

Auto-Calibration for Neonatal Condition Monitoring

Ioan-Anton Stanculescu



Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh

2010

Abstract

This project describes our approach for automating the calibration stage needed by the Factorial Switching Linear Dynamical System (FSLDS) for neonatal condition monitoring.

Using observations collected by the monitoring equipment, the FSLDS proposed by Prof Christopher Williams and Dr John Quinn [22] is highly successful in inferring the probability distributions of both physiological and artifactual hidden factors affecting the measurements. The detection of these factors presents vital importance to clinicians. However, the FSLDS must be calibrated using manually selected data segments corresponding to normal dynamics. Our goal is to automatically find such intervals.

The proposed solution uses a binary classifier able to discriminate normal sections from a set of fixed length measurement intervals. A “channel-based” classification is justified, together with a new labeling for the data, feature extraction solutions and a performance criterion especially developed for the task. Experimental results have demonstrated that our classifiers are able to correctly extract normality sections from the observations on hand. More importantly, a number of comparative tests between the auto- and manually-calibrated FSLDSs provide evidence that our approach is indeed successful in uncovering the hidden factors influencing the observational data.

Acknowledgements

I would like to thank my supervisor, Prof Christopher Williams, for the excellent theoretical insight and continuous support throughout the development of this project. Dr John Quinn has provided priceless technical information about the particularities of the condition monitoring application. My sincere gratitude to all the wonderful people I have met on both sides of Crichton Street during the past year. Finally, I would like to thank my family for being immensely supportive in the last twenty four years.

Declaration

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

(Ioan-Anton Stanculescu)

Table of Contents

1	Introduction	1
1.1	Outline of the dissertation	2
2	Physiological Monitoring Data	4
2.1	The measurement channels	5
2.2	Normal variation	6
2.3	Physiological and Artifactual events	7
3	Background	11
3.1	Previous work on probabilistic dynamical models	11
3.2	The Factorial Switching Linear Dynamical System	16
3.2.1	The structure of the FSLDS	16
3.2.2	Training the FSLDS. The calibration stage	18
3.2.3	Inference in the FSLDS	20
4	Building Classifiers	22
4.1	Exploratory Data Analysis	22
4.2	Finding an appropriate problem formulation	24
4.2.1	Chopping the data	24
4.2.2	Factors and Channels	25
4.2.3	Interval-Based vs. Channel-Based Approach	26
4.3	Feature extraction	27
4.4	Classifier setup	28
5	Evaluation	32
5.1	Evaluating the classifiers	32
5.1.1	The baseline feature set	33
5.1.2	Exploring other features	39
5.2	Evaluating the inferences produced by the auto-calibrated FSLDS	44

6	Conclusions and Future Work	56
6.1	Conclusions	56
6.2	Future Work	57
	Bibliography	59

List of Figures

2.1	Intervals of annotated normality on channels monitoring the cardiovascular system for two different babies (left column - Baby 10 and right column - Baby 13). Heart rate measurements clearly follow dissimilar patterns, while blood pressure measurements vary around baseline signals located at distinct levels.	6
2.2	Two instances of bradycardia: in the upper plot (Baby 1) starting around $t = 625$ and in the lower plot (Baby 7) starting around $t = 175$. Notice 2 other possible bradycardia events in the upper plot starting around $t = 125$ and $t = 875$. The annotator considered these are not significant enough to be (binary) classified as bradycardia.	7
2.3	Two instances of temperature probe disconnection (Baby 4 - upper plot and Baby 14 - lower plot). Incubator temperature measurements are shown in dashed lines. Core temperature measurements below the incubator temperature are likely to be caused by placing the core probe near the incubator doors.	8
2.4	Two instances of blood sampling events (left column - Baby 5 and right column - Baby 10). Note that the events have not only different length, but also the relationship between the median value of the channel and the extreme values on the ramps is hard to define. In both cases, heart rate readings cease during the procedure.	9
2.5	Two examples of opening the incubator doors (left column - Baby 13 and right column - Baby 8). In the left panel, we notice that the incubator was opened for taking a blood sample. In the second case, an unspecific pattern appears on the heart rate measurement when the incubator's doors are unclosed.	9
3.1	DAGs of various probabilistic models: SSM (upper-left corner), switching dynamics (upper-right corner), switching observations (lower-left corner) and Switching SSM (lower-right corner). Shaded variables are observed; circles represent continuous variables and squares represent discrete ones.	13

3.2	The DAG of a FSLDS with 2 factors (one physiological and one artifactual). The state is divided into dimensions that approximate the "true" physiology and dimensions that approximate the artifactual patterns. Shaded variables are observed; circles represent continuous variables and squares represent discrete ones.	17
5.1	Classifying heart rate (HR) intervals - ROC curves for three classifiers.	34
5.2	Classifying systolic blood pressure (BS) intervals - ROC curves for three classifiers.	35
5.3	Classifying core temperature (TC) intervals - ROC curves for three classifiers.	36
5.4	Classifying HR - ROC curves for three feature sets applied to ML logistic regression.	40
5.5	Classifying BD - ROC curves for three feature sets applied to ML logistic regression.	41
5.6	ROC curves showing the classification of the four factors of interest for three methods of calibration	45
5.7	Experiment 1: MANUAL calibration - Inferred distributions for Blood Sample and Bradycardia. Inference on both factors is correct. However, an inferred bradycardia instance around time 125 is in disagreement with the annotator's opinion.	50
5.8	Experiment 1: AUTO calibration - Inferred distributions for Blood Sample and Bradycardia. Inference on both factors is correct. However, an inferred bradycardia instance around time 125 is in disagreement with the annotator's opinion.	51
5.9	Experiment 2: MANUAL calibration - Inferred distributions for Incubator Open (i.e. Handling), Blood Sample and Bradycardia. The predictions are generally correct. Nevertheless, there are two exceptions: two separate handling instances are not discriminated and bradycardia is still predicted long after the event has ended.	52
5.10	Experiment 2: AUTO calibration - Inferred distributions for Incubator Open (i.e. Handling), Blood Sample and Bradycardia. The predictions are generally correct. Nevertheless, there are two exceptions: two separate handling instances are not discriminated and bradycardia is still predicted long after the event has ended.	53
5.11	Experiment 3: MANUAL calibration - Inferred distributions for Blood Sample and the X-factor (i.e. the factor for the Abnormal Class). Both factors are correctly inferred. The X-factor displays high sensitivity to any deviation from normality.	54

5.12 Experiment 3: AUTO calibration - Inferred distributions for Blood Sample and the X-factor (i.e. the factor for the Abnormal Class). Both factors are correctly inferred. In addition, the X-factor displays high sensitivity to any deviation from normality. 55

List of Tables

2.1	A list of known physiological and artifactual events and the channels they can influence	10
3.1	Overwriting ordering for 7 factors on 7 channels of interest for this project. A factor placed higher on the list overwrites a factor a with lower position.	19
4.1	Summary statistics of the incidence of various factors computed for the 360 hours of monitoring data on hand	23
4.2	The channels of interest and the factors that can influence them	25
4.3	The baseline set of features extracted from each channel.	27
4.4	Summary of data availability for the channels of interest. The right column contains the list of babies for whom the channel on the left column is unavailable.	29
5.1	Classifying heart rate (HR) intervals - Performance comparison of three classifiers in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.	34
5.2	Classifying systolic blood pressure (BS) intervals - Performance comparison of three classifiers in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets. The “-” mark signals a missing channel for the current baby (i.e. no classification to be made).	35
5.3	Classifying core temperature (TC) intervals - Performance comparison of three classifiers in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.	36
5.4	AUC comparison between the three classifiers for all seven channels of interest.	37

5.5	Classifying heart rate (HR) intervals - Performance comparison between logistic regression and Bayesian logistic regression in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.	37
5.6	Classifying diastolic blood pressure (BD) intervals - Performance comparison between logistic regression and Bayesian logistic regression in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets. The “-” mark signals a missing channel for the current baby (i.e. no classification to be made).	38
5.7	Classifying HR - Performance comparison of three sets of features applied to ML logistic regression in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.	40
5.8	Classifying BD - Performance comparison of three sets of features applied to ML logistic regression in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets. The “-” mark signals a missing channel (i.e. no classification to be made).	41
5.9	Classifying HR - Enhancing the feature set with gestation and post-natal age; comparison of ML logistic regression performance in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.	42
5.10	Classifying Incu. Air Humidity - Enhancing the feature set with gestation and post-natal age; comparison of ML logistic regression performance in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.	43
5.11	Summary statistics for the three methods of calibration	46
5.12	Summary statistics for manual- and auto-calibration when jitter has been added	47
5.13	Summary statistics comparing two auto-calibration methods: the standard “Auto” picking only the top prediction and “Auto-3” picking the top three predictions.	48
5.14	Experiment 1: MANUAL vs. AUTO calibration - Summary of the inferred distributions for Blood Sample and Bradycardia. The calibration methods perform equally well. However, they both flag a possible bradycardia instance around time 125, this being in disagreement with the annotator’s opinion.	49

5.15 Experiment 2: MANUAL vs. AUTO calibration - Summary of the inferred distributions for Incubator Open (i.e. Handling), Blood Sample and Bradycardia. The particularity of this experiment is factor overlapping, but the calibration methods deliver satisfactory performance. However, they both consider two separate handling annotations as a single interval.	49
5.16 Experiment 3: MANUAL vs. AUTO calibration - Summary of the inferred distributions for Blood Sample and the X-factor (i.e. the factor for the Abnormal Class). The quality of the predictions for Blood Sample is identical, but the manual version performs better for the X-factor.	49

Chapter 1

Introduction

It is often the case that people have to swiftly make critical decisions in order to respond to the behavior of complex systems. The situations when the exact state of the system cannot be directly observed are clearly the most challenging. One immediate action to be taken in such a scenario is to choose a set of relevant measurement channels for the analysed system and monitor their evolution through time.

In face of uncertainty, common-sense tells us to rely on models. In general, models tend to simplify the structure of a real-world system by making a set of assumptions. Sometimes, it can be useful to assume that the system switches between a certain number of modes of operation. *Condition monitoring* is the task of inferring which regime is being followed by using measurements taken from the system.

The present project is dedicated to performing condition monitoring on such a system, namely premature born babies receiving intensive care. Born 12-14 weeks before the full term gestation of 40 weeks, they have an extremely fragile physical condition. Their state of health cannot be directly observed, but a set of physiological measurements is collected on a second-by-second basis by cotside devices. A presentation of the vital aspects of a baby's physiology monitored in a neonatal unit will be given in Chapter 2. In broad terms, clinicians use their expert knowledge together with these recordings in order to determine the best course of action for each patient. Nevertheless, this is far from being a trivial task. Fortunately, the model introduced in the following comes into support.

Proposed by Dr John Quinn and Prof Christopher Williams [27, 22, 24], The Factorial Switching Linear Dynamical System (FSLDS) has the ability to model a system which switches between multiple modes of operation conditioned on a set of factors. The main focus of their work was on the inference problem. More precisely, given a sequence of observations, the FSLDS would output the filtering distribution of the switch setting at each time step. Since ex-

act inference is computationally intractable in such a model, approximate methods have been employed. A more detailed description of the FSLDS's structure, training procedure and the inference algorithms will be given in Chapter 3. The model has been especially tuned for the task of physiological condition monitoring. It has proved to be highly successful in determining the hidden factors that govern the observations collected by cotside computers.

As a crucial part of training the FSLDS, a manual *calibration* stage is needed. This requires finding an interval of normality for each examined baby. By normality, we generally understand a period in which the baby is in a stable condition. Once such an interval is detected, it is used to determine the parameters of the FSLDS associated with normal behavior. Subsequently, these parameters help in finding the dynamics of other known regimes.

The primary goal of this project is to *automate* this calibration stage. More exactly, we will build a binary classifier that predicts whether an interval of monitoring data is normal or not. The main reason for the feasibility of such a classification is that while the normal dynamics can be different for each baby, artifact is stereotypical. This assertion will be further explained throughout the dissertation. Crucial parts of our work were creating a more appropriate labeling for the data and extracting a relevant set of features to use as classifiers' input.

Performance evaluation for the auto-calibration procedure is done in two phases. First, we assess the quality of the predictions produced by the classifiers. Since only classifiers that have probabilistic outputs have been employed, we give our results as ROC curves. Furthermore, a performance criterion especially designed for the task is discussed. Second, we use the predictions of the classifiers in order to train the FSLDS and then use it to do inference. This analysis is much more interesting since it allows a direct comparison between the manual and auto-calibrated models. Our main conclusion is that the auto-calibrated FSLDS can be confidently used to perform neonatal condition monitoring. Examples of the auto-calibrated FSLDS in action are also presented.

1.1 Outline of the dissertation

Chapter 2 is an introduction into physiological monitoring in a neonatal intensive care unit. The most important aspects of the data used in this project are described. Explanations of the measurement channels and of common monitoring patterns are offered along with illustrative plots.

Chapter 3 is dedicated to the Factorial Switching Linear Dynamical System (FSLDS). It begins by reviewing the previous work in the area of probabilistic dynamical models that has allowed the construction of the FSLDS. Then we explain the structure of the FSLDS, its train-

ing procedure and the appropriate inference algorithms. A special emphasis is placed on the mandatory calibration stage required by the training procedure.

Chapter 4 shows our approach for automating the calibration stage needed by the FSLDS. Confronting the goals of the project with the results of the exploratory data analysis, we are able to articulate our design choices. The preference for a channel based classification of normality is motivated. Relying on the medical considerations reviewed in Chapter 2, we justify a new labeling for the data and our selections for feature extraction. Finally, the choice of the classifiers and the setup for training and testing are explained.

Chapter 5 shows the results obtained by employing the classifiers built in the previous sections. First, we study the probabilistic outputs of the classifiers. Then, we demonstrate that the auto-calibrated FSLDS is able to match the performance of the manually-calibrated model in terms of inference quality. We also plot various inferred distributions of the switch settings in support of our claim.

Chapter 6 presents and comments the main findings of the project and gives a set of recommendations for future work.

Chapter 2

Physiological Monitoring Data

This chapter is a short introduction into the medical aspects of the physiological condition monitoring project. The information presented in the following is essential in order to justify the design choices made for achieving the goal of automating the calibration stage of the FSLDS. A more detailed discussion of the topic can be found in [22].

In a neonatal intensive care unit, premature babies are continuously monitored by a set of devices placed next to their incubators. These patients have a very delicate health condition. This means that they may be treated for prematurity alone, having no other medical problem than their very short gestation term.

As already stated, the state of health of a baby cannot be observed directly, but there are a number of patterns appearing in the measurements which help clinicians formulate an appropriate diagnosis. More exactly, the medical staff looks at readings which include channels like heart rate, blood pressure or incubator humidity and infer whether there is any kind of pathology or whether everything is evolving normally.

Nevertheless, performing inference is a seriously complicated task for reasons which fall into two main categories: reasons due to the baby's physiology and reasons due to the monitoring equipment. The main issue in the first category is that human physiology is so complex that the number of factors which can affect it is hard to manage. In addition, the baby can display multiple combinations of physiological events occurring at the same time. To complicate things even further, some patterns appear quite frequently while others are particularly rare. At the same time, operating the monitoring equipment causes problems such as observation noise and corruption with artifact. A straightforward example of the latter is opening the incubator's doors, which usually produces measurements with increased variance on most of the observation channels.

Section 2.1 enumerates some of the most important measurement channels. We then show some intervals of normal variation in a baby's vital signs in Section 2.2. Examples of common physiological and artifactual events are presented in Section 2.3.

2.1 The measurement channels

Based on their preliminary beliefs about the physical condition of the baby, clinicians choose a set of observation channels to be monitored on a second-by-second basis. This means that it is highly probable for different babies to have recordings on different sets of measurement channels. In other words, it is common to see that some of the clinical parameters listed in the following are missing from the data belonging to a certain patient.

The physiological system can be thought of in terms of three partly independent sub-systems: the respiratory system, the cardiovascular system and the thermoregulatory system [22, §2.1.1]. Each of these systems has its associated set of measurement channels. The most significant of them are given in the next paragraphs.

The respiratory system is responsible for the transfer of various gases in the blood. The processes of oxygen absorption and carbon dioxide dispersion should be familiar even to non-medical staff. In order to monitor these processes a transcutaneous probe measures the partial pressure of O_2 (TcPO₂) and the partial pressure of CO_2 (TcPCO₂). At the same time, a pulse oximeter measures the O_2 saturation in the arterial blood (SO).

The flow of blood through the body is controlled by the cardiovascular system. This is traditionally monitored by obtaining heart rate (HR) measurements from an electrocardiogram. In addition, a pressure transducer records the evolution of the blood pressure on two different channels. When the heart is contracting, the first one accumulates the systolic blood pressure (BS); when the heart is resting, the second one accumulates the diastolic blood pressure (BD).

The thermoregulatory system does the job of keeping the body at an adequate temperature. This is monitored by two channels: core temperature (TC) and peripheral temperature (TP).

