonates. To that end they use a HSLDS for the purposes of predicting the onset of sepsis in neonates by adding a higher-level discrete variable with semantics sepsis/non-sepsis which enables the modelling of changes in lower-level, physiological factors that signal the presence of sepsis. Furthermore, in Lehman *et al.* (2015), a switching vector autoregressive process was used to extract features from vital signs which were used as input in a logistic regression classifier to predict patient outcome, while work by Nemati *et al.* (2013) was focussed on discriminatively training a SLDS for learning dynamics associated with patient outcome. These methods are primarily concerned with the classification of events of interest and they do not incorporate any additional information, such as administered drugs. Additionally, all of the aforementioned works use linear models as their core elements, while non-linearities might govern the observed data.

In Saria *et al.* (2010), a Bayesian non-parametric approach is developed for exploratory data analysis and feature construction in continuous time series for the task of condition monitoring in the case of neonates. Finally, methods based on Recurrent Neural Networks (RNNs) have recently started showing promising results such as the work by Lipton *et al.* (2015) who made use of RNNs for the task of classifying diagnoses from electronic health records.

There is also considerable work focussed on regression for physiological time-series. In Clifton *et al.* (2013), a model based on the Gaussian process (GP) regression framework is developed for deployment in wearable sensors for patient monitoring with an emphasis on handling artifacts and missing observations, where it is shown that a personalised medicine approach using GPs can result in a significant increase in the amount of warning time provided prior to the deterioration of a patient's physiology. In that work, an independent GP is used per vital sign and thus the dependences between observed vital signs are not taken into account. This independence assumption can be improved upon via the use of multi-task GPs, which can model multiple physiological time-series including their correlations in the modelling process as shown in Dürichen *et al.* (2015). Although GPs provide a very flexible probabilistic framework, they come at a computational complexity cost which can be prohibitive for large datasets since it scales cubically with the total length T of a vital sign. This is further exacerbated in the case of multi-task GPs where the computational complexity increases to $O\left(M^3T^3\right)$, where M is the number of "tasks", namely the number of vital signs to be modelled, although some approaches such as the

use of *inducing variables*(Quiñonero-Candela and Rasmussen, 2005) can be used to mitigate these concerns to some extent. Also, it should be noted that the GP formulation lends itself very naturally to the task of regression but is rather more complicated in the case of classification (see e.g. Rasmussen and Williams, 2006, sec. 3.3).

Finally, work more similar to this thesis' which also models the effect of infused drugs is presented by Enright *et al.* (2011) and Enright *et al.* (2013), where a non-linear dynamical system is developed to model the glucose levels of patients under the intravenous administration of glucose and insulin. However, their approach relies solely on converting existing systems of ODEs into a probabilistic model which is only a subset of this thesis' proposed work in Chapter 5.

A comprehensive review of the use of machine learning in critical care which provides references to a wide scope of relevant work can be found in Johnson *et al.* (2016).

3.5 Summary

In this chapter we have reviewed prior work that prepares the ground for the subsequent chapters that contain the original research contributions of this thesis. Section 3.1 briefly touched on some basic models that will be used mainly as components of more complex methods, while Section 3.2 presented in more detail various dynamical systems which are more tailored to the task at hand. In Section 3.3, the focus turned on prior work dealing explicitly with modelling drug effects via PK/PD models and Section 3.4 concluded the literature review by describing a number of machine learning methods that have been applied to the task of physiological condition monitoring.

Chapter 4

Discriminative Switching Linear Dynamical Systems

In Section 3.2.7.3, a generative approach via the FSLDS (Quinn *et al.*, 2009) was described for the task of physiological condition monitoring in ICUs. In this chapter, a different, discriminative process is developed which gives rise to a *discriminative switching linear dynamical system*, henceforth termed as DSLDS. Although a generative model can be more parsimonious, it is also restricted by its need to explicitly model the distribution over observed variables (and potential features thereof) which can be highly complex. Thus, it can be "very hard, and often impossible" to construct such a model (Koller and Friedman, 2009, Sec. 20.3.2). On the other hand, a discriminative approach sidesteps this challenge altogether by directly modelling the conditional probability of the hidden state given the observed variables. This approach enables one to directly focus on the desired task, which is the identification of the hidden states given the observed variables, while at the same time facilitates greatly the incorporation of prior knowledge into the model via the construction of arbitrarily complex (and potentially highly correlated) features. Indeed, building a generative process for the task at hand (via the FSLDS) required the construction of very detailed models of the events of interest. For the modelled events, especially artifactual ones, this can be non-trivial since some of them exhibit high variability which is hard to capture with a generative process, but can nonetheless contain informative features which can act as input to a discriminative process. Thus, the DSLDS makes use of a discriminative model for discrete-state inference and retains an almost identical inference scheme as in the FSLDS for the continuous latent variables conditional on

the inferred discrete state. The results show that using the DSLDS gives increased performance over the FSLDS in most cases of interest, and that a combination of the two methods via an α-mixture approach was able to achieve a higher performance than either of the two models separately.

The reader is reminded that the goals under the DSLDS remain the same as in the FSLDS, namely:

- Identifying artifactual processes (e.g blood samples), which will reduce the high false alarm rate in ICUs and facilitate the task of identifying physiological processes.
- Identifying physiological processes which can be of critical importance (e.g bradycardias).
- Providing an estimate of a patient's true physiological values when these are obscured by artifact.

The structure of the remainder of the Chapter is as follows: in Section 4.1 a description of the proposed model is given, and its graphical structure and inference methods are compared to those of the FSLDS. In Section 4.2 a description of the performed experiments is provided and results for the comparison between the DSLDS and the FSLDS are given. Finally, Section 4.3 concludes with general remarks about the proposed model and suggestions for future work. Parts of this chapter have been adapted from Georgatzis and Williams (2015).

4.1 Model description

The graphical model of the DSLDS is depicted in Figure 4.1 (left). The FSLDS is also depicted in Figure 4.1 (right) for the purpose of comparison[1]. The DSLDS, as the FSLDS, operates on three different sets of variables: The observed variables, $\mathbf{y}_t \in \mathbb{R}^{d_y}$ represent the patient's vital signs obtained from the monitoring devices at time t, which act as inputs to the model. The continuous latent variables, $\mathbf{x}_t \in \mathbb{R}^{d_x}$, track the evolution of the dynamics of a patient's underlying physiology. The discrete variable, s_t, represents the switch setting or regime which the patient is currently in (e.g. stable, a blood sample is

[1]This graphical model is equivalent to the one presented in Figure 3.5 but slightly modified for ease of comparison to the DSLDS's graphical model.

being taken etc.). The switch variable can be factorised according to the cross-product of M factors, so that $s_t = f_t^1 \otimes f_t^2 \otimes ... \otimes f_t^M$. Each factor variable, f_t^m, is usually a binary vector indicating the presence or absence of a factor, but in general it can take on $L^{(m)}$ different values and $K = \prod_{m=1}^{M} L^{(m)}$ is the total number of possible configurations of the switch variable, s_t.

It is perhaps convenient to consider the DLSDS by mapping its discriminative components to their generative FSLDS counterparts. We remind the reader that, conditioned on a particular regime, the FSLDS is equivalent to an LDS. The FSLDS can be seen then as a collection of LDSs, where each LDS models the dynamics of a patient's underlying physiology under a particular regime, and can also be used to generate a patient's observed vital signs. An LDS provides a generative framework for modelling our belief over the state space, given observations. Switching to the DSLDS's discriminative perspective, we start by modelling $p(s_t|\mathbf{y}_{t-l:t+r})$ with a discriminative classifier, where (features of) observations from the previous l and future r time steps affect the belief of the model about s_t. The inclusion of r frames of future context is analogous to fixed-lag smoothing in an FSLDS (see e.g. Särkkä, 2013, sec. 10.5). It is noted that inclusion of future observations in the conditioning set means that the DSLDS will operate with a delay of r seconds, since an output of the model at time t can be produced only after time $t+r$, as explained further in Section 4.2.1.1. The LDS can also be regarded from a similarly discriminative viewpoint which allows us to model $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t)$. This is similar to the MEMM (see Section 3.2.6) with the difference that the latent variable is continuous rather than discrete. The main advantage of this discriminative view is that it allows for a rich number of (potentially highly correlated) features to be used without having to explicitly model their distribution or the interactions between them, as is the case in a generative model. A combination of these two discriminative viewpoints gives rise to the DSLDS's graphical model. The DSLDS can be seen then as a collection of MEMMs, conditioned on s_t, where each MEMM in the DSLDS plays a role equivalent to that of each LDS in the FSLDS.

The DSLDS can be defined as

$$p(\mathbf{s}, \mathbf{x}|\mathbf{y}) = p(s_1|\mathbf{y}_{1:1+r}) p(\mathbf{x}_1|s_1, \mathbf{y}_1) \prod_{t=2}^{T} p(s_t|\mathbf{y}_{t-l:t+r}) p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t, \mathbf{y}_t) \ . \tag{4.1}$$

The simplest assumption we can make for the DSLDS is that $p(s_t|\mathbf{y}_{t-l:t+r})$ factorises, so that

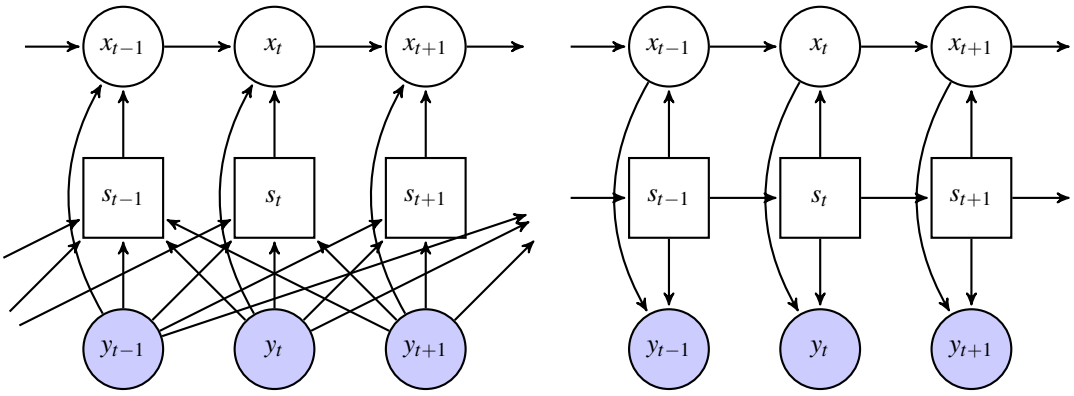

Figure 4.1: Graphical model of the DSLDS (left) and the FSLDS (right). The state-of-health and underlying physiological values of a patient are represented by s_t and $\mathbf{x}_t$ respectively. The shaded nodes correspond to the observed physiological values, $\mathbf{y}_t$. Note that in the case of the DSLDS the conditional probability $p(s_t|\mathbf{y}_{t-l:t+r})$ is modelled directly.

$$p(s_t|\mathbf{y}_{t-l:t+r}) = \prod_{m=1}^{M} p(f_t^{(m)}|\mathbf{y}_{t-l:t+r}) \quad . \tag{4.2}$$

However, one could as well make use of a structured output model to predict the joint distribution of different factors.

The DSLDS bears similarities to the model used by Lu *et al.* (2009) for the task of estimating true terrain elevation (potentially obscured by buildings, trees etc.) from Light Detection and Ranging (LiDaR) data. The idea behind their model, dubbed a "hybrid conditional random field", is that conditioned on an observation (which is the height of a surface point as measured by LiDaR), they first classify it as belonging to the ground or not (corresponding to our discrete variable, s_t) and then conditioned on this variable, they infer the true height of the terrain on that point (corresponding to our continuous variable, $\mathbf{x}_t$). Although there are similarities on the graphical representation of the model, their approach was used to model spatial relationships and they were only concerned with a binary discrete latent space (ground, non-ground). In our case, we are concerned with modelling temporal structure and we have a richer and more complex discrete latent space. More importantly, in their work the distribution maintained over the continuous latent space is a single multivariate Gaussian, whereas in our model, as described in Section 4.1.4, the belief over the continuous latent space is modelled as a mixture of K Gaussians. This allows us to keep track of multiple modes about the belief over a patient's underlying

physiology, since this is potentially affected by multiple factors.

In the remainder of this section, the models for the discrete underlying state of health (s_t) and the continuous underlying physiological state ($\mathbf{x}_t$) of a patient are presented in Sections 4.1.1 and 4.1.2 respectively. The learning process is documented in Section 4.1.3 and inference is explained in Section 4.1.4.

4.1.1 Predicting $\mathbf{s}_t$

The belief about the state of health[2] of a patient at time t is modelled by $p(s_t|\mathbf{y}_{t-l:t+r})$, the conditional probability of the switch variable given the observed vital signs. Following the factorisation of the switch variable in eq. 4.2, we model the conditional probability of each factor being active at time t given the observations with a probabilistic discriminative binary classifier, so that $p(f_t^{(i)} = 1|\mathbf{y}_{t-l:t+r}) = G(\phi(\mathbf{y}_{t-l:t+r}))$, where $G(\cdot)$ is a classifier-specific function, and $\phi(\mathbf{y}_{t-l:t+r})$ is the feature vector that acts as input to our model at each time step as described in Section 4.2.1. As is evident from Figure 4.1 (left) there is no explicit temporal dependence on the switch variable sequence. However, temporal continuity is implicitly incorporated in the model through the construction of the features. Thus, the classifier can make use of the additional information present in the temporal structure of the observations to infer the value of the switch variable.

4.1.1.1 An α-mixture of $\mathbf{s}_t$

The DSLDS model can be seen as complementary to the FSLDS, and they can be run in parallel. One way of combining the two outputs is to maintain an α-mixture over s_t. If $p_g(s_t)$ and $p_d(s_t)$ are the outputs for the switch variable at time t from FSLDS and the DSLDS respectively, then their α-mixture is given by:

$$p_\alpha(s_t) = c\left(p_g(s_t)^{(1-\alpha)/2} + p_d(s_t)^{(1-\alpha)/2}\right)^{2/(1-\alpha)}, \tag{4.3}$$

where c is a normalisation constant which ensures that $p_\alpha(s_t)$ is a probability distribution. The family of α-mixtures then subsumes various known mixtures of distributions

[2]Here, state-of-health is meant to be taken in a broader context so as to include events, such as various artifacts, that are potentially not related directly to the actual health of patients but could nonetheless affect their observed physiology.

and defines a continuum across them via the α parameter. For example, for $\alpha = -1$ we retrieve the mixture of experts (with equally weighted experts) framework, while for $\alpha \to 1$, eq. 4.3 yields $p_1(s_t) = c\sqrt{p_g(s_t)p_d(s_t)}$, rendering it equivalent to a product of experts viewpoint. In general, as α increases, the α-mixture assigns more weight to the smaller elements of the mixture (with $\alpha \to \infty$ giving $p_\infty(s_t) = \min\{p_g(s_t), p_d(s_t)\}$), while as α decreases, more weight is assigned to the larger elements (with $\alpha \to -\infty$ giving $p_{-\infty}(s_t) = \max\{p_g(s_t), p_d(s_t)\}$). The interested reader can find a thorough treatment of α-mixture models in Amari (2007).

4.1.2 Predicting $\mathbf{x}_t$

The model of the patient's physiology should capture the underlying temporal dynamics of their observed vital signs under their current health state. The idea is that the current latent continuous state of a patient should be dependent on (a) the latent continuous state at the previous time step, (b) the current state of health and (c) the current observed values. These assumptions are modelled as follows:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t, \mathbf{y}_t) \propto \exp\{-\frac{1}{2}(\mathbf{x}_t - \mathbf{A}^{(s_t)}\mathbf{x}_{t-1})^\top (\mathbf{Q}^{(s_t)})^{-1}(\mathbf{x}_t - \mathbf{A}^{(s_t)}\mathbf{x}_{t-1})\} \times \exp\{-\frac{1}{2}(\mathbf{C}^{(s_t)}\mathbf{x}_t - \mathbf{y}_t)^\top (\mathbf{R}^{(s_t)})^{-1}(\mathbf{C}^{(s_t)}\mathbf{x}_t - \mathbf{y}_t)\} \quad . \tag{4.4}$$

The first term on the RHS of eq. 4.4 is the system model for an LDS and captures the dynamics of a patient's latent physiology under state s_t. The second term can be seen as the discriminative counterpart of the observation model of an LDS. In our condition monitoring setting, the observed vital signs are considered to be noisy realisations of the true, latent physiology of a patient and thus, the observation model encodes our belief that $\mathbf{x}_t$ is a noisy version of $\mathbf{y}_t$. Under this assumption, $\mathbf{C}^{(s_t)}$ consists of 0/1 entries, which are set based on our knowledge of whether the observations $\mathbf{y}_t$ are artifactual or not under state s_t. In the FSLDS, the corresponding observation model encodes the belief that the generated $\mathbf{y}_t$ should be normally distributed around $\mathbf{x}_t$ with covariance $\mathbf{R}^{(s_t)}$, whereas in our discriminative version, the observation model encodes our belief that $\mathbf{x}_t$ should be normally distributed around $\mathbf{y}_t$ with covariance $\mathbf{R}^{(s_t)}$. The idea behind this model is that at each time step we update our belief about $\mathbf{x}_t$ conditioned on its previous value, $\mathbf{x}_{t-1}$,

and the current observation, $\mathbf{y}_t$, under the current regime s_t. For example, under an artifactual process, the observed signals which are obscured by artifact do not convey useful information about the corresponding latent variable of the underlying physiology of a patient. In that case, the connection between $\mathbf{y}_t$ and $\mathbf{x}_t$ is dropped (for the artifact-affected channels) which translates into setting the respective entries of $\mathbf{C}^{(s_t)}$ and the Kalman gain matrix to zero. Then, the latent state $\mathbf{x}_t$ evolves only under the influence of the appropriate system dynamics parameters $(\mathbf{A}^{(s_t)}, \mathbf{Q}^{(s_t)})$. Conversely, operation under a non-artifactual regime incorporates the information from the observed signals, effectively transforming the inferential process for $\mathbf{x}_t$ into a product of two "experts", one propagating probabilities from $\mathbf{x}_{t-1}$ and one from the current observations.

It should be noted that the step of conditioning on the current regime s_t in order to predict $\mathbf{x}_t$ is required for the task at hand, as no training data are available for the $\mathbf{x}$-state. Otherwise, one could imagine building a simpler model such as a conditional random field (CRF) (Lafferty *et al.*, 2001), to predict the $\mathbf{x}$-state directly from the observations. However, in our case, where only labels about the patient's regime are available, this is not possible.

4.1.3 Learning

The parameters that need to be learned are: $\mathbf{A}^{(s)}$, $\mathbf{Q}^{(s)}$, $\mathbf{C}^{(s)}$, $\mathbf{R}^{(s)}$. Given training data for each switch setting, these can be learned independently as LDS parameters for each configuration of s. Following Quinn *et al.* (2009) we use an independent ARI model with added observation noise for each channel. Casting such a model into state space form is a standard procedure, as was described in Section 3.2.5.4, and amounts into reformulating the parameters of the aforementioned model into the parameters of a state space model. Once the model is in state space form, $\mathbf{A}^{(s)}$, $\mathbf{Q}^{(s)}$, $\mathbf{C}^{(s)}$, $\mathbf{R}^{(s)}$ can be learned according to the maximum likelihood criterion by using numerical optimisation methods (like Newton–Raphson, Gauss–Newton), as presented in Shumway and Stoffer (2000, sec. 2.6) or the EM algorithm as presented in Section 3.2.5.2. We note that the vector ARMA (VARMA) representation is used, where for example a one-dimensional AR(p) process can be encoded as a $(p+1)$-dimensional VAR(1) process by maintaining a latent state representation of the form $\mathbf{x}_t = [x_t\ x_{t-1}\ \ldots\ x_{t-p}]$.

In the DSLDS, the same set of parameters needs to be learned. Also in our system, the

observation model encodes our belief that $\mathbf{x}_t$ is a noisy version of $\mathbf{y}_t$. This imposes a constrained form for the observation matrix $\mathbf{C}^{(s)}$ which consists of $0/1$ entries, which are set so as to pick the most recent value $\mathbf{x}_t$ under the VARMA representation and are also affected based on our knowledge of whether the observations $\mathbf{y}_t$ are artifactual or not under state s. A simple example is provided for clarity, where it is assumed that two channels which are both modelled as AR(1) processes are adequate to capture the event of interest. A further assumption is made that an artifactual event is being modelled, whose observed values of the first channel are artifactual, while the observed values of the second reflect the actual physiology of the patient. Under these assumptions:

$$\mathbf{C}^{(BS)} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{4.5}$$

The first row of $\mathbf{C}^{(BS)}$ sets the contribution of the first (artifact-obscured) observed channel to zero, while the second row picks the most recent value of the latent state of the second (unaffected by artifact) channel. Under this form of $\mathbf{C}^{(s)}$, the observation matrix $\mathbf{R}^{(s)}$ can be obtained as in the case of the LDS. In the LDS, the observation model encodes our belief that the generated $\mathbf{y}_t$ should be normally distributed around $\mathbf{x}_t$ with variance $\mathbf{R}$. In our discriminative version of the LDS, the observation model encodes our belief that $\mathbf{x}_t$ should be normally distributed around $\mathbf{y}_t$ with variance $\mathbf{R}$. Therefore, once the model is in state space form, $\mathbf{A}^s$, $\mathbf{Q}^s$ and $\mathbf{R}^s$ can be fit according to the maximum likelihood criterion by using the EM algorithm.

