Stony Brook University

Bayesian methods for feature extraction and classification of fetal heart rate signals

A Dissertation Presented
by
Shishir Dash

to
The Graduate School
in Partial Fulfillment of the
Requirements
for the Degree of
Doctor of Philosophy
in
Electrical Engineering
Stony Brook University

May 2014

Stony Brook University
The Graduate School

Shishir Dash

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Petar M. Djurić – Dissertation Advisor
Professor, Electrical and Computer Engineering

Mónica F. Bugallo – Chairperson of Defense
Assistant Professor, Electrical and Computer Engineering

Murali Subbarao
Professor, Electrical and Computer Engineering

J. Gerald Quirk
Professor, Obstetrics, Gynecology and Reproductive Medicine,
Stony Brook University Medical Center.

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Bayesian methods for feature extraction and classification of fetal heart rate signals

by

Shishir Dash

Doctor of Philosophy

in

Electrical Engineering

Stony Brook University

2014

A central problem in biomedical signal processing research is that of computer aided classification. We consider the problem of computer-aided diagnosis of intrapartum fetal status based on simultaneously recorded fetal heart rate (FHR) and uterine pressure (UP) signals. Clinically, visual diagnosis of the fetal heart rate signal is of critical importance when evaluating the status of pregnancy and delivery. This is because oxygen inadequacy, a major cause of adverse fetal outcomes, has a direct effect on the fetal heart rate. Purely visual assessment of fetal heart rate segments has, however, proven to have high intra- and inter observer variability, which has persisted despite the publication of standardized interpretation guidelines, such as those by the National Institute of Child Health and Human Development (NICHD). This has led to an alarming increase in the rate of caesarian sections and unnecessary litigation expenses in even simple cases. Thus, the development of automated algorithms for accurate classification of FHR-UP patterns is of paramount importance.

The main contribution of our research is the development of different Bayesian classification approaches that utilize two distinct paradigms for feature extraction: (a) summarizing patterns from long-duration data sets and (b) using sequences of features derived from short data-lengths. In most of the existing methods, feature-vectors for long-duration (10-20

minutes or more) datasets are mapped to scalar values. This approach ignores the inherent non-stationarity in, and the effects of short-term interactions between the two time series. In order to account for these factors, we develop methods to extract informative features from short time-series data, and then use sequences of such features as classifier input. We used these feature inputs in conjunction with classification methods based on density estimation using window-counting, Bayesian-network structure detection and generative mixture models. In particular, Bayesian network structure detection enables the discovery of novel correlations and causations amongst different features Generative mixture models turn out to be ideally suited for the modeling and classification of the feature-sequences described earlier, and to elegantly fuse information from both FHR and UP patterns. We explore a variety of features derived from expert-consensus guidelines and statistical metrics that quantify information about the series of beat-to-beat fractional changes in FHR. We also develop methods to accurately translate clinical guidelines for FHR categorization into algorithmic rules for decision-making. We describe the use of the NICHD guidelines to make such deterministic systems, which are compared to the aforementioned probabilistic classifiers.

The methods presented here have the potential to make accurate FHR-UP monitoring and automated decision support a possibility. We carried out rigorous performance evaluations of these techniques on several datasets acquired from real subjects. We show improvements in classification accuracy versus discriminative methods such as support vector machines and rule-based classifiers. Gold-standard labeling was done using manual physician interpretations of FHR-UP recordings or an objective fetal health metric (umbilical cord pH). This work also opens up several possibilities for future research, including unsupervised clustering of FHR-UP patterns using nonparametric Bayesian methods with generative models, development of fetal "risk scores" based on these discovered models, and the use of sampling techniques to automatically segment FHR-UP time-series data into distinct patterns.

Contents

List of Tables

List of Figures

Acknowledgements

First of all, I would like to thank my advisor Prof. Petar M. Djurić for his help and advice throughout my doctoral studies. His values, his extensive knowledge of and enthusiasm for research, his enjoyment of teaching and incredible curiosity for and knowledge of the world around him, have been an inspiration to me. I am especially appreciative of his belief in me whenever I felt discouraged by my lack of progress in work (and sometimes, in life as well).

I'm grateful to Dr. J. Gerald Quirk, who along with Prof. Djurić, introduced me to the fascinating field of fetal monitoring. His passion and support - both financial and otherwise - for my research, not to mention his patience with my endless questions on the clinical aspects, have been incredible. Without his ambition to make automated FHR classification a reality, his knowledge and desire to learn and his remarkable (and often hilarious) stories, our work would have been considerably diminished.

My thanks go to Prof. Mónica F. Bugallo, who I had the pleasure of working with on a research project on stem-cell evolution, in courses and during annual summer engineering camps, among others. Monica's intelligence, incredible work ethic, and people skills are something I have aspired to acquire. For her support and advice, I will always be grateful.

I would also like to thank Dr. Murali Subbarao for having taken the time to read, understand and offer invaluable advice about my research, and for agreeing to be on my defense committee. His experience, penetrating questions, and his patient answers to mine were invaluable.

My first advisor in graduate school was Prof. Ki Chon, who introduced

me to several exciting ways to think about biomedical signals. While working with him on biosignal analysis projects, I learnt the value of proper experimental design, working with real subjects' data, attention to detail and also the dirty work of "data-munging" (signal annotation, noise analysis, preprocessing etc) that is so important to any machine learning project and especially to biomedical data. I am grateful to him for passing down some of his knowledge and experience and for encouraging me.

I am thankful to Abhijit Patil and Seema Somani of GE Global Research in Bangalore, India, for a very exciting collaboration on acquisition of fetal heart rate signals via external non-invasive fetal ECG monitoring. Working with several patients in labor and with obstetric personnel gave me significant new knowledge on the intricacies of clinical monitoring. I am also grateful to them for providing part of the financial support for my own doctoral research.

Special thanks to Elizabeth Roemer, who was instrumental in researching opportunities for research grants and collaborations. Her experience and advice was invaluable to me during tough times. I also want to acknowledge the support from Darlene Swords and Rachel Ingrassia, without whose help in administrative affairs, I would have been lost.

I offer sincere thanks to my friends and colleagues in the lab - Çagla Taşdemir, Iñigo Urteaga, Yunlong Wang, Zhiyuan Weng, Li Geng, Zhe Shen, Jonathan Beaudeau, Isaac Manuel, and past members - for all the engaging discussions on research, soccer, food, cultures and more. They made my time in COSINE lab very memorable and rewarding. I'd like to also mention my sincere appreciation to friends outside of the lab (there are too many to list individually) for making life interesting and for always being supportive and honest.

Finally, I would like to thank my brother and my parents for always being my strongest pillars of support. Their love and encouragement have been the most important things in my life.

1 Introduction

Research in signal processing in the context of biomedical signals has a long and storied history. In fact, the term "signal processing" itself does not capture the richness of the field, which involves everything from monitoring and noise-free extraction of signals, feature extraction from the signals to capture informative deviations from "normality," pattern classification and machine learning to identify such areas as well as to do final diagnosis based on multiple signal features. Recently, Saeys *et al.* [90] have reviewed feature-selection techniques in the context of bioinformatics applications, while Lotte *et al.* [64] have described pattern-classification algorithms applied to brain-computer interfaces. Many other examples of such studies can be found in the current literature.

The primary goal of our research is to develop novel methods for more accurate and consistent classification of the fetal heart rate (FHR) and uterine pressure (UP) signals into categories that are meaningful to clinical implementation. I first explain the motivation behind the current research in Section 1.1. Section 1.2 describes the general physiological mechanisms behind the generation of these signals and also describes the pathophysiology associated with fetal distress. We provide a brief listing of the primary contributions of our research on FHR analysis in Section 1.6. Finally, a chapter outline for the rest of the dissertation is provided in Section 1.7.

1.1 Motivation

A 2005 report by Law *et al.* [59] found that the incidence of neonatal morbidity and mortality can vary widely around the world (high-income countries report rates as low as 4 deaths per 1000 live births while low-income

countries have an average of 33 deaths per 1000 live births), while the total neonatal mortality rate (NMR)[1] was an estimated 4 million deaths per year. A similar number of stillborn deaths were also reported. By 2009, this number had fallen to 3.3 million deaths in the first month[2] [76]. Lawn *et al.* [59] also report that the main causes of neonatal death are preterm births (28%), severe infections (26%) and asphyxia (23%), and that the highest risk for death is on the first day of life, and almost 75% of deaths occur in the first week of life.

A more direct assessment of fetal deaths is the fetal mortality rate, the assessment of which has been a goal of the National Vital Statistics System (NVSS) for many decades now. The latest such report, published in 2009 [65], reports on trends in fetal mortality from 1990 to 2005. In 2005, there were an estimated 26,000 fetal deaths at gestational ages of 20 weeks or more, (referred to as stillbirths or fetal deaths). The average number of fetal deaths per 1,000 live births was 6.22, which has not declined significantly from its previous reported value of 6.2 in 2004.

Thus it is clear that the assessment of fetal health is of paramount importance. This assessment can be done in a number of ways. Nowadays electronic fetal monitoring (EFM) has permeated obstetrical practice almost completely, as opposed to being used merely for complicated pregnancies when the technology was first introduced in the 1960s [23]. The most ubiquitous form of EFM involves the simultaneous electronic monitoring of the fetal heart rate (FHR) as well as the uterine pressure (UP) signal. A detailed description of the technicalities of fetal monitoring is provided in Section 1.3. In 2002, nearly 85% of live births in the US (3.5 million women) underwent EFM [70]. The technology offers obvious advantages compared to old methods like periodic auscultation with a fetoscope. Other methods include the use of umbilical or scalp blood-sampling to determine pH values, a direct marker of fetal acid-base balance, as well as non-established techniques

[1]The term "neonatal" refers to "within the first 4 weeks of life".

[2]In [76], the authors have stated that reliable civil registration data were available from only 389 countries, while a statistical model was used to estimate NMR's for 155 other countries. Nevertheless, it seems to be generally accepted that NMR has decreased over this period.

such as vibro-acoustic stimulation, fetal pulse oximetry and near-infrared spectroscopy [1].

However, in daily obstetric practice, interpretation of the readily available EFM data is only based on visual analysis [8, 36, 78] and therefore is subjective and thereby plagued with high variability and unreliability. Very little headway has been made using EFM to effectively identify the fetus/newborn at risk for a poor perinatal outcome since the introduction of this technology. The major impact of classifying EFM tracings into 3 categories [66] was to reach consensus on terminology. No progress has been made in improving intra/inter-observer consistency in identifying the fetus at risk for fetal encephalopathy [1, 82, 98]. At the same time, pregnant women are placed at an ever increasing risk for cesarean delivery, which is presently at 40%, with its attendant risks to the mother of hemorrhage, sepsis, pulmonary embolism, decreased reproductive potential, not to mention the significant rise in the cost of obstetric care. Not surprisingly, for many years, there has been a concerted effort to automate the analysis of fetal heart rate (FHR) rhythms to remove the arbitrariness that may arise while evaluating them [98]. However, despite important advances in biomedical signal analysis, there have not been significant improvements in automated decision support systems in this area.

For instance, one of the newest and most promising developments in automated EFM classification is the use of features from both FHR and fetal electrocardiogram (ECG) data for fetal risk stratification in the STAN system developed by Neoventa, Sweden. In recent years, this system has undergone significant clinical trials in Europe and in the US [1, 28, 36]. However, at the present time a major limitation to the use of this system is that, to access the fetal ECG signal, one must insert the electrode into the fetal scalp, as explained in Section 1.4. As a consequence, the patient must be in labor and the fetal membranes must be ruptured. Importantly, it cannot be used in patients not yet in labor and in those patients who are at risk for infecting their fetus via ascending infection.

There also exist some commercially available (but, to our knowledge, not very widely used) systems for automated, noninvasive FHR monitoring

and classification, including the ones developed by LMS/Perigen (Princeton, New Jersey, USA) [47,107], Omniview Sisporto (Alfragide, Portugal) [5] and Oxford Sonicaid (Oxford, UK) [80]. A good review of these technologies from a clinical perspective is provided in [83]. Other recent systems, developed in academic research settings, but not yet widely available for general use, include those using neural networks [39,75], nonlinear feature-classifiers [97], multiscale complexity [49], and time-frequency analysis [29]. One reason they have not permeated clinical use in any significant way is that they often summarise very long EFM datasets with scalar values for features, and in doing so, often ignore effects of nonstationarity and changes in fetal dynamics. Moreover, in most attempts at classification, discriminative methods are used, which does not easily allow for the estimation of confidence intervals (i.e., how sure is the system about the classification output?) or interpretability (i.e., do the reasons for a particular classification of a given dataset make clinical sense?). As is often the case with biomedical signals, noise, artifacts, fetal and maternal movements induce significant non-stationarity and the lack of beat-to-beat resolution in Doppler recordings greatly compromise the quality of the data. All these factors make the signal analysis challenging but potentially quite beneficial. In summary, there is a great need for consistent and accurate methods of classifying FHR-UP traces. They will add meaningful value to the daily practice in obstetrics and can lead to much more effective clinical decision support.

1.2 Physiological basis

The oxygen-delivery mechanism to the fetus is a pathway composed of maternal uterine arteries, the placenta and the umbilical vein. Starting in the uterine arteries, oxygen flows through to the uterus, then to the placenta and then via the umbilical vein through to the fetus itself. Intervillous space between the placental and uterine tissues facilitates oxygen and nutrient delivery mechanisms as well as systems for removal of carbon dioxide and other waste matter. Umbilical arteries bring the carbon dioxide and other fetal waste products to the placenta for removal via uterine veins. This is

summarized in Figure 1.

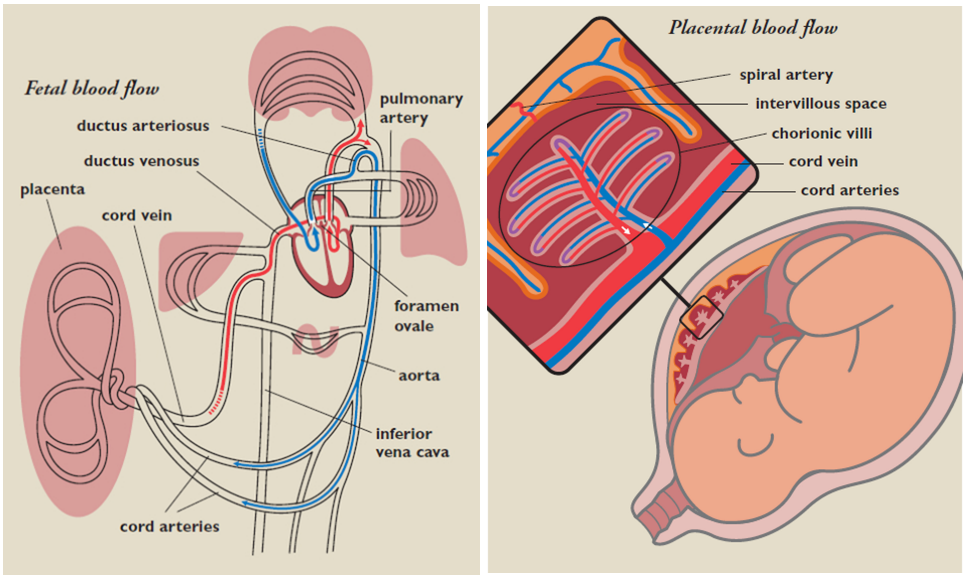

Figure 1: (a) Description of fetal circulation. (b) Intervillous space between the placental and uterine tissues facilitates oxygen and nutrient delivery mechanisms as well as waste removal. Figures taken from [100].

Uterine contractions cause reduction in uterine and placental blood flow. Consequently, oxygen exchange through the placenta is temporarily decreased, followed by normalization of the perfusion after relaxation of the uterus. Thus oxygenation of the fetus is directly dependent on 4 factors: maternal blood pressure, maternal oxygenation, available placental surface area for perfusion, and whether the umbilical cord is sufficiently open and unobstructed. If any one of these factors is compromised, it leads to compromised oxygen transfer, which is also termed uteroplacental insufficiency. During contractions, the fetus relies on appropriate residual perfusion of both maternal and fetal blood across the intervillous space (known as the placental reserve) for oxygen supplies.

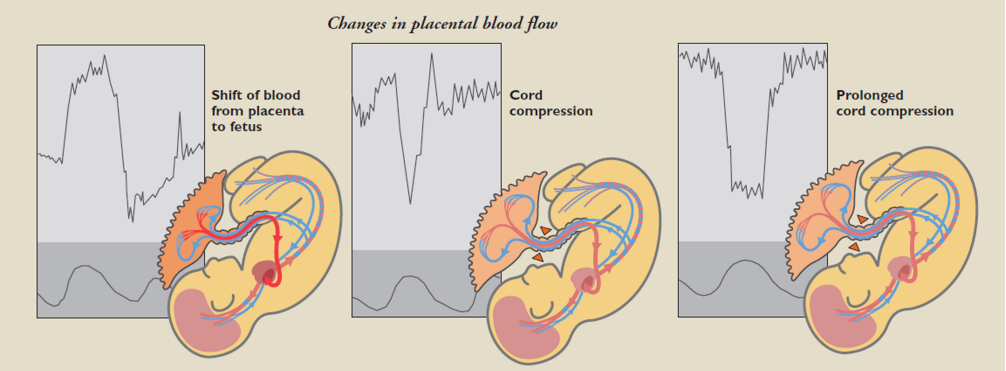

Figure 2: Fetal distress during umbilical cord constriction. Figure taken from [100].

1.2.1 Fetal oxygen insufficiency

Uteroplacental oxygen insufficiency may occur because of problems in one of 4 compartments.

1. If there is, say, an infarction in the *placenta*, it will result in poor oxygen transfer during relaxation of uterus, resulting in low oxygen levels (fetal hypoxemia).
2. Alternatively, if there are very deep contractions or too many contractions per unit time, the placenta will not have enough time to absorb oxygen from maternal blood between contractions.
3. The third mechanism is via decreased maternal perfusion of the placenta. This may occur if the mother is hypotensive, hypoxic, or has other complications. Even if the contractions are of normal length and size, the fetus may undergo distress because it cannot absorb enough oxygen even after the contraction is over.
4. Finally fetal oxygen supply may be insufficient if the umbilical cord is constricted, which decreases the delivery rate of oxygen to the fetus. This situation is displayed in Figure 2.

1.2.2 Relationship to the heart rate

Oxygenation levels in the fetal heart have a direct effect on the functioning of the fetal brain and nervous system. Since the cardiac contractions in the fetal heart are regulated by the autonomic nervous system (just as in adults), this will have a direct effect on cardiovascular functioning. The focus in the current work is on intrapartum (close to delivery) fetal heart rate analysis, by which time the cardiovascular system as well as the autonomic nervous system are sufficiently developed. Modulation of cardiac activity by the autonomic nervous system is carried out by 2 different mechanisms, namely parasympathetic and sympathetic.

Parasymathetic activation is mediated via the vagal nerve, and its main goal is rapid adaptation to changing physiological conditions. It is associated with decreases in fetal heart rate. Sympathetic activation, on the other hand, causes the release of stress hormones from the adrenal gland, usually causing an increase in fetal heart rates. Adaptation caused by sympathetic activation is in general slower than parasympathetic.

When a fetus is affected by hypoxia, acute or otherwise, chemo-receptors sensitive to the decrease in partial pressure of oxygen are activated. This, in turn stimulates both sympathetic and parasympathetic limbs of the nervous system, and causes an initial reduction in fetal heart rate if hypoxia is acute. Gradual hypoxia may in fact cause an increase in heart rate. Reduced placental blood flow during contractions can also cause similar decreases in heart rate via the same chain of chemo-receptor activation. These episodic decreases are called decelerations, and can also be triggered by the increase in blood pressure as part of the cardiovascular adaptation to the hypoxia. After the blood flow and oxygenation are returned to normal levels, the sympathetic activation is reactivated, causing increase in the FHR.

Despite a broad understanding of such physiological influences on the FHR, a complete understanding of these complex interactions remains elusive. However, the key takeaway is that any change in oxygen level directly affects the autonomic nervous system, and therefore the fetal heart rate. The objective of fetal heart rate monitoring by various methods is to assess this

influence and to solve the inverse problem of inferring whether there was a change in the oxygenation status in the first place, depending on the type of heart rate patterns observed.

1.3 Electronic fetal monitoring

Electronic fetal monitors typically measure two signals (both either externally or internally): the fetal heart rate signal, and maternal uterine activity.

1.3.1 The FHR time series

Internal monitoring of the fetal heart is achieved via direct acquisition of the ECG. A bipolar spiral electrode is attached directly to the fetal scalp. The wire electrode protrudes into the fetal scalp, while the metal wing on the electrode acts as the second pole. A saline electrical bridge is created by vaginal body fluids, which completes the circuit. This facilitates measurement of voltage differences between the two poles of the electrode. A third reference electrode is attached to the maternal thigh to eliminate all other electrical interference. The system setup is displayed in Figure 3.

In internal monitoring, FHR is derived from the ECG signal, which has distinct morphological features in each heart beat called fiducial points. These are denoted the P wave, QRS complex and the T and U waves. The R peak in the QRS complex is the most distinct and easiest to detect. The time interval between successive R peaks is called the RR interval, and the inverse of this is called the instantaneous heart rate.

In external fetal monitoring, ultrasound waves are used to detect the rhythmic movement of fetal heart valves and puslatile blood flow via Doppler shifts in frequency. The device consists of a transducer emitting ultrasound waves as well as a sensor to detect reflected waves. It is placed on the maternal abdomen with a coupling gel in between the transducer and skin to ensure proper conduction of the ultrasound waves. Reflected waves from many different moving sources are received. This creates a

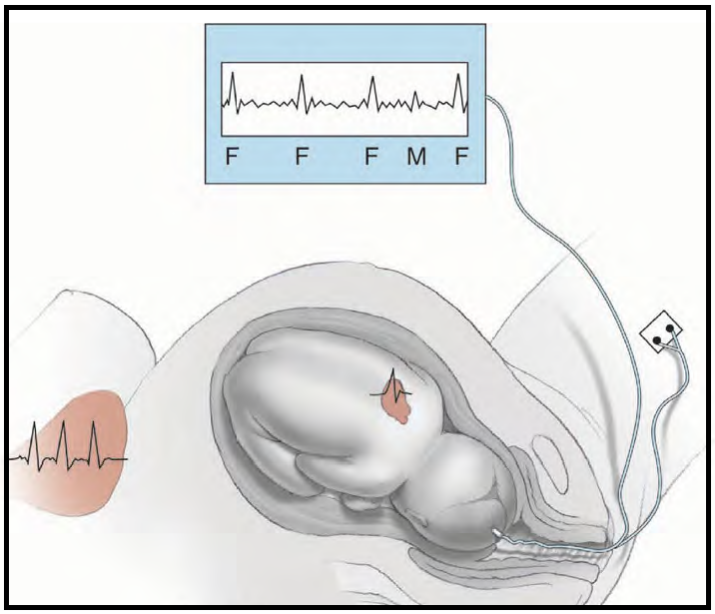

Figure 3: ECG acquisition setup in the case of internal fetal monitoring. Picture taken from [23], Chapter 18.

spectrum of many different frequency shifts corresponding to many different velocities. This velocity spectrum is repeated with each heart beat. An autocorrrelation method is used to detect periodicity in the spectrum as shown in Figure 4. In typical Doppler fetal monitoring units, significant post-processing is done on the signal after acquisition, in order to eliminate noise. In fact, since the autocorrelation process by definition requires at least a few heartbeats' duration of data for accurate rhythmicity detection, there is implicit smoothing which may obscure some high-frequency variations. The sampling rate in the case of Doppler data is 4Hz.

Despite the problems associated with Doppler monitoring, it is understandably the preferred choice of obstetric care providers for FHR monitoring, with scalp-electrode methods used only when warranted by necessity, such as when a cleaner signal is required for more detailed study.

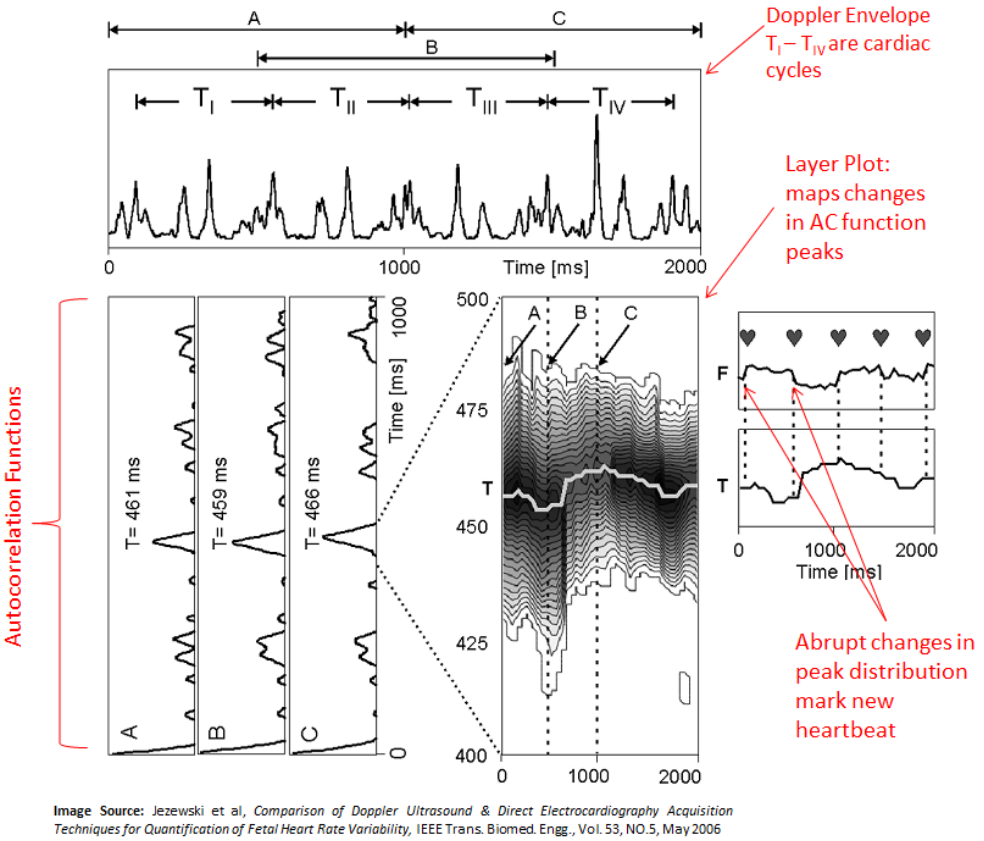

Figure 4: Doppler autocorrelation technique to detect heart period in the case of external monitoring. Figure based on figures in [52].

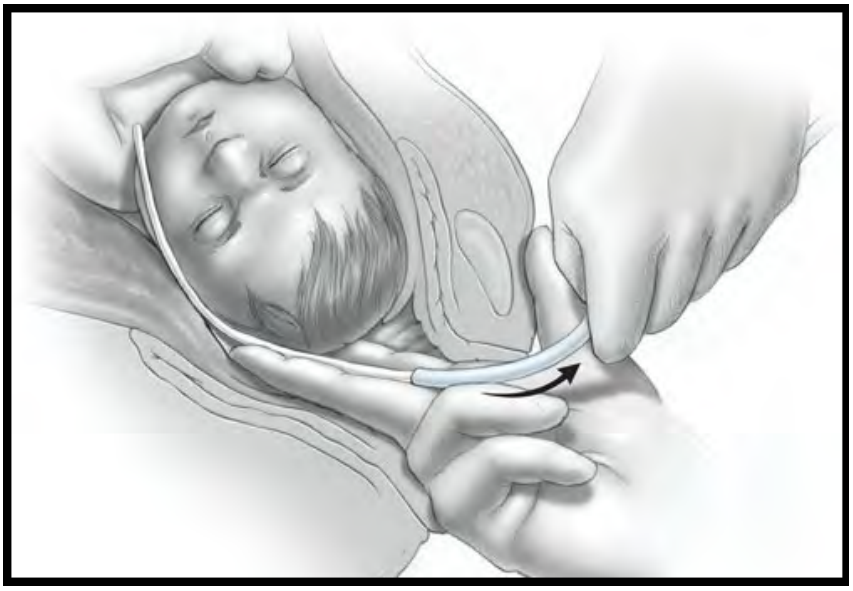

Figure 5: Placement of fluid-filled catheter for internal monitoring of uterine pressure variation. Figure taken from [23], Chapter 18.

1.3.2 Uterine activity

Internal monitoring of uterine activity involves the insertion of a catheter into the uterus along one side of the fetal head as shown in Figure 5. The catheter is made of plastic and contains fluid. It is connected to a strain-gauge pressure sensor calibrated to have the same level as the catheter tip in the uterus. Variation in pressure within the fluid system, caused by contraction activity in the uterus, creates a potential difference that is amplified and measured as a percentage signal.

In the case of external monitoring, a displacement transducer (in the form of a button or "plunger") is held against the abdominal wall. With each contraction, there is movement of the plunger proportional to the strength of the contraction. This movement can be converted into a measurable electrical signal, but it is a relative measure of contraction intensity. Thus, it is usually used to measure the onset, peak and return of the contraction.

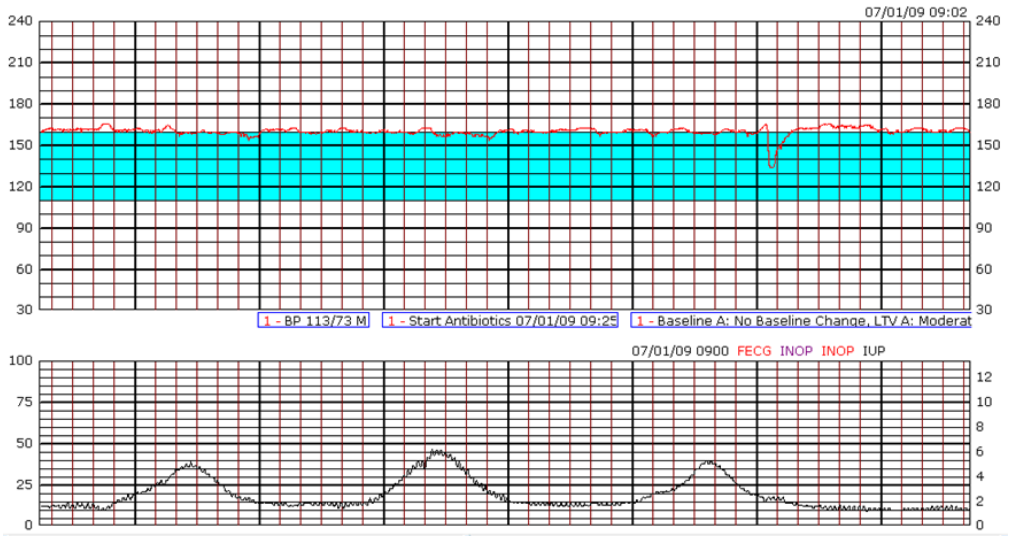

Figure 6: A typical recording of FHR (top) and uterine activity (bottom) as displayed on strip chart recorders. Blue area corresponds to the baseline FHR range considered to be "Normal". Information about whether there were baseline FHR changes or when antibiotics were started is also provided.