Along with this set of physiological measurements, clinicians also need to know that the environment inside the incubator functions in normal parameters. Two observation channels are allocated for the task: incubator temperature (Incu.Air Temp) and incubator humidity (Incu.Air Humidity).

2.2 Normal variation

Being asleep and motionless for most of the time, premature born babies usually have a stable health condition. By a stable condition, we understand a period that does not display any kind of pathology or other types of physiological events (see next section for examples). However, the fact the baby is in a stable condition does not rule out the existence of other long-term health problems [22, §2.2]. We classify an interval of monitoring data as normal if the baby is in a stable condition and none of the channels is corrupted by artifact (see next section for examples). One should bear in mind the point that such a period is well described by low variation in the measurement channels. (Also note that channels like heart rate or blood pressures inherently have a much higher variability than channels such as temperature.)

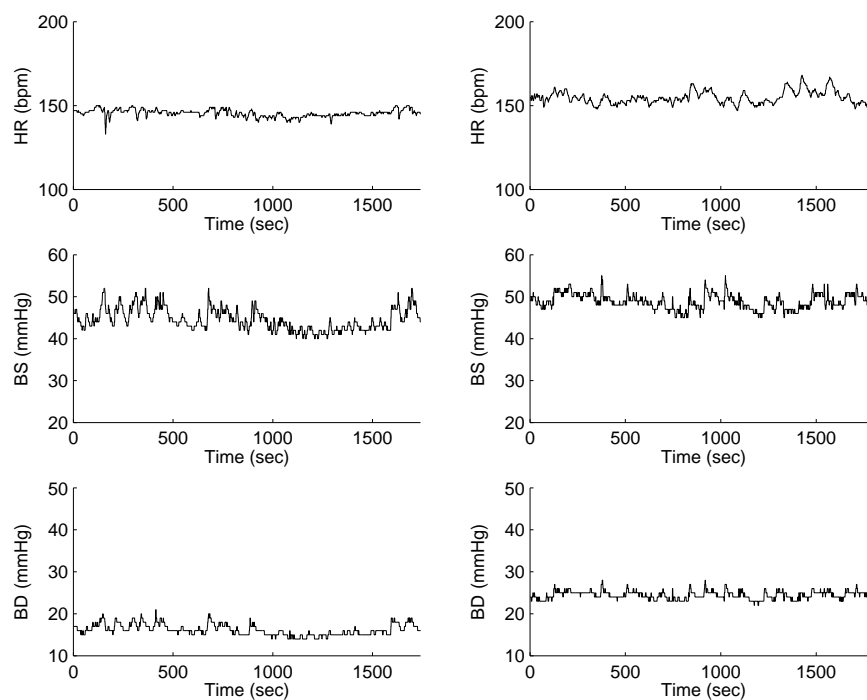


Figure 2.1: Intervals of annotated normality on channels monitoring the cardiovascular system for two different babies (left column - Baby 10 and right column - Baby 13). Heart rate measurements clearly follow dissimilar patterns, while blood pressure measurements vary around baseline signals located at distinct levels.

In the introductory chapter of this dissertation we highlighted the fact that different babies have different normal dynamics. From the medical’s staff point of view, the fact that each baby has a specific physiology is a perfectly reasonable assumption. Moreover, a simple visual inspection reveals that “low variation” is merely a general formulation for describing normality that fails to capture the particularities of the dynamics of the observation channels.

Focusing only on the channels associated with the cardiovascular system, Figure 2.1¹ shows the differences in normality between two distinct babies. The importance of these differences increases significantly when the FSLDS model needs to compute sets of autoregressive coefficients from intervals of normality on each channel (see Chapter 3).

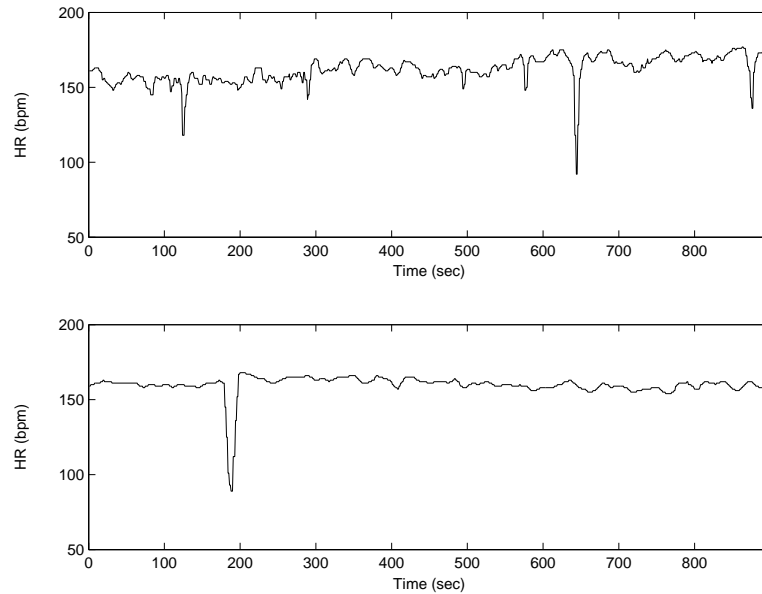


Figure 2.2: Two instances of bradycardia: in the upper plot (Baby 1) starting around $t = 625$ and in the lower plot (Baby 7) starting around $t = 175$. Notice 2 other possible bradycardia events in the upper plot starting around $t = 125$ and $t = 875$. The annotator considered these are not significant enough to be (binary) classified as bradycardia.

2.3 Physiological and Artifactual events

This section illustrates the physiological and artifactual events we plan to uncover by doing inference in the auto-calibrated FSLDS.

- **Bradycardia** is physiological event characterized by a temporary drop in the heart rate measurements. It can have a large variety of causes, but not all of them are of serious concern for the clinicians ([22, §2.4.1]). A couple of examples are shown in Figure 2.2. As all other labelings in this project, the annotation for bradycardia is a binary one (presence or absence). The shortcoming of this approach is that there is no way to quantify uncertainty in labeling; this becomes problematic in situations like the one in the upper plot of Figure 2.2.

¹For personal data protection reasons, the names of the babies are not available. We are using numbers instead. See Section 4.1 for a more detailed description of the dataset.

- **Probe disconnection** is a frequent artifactual event related to operating the monitoring equipment. Generally, when a probe is disconnected the measurements fall to zero. The exceptions are the temperature channels which drop to the temperature of the environment. Since temperature probe disconnection will be studied in the following chapters, we show a couple of examples in Figure 2.3.

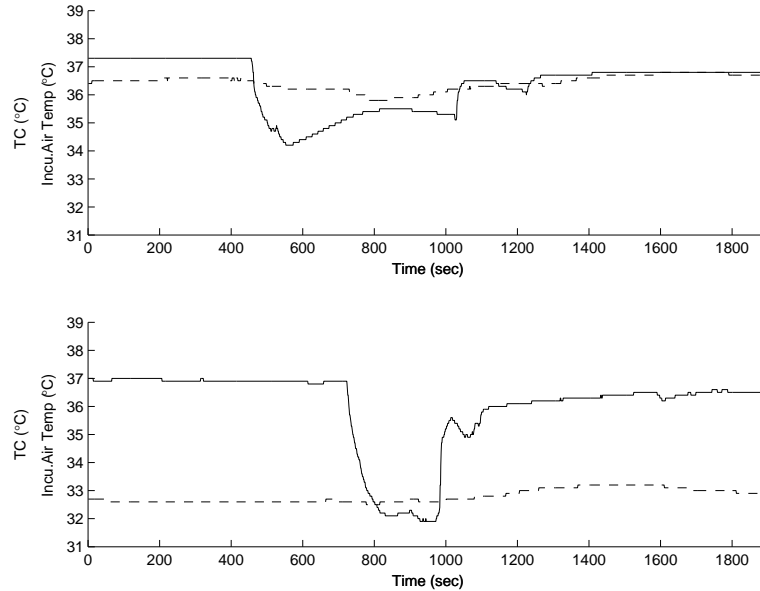


Figure 2.3: Two instances of temperature probe disconnection (Baby 4 - upper plot and Baby 14 - lower plot). Incubator temperature measurements are shown in dashed lines. Core temperature measurements below the incubator temperature are likely to be caused by placing the core probe near the incubator doors.

- Periodically taking a **blood sample** is another artifactual event. The procedure causes the emergence of an artifactual ramp in the blood pressure measurements (see Figure 2.4). Moreover, if the heart rate is also computed from the pressure sensor, readings will cease for the duration of the blood sampling event [22, §2.3.2].
- A common artifactual event is **opening the incubator's doors**. This is caused by various medical procedures that need to be performed on the patient. During this operation, we usually see an increased variance in the physiological measurement channels. At the same time the incubator's temperature and humidity slowly adjust to room values (see Figure 2.5 for examples).

It must be clearly stated that the number of different events is much larger than the one presented so far in this chapter. Other events worth mentioning include transcutaneous probe recalibration (artifactual), desaturation, or pneumothorax (both physiological) [22]. Trying to model all of them appears to be an infeasible task, since certain combinations of events might

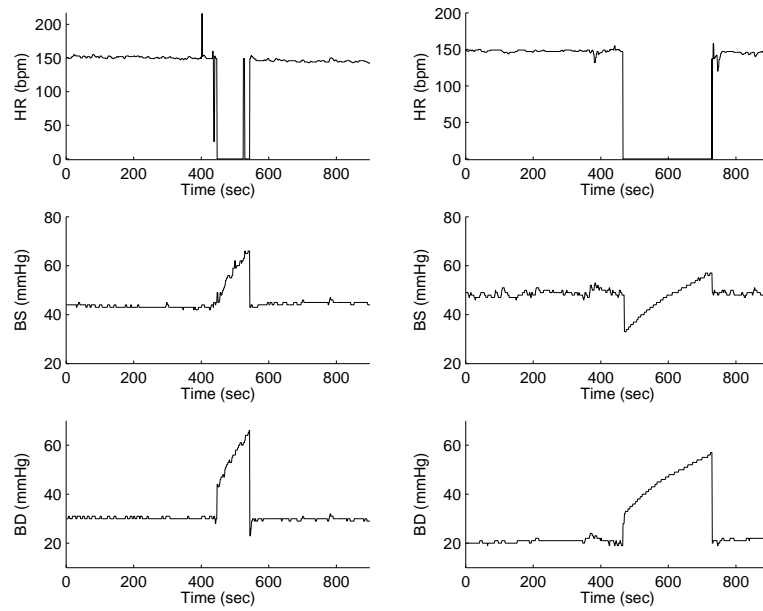


Figure 2.4: Two instances of blood sampling events (left column - Baby 5 and right column - Baby 10). Note that the events have not only different length, but also the relationship between the median value of the channel and the extreme values on the ramps is hard to define. In both cases, heart rate readings cease during the procedure.

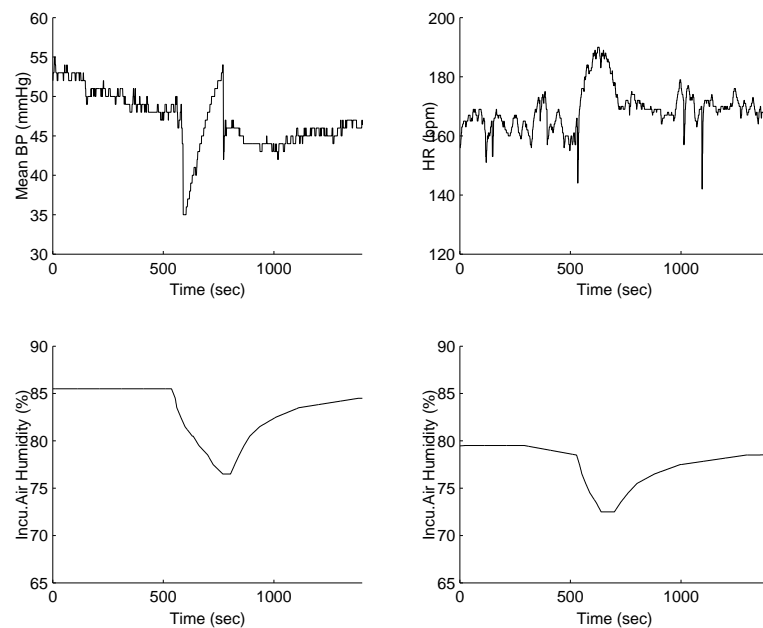


Figure 2.5: Two examples of opening the incubator doors (left column - Baby 13 and right column - Baby 8). In the left panel, we notice that the incubator was opened for taking a blood sample. In the second case, an unspecific pattern appears on the heart rate measurement when the incubator's doors are unclosed.

Table 2.1: A list of known physiological and artifactual events and the channels they can influence

Known event	Dependant measurement channels
Incubator Open	HR, BS, BD, SO, TC, Incu. Air Temp, Incu. Air Humidity
Blood Sample	HR, BS, BD
Bradycardia	HR
Temp Probe Disconnect	TC, Incu. Air Temp
Transcutaneous Probe Recal	TcPO ₂ , TcPCO ₂

even trigger patterns that have never been previously seen for specific babies. An elegant solution called the X-factor is proposed in [24, 22] and will be briefly discussed in the next chapter.

In the conclusion of this section, we are able to produce a summary as Table 2.1. The table answers the following question: “*For each analysed physiological or artifactual event, which are the measurement channels that are influenced by its occurrence?*”. This information will prove to be important for the design of the classifiers in Chapter 4.

Chapter 3

Background

After presenting the physiological monitoring data in the previous chapter, we are now able to discuss the probabilistic dynamical model developed for the condition monitoring application. The first part of the chapter has a more theoretical flavor, but is mandatory for understanding the remainder. We will also highlight the elegant manner in which the FSLDS manages to handle the particularities of the baby monitoring problem.

Section 3.1 presents the most important principles from the previous work on probabilistic dynamical models. We discuss autoregressive processes and state-space models together with generalizations of the latter. In Section 3.2, we show how the problem of neonatal condition monitoring is modeled by the Factorial Switching Linear Dynamical System (FSLDS) and what the main assumptions are. Training the FSLDS is then explained, with a special accent on the calibration stage. In the final part of the chapter, we study the inference problem, focusing on a method called the Gaussian sum approximation.

3.1 Previous work on probabilistic dynamical models

The most relevant theoretical concepts the FSLDS relies on are briefly described in the following. We begin by showing the ideas behind autoregressive processes and continue with the state-space model and its generalizations. Some important notions like switching and factorization are introduced. The intractability of exact inference in the generalizations of the state-space model is also discussed.

Autoregressive processes

The best known method to model a time series is probably the autoregressive process [4, §4]. In its simplest form, the autoregressive process ($AR(p)$) assumes that the current observation is

given by a weighted average of the observations at p previous time steps plus Gaussian noise. If we denote the observation sequence by y_t , then at any time t we have:

$$y_t - \mu = \sum_{k=1}^p \alpha_k (y_{t-k} - \mu) + n_t \quad (3.1)$$

where μ is the mean of the process and n_t is a zero-mean Gaussian noise with variance σ_n^2 . If the process is also assumed to be stationary, the solution is given by the Yule-Walker equations. These are a set of linear equations relating the AR coefficients, α_k , to the sample autocorrelation coefficients (for details see [4, §3.4.4]).

Popular generalizations of the $AR(p)$ process are the autoregressive moving average model ($ARMA(p, q)$ - when the observation also depends on a weighted sum of the noise from q previous time steps) and the autoregressive integrated moving average ($ARIMA(p, d, q)$ - an $ARMA(p, q)$ model applied to the signal after differentiating it d times). Note that the $ARIMA$ model is dedicated to non-stationary time series. The general approach is to assume that differencing for an appropriate number of times produces a stationary signal [4, §4.5]. An $ARMA$ model may now be fitted to the differenced data. For instance, differentiating an $ARIMA(p, d, 0)$ process d times results into a $AR(p)$ process that can be learnt using the Yule-Walker equations.

State-space models

In machine learning, a common model for time series is a two-layered directed acyclic graph (DAG), generically referred to as the state-space model (SSM) [3]. For each variable in the observed layer there is a corresponding state variable in the hidden layer, which is a first order Markov chain. There are two key conditional independence properties that hold in SSMs: (a) the current observation is independent of all other observations given the current state and (b) the past and future states are independent given the current state.

Two special cases of SSMs have been extensively studied: the Linear Dynamical System (LDS) [7, 3]¹, where all the variables are Gaussian and the Hidden Markov Model (HMM) [25, 3], where the hidden variables are discrete. The HMM has been widely applied in speech recognition applications. However, in the following we will only explore the LDS, since it constitutes the foundation of the condition monitoring application.