The task of determining the order of the respective ARI models is less straightforward. A practical approach was followed, as suggested in (Diggle, 1990, sec. 6.2). The partial autocorrelation function (PACF) of the stationary data (if a time series is not stationary, we make it stationary by successive differencing) was examined to provide an estimate of the appropriate model order. A clear cut-off at lag p in the PACF plot is suggestive of an AR(p) process. Clear cut-offs are rare in a real world application, in which case we looked for less clear tail-offs in the PACF plot as demonstrated in Figure 4.3 for two different channels.

4.1.4 Inference

In this work we are concerned with the task of computing the distribution $p(s_t, \mathbf{x}_t|\mathbf{y}_{1:t+r})$. According to our proposed model, $p(s_t|\mathbf{y}_{t-l:t+r})$ can be inferred at each time step via a classifier as described in Section 4.1.1. However, exact inference for $\mathbf{x}_t$ is still intractable, and we use the Gaussian Sum algorithm[3] as already described in Section 3.2.7.1. Therefore, at each time step t an approximation of $p(\mathbf{x}_t|s_t, \mathbf{y}_{1:t+r})$ is maintained as a mixture of J Gaussians. Moving one time step forward will result in the posterior $p(\mathbf{x}_{t+1}|s_{t+1}, \mathbf{y}_{1:t+r+1})$ having KJ components, which are again collapsed to J components (we use $J = 1$ for our experiments[4]) by matching moments (up to second order) of the distribution for each setting of s_t. Thus, inference in the DSLDS can be seen as a two-step process, where $p(s_t|\mathbf{y}_{t-l:t+r})$ is inferred by a discriminative classifier, and $p(\mathbf{x}_t|s_t, \mathbf{y}_{1:t+r})$ is inferred according to the Gaussian Sum algorithm. A derivation of the Gaussian Sum filtering updates for the DSLDS is given in Appendix B.

It should be noted that the FSLDS, being a generative model, can naturally handle missing observations. If a component of the observation vector is missing, then inference can proceed by ignoring its current contribution and relying only on the corresponding latent dynamics. In a discriminative model, such as the DSLDS there is no such natural mechanism. However, there are various ways in which missing data can be tackled. The simplest one perhaps is via imputation, which is deemed satisfactory in cases of few missing values. Other methods include the use of: a) *reduced models*, that effectively translates into building separate models for each pattern of missingness and b) response indicator variable augmentation, which augments the input space with indicator variables that function as an encoding mechanism for the observed variables. More details about these approaches can be found in Marlin (2008, Ch. 6). Finally, it should be mentioned that some discriminative models can handle missing data more naturally than others. For example , tree-based models can make use of surrogate splits during their construction phase, which allow for a different path to be taken during test time, should the primary split variable be missing (see e.g. Hastie *et al.*, 2009, sec. 9.2.4).

[3]The version of the Gaussian Sum algorithm presented here is also known as the second order Generalised Pseudo Bayes (GPB2) filter as mentioned in Murphy (2012, sec. 18.6.1.1).

[4]In a real-time application, speed of inference is a critical concern. Setting $J = 1$ results in an efficient approximation as this boils down to simple moment matching.

4.2 Experiments

In this section we describe experiments on two datasets comprising patients admitted to ICUs in two different hospitals, namely the NICU of ERI and the Neuro-ICU of SGH. It should be emphasised that it is highly non-trivial to obtain annotations for medical datasets as it requires the very scarce resource of experienced clinicians. Indeed, for the Neuro-ICU of SGH, the annotated data are the product of a one-year collaboration with that ICU. Physionet (Goldberger *et al.*, 2000), a freely available medical dataset, is not suitable for our identification (i.e. filtering and not prediction) task due to the fact that the only available time-series annotations are a limited set of life threatening/terminal events, for which identification would not be of practical use in the ICU, since by the time such an identification is made it would be already too late for an intervention.

For both datasets, we evaluate the performance of the DSLDS compared to the FSLDS. Note that the FSLDS has been shown in Quinn *et al.* (2009) to achieve superior results compared to more basic models such as a factorial hidden Markov model (FHMM) for the task of condition monitoring in ICUs. In Section 4.2.1 we provide a description of the various features that were used as input to the state-of-health model as described in Section 4.1.1, followed by an outline of the main characteristics of the two datasets. This section concludes by providing results on two tasks: a) inferring a patient's state of health and b) inferring a patient's underlying physiology in the presence of artifact corruption.

4.2.1 Features & Classifiers

As described in Section 4.1.1, the estimate of s_t is the output of a discriminative classifier. For both datasets, it was found (see Section 4.2.4) that using a random forest (Breiman, 2001) as our classification method (DSLDS$_{rf}$) yields the best performance. We followed suggestions for judicious selection of various tree construction parameters provided in (Hastie *et al.*, 2009, Ch. 15). For each tree in the random forest, a bootstrapped sample B of size n_{train} and dimension $\sqrt{d_{train}}$ (where n_{train} and d_{train} is the size and dimension of the feature matrix, constructed from the training dataset) was provided as input and the Gini index was used as the criterion for splitting nodes. The output of the random forest for a new test point is a prediction probability, obtained as an average of the predictions

produced by each tree, where the prediction of each tree is the proportion of the observations that belong to the positive class in the leaf node in which the test point belongs to. Experiments were also conducted with a logistic regression classifier (DSLDS$_{lr}$) but results were generally inferior compared to the random forest approach. Results (e.g. plots of specific examples, combinations of models etc.) are based on the DSLDS$_{rf}$ and whenever the subscript is omitted, the DSLDS$_{rf}$ model is implied.

A variety of features is used to capture interesting temporal structure between successive observations. At each time step, a sliding window of length $l+r+1$ is computed. For some features we also divide the window into further sub-windows and extract additional features from them. More precisely, the full set of features that are being used is: (i) the observed, raw values of the last l and future r time steps ($\mathbf{y}_{t-l:t+r}$); (ii) the slopes (calculated by ordinary least squares fitting) of the segments of that window that are obtained by dividing it in segments of length $(l+r+1)/k$; (iii) an exponentially weighted moving average of this sliding window of raw values (with a kernel, k_{mavg}, of width smaller than $l+r+1$); (iv) the minimum, median and maximum of the same segments; (v) the first order differences (of length $l+r$) of the original window; and (vi) differences of the raw values between different channels.

4.2.1.1 Acausal features

As already explained, experiments were conducted with the inclusion of a small number of future observations in the feature vector, which renders our model into an acausal system, meaning that the model's output is dependent on future observations. The reason behind this decision is that a lag was observed at the onset of an inferred regime due to the construction process of the feature vectors. Since a number, l, of previous time steps is always added to the feature vector, the classifiers require generally inputs which are already l steps within a specific regime before they can identify it as belonging to that regime's class. Feature vectors which correspond to inputs near the onset of a regime (i.e. less than l steps within it) will contain previous values corresponding (most likely) to stability and will thus be harder to classify correctly. One example of this effect can be seen in Figure 4.2, where the predictions about two blood sample events based on causal features (top bar under the plotted physiological values, labelled "causal") capture the events correctly, but with a small delay. One solution to this issue is to allow to our features a brief "peek" into the future by adding a small number, r, of future time steps. From a practical view-

point, this means that the model would operate with a delay of at least r seconds, since an output from the DSLDS at time t can be produced only after time $t+r$. Provided that r is small enough ($r \leq 10$ in experiments, see Section 4.2.4), this delay is negligible compared to the increase in performance. An example of how the issue is tackled is demonstrated in Figure 4.2 by the predictions which are based on the acausal features (middle bar under the plotted physiological values, labelled "acausal").

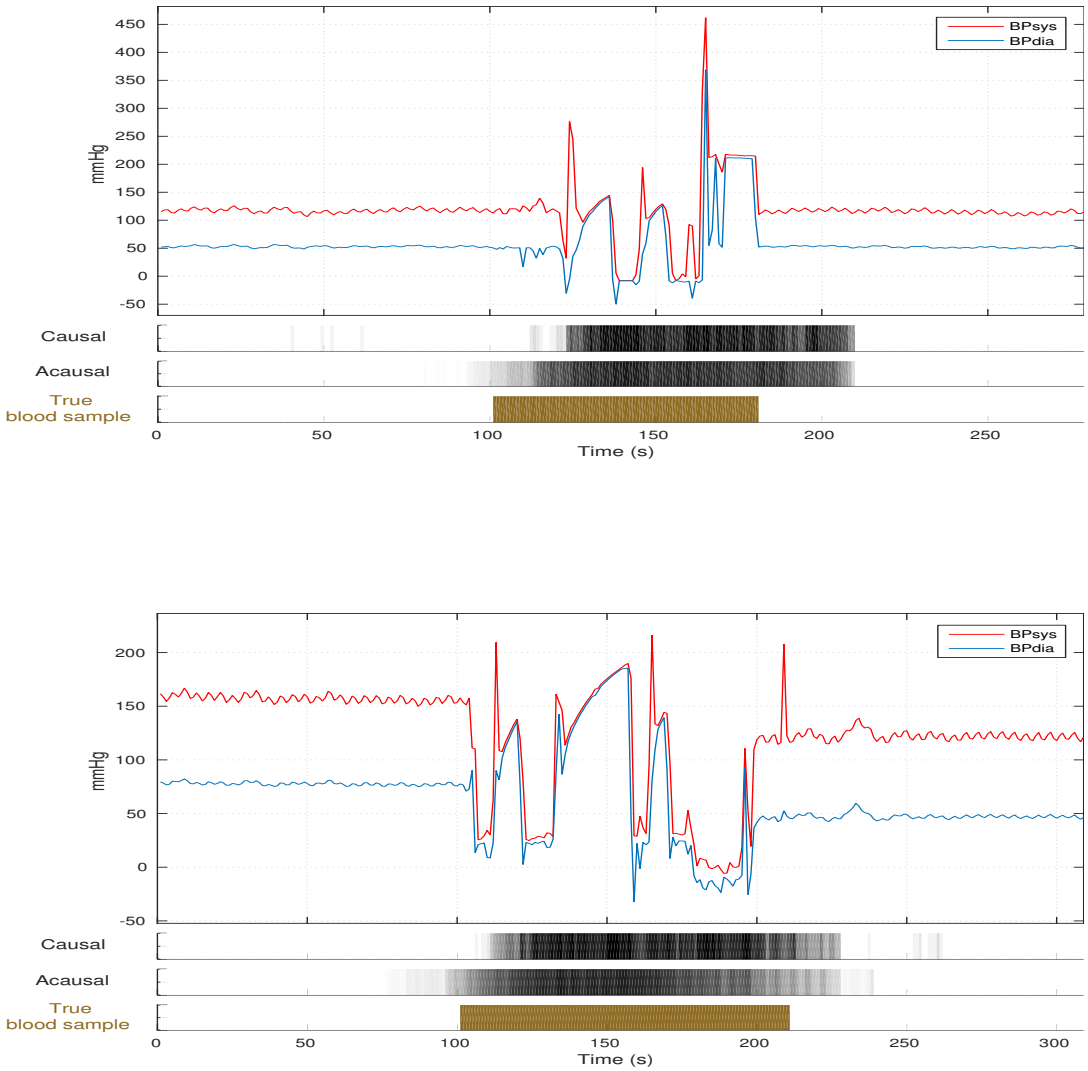

Figure 4.2: Example of DSLDS inference with causal and acausal features for two blood sample events.

4.2.2 NICU

The first dataset is obtained from the NICU of the ERI which was also the one used in Quinn *et al.* (2009)[5] and has been already described in Chapter 2. The modelled events of interest include: i) blood sample events (BS), ii) periods during which an incubator is open (IO), iii) core temperature probe disconnections (TD), iv) bradycardias (BR), and v) periods that are clearly not stable but no further identification was made by the clinicians (X). We used the same parameters for the underlying physiology model as the ones used in Quinn *et al.* (2009).

4.2.3 Neuro-ICU

The second dataset comprises data collected from the Neuro-ICU of the SGH in Glasgow and has already been described in Chapter 2. We give a brief description of the learning process for stability periods and modelled factors, which include blood sample events (BS), damped traces (DT), suction events (SC), and the X-factor (X).

Stable periods correspond to time periods when no annotation occurred from the experts, suggesting that the patient is in a stable condition. In Williams and Stanculescu (2011) it was found that a 15 minute period of stability provides an adequate amount of training data. We use the same time interval for our experiments. We found that ARI(2,1) models were adequate for all channels.

An example of a **blood sample** is shown in Figure 4.6 (bottom). Changes in BPsys and BPdia can be modelled as a four-stage process: i) the blood is diverted to a syringe for blood sampling, which causes an artifactual ramp in the observed measurements. This is similar to the blood sample model described in Quinn *et al.* (2009) and we follow the same approach here. ii) A recalibration stage follows, causing measurements to drop to zero which can be modelled similarly to a dropout event as in Quinn *et al.* (2009). iii) BP measurements continue as a stable period for a brief period. iv) The blood sample is concluded with a flushing event which causes a sharp increase in measurements. This stage is modelled as an AR(3) process for both the BPsys and BPdia channels. A total number of 64 blood sample events have been annotated, with an average duration of 1.6 minutes.

[5]The dataset has been anonymised and is available at: *www.cit.mak.ac.ug/staff/jquinn/software.html*

During a **suction event**, a significant increase in the values of all channels is observed. An AR(2) process models the HR channel (motivated by its PACF plot as shown in Figure 4.3 (left)), while AR(3) processes were used to model the remaining channels. A total number of 53 suction events have been annotated, with an average duration of 4.3 minutes.

A **damped trace**, an example of which is shown in Figure 4.6 (top), leads both BPsys and BPdia to converge to a similar mean value while at the same time the measurements exhibit high variability. Both channels were modelled with AR(3) processes (motivated by their PACF plots, of which the one corresponding to BPsys is shown in Figure 4.3 (right)). A total number of 32 damped trace events have been annotated, with an average duration of 14 minutes.

All remaining events, which are not explicitly modelled, are treated as belonging to the **X-factor** and we model them according to the X-factor model proposed in Quinn *et al.* (2009). A total number of 278 X-factor events have been annotated, with an average duration of 7.5 minutes.

Channels which are unaffected by an artifactual process (as shown in Table 4.1) are modelled as in the stable case. For example, during a blood sample event, only the BPsys and BPdia channels are affected. In that case, the latent state dynamics of these two channels evolve according to the parameters learned by the four-stage blood sample model, but the remaining channels' (i.e. HR and ICPsys) latent state dynamics continue to evolve under the model learned for stability. In every case, the parameters of the $\mathbf{x}$-state models were further optimised by EM.

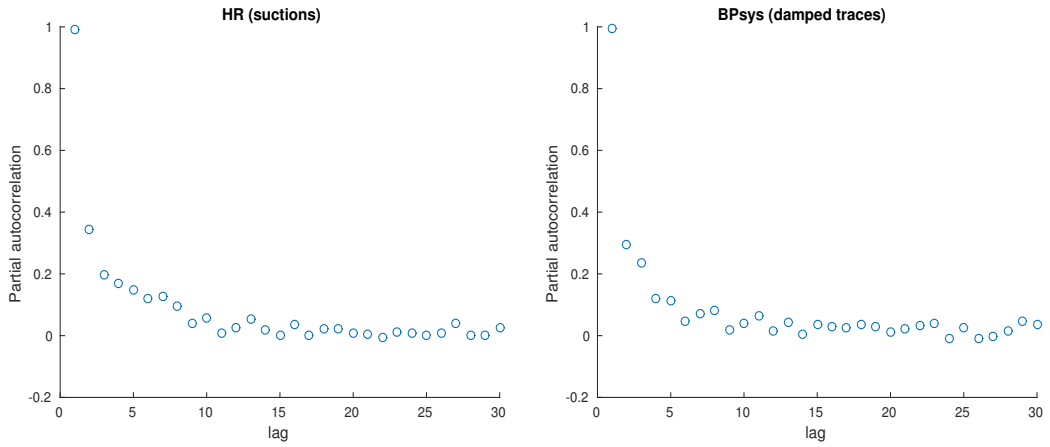

Figure 4.3: Example of partial autocorrelations for a series of lags in the case of the HR channel for suction events (left) and the BPsys channel for damped trace events (right).

Table 4.1: Channels affected by different processes for the Neuro-ICU are marked by •.

	HR	BPsys	BPdia	ICPsys
Blood sample		•	•	
Damped trace		•	•	
Suction	•	•	•	•
X-factor	•	•	•	•

4.2.4 Results

For both datasets we compare the performance of the DSLDS and the FSLDS for the task of inferring a patient's state of health. We measure the performance of the models by reporting the Area under the Receiver Operating Characteristic curve (AUC). Also, we provide plots of the Receiver Operating Characteristic (ROC) curves for the classification of the factors of interest for the DSLDS$_{rf}$, the DSLDS$_{lr}$, the FSLDS and an α-mixture of the DSLDS$_{rf}$ and the FSLDS.

We note that we treat the problem of inferring a patient's state of health as a second-by-second classification task. Given an event annotation with a start and end time, we treat each intermediate time point as belonging to that event and evaluate the models according to their classification performance for each elapsed second within the event of interest. Thus, ROC curves are constructed by treating each second-by-second observation as an instance that needs to be classified. A downside of this approach is that it treats all points within an event equally, whereas it could be argued that the start and/or end of an event are more important for their identification. An alternative approach would be to adopt an event-based evaluation, where we consider the whole period of an event as a single classification task and score our models with respect to their capability in correctly identifying whole events. Such an approach is taken in Stanculescu *et al.* (2014). However, other issues arise with this approach as well, such as what constitutes a correctly identified event as a whole (i.e. what percentage of an event needs to be identified in order to be considered as correctly classified in its entirety) and the issue of how to treat start/end times remains.

In the case of the DSLDS, the features described in Section 4.2.1 involve a number of hyperparameters that need to be chosen. Fitting them with a standard cross-validation

(CV) scheme when data are not abundant poses a non-negligible risk of overfitting. As is shown in Varma and Simon (2006), using CV to evaluate performance of a model when the model's hyperparameters have been themselves tuned using CV can lead to an optimistic bias of the estimate of the true performance. In that same work, a nested CV approach is shown to yield an almost unbiased estimate of the true performance, which is also followed in our experiments. In the outer loop the data are partitioned into P disjoint test sets. After choosing one of these partitions, the rest of the data are used in the inner loop in a standard CV setup to select the hyperparameters. The hyperparameters which yielded the highest performance (average cross-validated AUC across factors in our case) in the inner loop are then used to estimate the performance of the model on the partition (test set) in the outer loop. This process is repeated P times, once for each partition in the outer loop. For both datasets, we use leave-one-patient-out CV for the inner loop and 3-fold (with each patient's data belonging to only one fold) CV for the outer loop. In the inner loop, a grid search is performed over hyperparameters in the following sets: a) number of trees for random forest classifiers (DSLDS$_{rf}$) in $\{10, 25, 50, 100, 200\}$; b) l in $\{4, 9, 14, 19, 29, 49\}$; c) r in $\{0, 5, 10\}$. The sub-segments lengths (for slope features) were always set to $\max\{5, (l+r+1)/5\}$ and the kernel widths (for moving average features) were always set to $\max\{5, (l+r+1)/5\}$.

The FSLDS does not involve a similar hyperparameter selection process[6], and thus it is not necessary to follow the same procedure. We therefore use 3-fold CV to evaluate the FSLDS's performance.

To evaluate the α-mixture model, we have chosen the optimal α value as the one that maximises the average AUC across factors, via 3-fold CV. This also allowed us to explore the behaviour of the model as a function of α for both datasets.

4.2.4.1 NICU

In the case of the NICU we compare the two models on the full set of annotated factors reported in Quinn *et al.* (2009). Summary results are reported in Table 4.2[7]. The

[6] Except for the orders of the ARI processes which are chosen as described in Section 4.1.3.

[7] The FSLDS results were obtained using code provided by Quinn *et al.* (2009) with the same parameters as the ones mentioned there. The results are very close with the exception of the core temperature disconnection factor (for which the reported AUC in Quinn *et al.* (2009) was 0.79, while we obtained a value of 0.88), and the blood sample factor (for which the reported AUC in Quinn *et al.* (2009) was 0.96, while we obtained a value of 0.92).