1.4 Clinical assessment

After acquisition, the two signals are displayed either on a digital monitor or read out on paper strips. A typical recording looks like in Figure 6. Note that information about certain morphological changes is also calculated by the fetal monitoring unit. Real-time clinical assessment of electronic fetal recordings is described in detail in Chapter 2 as well as in the standardized guidelines developed by the National Institute of Child Health and Human Development (NICHD), last updated in 2008 [66]. There are five main features of interest obtained from the two signals: uterine contraction rate (frequency of contractions), baseline heart rate, magnitude of variation of heart rate around the baseline rate and presence of episodic deviations from the baseline termed "accelerations" or "decelerations". The most important indicators of fetal distress are the lack of variability, presence of repetitive

late decelerations, and persistently bradycardic (lower than normal) average heart rate. In addition, sinusoidal patterns are also considered potentially dangerous. All these aspects are described in more detail in Chapter 2.

1.5 Previous studies

In the context of medical diagnosis and classification in general, rule-based systems have been in use for many years. Since 2002, they have been applied to areas like diagnosis and classification of thyroid neoplasms [89], chromatography applications [108], measurement of liver and kidney toxicity [61], otology [45], urinary incontinence [63], assessment of hepatotoxic potential [69], and diagnosis of primary immunodeficiencies [91]. A 2002 review by McNeely *et al.* [71] reported on the effective use of such expert systems in the field of laboratory testing, in systems such as BloodLink and LAS, as well as in providing interpretations of results of said testing. It has also been used in applications like orthodontry [48] and epidemiology of sleep disorders [77].

In the specific case of FHR analysis, computerized rule-based systems have already been implemented previously, e.g. [3, 80]. Seufert *et al.* [94] reported that it was possible to use artificial intelligence techniques like rule-based systems or neural networks for interpreting EFM patterns, but that its effectiveness would be significantly enhanced only after inclusion of external factors like partial oxygen pressure. Other rule-based systems were developed by [2] (called the NST-EXPERT) as well as rule based system by Keith *et al.* [56] that uses a database of 400 rules.

A comprehensive review of features relevant for FHR has been provided in [17]. They usually include power spectral density estimates [46,86], morphological features such as number of "accelerations" (increases in FHR), "decelerations" and their corresponding sizes [4, 54, 68, 80], linear features such as mean and variance of FHR over some time period and nonlinear features such as approximate or sample entropy [21, 31, 35, 40–43, 75, 79, 86, 96, 97]. One very interesting approach by Warrick et al [105, 106]

has focused on the dynamic relationship between the fetal heart rate and the maternal uterine pressure signal, quantifying it as an impulse response function and using the associated gain and phase delay as features capable of discriminating between normal and abnormal cases.

In the case of fetal heart rate monitoring, many standard classification approaches have been applied in recent studies. Noguchi *et al.* [75] used neural networks to classify a feature set composed of standard FHR morphological features as well as novel features quantifying the effect of sinusoidal variations in FHR. A back-propagation method was used for neural-network learning, with a modest-sized training database of 20 traces. A 2008 study by Chudacek *et al.* [18] used three different techniques for feature selection (principal component analysis, Group of Adaptive Models Evolution and neural networks) followed by direct correlation of well-discriminating feature sets with FHR pathology (indicated by abnormal range of umbilical pH values). Candidate features included many standard morphological features as well as several adult HRV features. Binary particle swarm optimization has also been used in [41] to perform automatic feature selection, followed by classification using support vector machine (SVM) and neural network techniques. The authors state that one important modification in their method was the use of minority oversampling to prevent errors arising out of imbalanced segmentation of examples. Features used included time domain as well as standard frequency domain heart rate variability (HRV) features. The same groups has also used grammatical evolution for feature selection [40] followed by multilayer perceptron for classification. A genetic algorithm approach was used to train the neural network.

In 2006, Costa-Santos *et al.* [21] used a clustering approach for classification of fetal heart rate recordings taken from 4 different hospitals. Clustering was done using a compression approach. A novel distance metric for measuring differences between two recording was defined using the ratio between compression statistic of two data sets taken simultaneously and the minimum of compression statistics for the two individual data sets. For all possible pairs of datasets, this distance metric is calculated, and clustering

is done using the simple rule that the smaller the "compression distance" between the two recordings, the more similar they are.

One of the most popular classifiers these days is the SVM, mainly because of the relative simplicity of the theory, the convenient extension to nonlinear discriminators through the use of the kernel trick, and high accuracy in several real applications [15,16]. It has been used recently in studies by Warrick *et al.* [105,106] for classification of FHR-UP recordings using the novel features described in Section 5.3.1. Another rich group of classifiers includes Bayesian methods, which allow probabilistic formulations of the problem at hand and a natural way to derive confidence measures for classification outputs. These have been described in detail in several books and papers, for instance, in [30].

1.6 Contributions

During the course of our research on these challenging datasets, we have developed a number of methods to analyse them. In addition, our use of Bayesian methods has opened up the possibility of new approaches to such analyses, as we have detailed in the following chapters. We now provide brief introductions to the main contributions of our research.

1.6.1 Translation of clinical domain knowledge

As reported in the previous section, several attempts have been made to build software implementations of the expert consensus guidelines developed for consistent EFM interpretation. We have implemented the guidelines prescribed by the NICHD in 2008 [66], which describe methods to (a) assess and quantify the different patterns observed in both FHR and UP signals such as baseline and contraction rates, numbers of episodic deviations and variability (b) rules to map numeric features to types and (c) rules to decide whether an FHR-UP tracing is reassuring or not based on what combinations of these types exist. Although rule-based decision systems do not, in general perform as well as probabilistic ones, one of

the advantages of translating these guidelines is that we can obtain several informative features that are easily interpretable by the obstetric care community. We have used modified versions of these features in different classifiers, notably in generative model-based systems, in nonparametric window-counting methods, and in Bayesian networks. We show performance evaluations with several real datasets using both rule-based and other systems.

1.6.2 Use of generative models for data representation

Our proposed approach is probabilistic, where we aim at employing generative classifiers which learn the joint probability distribution of the data and their labels, followed by the use of Bayes' rule [53]. Generative models (GMs) are more flexible than discriminative models in expressing dependencies among different variables [11]. Furthermore, in contrast to discriminative models, they can naturally be used in unsupervised learning [11,53]. Even though GM are computationally demanding, they allow for estimation of marginal distributions of data, which in our scenario is very useful. More specifically, with GMs, one can achieve improved detection of new data that are outliers under the model and whose predictions have low accuracy [10, 101]. In our problem, we can view the abnormal traces as outliers because they simply are relatively rare, which further justifies the use of these models (in our preliminary study, less than 1% of all the data were classified as abnormal by physicians).

The GM methodology also naturally allows for accurate clustering using nonparametric Bayesian (NPB) methods when the number of clusters is unknown. This approaches uses the concept of infinite mixture models to provide degrees of freedom unmatched by other approaches where the number of classes has to be prespecified. In recent years, NPB methods have been steadily gaining in popularity, primarily because of the improved understanding of the theory behind them and the potency of present day computers for their implementation [50,72]. They are non-supervised and most

importantly, they do *not* require that the number of classes is predefined. Instead, the clustering of the data is "combined" with the task of determining the number of classes. When new data arrive and are classified, they can either join one of the existing classes or they can be considered to belong to a new class. This allows the number of classes to grow as new data are being acquired and are being classified. Bayesian approaches to inference and estimation can be applied very easily in this framework, which means one can specify priors for all the parameters in the system and obtain posterior distributions for them.

1.6.3 Feature identification

A critical step for supervised or unsupervised clustering is the identification of features from the FHR-UP signals. As part of our research, we have explored a number of new features, including several derived directly from physician's domain knowledge. They include baseline FHR trend and variability, contraction rates, and numbers and types of episodic variations such as accelerations and decelerations. Additionally, we have extracted novel informative features from time series of fractional changes in successive FHR samples, which were also found useful for classification using non-parametric window based classifiers.

We used extensive analysis using Bayesian networks to isolate subsets of features that were informative of the fetal status in a phenomenological sense. That is, we use a large database of real FHR-UP signals, extracted several candidate morphological features and used efficient network structure detection strategies to identify which subsets were directly related to fetal status. This method also gave us significant new insights into which feature-pair correlations are actually observed in real data, and can thus be important for fetal diagnosis. When refining our classification methods, such information allows us to ignore redundant features and can make the algorithms more efficient. In addition, we also used stratified crossvalidation strategies [57] in conjunction with our classification methods, with different feature subsets fed as input, to check which feature sets gave highest

classification accuracies for the various methods.

1.6.4 Exploiting dynamics of individual time series

An important feature of the used signals is their rich dynamic nature, and extracting information from it provides additional feature space for discrimination. Our approach to exploiting the time dynamics of the FHR-UP signals for classification relies on naïve Bayes and first-order Markov models for describing feature-sequence evolution. In particular, we demonstrate that using sequences of features derived from short segments of long stretches of data give a more complete picture of the variations in FHR-UP patterns. Compared to several existing methods of FHR analysis that map such long datasets (often 20-40 minutes or more) to scalar values for the features, our sequence-of-features input, when modeled using the GM methods described above, yields higher accuracies.

In fact GMs open many new possibilities for extracting and modeling such sequences of features. We first explore the use of fixed-size segmentation of the FHR-UP time series data, and provide ways to deal with the problem of estimating the segmentation duration using crossvalidation schemes. As shown in Chapter 6, however, we can also model feature sequences by describing the time series with a sequence of hidden states (which would be proxies for the features) and state-dependent observations, i.e., as the observation sequence in a hidden Markov model. Now, if we were to implement efficient, unsupervised algorithms for detecting the "best-fit" state (feature) sequences, we no longer have a need to prespecify the segmentation period, and thus avoid introducing unnecessary parameters. We demonstrate the use of novel NPB methods to perform such data-driven segmentation, in which we do not need to make any assumptions about the number or types of possible FHR-UP "states".

Considerable attention has also been devoted to NPB segmentation of switching autoregressive processes, notably by [7, 33, 102]. In [33], the authors have shown the use of a Baum-Welch type forward-backward (FB) recursion to block-sample the state sequence of the HMM ($x_{1:d}$) from

$P(x_{1:d}^{(t)}|y_{1:d}, \psi^{(t)}, A^{(t)})$ in order to increase the mixing rates in the direct Gibbs (DG) approach. In this study, we improved upon these results by Rao-Blackwellising the FB algorithm using a Monte-Carlo approach. By sampling ψ, A, multiple times from their respective posteriors and averaging the resultant posterior probability of the state-sequence, we are able to encourage the sampler to explore the posterior landscape more efficiently.

One can also extend this idea quite naturally to the case where overall FHR-UP categories as well as FHR-UP patterns can be described using infinite mixture models,. The underlying idea for using the NPB approach is to keep the number of possible patterns in the data flexible. The approach allows for discoveries of shared patterns across different time series.

1.6.5 Fusion of information from UP signals

Recent studies have shown that timing and magnitude of the heart rate responses to changes in maternal uterine pressure signals can be indicative of fetal distress [22, 100, 105, 106], which correlates well with physicians' experience. Information about whether a contraction occurs in the vicinity of some distinct pattern in the FHR time series can be important for classification, particularly when we try to exploit the short-term dynamics of the two series to derive features. In the studies by [105, 106], a linear relationship is assumed between the two signals, with the UP signal as input and FHR as output. System identification methods are used to quantify the magnitude, memory and time-delay of the corresponding impulse response function of the system, and these are used as features for classification. Once again, however the techniques used in these studies rely on parametric approaches and use long stretches of data, which fails to account for nonlinearities and nonstationarity, and thus fail to be effective for many cases.

In our research, we derive several informative features from the UP signals. We use a probabilistic approach to combine knowledge about contractions with the observed FHR patterns. It makes far fewer assumptions about the nature of the relationship between the two signals, and can allow for much more robust inference and estimation methods. Moreover, the

specific GMs used in our research use discretized versions of the features, and so can be very easily used to perform this fusion of information. Our primary idea is to jointly label the segments of both the FHR and UP signals. For example, we have labels like "Contraction Present - Acceleration Present" or "Contraction Absent - Deceleration Present", thus generating sequences of letters from the joint FHR-UP alphabet. Although this does double the number of symbols in the feature alphabet, it can stil summarize information very compactly and requires much less overload to estimate various model parameters (compared to say, the earlier system identification methods which have to analytically calculate or approximate inverses of the coefficient matrices).

1.7 Dissertation outline and references

In this chapter, we provided the motivation, main goals and contributions, a general overview of the biological problem, the EFM methodology and a review of the existing literature on FHR monitoring. In Chapter 2, we describe in detail the translation of the NICHD standardized guidelines for EFM interpretation. Performance evaluation on a small preliminary database consisting of 30 expert-annotated time-series is described. Evaluation of this method on a larger dataset is described in Chapter 5. Parts of Chapter 2 have been taken from our 2011 paper submitted to the Asilomar Conference on Signals Systems and Computers [25].

Chapter 3 shows how the NICHD features can be incorporated as random variables (nodes) in a directed acyclic graph (DAG), also called a Bayesian network (BN). Our goal here is to try to make the process of getting confidence intervals for classification more systematic. In addition we are able to use methods of BN structure detection to get insights on which feature-pairs show substantial correlations as evidenced in 830 real FHR datasets collecred from nine subjects, with gold-standard labeling done by visual annotation. . Parts of this chapter are taken from our 2012 paper presented in the IEEE Engineering in Medicine and Biology Conference [27].

We then describe, in Chapter 4, novel methods to extract features from

short-duration segments. We demonstrate their use as inputs to a novel window-based classification method that does not make any parametric assumptions about the data, and is able to improve classification compared to existing discriminative approaches such as SVMs. We show performance evaluation of these methods on a dataset consisting of 580 short 15-s epochs of FHR time-series data, with gold-standard labeling done by visual annotation. Parts of this chapter are taken from our 2012 paper presented in the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [26].

We describe our GM-based feature-sequence classification approach in detail in Chapter 5. This incorporates the descriptions of the uniform segmentation of the FHR-UP time series pair, discretization of the feature alphabets, their modeling using two different GMs and the classification methodology using maximum a-posteriori rules. Unlike previous evaluations which used data collected from a small number of patients, performance evaluation was performed on data collected from 83 patients. We used as a fetal health metric the post-delivery umbilical cord pH value, which is a direct measure of the fetal acid-base status (lower pH is indicative of acidemia which is a result of reduced oxygenation). Parts of this chapter are taken from a paper we recently (in 2013) submitted to the journal IEEE Transactions on Biomedical Engineering.

Finally, we conclude by describing our ongoing research on extensions of the GM methodology. Parts of this chapter dealing with the data-driven time-series segmentation have been taken from a paper we recently submitted to the 22nd European Signal Processing Conference (2014). We describe preliminary results and plans for four different goals: (a) the use of NPB methods to perform data-driven segmentation of and feature extraction from FHR-UP signals, (b) the use of NPB methods to perform unsupervised clustering of whole feature sequences, (c) the development of a fetal "Risk Score" that combines information from objective and visual fetal assessments to get a single gold-standard measurement of fetal health, and (d) the extension of the clustering methods to real time monitoring.

2 Rule-based classifier: the NICHD expert system

One way to deal with the problem of high intra- and inter-observer variability is to make a set of standardized diagnosis criteria for clinical diagnosis that is in line with consensus knowledge. Since 1997, the National Institute for Child Health and Human Development (NICHD) has been publishing periodically updated standardized guidelines for the interpretation of the EFM traces and for use in clinical evaluations, which have been in use in the US for some years. The latest version was published in 2008 [66]. It describes (a) general rules for visual assessment of morphological features in the FHR-UP recording, (b) a guide to qualitative clustering of these features into categories that are indicative of fetal health and (c) if-then rules for final diagnosis of the trace given the appearance of specific combinations of features. Figures 7, 8 and 9 show examples of FHR-UP tracings that were visually classified into the three different NICHD categories.

The term expert system (ES) refers to a type of classifier in which simple if-then rules are used to translate observed relationships between features for final diagnosis. They have the advantage of being very easy to interpret, and are most naturally used in database applications where information is encoded in logical relationships between observed variables [30]. An example (simplified) illustration in the case of, deciding whether I have a cold or allergy (assuming these are the only possible explanations for my symptoms) is shown in Algorithm 1.

In this regard, expert systems can be seen as rule-based systems. The most natural way to use such systems is to first decide which features can be used to encode such information. These features are denoted x. The

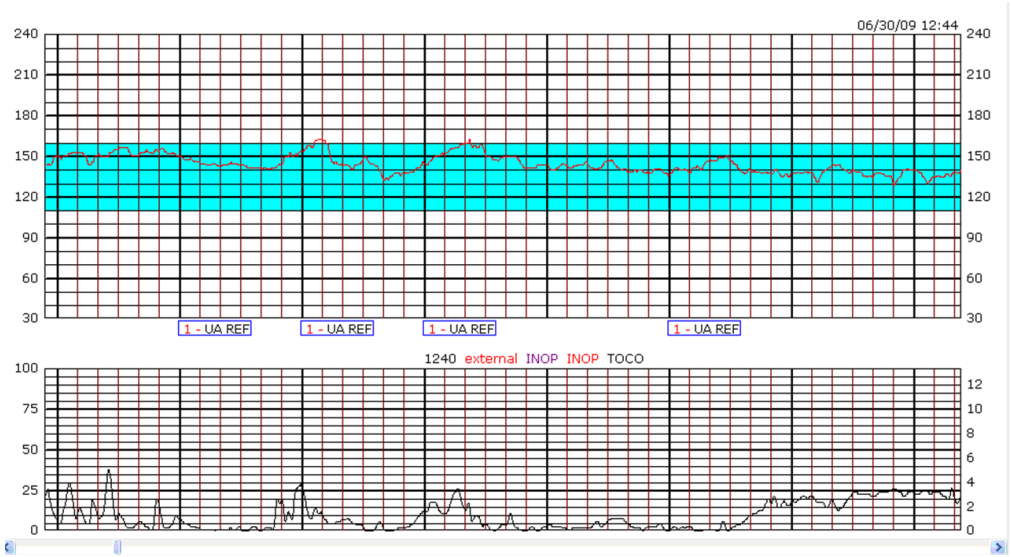

Figure 7: An example FHR-UP record that was visually interpreted as a category 1 tracing according to NICHD guidelines. The record shows typical characteristics of a *normal* FHR-UP record: normal-range baseline (about 140 bpm), moderate variability, the presence of accelerations present, and no decelerations.

Algorithm 1 Algorithm to decide whether a subject suffers from cold or allergy given knowledge of the symptoms.

```
procedure DECIDECOLD(sneezing,cough)
    cold = FALSE; allergy = FALSE;
    if (sneezing == TRUE) AND (cough == TRUE) then
        cold = TRUE
    else(sneezing == TRUE) AND (dust == TRUE)
        allergy = TRUE
    end if
end procedure
```

next step is to find how these features are instantiated, i.e., what their domains are. Next, we need to find which combination of instantiations is relevant for classification. In the cold example, our features $\{x_1, x_2, x_3\}$ are "sneezing", "cough" and "dust". Then the relevant combinations are

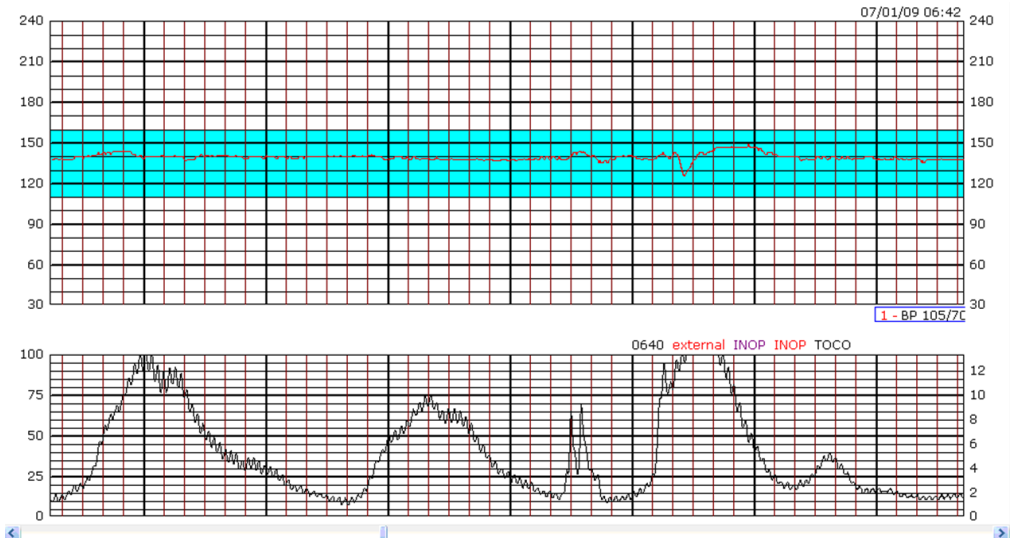

Figure 8: An example FHR-UP record that was visually interpreted as a category 2 tracing according to NICHD guidelines. The record shows typical characteristics of a *non-reassuring or indeterminate* FHR-UP record: tachycardic baseline (about 160 bpm), minimal (but not absent) variability, and the absence of accelerations and decelerations.

$\{x_1 = 1, x_2 = 1\}$ and $\{x_1 = 1, x_3 = 1\}$. Logical rules for classification are constructed by using predicates which return outputs of TRUE or FALSE based on which specific combinations of feature-instantiations are present. These are denoted $y(\mathbf{x})$ where $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}$; and the range of y is $\{0, 1\}$. These predicates are used in the final diagnosis in if-then trees.

In their book on pattern recognition [30], Duda *et al.* have differentiated between *propositional* and *first-order* if-then rules. In the former, the logical variables within the predicate can only take values of True or False (e.g., x_1, x_2, x_3 in the cold example), whereas predicates under the latter definition permit the use of variables in the within-predicate rules. The former has the disadvantage that one cannot represent general relations occurring in large numbers of instances. For instance, in the context of FHR analysis, one may wish to encode the general rule that the presence of decelerations that are both *variable* as well as *recurrent* is an indication of distress. Clearly, not every recurrent deceleration is variable (nor is the converse

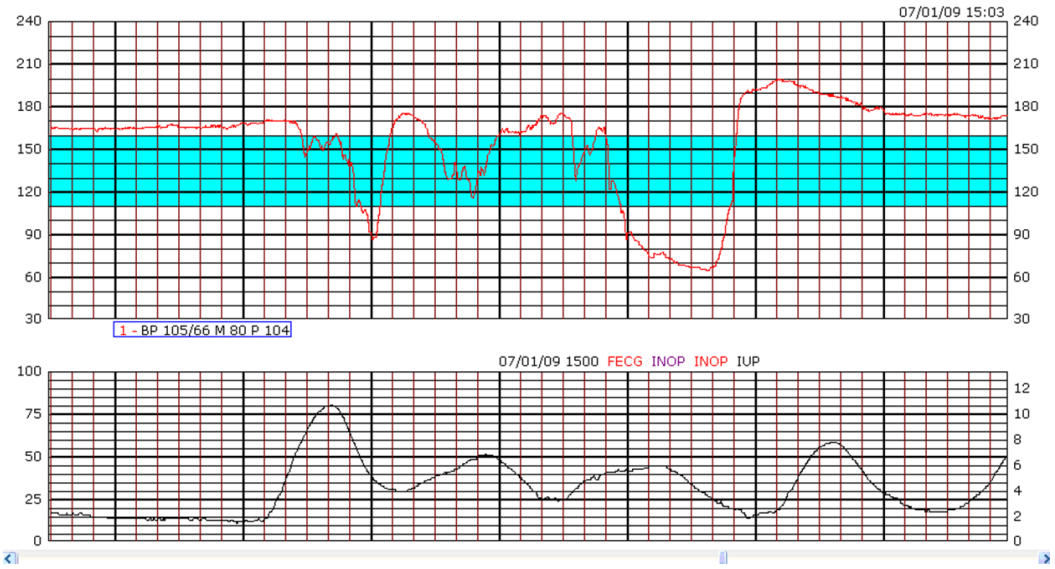

Figure 9: An example FHR-UP record that was visually interpreted as a category 3 tracing according to NICHD guidelines. The record shows typical characteristics of an *abnormal* FHR-UP record: tachycardic baseline (about 170 bpm), absent variability, the absence of accelerations and presence of variable decelerations.

true). Thus if only logical features such as $x_1 =$ anyRecurrentPresent and $x_1 =$ anyVariablePresent are used in the predicate, the rule may falsely output TRUE in some cases. To overcome this, we have to define first-order predicates which can use as input the location and size of each deceleration, and output TRUE only when all conditions of *Recurrent Variable* decelerations are satisfied. Thus, using both types of features may be better from classification point of view.

The main disadvantage of using expert systems is that there is no natural probabilistic formulation for them, which means it is difficult to get confidence measures for classification performance. However, from the point of view of ease of interpretation, they are arguably the best.

In this chapter, we describe a novel expert system built on the NICHD rules for performing consistent diagnoses and for emulation of physician

decision making in the clinic. Building such systems involves feature extractions that give equivalent information as that visually adjudged in the clinic. According to the NICHD rules there are five main informative features for diagnosing fetal health. They are the baseline rate, uterine contraction frequency, baseline FHR variability, presence and types of decelerations, and presence of accelerations. Algorithms that assess each of these are described in the sequel. In accordance with the NICHD rules, we use simple rules for categorizing a continuous valued feature (such as baseline variability) into "types" (such as "marked" or "minimal"). Finally, the algorithm performs diagnosis based on knowledge of which combinations of feature types occur for any given subject.

In the next section we describe the features and their extraction from the acquired signals. In Section 2.2, we describe the feature categorization and in Section 2.3, the diagnostic decision flow. Experimental results on some preliminary data are provided in Section 2.4 along with some discussion. Performance evaluations on a larger dataset are described in Chapter 5. We provide a list of all the symbols and their meanings in Table 1.

Table 1: List of symbols and definitions used in this chapter

Symbol	Definition
$u[n]$	UP at sample n
$h[n]$	FHR at sample n
f_s	Sampling rate 4 Hz
$b_u[n]$	Baseline of UP signal at sample n
$b_h[n]$	Baseline of FHR signal at sample n
$n_s^u(n_r^u)$	Sample number at which a uterine contraction first deviates upwards from (returns to) UP baseline
$n_s^A(n_r^A)$	Sample number at which an acceleration first deviates upwards from (returns to) FHR baseline
$n_s^D(n_r^D)$	Sample number at which a deceleration first deviates downwards from (returns to) FHR baseline
n_p^u	Sample number at which a UP contraction peak occurs
n_p^A	Sample number at which an acceleration peak occurs

Table 1 – continued from previous page

n_p^D	Sample number at which an deceleration nadir occurs
L_u	Duration of a given UP contraction = $n_r^u - n_s^u$
L_A	Duration of a given acceleration = $n_r^A - n_s^A$
L_D	Duration of a given deceleration = $n_r^D - n_s^D$
θ_L^u	Duration threshold (in samples) which needs to be exceeded for a UP contraction to be considered valid
θ_L^A	Duration threshold (in samples) which needs to be exceeded for an acceleration to be considered valid
θ_L^D	Duration threshold (in samples) which needs to be exceeded for an deceleration to be considered valid
θ_s^u	Upward deviation threshold (in percentage points) which needs to be exceeded for a contraction onset to be confirmed
θ_s^A	Upward deviation threshold (in bpm) which needs to be exceeded for an acceleration onset to be confirmed
θ_s^D	Downward deviation threshold (in bpm) which needs to be exceeded for a deceleration onset to be confirmed
θ_p^D	Duration threshold a deceleration downslope (onset to nadir time) needs to exceed to be confirmed as valid.
F_u	Number of detected contractions in a 20-minute period
$\bar{h}_p^A$	Peak deviation of acceleration from FHR baseline

2.1 Feature extraction

2.1.1 Preprocessing

Prior to carrying out feature extraction, some preprocessing is done to remove various artifacts including ones due to movement. We use a method that is similar to that employed in [3]. The FHR time series, which had

been acquired via the Doppler-autocorrelation or internal scalp-electrode method with sampling frequency $f_s = 4$Hz, are processed to remove so-called "spiky" artifacts. The latter are defined as FHR segments where successive HR differences greater than 25bpm were detected. Whenever such a beat combination is detected, linear interpolation is performed between the first detection and the first subsequent "stable" segment, defined as a group of five samples whose beat-to-beat difference did not exceed 10 bpm. The program stores the time information of the interpolated segments in order to isolate tracing areas with large amounts of noise. While looking for specific features such as decelerations, the system searches certain sub-segments of the 20-min epoch. If the total duration of interpolated periods during any sub-segment exceeds 30% of the sub-segment duration, the sub-segment is rejected from the search procedure entirely. The UP signal is generally a cleaner signal. It is smoothed by a simple averaging filter whose length is fixed 17 samples. The smoothed FHR and UP signals were passed to the feature extraction block.

In the following, the thresholds for various onset, return, and peak/nadir detections are denoted θ. The corresponding times of detection are denoted n, and the actual FHR and UP values, by h and u, respectively. Accelerations are denoted as A, while decelerations by D.

2.1.2 Uterine Contractions

The UP signal is first scaled to values between 0 and 100 (percentage scale) and then a reasonable "baseline" is estimated via mode estimation. A Gaussian kernel method is used to estimate the probability mass function (using bin centers at $\{0.5, 1.5, \ldots, 99.5\}$). The kernel widths are calculated using the method from [9] as follows (where $u[n]$ is the UP signal and $N = 20$ minutes):

$$\begin{aligned} \text{Kernel width } S &= 0.9 \min\{\sigma, 1.4826 M_u\}, \\ \text{Where, } \mathbf{u} &= \{u[1], \cdots, u[N f_s]\}, \\ \sigma^2 &= \text{Variance of } \mathbf{u}, \\ M_u &= \text{Mean absolute deviation of } \mathbf{u}. \end{aligned}$$

The u value at which the pmf is maximized is considered the baseline b_u. A uterine contraction onset is detected whenever $u[n]$ exceeded b_u by a minimum of $\theta_s^u = 3\%$, and that time instant is denoted n_s^u. For each such onset candidate the return time n_r^u is detected. If the duration of contraction $L_u = (n_r^u - n_s^u)$ exceeds the threshold $\theta_L^u = 185 f_s$, the mode for this candidate period is recalculated by the above procedure. The onset and return detections are recursively performed in this way until a valid contraction, if any, is detected. Finally, once a valid contraction is detected, the system calculates the peak time of the contraction as n_p^u. For diagnosis, the contraction frequency F_u, defined as the number of detected contractions in a 20-minute period, is calculated as the feature of interest. An example of contraction detection is shown in Fig. 10, where the recursive mode computation is also shown.

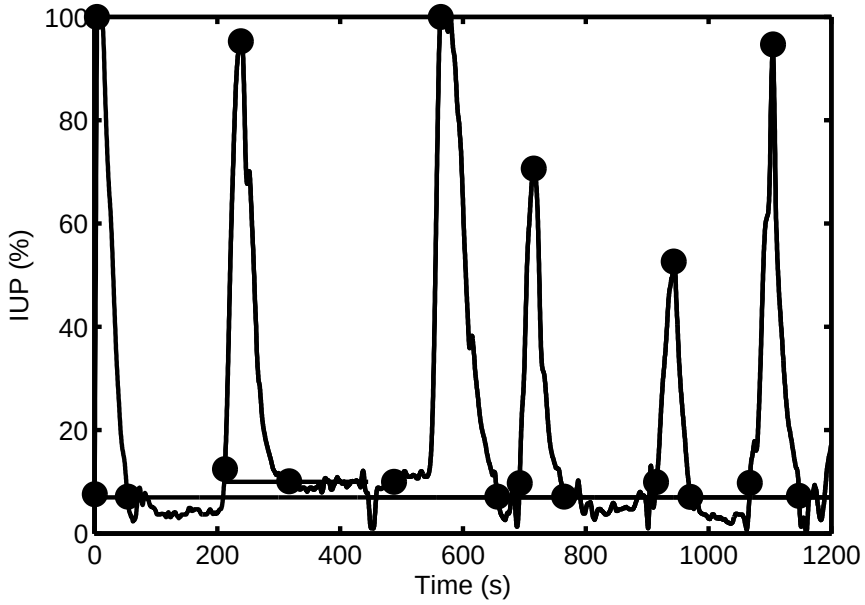

Figure 10: An example of uterine contraction detection with onsets and ends annotated by filled circles. The mode was calculated recursively for the second contraction as described in the text.