The LDS was developed in order to offer an elegant answer to the following question: “*How can one measure an unknown non-constant quantity using a noisy sensor?*”. The importance of this solution is tremendous considering today’s great variety of devices which rely on data

¹The Linear Dynamical System (LDS) is also referred to as the Kalman Filter (KF). For clarity reasons, only the first naming will be used throughout this dissertation.

coming from their sensors. We now briefly show the theory which supports the LDS. If one denotes the hidden states by $\mathbf{x}_t, t = 1, 2, \dots, T$ then the following linear and time invariant equation describes the dynamics of the system:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t \tag{3.2}$$

where \mathbf{A} is a square matrix called the dynamics matrix and \mathbf{w}_t denotes the Gaussian noise, $\mathbf{w}_t \sim N(0, \mathbf{Q})$. The observations, $\mathbf{y}_t, t = 1, 2, \dots, T$, are also governed by a linear and time invariant equation:

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{v}_t \tag{3.3}$$

where \mathbf{C} is the observation matrix and \mathbf{v}_t is the Gaussian observation noise, $\mathbf{v}_t \sim N(0, \mathbf{R})$. A corresponding directed acyclic graph (DAG) is shown in Figure 3.1 (upper-left corner). The factorization of the complete-data probability distribution is:

$$p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = p(\mathbf{x}_1)p(\mathbf{y}_1 | \mathbf{x}_1) \prod_{t=2}^T p(\mathbf{x}_t | \mathbf{x}_{t-1})p(\mathbf{y}_t | \mathbf{x}_t) \tag{3.4}$$

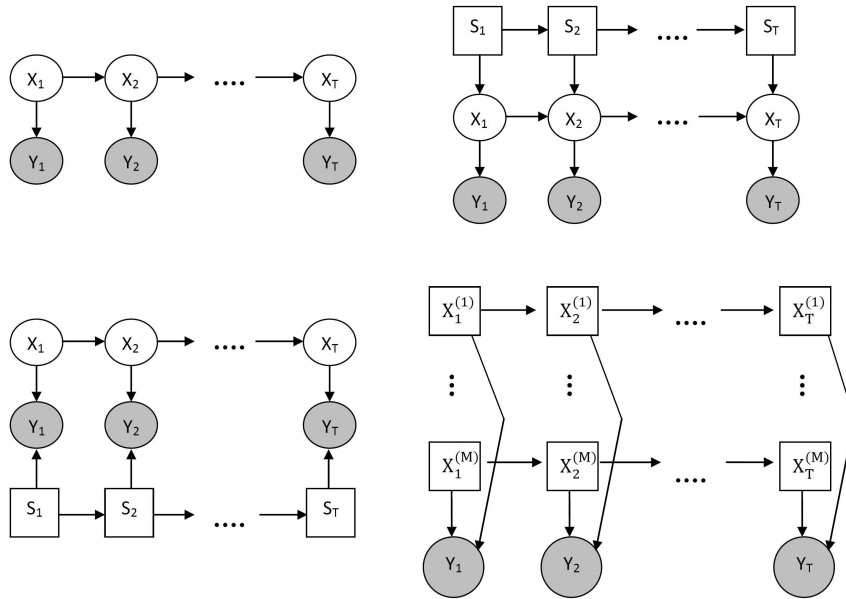


Figure 3.1: DAGs of various probabilistic models: SSM (upper-left corner), switching dynamics (upper-right corner), switching observations (lower-left corner) and Switching SSM (lower-right corner). Shaded variables are observed; circles represent continuous variables and squares represent discrete ones.

Once the model has been built, we can move on to the inference problem. In a graphical probabilistic model, inference is defined as computing the probability density distributions of the hidden variables conditioned on the visible ones. If the observations come from a temporal sequence of measurements, the following categorization of doing inference can be performed. When computing posterior probabilities at the current time, t , given the observations up to time t , the inference problem is called *filtering*. In the *smoothing* problem the conditioning is made on all the available observations, say up to time T . Furthermore, *prediction* consists of computing posterior probabilities at time t , given observations up to time t' , where $t > t'$. In the present project, we have only used filtering distributions.

In the LDS presented above, if $p(\mathbf{x}_1)$ is a Gaussian then, at any time step t , $p(\mathbf{y}_t | \mathbf{x}_t)$ and $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ would be Gaussians as well. In this restricted framework inference can be computed in closed form using the Kalman filtering (or smoothing) equations. In filtering, we are interested in computing the mean and variance of the following:

$$p(\mathbf{x}_t | \mathbf{y}_{1:t}) = N(\boldsymbol{\mu}_t, \mathbf{V}_t) \quad (3.5)$$

It can be shown that $\boldsymbol{\mu}_t$ is computed as the sum of a predicted mean derived directly from the dynamics equation and a weighted correction term derived from the observation at time t . The weight of the correction term is known as the Kalman gain matrix. (for details about inference in the LDS see either [3, §13.3], or [7]).

Generalizations of the SSM

As the HMM shares a analogous inference algorithm with the LDS, the parameters of both models can be trained in a similar fashion using expectation-maximization (EM) [3, §13]. Nevertheless, the assumptions they make about the probability distribution of the data proved to be too restrictive. In many real-world scenarios it is fruitful to think that the observations are generated by the mixture of a number of hidden processes evolving in parallel. Two of the ideas that made generalizations of the SSMs possible are switching and factorization.

The *switching* idea [26, 8, 18] implies introducing a discrete variable to determine the hidden process responsible for the current observation. For example, in the mixture of Kalman filters model [18] (see Figure 3.1, upper-right) the discrete switch variable S_t determines which matrix \mathbf{A}_t should be used by the dynamics equation at time t . Note that the switch variable follows a Markov chain. This model is sometimes referred to as the Switching Linear Dynamical System (SLDS). Another approach is to keep the dynamics equation linear and time invariant, but to relax these conditions for the observation equation. Such a model was proposed in [26] (see DAG in Figure 3.1, lower-left) and has an interesting motivation. The \mathbf{y} vectors represent

sensors, while the \mathbf{x} vectors are moving targets that need to be tracked. Since one is uncertain about which target is detected by any of the sensors, it is sensible to build a \mathbf{C} matrix for each existing possibility.

Factorization refers to transforming a discrete state variable into a cross product of factors, each linked to an a priori independent hidden process. In Figure 3.1, lower-right, we give the DAG of the Factorial Hidden Markov Model (FHMM) [9], a model that uses the idea of factorization to generalize the HMM. The FHMM shares the same complete-data probability distribution as the HMM:

$$p(\mathbf{S}_{1:T}, \mathbf{y}_{1:T}) = p(\mathbf{S}_1) p(\mathbf{y}_1 | \mathbf{S}_1) \prod_{t=2}^T p(\mathbf{S}_t | \mathbf{S}_{t-1}) p(\mathbf{y}_t | \mathbf{S}_t) \quad (3.6)$$

However, this time each state variable S_t is represented as the cross product of a set of discrete state variables (i.e. has a factorised representation):

$$S_t = S_t^{(1)} \otimes S_t^{(1)} \otimes \dots \otimes S_t^{(M)} \quad (3.7)$$

Each of these M state variables can take one of $K^{(m)}$ values. It is also natural to consider that each of them evolves a priori independently of the others:

$$p(S_{t+1} | S_t) = \prod_{m=1}^M p(S_{t+1}^{(m)} | S_t^{(m)}) \quad (3.8)$$

The particularity of the FHMM, is that at any time step t , the observation depends on all state variables. For instance, if the observed probability distribution is a Gaussian, then its mean may be a linear combination with weights given by the states of the M discrete variables. The model has been successfully used for modeling a collection of musical pieces in [9].

In contrast to the basic SSMs (LDSs and HMMs), performing exact inference in all the generalizations presented above is computationally intractable [14]. Theoretically, the same type of recursive equations can be applied, but this requires keeping track of a number of terms which grows exponentially with the number of time steps in the analysed sequence. The reason for this exponential growth is that all possible switch settings need to be taken into account.

Luckily, approximate inference methods have been formulated in order to obtain tractability. Variational inference [9, 8], Gaussian sum approximations [1, 18, 12, 2] or Rao-Blackwellised particle filtering [17] are probably the most important solutions so far. The main idea behind variational inference is to optimize the parameters of a simpler computationally tractable distribution that approximates the computationally intractable original distribution [16, §33]. The parameter optimization is done by minimizing the Kullback-Leibler divergence between the

approximated and the original distributions. The key of Gaussian sum approximation is to keep the number of terms needed at each time step constant. We will return to this method in the following section. Rao-Blackwellised particle filtering is a sampling based procedure whose primary advantage is speed.

3.2 The Factorial Switching Linear Dynamical System

In the previous section, we have seen that one major disadvantage of the probabilistic dynamical models discussed was their lack of flexibility. We have also presented two solutions, switching and factorization, also noting that they come at the cost of computationally intractable exact inference. The Factorial Switching Linear Dynamical System (FSLDS) [27, 22, 24] has been especially designed to address this lack of flexibility. The model has proved its ability to produce accurate inferences in the baby monitoring application.

As hinted in Chapter 2, there are a number of assumptions to be made in order to build a model for the condition monitoring task [22]. First, it is considered that each baby has its own normal dynamics. Furthermore, we assume that normality has a stationary behavior for each baby. Another assumption is that the dynamics of the known artifactual or physiological events is independent of the patient. For instance, we presume that all blood sampling events share the same dynamics irrespective of the baby.

In the following, we will discuss how the FSLDS solves the physiological condition monitoring task. We begin by giving the structure of the model. Then, training the FSLDS is explained emphasizing the importance of the calibration stage. In the end of the section, we show how inference is done utilizing the model.

3.2.1 The structure of the FSLDS

The FSLDS cleverly combines both switching and factorization ideas with the advantages of autoregressive processes to model baby monitoring. In the traditional Switching Linear Dynamical System (SLDS) [26], the discrete states evolve according to Markovian transition probabilities, $p(s_t | s_{t-1})$, and determine which set of parameters is used by the dynamics and observation equations at the current time step. The joint distribution of such a model is:

$$p(s_{1:T}, \mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = p(s_1)p(\mathbf{x}_1)p(\mathbf{y}_1 | \mathbf{x}_1, s_1) \prod_{t=2}^T p(s_t | s_{t-1})p(\mathbf{x}_t | \mathbf{x}_{t-1}, s_t)p(\mathbf{y}_t | \mathbf{x}_t, s_t) \quad (3.9)$$

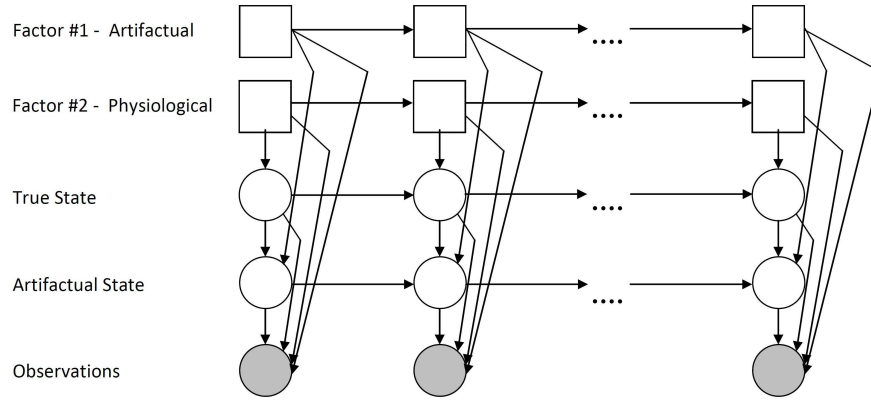


Figure 3.2: The DAG of a FSLDS with 2 factors (one physiological and one artifactual). The state is divided into dimensions that approximate the “true” physiology and dimensions that approximate the artifactual patterns. Shaded variables are observed; circles represent continuous variables and squares represent discrete ones.

However, in applications like physiological condition monitoring there is a large number of factors influencing the dynamics of the system. In fact, the union of all the distinct artifactual or physiological events forms the set of factors. We have already seen in Chapter 2 that even enumerating the possible events is a challenging (or even impossible) task. Moreover, in the SLDS model above we would need a switch variable that can take as many values as the number of distinct combinations of factor settings. Since this would have made inference awfully slow, the solution was to use a factorized representation of the switch variable. If all those discrete factors are assumed to be a priori independent, the transition probabilities can be written as:

$$p(s_t | s_{t-1}) = \prod_{m=1}^M p(f_t^{(m)} | f_{t-1}^{(m)}) \quad (3.10)$$

There is also an interesting interpretation for the discrete factors and continuous hidden states in the FSLDS model for baby monitoring (see Figure 3.2). Factors can be either linked to the physiology of the patient (e.g. bradycardia) or to artifact (e.g. incubator open, blood sample). At the same time, the continuous hidden state dimensions can be either associated with the true values of the channels or with artifact. Also note that artifactual factors can affect only artifactual states. It is important to keep track of the artifact too, because this allows an inspection of its patterns.

The previous paragraph has hinted a valuable representational choice that is permitted by the structure of the FSLDS. For each of the measurement channels, there is precisely one visible dimension in the observation vector at time t , y_t , and there are one or more hidden continuous state dimensions in the state vector at time t , \mathbf{x}_t . The goal is to increase the capability of the system to keep track of the “true” physiological signals. In addition, the dynamics matrices, \mathbf{A} ,

and dynamics noise matrices, \mathbf{Q} , are chosen to have a “block diagonal” structure [24]. This is a great advantage, since we have seen that the set of available observation channels usually varies from baby to baby. Further benefits of this representation will surface in the next paragraphs, where training of the FSLDS is discussed.

3.2.2 Training the FSLDS. The calibration stage

Learning in the FSLDS model is facilitated by the fact that part of the regimes in the data can be interpreted. In other words, annotated intervals for various known factors (e.g. bradycardia, incubator open) are available. The benefit is that when the hidden switch state is known, we can condition on it, making learning equivalent to the one in a regular LDS. If the annotated data hadn't been available, the entire FSLDS could have been trained by EM, in a similar fashion to the training of the switching models described in the previous section.

Furthermore, having labeled data allows us to learn the transition probabilities just by counting the number of transitions from one state to the other. In [22], the assumption that transitions are allowed only between normality and any other factor is also used for the purpose of reducing computational expense.

For the moment, training the full switching system has been reduced to training a simple LDS for each switch setting. The question now is if it is sensible to do a separate training for each possible switch setting. The answer given in [24] is no. It is always the case that some factors “overwrite” others. For example, if an instance of bradycardia occurs while a blood sample is taken from the baby (and consequently the heart rate readings have ceased), we would not be able to detect it. In addition, normality is overwritten by any other factor and a dropout (i.e. a factor for the probe disconnection event) overwrites everything else.

Table 3.1 structures the overwriting ordering for the factors and channels used in this MSc project. The list of factors includes a dropout factor, four known factors (Blood Sample, Incubator Open, Temperature Probe Disconnection and Bradycardia), a factor for Normality and one for Abnormality. Since it is hard to manage all possible events that may occur in the physiological condition monitoring problem, the solution proposed in [22] was to use a factor generically called “Abnormal” or the “X-factor”. This factor is supposed to handle every deviation from normality other than the ones modeled by the known factors enumerated above. Note that the Abnormal factor can affect all the measurement channels, including the ones which are not directly related to the baby's physiology like the incubator's temperature.

Once we have established the general framework, we continue with explaining the training procedure for the LDSs obtained by conditioning on the switch settings.

Table 3.1: Overwriting ordering for 7 factors on 7 channels of interest for this project. A factor placed higher on the list overwrites a factor a with lower position.

	HR	BS	BD	SO	TC	Incu.Air Temp	Incu.Air Humidity
Dropout	•	•	•	•	•	•	•
Blood Sample	•	•	•				
Temp Probe Dis					•	•	
Incubator Open	•	•	•	•	•	•	•
Bradycardia	•						
Abnormal	•	•	•	•	•	•	•
Normal	•	•	•	•	•	•	•

We begin by analyzing the dynamics under the normal regime. Technically, normality corresponds to the LDS obtained when all the other factors are off. This stage of learning is called *calibration* and needs to be performed for each baby. In fact we would not be able to produce accurate inferences about a certain baby's condition unless we *calibrate* the FSLDS using an interval of normality from that baby. Calibration is done separately on each observation channel,

The dynamics parameters for all the observation channels are obtained from training autoregressive processes. The parameters may be further tuned by EM updates. Since the channels are inherently different, the types of processes to best model each of them are also different. For measurements like heart rate (HR) or blood pressures (BS and BD), the modeling approach was to consider that the observations are the sum of a slowly fluctuating baseline, b_t , and a high frequency signal, x_t [24, V.C]. In terms of AR processes, we might have an $AR(p_1)$ signal and an $AR(p_2)$ baseline:

$$\begin{aligned} x_t - b_t &\sim N(\sum_{k=1}^{p_1} \alpha_k (x_{t-k} - b_{t-k}), \eta_1), \\ b_t &\sim N(\sum_{k=1}^{p_2} \beta_k b_{t-k}, \eta_2) \end{aligned} \quad (3.11)$$

where η_1 and η_2 are the noise variances. For example, appropriate choices for the heart rate (HR) channel are using an $AR(2)$ signal and an $ARIMA(1, 1, 0)$ baseline. The corresponding state-space representation for the HR channel is:

$$\mathbf{x}_t^{HR} = \begin{pmatrix} x_t \\ x_{t-1} \\ b_t \\ b_{t-1} \end{pmatrix}, \quad \mathbf{A}^{HR} = \begin{pmatrix} \alpha_1 & \alpha_2 & 1 - \alpha_1 & -\alpha_2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & \beta_1 & \beta_2 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad (3.12)$$

$$\mathbf{Q}^{HR} = \begin{pmatrix} \eta_1 + \eta_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \eta_2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{C}^{HR} = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \quad (3.13)$$

where we keep the usual LDS notation, but have added the *HR* superscript to indicate that those are only the blocks corresponding to HR measurements and not the whole matrices used by the model. The oxygen saturation (SO) and the incubator's humidity (Incu.Air Humidity) are well modeled by directly fitting *AR*(1) processes. An *ARIMA*(1,1,0) model is used for each of the temperature channels (TC and Incu.Air Temp). Summing up, the main application presented in Chapter 4 and evaluated in Section 5.2 uses 7 observation channels which are represented by 18 hidden “true” state dimensions.