DSLDS$_{lr}$ cannot match the performance of the other two models, with the exception of the blood sample factor, where its performance is higher than the FSLDS and on par with the DSLDS$_{rf}$. Its generally lower performance is not surprising since the model is linear. The DSLDS$_{rf}$ outperforms the FSLDS in three out of the four clinically identified factors. The difference in favour of the DSLDS$_{rf}$ is clear for bradycardias and blood samples, but less pronounced for core temperature disconnections. The FSLDS achieves slightly higher performance in the case of the incubator open factor, and clearly outperforms the DSLDS$_{rf}$ in the case of the X-factor. The FSLDS models the presence of outliers by the inclusion of an extra factor, which is essentially governed by the same parameters as stability with the only difference being that the system noise covariance is an inflated version of the respective covariance of the stability dynamics (for more details, see Quinn *et al.*, 2009). Such an approach has the potential to address the issue of outlier detection in a more general and thus more satisfactory way. In the case of the DSLDS, our approach is to collectively treat all abnormal events, other than the ones attributed to known factors, as an "X-class" and build a binary classifier to distinguish that class. As the training datapoints for this class are highly inhomogeneous in terms of shared discriminative features, and test points belonging to the X-class may not exhibit a high degree of similarity to the training set, it is not surprising that the DSLDS may perform rather poorly for the X-factor. However, by considering an α-mixture of the two models, we can combine the discriminative power of the DSLDS$_{rf}$ for known factors with the increased performance of the FSLDS for the X-factor, thus achieving a higher performance (bottom line of Table 4.2) compared to considering the two models separately. The behaviour of the α-mixture model as a function of α is shown in Figure 4.4. The optimal α-mixture ($\alpha = 0.5$) yields the best average AUC across factors (in fact, $\alpha = 0.5$ yields optimal performance for each factor separately except bradycardia, where it is almost optimal) compared to all other considered α values and also outperforms the DSLDS and the FSLDS in all cases except for the bradycardia factor, where the DSLDS$_{rf}$ performs slightly better.

In a real-world setting, a choice has to be made with respect to a threshold, according to which the model would classify an example as belonging to the positive or negative class. A ROC curve thus can be more informative than the AUC since it depicts a model's classification behaviour as this threshold varies. In Figure 4.5, the ROC curves for the classification of each factor are shown. The DSLDS$_{rf}$ dominates (i.e. always achieves higher classification accuracy regardless of the choice of threshold) almost entirely in the ROC curve space over the FSLDS for both the blood sample and bradycardia factors,

Table 4.2: Comparison of DSLDS, FSLDS and α-mixture performance for the NICU dataset. Optimal value of the α parameter is shown inside parenthesis.

AUC	BS	IO	TD	BR	X
DSLDS_{rf}	0.98	0.83	0.90	0.94	0.57
DSLDS_{lr}	0.98	0.75	0.73	0.81	0.51
FSLDS	0.92	0.87	0.88	0.85	0.66
$\alpha\text{-mixture}^{(0.5)}$	0.98	0.89	0.93	0.92	0.67

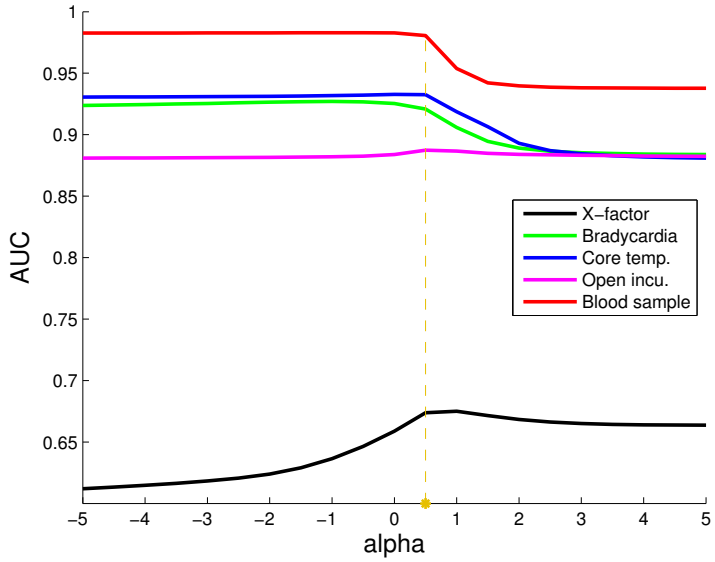

Figure 4.4: Performance of the α-mixture models as a function of α ($step = 0.25$) for the NICU dataset. The asterisk marks the optimal value for α.

while the FSLDS dominates in the case of the X-factor. In the case of the open incubator factor, the FSLDS dominates above a threshold choice which would correspond to a false positive rate (FPR) of approximately 0.1, while below this threshold the DSLDS_{rf} exhibits a slightly better performance. This picture is effectively reversed in the case of the temperature probe disconnection factor, where the DSLDS_{rf} dominates for threshold choices corresponding to FPRs ranging in 0.25–0.65, while the FSLDS dominates outwith that interval. In concordance to the AUC results, the α-mixture model dominates in ROC curve space over all other models, with the exception of bradycardias, where the DSLDS_{rf} is the better choice.

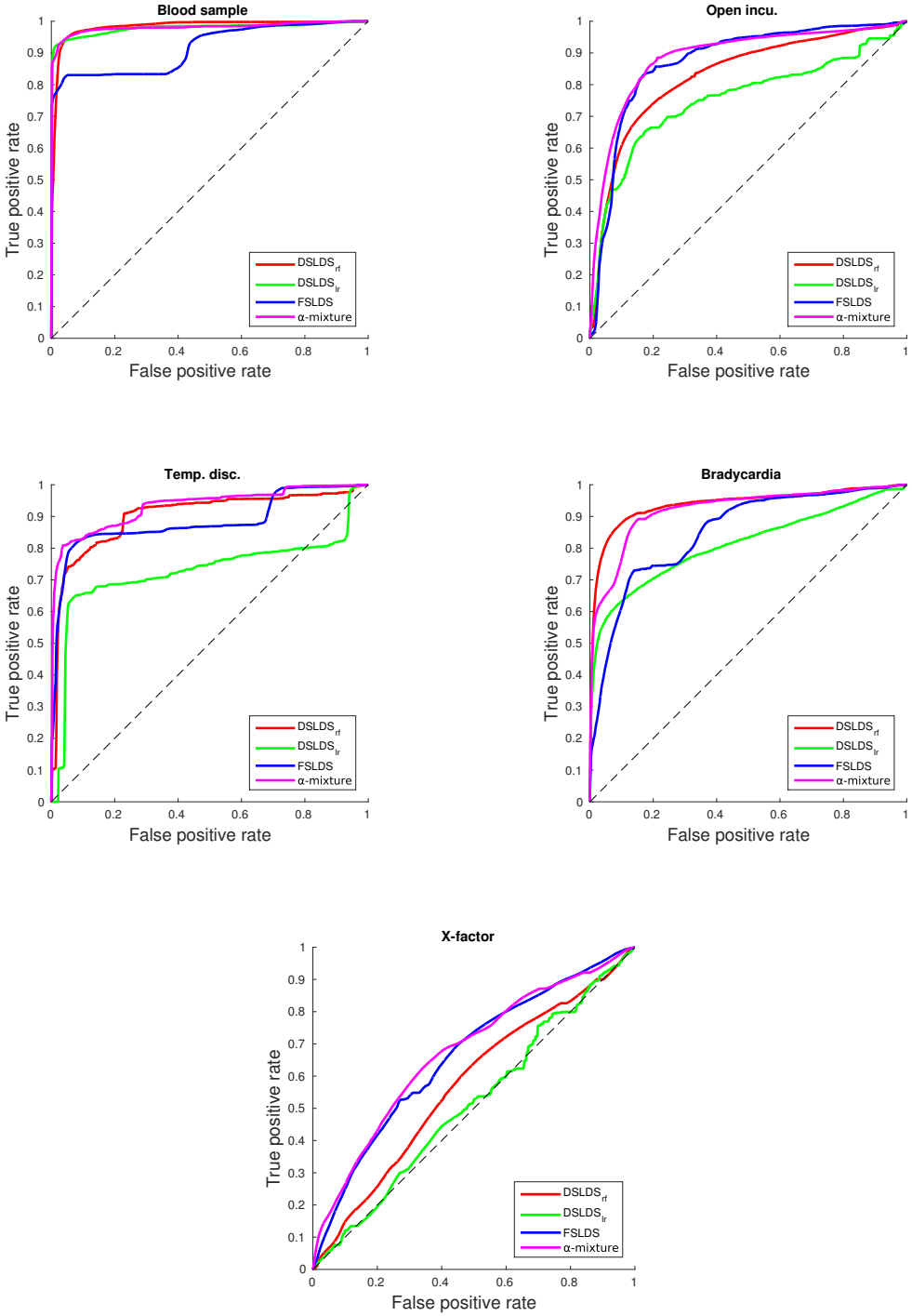

Figure 4.5: ROC curves for each modelled factor in the case of the NICU, using the DSLDS (red line for DSLDS$_{rf}$, green line for DSLDS$_{lr}$), the FSLDS (blue line), and their α-mixture (magenta line). The dashed diagonal line corresponds to a random classifier.

4.2.4.2 Neuro-ICU

In the case of the Neuro-ICU, inferences for two example events are shown in Figure 4.6. In the top, a damped trace event is shown, which lasts for almost one hour before being resolved by a flushing event (spiking of both channels). The DSLDS accurately identifies the damped trace event, while the FSLDS fails to detect it, but hypothesises several incorrect blood sample events instead. In the bottom panel a blood sample event is shown, where the multiple stages are clearly visible. The event starts with two artifactual ramps, followed by a flushing, a zeroing, and finally with another flushing. This is slightly different than the description we have already given, but slight deviations from the standard protocol due to human error is to be expected. In this case, both models manage to capture the event in a generally satisfactory manner. Summary results are reported in Table 4.3. The DSLDS$_{rf}$ outperforms the FLSDS on all of the known factors. The damped trace and suction events particularly are characterised by high variability which is hard to capture with a generative process. However, simple discriminative features are able to capture them with higher accuracy. As was expected, the FSLDS achieves a higher AUC for the X-factor. Similarly to the case of the NICU the DSLDS$_{lr}$ exhibits a lower performance compared to the DLSDS$_{rf}$, but manages to outperform the FSLDS in two out of the four factors, reinforcing the belief that (even simple, linear) discriminative models are more appropriate for the investigated tasks.

Table 4.3: Comparison of DSLDS, FSLDS and α-mixture performance for the Neuro-ICU dataset. Optimal value of the α parameter is shown inside parenthesis.

AUC	BS	DT	SC	X
DSLDS$_{rf}$	0.96	0.93	0.67	0.65
DSLDS$_{lr}$	0.88	0.91	0.62	0.54
FSLDS	0.95	0.79	0.57	0.74
α-mixture$^{(0)}$	0.99	0.94	0.70	0.71

Again, the optimal α-mixture ($\alpha = 0$) outperforms the DSLDS$_{rf}$ and the FSLDS in all cases except for the X-factor, where the FSLDS achieves a slightly higher AUC. Contrary to the NICU dataset, as shown in Figure 4.7 there are alternative α values which can yield higher AUC across different factors. For example, an X-factor AUC value of 0.76 can be obtained by setting $\alpha = 5$. However, apart from the superior (on average) performance

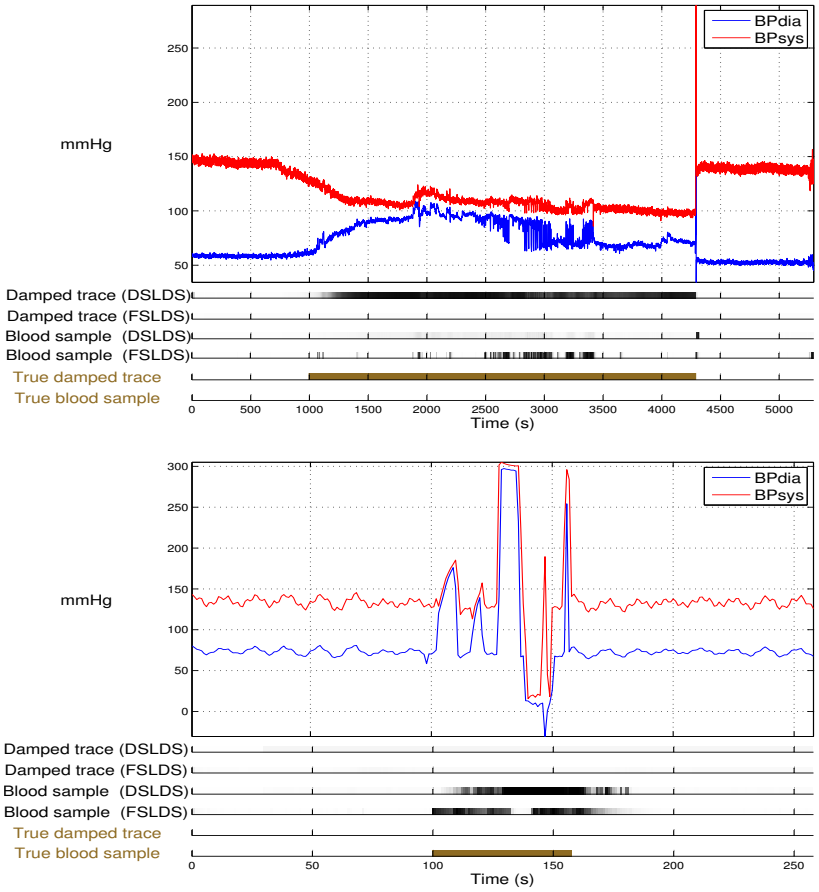

Figure 4.6: Example of DSLDS and FSLDS inferences for a damped trace event (top) and a blood sample event (bottom).

of the α-mixture, another appealing property is that α could be treated as a user-tunable parameter. In a practical setting, the model could be preset with the optimal α value, but a clinician could decide, for example, to make the model focus on maximising its predictive performance on the X-factor (or some important physiological factor like bradycardia) to the potential detriment of other factors. Then the model could adjust its α parameter in real-time based on training data results to maximise its performance on the desired factor.

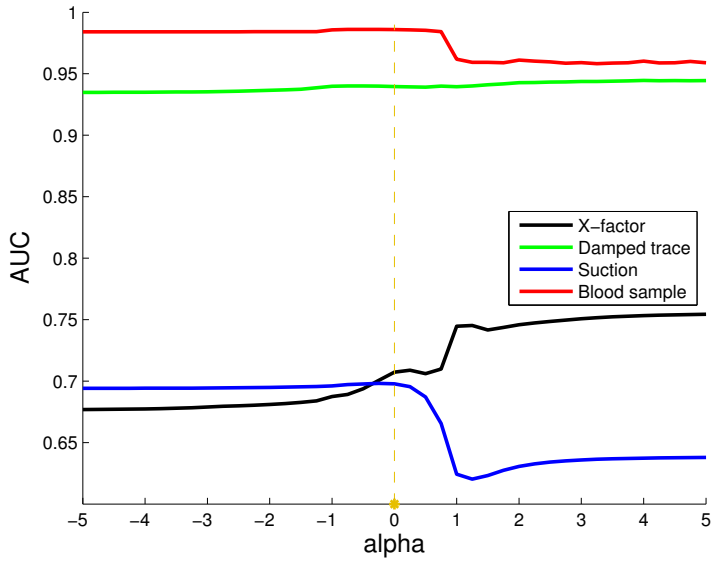

Figure 4.7: Performance of the α-mixture models as a function of α ($step = 0.25$) for the Neuro-ICU. The asterisk marks the optimal value for α.

In Figure 4.8, the ROC curves for the classification of each factor are shown. The DSLDS$_{rf}$ dominates almost entirely in the ROC curve space over the FSLDS for the suction (for a threshold choice corresponding to FPR > 0.15) and damped trace factors, while the FSLDS dominates in the case of the X-factor. In the case of the blood sample factor, the DSLDS$_{rf}$ dominates until a threshold choice which would correspond to an FPR of approximately 0.25, while over this threshold the results are reversed in favour of the FSLDS. As expected, the α-mixture model dominates in ROC curve space over all other models, with the exception of the X-factor (and the damped trace factor for thresholds corresponding to FPRs < 0.1, where interestingly, the DSLDS$_{lr}$ is the best choice), where the FSLDS achieves a higher performance.

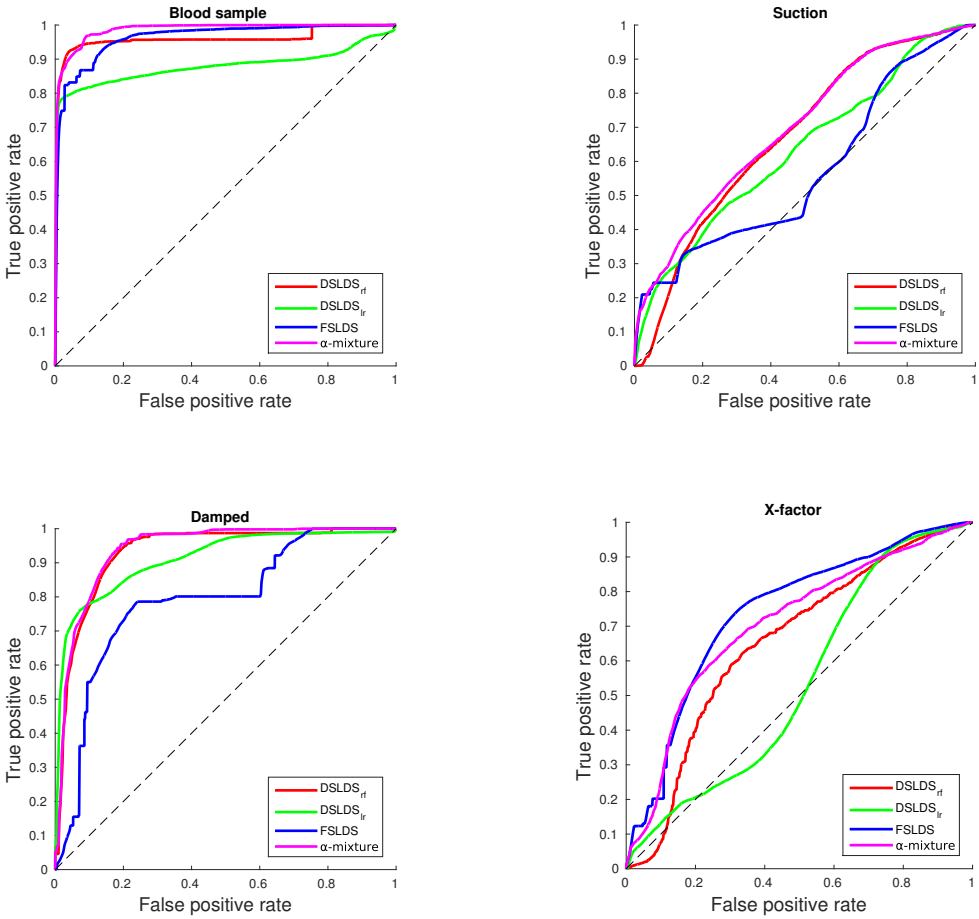

Figure 4.8: ROC curves for each modelled factor in the case of the Neuro-ICU, using the DSLDS (red line for DSLDS$_{rf}$, green line for DSLDS$_{lr}$), the FSLDS (blue line), and their α-mixture (magenta line). The dashed diagonal line corresponds to a random classifier.

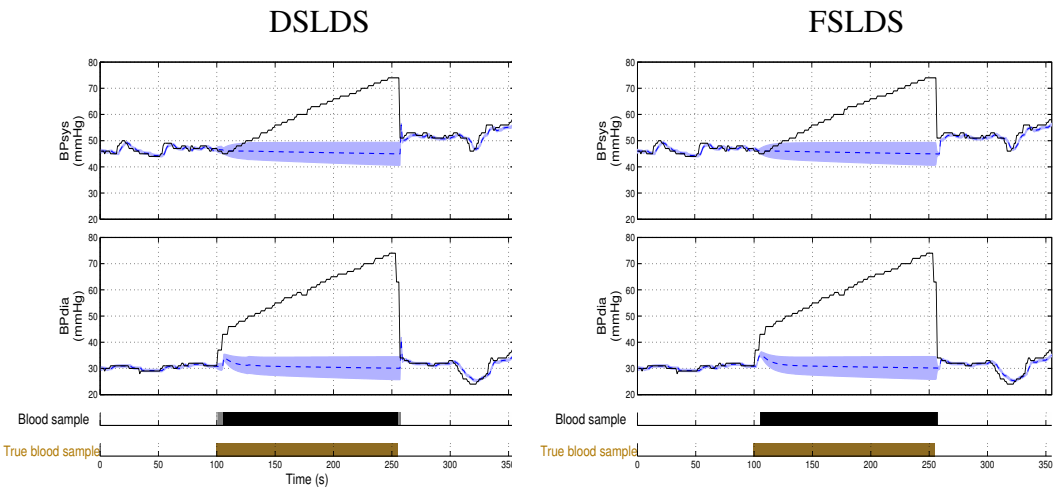

Figure 4.9: Example of the inferred underlying physiology in the presence of a blood sample in the case of the DSLDS (left) and the FSLDS (right). The solid line corresponds to the actual observations, while the estimated true physiology is plotted as a dashed line with the shaded area indicating two standard deviations.