2.1.3 Baseline rate

Clinically, the baseline is defined as the average heart rate over FHR periods free from episodic deviations such as accelerations, decelerations and marked variability-periods. However, the episodic deviations are defined with reference to a pre-calculated baseline FHR, which leads to a problem of definition of the baseline. Despite this, the baseline is estimated without much trouble during clinical assessment (where doctors are indifferent to its precise values). However, for a programmatic description one needs to provide more concrete rules.

In the current approach, the baseline FHR is estimated using a windowed median filtering method. Simulation methods are used to measure the performance of median filtering methods for various window lengths, and a five minute window length is found to be appropriate for accurate baseline estimation. The key for good estimates is in keeping the window short enough so that it does not miss important slow changes to the FHR trend (of periods free of episodes) while rejecting shorter episode-related deviations. The baseline signal is denoted $b_h[n]$ and was defined over the same time interval. The feature of interest from diagnostic perspective is the median baseline FHR denoted B_h. An example baseline estimation result can be seen in Fig. 11.

2.1.4 Accelerations

Clinically, accelerations are defined as "visually apparent *abrupt* increases from baseline". Once the baseline FHR is estimated, the onset times of accelerations are detected as the first sample indices n_s^A when the FHR $h[n]$ upwardly deviated from $b_h[n]$ by at least $\theta_s^A = 1$bpm. For each onset candidate, the system estimates the return time n_r^A and the duration $L_A = n_r^A - n_s^A$. If $L_A > \theta_L^A (= 15 f_s)$, the system estimates the location n_p^A of the peak deviation from baseline, which is denoted $\bar{h}_p^A$. Typically since the FHR is not a smooth signal, detecting an obvious peak is difficult. Hence, the system detects only the first "significant" peak, defined as the first local maximum within the top 20th percentile of the series of FHR deviations during the

acceleration. If there is no such local maximum, it simply calculates the global maximum during the acceleration duration. Finally, the candidate acceleration is said to be valid if it satisfies the following three conditions:

$$n_p^A - n_s^A < \theta_p^A = 30 f_s,$$
$$\bar{h}_p^A > \theta_h^A = 15 bpm,$$
$$L_A \in [15 f_s, 600 f_s].$$

2.1.5 Decelerations

Clinically, decelerations are defined as visually apparent *abrupt* or *gradual* decreases from baseline. Once the baseline FHR is estimated, the system detects the onset times of decelerations as the first sample indices n_s^D when the FHR $h[n]$ downwardly deviates from $b_h[n]$ by at least $\theta_s^D = 1$bpm. For each onset candidate, it then estimates the return time n_r^D and durations $L_D = n_r^D - n_s^D$. If $L_D > \theta_L^D (= 15 f_s)$, it is considered a likely deceleration candidate, and the nadir n_p^D and the corresponding deviation from the baseline $\bar{h}_p^D$ at the nadir location are then found. In order to detect only the first significant nadir, a procedure similar to the one for accelerations is used. In addition, it was observed that deceleration detection was particularly prone to false positives because of a higher degree of noise due to electrode movement/ drop-off. In such instances, the signal suddenly would dip below threshold and it could take it some time to come back to baseline, thus artificially increasing the "abruptness" of the episode. Hence a different threshold, $\theta_p^D = 3 f_s$, to differentiate true decelerations from such false episodes, is used. In other words, any candidate deceleration has to take at least 3s from onset to nadir to count as a valid deceleration. An example detection is shown in Fig. 11.

2.1.6 Baseline Variability

The variability of the FHR signal is considered one of the most important features for detection of fetal distress. Whereas there is a rich literature on adult heart rate variability with more-or-less agreed upon standards of

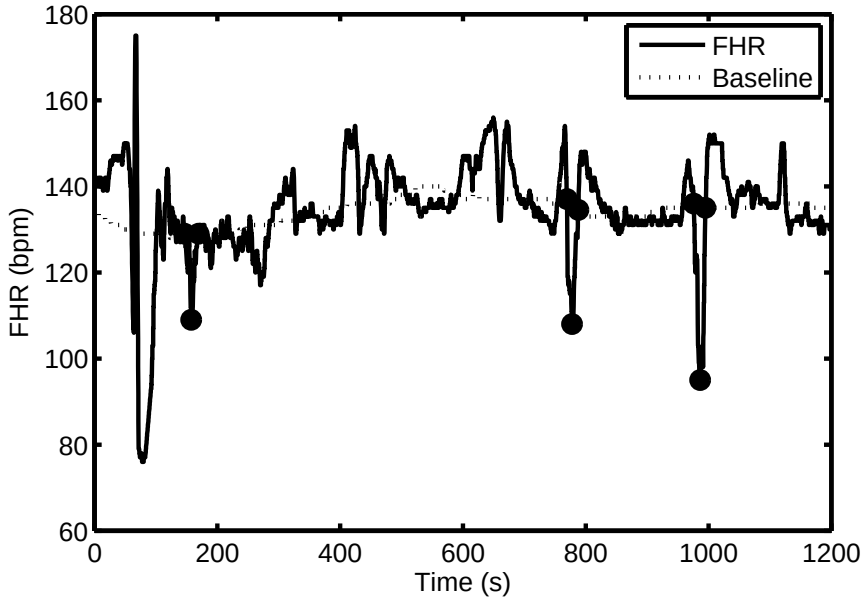

Figure 11: An example of deceleration detection and estimated baseline. As desired, the large artifact towards the beginning of the tracing was not treated as a deceleration.

measurement [67], there is no such agreement for fetal heart rate studies. In fact, the 2008 NICHD guidelines make no difference between beat-to-beat and long-term variability, "because in actual practice they are visually determined as a unit."

In order to stay close to physician interpretation of variability, and with a view towards keeping feature extraction as non-parametric as possible, traditional methods of estimating variability, such as power spectral densities or entropy measures, are not used. Instead a simple zero-crossing method is defined as follows. First, the system finds sub-segments in the FHR series $\mathbf{h} = \{h[1], \ldots, h[Nf_s]\}$ which are free of accelerations, decelerations and noise (as defined previously). Each such sub-segment is first de-baselined (using the b_h value) and further divided into non-overlapping one-minute segments. From the resulting signal $\bar{h}_v[n]$, the program estimates the number of times the signal went above (resp. below) the thresholds θ_δ (resp. $-\theta_\delta$), and this result is denoted k_v. This is taken as an estimate

of the number of FHR cycles around the baseline. If the per minute cycle frequency ($= k_v$) exceeds the clinical threshold (2 cycles/min) for a valid variability signal, the program then estimates a feature of interest as follows. For each detected cycle, the crest-to-trough range is estimated. The median of all these values is the variability $\tilde{V}_h$ for the one-minute sub-segment. Finally, in order to calculate a variability value for the full 20-minute signal, the median value of all the $\tilde{V}_h$'s over that period is calculated. This is denoted V_h. An example of variability estimation for a one-minute subsegment is shown in Fig. 12. In this example, there are three clear crest-to-trough crossings that clear the defined thresholds $\pm\theta_\delta = \pm 2$ bpm, which translates to a per-minute cycle frequency of 3 cycles per minute.

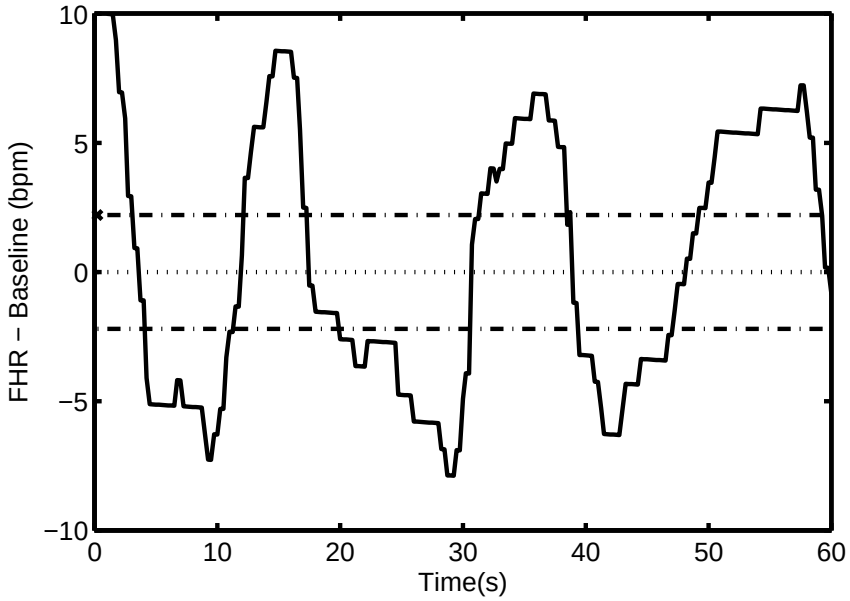

Figure 12: An example of variability estimation. The middle dashed line shows the zero value, and the two dot-dashed lines are the crossing thresholds ($\pm\theta_\delta$).

2.2 Feature categorization

Clearly, simply finding out numerical values for the above features is not very helpful for clinical diagnosis. In order to be of value to physicians, one needs to define appropriate categories which would map continuous valued features into "types" or "quality" measures. This has also been done using the guidelines in [66].

First, the uterine contraction rate is denoted *Tachysystole* whenever the contraction frequency F_u exceeds a value of 0.5 contractions per minute as calculated over the 20-minute segment; otherwise, it is called *Normal*.

Next, the baseline rate is categorized. If B_h is less than 110 bpm, it qualifies as *Bradycardia*, and if it is greater than 160 bpm, it is denoted *Tachycardia*. Baseline FHR in the intermediate range is considered *Normal*.

The value of baseline variability V_h is considered *Marked* if it exceeds a threshold of 25 bpm. Values of V_h between 5 and 25 bpm are called *Moderate*, while those between 2 and 5 bpm, *Minimal*. *Absent* variability corresponds to the situation when the variability detection algorithm cannot find a single valid "cycle" in most of the one-minute epochs in the data set (thus making the median V_h null valued). In clinical practice, when physicians see a so-called "flat-line" trace in the HR record, it is considered that variability is absent. However, the "flat-line" criterion does not seem to be strictly used. For example, in the current training database, there were instances in which physicians would classify a segment as having absent variability even when small fluctuations could be perceived. This is one of the reasons the thresholding approach is used in the variability algorithm, instead of just looking for zero-crossings.

Accelerations are classified as either *Normal* or *Prolonged* depending on whether the total duration of the episode (from onset to return) is less than two minutes or not. However, things are slightly more complicated in the case of decelerations. In clinical practice, four different types of deceleration-related features are assessed: (a) time until deceleration nadir, (b) timing of each deceleration with respect to associated uterine contractions, (c) number of decelerations associated with uterine contractions and

(d) total duration of each deceleration. Thus, for the expert system implementation, one first needs to define a set of rules to decide which uterine contractions are associated with the deceleration. This is done in the current algorithm by simply finding any contraction that has at least 25% overlap with the deceleration. In this way, there may be more than one deceleration for some contractions or vice versa.

The first classification (covering cases (a) and (b)) is shown in the pseudocode in Figure 2. Every detected deceleration is first classified into one of three types (*Early*, *Late* or *Variable*) using this algorithm. In addition, there may be some cases where none of the if conditions is satisfied by a given deceleration (for instance, if the deceleration is gradual but does not have any contraction associated with it). Since, clinical guidelines do not explicitly state how to deal with such cases, we classify such decelerations as type *Unknown*.

For deceleration-related information of type (c), the goal is to find whether each type of detected deceleration is (in clinical parlance) *Recurrent*. This is illustrated with an example. Let us assume that for some FHR trace, each of the detected decelerations is one of three types: *Early*, *Variable* or *Late*. The program will then find how many *Variable* decelerations were associated with contractions. This number is divided by the total number of contractions detected in the trace. If this fraction R_v^D exceeds a threshold $\theta_R^D(= 0.5)$, the program outputs a decision that *Recurrent Variable* decelerations were detected. Similarly, the program decides if *Recurrent Late* or *Recurrent Early* decelerations were detected using the corresponding fractions R_e^D and R_l^D respectively.

Finally, the program needs to decide if any *Prolonged* decelerations were detected. It does this by finding whether any decelerations had total onset-to-return duration L_D greater than the threshold $\theta_{prol}^D(= 120 f_s)$.

2.3 Diagnostic decision flow

Based on the clustering updates obtained from the procedure described in the previous section, one can use the NICHD diagnostic criteria to classify

Algorithm 2 Algorithm to classify decelerations depending on abruptness of FHR decrease and timing of deceleration nadir with respect to associated contraction's peak. $\mathbf{D}$ is a structure containing deceleration information while $\mathbf{U}$ contains onset, peak and return information for contractions associated with this deceleration.

procedure CLASSIFYDECEL($\mathbf{D}, \mathbf{U}$)
    $n_{dip} \leftarrow n_p^{\mathbf{D}} - n_s^{\mathbf{D}}$ ▷ Time to nadir
    $n_{coinc} \leftarrow |n_p^{\mathbf{U}} - n_p^{\mathbf{D}}|$ ▷ Time diff. between peaks
    $n_{dur} \leftarrow L_{\mathbf{D}}$ ▷ Time from decel onset to return
    if $n_{dip} > \theta_n^{dip}$ **then** ▷ If dip is gradual...
        if $n_{coinc} \leq \theta_n^{coinc}$ **then**
            dType $\leftarrow$ "Early"
        else
            dType $\leftarrow$ "Late"
        end if
    else ▷ If dip is abrupt...
        if $\bar{h}_p^{\mathbf{D}} < \theta$ AND $t_{dur} \in [\theta_{dur1}^D, \theta_{dur2}^D]$ **then** ▷ If its a big dip but has normal duration
            dType $\leftarrow$ "Variable"
        end if
    end if
end procedure

a given trace into one of three categories: Category 1 corresponds to *Normal* traces, Category 2 to *Indeterminate*, and Category 3 to *Abnormal*. From clinical perspective, detection of abnormalities is of paramount importance, followed by Category 2 (where there may be some evidence of compromise but not convincing enough) and then Category 1. This is the order of decision making in the program version as well.

2.3.1 Category 3 conditions

A trace is diagnosed as Category 3 when the following conditions are satisfied:

1. *Absent* baseline variability AND
   - Any *Recurrent Variable* OR *Recurrent Late* decelerations OR
   - Baseline Rate is *Bradycardia*.

2.3.2 Category 2 conditions

If the above symptom-combinations are not present, a check for category 2 conditions is done. When any one or more of the following conditions are satisfied, the tracing is categorized as Category 2.

1. Baseline rate is *Bradycardia* AND variability is not *Absent*
2. Baseline rate is *Tachycardia*
3. Baseline variability is *Minimal*
4. Baseline variability is *Absent* AND any *Recurrent* decelerations present
5. Baseline variability is *Marked*
6. Presence of *Recurrent Variable* decelerations AND variability is *Minimal* OR *Moderate*
7. Presence of *Recurrent Late* decelerations AND variability is *Moderate*
8. Presence of *Prolonged* decelerations.

2.3.3 Category 1 conditions

The last check is for Category 1 conditions. When all of the following conditions are satisfied, the tracing is categorized as Category 1:

1. Baseline rate is *Normal*
2. Baseline variability is *Moderate*
3. No *Recurrent Variable* or *Recurrent Late* decelerations detected.

2.4 Preliminary results

		Physician		
		1	2	3
ES	1	13	0	0
	2	3	9	1
	3	0	0	4

Table 2: Confusion matrix for expert system classification of 30 real data sets. 'ES' stands for "Expert System Classification", while 'Physician' denotes the true physician labelling.

Initial testing of this system was performed on a database of 30 20-minute FHR-UP recordings collected from 9 subjects at the Stony Brook University Hospital. All consent and approval guidelines were followed rigorously. Each record was independently labeled by two physicians, and it was observed that for all tracings except one, there were no disagreements in categorization between them. The only record whose diagnosis was disputed was diagnosed as Category 2 by one physician while the other diagnosed it a Category 3. Because of the dispute, it was agreed to take the gold-standard labeling for this record as Category 2. Table 2 shows the confusion matrix for the classification by the program, which shows that 81% of Category 1 recordings were detected as Category 1 while 80% of Category 3 tracings were detected as Category 3 by the program. Although the system performs well for this small database of records, the accuracies were found to be markedly lower for larger testing sets. In particular, when we compared the ES classification decisions to those made by more objective

fetal health metrics such as umbilical cord pH, we found large error rates. Details of these evaluations are provided in Chapter 5.

In the sequel, we describe how the features described here can be used in probabilistic classifiers. In particular, we use Bayesian networks to encode expert-guided relationships between discretized versions of the features, to construct a factored representation of the joint probability distribution and to use the same for probabilistic inference and hard-decision classification. Bayesian network structure detection can also be used to find whether the expert-guided structure is actually correct. That is, we attempt to answer the question: does the observed data support the hypothesis that any two features are probabilistically dependent? This can be hugely important from the point of view of continuous learning; the more data one collects, the higher the possibility of encountering never-before seen dependencies or independences. Structure detection offers us a systematic and intuitively sensible way to find these. Finally, we use segmented versions of the NICHD features to describe entire feature sequences, which are then modeled as specific observation instances of generative mixture models. These can also be used in both supervised and unsupervised settings to group fetal data in clinically useful ways.

3 Bayesian network classifiers

In this chapter, we describe a Bayesian network (BN) formulation to integrate the features from our expert system (ES) into a probabilistic framework. A BN [51] is a specific type of graphical model in which known (or hypothesized) causal relationships between nodes can be represented as conditional probability relationships. Edges between the nodes can be endowed with directions representing the flow of information from one node to the other. One well known use of BNs for medical diagnosis is the Quick Medical Reference (QMR) system [60], which has more than 4000 observable nodes and 600 unobservable nodes representing the presence or absence of specific diseases and their symptoms.

The first BN we describe is derived from expert guidance, i.e., from interpreting the NICHD consensus guidelines for FHR interpretation [66] described in Chapter 2. However, this may not be the best choice in terms of classification performance or elicitation of the most relevant causal dependencies. In general, in addition to improving the classification accuracy, we are also interested in enhancing our understanding of relationships between the different FHR-UP features. Are there correlations or independencies between given pairs of features that are actually supported by evidence in the observed time-series data? Is it possible to answer these questions in a systematic way and within a probabilistic and Bayesian framework? Answering these questions is crucial to the idea of continuous learning: the more data the system acquires, the more patterns it learns, and at least in theory, the more possibilities there are to improve classification performance. To this end, we use the well-studied approach of Bayesian network structure detection for establishing the presence of relationships between FHR-UP features.

There exists a very rich literature on the problem of efficient Bayesian structure learning methods, such as those developed in [20] (K2 algorithm) and in [38] (Markov chain Monte Carlo methods). For the current problem, we have used the K2 method to get an accurate representation of the probabilistic dependencies between FHR features described earlier, based on real-data evidence. Using this learnt structure, one can learn conditional probability table (CPT) parameters using traditional Bayesian or maximum likelihood techniques. Then, one can also get the posterior probabilities of the "class" variable conditioned on the instantiations of the attribute variables. This can be used as the classifier decision function. Though there exist many other sophisticated structure learning techniques for such applications, our purpose here is not a comparison of these techniques, but a demonstration that using such algorithms can tell us more about real FHR data than just expert guidance.

In the sequel, in Section 3.1.1 we present the K2 structure learning algorithm, in Section 3.1.2, the ES features, and in Section 3.1.3, the classification procedure. Results of classification performance using Leave-One-Out (LOO) procedure are provided in Section 3.3. We conclude the paper with a discussion of the results, and possible future work in Section 3.4. A list of the symbol definitions used in this chapter is provided in Table 3

3.1 Bayesian network formulation

The BN consists of two sets: a set of nodes U and a set of edges E. For the current application, we denote the N features extracted from the data set as random variables X_i, with $i \in \{1, 2, \ldots, N-1\}$. The set $U = \{X_1, \ldots, X_N\}$, contains the random variables representing the features and an additional variable representing the "true" fetal state (equivalent to the "class" variable) as labeled by a physician or some other objective diagnostic procedure. We refer to these variables as *nodes* of the graph. For full description of the graph, we also need the set of directed *edges* E, which represent conditional dependencies between the nodes. Thus, the graph is formally denoted by $G = (U, E)$. For every pair of variables connected as $Y \rightarrow X$, Y is called

Table 3: List of symbols and definitions used in this chapter

Symbol	Definition
U	Set of nodes in BN
E	Set of edges in BN
G	Random variable representing the BN structure $= (U, E)$
$\hat{G}$	Estimated BN structure
X_i	i-th node in BN = i-th feature
$x_{i,k}$	k-th possible instantiation of feature X_i
r_i	Number of possible instantiations of feature X_i
$\mathbf{Y}_i^G$	Set of parents of X_i in graph G
$\mathbf{y}_k^G$	k-th possible joint instantiation of the set of variables $\mathbf{Y}_i^G$
q_i^G	Number of possible joint instantiations of the set $\mathbf{Y}_i^G$
$\gamma(G)$	Score for the graph structure G
Φ	Input node ordering to the K2 structure learning algorithm
$\mathbb{D}$	Set of observed feature vectors
α_{ijk}^G	Prior psueudocount of the event $\{X_i = x_{i,k}\} \cap \{\mathbf{Y}_i^G = \mathbf{y}_{i,j}^G\}$
c_{ijk}^G	Number of observed instances of the event $\{X_i = x_{i,k}\} \cap \{\mathbf{Y}_i^G = \mathbf{y}_{i,j}^G\}$
N_{ij}^G	Prior pseudocount of the event $\{\mathbf{Y}_i^G = \mathbf{y}_{i,j}^G\} = \sum_{k=1}^{r_i} \alpha_{ijk}^G$
M_{ij}^G	Number of observed instances of the event $\{\mathbf{Y}_i^G = \mathbf{y}_{i,j}^G\} = \sum_{k=1}^{r_i} \alpha_{ijk}^G$

the *parent* of X, and X is the *child* of Y. Each feature X_i has a range of instantiations $\{x_{i,1}, \ldots, x_{i,k}, \ldots, x_{i,r_i}\}$, where r_i is the number of possible instantiations for that variable. For graph structure G, the set of parents of node X_i is denoted $\mathbf{Y}_i^G$ and this collection of variables can take values from the set $\{\mathbf{y}_{i,1}^G, \ldots, \mathbf{y}_{i,j}^G, \ldots, \mathbf{y}_{i,q_i^G}^G\}$, where q_i^G is the product of the cardinalities of all the variables in $\mathbf{Y}_i^G$.

The advantage of using BNs stems from the fact that given a specific BN structure, the joint probability distribution over all the nodes of the graph factorizes as

$$P(X_1, \ldots, X_N|G) = \prod_{i=1}^{N} P(X_i|\mathbf{Y}_i^G). \tag{1}$$

This enables very efficient characterization of the probability distribution of the features since the number of parameters can be greatly reduced when the conditional independencies encoded in the BN structure are taken into account. In addition, structure elicitation presents a precise method of detection of (possibly) causal dependencies between the various variables represented on the graph.

3.1.1 BN structure learning

It is well known that learning the structure of any general BN is NP-hard since the number of possible structures increases super exponentially with the number of nodes in the network. Thus, a variety of search heuristics have been developed to address this problem. We focus on the well-known K2 algorithm from [20].

The K2 method is a greedy hill climbing method for searching the space of directed acyclic graphs (DAGs). The method is constrained by a user-input ordering of the nodes, and it maximizes a chosen scoring metric γ (described below) that captures how well the DAGs represent the observed datasets. The input node-ordering Φ reflects the user knowledge of the total node ordering and not individual subgraph structure.

Initially the algorithm assumes that none of the nodes have any parents and calculates an "empty" score $\gamma(G)$ for the graph G having no edges.

Thereafter, for every node $X_i \in \Phi$, the algorithm searches for the single best parent Y from the set Φ_{-i} (consisting of all nodes preceding X_i in the total ordering) that, when connected to the node X_i, provides the greatest increase in γ. If it finds no such parent, it stops and goes to the next node in Φ. Otherwise, it (a) updates the graph G to have the edge $Y \rightarrow X_i$ and (b) restarts the search for other possible X_i parents. The procedure is repeated until all nodes have been explored. Since the total ordering is provided, the algorithm does not need to check for graph acyclicity at each step. We note that the cost of using this efficient heuristic is that the learnt BN structure is significantly influenced by the choice of topological ordering, and thus susceptible to bias. However, since the purpose is not to replace expert guidance but merely to refine it, we feel the benefits of the algorithm outweigh the costs.

There are several choices for the score function as reviewed in [38]. Since we used Dirichlet priors for the discrete features, we employed the Bayesian scoring criterion $\gamma(G) = P(\mathbb{D}|G)$ [74]:

$$P(\mathbb{D}|G) = \prod_{i=1}^{N} \prod_{j=1}^{q_i^G} \frac{\Gamma(N_{ij}^G)}{\Gamma(N_{ij}^G + M_{ij}^G)} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}^G + c_{ijk}^G)}{\Gamma(\alpha_{ijk}^G)}, \tag{2}$$

where $\mathbb{D}$ is the set of observed data values, α_{ijk}^G is the Dirichlet parameter associated with the event $\{X_i = x_{i,k} | \mathbf{Y}_i^G = \mathbf{y}_{i,j}^G\}$, $N_{ij}^G = \sum_{k=1}^{r_i} \alpha_{ijk}^G$, c_{ijk}^G is the number of data cases in which node X_i takes the value $x_{i,k}$ and X_i's parents take the value $\mathbf{y}_{i,j}^G$, and $M_{ij}^G = \sum_{k=1}^{r_i} c_{ijk}^G$. Since this scoring criterion decomposes into local frequency computations for each node, it is computationally quite efficient.

3.1.2 FHR Features

In order to stay close to physician guidelines, we restricted our feature set to those features recommended by the standard guidelines [66] as described in Chapter 2. We provide brief descriptions of these features here. In the following list, we represent the FHR features as random variables, and provide the range of possible instantiations for each of them:

1. Baseline FHR $B \in$ {Bradycardia, Normal, Tachycardia},
2. Baseline Variability $V \in$ {Absent, Minimal, Moderate, Marked},
3. Presence of Accelerations $A \in$ {No, Yes},
4. Presence of Decelerations $D \in$ {No, Yes},
5. Presence of Recurrent Decelerations $D_r \in$ {No, Yes},
6. Presence of Early Decelerations $D_e \in$ {No, Yes},
7. Presence of Late Decelerations $D_l \in$ {No, Yes},
8. Presence of Variable Decelerations $D_v \in$ {No, Yes},
9. Presence of Prolonged Decelerations $D_p \in$ {No, Yes},
10. Presence of Recurrent Early Decelerations $D_{re} \in$ {No, Yes},
11. Presence of Recurrent Late Decelerations $D_{rl} \in$ {No, Yes}, and
12. Presence of Recurrent Variable Decelerations $D_{rv} \in$ {No, Yes}.

The random variable S, which corresponds to the fetal state, can take values from the set $\{1, 2, 3\}$, which correspond to the subjective assessments $\{Normal, Indeterminate, Abnormal\}$.

The *expert-guided* BN structure is provided in Figure 13. The fetal status S is assumed to have a direct causal effect on *all* the "symptom" variables $\{B, V, A, \ldots, D_{rl}, D_{rv}\}$. The features B, V, A and D are pairwise conditionally independent given knowledge of S. In addition, we have 8 more variables representing deceleration types that are directly dependent on the presence of decelerations. For instance, if D takes the value "No", then all the other deceleration variables have to take the value "No"; however, if $D =$ "Yes", then it is not necessary that, say, recurrent decelerations are also present. The directed edge $D \rightarrow D_r$ encodes this intuitive notion.

3.1.3 FHR classification using BN

The CPTs for each parent-child pair are learnt from the training set of feature instantiations using maximum likelihood frequency updates. For any test data set d_i whose status variable S is unknown, we can derive the marginal posterior distribution for the S_i conditioned on knowledge of the features using simple sum-product techniques [37]. A maximum a-posteriori (MAP) criterion is used to find the S_i instantiation $s_i \in \{1, 2, 3\}$ having the highest

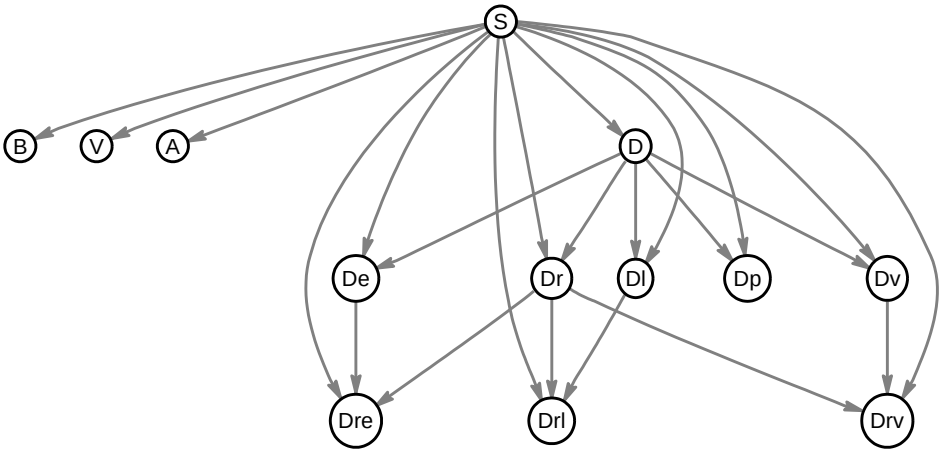

Figure 13: Expert-guided BN structure for categorization of FHR features. The fetal status S has a direct causal effect on all the other variables.

marginal probability mass. This is taken to be the classifier output for d_i.

3.2 Data

The program was tested on a database of 830 20-minute FHR records collected from 9 subjects during the antepartum period at the Stony Brook University Medical Center. All consent and approval guidelines were followed rigorously. The FHRs were continuously monitored using the Doppler technique via GE Corometrics devices. The usual method for extraction of FHR from the Doppler signal is to use autocorrelation functions to detect the periodic movements of the heart valves. Although this does impose some restrictions on the detection of short-term variability, for our purposes it was deemed to have sufficient resolution for effective tracing characterization.

Prior to carrying out feature extraction, FHR preprocessing was done to remove various artifacts including ones due to movement as described

in [25]. Each record was independently labeled as category 1, 2 or 3 by a physician who had access only to the raw and preprocessed noise-free versions of the FHR. It was observed that for some files heavily corrupted by tracing noise and for those dominated entirely by episodic variations, our ES would not be able to get values of certain features like variability or episode locations. For this study, we ignored such data sets from the training and testing procedures. We were then left with 754 out of 830 20-min traces from the original database.