After training normality, we can use this information to assist learning dynamics under known factors. Some of these factors (e.g. blood sampling) are associated with extra “artificial” hidden state dimensions. As mentioned before, they allow the system to keep track of the patterns that appear under artifact. Counting these states results into a total of 23 hidden state dimensions for our application. The details of learning dynamics under known factors are given in [22, 24] and will not be referred to in this dissertation.

3.2.3 Inference in the FSLDS

We have previously enumerated three methods of doing approximate inference in the generalizations of the SSM. For the FSLDS, filtering has been performed for both Gaussian sum approximation and Rao-Blackwellized particle filtering [22]. Since the results of the former have been better, we will only use this method in the auto-calibration tests.

The Gaussian sum approximation we have employed has been proposed in [18]. At every time step, we are interested in the following values

$$\begin{aligned} \mathbf{x}_{t|t}^j &= E[\mathbf{x}_t | \mathbf{y}_{1:t}, S_t = j] \\ \mathbf{V}_{t|t}^j &= Cov[\mathbf{x}_t | \mathbf{y}_{1:t}, S_t = j] \\ M_{t|t}(j) &= Pr[S_t = j | \mathbf{y}_{1:t}] \end{aligned} \quad (3.14)$$

where a subscript notation of type $t_1|t_2$, means computing the statistic at time t_1 based on the sequence of measurements $1 : t_2$.

The main idea of the method is to keep at any time step a constant number of terms, equal to the number of states. If the system has M states, then propagating the Kalman equations one

step forward produces M^2 terms of the form:

$$\begin{aligned}
 \mathbf{x}_{t+1|t+1}^{i(j)} &= E[\mathbf{x}_{t+1} \mid \mathbf{y}_{1:t+1}, S_{t+1} = j, S_t = i] \\
 \mathbf{V}_{t+1|t+1}^{i(j)} &= \text{Cov}[\mathbf{x}_{t+1} \mid \mathbf{y}_{1:t+1}, S_{t+1} = j, S_t = i] \\
 M_{t+1|t+1}(i, j) &= \text{Pr}[S_{t+1} = j, S_t = i \mid \mathbf{y}_{1:t+1}]
 \end{aligned} \tag{3.15}$$

where the notation $i(j)$ indicates that the state at the current moment is j and that at the previous moment the state was i . A collapsing procedure is now applied to reduce the number of terms back to M . This simple method is called *moment matching*. We just fix j , pick $\mathbf{x}_{t+1|t+1}^{i(j)}$, $\mathbf{V}_{t+1|t+1}^{i(j)}$ and $M_{t+1|t+1}(i, j)$ for all M values of i and compute the unconditional moments $\mathbf{x}_{t+1|t+1}^j$ and $\mathbf{V}_{t+1|t+1}^j$. It can be shown that a Gaussian distribution having these moments is the nearest Gaussian to the original distribution in terms of Kullback-Leibler divergence [13].

Chapter 4

Building Classifiers

This chapter explains our solutions for automating the calibration stage required by the FSLDS for physiological condition monitoring. As detailed in the previous chapter, the calibration stage consists of training the parameters corresponding to normal dynamics. In order to do this, an interval that correctly reflects this regime needs to be found for each baby. In Chapter 2, we have defined normal variation as a period satisfying two criteria: (a) the baby is in a stable condition (i.e. there is no evidence of any kind of pathology or of other types of physiological events) and (b) all the analysed measurement channels are free of artifact. Our primary goal is to automatically find such an interval.

The general approach we took for solving the auto-calibration task was to build a binary classifier able to predict whether a period of monitoring data is normal or not. For convenience, we will now define the Non-Normal class which will contain all the intervals that are not normal¹.

Section 4.1 shows the main results of the exploratory data analysis undertaken. The next part, Section 4.2, explains how we combine the existing labeling with clinical knowledge in order to obtain a detailed problem formulation for our task. Section 4.3 discusses various approaches for doing feature selection. The final section of the chapter, 4.4, explains how we have chosen our classifiers and how we set up the tests in accordance to the quantity of data on hand.

4.1 Exploratory Data Analysis

While Chapter 2 has analysed the physiological monitoring data from a medical perspective, the current section does the same thing from a data mining perspective.

¹In this project, Abnormal and Non-Normal are different terms that should not be confused. Abnormal is the class that contains all physiological or artifactual events other than the ones for which we have explicit annotations.

Table 4.1: Summary statistics of the incidence of various factors computed for the 360 hours of monitoring data on hand

	Number of Incidences	Total Duration	Avg. Duration (sec)	Std. Dev. (sec)	3 rd centile (sec)	97 th centile (sec)
Incubator Open	689	41 hrs	214	341	40	847
Bradycardia	272	161 min	36	21	8	76
Blood Sample	91	251 min	166	79	44	321
Temp Probe Dis	86	572 min	399	626	22	2615
Abnormal	726	37 hrs	185	300	9	736

The dataset consists of 24 hours recordings taken from fifteen premature born babies at Edinburgh Royal Infirmary. All the 360 hours of monitoring data will be used in our experiments. The babies were around 24 to 29 weeks gestation and aged 1 to 16 days post-natal (However, this kind of information is available for only thirteen of the patients.). Manually made expert annotations are available for five known factors (Bradycardia, Blood Sample, Incubator Open, Temperature Probe Disconnection and Transcutaneous Probe Recalibration) and for the Abnormal factor [24]. As in [22, 24], due to the scarcity of examples in the dataset we will not use the Transcutaneous Probe Recalibration factor in any of the following experiments. In addition, one period of Normal data is highlighted for each baby. These carefully chosen intervals have been used up to now to “manually” calibrate the FSLDS.

The data has been recorded on a second-by-second basis and the set of measurement channels varies from baby to baby. The motivation is that according to their doubts about the health condition of each baby, the medical staff decide which channels should be monitored. Dropouts can appear on all the observation channels evaluated in the project and there is no specific annotation for them. However, they can be easily detected by a trivial inspection of the data. A sequence of zeros on any channel other than temperature measurements indicates a dropout on that channel. For temperature channels the dropout value is 20 °C. We will return to the problem of handling dropouts in Section 4.4. It is also important to mention that we will be working with an already preprocessed version of the data. Preprocessing consisted in smoothing the incubator humidity measurement, a procedure needed in order to correctly train the autoregressive process for this channel (i.e. $AR(1)$).

Table 4.1 presents a few significant statistics about our data. Counting the number of incidences of each factor, we notice that factors such as opening the incubator and abnormality are far more frequent than taking a blood sample or a disconnection of the temperature probes. We also note that the mean durations of the various events are quite different as well. A much more important information is revealed by inspecting the standard deviations and the 3rd and

97th centiles. Although artifactual events always respect the same patterns, there is a great deal of variability in their duration. The consequence is that the feature extraction task (Section 4.3) becomes more challenging.

We also know that a period for which there are no annotations is automatically considered a period of normality. Thus we can compute the total duration of normality in our data, which is 283 hours (79% of the total 360 hours). Note that this computation cannot be performed by summing up the total durations of the factors in Table 4.1, since there is a significant overlap between factors.

The original plan was to use another dataset of physiological monitoring measurements, called Neonate [5], for validation purposes. It consists of 83 measurement sequences recorded from sixteen premature infants for a total of around 363 hours of monitoring data. However, major problems such as a fairly different set of observation channels and annotations available only for normality prevented us from using the Neonate data in this project.

4.2 Finding an appropriate problem formulation

This section is dedicated to explaining the most important design choices made for building the auto-calibrated inference system. We will begin with problems related to the proper length of the intervals we want to classify. A special accent is then placed on creating a more useful labeling for our task. This is also closely related to picking the appropriate set of measurement channels. In the final part of the section we assess the superiority of a channel-based classification to an interval-based one.

4.2.1 Chopping the data

Since the objective is to extract some periods of normality from continuously recorded data, we have to begin by choosing an appropriate length for these intervals. Given that there are no scientific grounds to opt for sections of variable length, we chose to utilise only fixed length intervals. Moreover, for simplicity, no overlapping is permitted between the intervals. However, this approach has its unavoidable disadvantages. Since the measurements are literally chopped, we often split the continuous events between different fixed length intervals. The effect is that the quantity of information provided by the event's pattern is seriously limited in such cases.

The proper choice for the duration of an interval is another aspect that needs to be discussed. The ideal length has to satisfy two conflicting requirements. First, the period should be long enough for the purpose of having autoregressive coefficients accurately reflecting the channels'

Table 4.2: The channels of interest and the factors that can influence them

Channel	“Governing” factor
HR	Bradycardia, BloodSample, IncubatorOpen, Abnormal
BS	BloodSample, IncubatorOpen, Abnormal
BD	BloodSample, IncubatorOpen, Abnormal
SO	IncubatorOpen, Abnormal
TC	Temp Probe Dis, IncubatorOpen, Abnormal
Incu. Air Humidity	IncubatorOpen, Abnormal
Incu. Air Temp	Temp Probe Dis, IncubatorOpen, Abnormal

dynamics. Second, utilizing longer intervals also means that the chance of finding such periods free of artifact severely decreases. The fifteen annotated Normal intervals have extents varying between 15 and 40 minutes, with an average of approximately 26 minutes. Furthermore, the (weighted) average of an event’s duration is around 3 minutes. Putting all this information together, choosing a length of 15 minutes (i.e. 900 seconds) seems to be a reasonable choice.

4.2.2 Factors and Channels

Examining our annotations, we have concluded that we can use at most four known factors (Bradycardia, Blood Sample, Incubator Open and Temperature Probe Disconnection) to assess the performance of the FSLDS.

We now return to the information summarized in Table 2.1. It indicates for each known factor, which are the channels influenced by that factor. In other words, the table points out the channels which need to be observed in order to infer the presence or absence of the known factor. Consequently, consider the union of all the channels corresponding to the four factors of interest:

$$\{\text{HR, BS, BD, SO, TC, Incu.Air Humidity, Incu.Air Temp}\}$$

This set is the necessary and sufficient set of channels that need to be observed in order to set up a FSLDS capable to infer the four discussed factors. Our original problem of finding an interval of normality by looking at all the available channels for a baby has just reduced to looking at all the channels enumerated above. Note that this does not imply that all the channels in the set above are on hand for all the babies. One may also notice that introducing an observation channel not influenced by any factor in the FSLDS will have no effect on inferences because of the block diagonal structure of the \mathbf{A} and \mathbf{Q} matrices (see Section 3.2.2).

Another useful observation is that the set of known factors we plan to infer coincides with

the set of known factors that can influence the set of channels we have just determined. This is because the fifth available factor, Transcutaneous Probe Recalibration, cannot affect any of these observations. However, the Abnormal factor can influence any channel and must be taken into account. All this information is summarized by Table 4.2.

4.2.3 Interval-Based vs. Channel-Based Approach

To this point, we have found the set of factors we want to monitor, the seven channels of interest, the five factors (four known plus the Abnormal factor) that affect those channels and the length of intervals. With all this into place, the following paragraphs introduce two classification approaches: a straight forward “interval-based” classification and a more elaborate “channel-based” classification. We then explain why the second approach was favored.

In an “interval-based” approach, an interval of monitoring data on all seven channels of interest is tested for normality. In other words, for a selected period of fifteen minutes we look at all the channels and make a single prediction referring to all of them. There is also a simple way to label an interval. If its intersection with any period marking the presence of any of the five factors is the empty set, then the interval is labeled as being Normal. On the other hand, if at least one of the factors has a non-empty intersection with the interval, it will be assigned to the Non-Normal class.

The “channel-based” approach does a separate classification for each channel. In other words, we predict normality at channel level. For instance, questions like “*Is this period of heart rate measurements normal or not?*” are asked. In order to label our data we return to the information in Table 4.2. For each channel, if an interval’s intersection with any period marking the presence of any of the factors listed in the right column of the table is the empty set, then the interval is labeled as being Normal. On the other hand, if at least one of the factors on this list has a non-empty intersection with the interval, it will be assigned to the Non-Normal class. In fact, we have broken our classification problem into seven smaller classification problems, one for each measurement channel.

Our reason for pursuing the “channel-based” approach is that it makes a better use of the (limited) amount of data on hand. In the first approach, an important quantity of normal data is wasted during the labeling procedure. The reason is that it is often the case that during a fifteen minute period, only a factor not affecting all channels is present. According to the rules above, the interval will be assigned to the NonNormal class. However, let us assume the factor was bradycardia. Such an event affects only the heart rate measurements. From different point of view, this means that all the other channels were evolving normally during this period. Summing up, we have just lost possibly valuable information about normality on

Table 4.3: The baseline set of features extracted from each channel.

Channel name	Extracted features
HR	minimum-mean, median, standard deviation
BS	maximum-median
BD	maximum-median
SO	median, median-mean
TC	minimum, maximum, standard deviation
Incu.Air Temp	median, standard deviation
Incu.Air Humidity	median-minimum, standard deviation

these channels.

In the end of this section, we make an essential observation about our preferred “channel-based” approach. Note that when active, the Incubator Open and Abnormal factors may not affect the whole set of factors they can influence (the ones given in Table 2.1), but only a subset of them. Our solution was to look at all available channels before making our predictions, even if we follow a “channel-based” approach. This aspect will be reiterated in Section 4.4.

4.3 Feature extraction

It is beyond doubt that extracting good features is an essential requirement for the success of any classification problem. For auto-calibration task there are two main difficulties that need to be dealt with.

First, we are working with time series and rely on annotations signaling occurrences of a number of events. Since there is no way to control the distribution of those events in time, we do not know when they might start or end in our selected fifteen minute intervals. Even worse, by “chopping” the monitoring data into fixed length sections, we have often strongly altered the factors’ patterns.

Second, there is a significant amount of variability in the patterns of the known factors. Table 4.1 provides strong evidence with respect to the distribution of events’ duration. In addition, other sources of variability are the magnitude of the events (see the examples of Bradycardia in Figure 2.2) or the correspondence between the median value of normality and the levels of artifactual measurements (see the examples of Blood Sample in Figure 2.4).

Our first efforts were directed to building a baseline set of features for each channel. The results of this process are summed up in Table 4.3 and will be clarified below.

Normal heart rate (HR) measurements usually display a low amplitude high frequency fluctuation around a slowly changing baseline. An event affecting this channel will generally result into a higher variance, so we picked the standard deviation as a feature. The level of the baseline heart rate signal is captured by the median feature. In order to detect bradycardia, we have chosen to record the difference between the minimum and average values of the observations. The most common event influencing blood pressure measurements (BS and BD) is taking a blood sample. The difference between the maximum and median values of these channels has been experimentally found to capture such variations. The oxygen saturation (SO) channel's dynamics can be recorded by computing the median and the difference between the median and the mean of the observations. Moving to the core temperature (TC) measurements, we are interested for these values to stay within some acceptable lower and upper limits. Thus we pick the minimum and maximum values of the channel as features. The standard deviation also offers valuable information about the baby's condition. When the incubator's doors are opened we usually see a drop in the humidity measurements (Incu.Air Humidity). Consequently, we keep track of the standard deviation of the channel and of the difference between the median and minimum values of the channel. A similar rationale is applied for the incubator temperature channel (Incu.Air Temp).

Apart from this set of features, several other ideas have been explored. Some of the most interesting are: using the parameters of an LDS trained on the selected interval as features and inputting information like gestation or post-natal age into the classifiers. These attempts will be further discussed in the next chapter.

4.4 Classifier setup

This final section of the chapter shows how the available measurements, newly built labeling methods and feature extraction procedures are put together in the “channel-based” classification approach. The option for basic “off-the-shelf” classifiers is also justified.

Labels and feature spaces

In Section 4.2.3, we made clear the fact that, even if we follow the “channel-based” approach, we still have to look at all the measurements of interest before performing the classification. It is actually the labeling which crucially differentiates the way we make our predictions for various channels. In other words, two classifiers corresponding to different channels might be working in the same feature space, but will use different definitions for the Normal and Non-Normal classes. Furthermore, every time we look at a channel the same features will be extracted.