4.2.4.3 Inference for x-state

Finally, Figure 4.9 shows the inferred distribution of underlying physiology during a blood sample taken from a neonate for both models. In both cases, estimates are propagated with increased uncertainty under the correctly inferred artifactual event. Note a small difference at the start of the event: The DSLDS partially identifies the event causing an increase in uncertainty, while the FSLDS (incorrectly) identifies this part as stable and thus its $\mathbf{x}$-state update exhibits lower uncertainty. Maintaining an estimate of the underlying vital signs in the presence of artifacts can then be used for data imputation. Another use, which has been deemed important by clinical experts, is that such an estimate can help doctors maintain an approximate view of a patient's underlying physiology during artifactual events that would otherwise completely obscure a patient's vital signs. This can be crucial during treatment of a patient under critical conditions, such as the ones found in an ICU.

4.3 Summary

In this chapter, a discriminative approach has been presented for the application of patient monitoring in ICUs. It was shown that the proposed discriminative approach is able to

outperform the previous generative approach used for the same task in most of the investigated cases. Additionally, the DSLDS provides a more data-driven framework within which it is more straightforward and less time consuming to model a patient's state of health compared to the generative approach for which a very detailed understanding of the associated factors based on expert knowledge is a prerequisite. It was also shown that an α-mixture of the two approaches yields better results than either model separately. In our approach we have assumed that the prediction of the switch variable factorises over the state space. However, one could use a structured output model to predict the joint distribution of different factors. For example, one could set up a graphical structure to model factors that are mutually exclusive, so that one factor's presence could be explained away in the presence of another, competing factor or potentially set up a hierarchy of factors motivated by domain knowledge where factors higher in the hierarchy could affect the dynamical behaviour of factors at the lower levels as is the case in Stanculescu *et al.* (2014).

Finally, another issue is the lack of explicit temporal continuity in the *s*-chain. Implicitly, this is handled by the feature construction process. However, a future direction could be to establish a Markovian connection on the *s*-chain too and compare with the current approach.

Chapter 5

Input-Output Non-Linear Dynamical Systems

In the previous chapter, the development of the DSLDS was motivated by the problem of identifying clinical events of interest (mainly artifactual) and capturing the underlying physiology of a patient in the presence of artifact. In this chapter, the focus is centred on the problem of predicting the effect of infused drugs on the vital signs of a patient. In the biomedical literature this problem is broken into pharmacokinetics (PK) and pharmacodynamics (PD). As already mentioned in Section 3.3, pharmacokinetics is concerned with the behaviour of drugs once they enter a patient's body, while pharmacodynamics addresses the biochemical and physiological effects of drugs on the body. As already described in the same section, PK models are typically expressed as sets of ordinary differential equations (ODEs), while PD models are typically nonlinear functions relating drug concentration to observed vital signs.

From a machine learning point of view, these models consist of a linear dynamical system with control inputs (the drug infusion rates), and a non-linear output model. The PK models are based on quite a number of assumptions that are arguably not highly accurate representations of a patient's physiological reaction to drug administration (see Section 3.3 for more details). Thus this chapter's contribution is to take a more "data-driven" approach to the problem, fitting an input-output non-linear dynamical system (IO-NLDS[1]) to predict the vital signs based on input drug infusion rates. A notable difference to the PK/PD approach is that the latent process is not constrained to be mapped to any physio-

[1] Inspired by the IO-HMM of Bengio and Frasconi (1995).

logically interpretable quantity and the model is free to learn any latent representation that might better explain the observed data. The results show clear improvements in performance over the PK/PD approach. As a second contribution, the IO-NLDS model is fitted using an unscented filter (Wan and Van Der Merwe, 2000) within an EM algorithm. To the best of the author's knowledge this is the first time such an approach has been taken for a system including control inputs on a real-world dataset with several parameters.

The structure of the remainder of this Chapter is as follows: in Section 5.1 a description of the PK/PD model and the IO-NLDS is provided, their structure is compared, and details about inference and learning are given. In Section 5.2 follows a description of experiments and results are provided on the comparison between the two methods. Finally, Section 5.3 concludes with general remarks about the proposed model and suggestions for future work. Parts of this chapter have been adapted from Georgatzis *et al.* (2016b).

5.1 Model description

In this section, details about the PK/PD model and the IO-NLDS are provided. In Section 5.1.1 the standard formulation of the PK/PD approach is described, motivated by the relevant literature. In Section 5.1.2 the IO-NLDS is described, while in Section 5.1.3 the PK/PD model is cast as a probabilistic NLDS. Sections 5.1.4 and 5.1.5 provide details about inference and learning in the IO-NLDS setting.

5.1.1 PK/PD model

A description of the standard PK/PD approach which is based on compartmental modelling has already been given in Section 3.3 and what follows here is a more detailed exposition of this approach focussing on the anaesthetic drug Propofol. As already explained, compartmental models are an abstraction used to describe the rate of change of a drug's concentration in a patient by accounting for the processes of absorption, distribution, metabolism and excretion of the drug in different parts (compartments) of the human body. This approach then involves building a system of ODEs that describe the evolution of drug concentrations at different compartments. The standard compartmental model for modelling the pharmacokinetical properties of Propofol is comprised of three compartments and dates back to Gepts *et al.* (1987). Based on that work, a model widely

used in practice is the one introduced by Marsh *et al.* (1991) which has been further improved upon by White *et al.* (2008). This line of work only addresses the PK aspect of the task. In order to quantify the effect of the drug on the observed vital signs, one needs to add an extra "effect" compartment for each observed vital sign, and link the concentration at the effect compartment with the observed physiology as in e.g. Bailey and Haddad (2005).

A graphical representation of this overall PK/PD approach is shown in Figure 5.1[2], where x_i is the concentration of drug in compartment i and k_{ij} is the drug's transfer rate from compartment i to j.[3]. Tne functional relationship between x_e and the observed vital sign y can be modelled via a generalised logistic function (also known as Richards' curve; see Richards, 1959) of the form:

$$g(x_e) = m + \frac{(M-m)}{(1+e^{-\gamma x_e})^{(1/\nu)}} \ , \tag{5.1}$$

where the parameters m, M govern the lower/upper asymptote respectively, γ controls the decrease rate and ν determines near to which asymptote maximum decrease occurs[4]. The whole model can be described by a system of linear ODEs with the addition of a generalised logistic function as follows:

$$\begin{aligned} dx_{1t}/dt &= -(k_{10}+k_{12}+k_{13})x_{1t} + k_{21}x_{2t} + k_{31}x_{3t} + u_t \ , \\ dx_{2t}/dt &= k_{12}x_{1t} - k_{21}x_{2t} \ , \\ dx_{3t}/dt &= k_{13}x_{1t} - k_{31}x_{3t} \ , \\ dx_{e_qt}/dt &= k_{1e_q}(x_{1t} - x_{e_qt}) \ , \\ y_{qt} &= g(x_{e_qt}) \ , \end{aligned} \tag{5.2}$$

where $q \in \{1,2,3\}$. An important aspect of this model is that the parameters associated with the PK part (i.e. the set of k_{ij}'s) are considered fixed and their values have been

[2]This representation is equivalent to the one presented in Figure 3.9 but with three effect sites instead of one.

[3]Parameter k_{1e} is known as k_{e0} in the PK literature.

[4]This form of the generalised logistic function with four parameters was used to mirror the Hill function (see Eq. 3.37). We also note that the generalised logistic function was used instead of the Hill function due to its ability to handle negative values.

determined by Marsh *et al.* (1991) based on principles of human physiology. In White *et al.* (2008) it was established that this PK model could be improved by allowing k_{10} to vary according to a patient's age and gender. Hence, in this work, the improved version of White *et al.* (2008) is followed.

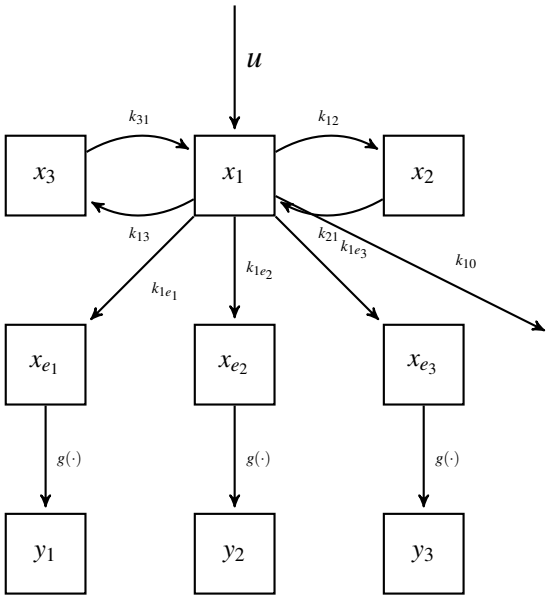

Figure 5.1: A three compartment model for Propofol with one added effect site compartment. See text for details.

5.1.2 IO-NLDS

In contrast to the PK/PD an alternative approach is adopted which makes no explicit use of expert physiological knowledge and is not restricted in associating the model's parameters with physiologically relevant quantities. Instead, the minimal assumptions made are that a latent temporal process with linear dynamics driven by control inputs (drug infusion rates), gives rise to observed quantities (vital signs) which are non-linearly dependent on the latent process. This gives rise to an input-output non-linear dynamical system (IO-NLDS), which can be thought of as an LDS with control inputs and a generalised logistic function (see Eq. 5.1) as its observation model and is further defined by the following joint distribution:

$$p(\mathbf{x},\mathbf{y}|\mathbf{u}) = p(\mathbf{x}_1|\mathbf{u}_1)p(\mathbf{y}_1|\mathbf{x}_1)\prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{u}_t)p(\mathbf{y}_t|\mathbf{x}_t)\ , \tag{5.3}$$

where $\mathbf{x}_t \in R^{d_x}$, $\mathbf{u}_t \in R^{d_u}$ and $\mathbf{y}_t \in R^{d_y}$ denote the latent states, control inputs and observed vital signs at time t, and T denotes the total length of the observed vital signs. It is further assumed that the random variables are distributed according to:

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}_t, \mathbf{Q})\,, \tag{5.4}$$

$$\mathbf{y}_t \sim \mathcal{N}(g(\mathbf{C}\mathbf{x}_t), \mathbf{R})\ . \tag{5.5}$$

The graphical model corresponding to these equations is shown in Figure 5.2. Equation (5.5) encodes the assumption that the latent state is linearly projected onto the observation space via matrix **C**, is subsequently non-linearly transformed via the generalised logistic function $g(\cdot)$, and corrupted by Gaussian observation noise with covariance **R** to give rise to the observations. Function $g(\cdot)$ is parameterised as in the PK/PD approach. The latent state itself is following linear dynamics, governed by matrix **A** and a linear transformation (via matrix **B**) of control inputs, and additive Gaussian noise with covariance **Q**. A notable difference to the PK/PD approach is that the latent process is not constrained to be mapped to any physiologically interpretable quantity and the model is free to learn any latent representation that might better explain the observed data. In contrast to the PK/PD model, the latent process exhibits a higher degree of flexibility, being unconstrained of any (simplifying) physiologically motivated assumptions and can thus be expected to model the patient-specific underlying dynamics in a more expressive way.

5.1.3 PK/PD model as NLDS

The PK/PD model as described in Section 5.1.1 does not incorporate any uncertainty while the IO-NLDS is a probabilistic model. In order to compare the two models, the PK/PD model is cast into the IO-NLDS form as described in eqs. (5.4), (5.5). This involves two steps: a) the discretisation of the continuous time dynamics as described by the system of ODEs in the first four equations of eq. (5.2) and b) the addition of

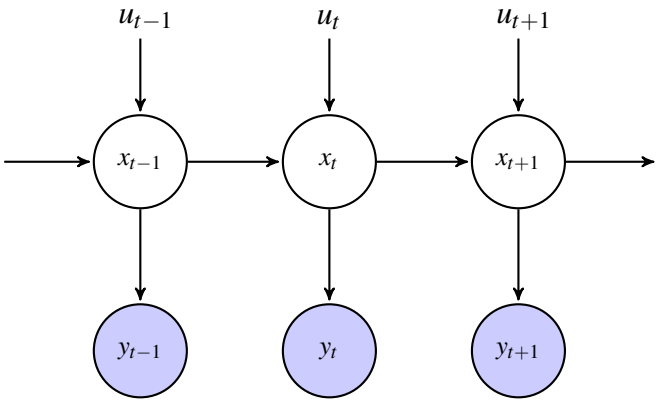

Figure 5.2: Graphical model of IO-NLDS. The latent physiological state of a patient and the drug infusion rates at time t are denoted by $\mathbf{x}_t$ and $\mathbf{u}_t$ respectively. The shaded nodes correspond to the observed physiological values, $\mathbf{y}_t$.

Gaussian noise on the discretised dynamics and on the non-linear output. This can be done by setting the following parameters (assuming here $d_y = 1$ for compactness) as $\mathbf{A} = \exp\{\mathbf{F}\Delta t\}, \mathbf{B} = [1\ 0\ 0\ 0]^\top, \mathbf{C} = [0\ 0\ 0\ 1]$, where:

$$\mathbf{F} = \begin{bmatrix} -(k_{10}+k_{12}+k_{13}) & k_{21} & k_{31} & 0 \\ k_{12} & -k_{21} & 0 & 0 \\ k_{13} & 0 & -k_{31} & 0 \\ k_{1e} & 0 & 0 & -k_{1e} \end{bmatrix}.$$

The dynamics matrix $\mathbf{A}$ is the discretised version of its continuous time counterpart $\mathbf{F}$, which involves computing the matrix exponential of $\mathbf{F}$ times the discretisation step Δt as described e.g. in Aström and Murray (2010, sec. 5.3). Matrix $\mathbf{F}$ is made of the continuous-time parameters that govern the system of ODEs presented in eq. 5.2. It is noted that the PK model provided by White *et al.* (2008) provides estimates for the parameters involved in $\mathbf{F}$, except for k_{1e}. To fit an appropriate k_{1e} a fine-grained one-dimensional grid search per observed channel around a clinically relevant value was performed. The noise matrices $\mathbf{Q}$ and $\mathbf{R}$ have the same interpretation as in the case of the IO-NLDS and can be learned in the same way, as described in Section 5.1.5. Under this form, the PK/PD model can be seen as an IO-NLDS with a constrained parametric form, where the parameters $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are constrained in such a way as to capture the physiological processes involved with

the infusion of an anaesthetic drug, as already described in Section 5.1.1.

5.1.4 Inference

If the observation model was linear, then the overall model would be a linear dynamical system (LDS) and exact filtering would be feasible via the well-known Kalman filter (Kalman, 1960). With the addition of the nonlinear observation function $g(\cdot)$, exact inference becomes intractable and one must resort to some form of approximation. Here, a sigma-point filtering approach is adopted, and more specifically the unscented Kalman filter (UKF) is used, as described in Särkkä (2013).

The main component of UKF is the unscented transform (UT) as presented by Julier and Uhlmann (1996) (although the term "unscented" was introduced later). The UT is also described in Appendix C. The idea behind it is that a non-linear transformation of a Gaussian distribution can be approximated by first deterministically selecting a fixed number of points (called sigma points) from that distribution to capture its mean and covariance, then computing the exact non-linear transformation of these points, and subsequently fitting a Gaussian distribution to the non-linearly transformed points. Under this procedure, the UKF reduces to applying the UT twice at each time step t: a) once to compute the predictive density $p(\mathbf{x}_t|\mathbf{y}_{1:t-1}, \mathbf{u}_{1:t})$, by using the UT on the previous filtered density $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}, \mathbf{u}_{1:t-1})$ and b) to compute the current filtered density $p(\mathbf{x}_t|\mathbf{y}_{1:t}, \mathbf{u}_{1:t})$ by using the UT on the predictive density to compute the current likelihood $p(\mathbf{y}_t|\mathbf{x}_t)$ and then using Bayes rule to obtain the required filtered (posterior) density. These two steps constitute the standard predict-update cycle that also forms the basis of inference in a LDS via the KF as seen in Section 3.2.5.1. Since the model is governed by linear dynamics, the first step can be calculated by the standard Kalman filter equivalent step with both methods yielding the same results.

The task of inference in non-linear dynamical models has been thoroughly explored and methods such as the extended Kalman filter (see e.g. Särkkä, 2013, sec. 5.2), the UKF (Wan and Van Der Merwe, 2000) and the particle filter (Gordon *et al.*, 1993) have been proposed. The UKF is shown empirically to outperform the extended KF (see e.g. Haykin, 2001, Ch. 7), and was chosen over the particle filter (PF) because the PF can require "orders of magnitude" (see Wan and Van Der Merwe, 2000) more sample points compared to the UKF's fixed, small number of sigma points to achieve high accuracy, rendering it

computationally prohibitive for a real-time application such as ours.

5.1.5 Learning

The parameters of the proposed model that need to be learned are $\boldsymbol{\theta} = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \boldsymbol{\eta}\}$, where it is assumed that $\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{\eta} = \{\mathbf{m}, \mathbf{M}, \boldsymbol{\gamma}, \boldsymbol{\nu}\}$ are the parameters of $g(\cdot)$. Maximum likelihood (ML) estimates of these parameters are learned by using the expectation maximisation (EM) algorithm (Dempster *et al.*, 1977). EM then maximises the likelihood via maximising the following surrogate function:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \mathbb{E}_{p(\mathbf{x}|\mathbf{y},\mathbf{u},\boldsymbol{\theta}^{old})}[\log p(\mathbf{x}, \mathbf{y}|\mathbf{u}, \boldsymbol{\theta})] \,, \tag{5.6}$$

which, due to the Markov properties of the model, can be further decomposed as:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = {} & \mathbb{E}_{p(\mathbf{x}_1|\mathbf{y}_{1:T},\mathbf{u}_{1:T},\boldsymbol{\theta}^{old})}[\log p(\mathbf{x}_1|\mathbf{u}_1, \boldsymbol{\theta})] \\ & + \mathbb{E}_{p(\mathbf{x}_t,\mathbf{x}_{t-1}|\mathbf{y}_{1:T},\mathbf{u}_{1:T},\boldsymbol{\theta}^{old})}[\log p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{u}_t, \boldsymbol{\theta})] \\ & + \mathbb{E}_{p(\mathbf{x}_t|\mathbf{y}_{1:T},\mathbf{u}_{1:T},\boldsymbol{\theta}^{old})}[\log p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta})] \,. \end{aligned} \tag{5.7}$$

Expectations with respect to the smoothing distributions $p(\mathbf{x}_t|\mathbf{y}_{1:T}, \mathbf{u}_{1:T}, \boldsymbol{\theta}^{old})$ and the pairwise joint smoothing distributions $p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{y}_{1:T}, \mathbf{u}_{1:T}, \boldsymbol{\theta}^{old})$ are involved in these terms. These distributions can be computed in general via the unscented Rauch–Tung–Striebel (URTS) smoother as described in Särkkä (2008). In this case, since the system's dynamics are linear, the backward smoothing step can be performed by the standard RTS smoother after obtaining the unscented filtered estimates during the forward filtering step via the UKF. Analogously to the UKF, where the UT is used to approximate the required expectations, the UT needs to be employed here to approximate the expectation appearing in the last term of the RHS of eq. (5.7). More details are given in Kokkala *et al.* (2014).

Computing the required distributions and expectations in order to calculate eq. (5.7) constitutes the E-step of EM. During the M-step the model parameters are set such that

$\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$. This maximisation step can be done analytically in the case of the LDS, but one needs to make use of numerical optimisation methods in the non-linear case. In this work, the subset of parameters $\boldsymbol{\theta}_L^* = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \mathbf{A}, \mathbf{B}, \mathbf{Q}\}$ which correspond to the linear part of the model can be computed in closed form as shown in Cheng and Sabes (2006), and $\mathbf{R}^*$ can be computed similarly to the linear case as shown in Särkkä (2013, sec. 12.3.3). These ML estimates are provided in Appendix C. The remaining set of parameters $\boldsymbol{\theta}_{NL}^* = \{\mathbf{C}, \boldsymbol{\eta}\}$ can be then optimised via numerical optimisation. For this work's experiments the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (see e.g. Fletcher, 2013) is used. It should be noted that EM is a natural choice in this case since the model involves a linear sub-component which can be exploited in the decomposition of eq. (5.7) to derive a subset of parameters in closed form. In a fully non-linear case however one could use a numerical optimisation procedure to directly maximise the likelihood function instead of a surrogate function, as argued in Kokkala *et al.* (2015). Finally, the subset $\{\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}\}$ of the IO-NLDS's parameters is initialised via the spectral learning approach of Siddiqi *et al.* (2007) which has been described in Section 3.2.5.2.