3.3 Results

Classification performance was analysed using Leave-One-Out (LOO) procedures, *i.e.*, for each data set d_i in the record database $\mathbb{D}$, we learn the CPTs from the database $\mathbb{D}_{-i}$ consisting of all data sets except d_i. However, structure learning was done using the entire data set $\mathbb{D}$ in order to ensure that we obtained exactly one learnt structure $\widehat{G}$ (as opposed to $|\mathbb{D}_{-i}|$ different structures) to compare against the expert-guided network G.

We first present the result of structure learning using the K2 algorithm on the 754-strong database in Fig. 14. The differences between $\widehat{G}$ and G (Fig. 13) are discussed further in Section 3.4. The total number of edges in $\widehat{G}$ is 16, as opposed to 23 in G. It was seen that for this particular database, the S node has a causal effect only on $\{V, D, A\}$ and the node B representing the average baseline value for the record is not connected to any other node in the network. As a result of the reduction in the number of edges, the total number of independent conditional probability distribution (CPD) parameters decreased from 89 for G to 60 (or 59 if we ignore the B node from the structure entirely) in $\widehat{G}$. In Table 4, we present confusion matrices for classifier performance when using posterior probabilities calculated by the expert and K2 BNs. Both networks yield similar performances ($\approx$80% sensitivity and $\approx$ 60% specificity, with classifier outputs of categories 2 or 3 treated as "positive" detections).

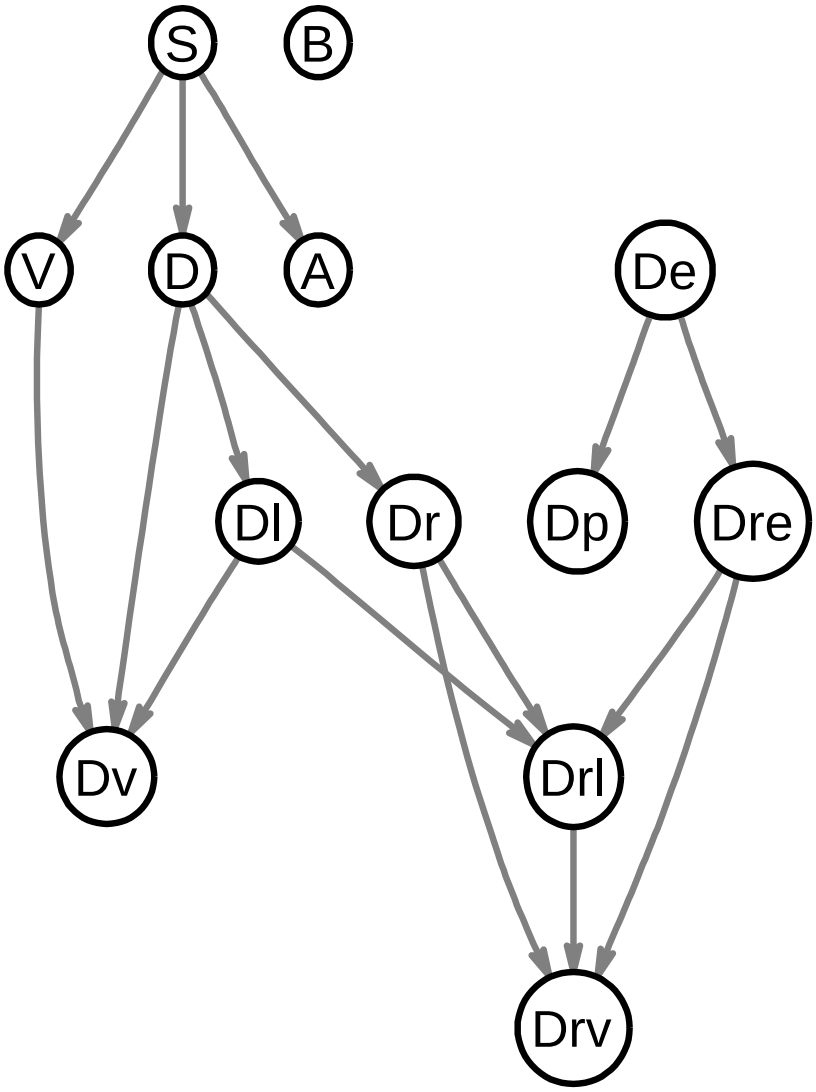

Figure 14: BN structure learnt from the K2 algorithm. The fetal status S has a direct causal effect on only $\{V, D, A\}$.

Table 4: Confusion matrix for BN classification of 754 real data sets.

		Expert BN			K2 BN		
		1	2	3	1	2	3
Physician	1	182	118	0	179	121	0
Labeling ↓	2	91	356	1	91	356	1
	3	5	1	0	5	1	0

3.4 Discussion

Obstetric care providers use the Doppler ultrasound monitors to get continuous recordings of FHR. Standardized clinical guidelines are used to interpret patterns of specific morphological features in the FHR signal such as decelerations (abrupt or gradual decreases in heart rate) or loss of variability (variation around a "baseline" FHR signal). These are explained in detail in Chapter 2. In this study we developed a method to (a) incorporate these features in a BN formulation, (b) to learn network structure from a given set of observed data, and (c) to measure classification performance using posterior probabilities. Although BN structure learning has been widely used in diverse fields such as fault diagnosis, image processing, and medical diagnosis, to our knowledge, this is its first application specific to FHR. The K2 structure learning technique reduces the redundancy in the graph and the total number of CPD parameters, while maintaining the same level of classification accuracy. This is an advantage in terms of efficiency, and it suggests that parameter learning from new data sets using the learnt structure may be more robust.

Prior to structure learning, we had prediscretized the FHR features. Although in principle, continuous features can provide better feature resolution, we worked with discrete features for several reasons including the facts that (a) our feature discretization [25] is very similar to clinical feature definitions as described in [66] and routinely used in obstetric care centers,

(b) using continuous features in the BN requires the introduction of (possibly non-Gaussian) parametric continuous distributions, necessitating the learning of many more hyperparameters, and (c) structure learning with continuous features is considerably more difficult, especially when the data lack diversity.

With the proposed approach, we are able to learn new correlations (or the lack thereof) present in the data. In Fig. 14, one can see that the variable B has no parents or children. Indeed, it was seen that for this database, the vast majority of FHR recordings (741 out of 754) had Normal baseline FHR (between 110 and 160 bpm), only 12 had tachycardic baseline (greater than 160 bpm), while only one data set had bradycardic baseline. This implies that for nearly all possible instantiations of its possible child-nodes, the baseline B node takes the same value; thus, the CPD remains indifferent to the value of B. Similar arguments in the case of the nodes D_e and D_p justify their disconnection from the "class" variable S. Another prominent difference is the inclusion of new edges $D_{rl} \rightarrow D_{rv}$ and $D_l \rightarrow D_v$, which suggests that there are strong correlations between the existence of late and variable decelerations. Moreover, in $\widehat{G}$ the "class variable" S is only connected to the variables $\{V, D, A\}$. This set is also S's Markov blanket, suggesting that for classification via evidential reasoning, it may only be required to look at $\{V, D, A\}$ instead of the entire gamut of possible morphological "symptoms". However, this needs to be tested with more datasets.

4 Density estimation classifiers

In Chapter 2, we showed how standardized guidelines on clinical interpretation of FHR patterns may be translated into features useful for automatic diagnosis using an expert system. However, such features may not be able to capture all the information available in the two input signals. In this chapter, we describe development of such new features. In principle, one can use these novel features in conjunction with ES features for potentially more useful classification performance. We also describe how we selected a group of informative features using hypothesis testing methods and a novel nonparametric classifier that can use these features as input, and improve accuracy compared to support vector machines (SVM) and rule-based systems.

We describe in Section 4.1, a precise formulation for the classification problem under consideration. In Section 4.2, we describe how we segmented our FHR data to isolate regions that display characteristics of reassuring and non-reassuring fetal states, using annotations from independent visual interpretation by experts. This was done to get a large training and testing database for crossvalidation purposes. The features extracted from the raw and transformed time-series data are then described in Section 4.3. Section 4.4 describes our proposed idea for nonparametric window-based classification procedure, which we term neighbor-counting. Details of the performance evaluations and comparisons with the SVM and rule based systems are provided in Section 4.5, followed by a discussion of the pros and cons of this method and future outlook in Section 4.6.

Table 5: List of symbols and definitions used in this chapter. Specific definitions of the individual features are provided in Section 4.3.

Symbol	Definition
y_n	FHR at sample n
$\tilde{y}_n$	n-th sample value of de-baselined FHR = $y_n - b_n$
r_n	FHR return series value at sample n
$\mathbf{x}_i$	Feature vector extracted from i-th FHR time series $\mathbf{y}_i = \{y_{i,1}, \ldots, y_{d_i}\}$
m	Number of features in feature vector
N	Number of training vectors
$N_+(N_-)$	Number of training vectors in the abnormal (normal) class
C_i	Class variable for i-th dataset
b_n	FHR baseline value at sample n
W	Hypervolume of the "immediate vicinity" of a test point in feature space
V	Hypervolume of the total support of the feature space
t_j	Width of window of immediate vicinity in a direction along the axis of the j-th feature.
S_j	Width of entire support of the j-th feature
μ	Window width parameter
ρ	Threshold parameter for the maximum a-posterior decision function

4.1 Problem formulation

We concentrate for now on binary classification of data (e.g., "Normal" vs. "Abnormal") in the supervised case. The FHR signal is denoted as y_n, where n is the sample number. The features in the database may be viewed as points in m-dimensional feature-space and will be denoted as $\mathbf{x}_i; i \in \{1, 2, \ldots, N\}$, where N is the number of training vectors. For each $\mathbf{x}_i$ we have a corresponding true label C_i which can take values from $\{+1, -1\}$. Training involves finding a function $f(\mathbf{x})$ that is able to separate the features in the two categories with minimum overall cost.

In the following, we describe first the features extracted from the data, including features from the so-called FHR "return" series. Then we outline the proposed non-parametric approach to binary classification via a density estimation method. SVM's and parametric Bayesian methods based on Gaussian assumptions are very well studied and are thus not described here. In the Results section, we demonstrate that these can be used to classify short segments of data in an efficient way.

4.2 Data segmentation

Our real database consists of 580 short 15-s epochs of fetal heart rate time series data extracted from clinical recordings of 11 different subjects. The original 20-min Doppler FHR segments, collected in the Department of Obstetric/Gynaecology at Stony Brook University hospital at 4Hz sampling rate, were labeled as normal or abnormal by independent physicians. Out of these 11 recordings, regions of FHR patterns were carefully isolated and labeled as such. Each labeled region was segmented into non-overlapping 15-s epochs. Abnormalities may include decelerating heart rate or low variability, and an epoch was labeled as "abnormal"(the target class, denoted +) if it had either one or both of these attributes. Figure 15 shows a comparison of FHR regions from traces of two different patients that showed the characteristics typical of reassuring and nonreassuring (abnormal) fetal health. In particular, note the presence of several accelerations in the FHR series

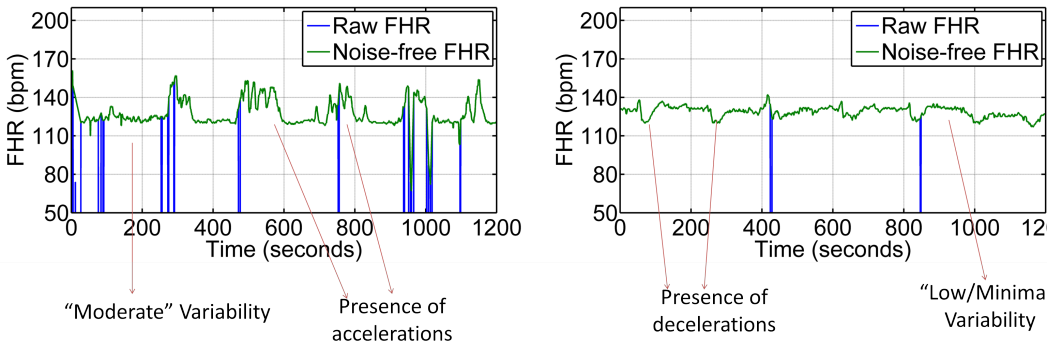

Figure 15: Examples of (left) reassuring and (right) non-reassuring FHR traces showing differences in FHR variability and types of accelerations and decelerations.

on the left and the absence of decelerations. In addition, for regions free of these episodic variations, we see that the variability of the FHR around the baseline (DC) component is relatively high compared to the series on the right. On the other hand, we see much reduced variability in the FHR series on the right, along with the presence of decelerations that are *variable* and *recurrent* according to the NICHD consensus definitions as explained in Chapter 2.

Each epoch was denoised using an algorithm similar to [4]. FHR baseline detection was performed using a median-filtering algorithm. We made no differentiation between recordings taken at different gestational ages or stages of delivery. From each 15-s (60 sample) epoch of data denoted $\mathbf{y} = [y_1, h_2, \ldots, y_{60}]$, several different types of features can be extracted. We focus here on ten specific features as described in the sequel.

4.3 Features

4.3.1 From the raw FHR series

From the raw FHR series, we can extract features to quantify the average time the FHR decelerates or accelerates. We consider the following features:

1. Number of FHR samples out of 60, that are above an acceleration threshold ξ_U.
2. Number of FHR samples out of 60 that are below a deceleration threshold ξ_L. Both acceleration and deceleration thresholds were set according to the prescribed guidelines set by the NICHD [66].
3. Standard deviation of the FHR series σ_y.

4.3.2 Features from the FHR return series

The return series r_n is computed as a time series of sample-to-sample percentage changes in the FHR signal y_n. In order to standardize the range of the possible r_n signals, we first center the original signal y_n around a reasonable constant FHR value (here 140 bpm) and then calculate the baseline (b_n) of the centered signal ($\hat{y}_n$). The "unbiased" FHR signal $\tilde{y}_n$ is then obtained by subtracting b_n from y_n. The return series is then obtained for $\tilde{y}_n$ according to:

$$r_n = \frac{\tilde{y}_n - \tilde{y}_{n-1}}{\tilde{y}_n}. \tag{3}$$

From the return series $\mathbf{r} = [r_1, \ldots, r_{60}]$, we can obtain several features. In a separate, as yet unpublished study, we explored the class-separation performance of several different types of features when used individually. These included several statistical moments (since they can quantify the r_n probability distributions) as well as nonlinear features which have been used previously in adult heart rate variability studies. Hypothesis testing was done via the Kolmogorov-Smirnov test. The statistics for the following 6 features were found to be significantly different for the two classes at the $p = 0.05$ level:

1. Total return $S_r = \sum_{n=1}^{60} r_n$.
2. Variance of return data σ_r^2.
3. Skewness of return data $\gamma_r = (E(r - \mu_r)^3)/\sigma_r^3$, where $\mu_r =$ mean of the return values.
4. Kurtosis of return data $K_r = (E(r - \mu_r)^4)/\sigma_r^4$.

5. Runs ratio of return data ρ_r. This is the number of distinct runs of consecutive increases or decreases of the return series from zero. For instance, the sequence $\{+,+,-,-,+,-,+\}$ has five runs. A higher number of runs indicates higher variability.
6. Shannon Entropy of return data ϵ_r. This feature summarizes the complexity in the return series. We find the histogram of the sequence of r_n's in N_b bins, with frequency in bin k denoted p_k, and then compute ($N_b = 16$ in our implementation):

$$\epsilon_r = -\sum_{k=1}^{N_b} \ln(p_k/60)p_k/60. \tag{4}$$

7. Turning Point Ratio of return data τ_r: A sample z_k from any given sequence $\{z_1, \ldots, z_N\}$ is denoted a turning point if the samples z_{k-1} and z_{k+1} are either both greater than or both smaller than z_k. The turning point ratio τ_z is then defined as the ratio of the number of turning points to the length of the sequence N [24]. To obtain τ_r we first map the r_n data so that for a positive (negative) r_n, $r_n^* = +1(-1)$. Then the number of points in r_n^* preceded and succeeded by r^*'s of opposite signs is computed. This number expressed as a fraction of the total number of samples is τ_r. For example, the sequence $\{+,+,-,-,+,-,++\}$, has two turning points, and the TPR is 2/8 = 0.25.

Out of these 10 features, we use m at a time for classification. For each value of m there may be many combinations of features, each yielding a different performance. For training and testing, we used the method of 10-fold cross-validation on the full set of 580 feature sets. We utilized the receiver operating characteristic (ROC) method to get classification performance measures for the best combinations of m features. As a performance metric, we used the area under the ROC curve (AUC).

4.4 Classification

We developed a Bayesian formulation for the pattern classification problem, but without assuming any parametric model for the class-conditional likelihoods. After mapping the training feature sets, the test vector to be classified is mapped into the same feature space. To get an estimate of the target (control) likelihood of the test feature vector, we simply estimate the number of target (control) training samples in the immediate vicinity of the test point and divide this number by the total number of training data from target (control) class $N_+(N_-)$. Figure 16 shows an example of this process. Data points from the target (control) class are represented by the "+"("-") symbol. The test point, denoted by a solid circle, has in its immediate neighborhood, more target points than control points, and would be classified as a positive class (assuming the number class-frequencies in the training set are roughly equal). However, we need to appropriately define the term "immediate vicinity".

The simplest assumption one can make about any feature data is that they arise from a uniform distribution. That is, given any region of hypervolume T in the feature space, the probability that a given feature vector falls in this region is $p = T/V$, where V is the total support hypervolume of the distribution. We define the region as a rectangular cuboid (in m dimensions) whose volume W is directly dependent on the support of each of the feature vectors, as estimated from the training database. If the training feature vector is $\mathbf{x}_i = [x_1^i, x_2^i, \ldots, x_m^i]^T$, for $i \in \{1, \ldots, N\}$, the feature supports and total volume V can be defined as

$$S_j = \max(x_j^1, \ldots, x_j^N) - \min(x_j^1, \ldots, x_j^N); \quad j = 1, \ldots, m \tag{5}$$

$$V = \prod_{j=1}^{m} S_j. \tag{6}$$

Assuming the window of immediate vicinity has width t_j in a direction along the axis of the j-th feature, the corresponding hypervolume is $T = \prod_{j=1}^{m} t_j$. We assume each t_j is directly proportional to the support S_j of the corresponding feature. Given N training vectors, the average number of

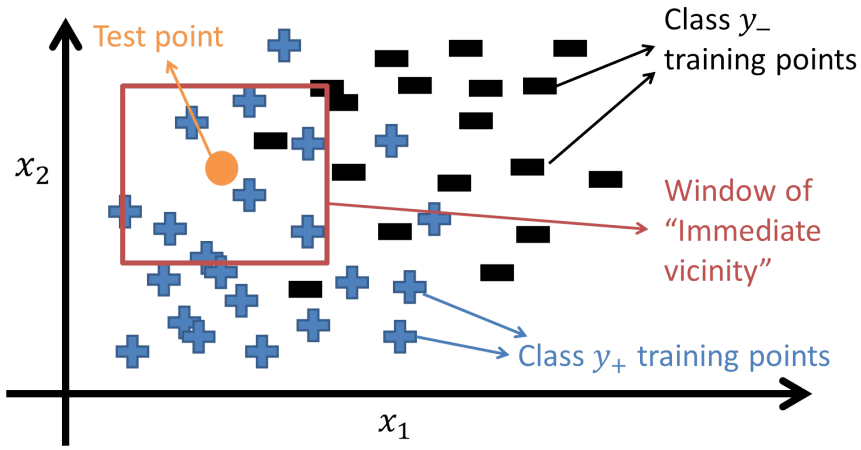

Figure 16: A graphical demonstration of the neighbour counting method of classification. There are two input features x_1 and x_2. Feature vectors in the positive and negative classes are shown with symbols "+" and "-". The box around the test point (solid circle) denotes its "immediate vicinity". Based on the counts of the training vectors in the two categories, this test point would be classified as belonging to the positive class.

training vectors mapping inside T is Np. Thus, given some choice of μ, we can calculate the widths as

$$t_j = \gamma S_j \tag{7}$$

$$= S_j \left(\frac{\mu}{N - k\mu} \right)^{1/m}, \quad \forall j \in \{1, 2, \ldots, m\}, \tag{8}$$

where μ is simply a convenient way to define the window widths for estimating the class-conditional probability of the feature vector, and $0 < k < (1-p)/p$. Given these widths and knowledge of the prior class probability $P(C_i)$, we can define the class-conditional and posterior probabilities for some unlabeled test feature vector $\mathbf{x} = [x_1, \ldots, x_m]^T$ as

$$P(\mathbf{x}|C = c) = K_x / K_c, \tag{9}$$

$$P(C = c|\mathbf{x}) \propto P(\mathbf{x}|C = c)P(C = c); \quad c \in \{+1, -1\} \tag{10}$$

where K_x denotes the number of training vectors $\mathbf{x}_i$ in class c satisfying $|x_j - x_j^i| \leq t_j, \forall j$, and K_c denotes the total number of training vectors in

class c. A threshold dependent decision function can now be defined as follows:

$$f_{NC}(\mathbf{x};\mu,\rho) = \begin{cases} +1, & \frac{P(C=+1|\mathbf{x},\mu)}{P(C=-1|\mathbf{x},\mu)} - \rho > 0. \\ -1, & \text{otherwise} \end{cases} . \tag{11}$$

4.5 Results

Empirical performance analysis was done across all possible feature combinations to find good feature sets. The area under the receiver operating characteristic (ROC) curve, also called AUC, was found for each decision function by varying the corresponding $\boldsymbol{\theta}$ (parameters of the decision function) in a way that the full ranges of sensitivity and specificity are explored. It is denoted $A_{\boldsymbol{\theta}}$. It has been established that the AUC metric is an estimate of the total probability that the value of the decision function of a randomly chosen feature vector from the target class is greater than that for a randomly chosen feature vector from the control class [14]. Thus, the higher the AUC value, the better the method performs (in an average sense).

For the SVM method, we kept the radial basis function scaling factor fixed at 1 and varied the box constraint for soft margins, i.e., $\boldsymbol{\theta} = C$. When using the Bayesian method with Gaussian assumption (n-dimensional Gaussian (NDG) method), we varied the likelihood ratio threshold, i.e., $\boldsymbol{\theta} = \rho$. For the neighbour counting method f_{NC}, we had to vary two parameters, μ and the likelihood ratio threshold, i.e., $\boldsymbol{\theta} = [\mu, \rho]$. However, the analysis was further complicated by the fact that there were a number of possible feature combinations to explore for each value of m. For instance, for $m = 2$, we had a total of $\binom{10}{2} = 45$ feature combinations to sift through in finding the best performance.

We provide ROC curves for two different feature sets of lengths $m = 2$ ($\mathbf{x} = \{\xi_U, \sigma_f\}$) and $m = 9$ ($\mathbf{x} = \{\xi_L, \xi_U, \sigma_f, S_r, \epsilon_r, \rho_r, \sigma_r^2, \tau_r, K_r\}$), respectively. These two feature combinations were found to give good classification performance using all three methods. The corresponding receiver operating characteristic curves are shown in Figure 17.

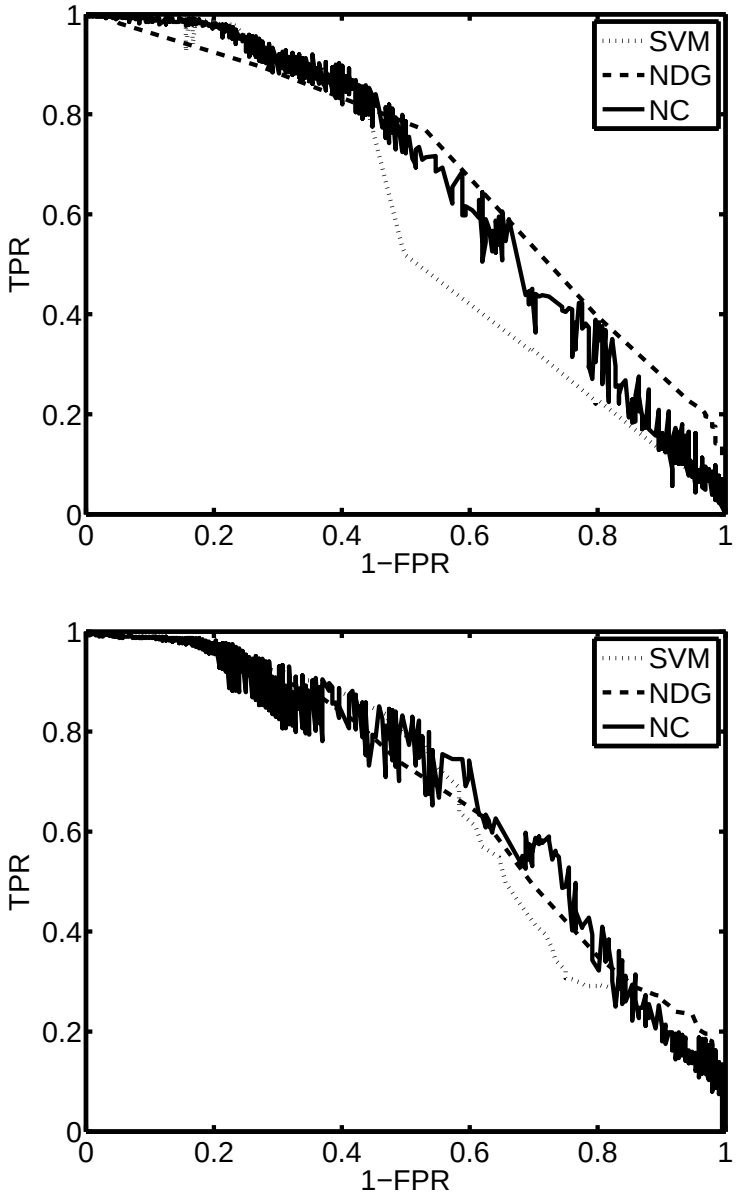

Figure 17: Receiver operation characteristic curve for the m-feature combinations from analysis of FHR data for all 3 methods, with (top) $m = 2$ and (bottom) $m = 9$. TPR = True positive rate (sensitivity); FPR = False Positive Rate.

Table 6: Comparison of classification performance. Higher $A_{\boldsymbol{\theta}}$ values imply better average classifier performance.

Method	m	$A_{\boldsymbol{\theta}}$	Best TPR	Best (1-FPR)
SVM	2	0.59	0.85	0.43
NDG	2	0.60	0.77	0.53
NC	2	0.66	0.69	0.59
SVM	9	0.66	0.78	0.53
NDG	9	0.68	0.64	0.61
NC	9	0.68	0.74	0.60

Table 6 summarizes the results. TPR denotes True Positive Rate (also called Sensitivity) while FPR is the False Positive Rate (also called the rate of false alarm). The terms "Best TPR" and "Best $(1-FPR)$" denote the pair of coordinates $(TPR, 1-FPR)$ that maximizes the product $TPR(1-FPR)$.

4.6 Discussion

The above results demonstrate the feasibility of using short epochs for classification of FHR signals. We note that making a final decision on the status of the FHR series incorporates many other factors including the presence of accelerations, variability around the baseline, presence and frequency of contractions (obtained from the uterine contraction signal) along with the use of long segments of cardiotocographic signals. In practice, anywhere between 10-40 minutes of FHR signal may be used by doctors to do a classification. While this has the advantage of utilizing more information, from a signal-processing perspective, it is not very advantageous since the FHR signal, like most biomedical signals, can have significant nonstationarities across long time-scales.

In our proposed approach, we first classify short segments of FHR series, followed by ensemble classification of the sequence of short-classifications. Additionally, we have shown here the possibility of using several features extracted from the FHR return series instead of the raw

FHR. Most of the statistical features such as variance, skew and entropy are usually applied on the raw signal. We can see from the results that for a false-alarm rate of 40%, it is possible to achieve 74% sensitivity using 9 out of the 10 considered features.

When we studied class-separation performance using hypothesis testing methods on individual features, we also analyzed the effect of using segment lengths varying from 10s to 1 minute. It was observed that the features from the raw FHR signal (i.e., ξ_U, ξ_L, σ_h) were significantly separated for all segment lengths. For return features, different results were obtained for different segment lengths. The two nonlinear features were significant separators for all segment lengths, while the statistical moments like total return and variance were significant separators of shorter segment lengths (up to 30s). However, since we were using non-overlapping epochs and a fixed amount of real data, the total number of feature sets was different for different segment lengths (fewer training sets of 1 minute length). In our judgement, it makes sense to use results from bigger training sets as a basis for choosing features, which is why we used 15s epoch lengths.

To further improve the results, we need to study (a) feature extraction from different durations of FHR and (b) a bigger database of supervised training data. However, we note that complete visual annotation of large durations of FHR data segmented into many shorter segments may not be feasible, and thus we also need to develop algorithms for unsupervised or semi-supervised training. In addition, need to include features extracted from the uterine pressure, as input to the classifier.

In the next chapter, we work with sequences of features from short-duration FHR-UP time series. We will try to also address the problem of significant inter- and intra-observer variability in the gold-standard classification by using a dataset that uses an objectively measured umbilical cord blood pH value (after delivery) to assess fetal health. Another common problem in FHR analysis is the extreme rarity of FHR tracings showing true fetal distress (category 3). For instance, in the data used in Chapter 3, category 3 inputs compose only 0.8% of the training set. This lack of diversity

leads to poor performance when classifying category 3 recordings. We address this issue by collecting data from a lot more patients, which increases the number of datasets in each category studied. The primary framework used for classification next is that of GMs and we compare the results to existing discriminative approaches such as SVMs and ES. The results obtained in this chapter and the next indicate that there is value in the use of short-durartion FHR features, and exploiting the unique dynamics of the FHR variations can improve the classification performance. Furthermore, these models can incorporate UP dynamics quite elegantly, and also enable efficient unsupervised clustering.

5 Generative model classifiers

The FHR-UP classification approaches described until now, and also used in the state-of-the-art, extract single values for each feature-value from long-duration datasets. For instance, if the feature we are interested in is "variability", such an approach would involve finding the variability (estimated using, say, a zero-crossing approach described in Chapter 2) for each separate episode-free segment of a record of length 20-minutes or more, and then summarizing it by taking the mean or median of all these values. As a result, they typically do not account for the natural non-stationarity and variation present in FHR dynamics. Even more sophisticated approaches to account for the interactions between FHR and UP signals and the influence of past FHR values on the present, such as the system identification approach in [106], are often unable to find a direct relationship between FHR and UP signals for a large number of records and rely on a linear model for system identification that does not account for nonlinear interactions such as higher frequency variations from sharp decelerations or variability coincident with decelerations.

As discussed in Chapters 2 and 4, and in our previous studies [25–27], we have explored a number of features, including some derived directly from physician's domain knowledge. They include baseline FHR trend and variability, contraction rates, and numbers and types of episodic variations such as accelerations and decelerations. An important characteristic of these signals is their rich dynamic nature and extracting information from it provides additional feature space for discrimination. In this chapter, we propose the idea of partitioning each T-second FHR record into sequences of much shorter (t-second duration) segments and describing each segment

by discrete features including "FHR variability" or "presence of accelerations." Each FHR record can thus be represented by a *sequence* of feature-values, which is then used as an input to a Bayesian classification system capable of finding shared patterns among feature sequences from different records. For modeling the feature sequences, we use the framework of generative models (GMs). Similar ideas have been used previously in several fields including text classification [13,103], image processing [99], representation of electroencephalogram and electrocardiogram data [104], and adult heart rate data [92]. Our method implements learning of the needed probability distributions and follows up with the use of Bayes' rule. Although several state-of-the-art machine learning techniques have already been used for automated FHR-UP classification, the predominant paradigm has been discriminative. Unlike GMs, the discriminative models do not naturally allow for the estimation of a posteriori distributions of all the variables of interest. The GMs are also much more flexible in expressing dependencies among different variables [11], but their implementations are computationally more intensive.