Table 4.4: Summary of data availability for the channels of interest. The right column contains the list of babies for whom the channel on the left column is unavailable.

Channel name	Baby indices
HR	-
BS	2, 3, 6, 7, 11
BD	2, 3, 6, 7, 11
SO	-
TC	-
Incu. Air Temp	-
Incu. Air Humidity	11,15

Handling missing data

We now compare our data with the set of seven observation channels of interest determined in Section 4.2.2. Unfortunately, we have to face the problem that there are babies for whom we do not have all these measurements on hand. The problematic channels are shown in Table 4.4. Our solution was to always input as much information as possible into our classifiers. Theoretically, we are making a Missing at Random (MAR) [15] kind of assumption about the absent measurement channels. However, this means we will have to train separate classifiers that work on different feature sets. Inspecting Table 4.4, we conclude that for the baseline set of features there are four types of classifiers. An inconvenience of this solution is that the amount of data available to test and train classifiers is reduced, especially for babies that have all the channels on hand (i.e. babies 1, 4, 5, 8, 9, 10, 12, 13 and 14).

Handling dropouts

Considering the manner in which we have chosen our (baseline) set of features, dropout measurements may raise serious problems. However, as hinted in Section 4.1, they can be trivially detected. Since we clearly don't want to calibrate the FSLDS using an interval that contains dropouts, we will dispose of periods containing such artifact from the very beginning. Alternatively, we could have used numerical interpolation to eliminate some of the shorter dropout sections.

“Off-the-shelf” classifiers

The classifiers employed for the task were logistic regression, Naïve Bayes and decision trees. These choices are mainly motivated by the simplicity, the easier interpretation of results and the reduced number of parameters associated with these classifiers. Furthermore, due to the amount of data on hand, we concluded that we could not afford keeping a separate validation

set.

Logistic regression is a probabilistic discriminative linear classification model [3, §4.3]. In our work we have used only the binary classification model, which implies using the logistic sigmoid function, σ . If the two classes are denoted C_1 and C_0 (notations according to the values of the targets: 1 and 0 respectively), then the idea is to model the posterior probability of class C_1 as:

$$p(C_1 | \mathbf{x}) = \sigma(b + \mathbf{w}^T \phi(\mathbf{x})) \quad (4.1)$$

where $\phi(\mathbf{x})$ is a vector of basis functions (in our implementation $\phi(\mathbf{x}) = \mathbf{x}$). In the maximum likelihood approach, the optimization of the parameter set, $\{\mathbf{w}, b\}$, cannot be performed explicitly [3, §4.3.2]. A recommended solution is the Iterative Reweighted Least Squares Algorithm (IRLS) [3, §4.3.3]. Furthermore, convergence guaranties are given by the quadratic form of the error function.

We have also run comparative tests with the Bayesian version of logistic regression. Since exact Bayesian inference is intractable [3, §4.5], we are using a Laplace approximation for the posterior distribution over the parameter set. The Laplace approximation is a Gaussian with mean given by the MAP solution and covariance equal to the curvature of the Hessian also computed at the MAP solution. We will sample parameter sets from this Gaussian and then average the predictions over these samples.

Naïve Bayes is a generative probabilistic model. The key simplifying assumption of this method is that, given a class label, all the attributes are conditionally independent. For the present binary classification task, we have used the binomial distribution to model prior class probabilities and univariate Gaussians for each of the attributes. Thus, each of the two class conditional probability distributions is written as a product of Gaussians.

Decision trees are a very intuitive class of discriminative probabilistic models. They create a tree by recursively partitioning the feature space in order to maximize a class purity score. The size of this tree is a very important measure of model complexity. Various tree building algorithms have been developed, but in this project we are using a method relying on the entropy criterion, C4.5. The entropy criterion [11, §10.5] finds a split such that the average entropy computed from the two resulting branches is minimized.

In order to avoid over-fitting, all the experiments in the next chapter are performed in a 3-fold cross-validation setting. For each of the three tests, ten babies are used for training and the remaining five are left for testing. More precisely, the three test sets are: babies 11 : 15, babies 1 : 5 and babies 6 : 10.

Due to the complication of needing classifiers trained on different sets of features, the number of classifiers to be trained will be different than three each time we perform a cross-validation test on a certain channel. For instance, to perform a 3-fold cross-validation test for the heart rate (HR) channel we will train seven classifiers. This can be checked by returning to Table 4.4. Let us consider our first fold having the test set consisting of babies 11 : 15. Because of channel availability we need one classifier for babies 12 : 14, one for baby 11 and another one for baby 15, for a total of three classifiers only for this fold. A similar rationale produces two more classifiers for each of the other two folds.

Chapter 5

Evaluation

This chapter describes the experiments done in order to assess the performance of the auto-calibrated FSLDS for neonatal condition monitoring. We have decided that it would be sensible to separate the evaluation task into two phases. First, we measure the success of the classifiers built in Chapter 4 and show our results in Section 5.1. The second phase of our analysis consists of comparing the quality of the inferences produced by the auto-calibrated FSLDS to the inferences obtained by the manually-calibrated system. The most relevant results of this evaluation are given in Section 5.2.

5.1 Evaluating the classifiers

We begin this section by explaining our approach for assessing the classifiers. As described in Section 4.4, all the experiments are performed in a 3-fold cross-validation setting. The main goal is to find a good set of features together with an appropriate classifier.

The quality of our predictions is measured by two criteria. First, we draw Receiver Operating Characteristics (ROC) curves for each employed classifier. A ROC curve shows the True Positive Rate (TPR) as a function of the False Positive Rate (FPR), when one varies a threshold applied on the classifier's outputs [6]. A widely accepted method for comparing classifiers using ROC curves is to compute the Area Under the ROC curve (AUC). The value of the AUC is the first criterion, with the observation that the larger the better.

However, our primary objective is to extract some intervals of normality from the data. This means that we do not necessarily look for the most accurate classification between Normal and Non-Normal intervals. In other words, it is sufficient for the employed classifiers to deliver some intervals that we can confidently consider to be typical for the Normal dynamics of a baby and then utilize them to calibrate the FSLDS. This consideration endorses our second criterion.

We will compare the classifiers based on how well they answer the following question: “*On a per baby basis, for how many positive instances (i.e. Normal intervals) does the classifier output a posterior probability of belonging to class Normal, $P(C = \text{Normal} | x)$, higher than the largest posterior of a negative instance (i.e. Non-Normal interval)?*”. For clarity, we will call this criterion the Interval Ranking Criterion (IRC).

In the following, we commence by discussing the results obtained using the baseline set of features (Section 5.1.1) and then we compare those with the ones obtained when we have used other features (Section 5.1.2).

5.1.1 The baseline feature set

In Chapter 4 we have built a baseline set of features primarily relying on medical considerations concerning the seven channels of interest. A summary of this set can be found in Table 4.3. Keeping the features fixed, we now compare the performance of three off-the-shelf classifiers: logistic regression¹, Naïve Bayes and a decision tree (C4.5²). Since the features we are using display intrinsically different ranges and variances, from this point on we will be exclusively working with standardized data (i.e. zero mean, unit variance). It is also important to mention that the decision tree was applied with its default parameters.

Since we have settled on having a different classification task for each channel, we naturally present our results separately as well. We illustrate the performance of heart rate classification in Figure 5.1 and Table 5.1, of systolic blood pressure classification in Figure 5.2 and Table 5.2 and of core temperature classification in Figure 5.3 and Table 5.3. A summary regarding all seven channels of interest is given in Table 5.4. Nevertheless, the observations made in the following are valid for all seven classification tasks.

The general conclusion is that logistic regression always outperforms the other two methods on both criteria of comparison. A distinguishing observation about logistic regression is that it has always found, for each baby, at least one positive interval with higher posterior probability of being normal than any negative instance. In any classification problem, it is useful to understand which attributes are relevant. With logistic regression we only have to inspect the magnitude of the learnt weights for each attribute. For instance, in the classification of the HR channel we have discovered that features like $\min(HR) - \text{mean}(HR)$, $\text{standard deviation}(TC)$ and $\text{standard deviation}(Incu.AirTemp)$ are most relevant. On the other hand, in the classification of the SO channel features like $\text{standard deviation}(HR)$ and $\text{standard deviation}(TC)$ proved to be the most significant.

¹Netlab’s [20, 21] implementation of logistic regression has been utilized. The optimization algorithm employed is IRLS.

²Weka’s [10] implementation of the C4.5 has been utilized.

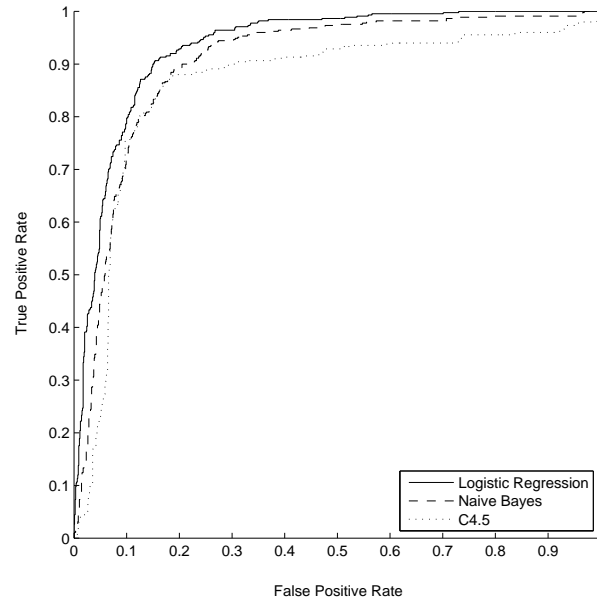


Figure 5.1: Classifying heart rate (HR) intervals - ROC curves for three classifiers.

Table 5.1: Classifying heart rate (HR) intervals - Performance comparison of three classifiers in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.

	Logistic regression	Naive Bayes	Decision tree (C4.5)
AUC	0.9317	0.9011	0.8605
baby 1 (24 positives)	4	0	0
baby 2 (49 positives)	25	26	6
baby 3 (28 positives)	4	5	0
baby 4 (16 positives)	5	1	0
baby 5 (40 positives)	5	1	0
baby 6 (33 positives)	14	6	0
baby 7 (45 positives)	28	10	0
baby 8 (11 positives)	3	3	8
baby 9 (37 positives)	5	13	15
baby 10 (35 positives)	2	7	0
baby 11 (43 positives)	1	7	2
baby 12 (14 positives)	8	4	7
baby 13 (21 positives)	3	1	0
baby 14 (31 positives)	19	19	0
baby 15 (23 positives)	5	0	0

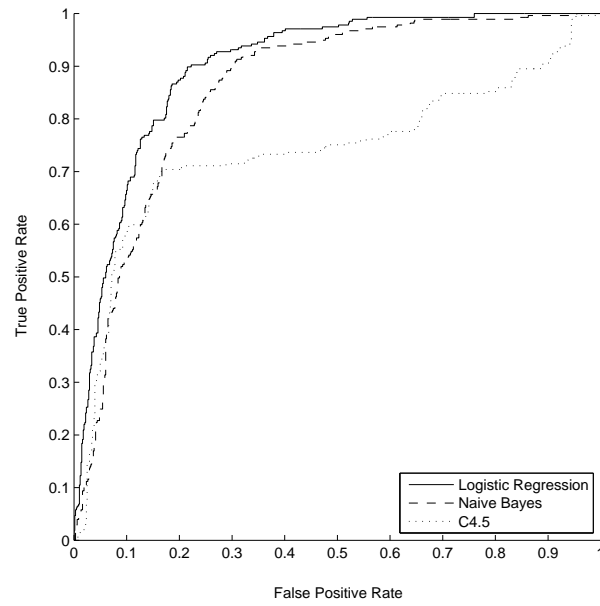


Figure 5.2: Classifying systolic blood pressure (BS) intervals - ROC curves for three classifiers.

Table 5.2: Classifying systolic blood pressure (BS) intervals - Performance comparison of three classifiers in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets. The “-” mark signals a missing channel for the current baby (i.e. no classification to be made).

	Logistic regression	Naive Bayes	Decision tree (C4.5)
AUC	0.8991	0.8570	0.7392
baby 1 (29 positives)	4	0	0
baby 2 (- positives)	-	-	-
baby 3 (- positives)	-	-	-
baby 4 (8 positives)	1	1	0
baby 5 (41 positives)	20	2	2
baby 6 (- positives)	-	-	-
baby 7 (- positives)	-	-	-
baby 8 (15 positives)	5	0	0
baby 9 (39 positives)	4	11	0
baby 10 (37 positives)	6	4	0
baby 11 (- positives)	-	-	-
baby 12 (19 positives)	9	6	0
baby 13 (24 positives)	2	2	13
baby 14 (40 positives)	10	11	1
baby 15 (25 positives)	1	0	0

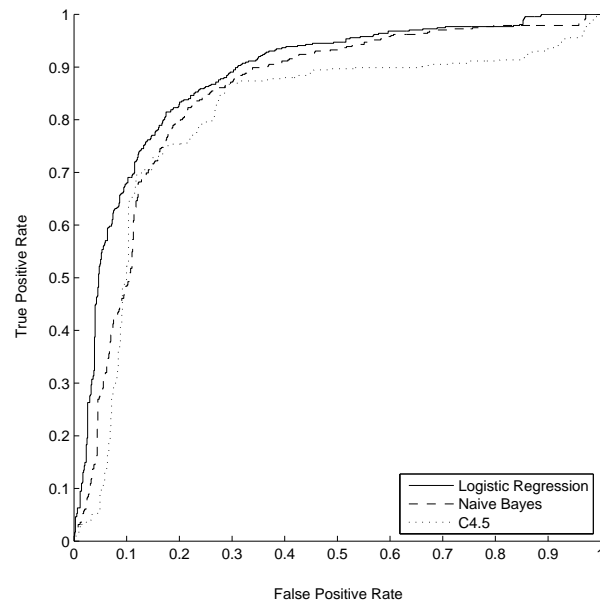


Figure 5.3: Classifying core temperature (TC) intervals - ROC curves for three classifiers.

Table 5.3: Classifying core temperature (TC) intervals - Performance comparison of three classifiers in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.

	Logistic regression	Naive Bayes	Decision tree (C4.5)
AUC	0.8927	0.8406	0.8152
baby 1 (29 positives)	2	0	0
baby 2 (53 positives)	1	1	0
baby 3 (43 positives)	7	1	0
baby 4 (18 positives)	8	1	0
baby 5 (41 positives)	27	1	0
baby 6 (56 positives)	6	5	0
baby 7 (47 positives)	29	20	0
baby 8 (15 positives)	8	6	0
baby 9 (39 positives)	4	15	0
baby 10 (36 positives)	9	6	0
baby 11 (46 positives)	2	14	-
baby 12 (18 positives)	9	5	0
baby 13 (24 positives)	3	3	0
baby 14 (41 positives)	11	18	15
baby 15 (29 positives)	2	1	0

Table 5.4: AUC comparison between the three classifiers for all seven channels of interest.

Measurement Channel	Logistic regression	Naive Bayes	Decision tree (C4.5)
HR	0.9317	0.9011	0.8605
BS	0.8991	0.8570	0.7392
BD	0.9020	0.8602	0.7252
SO	0.9210	0.9038	0.7989
TC	0.8927	0.8406	0.8152
Incu.Air Humidity	0.8803	0.8457	0.8025
Incu.Air Temp	0.8752	0.8381	0.7769

Table 5.5: Classifying heart rate (HR) intervals - Performance comparison between logistic regression and Bayesian logistic regression in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.

	ML Logistic regression	Bayesian logistic regression
AUC	0.9317	0.9328
baby 1 (24 positives)	4	4
baby 2 (49 positives)	25	29
baby 3 (28 positives)	4	3
baby 4 (16 positives)	5	7
baby 5 (40 positives)	5	5
baby 6 (33 positives)	14	14
baby 7 (45 positives)	28	28
baby 8 (11 positives)	3	3
baby 9 (37 positives)	5	6
baby 10 (35 positives)	2	2
baby 11 (43 positives)	1	1
baby 12 (14 positives)	8	8
baby 13 (21 positives)	3	2
baby 14 (31 positives)	19	18
baby 15 (23 positives)	5	4

Table 5.6: Classifying diastolic blood pressure (BD) intervals - Performance comparison between logistic regression and Bayesian logistic regression in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets. The “-” mark signals a missing channel for the current baby (i.e. no classification to be made).

	ML Logistic regression	Bayesian logistic regression
AUC	0.9020	0.9032
baby 1 (29 positives)	4	3
baby 2 (- positives)	-	-
baby 3 (- positives)	-	-
baby 4 (8 positives)	1	1
baby 5 (41 positives)	20	23
baby 6 (- positives)	-	-
baby 7 (- positives)	-	-
baby 8 (15 positives)	5	7
baby 9 (39 positives)	4	4
baby 10 (36 positives)	6	6
baby 11 (- positives)	-	-
baby 12 (14 positives)	9	9
baby 13 (24 positives)	3	2
baby 14 (41 positives)	9	8
baby 15 (25 positives)	1	1

Naïve Bayes offers a somewhat worse performance, clearly limited by the conditional independence assumption it places on the attributes. By returning to the information in Table 2.1, it is easy to see that such an assumption is violated. The table offers evidence that the channels do not evolve independently. Subsequently, the features we have extracted are not independent as well.