In general, learning is not as thoroughly explored as inference in NLDSs but in Hagenblad *et al.* (2008) a gradient-based iterative search method combined with numerical integration is used for deriving maximum likelihood estimates of the parameters of a *Wiener*[5] model, while in Schön *et al.* (2011) and Wills *et al.* (2011) the EM algorithm is used in conjunction with PF for parameter estimation. In Wills *et al.* (2013) the same method is used for parameter estimation in a *Hammerstein-Wiener*[6] model. Finally in Gašperin and Juričíc (2011) and Kokkala *et al.* (2014) the EM is used in conjunction with the UKF for the same task, but they do not include control inputs in their formulation as we do.

5.2 Experiments

This section contains an exposition of experiments that were conducted in order to establish if the newly proposed approach can model accurately the effect of drug infusions on the observed physiology of patients in ICUs. To this end, the IO-NLDS and the PK/PD

[5] In the control theory field, LDSs with nonlinear observation models are known as Wiener models (Schetzen, 1980).

[6] A static non-linearity connecting the input to the latent state of the model is known as a Hammerstein model (Narendra and Gallman, 1966). A Hammerstein-Wiener model is the combination of a Wiener and a Hammerstein model.

model are compared with respect to the task of predicting the effect of Propofol on patients' vital signs. Also, an additional comparison is conducted between the IO-NLDS and a more direct approach that does not make use of a latent space.

5.2.1 Data description

A general description of vital signs data under the effect of drug infusions has already been given in Section 2.3.2. Here, a more detailed description is provided pertaining to the ensuing experiments. The dataset comprises of 40 Caucasian patients admitted in the Neuro-ICU of the GJNH. All patients were spontaneously breathing with the maximum airway intervention being an oropharyngeal airway, and none was classified higher than Class 2 with respect to illness severity according to the American Society of Anesthesiologists.

Two controlled Propofol infusion protocols were investigated pre-operatively on these patients during a period of approximately 45 minutes. This investigation was part of an independent clinical study and the data have been anonymised. Each patient was randomly assigned to one of the two protocols. The first protocol involved a target Propofol concentration of 2 μg/ml for the first 15 minutes followed by a target concentration of 5 μg/ml for the next 15 minutes and 2 μg/ml for the last 15 minutes. The second protocol was the inverse of the first with a 5–2–5 μg/ml target sequence. The drug pumps which automatically administer Propofol calculate internally the desired infusion rates and these rates were used as control inputs of the model. Propofol doses are at peak during the early phase of the protocol making it very likely that the maximum haemodynamic effects will be seen during the study period and not after discontinuation of the drug. Also, during the study very few interventions were required in the case of hypotension (treated by small incremental doses of either ephedrine or Metaraminol) and bradycardia (treated by Glycopyrolate). The frequency and duration of these interventions were deemed sufficiently low so as to treat the relevant periods as unaffected by them.

During the duration of the protocol, six vital signs were recorded, namely systolic, mean and diastolic blood pressure (BPsys, BPmean, BPdia), heart rate (HR), respiratory rate (RR) and the bispectral index (BIS). From these values, BPsys, BPdia and BIS were expected to be affected by Propofol administration and the signals were of adequate quality so that they could be further analysed.

5.2.2 Model fitting

For both models EM was used as described in Section 5.1.5 to learn the parameters on each patient separately and then obtain predicted values using those fitted parameters. For both models, 100 iterations of EM where used and the BFGS algorithm was run with 1000 function evaluations during the first 10 iterations and 100 evaluations during the remaining iterations. An important advantage of the models' linear dynamics is that stability constraints can be enforced which are of paramount importance in a real-world setting. If the system dynamics matrix, $\mathbf{A}$, becomes unstable (i.e. if the modulus of its largest eigenvalue is greater than one) this matrix is projected back to the space of stable matrices such that it is also closer (in a least-squares sense) to the originally learned matrix. The approach proposed by Siddiqi *et al.* (2007) is followed to achieve this, which involves solving for a quadratic program inside EM as has already been mentioned in Section 3.2.5.3. This process is very fast and is performed only if an instability is observed. Also, the noise covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ are constrained to be diagonal.

Furthermore, a number of different $d_x = \{2,3,4,5\}$ is explored for the IO-NLDS. Under these assumptions, the IO-NLDS has a variable number of parameters which is higher or equal than the PK/PD model's parameters for $d_x = \{3,4,5\}$ and lower for $d_x = 2$.[7] This variable number of parameters in the case of the IO-NLDS reflects its greater expressive power since it is not constrained by physiological assumptions. In contrast with the PK/PD model, one has the flexibility to decide on alternative latent space dimensionalities by making use of an information criterion to determine d_x under a more formal measure of optimality, turning the whole process into a standard model order selection procedure. In the PK/PD case, a higher latent space dimension would correspond to an increased number of compartments, which could require years of additional research to validate physiologically appropriate extensions of the already existing literature. However, to perform a more fair model comparison between the two models and to make an informed decision about the optimal d_x from an information theoretic perspective, Bayesian Information Criterion (BIC) scores are also computed.

Finally, the UT involves determining three parameters α, β, κ (see Appendix C for details). Following Haykin (2001, sec. 7.3), these are set to $\alpha = 0.5$, $\beta = 2$, and $\kappa = 3 - d_x$. The choices for β and κ are optimal under the assumption of Gaussianity and α, which

[7] Calculations of the number of parameters for the IO-NLDS and the PK/PD model are provided in Appendix C

determines the sperad of the sigma points around the mean, is recommended in that work to be set to a small positive value in the range of 10^{-4}–1.

5.2.3 Results

The standardised mean squared error (SMSE)[8] between predictions and actual observations is used as the evaluation metric, and also representative examples of curve fits on the observed data produced by the two models are provided. The SMSE takes into account the variance of the observed data and provides a natural baseline with SMSE $= 1$ denoting the mean prediction. For both models the SMSE is calculated between model predictions and measured outputs of BPsys, BPdia and BIS across a number of different prediction horizons; namely 1, 10 and 20-step ahead predictions, (corresponding to 15 seconds, 2.5 minutes and 5 minutes respectively). The SMSE is also calculated for the "free-running" case, which corresponds to predictions for the whole duration of the protocol. These time intervals, apart from the 1-step interval, were decided as clinically relevant. Results on the 1-step ahead prediction task are included as well, since this is a standard evaluation task when assessing predictive performance of dynamical models.

A summary of the mean SMSEs of the two models per prediction horizon, per channel and for all investigated values of d_x for the IO-NLDS is shown in Table 5.1. The IO-NLDS's prediction errors are consistently lower than the PK/PD model's and by a large margin in most cases. Both models' errors are comparably low at the 1-step ahead prediction level. Moreover, the difference in favour of the IO-NLDS becomes more obvious as the prediction horizon increases. Also, in most cases the predictive errors for both models are increased as the prediction horizon increases, as expected. Another expected behaviour is that the predictive errors for the IO-NLDS exhibit in general (with a few exceptions) a downward trend as d_x increases since more parameters are fitted. In the same table, BIC scores are also provided per prediction horizon for all models to account for IO-NLDS's generally higher number of parameters. The BIC is defined as: $BIC = -2\ln(L) + b\ln(N)$, where L is the data likelihood under the model, b is the number of free parameters[9] and N is the number of data points. The IO-NLDS achieves a lower (better) score in most

[8] $SMSE = MSE/var_y$, where MSE is the mean squared error and var_y is the variance of the observed vital sign. The SMSE is also known as the fraction of unexplained variance (FUV) and is also equal to 1 minus the coefficient of determination $(1 - R^2)$.

[9] In the case of the PK/PD model the k_{ij} $(i, j = \{1,2,3\})$ parameters are treated as the result of a separate optimisation process carried out in the relevant literature and are thus included in the calculation of b.

cases and thus demonstrates that the increase in likelihood is not just an effect of higher statistical capacity but rather that the IO-NLDS manages to model the observed temporal structure better than the PK/PD model.

The results from Table 5.1 suggest that the IO-NLDS achieves higher performance (in terms of the combination of lower SMSE and BIC) when $d_x = 4$, thus the remaining results will refer to the IO-NLDS with a latent dimension of four (IO-NLDS$^{(4)}$). Thus, in Figures 5.4–5.7, four representative examples of observed vital signs and the fitted traces produced by the two models for each of the four prediction horizons are presented. In Figure 5.4, four examples on the 1-step ahead prediction task are given. The performance of the two models is comparable and both models manage to stay quite close to the observed signals. In Figure 5.5, four examples of the 10-step ahead task are given. Again, both models seem to have captured the temporal structure of the observed signals with satisfactory accuracy although the IO-NLDS$^{(4)}$ seems to be able to stay closer to the true signal. In Figure 5.6, four examples of the 20-step ahead task are given. Here the difference in performance is clearer, with the IO-NLDS$^{(4)}$ tracking the temporal evolution of the observed signals more accurately especially during the time that the vital signs decrease more rapidly which is an important haemodynamic event (as captured by BP-sys/dia) and also corresponds to a rapid increase in the patient's depth of anaesthesia (as captured by BIS). Finally, in Figure 5.7, the "free-running" predictions of the two models are shown. The IO-NLDS$^{(4)}$ manages in general to stay close to the actual signal, while on the other hand the PK/PD model showcases a considerably higher degree of error in its predictions. Especially in the bottom right panel, it seems that the PK/PD model has failed to capture the temporal dynamics of the signal and has underestimated its decrease rate.

Overall, in all cases the IO-NLDS$^{(4)}$ manages to stay closer to the observed signal especially at the beginning of the protocol during the steepest decrease of the observed signal. This constitutes the most critical phase of the protocol during which the uncertainty of an anaesthetist about the patient's reaction to the drug is considerable and the risk of an undesirable episode (e.g. hypotension) is at its highest. Therefore, accurate predictions during that phase are much more critical compared to the rest of the signal.

Retaining the focus on the IO-NLDS$^{(4)}$, results for all three channels and the four prediction horizons are shown as boxplots in Figure 5.3, where the central mark denotes the median, the edges of the box are the lower and upper quartiles and the whiskers extend to

outliers with extreme outliers being denoted separately by a red cross. For all three channels, the IO-NLDS's predictions are consistently more accurate than the PK/PD model's across all four prediction horizons[10].

The output of a more fine-grained comparison is displayed on Figure 5.8 which reveals that in 92% of all investigated cases the accuracy of the IO-NLDS$^{(4)}$'s predictions is better than the PK/PD model's. Finally, twelve paired, right-tailed t-tests (one per channel and per prediction horizon) were performed on the differences of the obtained SMSEs $(SMSE_{PKPD} - SMSE_{IO-NLDS^{(4)}})$ across the 40 patients. The null hypothesis that the mean of this difference is zero, is rejected in 10 out of 12 cases with p-values ranging from 1.6×10^{-2} to 4.8×10^{-17}. In the case of the 1-step ahead predictions on BPdia and BIS, p-values of 0.14 and 0.37 respectively were obtained and thus the null hypothesis could not be rejected at the 95% confidence level.

The experiments which were conducted so far aim to answer the question of whether a data driven approach (IO-NLDS) is more appropriate than a knowledge-based one (PK/PD model) which is well established in the relevant field. However, the focus is now shifted on the question of whether the latent space along with its temporal dynamics which is present in the IO-NLDS is crucial to its performance or whether perhaps a more direct method that does not make use of a temporal latent space could model the vital signs equally satisfactorily.

To this end, for each vital sign, a linear regression model coupled with a generalised logistic function was used to predict y directly from $\mathbf{u}$. Thus a prediction made by this model would be of the form $\hat{y}_t = g(\hat{y}_{lr,t})$, where $\hat{y}_{lr,t} = \mathbf{w}^\top \mathbf{u}_{t-l:t}$ is the prediction made at time t by a linear regression model trained via ordinary least squares (with target outputs $\mathbf{y}$) and learned weights $\mathbf{w}$ and l is a lag value such that a sliding window of drug infusions $\mathbf{u}$ is used as input at each time step t similarly to the DSLDS's classification approach presented in Section 4.1.1. The function $g(\cdot)$ was fitted as previously via the BFGS algorithm in order to minimise the residual sum of squares $RSS(\boldsymbol{\eta}) = \Sigma_{t=1}^{T}(y_t - g(\hat{y}_{lr,t}))^2$. This regression approach (henceforth called LR_g) was tried for different values of $l = \{6, 11, 17, 24\}$ so that the overall sum of the parameters per fitted channel would follow very closely the number of parameters learned by the IO-NLDS for the different values of d_x. Since the LR_g does not make use of any past values of $\mathbf{y}$ for its predictions, the comparison is made with respect to the free-running results for the IO-NLDS.

[10] With the exception of the 1-step predictions for BIS where the two models are effectively tied.

The results from this comparison, which are presented in Table 5.2, show that the IO-NLDS clearly achieves a higher performance in all investigated cases, suggesting that for these two fully data-driven approaches, the presence of a latent space and the use of a model which captures explicitly the temporal continuity of the investigated signals seem to result in a higher expressive power.

Table 5.1: Comparison of IO-NLDS and PK/PD model with respect to mean SMSE per channel and prediction horizon and mean BIC per prediction horizon. Lower numbers are better. Numbers inside parenthesis denote the latent dimensionality of the IO-NLDS.

SMSE/BIC	1-step				10-step				20-step				free-running			
	BPs	BPd	BIS	BIC	BPs	BPd	BIS	BIC	BPs	BPd	BIS	BIC	BPs	BPd	BIS	BIC
IO-NLDS$^{(2)}$	0.14	0.23	0.13	2534	0.24	0.32	0.22	4932	0.29	0.36	0.28	7286	0.45	0.52	0.40	19970
IO-NLDS$^{(3)}$	0.13	0.19	0.06	2494	0.22	0.29	0.20	5049	0.25	0.31	0.27	7909	0.31	0.38	0.33	21047
IO-NLDS$^{(4)}$	0.13	0.20	0.08	2592	0.19	0.27	0.21	4458	0.20	0.27	0.25	6656	0.25	0.36	0.32	16588
IO-NLDS$^{(5)}$	0.12	0.24	0.10	2786	0.19	0.31	0.18	4821	0.24	0.30	0.25	7385	0.23	0.33	0.23	19959
PK/PD	0.30	0.23	0.10	2709	0.52	0.53	0.41	5314	0.73	0.75	0.74	7919	0.84	0.82	0.89	19008

Table 5.2: Comparison of IO-NLDS and LR$_g$ model with respect to mean SMSE per channel for the free-running case. Lower numbers are better. Numbers inside parenthesis denote the latent dimensionality in the case of the IO-NLDS and the lag length of the input vectors in the case of the LR$_g$ model.

SMSE	BPs	BPd	BIS
LR$_g^{(6)}$	0.75	0.80	0.83
IO-NLDS$^{(2)}$	0.45	0.52	0.40
LR$_g^{(11)}$	0.69	0.73	0.76
IO-NLDS$^{(3)}$	0.31	0.38	0.32
LR$_g^{(17)}$	0.64	0.68	0.68
IO-NLDS$^{(4)}$	0.30	0.38	0.33
LR$_g^{(24)}$	0.60	0.63	0.58
IO-NLDS$^{(5)}$	0.18	0.31	0.34

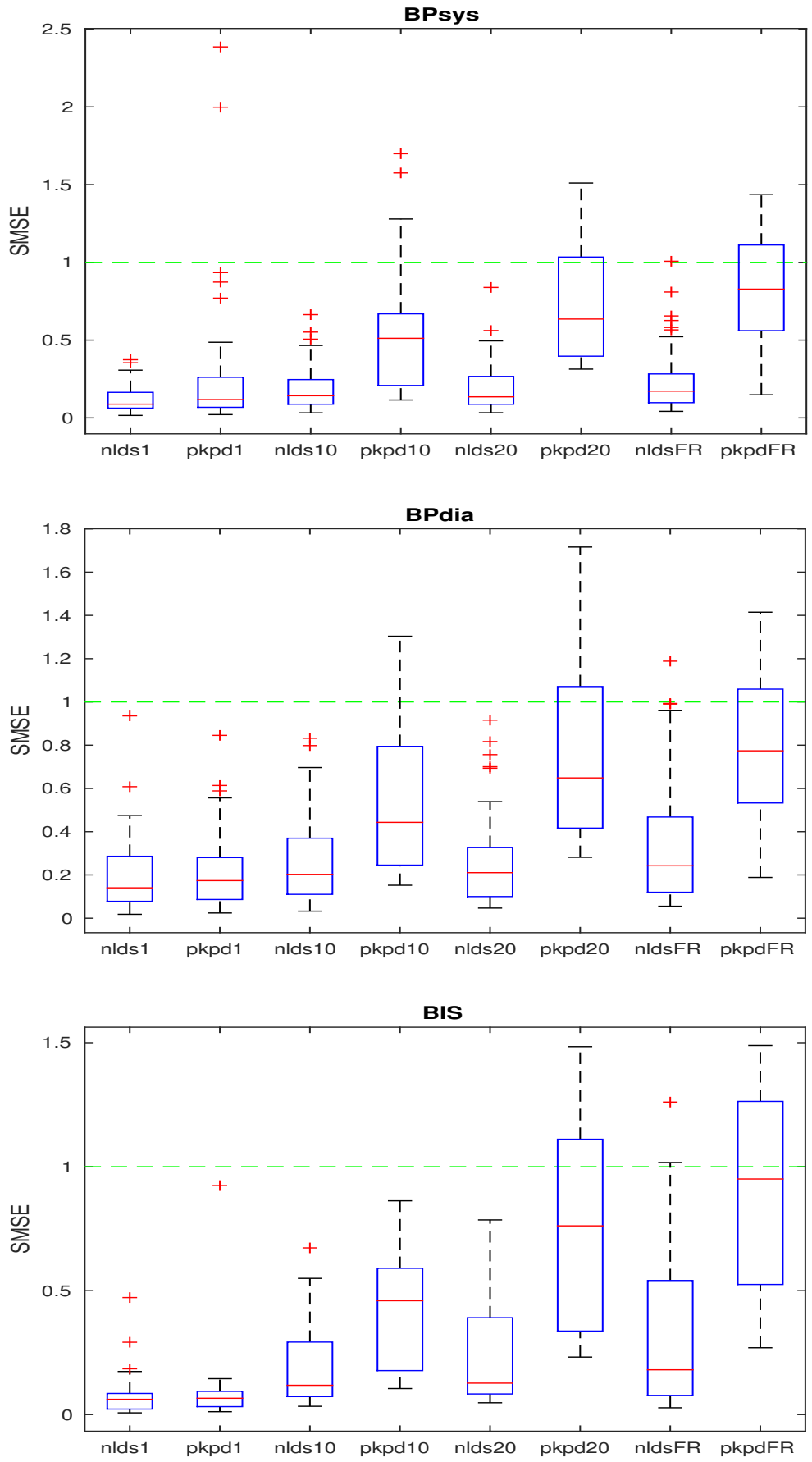

Figure 5.3: Summarised SMSE results for IO-NLDS$^{(4)}$ compared to PK/PD for BPsys, BPdia and BIS across the four prediction horizons for all patients. The green dashed line corresponds to the mean predictions. The x-axis is labelled according to the model name followed by the prediction horizon, where FR stands for free-running.