More specifically, with GMs, one can achieve improved detection of new data that are outliers under the model and whose predictions have low accuracy [10]. In our problem, we can view the abnormal traces as outliers because they are simply relatively rare, which further justifies the use of these models. This problem has long been recognized as a significant obstacle to FHR analysis.

Finally, GMs can naturally be used in unsupervised learning [11, 53]. This is very important since most supervised FHR classification methods are constrained by the absence of a reliable gold-standard. Possible reasons for the lack of it are (a) an abnormal tracing may be classified as gold-standard "normal" after interventions had been made to improve fetal status and the criterion for normality is avoiding injury/death [32], and (b) objective markers, like umbilical/arterial pH or Apgar score, can show poor correlation with FHR patterns [81]. The other issue is related to category 2, which is notoriously broad and which in practice includes everything that is deemed not to fall in categories 1 or 3 [19]. Unsupervised classification of

signals that belong to neither category 1 nor 3 may provide information for improved diagnosis, for which GMs are ideally suited.

There are unique challenges to describing fetal data using sequences of features. Some questions we attempt to answer in this chapter include: should one use uniform or data-driven time series segmentation? How can we combine information from the UP signal in this approach? What kind of probabilistic models are best suited for such data? In the sequel, we provide a detailed description of the uniform segmentation of FHR-UP time series, feature extraction and discretization procedure in Section 5.1. We then describe, in Section 5.2 how to use two types of generative models (naïve Bayes and first-order Markov chain) to model sequences of the above features. Section 5.3 provides brief descriptions of the support vector machine and rule-based approaches we compare our GM classifier to, while performance comparisons using stratified crossvalidation are described in Section 5.4. We provide a list of all the symbols used in this chapter and their definitions in Table 7.

Table 7: List of symbols and definitions used in this chapter

Symbol	Definition
y	A given FHR time series
u	A given UP time series
f_s	Sampling rate
Tf_s	Number of samples in the time series input
tf_s	Duration (in samples) of the prespecified segmentation period
x_j	Final feature value for the j-th segment
v_j	Normalized FHR variability in the j-th segment
H_v	Alphabet size of the variability symbol
b	Granularity of the discretization $= 1/H_v$
H_A	Alphabet size of the acceleration symbol = number of unique acceleration types
H_D	Alphabet size of the deceleration symbol = number of unique deceleration types

Table 7 – continued from previous page

H_x	Alphabet size of the final feature = $= H_v + H_A + H_D$
d_i	Number of segments in the i-th time series
C_i	Class variable of the i-th time series
k	k-th possible instantiation of the class variable
K	Total number of possible classes
γ	Set of (known) hyperparameters
π_k	Class k's prior probability mass
$\boldsymbol{\theta}_k$	Pmf of feature x in class k, defined over the support $\{1, 2, \ldots, H_x\}$
$\boldsymbol{\Psi}_k$	Transition probability matrix associated with class k. Each row is denoted $\boldsymbol{\psi}_{k,h} = \{\psi_{k,h,1}, \ldots, \psi_{k,h,H}\}$
$\boldsymbol{\phi}_k$	Initial feature value pmf associated with class k
$\boldsymbol{\alpha}$	Hyperparameter for the Dirichlet distribution prior for the class pmf $\boldsymbol{\pi}$
$\boldsymbol{\lambda}$	H-dimensional hyperparameter vector for the Dirichlet distribution prior for the feature-alphabet pmf $\boldsymbol{\theta}_k, k \in \{1, \ldots, K\}$
$\boldsymbol{\eta}$	H-dimensional hyperparameter vector for the Dirichlet distribution prior for the initial feature value pmf $\boldsymbol{\phi}_k, k \in \{1, \ldots, K\}$
$\boldsymbol{\beta}$	H-dimensional hyperparameter vector for the Dirichlet distribution prior for each row of the transition matrix $\boldsymbol{\psi}_{k,h}, k \in \{1, \ldots, K\}, h \in \{1, \ldots, H\}$
s_k	Number of training records in class k
$q_{k,h}$	Number of occurrences of feature value h in all training records of class k
$q_{*,h,n}$	Number of occurrences of feature value h in n-th feature sequence
$r_{k,h}$	Number of occurrences of feature value h in the first segments of class-k training records
$z_{k,g,h}$	Number of transitions from feature value g to h in all class-k training records

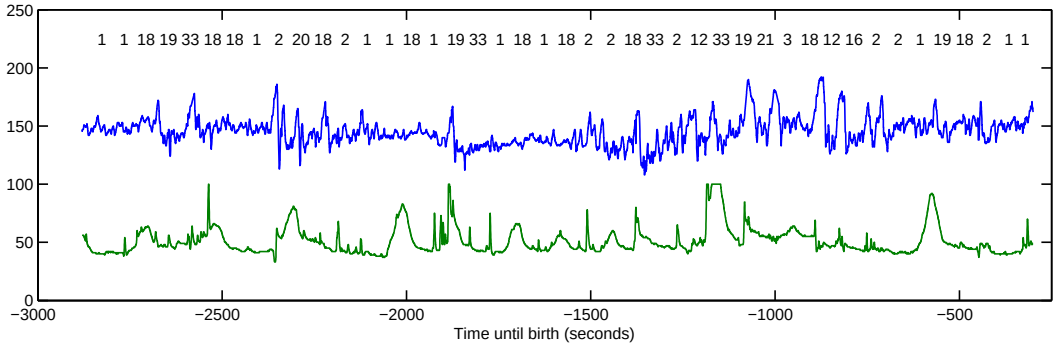

Figure 18: An example of a feature sequence (displayed as the row of symbol numbers at the top) extracted from an FHR (in blue)-UP (in green) record. The size of the alphabet H_x was 34 and the segment length was $t = 60$ seconds. FHR units are beats per minute (bpm) while UP is scaled to percentage values.

Table 7 – continued from previous page

$z_{*,g,h,n}$	Number of transitions from feature value g to h in n-th feature sequence
N_k	Total number of segments in all training records of class k
$\{C, \sigma\}$	Box-constraint and radial basis function kernel scaling factor for SVM soft margin

5.1 Feature extraction

For the proposed GM classification methods, we employ segmented versions of features described in Chapter 2, namely of accelerations, decelerations, and variability. As a reminder, the NICHD consensus guidelines describe general rules for visual assessment of morphological features in the FHR-UP recording, and a guide to mapping numerical features into qualitative "symptoms". Prior to carrying out feature extraction, some preprocessing is performed to remove various artifacts including ones due to movement, and to interpolate over missing beats if the gaps are small. Firstly,

from the UP signal, we detect the presence of maternal contractions using a mode estimation technique. We then estimate the baseline heart rate using a windowed median-filtering algorithm on the FHR time series, with a 5-minute moving window, which is advanced on a sample-to-sample basis. Then, we identify episodic deviations from this baseline and annotate any significant upward (downward) divergence-and-return to baseline as an acceleration (deceleration). For FHR regions free of accelerations and decelerations, we estimate FHR "variability" by calculating the interquartile range of the de-baselined FHR time series. All thresholds are derived from the NICHD guidelines, with minor tuning performed using a small subset of FHR records as described in Chapter 2.

We now describe the process of segmenting the dataset used, so as to enable the extraction of sequences of feature values indicative of the morphological changes in it, and the subsequent discretization of the features into a finite-sized "feature alphabet". The latter step is performed to enable efficient modeling of the feature sequences using multinomial distributions, which makes the job of parameter update and inference much easier.

5.1.1 Segmentation and feature discretization

Let a given contiguous input time-series pair (FHR and UP) be denoted $\{y, u\}$, and have length Tf_s samples, where f_s is the sampling frequency. In order to make feature sequences amenable to analysis using GMs, we first partition both y and u into synchronized segments of length tf_s samples with no overlap. For each segment (indexed by j), a discretization module assigns a feature value x_j by using an H_x-sized alphabet. We illustrate this process below.

- Let the (normalized) FHR variability of the jth segment be denoted by v_j, where $v_j \in [0, 1]$. If this segment is not classified as an acceleration or deceleration, then we discretize this to a label ξ_j in the following

way:

$$\xi_j = \begin{cases} 1, & v_j \in [0, b), \\ 2, & v_j \in [b, 2b), \\ \dots & \\ H_v, & v_j \in [1-b, 1], \end{cases} \tag{12}$$

where $b = 1/H_v$ is a bin width which controls the granularity of discretization.

- If segment j has at least 50% of samples classified as being part of an acceleration, the label ξ_j takes a value depending on the type of acceleration. Its values are quantified as $\xi_j = H_v + 1$ (normal), $\xi_j = H_v + 2$ (prolonged) or $\xi_j = H_v + 3$ (baseline change). Thus, $H_A = 3$.
- If segment j has at least 50% of samples classified as being part of a deceleration, the label ξ_j takes a value depending on the type of deceleration, that is, $\xi_j = H_v + H_A + 1$ (early), $\xi_j = H_v + H_A + 2$ (late), $\xi_j = H_v + H_A + 3$ (variable), $\xi_j = H_v + H_A + 4$ (no associated contraction) or $\xi_j = H_v + H_A + 5$ (baseline change). Thus, $H_D = 5$.
- We assign the final feature value x_j as follows. If segment j of the UP signal u has at least 50% of samples classified as part of a contraction, then the feature $x_j = \xi_j + H_v + H_A + H_D$; otherwise, $x_j = \xi_j$. Therefore, $H_x = 2(H_v + H_A + H_D)$.

Thus, the two important parameters for controlling the feature resolution are the segmentation period t, and the bin width b. Note that through this procedure we have combined available information from both FHR and UP signals in a single discrete feature type x. An example of a feature sequence is shown in Fig. 18. We recognize here that, unlike other existing FHR feature extraction methods (such as in rule-based systems) which also consider the baseline FHR as a feature, the restriction of one label per segment forces us to choose between the baseline FHR value and the variability for all segments not classified as acceleration/deceleration episodes. We chose to stick to FHR variability instead of baseline because this is recognized as being more informative of fetal health in previous studies, including our own. We denote the feature sequence from the ith FHR record

as $\mathbf{x}_i = \{x_{i,1}, \ldots, x_{i,d_i}\}$, where the second subscript identifies the segment and d_i is the total number of segments in the ith record. We model these feature sequences using GMs as follows.

5.2 Generative model classification

Let each data record be generated by one of K possible classes. Given a set $\mathbb{D}$ of N feature-class pairs $\{\mathbf{x}_i, C_i\}$, our goal is to infer, for some newly observed feature vector $\mathbf{x}_n$, its unobserved (hidden) class $C_n = k$, $k = 1, 2, \cdots, K$. An alternative perspective on the same problem is to get a representation of the posterior probability of the class variables for all the test datasets, $P(C_{1:N}|\mathbf{x}_1, \ldots, \mathbf{x}_N, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ denotes the set of known hyperparameters associated with the class-conditional models. The goal of the former approach is to get a hard classification decision for each dataset, although one can always simultaneously report the probability of the decision as a measure of uncertainty. The latter approach naturally leads us to the use of sampling approaches for unsupervised clustering, very useful when gold-standard labeling is not available. This is dealt with in Section 6.1.

We now describe the two models used to explain the generation of the feature sequences. The idea in supervised classification is to use the training data sequences to infer model parameters of the chosen model for each class, and then use these parameters to approximate the posterior probability mass function of the generating class variable. This can be used for making a classification decision in either unsupervised or supervised settings as shown in the sequel.

5.2.1 Naïve Bayes GM

A common simplifying assumption when studying sequential data is that the probability of a feature instantiation in a certain segment is independent of the feature instantiations in other segments. This is referred to as the *naïve*

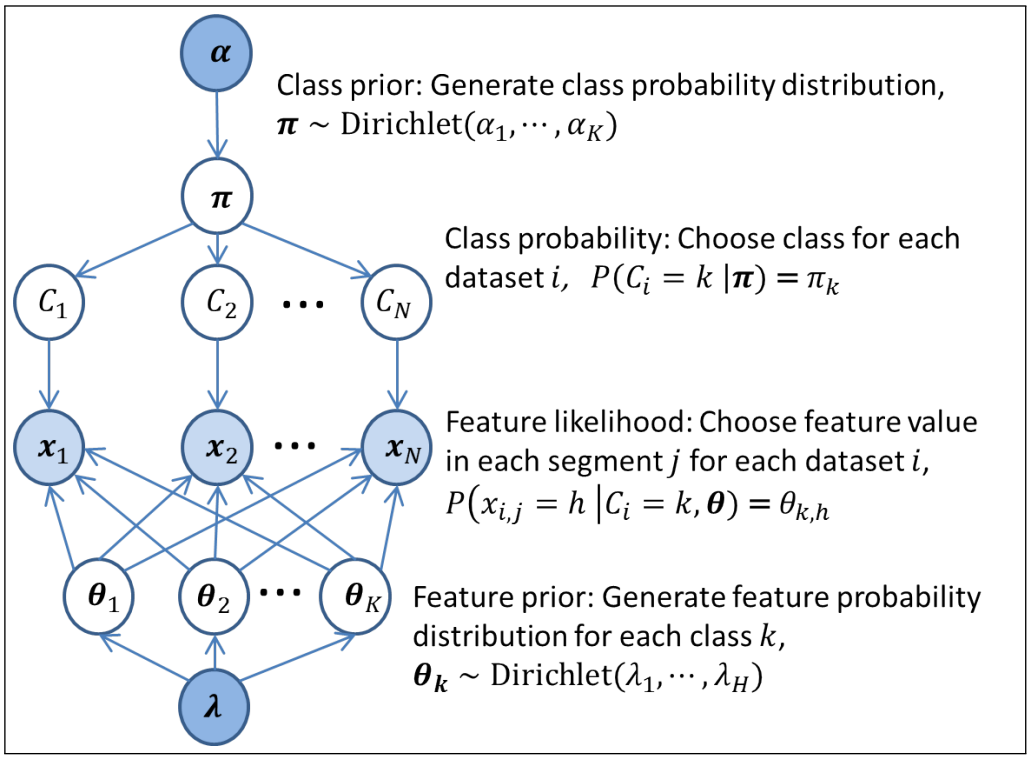

Figure 19: Directed acyclic graph with N data sets and K possible classes of the generative naïve Bayes model.

Bayes assumption. The feature order does not matter, and the feature sequence can be encoded easily via a multinomial distribution. Then, for the ith record, given its class $C_i = k$, an attribute value for each of the d_i segments is generated by the probability distribution $\boldsymbol{\theta}_k = [\theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,H}]$ whose support is the range of possible feature values $1, \ldots, H$.

Next, for the prior of class probability $\boldsymbol{\pi}$ and the distribution parameters $\boldsymbol{\theta}_k$ we use respectively Dirichlet distributions. The reason is that the Dirichlet distributions are conjugate to the multinomial distribution, and thus enable analytical solutions for integrating out parameters when calculating a posteriori probabilities. The respective hyperparameters for these priors are $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ and are assumed known. Thus, the GM for the N data records is (the graphical model is shown in Fig. 19)

1. Generate the class probability $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K] \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$.
2. For each class k, generate $\boldsymbol{\theta}_k = [\theta_{k,1}, \ldots, \theta_{k,H}] \sim \text{Dirichlet}(\lambda_1, \lambda_2, \ldots, \lambda_H)$.
3. For each data record $i = 1, \ldots, N$, draw the class $C_i \sim \text{Categorical}(\boldsymbol{\pi})$, i.e., $P(C_i = k|\boldsymbol{\pi}) = \pi_k$.
4. For each data record $i = 1, \ldots, N$, for each segment $j = 1, \ldots, d_i$, and given $C_i = k$, draw the feature value $x_{i,j}$, $x_{i,j} \sim \text{Categorical}(\boldsymbol{\theta}_k)$, that is, $P(x_{i,j} = h|C_i = k, \boldsymbol{\theta}_k) = \theta_{k,h}$.

Given a corpus of training records $\mathbb{D}$, one can estimate the class and feature probabilities of the model as follows:

$$\hat{\pi}_k \propto (\alpha_k + s_k), \quad \hat{\theta}_{k,h} = \frac{q_{k,h} + \lambda_h}{N_k + \sum\limits_h \lambda_h}, \tag{13}$$

where s_k is the number of training records in class k, $q_{k,h}$ is the number of occurrences of feature value h in all training records of class k, and N_k is the total number of segments, also in all training records of class k. Given the parameter estimates, we can calculate the likelihood of a feature sequence being generated according to the naïve Bayes model as

$$P(\mathbf{x}_n|C_n = k, \hat{\boldsymbol{\gamma}}) = \hat{\pi}_k \prod_{h=1}^{H_x} \hat{\theta}_{k,h}^{q_{*,h,n}}, \tag{14}$$

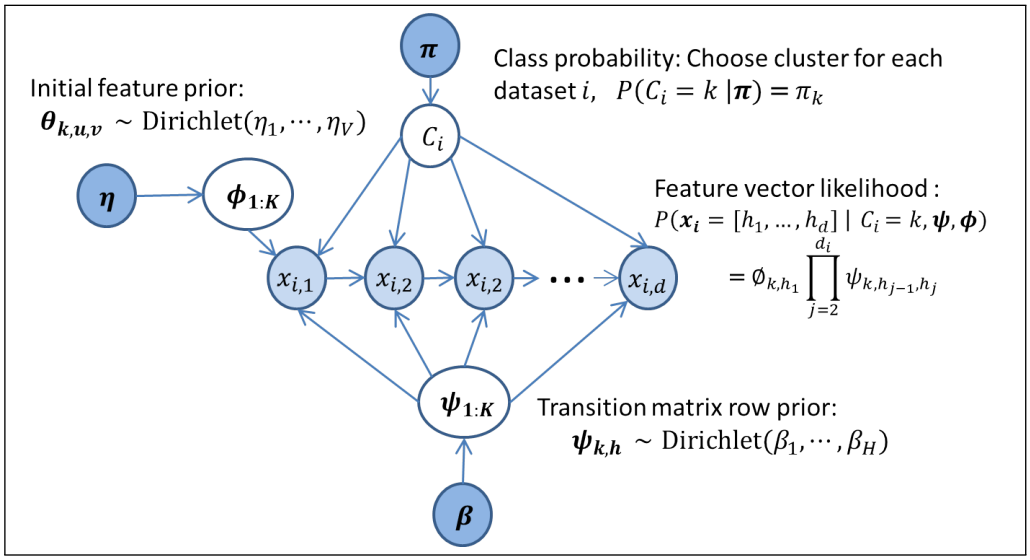

Figure 20: Directed acyclic graph of the generative model with a Markov structure encoding time-dependence in each (ith) FHR-UPrecord.

where $q_{*,h,n}$ denotes the number of occurrences of feature value h in the nth feature sequence.

5.2.2 First order Markov-chain GM

The naïve assumption that the feature values of each segment are independent of the segment index is a considerable leap of faith, since in general, FHR signals (like most real-world time series) exhibit correlations across time. A simple first-order Markov assumption encodes this via the following framework: given knowledge of class $C_i = k$, a feature value for each of the d_i segments of record i, except the first segment, is generated using the $H \times H$ transition probability matrix $\boldsymbol{\Psi}_k$. The feature value of the first segment is generated by a probability distribution $\boldsymbol{\phi}_k$.

Once again, we endow the model with Dirichlet priors for the cluster probability $\boldsymbol{\pi}$ and the parameters $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$ (each row of the transition matrix). The respective hyperparameters for these priors are denoted $\boldsymbol{\alpha}, \boldsymbol{\eta}$ and $\boldsymbol{\beta}$, and are assumed known. Thus, the GM for N data records is (the associated graphical model for the ith record is given in Fig. 20):

1. Generate the class probability $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K] \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$.
2. For each class k, generate the probabilities of initial feature value $\boldsymbol{\phi}_k = [\phi_{k,1}, \ldots, \phi_{k,H}] \sim \text{Dirichlet}(\eta_1, \ldots, \eta_H)$.
3. For each class k and each feature value, generate the feature value transition probabilities $\boldsymbol{\psi}_{k,h} = [\psi_{k,h,1}, \ldots, \psi_{k,h,H}] \sim \text{Dirichlet}(\beta_1, \ldots, \beta_H)$.
4. For each data record $i = 1, \ldots, N$, draw the class $C_i \sim \text{Categorical}(\boldsymbol{\pi})$, *i.e.*, $P(C_i = k|\boldsymbol{\pi}) = \pi_k$.
5. For each data record $i = 1, \ldots, N$, draw the feature sequence $\mathbf{x}_i$ according to: $P(\mathbf{x}_i = [h_1, \ldots, h_{d_i}]|C_i = k, \boldsymbol{\phi}, \boldsymbol{\psi}) = \phi_{k,h_1} \prod_{j=2}^{d_i} \psi_{k,h_{j-1},h_j}$.

Given a corpus of training records $\mathbb{D}$, one can estimate $\boldsymbol{\pi}$ by (13) and $\phi_{k,h}$ and $\psi_{k,g,h}$ as follows:

$$\hat{\phi}_{k,h} = \frac{\eta_h + r_{k,h}}{\sum_h \eta_h + s_k}, \quad \hat{\psi}_{k,g,h} = \frac{\beta_{g,h} + z_{k,g,h}}{\sum_h \beta_{g,h} + q_{k,g}}. \tag{15}$$

Here, $r_{k,h}$ is the number of occurrences of h in the first segments of class-k records in $\mathbb{D}$ and $z_{k,g,h}$ is the number of transitions from g to h in class-k records in $\mathbb{D}$. Given the parameter estimates, we can calculate the likelihood of a feature sequence being generated according to a given category-specific first-order Markov model as

$$P(\mathbf{x}_n = \{x_{n,1}, \ldots, x_{n,d_n}\}|C_n = k, \hat{\boldsymbol{\gamma}}) = \hat{\pi}_k \hat{\phi}_{k,x_{n,1}} \prod_{h=1}^{H_x} \hat{\psi}_{k,g,h}^{z_{*,g,h,n}}, \tag{16}$$

where $z_{*,g,h,n}$ denotes the number of occurrences of feature transitions $g \to h$ in the nth feature sequence.

5.2.3 Maximum a posteriori (MAP) decision

Once the parameters of the assumed model are estimated from $\mathbb{D}$, we can compute the approximated a posteriori probability distribution of the class of the nth test data record given the obtained feature sequence $\mathbf{x}_n$. We have

$$\begin{aligned} P(C_n = k|\mathbf{x}_n, \mathbb{D}) &\propto P(C_n = k|\hat{\boldsymbol{\pi}}) P(\mathbf{x}_n|C_n = k, \hat{\boldsymbol{\gamma}}_k) \\ &\propto (\alpha_k + s_k) P(\mathbf{x}_n|C_n = k, \hat{\boldsymbol{\gamma}}_k), \end{aligned} \tag{17}$$

where $P(\mathbf{x}_n|C_n = k, \hat{\boldsymbol{\gamma}}_k)$ is the likelihood of class k, $\hat{\boldsymbol{\gamma}}_k$ is the set of estimated parameters (e.g., $\{\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\phi}}_k\}$ for the naïve Bayes model), and $P(\mathbf{C}_n = k|\hat{\boldsymbol{\pi}})$ is the prior probability of class k. The likelihood of the feature sequence can be calculated from (14) and (16).

If we make a decision based on the MAP rule, the estimated class is obtained from

$$\hat{C}_n = \arg\max_k P(C_n = k|\mathbf{x}_n = \{x_{n,1}, \ldots, x_{n,d_n}\}, \mathbb{D}). \tag{18}$$

5.3 Other approaches

We compared our proposed approach to the recently developed system-identification and heart-rate variability based approach by [106] and to the NICHD expert system we described in Chapter 2. In this section, we provide an outline of the key components of the SVM method. For details of the expert system approach, we refer the reader to the relevant chapter.

5.3.1 Features

In [106], it was assumed that FHR (denoted y) is composed of three different components: (a) baseline heart rate corresponding to average cardiac output (the DC component), (b) response to changes in maternal uterine pressure signals, and (c) variability due to sympathetic-parasympathetic modulation. These three components are modeled as follows:

1. the FHR baseline signal y_{BL} in the 0-4.5 mHz frequency range as a linear trend,
2. the FHR response to UP y_{SI} in the 4.5-30 mHz frequency range as the output of filtering the maternal UP with a filter whose impulse response function is estimated from training data, and
3. the FHR response to the autonomic nervous system dynamics y_{HRV} in the 30 - 1000 mHz frequency range as the output of an autoregressive model driven by a white Gaussian noise input.

From these components, five different features are obtained: (1) the offset of the linear fit to y_{BL}, (2) the gain and (3) delay of the impulse response

function used to model the UP-y_{SI} dynamics, (4) the low frequency (30-150 mHz) and (5) movement-frequency (150-500 mHz) components of the power spectral density of y_{HRV}.

5.3.2 Discriminative classification

In [106], the SI-HRV features were used as input to an SVM, which is a discriminative approach, using a Gaussian kernel to allow for nonlinear boundaries between groups. The authors used two different SVMs for the SI and HRV features respectively and combined their results using OR conditions to get a final categorization. That is, if any one of the SI or HRV classifiers classified the data as "abnormal", the final classification was also "abnormal". We used the same method for our data, too. For each fold of crossvalidation, we obtained separate SVMs using (a) only the SI features, (b) only the HRV features, and (c) all five SI-HRV features, yielding three different arrays of test classification results. An additional result was obtained by combining the classifier output from SI and HRV SVMs using the OR condition. In order to be consistent with the authors' methods, we obtained accuracy values for all four by using a wide range of values for the SVM parameters (the scaling factor σ for radial basis function kernel, and the box constraint C for the SVM soft margin [16]).

5.4 Performance comparisons

5.4.1 Empirical setup

Our database consisted of deidentified EFM files from 201 different babies (admitted to the neonatal intensive care unit at the Stony Brook University Medical Center, Stony Brook, NY) for which post-delivery umbilical cord pH values were available for 111. FHR and UP monitoring was done for each of them using either internal (fetal scalp ECG) or external (Doppler) monitors. All consent and approval guidelines were followed rigorously.

Out of these 111 files, we found that 83 had usable EFM data from singleton pregnancies. We restricted analysis to the last half an hour of available EFM data. For this time period, each baby could have one or more contiguous EFM epochs[1] with breaks in between (resulting from excessive noise or electrode drop-off). Automated preprocessing, as explained in Chapter 2, was always followed by a visual inspection, in order to confirm that the automatically "cleaned-up" record was actually usable for further analysis. In consultation with our collaborating physician, we fixed the threshold for fetal distress to be pH ≤ 7.15. That is, if the cord pH was greater than 7.15, the baby was assumed to be "healthy" (category 1), otherwise "not healthy" (category 2). This labeling was treated as the gold-standard classification. This yielded 23 datasets in category 2 and 60 in category 1. There were an average of 2.4 non-overlapping epochs in the each dataset.

We assessed classification performance using the 10-fold stratified cross-validation method. This is acknowledged to be an effective method for measuring classifier performance [57]. Under this method, we partition the dataset into 10 different non-overlapping groups, with the proportions of category 1 and 2 datasets being the same in each group. We then treat each group as a test dataset, using the datasets in the remaining 9 as training data. Each such group-testing is called one "fold" of crossvalidation. Thus, in each fold, we use approximately 90% of the data for training and 10% for testing, and the partitioning ensures that each file is treated as a test dataset exactly once when all 10 folds have been evaluated. We calculate a confusion matrix by comparing the class-labels obtained for all datasets with the gold-standard categorization. We repeat this process for 10 different random partitions of the dataset, to get an empirical confidence measure. We note that, when comparing different classifiers for a given run, we test all methods using exactly the same cross-validation partition, so as to ensure a fair comparison. We compare the classifier performance (measured using the cost-insensitive weighted relative accuracy (WRA) [57, 58]) for all GM methods using a range of parameter values for data segment length t and bin width b, and for all SVM methods by varying the scale parameter

[1] We define an epoch as a sequence of nonoverlapping segments.

Table 8: Classification performance evaluations for all GM methods. The second column shows the parameter values yielding highest performance in terms of WRA. Best performances are in bold.

GM Method	t^*(s), b^*	Best TNR, TPR	Best WRA
GM-MM-C	95, 0.0100	0.817, 0.478	0.295
GM-NB-C	115, 0.1000	0.842, 0.348	0.189
GM-MM-E	45, 0.0158	0.820, 0.436	0.256
GM-NB-E	15, 0.0501	0.757, 0.300	0.057
GM-MM-CT	45, 0.0316	0.725, 0.587	0.312
GM-NB-CT	45, 0.0794	0.800, 0.435	0.235
GM-MMorNB-C	**115, 0.0200**	**0.817, 0.609**	**0.425**
GM-MMorNB-E	45, 0.0158	0.768, 0.473	0.241
GM-MMorNB-CT	115, 0.1259	0.742, 0.500	0.242

values σ and box constraint C. WRA is an unbiased accuracy measure that subtracts the component of true positive positive score that is attributable to chance, and is defined as WRA $= 4 \times \text{cost} \times (\text{TPR-FPR})/(1+\text{cost})^2$, where TPR and FPR denote the true and false positive rates, respectively. When the cost for misclassification is zero-one, as in this study, the range of WRA is $[-1, 1]$. Then WRA = TPR $-$ FPR.

5.4.2 Results

We tested several modifications of the GM and SVM classifiers on our data, and have shown the results in Tables 8 and 9. The first column for each subtable in (a) and (b) contains the type of classifier used. The suffix C, E or CT at the end of the type denotes "cumulative", "epoch-level" or "threshold-based cumulative" classification respectively. In *cumulative* classification, all the features obtained from all the EFM epochs for a given baby were simply concatenated together without regard for breaks in the data, making a single long feature sequence for each subject. This yielded 83 discrete x sequences,

Table 9: Classification performance evaluations for all non-GM methods. The second column shows the parameter values yielding highest performance in terms of WRA. Best performances are in bold.

Method	σ^*, C^*	Best TNR, TPR	Best WRA
SVM-SI-E	0.25, 1024.00	0.566, 0.436	0.003
SVM-HRV-E	**0.50, 0.50**	**0.710, 0.436**	**0.146**
SVM-SIHRV-E	0.50, 0.13	0.706, 0.345	0.051
SVM-SIorHRV-E	0.25, 512.00	0.415, 0.508	-0.077
SVM-SIorHRV-CT	0.25, 1024.00	0.175, 0.826	0.001
NICHD-E	–	0.163, 0.883	0.047
NICHD-CT	–	0.0833, 0.9130	-0.004

corresponding to the 83 babies. In *epoch-level* classification, we did not perform this concatenation, but instead treated each contiguous epoch of EFM data as a separate dataset. However, the gold-standard diagnosis for each such epoch was still assumed to be equal to the baby's final diagnosis (based on pH). There was a total of 201 discrete x sequences corresponding to 201 EFM epochs obtained from 83 babies. Lastly, in *threshold-based-cumulative* (CT) classification, we first performed epoch-level classification. Then, for a given test baby i^*, we classified it as category 2 only if at least one of baby i^* 's epochs was classified as category 2 (thus the term threshold-based, where "1 epoch" is the threshold for deciding abnormality).