The decision tree delivers the poorest performance, especially if we analyse the second criterion. An explanation is that due to the inherent instability of the method, it was usual to have some negative instances getting a very high posterior probability of being normal. Moreover, no parameter optimization has been performed.

So far it has been demonstrated that logistic regression seems to be the method of choice for our classification tasks. We are now interested to determine if the Bayesian approach on logistic regression can outperform the standard method. Since exact inference of the posterior distribution over the parameters is intractable, we use the Laplace approximation as introduced

in Section 4.4³. The number of parameter sets sampled from the posterior distribution was chosen to be 100. Comparative results for classifying the heart rate intervals are given in Table 5.5, while for classifying the diastolic blood pressure in Table 5.6. The performance improvement of the Bayesian approach is minimal. In fact, we don't give the ROC curves because they nearly overlap. In this case, we favor the simpler non-Bayesian logistic regression.

5.1.2 Exploring other features

The variety of choice for tackling the problem of feature extraction is definitely unlimited. This section discusses two ideas considered to be particularly relevant for our task: using the LDS parameters as features and using information such as gestation and post-natal age as features. The experimental results shown in the following are limited to ML logistic regression. The primary reason is that Naïve Bayes and decision trees have offered poorer performance in the previous section, especially on our IRC criterion.

In order to justify the first idea, an explanation of what we mean by LDS parameters is needed. Calibrating the FSLDS denotes learning the parameters associated to the normal regime. In terms of our model, the normal regime is an LDS obtained by conditioning the FSLDS on its switch variable set to the value corresponding to normality. Calibration is precisely the process of determining the parameters of this LDS. Since getting these parameters characterizing normality is the ultimate goal of our project, the idea is to use them as features. In other words, feature extraction from a fifteen minute section on each channel means training an appropriate autoregressive process followed by EM updates.

We now apply this idea for heart rate classification. Heart rate measurements are modeled as the sum of a slowly fluctuating baseline and a high frequency signal. Examining the state-space representation given by Equations 3.12 and 3.13, we can build the following feature vector:

$$\left(\alpha_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2 \quad \eta_1 + \eta_2 \quad \eta_1 \quad r \right) \quad (5.1)$$

where r is the observation noise. However, there is problem with applying logistic regression with this feature set. Since the relation $\beta_2 = \text{constant} - \beta_1$ holds for all intervals and all babies (see [24, V.C]), after standardizing the data we will get two linearly dependant features. This dependence causes trouble when matrix inversion is performed by the iterative reweighted least squares (IRLS) algorithm employed for finding the ML solution of logistic regression [3, §4.3.3]. Our solution was to remove β_2 from the feature set.

³Again, we have used Netlab's [20, 21] implementation

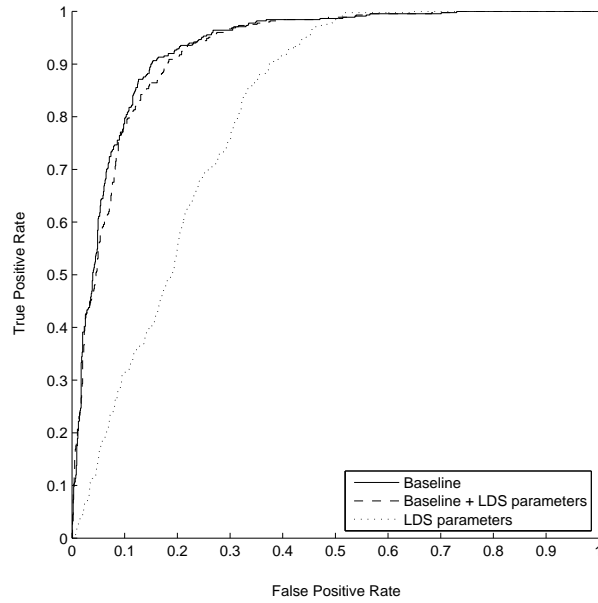


Figure 5.4: Classifying HR - ROC curves for three feature sets applied to ML logistic regression.

Table 5.7: Classifying HR - Performance comparison of three sets of features applied to ML logistic regression in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.

	Baseline	Baseline + LDS params	LDS params
AUC	0.9317	0.9256	0.8049
baby 1 (24 positives)	4	5	0
baby 2 (49 positives)	25	22	2
baby 3 (28 positives)	4	5	2
baby 4 (16 positives)	5	2	2
baby 5 (40 positives)	5	4	0
baby 6 (33 positives)	14	14	4
baby 7 (45 positives)	28	23	0
baby 8 (11 positives)	3	1	0
baby 9 (37 positives)	5	4	0
baby 10 (35 positives)	2	0	1
baby 11 (43 positives)	1	0	3
baby 12 (14 positives)	8	9	6
baby 13 (21 positives)	3	3	0
baby 14 (31 positives)	19	16	3
baby 15 (23 positives)	5	5	4

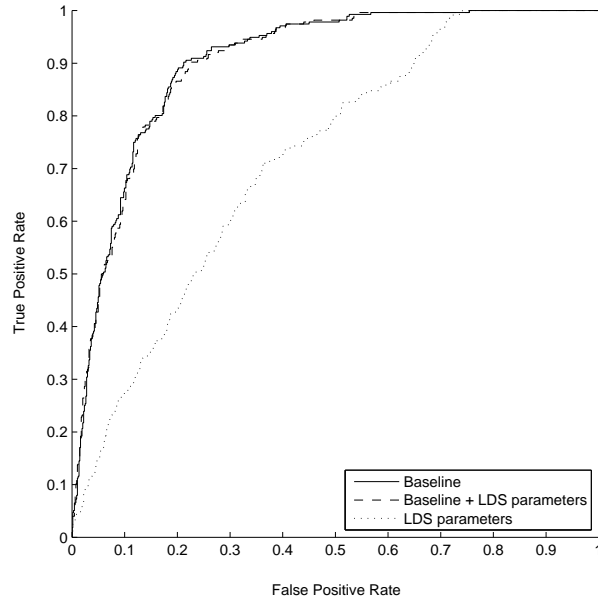


Figure 5.5: Classifying BD - ROC curves for three feature sets applied to ML logistic regression.

Table 5.8: Classifying BD - Performance comparison of three sets of features applied to ML logistic regression in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets. The “-” mark signals a missing channel (i.e. no classification to be made).

	Baseline	Baseline + LDS params	LDS params
AUC	0.9020	0.8999	0.7193
baby 1 (29 positives)	4	4	0
baby 2 (- positives)	-	-	-
baby 3 (- positives)	-	-	-
baby 4 (8 positives)	1	1	0
baby 5 (41 positives)	20	10	4
baby 6 (- positives)	-	-	-
baby 7 (- positives)	-	-	-
baby 8 (15 positives)	5	1	0
baby 9 (39 positives)	4	3	7
baby 10 (36 positives)	6	6	0
baby 11 (- positives)	-	-	-
baby 12 (14 positives)	9	9	0
baby 13 (24 positives)	3	2	0
baby 14 (41 positives)	9	6	1
baby 15 (25 positives)	1	0	1

Table 5.9: Classifying HR - Enhancing the feature set with gestation and post-natal age; comparison of ML logistic regression performance in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.

	Baseline	Baseline + Gestation + Post-natal age
AUC	0.9317	0.9111
baby 1 (24 positives)	4	5
baby 2 (49 positives)	25	13
baby 3 (28 positives)	4	4
baby 4 (16 positives)	5	6
baby 5 (40 positives)	5	0
baby 6 (33 positives)	14	13
baby 7 (45 positives)	28	24
baby 8 (11 positives)	3	3
baby 9 (37 positives)	5	7
baby 10 (35 positives)	2	2
baby 11 (43 positives)	1	1
baby 12 (14 positives)	8	9
baby 13 (21 positives)	3	3
baby 14 (31 positives)	19	14
baby 15 (23 positives)	5	5

Heart rate classification results are given in Figure 5.4 and Table 5.7 for logistic regression. Since the same baseline and signal model is used for the diastolic blood pressure, an analogous experiment on this channel has produced the results shown in Figure 5.5 and Table 5.8. The general conclusion is that the LDS parameters are not as good features as the ones in the baseline set for the task of discriminating between the Normal and Non-Normal intervals. In fact, when we have used the LDS features as our only features the performance was clearly worse.

Due to the nature of ML parameter optimization, adding more features into a logistic regression classifier should result into an increased likelihood for the training data. However, there are also issues of over-fitting which might cause inferior performance on a separate test set. This is exactly what happened when we have tried to add the LDS parameters to our baseline set of features (e.g. in the case of HR classification for the baseline set we have a log likelihood $L = -2720.2$, but adding the LDS parameters we get $L = -2693.5$ and a poorer classification; same for BD with values of $L = -1186.7$ and $L = -1154.7$ respectively).

Table 5.10: Classifying Incu. Air Humidity - Enhancing the feature set with gestation and post-natal age; comparison of ML logistic regression performance in terms of: (i) AUC computed for all babies and (ii) IRC computed on a per baby basis; the number of positive instances for each baby is given between brackets.

	Baseline	Baseline + Gestation + Post-natal age
AUC	0.8803	0.8791
baby 1 (29 positives)	2	3
baby 2 (55 positives)	9	10
baby 3 (43 positives)	4	5
baby 4 (19 positives)	3	3
baby 5 (42 positives)	17	16
baby 6 (56 positives)	8	8
baby 7 (46 positives)	1	0
baby 8 (15 positives)	2	0
baby 9 (40 positives)	4	5
baby 10 (36 positives)	8	8
baby 11 (- positives)	-	-
baby 12 (30 positives)	10	9
baby 13 (23 positives)	1	1
baby 14 (41 positives)	19	15
baby 15 (- positives)	-	-

The second idea was to incorporate other existing information about the baby into our classifiers. Gestation and post-natal age (i.e. days of life) are available for thirteen of the fifteen babies (missing for babies number 13 and 15). We remain faithful to the approach of using as much information as possible anytime when we try to predict the probability of an interval being normal. Comparative performance results are given in Table 5.9 for heart rate classification and in Table 5.10 for classifying the incubator humidity channel. As expected, the results are very similar in both cases. Consequently, we don't show the ROC curves because they severely overlap. In this case, it is useful to look at the weights associated by the logistic regression method to gestation and post-natal age. Comparing those to the ones of the attributes in the baseline set, we saw they have never been amongst the relevant ones (i.e the ones having weights with large absolute values), regardless of the classified channel or the analysed baby.

In the end of the section, we emphasize the fact that we do not need to fix a threshold in order to use any of the classifiers above in practice. This is because we are not directly interested in an excellent classification between Normal and Non-Normal instances on all channels. Since we just need some Normal intervals for each observed channel, we simply sort the probabilities for all the intervals corresponding to a baby and pick the top k predictions.

Based on the experiments presented so far, we can safely claim that the baseline set of features we have extracted is an appropriate choice for our classification task. In addition, logistic regression delivers a good performance on this feature set, especially on our tailored criterion, IRC. Concluding, we pick the classifier consisting of the baseline set of features and logistic regression as our choice for the tests done on the auto-calibrated model in the following section.

5.2 Evaluating the inferences produced by the auto-calibrated FSLDS

This section is dedicated to a comparative analysis between the inferences produced by the manually-calibrated FSLDS and its auto-calibrated version⁴. As previously explained, we set up a FSLDS able to infer the posterior probability distribution for four hidden factors: Incubator Open, Bradycardia, Temperature Probe Disconnection and Blood Sample.

The quality of the inferences will be assessed by the same two criteria as in [22, 24]. The first one is the AUC already introduced in the previous section and the second one is the equal error rate (EER). The EER is the error rate computed for the threshold value at which the false positive rate (FPR) is equal to the false negative rate (FNR)⁵. A graphical interpretation of the EER would be that it is the error rate corresponding to the intersection of the ROC

⁴All the experiments in this section made use of John Quinn's code for the FSLDS [23]. This has also implied employing the Bayes Net Toolbox for MATLAB [19].

⁵ $FNR = 1 - TPR$, where TPR is the true positive rate

curve with the diagonal connecting the upper-left and lower-right corners of the ROC plot. The EER is in fact proportional to the distance between this intersection point and the upper-left corner by a factor of $\sqrt{2}$. Analysing the EER is particularly useful when the number of instances in the two classes is unbalanced. Keeping in mind that the statistics are computed for a total of 360 hours of monitoring data and considering the total duration for the four factors of interest (Table 4.1), we immediately see that computing the EER is highly recommended in our problem. In addition, computing the EER can be regarded as a principled way to select a good threshold from an ROC curve. Since the EER is an error, the smaller the value the better.

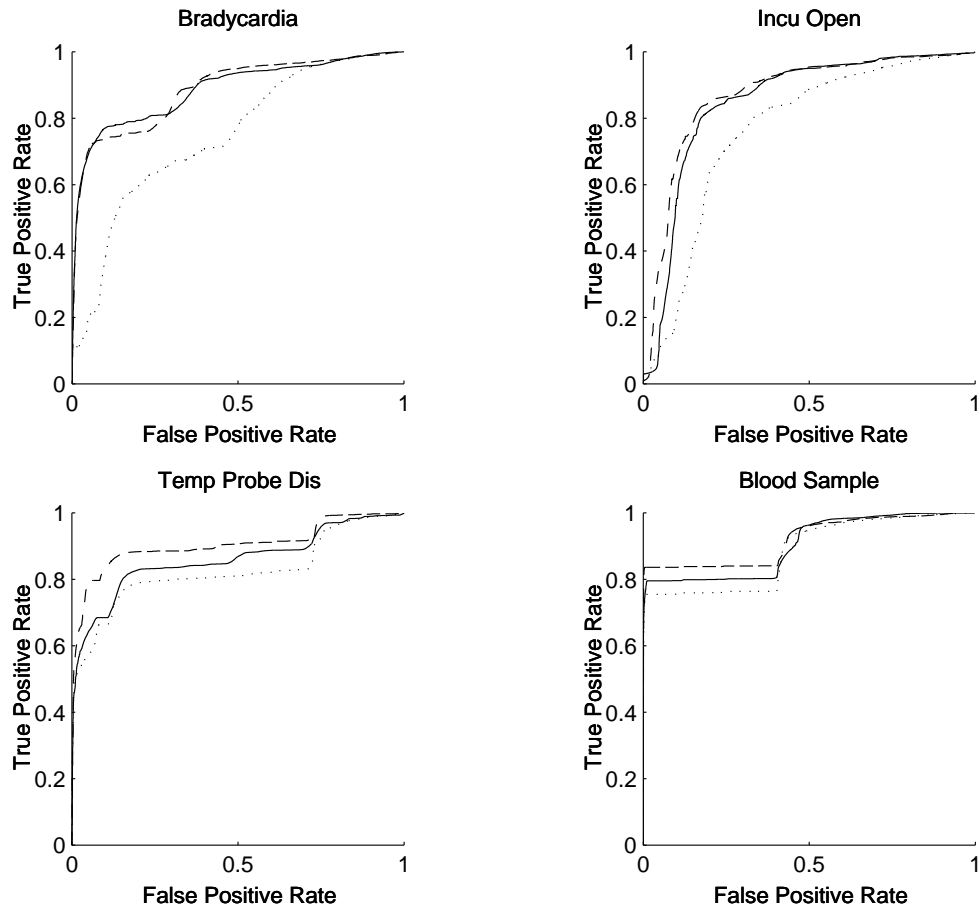


Figure 5.6: ROC curves showing the classification of the four factors of interest for three methods of calibration

For evaluation, we use the same setting as the one described in [24] and all the 360 hours of physiological monitoring data on hand. The experiment is done with three-fold cross-validation: ten babies are used for training and the remaining five for testing. For the moment, the calibration system selects the only top prediction outputted by the classifier (i.e. the instance with the highest posterior probability) and uses this interval to learn the normal dynamics. As argued in Section 3.2.3 the inference method employed is the Gaussian Sum (GS)

Table 5.11: Summary statistics for the three methods of calibration

Calibration	Statistic	Bradycardia	Incu Open	Temp Probe Dis	Blood Sample
Manual	AUC	0.89	0.87	0.90	0.92
	EER	0.24	0.17	0.13	0.16
Auto	AUC	0.89	0.85	0.86	0.91
	EER	0.21	0.18	0.18	0.20
Random	AUC	0.75	0.76	0.82	0.88
	EER	0.33	0.27	0.22	0.24

approximation. A single run processing all the data for the fifteen babies took 7.5 hours to complete.