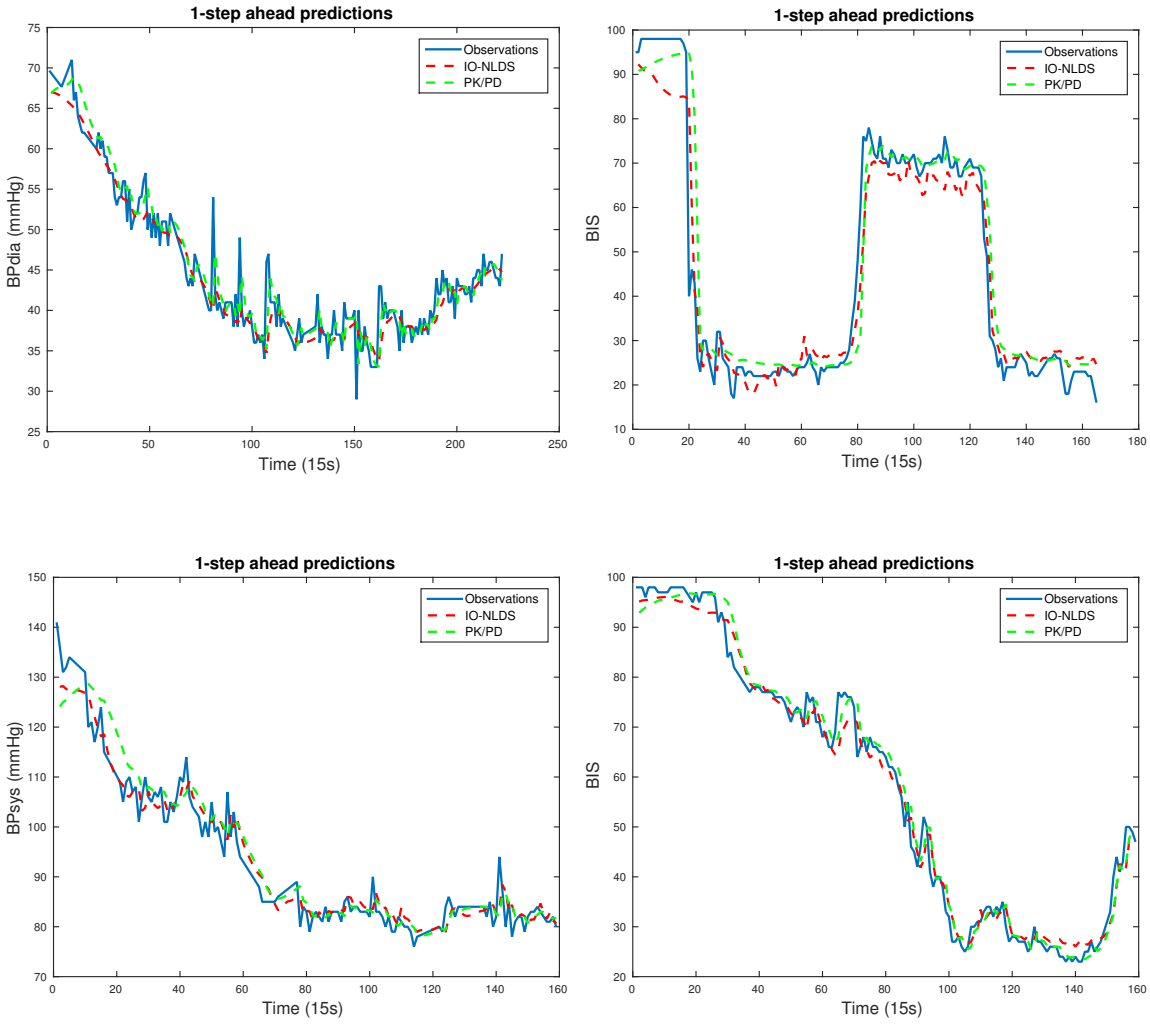

Figure 5.4: Examples of observed vital sign traces (blue solid line) and the predictions made by the IO-NLDS$^{(4)}$ (red dashed line) and the PK/PD model (green dashed line) for the 1-step ahead prediction task.

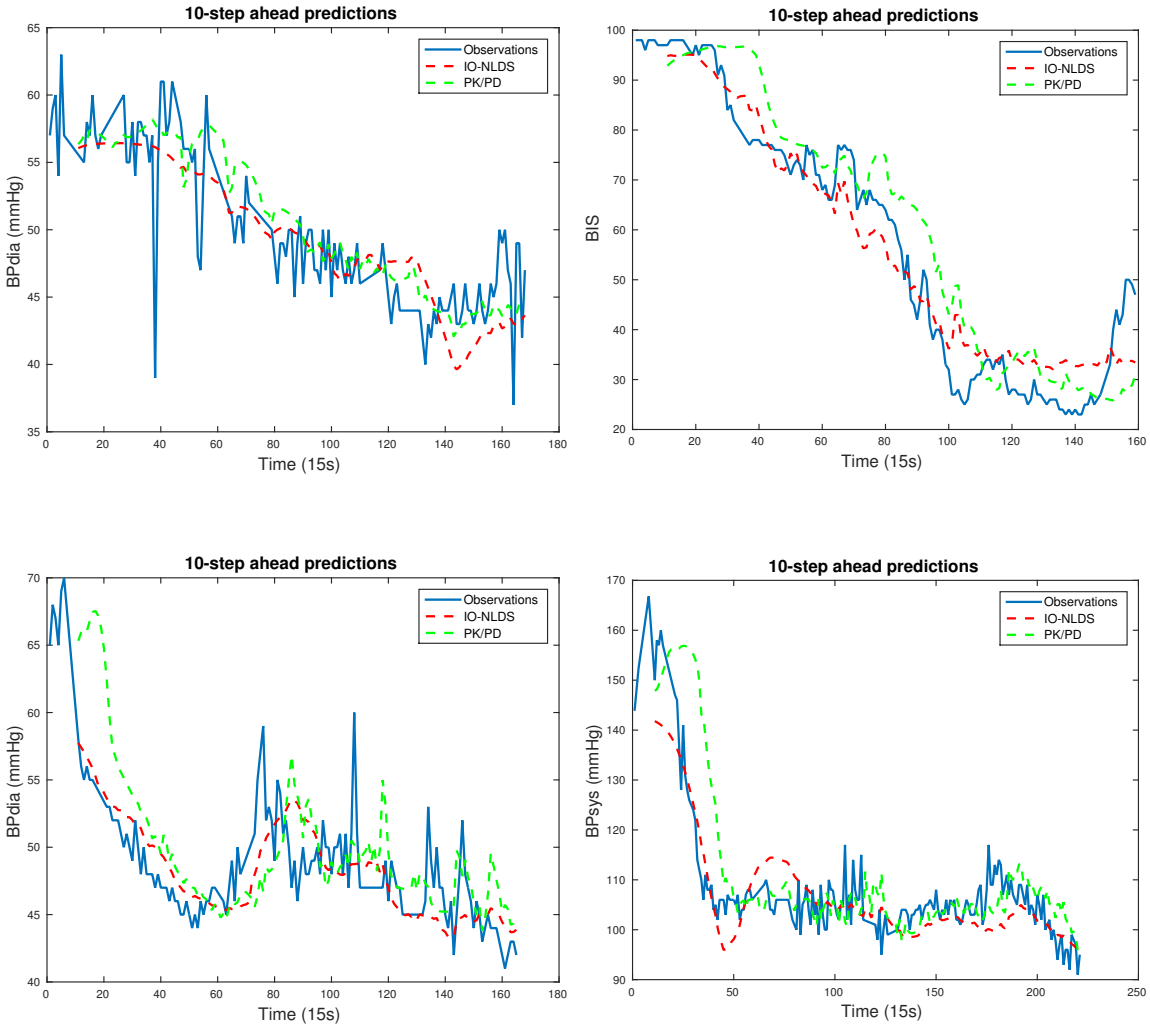

Figure 5.5: Examples of observed vital sign traces (blue solid line) and the predictions made by the IO-NLDS$^{(4)}$ (red dashed line) and the PK/PD model (green dashed line) for the 10-step ahead prediction task.

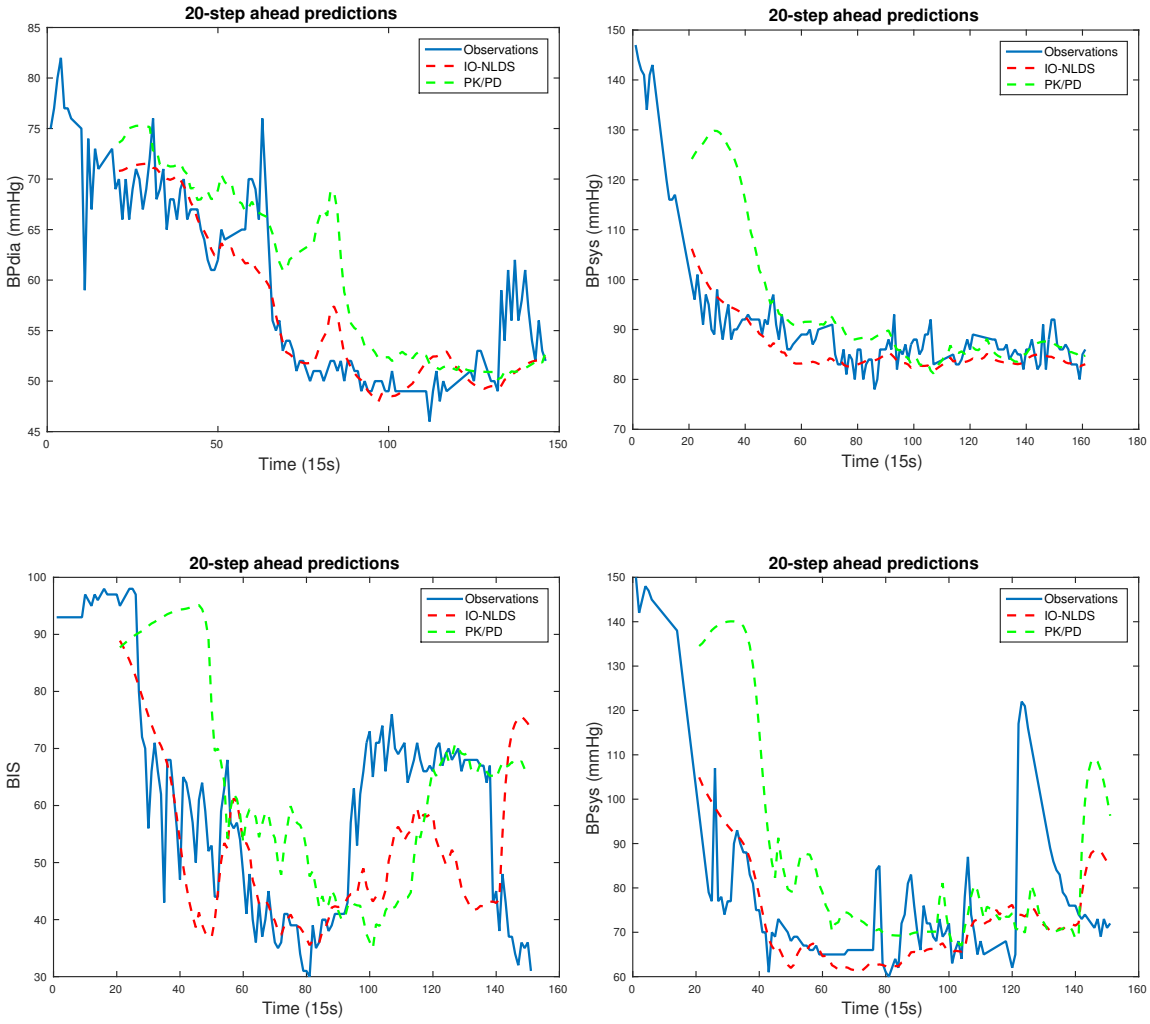

Figure 5.6: Examples of observed vital sign traces (blue solid line) and the predictions made by the IO-NLDS$^{(4)}$ (red dashed line) and the PK/PD model (green dashed line) for the 20-step ahead prediction task.

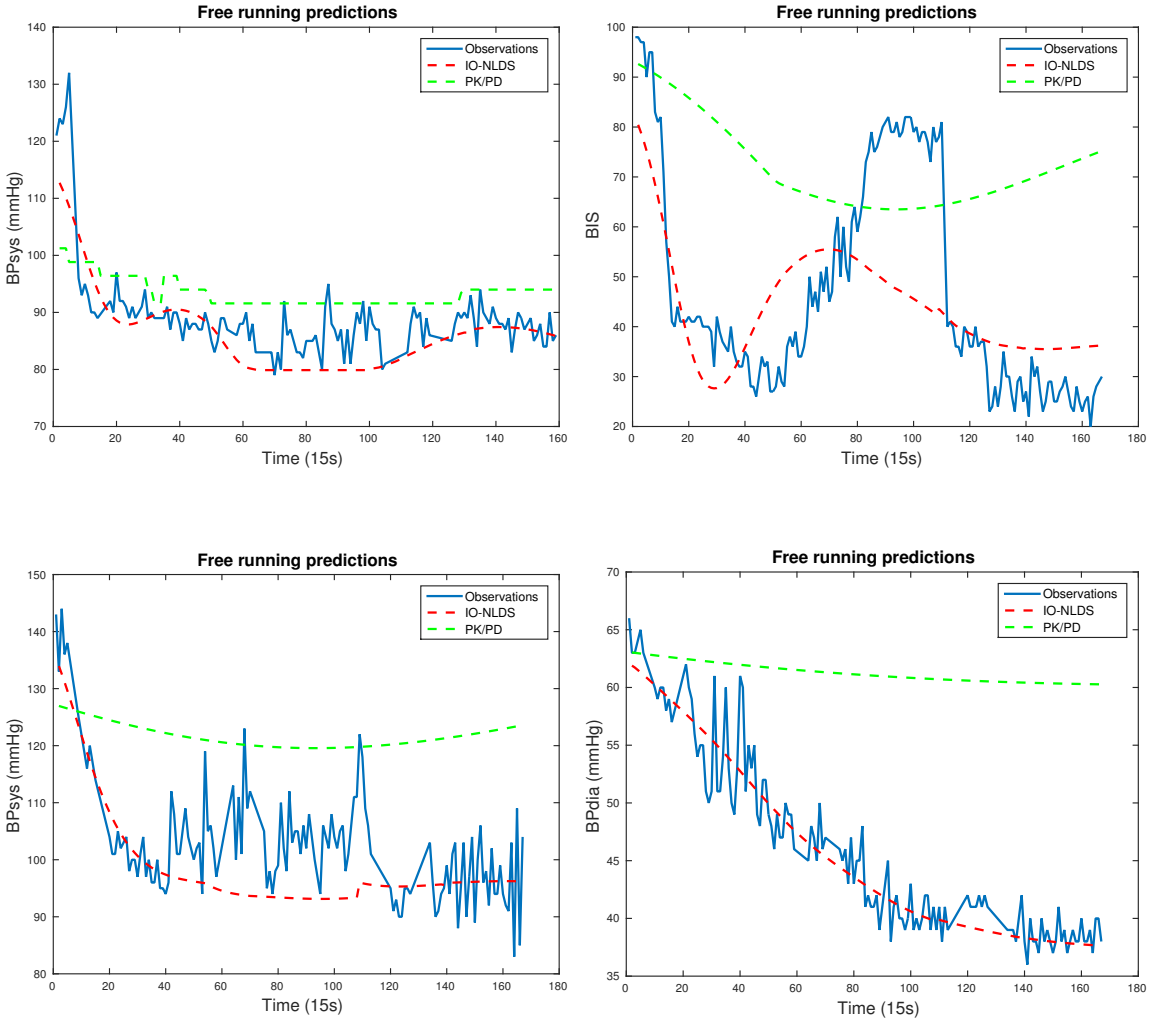

Figure 5.7: Examples of observed vital sign traces (blue solid line) and the predictions made by the IO-NLDS$^{(4)}$ (red dashed line) and the PK/PD model (green dashed line) for the free-running case.

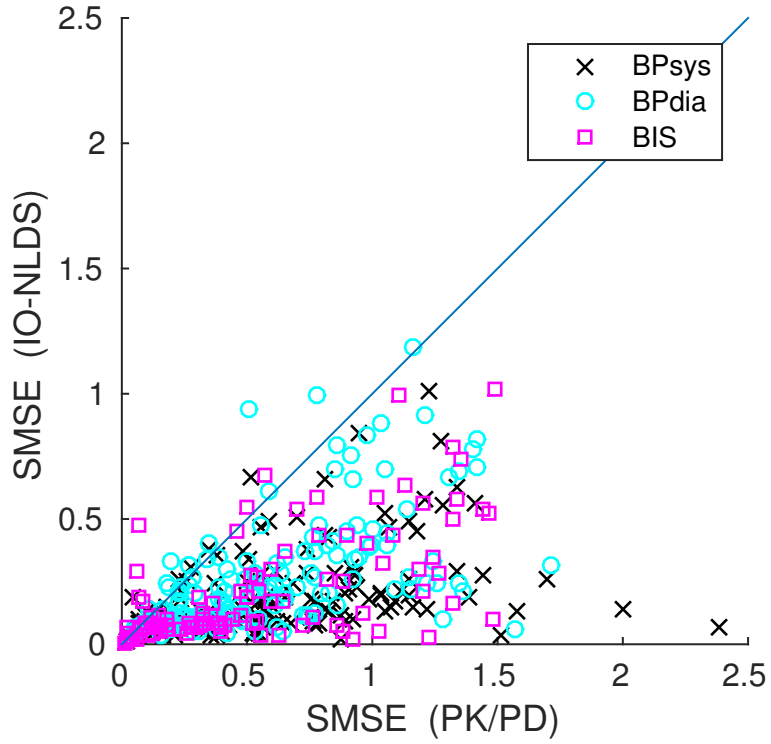

Figure 5.8: SMSE results for all patients across all four prediction horizons and all three measured channels. Points below the 45 degree line correspond to lower errors for the IO-NLDS$^{(4)}$ on the same task and points above that line correspond to lower errors for the PK/PD model. The diagonal line represents equal performance between the two models.

5.3 Summary

In this chapter, a purely data-driven approach for the important application of modelling drug effects on the physiology of patients in ICUs has been presented and it was shown that the new data-driven approach outperforms the previous expert-knowledge based one. To the best of the author's knowledge, this is the first time that a complex, real-world model has been fully learned via a combination of unscented filters and EM, making this a promising direction for related tasks. The increased performance of the IO-NLDS does not come as a surprise, since the IO-NLDS has (genrally) more parameters that are learned based on the present patient cohort. However, this flexibility is also the key idea behind the conception of the IO-NLDS: that a more personalised and data-driven approach should be a more suitable approach for the explored setting than a method like the PK/PD model which is based on a model that has been mostly fitted on a different patient cohort with the hope that this would generalise across all future patients.

Also, a use case scenario with a real-world application is now presented: The current practice in an ICU is that an anaesthetist will start drug administration on a patient with an initial target dosage and then readjust it according to the evolution of the patient's vital signs. This implies that the anaesthetist needs to be at bedside continuously and for the whole duration of the drug administration, and by definition can only be reactive to potential vital sign deteriorations. Also, this monitoring is not always straightforward. For example, during drug induction when the risk of hypotension is high, the anaesthetist's attention is divided as they will be also managing the airway (e.g. by inserting an endotracheal tube). Therefore a use case for the system is to take a clinician-specified drug infusion protocol, and make predictions for the time-course of the vital signs. The clinician could then inspect these forecasts, and either continue with the original protocol or adjust it accordingly. (In that case the model could also automatically notify the clinician, e.g. if a threshold is crossed by the predicted values). This extra functionality would help the anaesthetist to be proactive and to avert undesirable episodes for the patient (e.g. an episode of hypotension). The PK/PD literature has become very focussed on compartmental models, but more data-driven models may give rise to better, more personalised predictions which could thus translate to better clinical outcomes.

In the presented experimental setting, a separate NLDS was fitted per patient. However, the model would be more powerful if its parameters, $\boldsymbol{\theta}$ could be predicted from patient

covariates $\mathbf{z}$ (e.g. age, gender, weight etc.) and this could be the target of future work. In that case, a functional relationship could be learned in a supervised manner between $\mathbf{z}$ and $\boldsymbol{\theta}$ from existing training data and personalised model parameters $\boldsymbol{\theta}_{new}$ could be assigned to each new patient during test time such that $\boldsymbol{\theta}_{new} = f(\mathbf{z}_{new})$, where $f(\cdot)$ can be any appropriate regression algorithm. Finally, as the model is drug-agnostic, it can be applied to other drugs as well in a straightforward manner.

Chapter 6

Conclusion

This thesis has explored the development of dynamical probabilistic graphical models and their application to the critical task of physiological condition monitoring of patients admitted in ICUs. We conclude this document by summarising the research contributions made so far and providing suggestions for extending the presented material further.

6.1 Contributions

The main contributions of this thesis are as follows:

- The formulation, development and validation of the Discriminative Switching Linear Dynamical System (DSLDS). The DSLDS adopts a discriminative approach to the task of physiological condition monitoring and was shown to compare favourably to a prior generative approach (FSLDS). Furthermore, an α-mixture of the two approaches was developed and was shown to achieve higher performance compared to both the DSLDS and the FSLDS.

- The formulation, development and validation of the Input-Output Non-Linear Dynamical System (IO-NLDS). The IO-NLDS was developed with the goal of explicitly incorporating drug effects into the modelling process; something that the DSLDS (or the FSLDS) do not take into account. The IO-NLDS was shown to outperform prior work which is based on domain-specific, expert knowledge and provides a "plug-and-play", purely data-driven approach.

- As a secondary, minor contribution, prior work (FSLDS) which was successfully applied to the case of neonatal ICUs, was extended to the case of adult ICUs.

6.2 Future work

The nature of research is such that no project could ever be deemed fully completed. Here, we attempt to identify a number of approaches that could be fruitful avenues of exploration for researchers that might wish to carry on the work presented so far. We start by making suggestions which are specific to the DSLDS, then the IO-NLDS and then to a potential fusion of the two.