Another level of classifier modification was introduced by combining classifier results using OR conditions. For instance, for the cumulative GM paradigm, we applied the OR logic to the decisions made by the GM-MM-C and GM-NB-C classifiers. This is denoted GM-MMorNB-C. That is, we first used each of the MM and NB models to make an initial decision whether an epoch was category 2, and if at least one of these two decisions was positive, we called the dataset "category 2". This is similar to the method used in [106], in which classifier decisions from two different SVMs (one using system identification (SI) features, and the other using only HRV features)

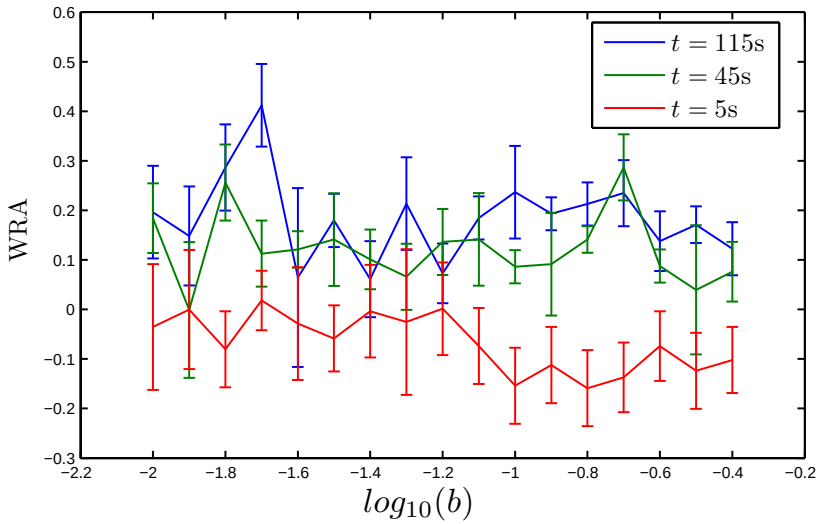

Figure 21: Error bars representing median $\pm$ interquartile range of WRA of classification using the GM-MMorNB-C method as a function of the bin width b and for three different t values.

Table 10: Confusion matrix for NICHD-ES classification.

		NICHD-ES		
		1	2	3
pH-based	1	23	117	1
labeling	2	7	53	0

were ORed to yield the final classification. Similar combinations were performed for the SVM methods as well.

Figure 21 shows the variation of the WRA when using the GM-MMorNB-C method, as a function of the bin width and for three different values of t. In a similar vein, Figure 22 shows the variation of the WRA when using the SVM-HRV-E method for different values of the C and σ parameters.

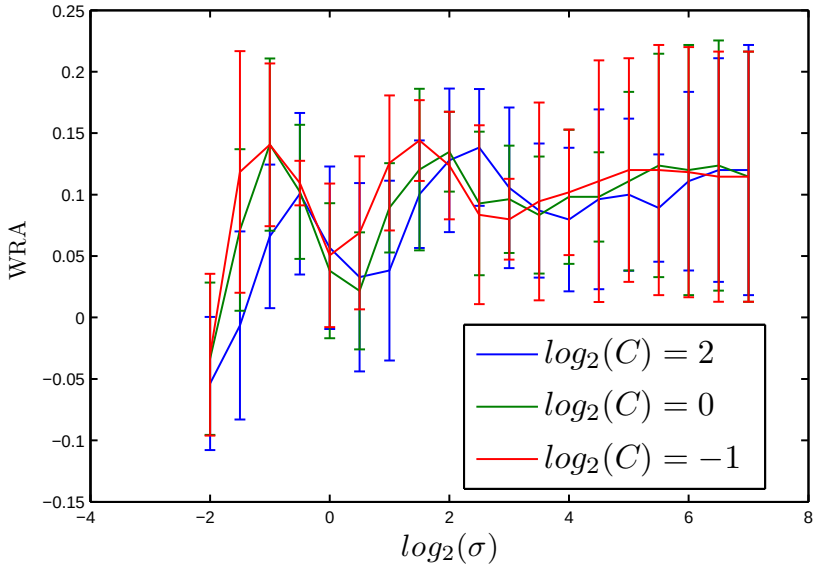

Figure 22: Error bars representing median $\pm$ interquartile range of WRA of classification using the SVM-WHRV-C method as a function of σ and for three different values of C.

As explained in Chapter 2, the NICHD-ES classifier yields, as a decision for a given FHR dataset, one of three category values. Table 10 shows the confusion matrix obtained when comparing this to the pH categorization. In order to compare this method's performance to the GM and SVM approaches, we transform the results such that NICHD-ES categories 2 and 3 are mapped to "not healthy". The best result from this is shown in the second to last row of Table 9.

TPR and TNR denote the median of the true and false positive rates across all crossvalidation runs. For each method, we varied the corresponding parameter values and looked for all (TPR, TNR) pairs such that TNR was at least 0.7 (similar to the search performed in [106]), assuming that would be a clinically reasonable rate. If a particular method did not yield a TNR >= 0.7 for any of its parameter values, we decreased the TNR threshold by 0.01 in steps, and repeated the search. Finally, out of these candidates, we selected the best performing parameters as that which yielded the best WRA.

As mentioned previously, the study by [106] used the SI-HRV features as inputs to a classification system, but without directly accounting for local variations. The authors had found that for "pathological" data sets,[2] the gain and delay of the impulse response function between UP (as input) and FHR (as output) were increased. This finding is consistent with previous research on timing of acute intrapartum hypoxic injury (e.g., in [22]) and by the physicians' practice of regularly looking for the influence (or lack thereof) of contractions on episodic variations such as decelerations [100]. In the present study, we found that our methods consistently performed at par with or better than the SI-HRV method (see Tables 8 and 9). Figures 21 and 22 show that in general, the SVM approach is less sensitive to the parameter settings than the GM approach. However, in terms of absolute performance, this insensitivity does not allow it to reach the WRA values achieved by GM approaches. Additionally, we note that the original approach by [106] called for the use of an OR condition on the results of two separate SVM classifications, followed by a threshold-based cumulative classification. This method, in our current study was found to have a worse performance than simply using the HRV features in a single SVM. In other words, the system-identification features do not seem to yield much information for this dataset.

The marked decrease in the ES classifier performance may be due to the hard if-then conditions and to failures when "symptoms" of more than one category are detected. We also observed that for several data sets, the ES classifier found symptom combinations that did not fit to any of the three NICHD categories. Although one can design the ES classifier to pool all such outcomes into category 2 ("indeterminate"), this increases the rate of false category 2 detection.

In contrast to our findings in Chapter 3 and our previous study [27], we found that binning the acceleration and deceleration features into different types, did increase performance. One reason for this could be that, unlike the study in [27] that used features extracted from long stretches of

[2]Defined as post-delivery arterial umbilical cord base deficit > 12 mmol/L, or showing evidence of hypoxic ischemic encephalopathy, or death.

data, we are using collections of short-segment-features. In [27], a discrete feature of, say, a deceleration subtype D_l ("existence of late decelerations") was assumed to take a truth value if there existed even one late deceleration in the full 20-min record. This is different from saying, as in the present study, that there existed a sequence of deceleration-segments, with, say 20 early deceleration-segments and 5 late deceleration-segments. Thus, we believe, this approach gives a more complete picture of the FHR record and improves the classification performance.

Computational issues in the GM methods studied here are also lessened. Since the GM classifiers are based on Bayesian parameter learning of discrete feature sequences, they can be interpreted as simple counting operations. Thus, they are considerably faster that, say the SVM classifiers, which have to solve an optimization problem every time training is performed on a new dataset. However, rule-based classification is the fastest, since after the set of features has been found, it simply goes through a checklist of if-then-else conditions, and does not involve any minimization or counting operation.

One limitation of the current study is dealing with noisy segments. In a previous version of the study, we had adopted a rule of removing noisy segments (defined as any segment with more than 30% of the samples had been interpolated over during preprocessing) entirely from the analysis. This process, however, can create complications because it forces feature segments before and after a removed region to become adjacent, potentially confounding the parameter updates if there are long stretches of noise. To tackle this in the present study, we chose simply to not designate any segment as noisy, thus keeping all the segments connected. Instead, we went back and made the preprocessing step more robust, and additionally did detailed visual inspection/correction to make sure the number of truly noisy segments in the final feature sequence was as low as possible. This process was kept the same for all the classification methods studied. We acknowledge that this may not be the best way to deal with noisy segments, especially in real-time settings, and are working to improve this aspect in our ongoing research.

Another limitation is that in clinical situations, data are acquired consecutively in time, and the classification of entire epochs is not very useful in such situations. One would, instead, like to have a system that can raise an alarm whenever an "at-risk" pattern of feature sequences is raised. One possibility is to analyze T-second epochs of FHR-UP data updated every t seconds to include a new incoming segment. Then, if more than E such epochs are deemed to fall in the "at-risk" category, one can raise an alarm. At present, we are exploring this possibility. We note that the "threshold-based-cumulative" (CT) classification paradigm addressed in Section 5.4.2 can give us some clues about the efficacy of this method. In particular, we can see from Table 8 that the GM-MM-CT method provides comparable performance to the best classifier (GM-MMorNB-C), but has a decreased specificity (72% v 81%).

6 Future work and conclusions

In this chapter, we describe possible extensions to our current research on fetal heart rate and maternal uterine pressure analysis and classification. As described in Chapter 5, we used the generative mixture model approach to bring improvements in classification accuracy. We are currently working on further extensions of this method for four related objectives: (a) decreased reliance on fuzzily defined gold-standard clinical annotations via the use of unsupervised learning, (b) further decreases in classification error rates (increases in sensitivity and specificity) by obtaining better, more data-driven features as opposed to those obtained from more ad-hoc rules for episodic variations, (c) defining fetal risk scores that can be used as true gold-standard and that sensibly combine the gold-standard classifications from physician interpretation and objective fetal status metrics, and (d) extension of the clustering methodology to the real-time monitoring case. In the sequel, we describe in more detail the results obtained for these goals.

6.1 Extension to unsupervised clustering

In the unsupervised clustering case, we no longer have any training data from possible classes to build specific class-conditional models from, and to compare with individual datasets. As mentioned previously, the lack of a clear gold-standard annotation hampers any supervised learning mechanism. This can happen even if we select a clear, objective criterion of abnormality such as “fetal death or injury", because when there is any indication that this future is possible for the unborn fetus during labor, immediate and drastic steps are taken to intervene in the normal course and prevent

it. Thus, even highly suspicious fetal patterns can be associated with “normal" fetal outcomes which compromises the quality of the training labels. Other non-visual fetal health metrics such as umbilical pH, base deficit values and Apgar scores can often show poor correlations with FHR patterns, while the unreliability of visual interpretation of fetal patterns has previously been discussed. Moreover, it is a well-accepted notion amongst the obstetric community that far too many cardiotocographic recordings are labeled as category 2 because of the understandable caution exercised by intrapartum care personnel to prevent any risk to the fetus. In practice, the rule boils down to putting all traces not classified as category 1 (definitely normal) or 3 (certainly abnormal) into category 2 (indeterminate). This increases the false positive rate for diagnosing fetal distress. Finally, the problem of obtaining a diverse training database is compounded by the fact that FHR tracings showing patterns definitively proving fetal distress are very rare (recall our original 830-strong database of signals, which had less that 1% of signals physician-labelled as category 3). All these reasons encourage us to try and obviate the need for gold-standard labeling, i.e., to try unsupervised learning methods capable of partitioning an input consisting of several datasets into meaningful groups. We choose a Bayseian formulation of this problem, where we try to estimate the posterior joint probabilty mass function of the entire set of class variables for the N input time-series records. Although there exist several discriminative approaches for unsupervised classification, we choose a generative model approach because of its considerable flexibility with regard to definition of the models, ease of calculation of entire joint distributions, elegant use of priors and hyperpriors for all the different parameters in the system, and robustness with respect to noise. This unsupervised approach is a very natural extension of the supervised classification methods developed in Chapter 5, as explained in the sequel.

Once again, we assume that we have access to a database of N different FHR-UP records, and from each of these we derive a feature sequence, for instance as explained in 5.1. A feature sequence x_i is extracted from the dataset y_i. We wish to find the class label C_i associated with each of

these feature sequences. Recall from Chapter 5 that our target distribution is $P(C_{1:N}|\mathbf{x}_{1:N}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ denotes the set of hyperparameters.

One can get an approximate form for this distribution through one of several approaches. Here, we propose a Markov chain Monte Carlo (MCMC) approach. In this approach, we first construct a Markov chain on the state space of clusters, whose stationary distribution is also $P(\mathbf{C}_{1:N}|\mathbf{x}_{1:N}, \boldsymbol{\gamma})$. Then, we perform a random walk in this state space such that the fraction of time spent in each state $\mathbf{C}_{1:N}$ is proportional to $P(\mathbf{C}_{1:N}|\mathbf{x}_{1:N}, \boldsymbol{\gamma})$. We then take lots of samples from the chain when it converges to this stationary distribution. Finally, we can get the "best" estimate of $C_{1:N}$ from these samples by appropriately "summarizing" the samples.

6.1.1 Gibbs sampling

Since the target distribution is over a very high-dimensional variable ($C_{1:N}$), we need a method capable of sampling efficiently from this. In general, this is not trivial. Because of the considerable size of the random vector to be sampled (one can easily imagine hundreds or thousands of individual time series in the input to the clustering module), it is easier to sample from conditional distributions instead of the full joint distributions, *i.e.*, from $P(C_i|\mathbf{C}_{-i}, \mathbf{x}_{1:N}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$.[1] So our problem could become considerably easier if one could somehow use the conditional distributions to sample from the full joint distribution. This can be achieved via a technique called Gibbs sampling. The basic idea is to sample each unobserved variable in turn, conditioned on all other variables in the system. If a certain variable is observed, its value is known and so, not sampled. This sampling procedure is repeated for a large number of iterations, with the assumption that after sufficient time, the chain will have mixed well, and the joint distribution will have converged to the target distribution.

Consider a system with variables $\mathbf{z} = \{z_1, \ldots, z_N\}$, and let our goal be eventual convergence to the target joint distribution $P(z_1, \ldots, z_N)$. We use the Gibbs sampling method to achieve this. Assuming that after some

[1] The subscript term $-i$ or $\backslash i$ denotes "all datasets except the i-th".

iteration t, we have the samples $\mathbf{z}^t$ for these variables, we can generate a new sample $\mathbf{z}^{t+1}$ (assuming $\mathbf{z}$ has dimension N):

$$\begin{aligned} z_1^{t+1} &\sim P(z_1|z_{2:N}^t) \\ z_2^{t+1} &\sim P(z_2|z_1^{t+1}, z_{3:N}^t) \\ &\vdots \\ z_N^{t+1} &\sim P(z_N|z_{1:N-1}^{t+1}). \end{aligned}$$

The distribution $P(z_i|\mathbf{z}_{-i})$ is called the *full conditional* for the i-th variable. In the case of class-sampling as desired in the case of FHR analysis, we must use $P(C_i|\mathbf{C}_{-i}, \mathbf{y}_{1:N}, \boldsymbol{\gamma})$ in place of $P(z_i|\mathbf{z}_{-i})$. The set of variables z in the argument of the full conditional would have to include the model paramaters (such as $\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\eta}$). Often however, we can use a slight modification of the sampler, called Collapsed or Rao-Blackwellized Gibbs sampling to integrate out some or all of these parameters before sampling. There are two advantages to this procedure, which stem from the Rao-Blackwell theorem. This theorem states that if $\mathbf{z}$ and $\mathbf{y}$ are dependent random variables, and $f(\mathbf{z}, \mathbf{y})$ is a scalar function, then $\text{Var}_{\mathbf{z},\mathbf{y}}[f(\mathbf{z}, \mathbf{y})] \geq \text{Var}_{\mathbf{z}}[E_{\mathbf{y}}[f(\mathbf{z}, \mathbf{y})|\mathbf{z}]]$. That is, marginalization of "nuisance variables" always leads to more robust estimates of the variables of interest. Furthermore, if one desires estimates of the model parameters, one need not sample from the joint distribution, but instead simply perform a maximum likelihood estimate *after* the sampling procedure has been completed for the variables of interest.

Coming back to our current problem, we can rewrite the expression for the full conditional $P(C_i|\mathbf{C}_{-i}, \mathbf{x}_{1:N}, \boldsymbol{\gamma})$ as follows:

$$P(C_i|\mathbf{C}_{-i}, \mathbf{x}_{1:N}, \boldsymbol{\gamma}) \propto P(C_i|C_{-i}, \boldsymbol{\gamma}_1) P(\mathbf{x}_i|\mathbf{C}_{1:N}, \mathbf{x}_{-i}, \boldsymbol{\gamma}_2), \tag{19}$$

where $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ denote the hyperparameters associated with the cluster and feature sequence likelihood model parameters respectively. The first term on the right hand side arises from the marginalization of the mixture weights π (the class proportions as described in Sections 5.2.1 and 5.2.2 and Figure 19). Since we have endowed these weights with a Dirichlet prior, we can write this term explicitly as:

$$P(C_i = k|C_{-i}, \boldsymbol{\alpha}) \propto (\alpha_k + s_{k,-i}), \tag{20}$$

where $s_{k,-i}$ denotes the number of data records from the set $\mathbb{D}_{-i}$ which in the current iteration have class-assignment k. α_k represents the concentration parameter of the Dirichlet prior.

The second term on the right hand side in (19) is the likelihood term. We can rewrite this term as:

$$P(\mathbf{x}_i|C_i = k, \mathbf{C}_{-i}, \mathbf{x}_{-i}, \boldsymbol{\gamma}_2) = P(\mathbf{x}_i|\{x_j|C_j = k, j \neq i\}, \boldsymbol{\gamma}_2)). \tag{21}$$

In other words, we use the datasets currently assigned to class k to "estimate" model parameters and find how well the i-th feature-sequence is "explained" by the cluster k. This is done for all possible cluster assignments $k \in \{1, \ldots, K\}$. One can use the product of the cluster and likelihood probabilities as a proportional estimate of the a-posteriori conditional probability, which is then used for sampling the cluster value.

Thus, in the case of the näive Bayes feature model, one can write the expression for the likelihood part as:

$$P(\mathbf{x}_i = \{h_1, h_2, \ldots, h_{d_i}\}|C_i = k, \mathbf{C}_{-i}, \mathbf{x}_{-i}, \boldsymbol{\lambda}) = \prod_{h=1}^{H} \hat{\theta}_{k,h}^{q_{*,h,i}}, \tag{22}$$

$$\hat{\theta}_{k,h} = \frac{q_{k,h,-i} + \lambda_{k,h}}{N_k + \sum\limits_{h} \lambda_{k,h}}. \tag{23}$$

where N_k is the number of datasets in $\mathbb{D}_{-i}$with current assignment = k, $w_{i,v} = |\{j : x_{i,j} = v; j \in \{1, \ldots, d\}\}|$ (the number of occurrences of feature value v in the dataset i) and $q_{m,v}^{-i} = |\{(p, j) : C_p = m, x_{p,j} = v\}|$ (the number of occurrences of feature being v across all the class-m datasets in $\mathbb{D}_{-i}$). Thus the expression for the full conditional is given by:

$$P(C_i = k|\mathbf{C}_{-i}, \mathbf{x}_{1:N}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) \propto (\alpha_k + s_{k,-i}) \prod_{h=1}^{H} \left(\frac{q_{k,h,-i} + \lambda_{k,h}}{N_k + \sum\limits_{h} \lambda_{k,h}} \right)^{q_{*,h,i}} \tag{24}$$

In the above, the count variable $q_{k,h,-i}$ represents the number of occurrences of the feature value h in the set of datasets $\mathbb{D}_{-i}$ that have the current class-variable sample k.

In the case of the first-order Markov chain model, the cluster probability expression remains the same as in (20), while the likelihood density

depends on the "estimated" values of the initial value probability ϕ and transition matrix $\boldsymbol{\theta}$. These estimates are given by:

$$\hat{\phi}_{k,h} = \frac{\eta_h + r_{k,h,-i}}{\sum_h \eta_h + s_{k,-i}}, \quad \hat{\psi}_{k,g,h} = \frac{\beta_{g,h} + z_{k,g,h,-i}}{\sum_h \beta_{g,h} + q_{k,g,-i}}. \tag{25}$$

Here, $r_{k,h,-i}$ is the number of occurrences of h in the first segments of class-k records in $\mathbb{D}_{-i}$ and $z_{k,g,h,-i}$ is the number of transitions from g to h in class-k records in $\mathbb{D}_{-i}$.

Given these update parameters, the sampling density can be calculated as:

$$P(C_i = k|\mathbf{C}_{-i}, \mathbf{x}_i = [h_1, \ldots, h_{d_i}], \mathbf{x}_{-i}, \alpha, \boldsymbol{\lambda}, \boldsymbol{\eta}) \propto (\alpha_k + s_{k,-i})\hat{\phi}_{k,h_1} \prod_{j=2}^{d} \hat{\theta}_{k,h_{j-1},h_j}. \tag{26}$$

Thus, at each iteration, for each dataset $\mathbf{x}_i$, one needs to calculate the sampling density based on the observed feature values in the i-th dataset, the current assignments and the feature values for all the other datasets. The sampling density is calculated for each possible cluster k, and we then get a vector of probabilities of length K. We then sample a value for C_i for this iteration, and update our current assignment dataset $\mathbb{D}$ to include the i-th dataset's new assignment. In this way we proceed to sample a cluster-value for each dataset for the iteration, and repeat this process for T iterations, which must be kept sufficiently large to ensure the chain mixes well. After sampling for T iterations, we need to summarize the cluster samples for each data record in order to produce a "best" clustering for each record. Typically, it is also recommended to exclude the first b iterations (e.g., the first 25% of iterations; this is also called the burn-in period) in order to ensure the chain has converged before we summarize. One can use several methods to summarize the samples, e.g., we may denote the best clustering for i-th dataset to be the mode of the cluster samples for this dataset for all iterations.

6.1.2 Extension to unknown number of clusters

The above sampling density calculations are applicable to the case when the number of clusters is known. Alternatively, we may be interested in finding out clinically relevant categories in the data, but may not sure how many such categories are present. For instance, such a situation may arise when we want to analyse what are the different kinds of fetal heart rate sub-categories whose recommended classification is "Indeterminate" (category 2). There may thus be potentially infinite number of clusters. However, in order to make sense of the data, we need to incorporate a preference for compact representations, i.e., we want the number of clusters to remain as few as possible but without shoehorning radically dissimilar records into the same cluster just for the sake of compactness. This can also be achieved as an extension of the Gibbs sampling procedure, by using a special form of the prior for the mixture weights π.

Thus, in the generative stories outlined in Section 5.2.1 and 5.2.2, we need to encode the fact that there are potentially infinite number of mixture components. That is,

$$\text{Likelihood parameters} : \boldsymbol{\gamma}_m \sim G; \quad G(\boldsymbol{\gamma}) = \sum_{k=1}^{\infty} \pi_k \delta(\boldsymbol{\gamma}, \boldsymbol{\gamma}_k), \tag{27}$$

In effect, the above procedure amounts to sampling the mixture weights according to:

$$\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m, \ldots, \pi_K] \sim GEM(\alpha), \tag{28}$$

where $GEM(\boldsymbol{\alpha})$ denotes the Griffiths, Engen and McCloskey process (also known as the stick-breaking construction). In this construction, the infinite number of $\boldsymbol{\pi}$ components are generated according to the following process:

$$\beta_k \sim \text{Beta}(1, \alpha), \tag{29}$$

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) = \beta_k (1 - \sum_{l=1}^{k-1} \pi_l), \tag{30}$$

where α is called the concentration parameter of the Dirichlet process $G(\alpha)$.

Thus, the two components of the sampling density in (24) and (26) need to be modified to include the possibility of new, previously unseen clusters. Recall that at each iteration, for each dataset, we need to sample the dataset cluster C_i using the distribution $P(C_i|C_{-i}, \mathbf{x}_{1:N}, \boldsymbol{\gamma}) \propto P(C_i|C_{-i}, \alpha)P(\mathbf{x}_i|C_{1:N}, \mathbf{x}_{-i}, \boldsymbol{\gamma}_2)$.

The mixture component probability can then be written out as:

$$P(C_i|C_{-i}, \boldsymbol{\alpha}) = \frac{1}{\alpha + N - 1}\left(\sum_{k=1}^{K} s_{k,-i}\delta(C_i, k) + \alpha\delta(C_i, \bar{k})\right), \tag{31}$$

where $\delta(C_i, k)$ denotes the Kronecker-delta function (equals 1 whenever $C_i = k$, 0 otherwise), K is the number of distinct clusters present in the current iteration in the collection of datasets excluding the i-th (denoted $\mathbb{D}_{-i}$), $\bar{k}$ denotes a new, previously unseen cluster, and α denotes the concentration parameter of the Dirichlet process G.

The cluster-conditional feature likelihood remains the same as in (24) and (26) for those clusters that are already seen in $\mathbb{D}_{-i}$. For a possible new cluster $\bar{k}$, the likelihood probability can be obtained by using only the Dirichlet prior parameters for "estimating" $\boldsymbol{\gamma}$ parameters as before. Thus, for the näive Bayes model, the expression can be written out as:

$$\hat{\theta}_{\bar{k},h} = \frac{\lambda_h}{\sum\limits_{h} \lambda_h}. \tag{32}$$

For the Markov model, the corresponding expressions are:

$$\hat{\phi}_{\bar{k},h} = \frac{\eta_h}{\sum\limits_{h=1}^{H} \eta_h}; \hat{\psi}_{k,g,h} = \frac{\lambda_{g,h}}{\sum\limits_{h=1}^{H} \lambda_{g,H}}; \tag{33}$$

So the sampling density is given by:

$$P(C_i|C_{-i}, \mathbf{x}_{1:N}, \alpha, \boldsymbol{\lambda}) \propto \Big(\alpha P(\mathbf{x}_i|\boldsymbol{\gamma}_{\bar{k}})\delta(C_i, \bar{k}) + \sum_{k=1}^{K} s_k^{-i} P(\mathbf{x}_i|C_i = k, \mathbf{C}_{-i}, \mathbf{x}_{-i}, \hat{\boldsymbol{\gamma}}_k)\delta(C_i, k)\Big) \tag{34}$$

Thus, at each iteration, for each feature-vector $\mathbf{x}_i$, one needs to calculate the sampling density based on the observed feature values in the i-th

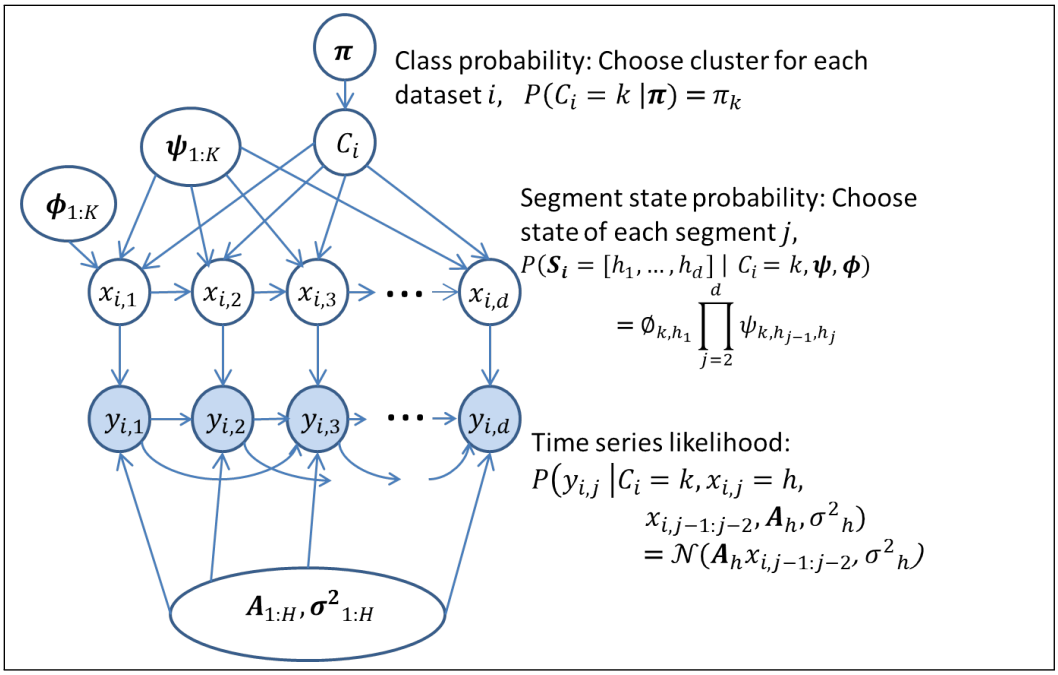

Figure 23: DAG for complete category-specific switching autoregressive model with K overall categories, H possible symbols in feature alphabet and autoregressive order $\rho = 2$.

dataset, the current assignments, and the feature values for all the other datasets. The sampling density is calculated for each possible cluster k observed in $\mathbb{D}_{-i}$ and for some new cluster $\bar{k}$, and we then get a vector of probabilities of length $K + 1$, where K is the number of different cluster-assignments currently seen in $\mathbb{D}$. Thus at each iteration, new clusters may be born or observed clusters may die out, and this depends on the prevalence of radically different datasets.

6.1.3 Outlook

One reason why we may be interested in inferring a clustering of a database of FHR-UP signals is that we may be interested in making a graded categorization of fetal heart rate data (from "fully normal" to "definitely abnormal"), with the number of classes K controlling the granularity of the

gradation. It is possible, but certainly not necessary, that K turns out to be very small. Making a small-K assumption prior to the analysis would be a considerable leap of faith, and the Bayesian approach with suitably defined priors for this number can be used to make the analysis robust to the granularity of categorization. The use of NPB methodology via generative mixture models will allow us to operate with unknown K and yet find the a posteriori probability distributions of interest, $P(C_i = k|\mathbf{x}_i, \boldsymbol{\gamma})$ in addition to the likelihood $P(\mathbf{x}_i|C_i = k, \boldsymbol{\gamma})$. By contrast, discriminative models are limited to estimation of the latter.

Another interesting possibility is the use of unsupervised methods to jointly detect the segmentation points, different FHR patterns and the overall clustering of entire feature sequences. The automatic segmentation of the time-series to detect features in an unsupervised fashion is discussed in Section 6.2. One possible model for this problem could be that shown in Figure 23. The time series data are represented with a second-order AR model with hidden state-dependent switches. The number of possible states is H. However, we have a higher level clustering present, too, in which each possible category k of FHR-UP data is associated with separate transition matrices $\boldsymbol{\Psi}_k$ and initial-value pmfs $\boldsymbol{\phi}_k$. The unknown parameters are the class-proportions $\boldsymbol{\pi}$, class-specific state-transition matrices, and the state-specific AR parameters A and σ^2. It should be possible to extend the MCMC sampler to this model, and infer the pdfs of both the pattern-sequences (state-sequences) and clusters simultaneously. Finally, one can extend this model also to the case where the alphabet size H and the number of states K are unknown via the use of Dirichlet process priors.