In Figure 5.6, we plot the ROC curves corresponding to the four inferred factors and for three methods of doing calibration: Auto, Manual⁶ and Radom. The last one will be explained below. The Values of the AUC and EER are organized in Table 5.11. The quality of the inferences produced by the auto-calibrated FSLDS is very close to the one of the inferences produced by the manually calibrated version for three of the factors: Incubator Open, Temperature Probe Disconnection and Blood Sample. Luckily, for the remaining factor, Bradycardia, the AUC values are identical in both cases. Moreover, for this factor the auto-calibrated FSLDS manages to outperform the manual version in terms of EER. At the same time, the largest difference between the quality of manual and automatic inferences is on Temperature Probe Disconnection. We will return to this issue below.

Until now we have seen that the auto-calibrated system offers a comparable performance to the manually-calibrated one. The question to be asked is what has been gained by employing a classifier that predicts normality with respect to a random calibration procedure. This implies sampling some arbitrary interval from the analysed baby and using it to learn the normal dynamics.

In order to perform this experiment a number of design choices need to be considered. We still kept the fifteen minute interval length, but added the constraint that at least one known factor is present in any randomly selected interval. Because the average event durations are clearly shorter than fifteen minutes and considering the effects of "chopping" the data our constraint is reasonable. On the other hand, intervals containing only dropout measurements on any channel have been ignored. Furthermore, the temperature and humidity channels often display constant values for long periods of time. The primary cause is the quantization performed by the mea-

⁶The results obtained with the manually-calibrated FSLDS are not identical to the ones in [24] due to using an updated version of both code and data annotations.

Table 5.12: Summary statistics for manual- and auto-calibration when jitter has been added

Calibration	Statistic	Incu Open	Temp Probe Dis
Manual	AUC	0.87	0.90
	EER	0.17	0.13
Manual + jitter	AUC	0.87	0.91
	EER	0.17	0.13
Auto	AUC	0.85	0.86
	EER	0.18	0.18
Auto + jitter	AUC	0.86	0.84
	EER	0.18	0.20

surement devices. Since we don't want to train the AR processes on constant intervals, we have added jitter on these channels (Gaussian noise - zero mean, 0.02 variance due to a quantization step $q = 0.1$). In addition, if the trained AR coefficients did not satisfy the convergence criteria (e.g. $|\alpha_1| < 1$ for an $AR(1)$ process, see [4, §3.4.4]), then the sample is rejected. If not, our experiments have shown that the inferred covariance of the hidden state ($\mathbf{V}_{t|t}^j$ in Equation 3.13) diverges. A number of 10 different sets of samples, one extracted from each baby, have been averaged to draw the "Random" curve in Figure 5.6.

Our results show that using a classifier predicting normality is clearly superior to randomly picking an interval when we do inferences for the Bradycardia and Incubator Open factors. This conclusion is valid for the manual calibration as well. However, for the two remaining factors, while still noticeable, the improvement is less significant. A possible explanation is that the random intervals are picked from the current test baby, so they still exhibit a certain amount of normality.

An interesting observation is that even in both the annotated and the predicted sections of normality there are some constant intervals on the temperature or humidity channels. We also added jitter to these intervals and ran inference again. The results are presented in Table 5.12. Only the known factors that can affect the temperature and humidity measurements have been included. From our results, clear benefits of this procedure are hardly noticeable. Probably, a more elaborate approach on reducing the effects of quantization on these channels would be conclusive.

So far, we have been using only the top prediction of our classifier for calibrating the FSLDS. As an alternative we have experimented picking the top three predictions. From the previous section we already know that among the top three predictions there might be some false positives. At the same time, the increased quantity of data can help towards a better learning

Table 5.13: Summary statistics comparing two auto-calibration methods: the standard “Auto” picking only the top prediction and “Auto-3” picking the top three predictions.

Calibration	Statistic	Bradycardia	Incu Open	Temp Probe Dis	Blood Sample
Auto	AUC	0.89	0.85	0.86	0.91
	EER	0.21	0.18	0.18	0.20
Auto-3	AUC	0.88	0.80	0.85	0.92
	EER	0.23	0.25	0.20	0.17

of the normal dynamics. Naming this procedure Auto-3, we show the summary statistics in Table 5.13. On average, the performance is lower compared to the standard auto-calibration procedure, but clearly better than the random approach.

In the final part of this chapter we illustrate some comparative examples of inferences done with the manually- and auto-calibrated FSLDS’s for physiological condition monitoring⁷. The horizontal bars in the lower part of the figures indicate the posterior distributions of factors. Levels of grey from white to black indicate values from zero to one respectively. In a first experiment (Table 5.14, Figure 5.7 and Figure 5.8), we see that our two systems perform equally well at inferring Bradycardia and Blood Sample. The second experiment (Table 5.15, Figure 5.9 and Figure 5.10) shows how the manual and auto FSLDSs function when the factors overlap. Here, it is interesting to note that both systems not only detect events with similar accuracy but also seem to make the same mistakes. In the last one (Table 5.16, Figure 5.11 and Figure 5.12), we also introduce the X-factor, the factor corresponding to the Abnormal class (see §3.2.2). The quality of the predictions for Blood Sample is identical for both calibration methods, while the automatic system tends to produce more false alarms when inferring the X-factor.

⁷The visualization code in [23] has been used.

Table 5.14: Experiment 1: MANUAL vs. AUTO calibration - Summary of the inferred distributions for Blood Sample and Bradycardia. The calibration methods perform equally well. However, they both flag a possible bradycardia instance around time 125, this being in disagreement with the annotator's opinion.

Statistic	Bradycardia		Blood Sample	
	AUC	EER	AUC	EER
Manual	0.885	0.260	0.970	0.05
Auto	0.905	0.143	0.968	0.05

Table 5.15: Experiment 2: MANUAL vs. AUTO calibration - Summary of the inferred distributions for Incubator Open (i.e. Handling), Blood Sample and Bradycardia. The particularity of this experiment is factor overlapping, but the calibration methods deliver satisfactory performance. However, they both consider two separate handling annotations as a single interval.

Statistic	Bradycardia		Blood Sample		Blood Sample	
	AUC	EER	AUC	EER	AUC	EER
Manual	0.997	0.031	0.996	0.026	0.907	0.147
Auto	0.998	0.031	0.995	0.026	0.922	0.145

Table 5.16: Experiment 3: MANUAL vs. AUTO calibration - Summary of the inferred distributions for Blood Sample and the X-factor (i.e. the factor for the Abnormal Class). The quality of the predictions for Blood Sample is identical, but the manual version performs better for the X-factor.

Statistic	X-factor		Blood Sample	
	AUC	EER	AUC	EER
Manual	0.839	0.176	0.998	0.016
Auto	0.792	0.307	0.998	0.016

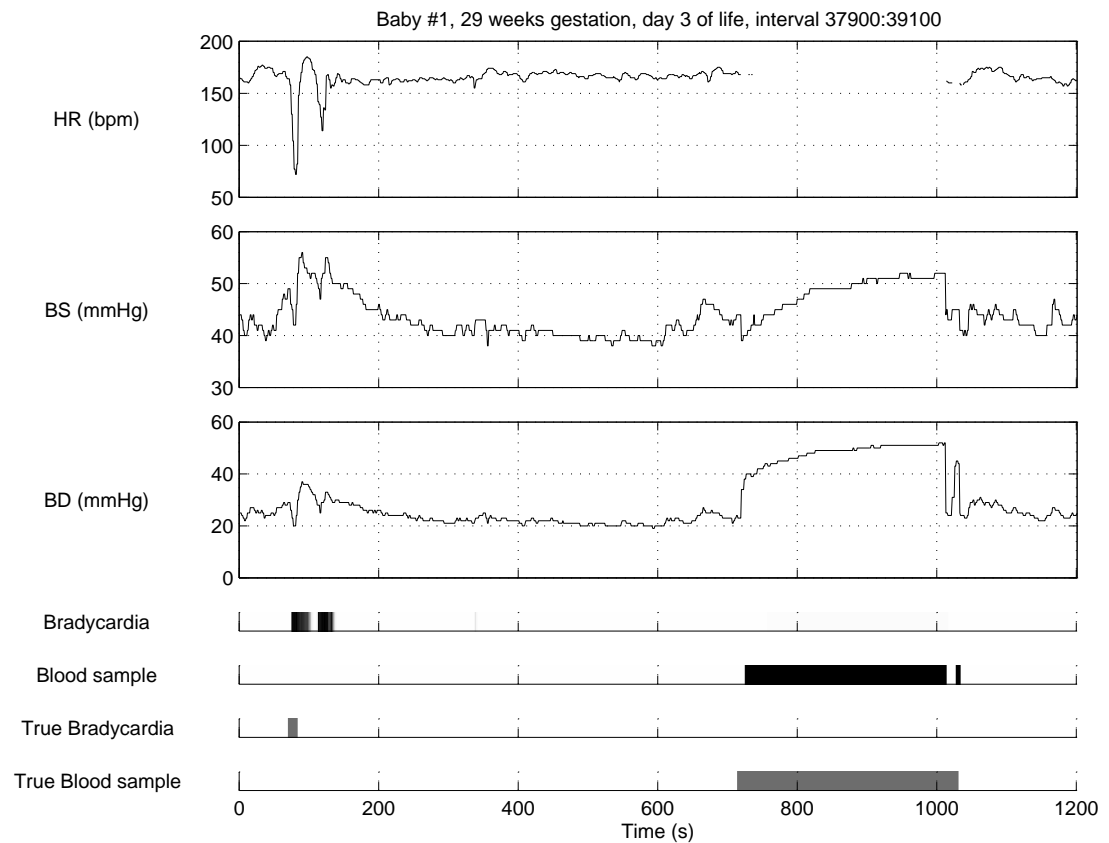


Figure 5.7: Experiment 1: MANUAL calibration - Inferred distributions for Blood Sample and Bradycardia. Inference on both factors is correct. However, an inferred bradycardia instance around time 125 is in disagreement with the annotator's opinion.

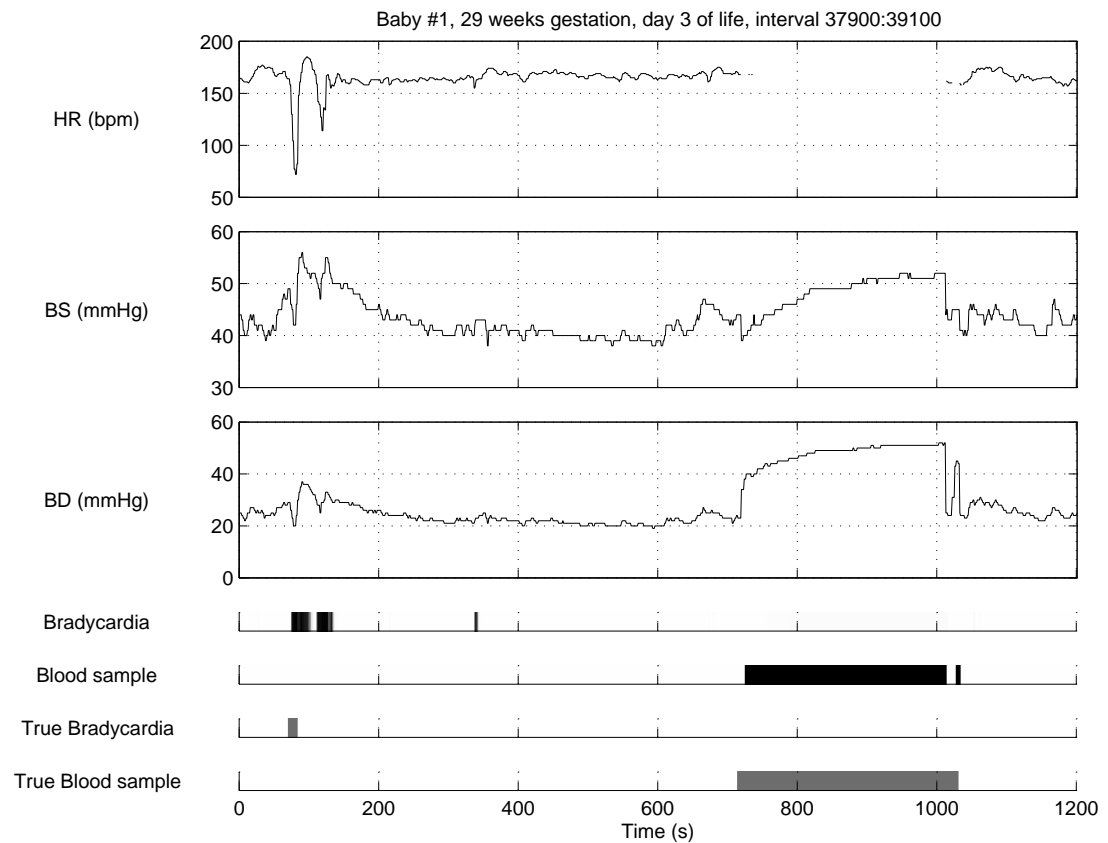


Figure 5.8: Experiment 1: AUTO calibration - Inferred distributions for Blood Sample and Bradycardia. Inference on both factors is correct. However, an inferred bradycardia instance around time 125 is in disagreement with the annotator's opinion.

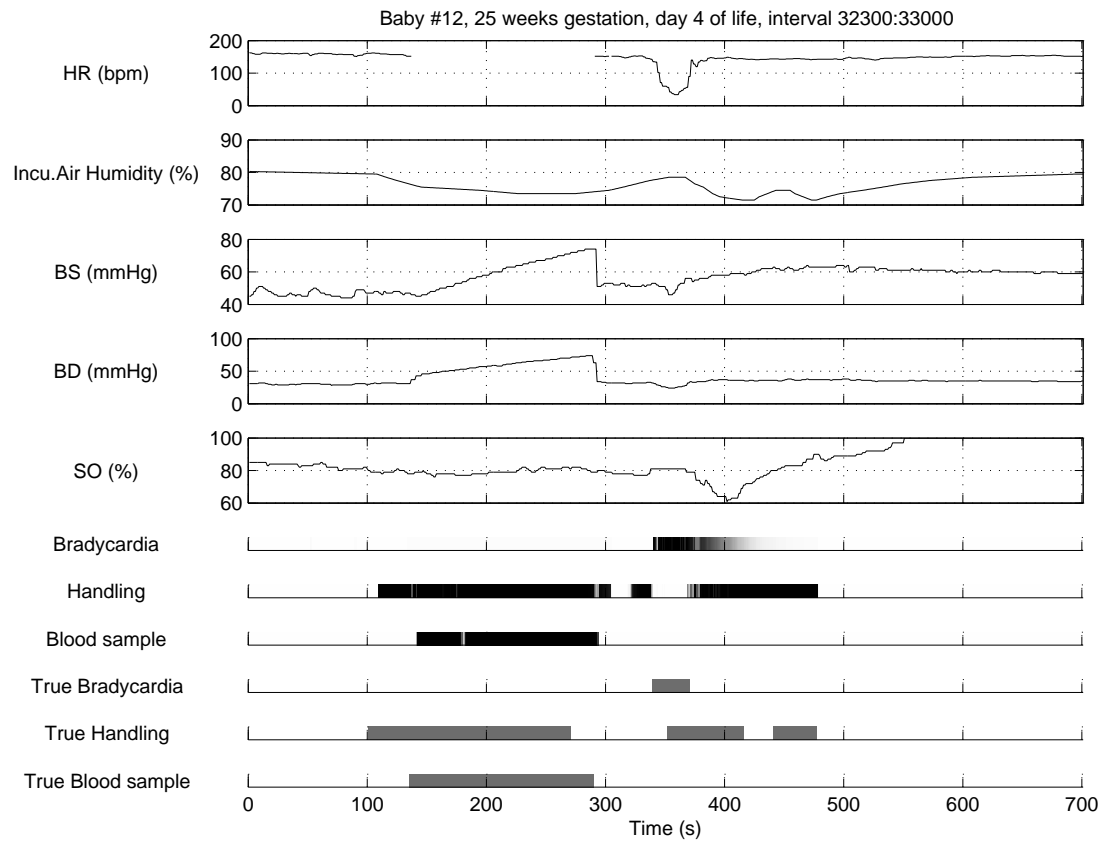


Figure 5.9: Experiment 2: MANUAL calibration - Inferred distributions for Incubator Open (i.e. Handling), Blood Sample and Bradycardia. The predictions are generally correct. Nevertheless, there are two exceptions: two separate handling instances are not discriminated and bradycardia is still predicted long after the event has ended.