DSLDS

The DSLDS makes the simplifying assumption that the joint distribution of the latent factors is fully factorised as in the case of the FSLDS. It could be the case that more complex interactions are present between different factors in which case a different graphical structure could be more appropriate. Such a structure could be potentially motivated by the available domain knowledge. Furthermore, the DSLDS in its current formulation models temporal continuity across the factors in an implicit manner, via the construction of features derived from the observations. An explicit link could also be added between the factors at different times and investigate if that impacts performance in a significant manner.

IO-NLDS

The current learning method for the IO-NLDS is carried out per patient and thus does not allow for the modelling of a new patient. This can be achieved by establishing a functional relationship between the learned parameters of the IO-NLDS per patient and their demographic characteristics (e.g. age, gender etc.). In that case a new patient's demographic information can be used to construct an appropriate IO-NLDS. Perhaps a more principled approach would be to have a hierarchical IO-NLDS. One could form a prior over the IO-NLDS's parameters (potentially conditioned on demographics) and then derive posterior beliefs via the hierarchical structure, so that statistical strength can be shared between (similar) patients. However that would further complicate inference and it is not guaranteed that it would yield better results compared to the previous approach. Finally,

it should be mentioned that the IO-NLDS and the PK/PD represent the two ends on the data-driven vs. expert-knowledge spectrum. A combination of the two is possible as well. One could initialise the IO-NLDS via the PK/PD parameterisation and allow its parameters to be learned (e.g. via EM) while retaining the structural zeros imposed by the initial PK/PD parameterisation. It would be then very interesting to examine whether the learned model would potentially retain the interpretability of the PK/PD model while exhibiting a similarly high performance as in the case of the fully data-driven IO-NLDS.

IO-(D)S(N)LDS

A more complete monitoring system would have the capacity to combine the DSLDS's and IO-NLDS's capabilities. That is, it would be able to identify artifactual and physiological processes of interest, model the effect of drug infusions on the observed physiology and maintain beliefs about the underlying physiology of a patient. This can be achieved via a fusion of the two models that would lead to an Input-Output Discriminative Switching (Non-)Linear Dynamical System (IO-DS(N)LDS). The graphical model for the proposed IO-DS(N)LDS is shown in Figure 6.1 and the model is defined by the following joint distribution:

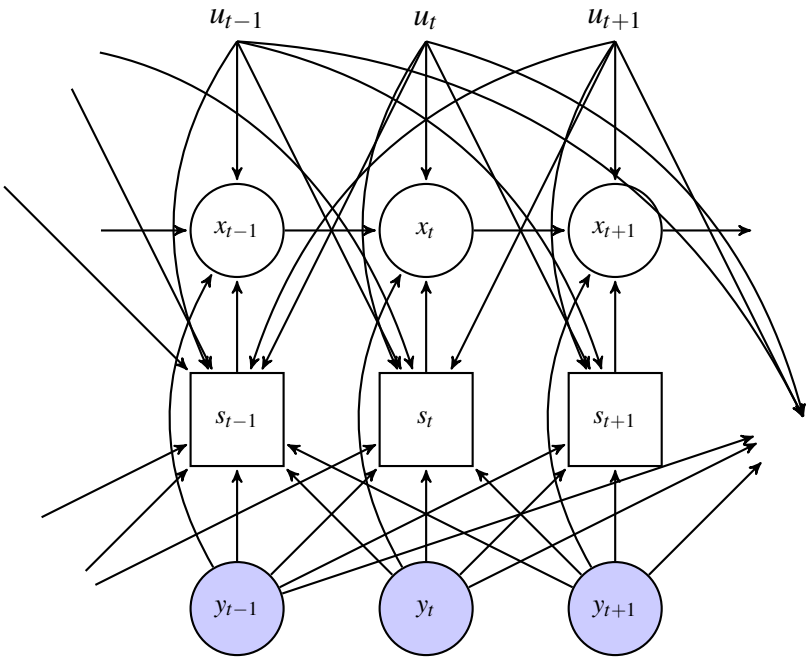

Figure 6.1: Graphical model of the IO-DS(N)LDS. Comapred to the DSLDS, past and future values of $\mathbf{u}$ can now affect inference about s. Inference about $\mathbf{x}_t$ now additionally depends on $\mathbf{u}_t$.

$$p(\mathbf{s},\mathbf{x}|\mathbf{y},\mathbf{u}) = p(s_1|\mathbf{y}_{1:1+r},\mathbf{u}_{1:1+r})p(\mathbf{x}_1|s_1,\mathbf{y}_1,\mathbf{u}_1) \prod_{t=2}^{T} p(s_t|\mathbf{y}_{t-l:t+r},\mathbf{u}_{t-l:t+r})p(\mathbf{x}_t|\mathbf{x}_{t-1},s_t,\mathbf{y}_t,\mathbf{u}_t) \ . \tag{6.1}$$

Such a model could be seen as an extension to the DSLDS so that drug effects on a patient's physiology can be captured. Inference about the state-of-health of a patient can now incorporate drug information; for example, the model could learn that a hypotension episode would be more likely under a higher dosage of an anaesthetic drug. Similarly, the underlying continuous physiology could be more accurately modelled so that if an artifact obscures the observed physiology while a drug is being infused, the $\mathbf{x}$-chain could be updated solely under the influence of the drug, ignoring the incorporation of artifactual observations. For example, if a damped trace obscures BP readings while an anaesthetic is being infused, the model could still alert clinicians of a potential upcoming hypotension if such an event is deemed likely by the learned dynamics, even if the (artifactual) readings might be suggesting otherwise.

Finally, a generative approach has still its own merits, especially with respect to modelling previously unseen abnormalities via the X-factor. It is worth mentioning that the IO-NLDS can be combined with the (generative) SLDS as well, giving rise to the IO-S(N)LDS. The graphical model of this generative approach can be seen in Figure 6.2 and is defined by the following joint distribution:

$$p(\mathbf{s},\mathbf{x},\mathbf{y}|\mathbf{u}) = p(s_1|\mathbf{u}_1)p(\mathbf{x}_1|s_1,\mathbf{u}_1)p(\mathbf{y}_1|\mathbf{x}_1,s_1,\mathbf{u}_1)\times \prod_{t=2}^{T} p(s_t|s_{t-1},\mathbf{u}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1},s_t,\mathbf{u}_t)p(\mathbf{y}_t|\mathbf{x}_t,s_t,\mathbf{u}_t) \ . \tag{6.2}$$

It should be noted though that as data become more readily available, the discriminative approach should be able to outperform the generative approach with increasing gains and thus could be a more fruitful avenue of research. Finally, it should be mentioned that the models developed within this thesis aim at operating on a real-time basis. As such, all methods are formulated in such a way so as to admit on-line inference procedures. Thus, advancing from the research prototyping stage into deploying them on-site in an ICU should not pose a significant challenge.

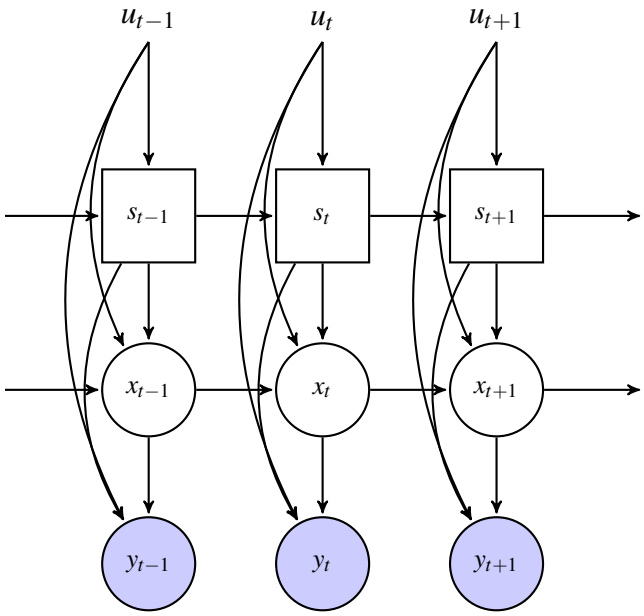

Figure 6.2: Graphical model of the IO-SLDS. Compared to the SLDS, the drug infusion at time t, $\mathbf{u}_t$, affects now the state of health, s_t, the underlying physiology, $\mathbf{x}_t$, and the observed values, $\mathbf{y}_t$.

Appendix A

EM algorithm for the LDS

Here, details are provided about the EM algorithm used for learning in the case of the LDS. The parameters to be estimated are $\boldsymbol{\theta} = \{\boldsymbol{\pi}_1, \mathbf{V}_1, \mathbf{A}, \mathbf{Q}, \mathbf{C}, \mathbf{R}\}$, where $\boldsymbol{\pi}_1, \mathbf{V}_1$ are the mean and covariance of the initial state and the rest of the parameters are as described in Section 3.2.5.

A.1 EM

The EM algorithm is a two-step process which is guaranteed to maximise the likelihood by maximising a surrogate function Q:

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

which is called the *expected complete data log-likelihood*. During the E-step, moments of the latent posterior density $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{old})$ are calculated as shown below and during the M-step, new estimates $\hat{\boldsymbol{\theta}} \leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$, are produced.

E-step

The E step requires access to the smoothed posterior densities. Smoothing is an off-line inference algorithm that updates the filtered posterior densities in a recursive manner

after the entirety of a temporal signal of total length T has been observed. The smoothed posteriors are then given by the Rauch–Tung–Striebel (RTS) smoothing algorithm (Rauch *et al.*, 1965) as follows:

$$p(\mathbf{x}_t|\mathbf{y}_{1:T}) \sim \mathcal{N}(\boldsymbol{\mu}_{t|T}, \mathbf{V}_{t|T}) \,, \tag{A.2}$$

$$\boldsymbol{\mu}_{t|T} = \boldsymbol{\mu}_{t|t} + \mathbf{J}_t(\boldsymbol{\mu}_{t+1|T} - \boldsymbol{\mu}_{t+1|t}) \,, \tag{A.3}$$

$$\mathbf{V}_{t|T} = \mathbf{V}_{t|t} + \mathbf{J}_t(\mathbf{V}_{t+1|T} - \mathbf{V}_{t+1|t})\mathbf{J}_t^\top \,, \tag{A.4}$$

$$\mathbf{J}_t = \mathbf{V}_{t|t}\mathbf{A}_{t+1}^\top \mathbf{V}_{t+1|t}^{-1} \,, \tag{A.5}$$

where $\mathbf{J}_t$ is called the backwards Kalman gain matrix. It is noted that this smoothing process does not need access to the observations $\mathbf{y}$ but only to the filtered means and covariances calculated via the Kalman filter. We also require the following two quantities:

$$\mathbf{P}_t = \mathbf{V}_{t|T} + \boldsymbol{\mu}_{t|T}\boldsymbol{\mu}_{t|T}^\top \,, \tag{A.6}$$

$$\mathbf{P}_{t,t-1} = \mathbf{V}_{t,t-1|T} + \boldsymbol{\mu}_{t|T}\boldsymbol{\mu}_{t-1|T}^\top \,, \tag{A.7}$$

where $\mathbf{P}_{t,t-1}$ can be obtained through the backward recursion:

$$\mathbf{V}_{t-1,t-2|T} = \mathbf{V}_{t-1|t-1}\mathbf{J}_{t-2}^\top + \mathbf{J}_{t-1}(\mathbf{V}_{t,t-1|T} - \mathbf{A}\mathbf{V}_{t-1|t-1})\mathbf{J}_{t-2}^\top \,, \tag{A.8}$$

which is initialised by $\mathbf{V}_{T,T-1|T} = (\mathbf{I} - \mathbf{K}_T\mathbf{C})\mathbf{A}\mathbf{V}_{T-1|T-1}$.

M-step

Having computed the required quantites and defining $\boldsymbol{\mu}_{t|T} = \hat{\boldsymbol{\mu}}_t$ for brevity, we can continue with the M step. For each parameter of the LDS, the M-step involves taking the corresponding partial derivative of eq. (A.1), setting them to zero and solving them. The results are listed below:

$$\hat{\mathbf{A}} = \left(\sum_{t=2}^{T} \mathbf{P}_{t,t-1} \right) \left(\sum_{t=2}^{T} \mathbf{P}_{t-1} \right)^{-1} , \tag{A.9}$$

$$\hat{\mathbf{Q}} = \frac{1}{T-1} \left(\sum_{t=2}^{T} \mathbf{P}_t - \hat{\mathbf{A}} \sum_{t=2}^{T} \mathbf{P}_{t-1,t} \right) , \tag{A.10}$$

$$\hat{\mathbf{C}} = \left(\sum_{t=1}^{T} \mathbf{y}_t \hat{\boldsymbol{\mu}}_t \right) \left(\sum_{t=1}^{T} \mathbf{P}_t \right)^{-1} , \tag{A.11}$$

$$\hat{\mathbf{R}} = \frac{1}{T} \sum_{t=1}^{T} (\mathbf{y}_t \mathbf{y}_t^\top - \hat{\mathbf{C}} \hat{\boldsymbol{\mu}}_t \mathbf{y}_t^\top) , \tag{A.12}$$

$$\hat{\mathbf{V}}_1 = \mathbf{P}_1 - \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1^\top , \tag{A.13}$$

$$\hat{\boldsymbol{\pi}}_1 = \hat{\boldsymbol{\mu}}_1 . \tag{A.14}$$

More details are provided in Ghahramani and Hinton (1996).

Appendix B

Gaussian Sum approximation for the DSLDS

Here, we derive the updates of the approximate inference scheme used in the DSLDS. The exposition follows closely Murphy (1998), with appropriate modifications.

We define the following quantities:

$$\begin{aligned}
\mathbf{x}_{t|\tau}^{ij} &= E[\mathbf{x}_t|\mathbf{y}_{1:\tau}, s_{t-1}=i, s_t=j] \\
\mathbf{x}_{t|\tau}^{j} &= E[\mathbf{x}_t|\mathbf{y}_{1:\tau}, s_t=j] \\
\mathbf{V}_{t|\tau}^{j} &= Cov[\mathbf{x}_t|\mathbf{y}_{1:\tau}, s_t=j] \\
\mathbf{V}_{t,t-1|\tau}^{j} &= E[\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{y}_{1:\tau}, s_t=j] \\
w_{t|\tau}^{j} &= p(s_t=j|\mathbf{y}_{1:\tau})
\end{aligned}$$

Filtering then amounts to performing the following steps:

$$[\mathbf{x}_{t|t}^{ij}, \mathbf{V}_{t|t}^{ij}] = filter(\mathbf{x}_{t-1|t-1}^{i}, \mathbf{V}_{t-1|t-1}^{i}, \mathbf{y}_t; \mathbf{A}^j, \mathbf{C}^j, \mathbf{Q}^j, \mathbf{R}^j) \tag{B.1}$$

$$w^i = p(s_{t-1}=i|\mathbf{y}_{1:t-1}) \tag{B.2}$$

$$[\mathbf{x}_{t|t}^{j}, \mathbf{V}_{t|t}^{j}] = collapse(\mathbf{x}_{t|t}^{ij}, \mathbf{V}_{t|t}^{j}, w^i) \tag{B.3}$$

The $filter$ and $collapse$ operators are defined below.

B.1 Filter

We first compute the predicted mean and covariance under state j.

$$\mathbf{x}^j_{t|t-1} = \mathbf{A}^j \mathbf{x}^i_{t-1|t-1}$$
$$\mathbf{V}^j_{t|t-1} = \mathbf{A}^j \mathbf{V}^i_{t-1|t-1} (\mathbf{A}^j)^\top + \mathbf{Q}^j$$

We then compute the error in our prediction (known also as innovation), the covariance of the error term and the Kalman gain.

$$\mathbf{e}^j_t = \mathbf{C}^j \mathbf{x}^j_{t|t-1} - \mathbf{y}_t$$
$$\mathbf{S}^j_t = \mathbf{C}^j \mathbf{V}^j_{t|t-1} (\mathbf{C}^j)^\top + \mathbf{R}^j$$
$$\mathbf{K}^j_t = \mathbf{V}^j_{t|t-1} (\mathbf{C}^j)^\top (\mathbf{S}^j_t)^{-1}$$

Finally, we compute the updated mean and covariance that correspond to the transition from state i to state j.

$$\mathbf{x}^{ij}_{t|t} = \mathbf{x}^j_{t|t-1} + \mathbf{K}^j_t \mathbf{e}^j_t$$
$$\mathbf{V}^{ij}_{t|t} = \mathbf{V}^j_{t|t-1} - \mathbf{K}^j_t \mathbf{S}^j_t (\mathbf{K}^j_t)^\top$$

B.2 Collapse

We first compute the collapsed mean for state j.

$$\mathbf{x}^j_{t|t} = \sum_i w^i \mathbf{x}^{ij}_{t|t}$$

We then compute the collapsed covariance for state j.

$$\begin{aligned}
\mathbf{V}^j_{t|t} &= \sum_i w^i Cov[\mathbf{x}_t | s_{t-1} = i] \\
&= \sum_i w^i E[(\mathbf{x}_t - \mathbf{x}^j_{t|t})(\mathbf{x}_t - \mathbf{x}^j_{t|t})^\top | s_{t-1} = i] \\
&= \sum_i w^i E[(\mathbf{x}_t - \mathbf{x}^{ij}_{t|t} + \mathbf{x}^{ij}_{t|t} - \mathbf{x}^j_{t|t})(\mathbf{x}_t - \mathbf{x}^{ij}_{t|t} + \mathbf{x}^{ij}_{t|t} - \mathbf{x}^j_{t|t})^\top | s_{t-1} = i] \\
&= \sum_i w^i E[(\mathbf{x}_t - \mathbf{x}^{ij}_{t|t})(\mathbf{x}_t - \mathbf{x}^{ij}_{t|t})^\top | s_{t-1} = i] + \sum_i w^i (\mathbf{x}^{ij}_{t|t} - \mathbf{x}^j_{t|t})(\mathbf{x}^{ij}_{t|t} - \mathbf{x}^j_{t|t})^\top \\
&= \sum_i w^i \mathbf{V}^{ij}_{t|t} + \sum_i w^i (\mathbf{x}^{ij}_{t|t} - \mathbf{x}^j_{t|t})(\mathbf{x}^{ij}_{t|t} - \mathbf{x}^j_{t|t})^\top
\end{aligned}$$

Appendix C

Inference and learning details for the IO-NLDS

Here, details are provided about the unscented transform (UT) and the EM algorithm used for inference and learning in the case of the IO-NLDS.

C.1 Unscented transform

Calculations of integrals of the form $\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] = \int g(\mathbf{x})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma)d\mathbf{x}$ are involved in the UKF and EM. Since $g(\cdot)$ is non-linear, these integrals are approximated via Gaussian cubature (i.e. multidimensional Gaussian quadrature) as: $\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] \approx \sum_i w_i g(\mathbf{x}_i)$. Following similar notation to Murphy (2012, sec. 18.5.2.1), we first define a set of $2d+1$ sigma points $\mathbf{x}_i$:

$$\mathbf{x} = \{\boldsymbol{\mu}, \{\boldsymbol{\mu} + (\sqrt{(d+\lambda)\boldsymbol{\Sigma}})_i\}, \{\boldsymbol{\mu} - (\sqrt{(d+\lambda)\boldsymbol{\Sigma}})_i\}\} \ ,$$

where $\boldsymbol{\Sigma}_i$ denotes the i'th column of matrix $\boldsymbol{\Sigma}$ and $i = 1,...,d$. This set of sigma points is propagated via $g(\cdot)$, producing a new set of transformed points $\mathbf{y}_i$ with mean and covari-

ance:

$$\boldsymbol{\mu}_y = \sum_{i=0}^{2d} w_m^i \mathbf{y}_i \ ,$$
$$\boldsymbol{\Sigma}_y = \sum_{i=0}^{2d} w_c^i (\mathbf{y}_i - \boldsymbol{\mu}_y)(\mathbf{y}_i - \boldsymbol{\mu}_y)^\top \ ,$$

with weights w's defined as:

$$w_m^0 = \frac{d}{d+\lambda} \ ,$$
$$w_c^0 = \frac{d}{d+\lambda} + (1 - \alpha^2 + \beta) \ ,$$
$$w_m^i, w_c^i = \frac{1}{2(d+\lambda)} \ ,$$

where $\lambda = \alpha^2(d+\kappa) - d$ and α, β, and κ are method-specific parameters.