We expect that a dynamic model will represent feature sequences more accurately than a static one (although the optimal model order n may vary for different features). In the context of dynamic models, an important challenge is dealing with the noise. When the model does not have dynamics, it is less difficult to account for it; the order of occurrence of feature values is irrelevant to the likelihood computation, and so the noise segments can simply be ignored. However, when temporal effects are included, we will need a way to handle the segments whose features are missing due

to noise. One possible solution is to incorporate marginalization. In other words, when calculating the feature vector likelihood, we can identify the noise segments and simply sum over all the state values that the segments could have taken. Another problem is the fact that often it is not clear which outcome measure to use for gold-standard classification, and this can compromise the evaluation of the clustering. This problem can be addressed by defining an appropriate "risk-score" (Section 6.3).

6.2 Data-driven segmentation of FHR-UP records

In previous chapters, we described the use of probabilistic approaches to describe fetal heart rate patterns. In Chapter 5, we proposed the use of supervised clustering algorithms to partition a set of FHR data records into clinically useful categories. The supervised classifier combines three major components: (a) extraction of *sequences* of discrete valued features from fixed-length-segmented FHR time series data, (b) modeling these feature-sequences as observations from finite or infinite Dirichlet mixtures, and (c) hard decision-making using maximum a-posteriori rules. We use näive Bayes as well as Markov-time-dependence models for the evolution of feature sequences.

The performance of the first of these system components (feature extraction for each segment in an FHR record) depends strongly on the quality of the segmentation. In fixed-length segmentation, the time-series is divided into a sequence of non-overlapping t-s segments, each of which is input to the feature extraction module. This has a significant disadvantage; often, features extracted for a particular segment do not take into account the "context". For instance, lets say the feature we are looking for is "presence of an acceleration (an upward increase and return to the FHR baseline)", i.e., for each segment, an "acceleration detector" module assesses whether it actually contains one, and returns a value "Yes (= 1)" or "No (= 0)" depending on the result. But consider the case where a certain segment

contains an acceleration for some time, followed by a deceleration (a downward decrease and return to FHR baseline). Should this segment then yield the value "Yes", "No" or "Unknown"? Should one define a certain threshold in order to decide when a segment is an acceleration or not? Trying to solve such questions using ad-hoc methods such as majority decisions or time thresholds introduces more parameters into our algorithm. In general, the more features we try to detect, the more data we need to decide appropriate values for such tuning parameters.

In order to obviate the need for many such arbitrarily defined parameters, we aim to perform automatic, data-driven segmentation of the time series. The aim of this part of our current research is to use the frameworks of switching autoregressive (AR) processes and Monte Carlo sampling to perform this segmentation. At any given time point, the observed data (in this case, the FHR or FHR-UP pair) are modeled as being generated by one of H possible underlying modes/states. These states are assumed to evolve according to a Markov chain. The parameters of the observation probability distribution function (pdf) are assumed to be dependent only on the value of the hidden state. This structure is commonly referred to as a hidden Markov model (HMM), since the underlying states cannot be observed directly. It has been used very successfully to solve many different types of real-world state-detection problems. Some well-known examples are: (a) detecting the order of nucleotides in a DNA sequence, (b) finding the sequence of speech units (words or phonemes) from a particular speech signal, (c) detecting the sequence of human motions from measurements of positions, velocities or other data, and (d) finding the effects of major world events on stock-market returns. In the specific case of segmenting FHR-UP data, we are interested in finding a sequence of segments that can describe all the different fetal dynamics observed in any given FHR-UP time series. If we can do this using a data-driven, unsupervised approach, we no longer need to prespecify the time divisions that define features, and it does away with the need for somewhat arbitrarily defined threshold and windowing parameters for detection of FHR-specific patterns. In essence, we aim to let our data do the talking, instead for imposing any apriori assumptions about

expected patterns.

6.2.1 Background

In a first-order HMM, the state-evolution is governed by the pdf $P(x_j = h|x_{j-1} = g, \psi) = \psi_{g,h}$, where x_j denotes the state at time-point j, ψ denotes the transition probability matrix, and the element in ψ's g-th row and h-th column denotes the probability of transitioning from state g to state h. The observations y_j are assumed to be governed by $P(y_j|x_j = h, \theta) = L(y_j|\theta_h)$, where θ denotes the set of likelihood parameters for state h. Thus, the observations are conditionally independent of each other given the state, while each state is conditionally independent of all other states given knowledge of the previous state's instantiation. In the case of time series data, the observations can often be assumed to have an AR model structure, i.e., $P(y_j|x_j = h, y_{\setminus j}, \theta) = L(y_j|y_{j-1}, \ldots, y_{j-\rho}; \theta_h)$, where ρ denotes the (possibly unknown) model order. This simple Markov structure makes it possible to devise learning and inference algorithms to solve problems of research interest, be it the efficient computation of the probability of the observation sequence $y_{1:d}$ (given knowledge of parameters and state-sequence $x_{1:d}$, where d is the total number of observed data-points in the sequence) or the estimation of $x_{1:d}$ that "best-explains" the observations.

When one knows or can accurately estimate (a) the number of possible states H, (b) the transition probabilities, and (c) the functional forms and parameters of the observation pdfs for each state, one can use the Viterbi algorithm to solve this problem [85]. This method is a dynamic programming approach to maximizing the probability $P(x_{1:d}|y_{1:d}, \psi, \theta)$, and has been very successful in practice. Variations of this approach exist to incorporate more complicated dynamics such as AR processes. When (b) and (c) are not known, it is still possible to use the Viterbi algorithm via a cross-validation/tuning approach [85].

Recently, several efforts have been directed towards solving the same problem in the case when the total number of states is not known or may be potentially infinite [7, 33, 44, 102]. For instance, financial time series data

may be governed by any one of a variety of statistical regimes, depending on the vagaries of the wider world. In our problem, fetal heart rate patterns can change depending on a variety of external and internal conditions such as the maternal health, the presence and depth of contractions, movements by both the mother and baby, specific cardiac conditions and so on. Additionally, there may be several distinct variations of specific patterns. For instance, decelerations can occur with or without "shoulders" (a small upward-deviation-and-return at the end of the deceleration), or slow returns to baseline. Thus, one needs to incorporate the fact that new regimes may arise at any time. Nonparametric Bayesian (NPB) methods offer an elegant solution to such problems by using hierarchical Dirichlet process (DP) priors for state transition probability matrices. By construction, these are defined over countably infinite supports, while the hierarchical structure still allows for a finite probability of transition between any pair of states. This allows users to devise efficient Markov chain Monte Carlo (MCMC) sampling schemes to get good representations of the posterior probabilities of the state-sequences, the most common approaches being variations of Gibbs sampling [88] or Metropolis-Hastings methods [73]. These approaches work as long as the Markov structure of the state-evolution is maintained, which means that more complex observation models like linear dynamical systems (LDS) can be utilized without much overhead [33].

In the specific case of switching vector autoregressive (VAR) processes with unknown number of states, NPB methods [33, 34] have been formulated to sample the state-sequences, parameters and transition probabilities of the HMM. One highly effective segmentation method reported in [33] is a variant of the forward-backward (FB) algorithm that uses a truncated approximation of the HDP as a prior for the transition matrix parameters and then samples, at each iteration, the entire state sequence $x_{1:d}$ from $P(x_{1:d}|\psi, \theta, y_{1:d})$. In contrast to the simpler direct Gibbs (DG) approach, which samples each x_j from its full conditional $P(x_j|x_{\setminus j}, y_{1:d}, \theta, \psi)$, this block-sampling method completely uncouples the state-sequence samples in iterations t and $t-1$ [93]. This leads to faster mixing of the Markov

chain, and gives better representations of the posterior after the burn-in period. However, the segmentation error between the state-sequence sample and the true sequence still has considerable spread. In the present method, we have used a variant of the FB algorithm, in which we approximate "integrating out" some or all of the variables in the set $\{\psi, \theta\}$; in other words, we try to Rao-Blackwellize this sampling strategy. The approximation is necessary because it is not possible to run an FB procedure to find the sampling probability $P(x_{1:d}|y_{1:d})$ after integrating out the parameters analytically.

In the sequel, we describe the problem formulation, priors and sampling methods in detail in Section 6.2.2. We then provide, in Section 6.2.3, results of applying this method to two simulated time series that had switching AR dynamics. The advantages and potential pitfalls of this method are discussed in Section 6.2.4

6.2.2 Methods

We focus, for now, on switching AR models for scalar continuous-time observations, and assume that the state-evolution is governed by a Markov chain, i.e.,

$$P(x_j = h|x_{j-1} = g) = \psi_{g,h}, \quad y_j = \sum_{p=1}^{\rho} A_{h,p} y_{j-p} + w_j, \tag{35}$$

where $A_{h,p}$ denote, for state h, the AR coefficients, ρ the known model order and w the driving noise, which we assume has a Gaussian pdf with zero mean and known variance σ^2. The directed acyclic graph (DAG) for an example second-order switching AR model is shown in Figure 24. Note that, conditioned on knowledge of the past and future state variables $x_{i,j-1}, x_{i,j+1}$ the observation $y_{i,j}$ and the transition matrix $\boldsymbol{\psi}$ (or initial value pmf $\boldsymbol{\phi}$), the state variable $x_{i,j}$ is independent of all other states. Thus, the set $\{x_{i,j-1}, x_{i,j+1}, y_{i,j}, \boldsymbol{\psi}(\text{or } \boldsymbol{\phi})\}$ is called the Markov blanket of $x_{i,j}$.

Just as in [33], we assume that A_h are distributed according to multivariate Gaussian pdfs with hyperparameters μ_{A_h}, Σ_{A_h}. In order to allow for possibly infinite number of states, we can use the hierarchical Dirichlet

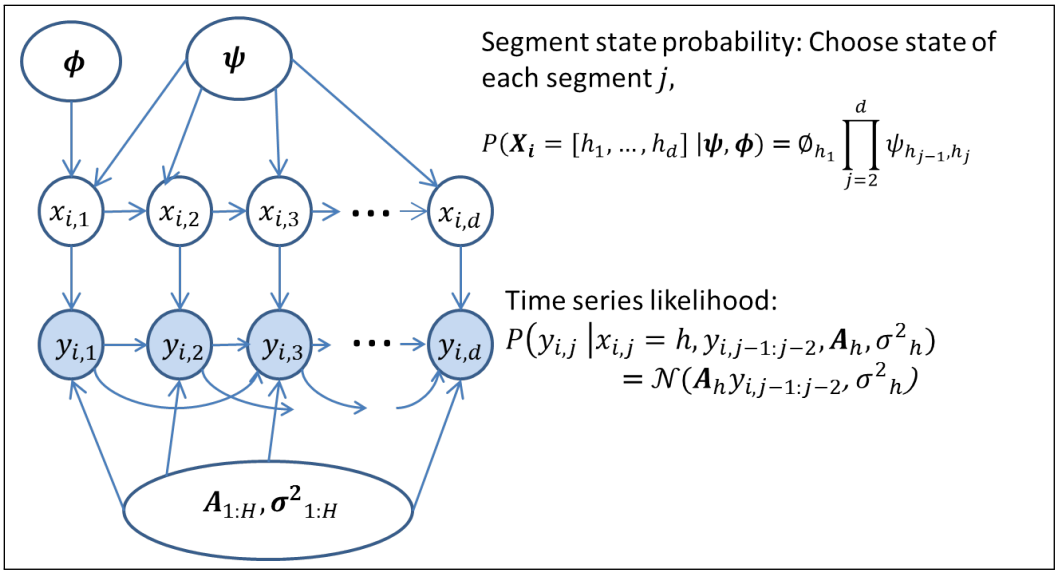

Figure 24: DAG for the switching autoregressive model used for modeling nonstationary time series data. Here, the autoregressive process is of second order, i.e., $\rho = 2$.

process (HDP) as a prior for the rows of the transition matrix ψ, i.e.,

$$G_0 = \sum_{h=1}^{\infty} \beta_k \delta_{A_h}, \quad \beta|\gamma \sim \text{GEM}(\gamma), \tag{36}$$

$$G_h = \sum_{l=1}^{\infty} \psi_{g,h} \delta_{A_g}, \quad \psi_h|\alpha, \beta, \kappa \sim \text{DP}\left(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_h}{\alpha + \kappa}\right), \tag{37}$$

$$A_h|\mu_A, \Sigma_A \sim \mathcal{N}(\mu_A, \Sigma_A), \quad \forall h \in \{1, 2, \ldots\}, \tag{38}$$

where β_h denotes the global probability mass associated with the state h and is distributed as a Griffiths-Engen-McCloskey (GEM) process with hyperparameter γ (also called a stick-breaking process) [102]. Thus, G_0 represents the prior pmf of the global frequency of each state. Similarly, G_h represents the prior probability of transition from state h to all other states, and is itself a Dirichlet process with concentration parameter α and base distribution β. This hierarchical construction ensures that each possible state-transition has a finite probability, since the global DP G_0 essentially "ties together" each state-associated DP G_h. Additionally, each self transition $h \to h$ is assumed

to have an extra prior probability mass κ, which is called the stickiness mass and encourages the samplers to learn models that have persistence (since real-world data usually exhibit slower transition dynamics). The variables α, γ and κ are also provided their own Gamma priors, and are learned from the data as in [34]. Henceforth, we will refer to the set of hyperparameters $\{\beta, \gamma, \alpha, \mu_A, \Sigma_A\}$ as Φ.

6.2.2.1 Direct Gibbs (DG) sampling

The goal of Gibbs sampling in the context of time-series segmentation is to get a representation of the required posterior distribution of the state-sequence, i.e., to find $P(x_{1:d}|y_{1:d}, \Phi)$. For each iteration t, a typical sampling scheme would be, (symbols with superscript (t) denote the value of the corresponding variable at iteration t):

1. Sample transition and AR parameters ψ and A from their conditionals $P(\psi^{(t)}, A^{(t)}|y_{1:d}, x_{1:d}^{(t-1)}, \Phi^{(t-1)})$.
2. Sample the state sequence $s_{1:d}$ from $P(x_{1:d}^{(t)}|y_{1:d}, \psi^{(t)})$ (either in a block or sequential fashion).
3. Sample the hyperparameters Φ from the conditional pdf given by $P(\Phi^{(t)}|y_{1:d}, x_{1:d}^{(t)}, A^{(t)}, \psi^{(t)})$.
4. Update the sufficient statistics and obtain updated posteriors for ψ and A.

In the direct Gibbs sampler, in step 2, we sample each state from its full conditional distribution, which factors as

$$\begin{aligned} P(x_j^{(t)} = h \mid x_{j-1}^{(t)} = g, x_{j+1}^{(t-1)} = h_1, y_{1:d}, \psi^{(t)}, A^{(t)}) \\ \propto \psi_{g,h}^{(t)} \psi_{h,h_1}^{(t)} L(y_j|y_{j-1:j-\rho}, A_h^{(t)}, \sigma^2). \end{aligned} \tag{39}$$

This approach necessarily couples the samples at adjacent iterations, which, compounded by the presence of correlated observations from AR processes, results in very slowly mixing Markov chains. Thus, an alternative method, which block-samples the entire state sequence from the pdf $P(x_{1:d}^{(t)}|y_{1:d}, A)$, is used here (as in [33]). We call it the full FB algorithm to distinguish it from our Monte Carlo Rao-Blackwellised version described later.

6.2.2.2 Full FB algorithm

The forward-backward (FB) sampler is a modification of the forward-backward method of [85] that was originally used for inferring the most probable state-sequences in an HMM. Since the original FB recursions work only for finite state supports, we have to limit the maximum number of possible states to $H'(> H)$, where H is the true number of unique states. The resulting prior is a hierarchical sticky Dirichlet *distribution*, and is a finite approximation to the sticky-HDP:

$$P(\beta|\gamma) \sim \text{Dirichlet}(\beta; \gamma/H', \ldots, \gamma/H'), \tag{40}$$

$$P(\psi_h|\beta) \sim \text{Dirichlet}(\psi_h; \alpha\beta_1, \ldots, \alpha\beta_h + \kappa, \ldots, \alpha\beta_{H'}). \tag{41}$$

Given $\psi^{(t)}, A^{(t)}$, the FB method for sampling $x_{1:d}$ proceeds as follows:

1. We initialize an array of *messages* $m_{j,j-1}(h)$ to 1.
2. We compute, $\forall j \in d, d-1, \ldots, 1$ and $\forall h \in 1, \ldots, H'$,

$$m_{j,j-1}(h) := \sum_{l=1}^{H'} \psi_{g,h}^{(t)} m_{j+1,j}(h) \mathcal{N}(y_j; A_h^{(t)} y_{j-1:j-\rho}, \sigma^2). \tag{42}$$

3. We initialize the state-transition counts $z_{g,h} = 0, \forall g, h \in \{1, \ldots, H'\}$. For each h, we compute the probability $L_h(y_j) = \mathcal{N}(x_j; A_h^{(t)} y_{j-1:j-\rho}, \sigma^2) m_{j+1,j}(h)$.
4. We then sample a state assignment

$$x_j^{(t)} \sim \sum_{h=1}^{H'} \psi_{x_{j-1},h}^{(t)} L_h(y_j) \delta(x_j, h), \tag{43}$$

   and increment transition counts $z_{g,h}$ accordingly.

The state transition counts are used as sufficient statistics (in addition to the observation sufficient statistics) to update the posteriors of ψ, A, Φ. Details of these computations have been provided in [34].

6.2.2.3 FB with Monte-Carlo-based Rao-Blackwellisation (RBFB)

The Rao-Blackwell theorem [55,62] (Section 6.1.1) suggests that, if one were to integrate out the "nuisance" parameters ψ, A from the joint distribution,

one would obtain samplers that would give more accurate representations of the posterior. Analytically marginalizing out some variables from a joint pdf always reduces the spread of any estimate dependent on it. Thus, we would ideally want to run a sampler that uses the conditional $P(x_{1:d}|y_{1:d})$ instead of $P(x_{1:d}|y_{1:d}, \psi, A)$.

Note, however, that in any HMM with Dirichlet priors, analytically marginalizing out ψ, A from the joint pdf leaves us with a Pólya urn process [12], in which the probability of any transition $g \rightarrow h$ depends on the number of times this transition has already occurred. In effect, not sampling (collapsing out) ψ, A makes every pair of states (x_i, x_j) dependent, and we lose the Markov independence structure that enables FB to work when the parameters are instantiated.

In order to overcome this problem, we use a simple Monte-Carlo integration procedure to "approximately" Rao-Blackwellize the FB method. Prior to step 1 (in the full FB procedure of Section 6.2.2.2), instead of sampling only once from the conditional probability distribution of the parameters, we sample M times. For each of the M parameter samples, we perform steps 1 to 3 of the FB process separately and store, for each h, m, the obtained $L_h^{(m)}(y_j)$ terms. Finally, in step 4, we sample each state from

$$x_j^{(t)} \sim \sum_{h=1}^{H'} \psi_{x_{j-1},h}^{(t)} L'_h(y_j)\delta(x_j, h) \tag{44}$$

$$= \frac{1}{M} \sum_{h=1}^{K'} \sum_{m=1}^{M} \psi_{x_{j-1},h}^{(t,m)} L_h^{(m)}(y_j)\delta(x_j, h), \tag{45}$$

In this way, we can preserve the advantages offered by the FB approach (faster mixing) while disposing of the layer of stochastic variability introduced by sampling the model parameters.

6.2.3 Preliminary results

In order to test the RBFB method, we constructed two simulated time series from switching AR processes of order 1 (Dataset 1) and 2 (Dataset 2) respectively. For each dataset, the true number of states was $H = 3$. Exactly

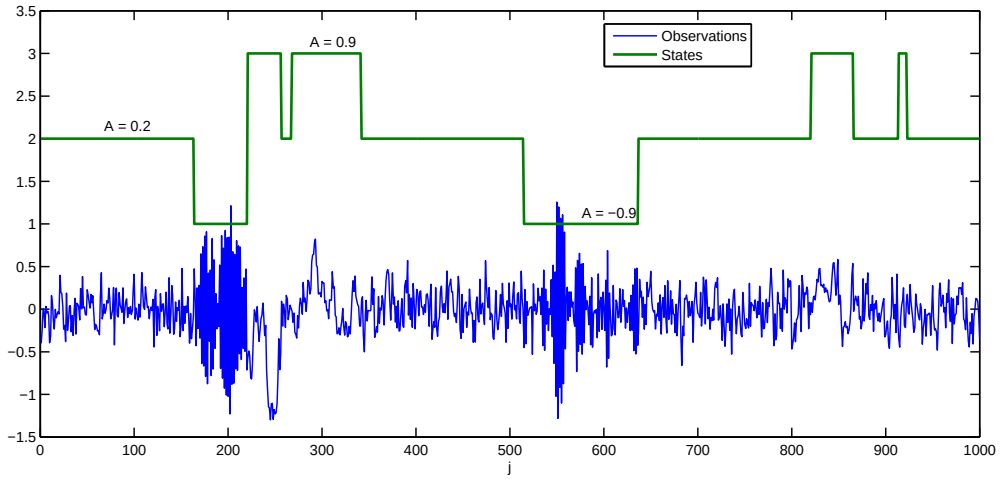

Figure 25: Simulated observations and state sequence in Dataset 1, generated from a 1-parameter autoregressive HMM, with known noise variance = 0.04 and 3 states.

one time series, with $d = 1000$, was generated from each model. These are shown in Figures 25 and 26, respectively.

In **Dataset 1**, we fixed the noise variance σ^2 at 0.04 and the true AR parameters to $A_1 = -0.9, A_2 = 0.2, A_3 = 0.9$. The true transition probability matrix, with diagonal elements kept much higher than non-diagonal ones in order to simulate persistent state dynamics, was fixed at

$$\boldsymbol{\psi} = \begin{pmatrix} 0.9900 & 0.0051 & 0.0049 \\ 0.0060 & 0.9896 & 0.0044 \\ 0.0056 & 0.0128 & 0.9817 \end{pmatrix}. \tag{46}$$

In **Dataset 2**, we kept $\sigma^2 = 1$, and true AR parameters for the three states as $A_1 = [0.49, 0.49], A_2 = [1, -0.5], A_3 = [-1, -0.5]$. The transition probability matrix was fixed at

$$\boldsymbol{\psi} = \begin{pmatrix} 0.9770 & 0.0037 & 0.0193 \\ 0.0085 & 0.9889 & 0.0026 \\ 0.0029 & 0.0564 & 0.9408 \end{pmatrix}. \tag{47}$$

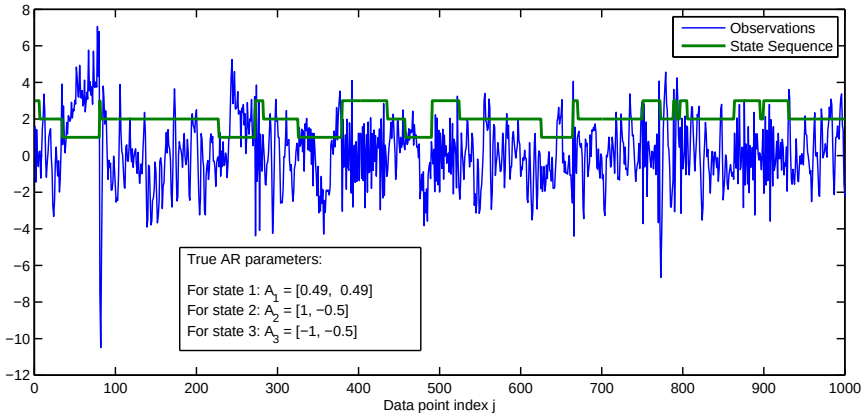

Figure 26: Simulated observations and state sequence in Dataset 2, generated from a 2-parameter autoregressive HMM, with known noise variance = 1 and 3 states.

Each time series y was then used as input to the three different Gibbs samplers. Each sampler was run for $T = 3000$ iterations, with each iteration containing 100 hyperparameter sampling steps. The $\alpha + \kappa$ and γ hyperparameters were both given the same Gamma priors, with the A and B parameters for both priors fixed at 1 and 0.01, respectively. Additionally, the hyperparameter $\kappa/(\alpha + \kappa)$ was used to sample the stickiness parameter, and was given a Beta prior with C and D parameters values of 10 and 1, respectively. Details on updates of the hyperparameters can be found in [34]. For ease of visualization, the maximum number of possible states H' was limited to 5. The state-sequence initialization was $s_j^{(0)} = 1, \forall j = 1, \ldots, d(= 1000)$. For the A parameter, a Gaussian prior was used with mean and standard deviation parameters $\mathbf{0}$ and $\mathbf{I}$, respectively. When sampling A, truncation was enforced in order to ensure that only stable AR processes were sampled [6, 95]. For analysis, after obtaining all the samples, we rejected the first 1000 as burn-in samples.

We first consider the results obtained from using the simulated first-order AR time series (Dataset 1). For this data, Figures 27, 28 and 29 show the posterior updates (at the end of each iteration) for the mean and standard deviations of the autoregressive coefficients for each of the $H' = 5$

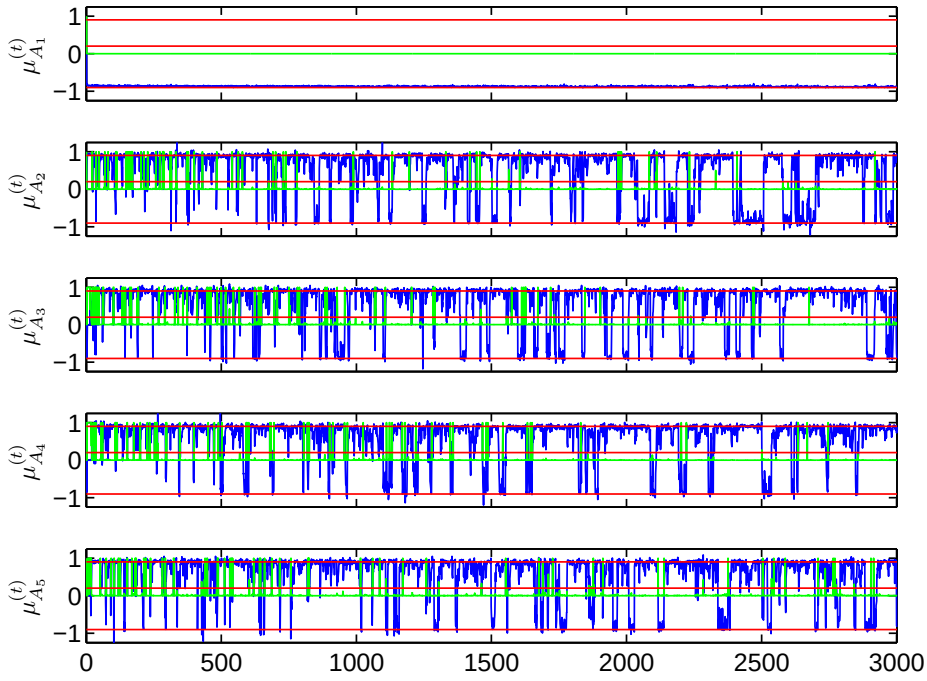

Figure 27: Posterior updates at the end of each iteration for mean (blue) and standard deviations (green) of AR parameters for each of the H = 5 possible states considered in the DG sampling strategy, when using Dataset 1 as the sampler input time series. Red lines indicate the true AR coefficients for each of the three true states.

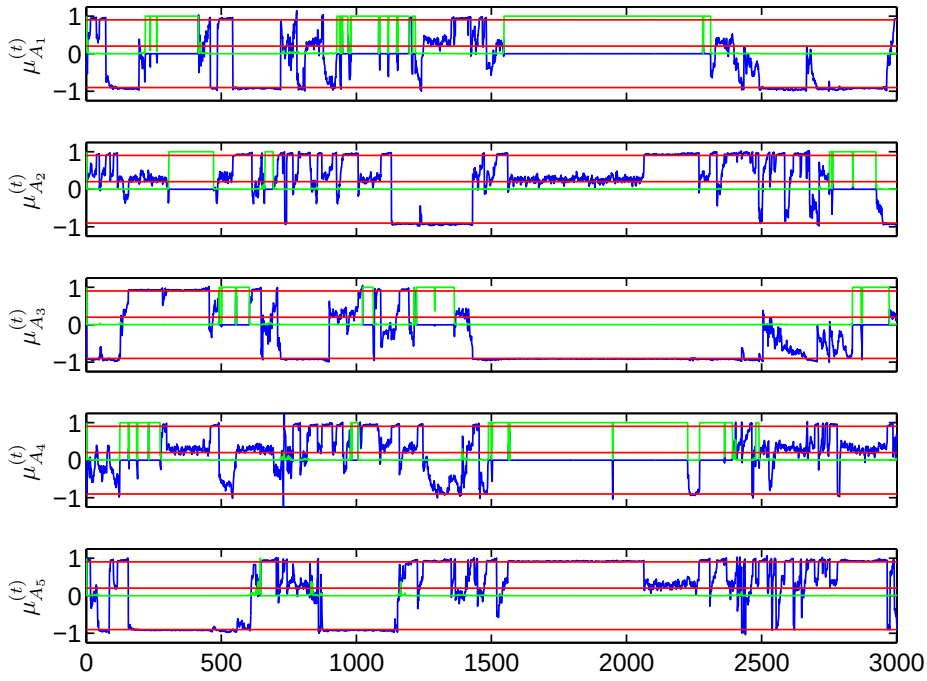

Figure 28: Posterior updates at the end of each iteration for mean (blue) and standard deviations (green) of AR parameters for each of the H = 5 possible states considered in the Full FB sampling strategy, when using Dataset 1 as the sampler input time series. Red lines indicate the true AR coefficients for each of the three true states.

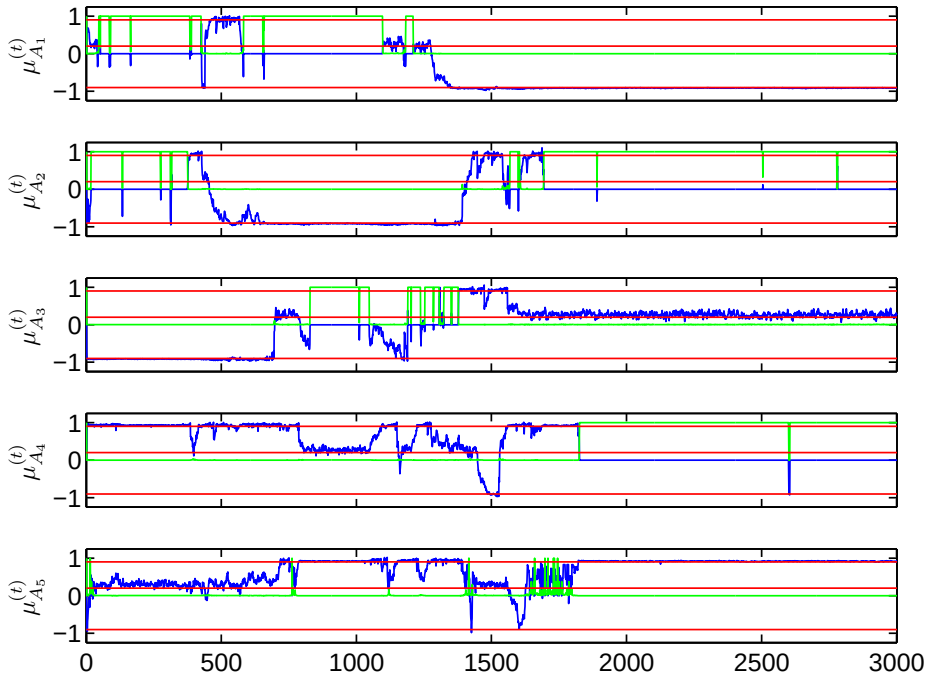

Figure 29: Posterior updates at the end of each iteration for mean (blue) and standard deviations (green) of AR parameters for each of the H = 5 possible states considered in the RBFB sampling strategy, when using Dataset 1 as the sampler input time series. Red lines indicate the true AR coefficients for each of the three true states.

states under consideration for the three sampling strategies studied. We can see that, using the DG sampling strategy (Figure 27), posterior-means for the AR coefficient of sampled-state 1 are very close to -0.9. However, the AR coefficient means for the other four sampled-states seem to exchange their allegiances from iteration to iteration. Thus, the coefficient-mean updates at the end of each iteration for states 2 to 5 vary widely, and so do the class-specific coefficient-standard deviations. If we were tasked with using these 3000 samples to estimate a state-value for each of the 1000 time points, and we chose, say the mode of the samples at each time point as the estimate, the only state we could uniquely identify would be the one associated with A_1. This problem is often referred to in the literature as "label switching". It is a feature of all sampling approaches to mixture component identification problems, and is what leads to erroneous segmentation performance when using it in the context of hidden Markov models. By comparison, when we use the FB strategy (Figure 28), we are able to reduce this rapid switching in AR coefficient parameters considerably. The least amount of switching occurs when using the RBFB sampler (Figure 29). If we were to use the same mode estimation strategy to identify the state-specific AR coefficients associated with each time point, we would have a much easier time.