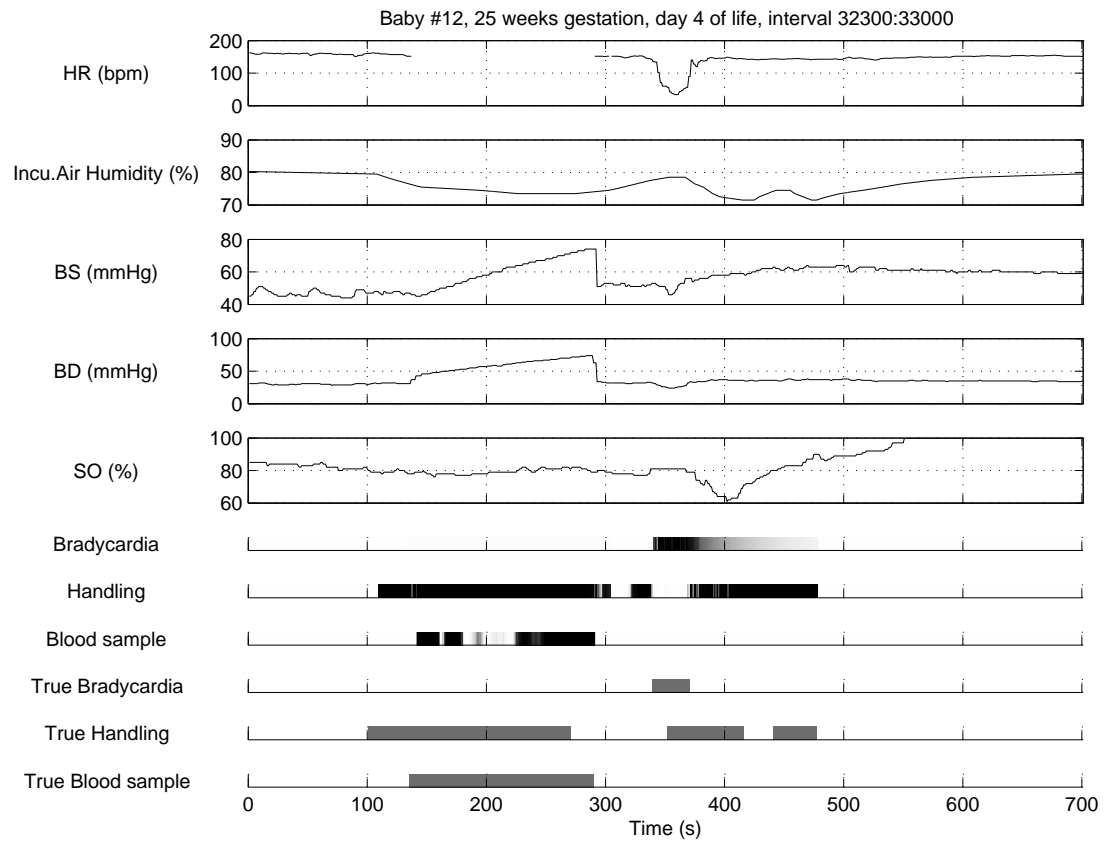


Figure 5.10: Experiment 2: AUTO calibration - Inferred distributions for Incubator Open (i.e. Handling), Blood Sample and Bradycardia. The predictions are generally correct. Nevertheless, there are two exceptions: two separate handling instances are not discriminated and bradycardia is still predicted long after the event has ended.

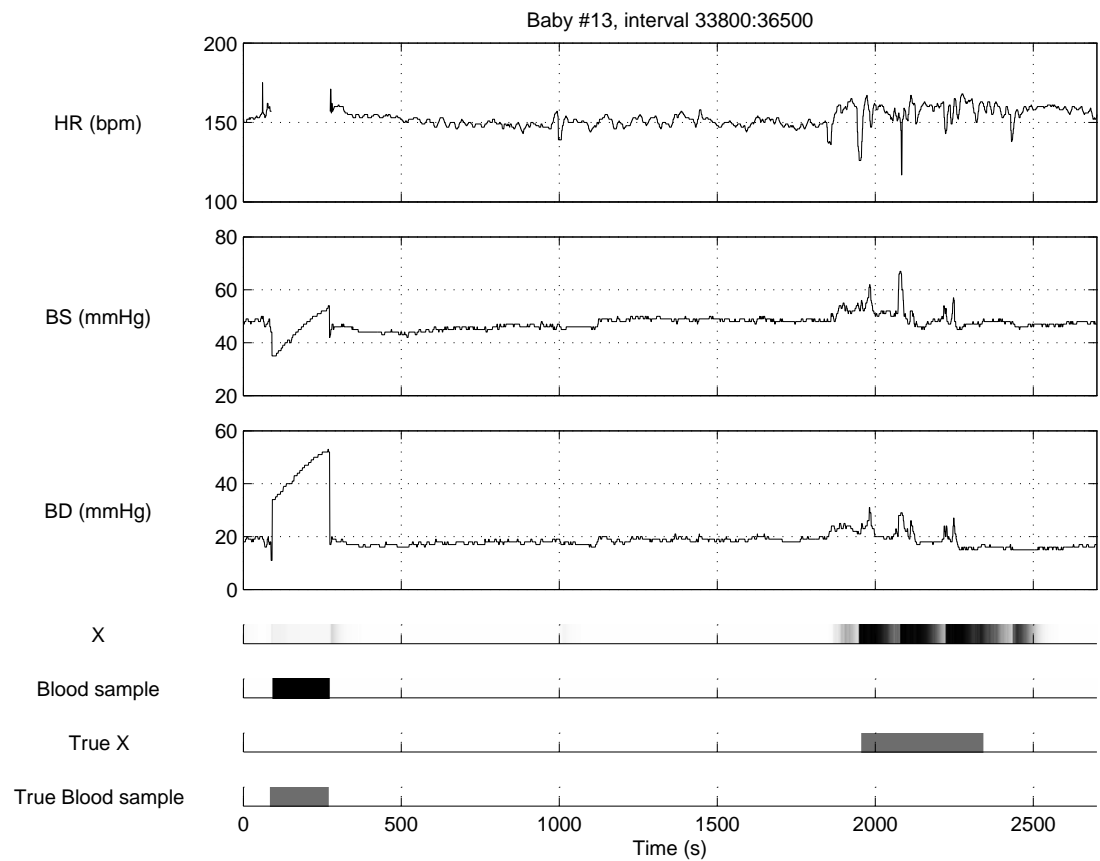


Figure 5.11: Experiment 3: MANUAL calibration - Inferred distributions for Blood Sample and the X-factor (i.e. the factor for the Abnormal Class). Both factors are correctly inferred. The X-factor displays high sensitivity to any deviation from normality.

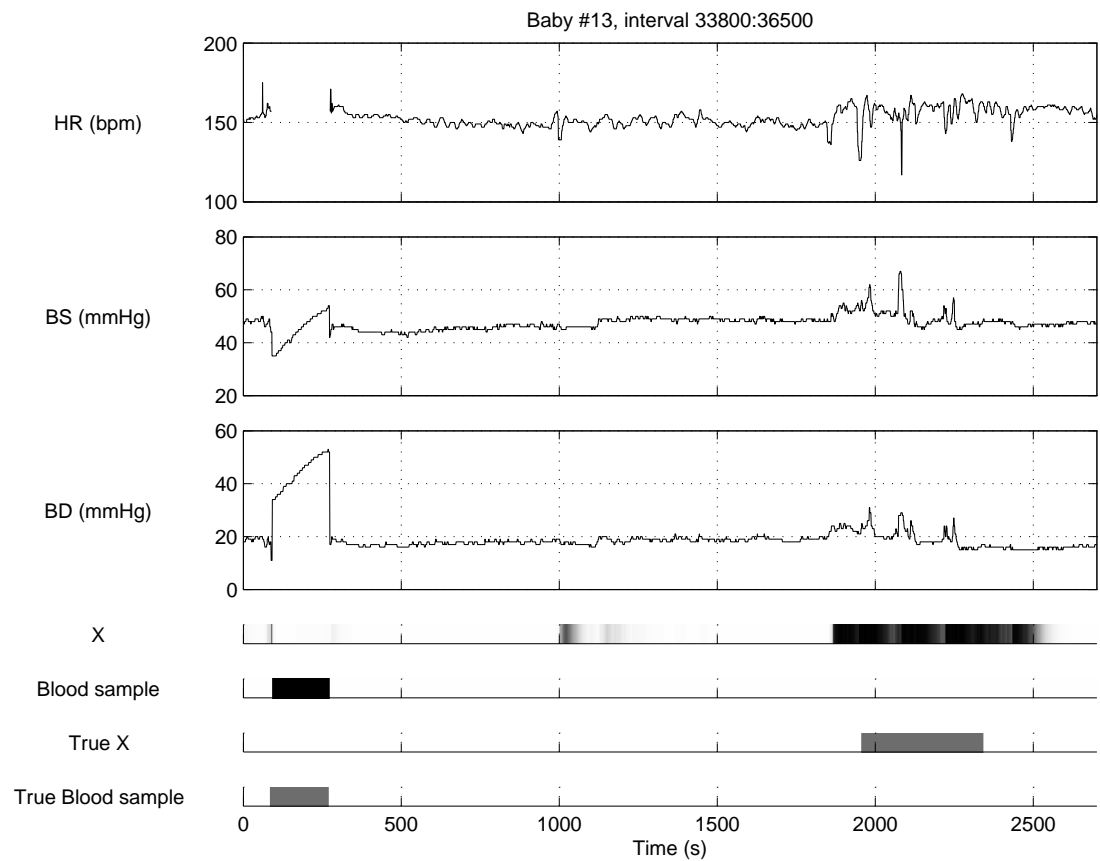


Figure 5.12: Experiment 3: AUTO calibration - Inferred distributions for Blood Sample and the X-factor (i.e. the factor for the Abnormal Class). Both factors are correctly inferred. In addition, the X-factor displays high sensitivity to any deviation from normality.

Chapter 6

Conclusions and Future Work

The current project has described our approach for automating the calibration stage of the FSLDS for neonatal condition monitoring. Section 6.1 reviews the key aspects of our solution and highlights the most important results. Based on those findings, we are able to formulate some interesting recommendations for extending our work in Section 6.2.

6.1 Conclusions

The main objective of this project was to reduce or even eliminate the need for a manual calibration stage in the FSLDS for condition monitoring. Our work was divided into two stages. In a first stage we have built a probabilistic method for extracting sections of normality from physiological monitoring data. The second stage consisted of assessing the performance of the auto-calibrated system with respect to its manually-calibrated version.

We chose to extract intervals of normality by employing a binary classifier able to discriminate between Normal and Non-Normal Sections. Exploratory data analysis (Section 4.1) has revealed some of the main challenges of our task: normality being specific to each baby, known events displaying variability in patterns, variability in occurring frequencies and durations, missing data and uncertainty in the “factor - influenced channel” relationship.

Using summary statistics and knowledge engineering, we articulated our choices for designing the classifiers. First, we found an appropriate length for the sections to be classified (Section 4.2.1). Then we have justified our choice for a channel-based approach of the task. Accordingly, a more relevant labeling for the data was introduced (Section 4.2.3). Afterwards, we have proposed feature extraction solutions (Section 4.3). Our experiments have shown that a baseline set of simple features (e.g. means, standard deviations or median values) carefully

chosen relying on medical considerations is a proper choice for the problem. With all this in place, a judicious classifier setup is explained (Section 4.4). The performance of the employed “off-the-shelf” classifiers (ML and Bayesian logistic regression, Naïve Bayes and decision tree) has been assessed in a cross-validation setting on two criteria, including one especially designed for the task. Analysing the experimental results (Section 5.1), we have concluded that the combination of the baseline set of features and logistic regression is the best choice for the auto-calibration problem.

The most important test in this project was running inference in the auto-calibrated FSLDS for the whole amount of physiological data (Section 5.2). An immediate conclusion is that the auto-calibrated FSLDS is clearly successful in uncovering the distribution of all four analysed factors: Brdaycardia, Incubator Open, Temperature Probe Disconnection and Blood Sample. In a direct comparison with its manual version, our system delivers a marginally better performance at inferring Bradycardia and is almost as good on the other three factors (Figure 5.6 and Table 5.11). From another point of view, our experiments have shown that employing the classifier for normality is definitely superior to randomly choosing a data section from the analysed baby for calibration. In addition, we provide evidence that randomly selected sections can produce diverging estimates for hidden state variances. Other ideas we have explored are: jittering channels displaying long periods of constant measurements and selecting more than one interval for calibration. Plots of inferred distributions obtained by the auto and manually calibrated FSLDS’s are given. The observation is that both systems not only detect events in a similar fashion, but also tend to make the same errors.

6.2 Future Work

This section enumerates some interesting extensions for our work. The focus is on ideas for improving the extraction of normality sections from the physiological monitoring data, but some recommendations refer to the study of inferences produced by the FSLDS.

In the current project the same dataset has been used for both training the classifier for Normality and learning the FSLDS. A more rigorous test can be performed if the classifier can be trained on a separate dataset of the same kind. Moreover, this might allow holding out a validation set, so more sophisticated classifiers can be utilized. At the same time, more experiments with Bayesian versions of the classifiers can be done. Our results in Section 5.1.1 also suggest combining the outputs of different classifiers (i.e. ensemble methods).

We have chosen to select fifteen minute non-overlapping intervals as instances to be tested for normality. However, many different approaches might be taken. For instance, normality may

be classified on a second-by-second basis. The idea is to use a window centered on the current time step to select a much longer section from which the features will be extracted. Modifying our labeling and feature extraction methods might also be required. In addition, the shape of the window can be used to weight the importance of the observations at various distances from the current time step.

Another idea is a thorough study of the missing data problem. We have seen that measurements on various channels are missing, because clinicians usually select for monitoring only the channels they consider most informative about a certain patient's condition. One might attempt to estimate the values on the missing channels by modeling their probability distribution conditioned on the data on hand (see [15] for details).

The comparison between the manually- and auto-calibrated FSLDSs has been performed based on second-by-second inferences. However, it is also practical to analyse only if an event has been detected or not. Results of such an approach are already available for the manually-calibrated system [22], but not for its automated version. In addition, both the FSLDS and the classifier built in this project can easily be adapted to include more factors (e.g. transcutaneous probe recalibration, sepsis).

Bibliography

- [1] Daniel L. Alspach and Harold W. Sorenson. Nonlinear Bayesian Estimation Using Gaussian Sum Approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972.
- [2] David Barber and Bertrand Mesot. A Novel Gaussian Sum Smoother for Approximate Inference in Switching Linear Dynamical Systems. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 89–96. MIT Press, Cambridge, MA, 2007.
- [3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.
- [4] Chris Chatfield. *The Analysis of Time Series – An Introduction*. Chapman & Hall/Crc, 6th edition, 2004.
- [5] G Ewing, L Ferguson, Y Freer, J Hunter, and N McIntosh. Observational Data Acquired on a Neonatal Intensive Care Unit. Technical report, University of Aberdeen, 2002. Computing Science Departmental Technical Report: TR 0205.
- [6] Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, 2004.
- [7] Zoubin Ghahramani and Geoffrey E. Hinton. Parameter Estimation for Linear Dynamical Systems. Technical report, University of Toronto, 1996.
- [8] Zoubin Ghahramani and Geoffrey E. Hinton. Variational Learning for Switching State-Space Models. *Neural Computation*, 12(4):831–864, 2000.
- [9] Zoubin Ghahramani and Michael I. Jordan. Factorial Hidden Markov Models. *Machine Learning*, 29:245–273, 1997.
- [10] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

- [11] David Hand, Heikki Mannila, and Padharic Smyth. *Principles of Data Mining*. MIT Press, 2001.
- [12] C-J. Kim. Dynamic Linear Models with Markov-Switching. *J.Econometrics*, 60:1–22, 1994.
- [13] S. L. Lauritzen. *Graphical Models*. Oxford Univeristy Press, 1996.
- [14] Uri Lerner and Ronald Parr. Inference in hybrid networks: Theoretical limits and practical algorithms. In *UAI*, pages 310–318, 2001.
- [15] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. New York, Wiley, 1987.
- [16] David J. C. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.
- [17] Kevin Murphy and S Russell. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In N. de Freitas A. Doucet and N. Gordon, editors, *Sequential Monte Carlo in Practice*. Springer-Verlag, 2001.
- [18] Kevin P. Murphy. Switching Kalman Filters. Technical report, U.C. Berkeley, 1998.
- [19] Kevin P. Murphy. The Bayes Net Toolbox for MATLAB. *Computing Science and Statistics*, 33:2001, 2001.
- [20] Ian T. Nabney. *Netlab: Algorithms for Pattern Recognition*. Springer, 2001.
- [21] Ian T. Nabney. Netlab Neural Network Software, 2004. <http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/>.
- [22] John Quinn. *Bayesian Condition Monitoring in Neonatal Intensive Care*. PhD thesis, University of Edinburgh, 2007. <http://hdl.handle.net/1842/2144>.
- [23] John Quinn. Neonatal condition monitoring demonstration code, 2008. <http://omnipresence.org/jq/software.html>.
- [24] John A. Quinn, Christopher K. I. Williams, and Neil McIntosh. Factorial Switching Linear Dynamical Systems Applied to Physiological Condition Monitoring. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(9):1537–1551, 2009.
- [25] Lawrence Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.
- [26] R. Shumway and D. Stoffer. Dynamic linear models with switching. *J. of the American Statistical Association*, 86:763–769, 1991.

- [27] Christopher K. I. Williams, John A. Quinn, and Neil McIntosh. Factorial Switching Kalman Filters for Condition Monitoring in Neonatal Intensive Care. In Y Weiss, B Schölkopf, and J Platt, editors, *Advances in Neural Information Processing Systems 18*. MIT Press, 2006.