C.2 EM

We provide ML estimates for the parameters involved in eq. (5.7) for the i'th iteration of EM that can be derived in closed form. First, the following quantities are defined:

$$S_{x-x-} = \sum_{t=2}^{T} \mathbf{V}_{t-1|T} + \boldsymbol{\mu}_{t-1|T} \boldsymbol{\mu}_{t-1|T}^\top \ ,$$
$$S_{xx-} = \sum_{t=2}^{T} \mathbf{V}_{t,t-1|T} + \boldsymbol{\mu}_{t|T} \boldsymbol{\mu}_{t-1|T}^\top \ ,$$
$$S_{xx} = \sum_{t=2}^{T} \mathbf{V}_{t|T} + \boldsymbol{\mu}_{t|T} \boldsymbol{\mu}_{t|T}^\top \ ,$$
$$S_{x-u} = \sum_{t=2}^{T} \boldsymbol{\mu}_{t-|T} \boldsymbol{u}_t^\top \ ,$$
$$S_{xu} = \sum_{t=2}^{T} \boldsymbol{\mu}_{t|T} \boldsymbol{u}_t^\top \ ,$$
$$S_{uu} = \sum_{t=2}^{T} \boldsymbol{u}_t \boldsymbol{u}_t^\top \ ,$$

where $\boldsymbol{\mu}_{t|T}$, $\mathbf{V}_{t|T}$ denote the smoothed mean and covariance estimates at time t and $\mathbf{V}_{t,t-1|T}$ denotes the smoothed pairwise covariance estimates at times $t-1,t$, as outputted by the URTS smoother. What follows is:

$$\begin{aligned}
\boldsymbol{\mu}_1^i &= \boldsymbol{\mu}_{1|T} \ , \\
\boldsymbol{\Sigma}_1^i &= \mathbf{V}_{1|T} \ , \\
[\mathbf{A}^i \ \mathbf{B}^i] &= [S_{xx-} \ S_{xu}] \begin{bmatrix} S_{x-x-} & S_{x-u} \\ S_{x-u}^\top & S_{uu} \end{bmatrix}^{-1} \ , \\
\mathbf{Q}^i &= \frac{1}{T-1}(S_{xx} - \mathbf{A}^i S_{xx-}^\top - \mathbf{B}^i S_{xu}^\top) \ .
\end{aligned}$$

Finally:

$$\mathbf{R}^i = \frac{1}{T}\sum_{t=2}^{T} \mathbb{E}_{p(\mathbf{x}_t|\mathbf{y}_{1:T},\mathbf{u}_{1:T},\boldsymbol{\theta}^{i-1})}[(\mathbf{y}_t - g(\mathbf{x}_t,\boldsymbol{\theta}_{NL}))(\mathbf{y}_t - g(\mathbf{x}_t,\boldsymbol{\theta}_{NL}))^\top] \ ,$$

where the expectation is another Gaussian integral that can be computed via the unscented approximation as described in section C.1.

C.3 Parameter counts

Here, we provide calculations for counting the number of parameters involved during the learning process of the IO-NLDS and the PK/PD model. The parameters to be fitted are: $\boldsymbol{\theta} = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \boldsymbol{\eta}\}$. In the case of the IO-NLDS, the number of parameters is:

$$b_{IONLDS} = d_x + \frac{d_x(d_x+1)}{2} + d_x^2 + d_x + d_y(d_x + d_\eta) + d_x + d_y \ , \tag{C.1}$$

where d_x and d_y are the latent space and observations space dimensionalities respectively and d_η is the number of parameters associated with the generalised logistic function as defined in eq. (5.1) and is always equal to four. Each term in the above formula reflects the contribution to the parameter counting made by each respective element in $\boldsymbol{\theta}$.

The PK/PD model follows a more constrained parameterisation. All elements of $\boldsymbol{\mu}_1$ are set to 0 to reflect the fact that the initial drug concentration in all compartments equals zero. Matrix $\mathbf{B}$ is a vector with one element set to 1 and the rest set to 0 to reflect the drug quantity which is infused solely into the central compartment. Matrix $\mathbf{C}$ is similarly constrained to have 1/0 entries so as to link the appropriate effect site concentration with the respective observed signal. Therefore, these parameters do not contribute to parameter count for the PK/PD model. The contributions in the parameter count stem from the fitted elements in $\boldsymbol{\Sigma}_1, \mathbf{A}, \boldsymbol{\eta}, \mathbf{Q}, \mathbf{R}$. Therefore, in the case of the PK/PD model, the number of parameters is:

$$b_{PKPD} = \frac{d_x(d_x+1)}{2} + (d_{PK} + d_y) + d_x + d_y d_\eta + d_y \ , \tag{C.2}$$

where d_{PK} refers to the parameters of $\mathbf{A}$ associated with the pharmacokinetic aspect of the model ($k_{12}, k_{21}, k_{13}, k_{31}, k_{10}$) and $d_x = 6$ in the case of the PK/PD model.

Bibliography

Aleks, N., Russell, S. J., Madden, M. G., Morabito, D., Staudenmayer, K., Cohen, M., and Manley, G. T. (2009). Probabilistic detection of short events, with application to critical care monitoring. In *Advances in Neural Information Processing Systems 21*, pages 49–56.

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

Amari, S.-i. (2007). Integration of Stochastic Models by Minimizing α-Divergence. *Neural Computation*, **19**(10), 2780–2796.

Andersen, K. E. and Højbjerre, M. (2003). A Bayesian Approach to Bergman's Minimal Model. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pages 236–243.

Aström, K. J. and Murray, R. M. (2010). *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press.

Bailey, J. M. and Haddad, W. M. (2005). Drug Dosing Control in Clinical Pharmacology. *Control Systems, IEEE*, **25**(2), 35–51.

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

Bengio, Y. and Frasconi, P. (1995). An Input Output HMM Architecture. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 427–434. MIT Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Boots, B. (2009). *Learning Stable Linear Dynamical Systems*. Master's thesis, Carnegie Mellon University.

Box, G. and Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**(1), 5–32.

Brigham, E. O. (1988). The Fast Fourier Transform and its Applications.

Brockwell, P. J. and Davis, R. A. (2009). *Time Series: Theory and Methods*. Springer.

Caruana, R. and Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 161–168. ACM.

Chambrin, M.-C., Ravaux, P., Calvelo-Aros, D., Jaborska, A., Chopin, C., and Boniface, B. (1999). Multicentric Study of Monitoring Alarms in the Adult Intensive Care Unit (ICU): A Descriptive Analysis. *Intensive care medicine*, **25**(12), 1360–1366.

Chatfield, C. (2003). *The Analysis of Time Series: an Introduction*. Chapman and Hall, CRC Press.

Cheng, S. and Sabes, P. N. (2006). Modeling Sensorimotor Learning with Linear Dynamical systems. *Neural Computation*, **18**(4), 760–793.

Clifford, G. D. (2002). *Signal Processing Methods for Heart Rate Variability*. Ph.D. thesis, Department of Engineering Science, University of Oxford.

Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J., and Tarassenko, L. (2013). Gaussian Processes for Personalized e-Health Monitoring With Wearable Sensors. *IEEE Transactions on Biomedical Engineering*, **60**(1), 193–197.

Colburn, W. A. (1981). Simultaneous Pharmacokinetic and Pharmacodynamic Modeling. *Journal of Pharmacokinetics and Biopharmaceutics*, **9**(3), 367–388.

Cooper, T. R., Berseth, C. L., Adams, J. M., and Weisman, L. E. (1998). Actuarial Survival in the Premature Infant Less Than 30 Weeks' Gestation. *Pediatrics*, **101**(6), 975–978.

De Freitas, N. (2002). Rao-Blackwellised Particle Filtering for Fault Diagnosis. In *Aerospace Conference Proceedings, 2002, IEEE*, volume 4, pages 4–1767. IEEE.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Diggle, P. (1990). *Time Series: a Biostatistical Introduction*. Oxford University Press.

Donald, R. (2014). *Predicting Hypotensive Episodes in the Traumatic Brain Injury Domain*. Ph.D. thesis, University of Glasgow.

Donnet, S. and Samson, A. (2013). A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. *Advanced Drug Delivery Reviews*, **65**(7), 929–939.

Doucet, A. and Johansen, A. M. (2009). A Tutorial on Particle Filtering and Smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, **12**(656-704), 3.

Dräger Medical (2007). SC7000 patient monitor datasheet.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Dürichen, R., Pimentel, M. A., Clifton, L., Schweikard, A., and Clifton, D. A. (2015). Multitask Gaussian Processes for Multivariate Physiological Time-Series Analysis. *IEEE Transactions on Biomedical Engineering*, **62**(1), 314–322.

Enright, C. G., Madden, M. G., Madden, N., and Laffey, J. G. (2011). Clinical Time Series Data Analysis Using Mathematical Models and DBNs. In *Artificial Intelligence in Medicine*, pages 159–168. Springer.

Enright, C. G., Madden, M. G., and Madden, N. (2013). Bayesian networks for mathematical models: Techniques for automatic construction and efficient inference. *International Journal of Approximate Reasoning*, **54**(2), 323–342.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research (JMLR)*, **15**(1), 3133–3181.

Fletcher, R. (2013). *Practical Methods of Optimization*. John Wiley & Sons.

Fuseau, E. and Sheiner, L. B. (1984). Simultaneous Modeling of Pharmacokinetics and Pharmacodynamics With a Nonparametric Pharmacodynamic Model. *Clinical Pharmacology & Therapeutics*, **35**(6), 733–741.

Ganong, W. F. and Barrett, K. E. (1995). *Review of Medical Physiology*. Appleton & Lange Norwalk, CT.

Gašperin, M. and Juričíc, D. (2011). Application of Unscented Transformation in Nonlinear System Identification. In *Proceedings of the 18th IFAC World Congress*, volume 18, pages 4428–4433.

Georgatzis, K. and Williams, C. K. I. (2015). Discriminative Switching Linear Dynamical Systems applied to Physiological Condition Monitoring. In *Proceedings of the Thirty-first Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 306–315.

Georgatzis, K., Lal, P., Hawthorne, C., Shaw, M., Piper, I., Tarbert, C., Donald, R., and Williams, C. K. I. (2016a). Artefact in Physiological Data Collected from Patients with Brain Injury: Quantifying the Problem and Providing a Solution Using a Factorial Switching Linear Dynamical Systems Approach. *Intracranial Pressure and Brain Monitoring XV, Acta Neurochirurgica Journal Supplement*, **122**(1), 301–305.

Georgatzis, K., Williams, C. K. I., and Hawthorne, C. (2016b). Input-Output Non-Linear Dynamical Systems applied to Physiological Condition Monitoring. *Proceedings of Machine Learning in Healthcare, JMLR W&C Track, Volume 56*.

Gepts, E., Camu, F., Cockshott, I., and Douglas, E. (1987). Disposition of Propofol Administered as Constant Rate Intravenous Infusions in Humans. *Anesthesia & Analgesia*, **66**(12), 1256–1263.

Ghahramani, Z. and Hinton, G. E. (1996). Parameter Estimation for Linear Dynamical Systems. Technical report, Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science.

Ghahramani, Z. and Hinton, G. E. (2000). Variational Learning for Switching State-Space Models. *Neural computation*, **12**(4), 831–864.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial Hidden Markov Models. *Machine learning*, **29**(2-3), 245–273.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, **101**(23), e215–e220.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *Radar and Signal Processing, IEE Proceedings of*, volume 140, pages 107–113. IET.

Goutelle, S., Maurin, M., Rougier, F., Barbaut, X., Bourguignon, L., Ducher, M., and Maire, P. (2008). The Hill equation: a review of its capabilities in pharmacological modelling. *Fundamental & Clinical Pharmacology*, **22**(6), 633–648.

Grewal, M. and Andrews, A. (2001). *Kalman Filtering: Theory and Practice Using MATLAB*. John Wiley, New York.

Hagenblad, A., Ljung, L., and Wills, A. (2008). Maximum Likelihood Identification of Wiener Models. *Automatica*, **44**(11), 2697–2705.

Harrison, J. and West, M. (1999). *Bayesian Forecasting & Dynamic Models*. Springer.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Haykin, S. S. (2001). *Kalman Filtering and Neural Networks*. Wiley Online Library.

Heckerman, D. (1998). A Tutorial on Learning With Bayesian Networks. Technical report.

ixellence GmbH (2015). ixTrend. `https://www.ixellence.com/index.php/en/products/ixtrend`.

Johnson, A. E., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., and Clifford, G. D. (2016). Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE*, **104**(2), 444–466.

Jones, P. A., Andrews, P. J., Midgley, S., Anderson, S. I., Piper, I. R., Tocher, J. L., Housley, A. M., Corrie, J., Slattery, J., and Dearden, N. (1994). Measuring the Burden of Secondary Insults in Head-Injured Patients During Intensive Care. *Journal of Neurosurgical Anesthesiology*, **6**(1), 4–14.

Julier, S. J. and Uhlmann, J. K. (1996). A General Method for Approximating Nonlinear Transformations of Probability Distributions. Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, **82**(1), 35–45.

Kokkala, J., Solin, A., and Särkkä, S. (2014). Expectation Maximization Based Parameter Estimation by Sigma-Point and Particle Smoothing. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1–8. IEEE.

Kokkala, J., Solin, A., and Särkkä, S. (2015). Sigma-Point Filtering and Smoothing Based Parameter Estimation in Nonlinear Dynamic Systems. *arXiv:stat.ME/1504.06173v2*.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *International Conference on Machine Learning (ICML)*.

Lal, P., Williams, C. K. I., Georgatzis, K., Hawthorne, C., McMonagle, P., Piper, I., and Shaw, M. (2015). Detecting Artifactual Events in Vital Signs Monitoring Data. Technical report, University of Edinburgh.

Lehman, L.-w. H., Adams, R. P., Mayaud, L., Moody, G. B., Malhotra, A., Mark, R. G., and Nemati, S. (2015). A Physiological Time Series Dynamics-Based Approach to Patient Monitoring and Outcome Prediction. *Biomedical and Health Informatics, IEEE Journal of*, **19**(3), 1068–1076.

Lerner, U. and Parr, R. (2001). Inference in Hybrid Networks: Theoretical Limits and Practical Algorithms. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 310–318. Morgan Kaufmann Publishers Inc.

Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzell, R. (2015). Learning to Diagnose with LSTM Recurrent Neural Networks. *arXiv preprint arXiv:1511.03677*.

Lu, W.-L., Murphy, K. P., Little, J. J., Sheffer, A., and Fu, H. (2009). A Hybrid Conditional Random Field for Estimating the Underlying Ground Surface from Airborne LiDAR Data. *Geoscience and Remote Sensing, IEEE Transactions on*, **47**(8), 2913–2922.

Maas, A. I., Stocchetti, N., and Bullock, R. (2008). Moderate and Severe Traumatic Brain Injury in Adults. *The Lancet Neurology*, **7**(8), 728–741.

Marlin, B. M. (2008). *Missing Data Problems in Machine Learning*. Ph.D. thesis, University of Toronto.

Marsh, B., White, M., Morton, N., and Kenny, G. (1991). Pharmacokinetic Model Driven Infusion of Propofol in Children. *British Journal of Anaesthesia*, **67**(1), 41–48.

McCallum, A., Freitag, D., and Pereira, F. C. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *International Conference on Machine Learning (ICML)*, pages 591–598.

Murphy, K. P. (1998). Switching Kalman Filters. Technical report, University of California, Berkeley.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Narendra, K. and Gallman, P. (1966). An Iterative Method for the Identification of Nonlinear Systems Using a Hammerstein Model. *IEEE Transactions on Automatic control*, **11**(3), 546–550.

Nemati, S., Lehman, L.-W., and Adams, R. P. (2013). Learning Outcome-Discriminative Dynamics in Multivariate Physiological Cohort Time Series. In *Engineering in Medicine and Biology Society (EMBS), 2013 35th Annual International Conference of the IEEE*, pages 7104–7107. IEEE.

Øksendal, B. (2013). *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media.

Peck, C. C., Beal, S. L., Sheiner, L. B., and Nichols, A. I. (1984). Extended Least Squares Nonlinear Regression: A Possible Solution to the Choice of Weights Problem in Analysis of Individual Pharmacokinetic Data. *Journal of Pharmacokinetics and Biopharmaceutics*, **12**(5), 545–558.

Penrose, R. (1955). A Generalized Inverse for Matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 51, pages 406–413. Cambridge University Press.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1996). *Numerical recipes in C*, volume 2. Cambridge University Press.

Pronovost, P. J., Angus, D. C., Dorman, T., Robinson, K. A., Dremsizov, T. T., and Young, T. L. (2002). Physician Staffing Patterns and Clinical Outcomes in Critically Ill Patients: a Systematic Review. *Jama*, **288**(17), 2151–2162.

Quinn, J. (2007). *Bayesian Condition Monitoring in Neonatal Intensive Care*. Ph.D. thesis, University of Edinburgh.

Quinn, J. A. and Williams, C. K. I. (2007). Known Unknowns: Novelty Detection in Condition Monitoring. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 1–6. Springer.

Quinn, J. A., Williams, C. K. I., and McIntosh, N. (2009). Factorial Switching Linear Dynamical Systems applied to Physiological Condition Monitoring. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **31**(9), 1537–1551.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A Unifying View of Sparse Approximate Gaussian Process Regression. *Journal of Machine Learning Research*, **6**(Dec), 1939–1959.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, **77**(2), 257–286.

Ramanathan, M. (1999). An Application of Ito's Lemma in Population Pharmacokinetics and Pharmacodynamics. *Pharmaceutical Research*, **16**(4), 584–586.

Rampil, I. J. (1998). A Primer for EEG Signal Processing in Anesthesia. *The Journal of the American Society of Anesthesiologists*, **89**(4), 980–1002.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rauch, H. E., Striebel, C., and Tung, F. (1965). Maximum Likelihood Estimates of Linear Dynamic Systems. *AIAA journal*, **3**(8), 1445–1450.

Richards, F. (1959). A Flexible Growth Function for Empirical Use. *Journal of Experimental Botany*, **10**(2), 290–301.

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

Saria, S., Koller, D., and Penn, A. (2010). Discovering Shared and Individual Latent Structure in Multiple Time Series. *arXiv preprint arXiv:1008.2028*.

Särkkä, S. (2008). Unscented Rauch–Tung–Striebel Smoother. *Automatic Control, IEEE Transactions on*, **53**(3), 845–849.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.

Särkkä, S. and Solin, A. (2014). Applied Stochastic Differential Equations. Lecture notes of the course Becs-114.4202 Special Course in Computational Engineering II, Aalto University.

Schetzen, M. (1980). The Volterra and Wiener Theories of Nonlinear Systems.

Schnider, T., Minto, C., Gambus, P., Andresen, C., Goodale, D., Shafer, S., and Youngs, E. (1998). The Influence of Method of Administration and Covariates on the Pharmacokinetics of Propofol in Adult Volunteers. *Anesthesiology*, **88**(5), 1170–1182.

Schön, T. B., Wills, A., and Ninness, B. (2011). System Identification of Nonlinear State-Space Models. *Automatica*, **47**(1), 39–49.

Sendelbach, S. and Funk, M. (2013). Alarm Fatigue: A Patient Safety Concern. *AACN advanced critical care*, **24**(4), 378–386.

Shaw, M. (2013). A concise description of the clinical waveform pre-processing workflow. Unpublished manuscript.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications*. Springer New York.

Siddiqi, S. M., Boots, B., and Gordon, G. J. (2007). A Constraint Generation Approach to Learning Stable Linear Dynamical Systems. In *Advances in Neural Information Processing Systems 20*, pages 1329–1336.

Smith, G. and Robinson, A. (2000). A Comparison Between the EM and Subspace Identification Algorithms for Time-Invariant Linear Dynamical Systems. Technical report.

Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E., and Featherstone, P. I. (2013). The Ability of the National Early Warning Score (NEWS) to Discriminate Patients at Risk of Early Cardiac Arrest, Unanticipated Intensive Care Unit Admission, and Death. *Resuscitation*, **84**(4), 465–470.

Stanculescu, I., Williams, C. K. I., and Freer, Y. (2014). A Hierarchical Switching Linear Dynamical System Applied to the Detection of Sepsis in Neonatal Condition Monitoring. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 752–761.

Steiner, L. and Andrews, P. (2006). Monitoring the Injured Brain: ICP and CBF. *British Journal of Anaesthesia*, **97**(1), 26–38.

Tarbet, C. (2012). Internal Report. Department of Clinical Physics, NHS Greater Glasgow and Clyde, Unpublished manuscript.

Van Overschee, P. and De Moor, B. (1996). *Subspace Identification for Linear Systems: Theory, Implementation, Applications*, volume 1. Kluwer Academic Publishers.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**(1), 91.

Wan, E. A. and Van Der Merwe, R. (2000). The Unscented Kalman Filter for Nonlinear Estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, pages 153–158. Ieee.

White, M., Kenny, G. N., and Schraag, S. (2008). Use of Target Controlled Infusion to Derive Age and Gender Covariates for Propofol Clearance. *Clinical Pharmacokinetics*, **47**(2), 119–127.

Williams, C. K. I. and Stanculescu, I. (2011). Automating the Calibration of a Neonatal Condition Monitoring System. In *Artificial Intelligence in Medicine*, pages 240–249. Springer.

Wills, A., Schön, T. B., Ljung, L., and Ninness, B. (2011). Blind Identification of Wiener Models. *IFAC Proceedings Volumes*, **44**(1), 5597–5602.

Wills, A., Schön, T. B., Ljung, L., and Ninness, B. (2013). Identification of Hammerstein-Wiener Models. *Automatica*, **49**(1), 70–81.