We calculated the segmentation error (SE) for each iteration's sampled state sequence in the following way. Since each state is characterized by the corresponding AR parameters, we mapped any given sequence of state-indices $x_{i:j}$ to a sequence of corresponding AR coefficient (or AR coefficient-mean) vectors. For the 1st order AR process, SE for some iteration t is the mean squared error between sequences of AR means obtained from $x^t_{1:d}$ and from the true sequence $x_{1:d}$. For the 2nd-order case, we mapped each state's AR parameter vector to a set of characteristic-equation roots on the z-plane and calculated the corresponding complex angle. Thus we obtained a sequence of z-plane complex angles for each state-sequence. For instance, a section of the true state-sequence $s_{80:82} = [1, 3, 2]$ was mapped to the A sequence $[[0.49, 0.49], [-1, -0.5], [1, -0.5]]$, then to the z-plane as

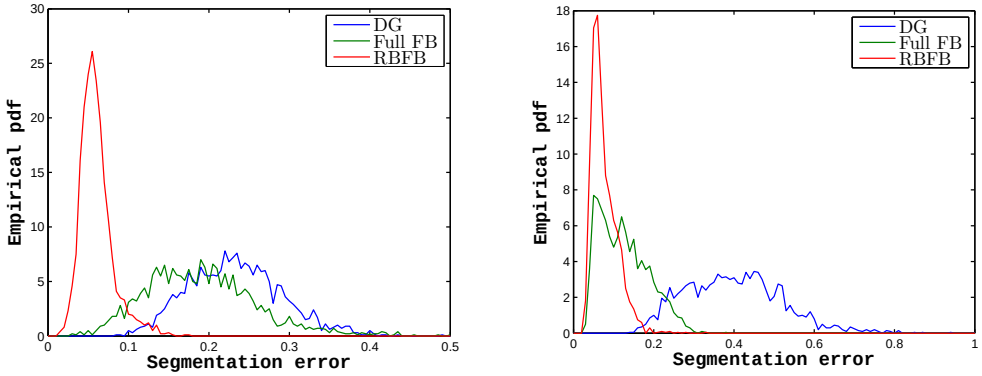

Figure 30: Empirical pdfs of the segmentation error between sampled and true state-sequences, when using various sampling strategies on the (top) 1st and (bottom) 2-nd order AR data. SE was calculated as explained in the text.

$[[0.987, -0.497], [-0.5 \pm 0.5i], [0.5 \pm 0.5i]]$, which yielded a sequence of complex angles (in radians) $[0, 2.356, 0.785]$. Similarly, we obtained sequences of z-plane angles for each iteration's sampled state-sequence $x_{1:d}^{(t)}$ and calculated the mean squared error between this and the true z-plane argument sequence. Figure 30 shows comparisons of the empirical pdfs constructed from the histograms of the SEs obtained from the DG, full FB and RBFB strategies' respective samples after burn-in. Using the full FB method yields lower SE on average compared to DG, and the RBFB further improves on this.

Figure 31 shows empirical pdfs constructed from histograms of parameter mean updates at the end of each sampling iteration (i.e., after step 4 in Section 6.2.2.1) for each state considered in the DG, full FB and collapsed FB strategies when applied to the simulated datasets. Note how in the bottom-left panel of Figure 31, which show results for the RBFB method when applied to a first-order switching AR process time-series, the means for state 3 (in red) are concentrated around 0. This is because for almost all iterations, this state was not instantiated when sampling for $x_{1:d}$, and the sampler generates A values from the prior, which is centered around 0. The same is true for state 2 (cyan) for a smaller number of iterations. For both datasets, when

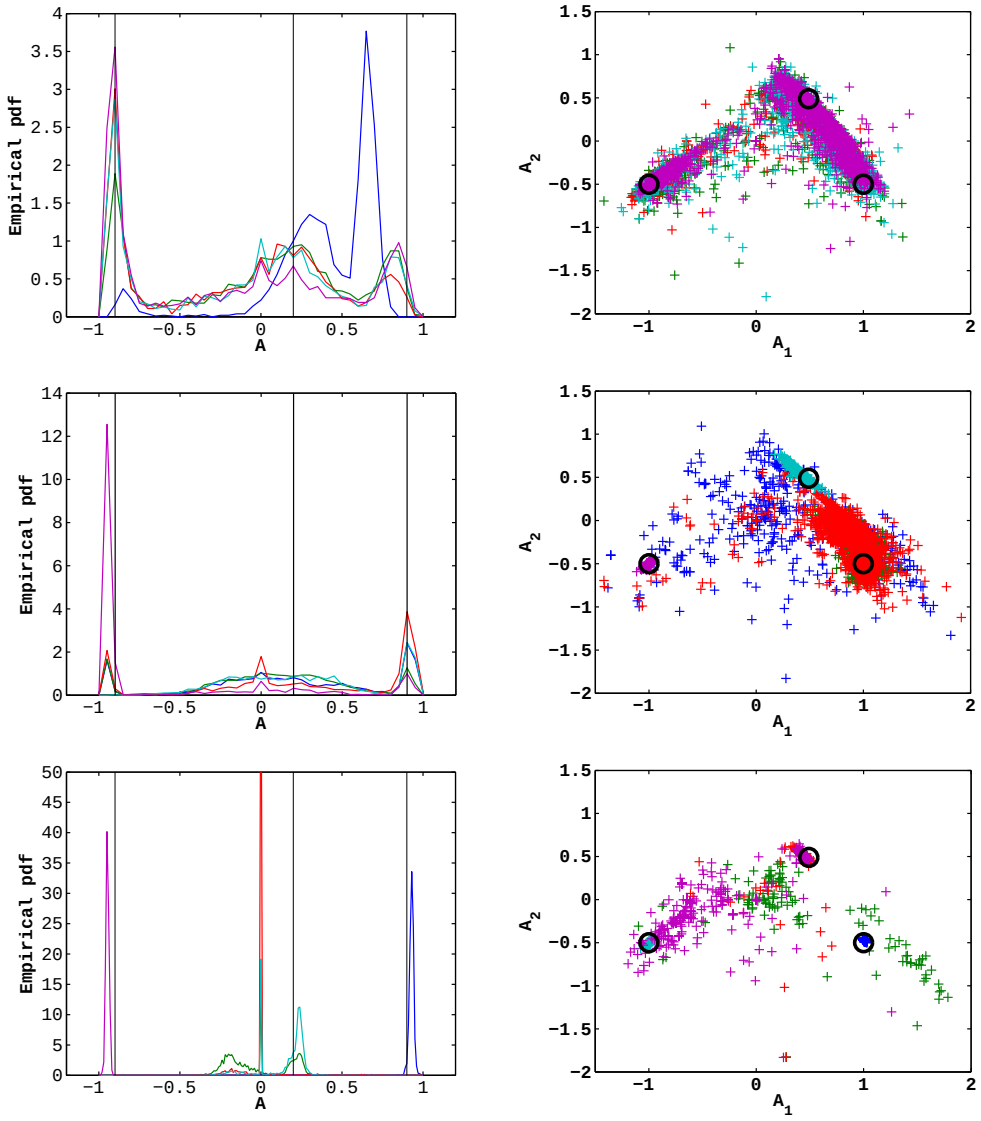

Figure 31: Empirical pdfs of posterior updates at the end of each iteration for mean of AR parameters for each of the $K' = 5$ possible states ($k = 1, 2, 3, 4, 5$) considered in (top panels) the DG, (center panels) the full FB and (bottom panels) RBFB sampling strategies. The left column shows results from Dataset 1 (switching AR process of order 1) and the right, those from Dataset 2 (switching AR process of order 2). Solid black lines in left-column figures and solid black circles in right-column figures indicate the true AR coefficient values.

using the RBFB method, the parameter means are scattered around the true AR coefficients, and it yields the least spread of all the samplers, while accurately identifying the three existing AR models in most iterations.

6.2.4 Outlook

In Figure 31, one can see that when using the DG method on 1-st order AR data, each considered state wandered through a wide region in the A space. Also, while the $A = -0.9$ state was identified in most iterations by all samplers, none of the considered states in the DG method could uniquely pick out the $A = 0.9$ or $A = 0.2$ states. When using the full FB method, it's clear that true states 1 and 3 are picked out by more than one of states considered, while the $A = 0.2$ state is ignored. Finally, the RBFB sampler yields much narrower empirical pdfs for each considered state, and is the only one that is able to pick out the $A = 0.2$ state consistently. Moreover, for most iterations, each state considered by the RBFB samples from unique regions in the A space, leading to less uncertainty in state-identification.

One limitation of the RBFB method is the significant overhead in terms of computational complexity; the present method performs the entire sampling chain M times for each sampled value of the parameter set, which increases the time taken for sampling M times. However, vectorizing the code (programs were implemented in Matlab) for the parameter sampling step enabled us to make the RBFB method more efficient. We performed some small simulations to test the computational savings from vectorization. Using observed time-series from dataset 1, we found that on average, the vectorized RBFB method completed the entire sampling chain in about 1.5 s per iteration, whereas code without vectorization took about 2.5 s per iteration on average. As expected, the direct Gibbs and FB methods outperformed the RBFB method in computation time. However, we note that further studies need to be done to analyze the computational complexity of the samplers, and to decrease it if possible. In addition, we are in the process of testing the effect of Rao-Blackwellization on the segmentation error with different model complexities (higher order AR or more general linear

dynamical systems), higher noise variance and for varying numbers of M.

6.3 Fetal risk scores

An overall important pitfall in FHR analysis is the lack of undisputed outcome measures for evaluating the proposed methods. Often, there seem to be no clear answers to the questions: is the fetal state best described by the base deficit, umbilical pH and Apgar scores (post-birth) or the physician categorization (pre-birth)? Is there a correlation between the found clusters and the available measures? We can attempt to answer these questions by (a) finding the best predictor models of post-birth and pre-birth fetal status via model-selection, and (b) constructing a "meta-score" that will weight the classification by the pH, Apgar and NICHD categorization appropriately and using this as a continuous output variable. The problem can thus be transformed from classification to regression, and finding the optimal weights could be cast as a parameter estimation problem.

6.4 Extensions to real-time monitoring.

The ultimate goal of our research is to demonstrate the feasibility of real-time classification of FHR-UP signals. One would like to have a system that can raise an alarm whenever an "at-risk" pattern of feature sequences is raised. Once we engage machine learning for the adopted models and learn them from the available data, we can implement them in real-time scenarios where classification is conducted by sequential processing. Our applied methodology allows for sequential computation of complete posterior distributions, which provide comprehensive pictures of evolutions of fetal state probabilities and a platform for informative visual and easily interpreatble displays for end-users.

A simple extension is to follow a sliding window approach where the system only classifies the latest T second window composed of a sequence of nonoverlapping t second segments. When the features of a new t second

segment are computed, the system removes the oldest t second segment and appends the latest segment. Then this T second window is classified anew. For any series of T-second epochs of FHR-UP data updated every t seconds, if more than E such epochs are deemed to fall in the "at-risk" category, one can raise an alarm. The "threshold-based-cumulative" classification methods described in Section 5.4.2 can give us some clues about the efficacy of such methods. There we saw that the epoch-threshold methods used with GM methods performed better than with the SVM and ES approaches, although it wasn't able to achieve the highest performance standards (there was a 9% drop in specificity to 72% compared to the cumulative approach).

One can also pursue other approaches to perform online clustering based on the GM and NPB frameworks. In particular, dynamic hierarchical Dirichlet processes [84, 87] seem to be promising alternatives. One crucial aspect of all these possibilities is the need to address computational issues because the applied techniques are computationally intensive and real-time processing requires that all the computations are completed on time.

6.5 Conclusion

The proposed research will ultimately benefit the end-users - patients and obstetric care professionals. Our main goal in this research was to explore the feasibility of the proposed methods for accurate clustering. The idea is to use powerful probabilistic machine learning techniques to stratify fetal risk from retrospective data, something existing obstetric practice and automated monitoring fail to do effectively. Eventually, we hope that the results of this project will be the building blocks of a full-scale implementation of real-time monitoring that will leverage the constant streams of patient data available in obstetric clinics to improve fetal risk prediction.

Bibliography

[1] Z. Alfirevic, D. Devane, and G. Gyte. *Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour (Review)*. John Wiley & Sons, Inc., 2007.

[2] A. Alonso-Betanzos, B. Guijarro-Berdiñas, V. Moret-Bonillo, and S. Lóez-Gonźalez. The NST-EXPERT project: the need to evolve. *Artificial Intelligence in Medicine*, 7(4):297–313, Aug 1995.

[3] D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de-Sá, and L. Pereira-Leite. SisPorto 2.0: A program for automated analysis of cardiotocograms. *The Journal of Maternal-Fetal Medicine*, 9(5):311–318, 2000.

[4] D. Ayres-de Campos, C. Costa-Santos, J. Bernardes, et al. Prediction of neonatal state by computer analysis of fetal heart rate tracings: the antepartum arm of the SisPorto® multicentre validation study. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 118(1):52–60, 2005.

[5] D. Ayres-de Campos, P. Sousa, A. Costa, J. Bernardes, et al. Omniview-SisPorto® 3.5-a central fetal monitoring station with online alerts based on computerized cardiotocogram+ ST event analysis. *Journal of Perinatal Medicine*, 36(3):260, 2008.

[6] E. R. Beadle and P. M. Djurić. Uniform random parameter generation of stable minimum-phase real ARMA(p, q) processes. *Signal Processing Letters, IEEE*, 4(9):259–261, 1997.

[7] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, pages 577–584, 2001.

[8] J. Bernardes, A. Costa-Pereira, D. Ayres-de Campos, H. P. v. Geijn, and L. Pereira-Leite. Evaluation of interobserver agreement of cardiotocograms. *International Journal of Gynecology & Obstetrics*, 57(1):33–37, Apr. 1997.

[9] D. R. Bickel. Robust and efficient estimation of the mode of continuous data: The mode as a viable measure of central tendency. *Journal of Statistical Computation and Simulation*, 73(12):899–912, Dec. 2003.

[10] C. M. Bishop. Novelty detection and neural network validation. In *Vision, Image and Signal Processing, IEEE Proceedings-*, volume 141, pages 217–222. IET, 1994.

[11] C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 1. Springer New York, 2006.

[12] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[14] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997.

[15] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[16] C. Campbell and Y. Ying. *Learning with Support Vector Machines*. Morgan & Claypool Publishers, 2011.

[17] V. Chudáček. *Automatic Analysis of Intrapartum Fetal Heart Rate*. PhD thesis, Czech Technical University, Prague, 2011.

[18] V. Chudáček, J. Spilka, B. Rubackova, M. Koucky, G. Georgoulas, L. Lhotska, and C. Stylios. Evaluation of feature subsets for classification of cardiotocographic recordings. In *Computers in Cardiology, 2008*, pages 845–848, 2008.

[19] S. L. Clark, M. P. Nageotte, T. J. Garite, R. K. Freeman, D. A. Miller, K. R. Simpson, M. A. Belfort, G. A. Dildy, J. T. Parer, R. L. Berkowitz, et al. Intrapartum management of category II fetal heart rate tracings-towards standardization of care. *American Journal of Obstetrics and Gynecology*, 2013.

[20] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[21] C. Costa Santos, J. Bernardes, P. Vitanyi, and L. Antunes. Clustering fetal heart rate tracings by compression. In *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on*, pages 685–690. IEEE, 2006.

[22] F. Cowan, M. Rutherford, F. Groenendaal, P. Eken, E. Mercuri, G. M. Bydder, L. C. Meiners, L. Dubowitz, and L. S. de Vries. Origin and timing of brain lesions in term infants with neonatal encephalopathy. *The Lancet*, 361(9359):736–742, 2003.

[23] F. G. Cunningham, K. J. Leveno, S. L. Bloom, J. C. Hauth, D. J. Rouse, and C. Y. Spong. *Williams Obstetrics*. McGraw-Hill Medical, 2010.

[24] S. Dash, K. Chon, S. Lu, and E. Raeder. Automatic real time detection of atrial fibrillation. *Annals of biomedical engineering*, 37(9):1701–1709, 2009.

[25] S. Dash, J. Muscat, J. G. Quirk, and P. M. Djurić. Implementation of NICHD diagnostic criteria for feature extraction and classification of fetal heart rate signals. In *Signals, Systems and Computers (ASILOMAR)*, pages 1684–1688, 2011.

[26] S. Dash, J. Muscat, J. G. Quirk, and P. M. Djurić. Classification of fetal heart rate series. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 629–632. IEEE, 2012.

[27] S. Dash, J. G. Quirk, and P. M. Djurić. Learning dependencies among fetal heart rate features using Bayesian networks. In *Engineering in Medicine and Biology Society (EMBC), 2012*, pages 6204–6207. IEEE, 2012.

[28] L. D. Devoe, M. Ross, C. Wilde, M. Beal, A. Lysikewicz, J. Maier, V. Vines, I. Amer-Wrahlin, H. Lilja, H. Norén, et al. United states multicenter clinical usage study of the STAN 21 electronic fetal monitoring system. *American journal of obstetrics and gynecology*, 195(3):729–734, 2006.

[29] S. Dong, B. Boashash, G. Azemi, B. E. Lingwood, and P. B. Colditz. Detection of perinatal hypoxia using time-frequency analysis of heart rate variability signals. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 939–943. IEEE, 2013.

[30] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2000.

[31] J. Echeverria, B. Hayes-Gill, J. Crowe, M. Woolfson, and G. Croaker. Detrended fluctuation analysis: a suitable method for studying fetal heart rate variability? *Physiological measurement*, 25:763, 2004.

[32] C. Elliott, P. A. Warrick, E. Graham, and E. F. Hamilton. Graded classification of fetal heart rate tracings: association with neonatal metabolic acidosis and neurologic morbidity. *American Journal of Obstetrics and Gynecology*, 202(3):258–e1, 2010.

[33] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Bayesian nonparametric inference of switching dynamic linear models. *Signal Processing, IEEE Transactions on*, 59(4):1569–1585, 2011.

[34] E. B. Fox. *Bayesian nonparametric learning of complex dynamical phenomena*. PhD thesis, Massachusetts Institute of Technology, 2009.

[35] B. Frank, B. Pompe, U. Schneider, and D. Hoyer. Permutation entropy improves fetal behavioural state classification based on heart rate analysis from biomagnetic recordings in near term fetuses. *Medical and Biological Engineering and Computing*, 44(3):179–187, 2006.

[36] R. K. Freeman, T. J. Garite, M. P. Nageotte, and L. A. Miller. *Fetal heart rate monitoring*. Lippincott Williams & Wilkins, 2012.

[37] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2):131–163, 1997.

[38] N. Friedman and D. Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50(1):95–125, 2003-01-01.

[39] A. Georgieva, S. Payne, M. Moulden, and C. Redman. Artificial neural networks applied to fetal monitoring in labour. *Neural Computing & Applications*, pages 1–9.

[40] G. Georgoulas, D. Gavrilis, I. Tsoulos, C. Stylios, J. Bernardes, and P. Groumpos. Novel approach for fetal heart rate classification introducing grammatical evolution. *Biomedical Signal Processing and Control*, 2(2):69–79, 2007.

[41] G. Georgoulas, C. Stylios, V. Chudacek, M. Macas, J. Bernardes, and L. Lhotska. Classification of fetal heart rate signals based on features selected using the binary particle swarm algorithm. In *World Congress on Medical Physics and Biomedical Engineering 2006*, pages 1156–1159. 2007.

[42] G. Georgoulas, C. Stylios, G. Nokas, and P. Groumpos. Classification of fetal heart rate during labour using hidden Markov models. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 3, pages 2471–2475vol.3, july 2004.

[43] G. Georgoulas, D. Stylios, and P. Groumpos. Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. *Biomedical Engineering, IEEE Transactions on*, 53(5):875 –884, May 2006.

[44] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer, 2003.

[45] L. Goggin, R. Eikelboom, and M. Atlas. Clinical decision support systems and computer-aided diagnosis in otology. *Otolaryngology–Head and Neck Surgery*, 136(4 suppl):s21, 2007.

[46] H. Goncalves, A. Rocha, D. Ayres-de Campos, and J. Bernardes. Frequency domain and entropy analysis of fetal heart rate: appealing tools for fetal surveillance and pharmacodynamic assessment of drugs. *Cardiovascular & Haematological Disorders-Drug Targets*, 8(2):91–98, 2008.

[47] E. F. Hamilton, M. C. Glaude, M. Macieszczak, and P. A. Warrick. Apparatus for monitoring the condition of a fetus, Dec. 19 2012. EP Patent 1,489,960.

[48] R. Hammond and T. Freer. Application of a case-based expert system to orthodontic diagnosis and treatment planning: a review of the literature. *Australian orthodontic journal*, 14(3):150, 1996.

[49] H. Helgason, P. Abry, P. Gonçalv́es, C. Gharib, P. Gaucherand, and M. Doret. Adaptive multiscale complexity analysis of fetal heart rate. *Biomedical Engineering, IEEE Transactions on*, 58(8):2186–2193, Aug. 2011.

[50] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors. *Bayesian Nonparametrics*. Cambridge University Press, 2010.

[51] F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer, 2007.

[52] J. Jezewski, J. Wrobel, and K. Horoba. Comparison of Doppler ultrasound and direct electrocardiography acquisition techniques for quantification of fetal heart rate variability. *Biomedical Engineering, IEEE Transactions on*, 53(5):855–864, 2006.

[53] M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes. *Advances in Neural Information Processing Systems*, 14:841, 2002.

[54] K. Kaluzynski, M. Berson, L. Pourcelot, and T. Palko. Detection of fetal breathing and cardiac movements and rhythms in ultrasonic doppler signal recorded on the surface of the maternal abdomen. *Medical and Biological Engineering and Computing*, 31(4):405–411, 1993.

[55] S. M. Kay. *Fundamentals of Statistical Signal Processing*. Prentice-Hall, 1993.

[56] R. Keith and K. Greene. Development, evaluation and validation of an intelligent system for the management of labour. *Baillière's clinical obstetrics and gynaecology*, 8(3):583–605, 1994.

[57] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145, 1995.

[58] N. Lavrač, P. Flach, and B. Zupan. *Rule Evaluation Measures: A Unifying View*. Springer, 1999.

[59] J. Lawn, S. Cousens, J. Zupan, et al. 4 million neonatal deaths: When? where? why? *The Lancet*, 365(9462):891–900, 2005.

[60] J. Lemaire, J. Schaefer, L. Martin, P. Faris, M. Ainslie, and R. Hull. Effectiveness of the Quick Medical Reference as a diagnostic tool. *Canadian Medical Association Journal*, 161(6):725–728, 1999.

[61] J. C. Lindon, E. Holmes, and J. K. Nicholson. Metabonomics in pharmaceutical R&D. *FEBS J*, 274(5):1140–1151, Mar 2007.

[62] J. S. Liu, W. H. Wong, and A. Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1994.

[63] M. Lopes, H. Marin, N. Ortega, and E. Massad. The use of expert systems on the differential diagnosis of urinary incontinence. *Revista da Escola de Enfermagem da USP*, 43(3):704–710, 2009.

[64] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi. A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 4(2):R1–, 2007.

[65] M. MacDorman and S. Kirmeyer. Fetal and perinatal mortality, United States, 2005. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 57(8):1, 2009.

[66] G. A. Macones et al. The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: Update on definitions, interpretation, and research guidelines. *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, 37(5):510–515, 2008.

[67] M. Malik. Heart rate variability. *Annals of Noninvasive Electrocardiology*, 1(2):151–181, 1996.

[68] R. Mantel, H. Van Geijn, F. Caron, J. Swartjes, E. Van Woerden, and H. Jongsma. Computer analysis of antepartum fetal heart rate: 1. baseline determination. *International Journal of Bio-medical Computing*, 25(4):261–272, 1990.

[69] C. Marchant, L. Fisk, M. Patel, and D. Suárez. An expert system approach to the assessment of hepatotoxic potential. *Chemistry & Biodiversity*, 6(11):2107–2114, 2009.

[70] J. A. Martin, B. E. Hamilton, P. D. Sutton, S. J. Ventura, F. Menacker, and M. L. Munson. Birth: Final data for 2002. *National Vital Statistics Report*, 52:1, 2003.

[71] M. D. D. McNeely. The use of expert systems for improving test use and enhancing the accuracy of diagnosis. *Clinics in Laboratory Medicine*, 22(2):515–528, Jun 2002.

[72] P. Müller and R. Mitra. Bayesian nonparametric inference: Why and How. *Bayesian Analysis*, 8(2):1–34, 2013.

[73] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[74] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, 2004.

[75] Y. Noguchi, F. Matsumoto, K. Maeda, and T. Nagasawa. Neural network analysis and evaluation of the fetal heart rate. *Algorithms*, 2(1):19–30, 2009.

[76] M. Oestergaard, M. Inoue, S. Yoshida, W. Mahanani, F. Gore, S. Cousens, J. Lawn, and C. Mathers. Neonatal mortality levels for 193 countries in 2009 with trends since 1990: A systematic analysis of progress, projections, and priorities. *PLoS Med*, 8(8):e1001080, 2011.

[77] M. M. Ohayon. Improving decisionmaking processes with the fuzzy logic approach in the epidemiology of sleep disorders. *Journal of Psychosomatic Research*, 47(4):297–311, Oct 1999.

[78] O. Palomäki, T. Luukkaala, R. Luoto, and R. Tuimala. Intrapartum cardiotocography–the dilemma of interpretational variation. *Journal of Perinatal Medicine*, 34(4):298–302, 2006.

[79] S. Papadimitriou and A. Bezerianos. Nonlinear analysis of the performance and reliability of wavelet singularity detection based denoising for doppler ultrasound fetal heart rate signals. *International Journal of Medical Informatics*, 53(1):43–60, 1999.

[80] J. Pardey, M. Moulden, and C. W. Redman. A computer system for the numerical analysis of nonstress tests. *American Journal of Obstetrics and Gynecology*, 186(5):1095–1103, 2002.

[81] J. Parer, T. King, S. Flanders, M. Fox, and S. Kilpatrick. Fetal acidemia and electronic fetal heart rate patterns: Is there evidence of an association? *Journal of Maternal-Fetal and Neonatal Medicine*, 19(5):289–294, 2006.

[82] J. T. Parer and E. F. Hamilton. Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation. *American Journal of Obstetrics and Gynecology*, 203(5):451–e1, 2010.

[83] D. T. Parry and E. Parry. *Medical informatics in obstetrics and gynecology*. Medical Information Science Reference, 2009.

[84] A. Rabaoui, N. Viandier, E. Duflos, J. Marais, and P. Vanheeghe. Dirichlet process mixtures for density estimation in dynamic nonlinear modeling: Application to GPS positioning in urban canyons. *Signal Processing, IEEE Transactions on*, 60(4):1638–1655, 2012.

[85] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[86] A. Reddy, M. Moulden, and C. W. G. Redman. Antepartum high-frequency fetal heart rate sinusoidal rhythm: computerized detection and fetal anemia. *American Journal of Obstetrics and Gynecology*, 200(4):407.e1–407.e6, 2009.

[87] L. Ren, D. Dunson, S. Lindroth, and L. Carin. Dynamic nonparametric bayesian models for analysis of music. *Journal of the American Statistical Association*, 105(490), 2010.

[88] P. Resnik and E. Hardisty. Gibbs sampling for the uninitiated. Technical report, DTIC Document, 2010.

[89] D. Rotin, N. Petrovichev, A. Pavlovskaia, V. Nikitaev, E. Berdnikovich, A. Pronichev, and D. Popov. [expert system for the diagnosis of thyroid neoplasms]. *Arkhiv patologii*, 66(2):47.

[90] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507, 2007.

[91] C. Samarghitean and M. Vihinen. Bioinformatics services related to diagnosis of primary immunodeficiencies. *Current Opinion in Allergy and Clinical Immunology*, 9(6):531, 2009.

[92] S. Saria, D. Koller, and A. Penn. Learning individual and population level traits from clinical temporal data. In *Proc. Neural Information Processing Systems (NIPS), Predictive Models in Personalized Medicine workshop*, 2010.

[93] S. L. Scott. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):pp. 337–351, 2002.

[94] R. Seufert, F. Woernle, and F. Casper. [Computer-assisted cardiotocogram analysis–from descriptive to perinatal expert system]. *Zentralblatt für Gynäkologie*, 122(6):328, 2000.

[95] P. Shcherbakov and F. Dabbene. On random generation of stable polynomials. In *Control Applications,(CCA) & Intelligent Control,(ISIC), 2009 IEEE*, pages 406–411. IEEE, 2009.

[96] M. Signorini, G. Magenes, S. Cerutti, and D. Arduini. Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. *Biomedical Engineering, IEEE Transactions on*, 50(3):365–374, 2003.

[97] J. Spilka, V. Chudacek, M. Koucky, L. Lhotska, M. Huptych, P. Janku, G. Georgoulas, and C. Stylios. Using nonlinear features for fetal heart rate classification. *Biomedical Signal Processing and Control*, In Press, Corrected Proof, 2011.

[98] P. J. Steer. Has electronic fetal heart rate monitoring made a difference? In *Seminars in Fetal and Neonatal Medicine*, volume 13, pages 2–7. Elsevier, 2008.

[99] E. Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, Massachusetts Institute of Technology, 2006.

[100] A. Sundström, D. Rosén, and K. Rosén. *Fetal surveillance*. Göteborg: Neoventa, 2000.

[101] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Artificial Neural Networks, 1995., Fourth International Conference on*, pages 442–447. IET, 1995.

[102] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[103] C. Wang, D. M. Blei, and D. Heckerman. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*, 2012.

[104] J. Wang, P. Liu, M. F. She, S. Nahavandi, and A. Kouzani. Bag-of-words representation for biomedical time series classification. *Biomedical Signal Processing and Control*, 8(6):634–644, 2013.

[105] P. Warrick, E. Hamilton, D. Precup, and R. Kearney. Identification of the dynamic relationship between intrapartum uterine pressure and fetal heart rate for normal and hypoxic fetuses. *Biomedical Engineering, IEEE Transactions on*, 56(6):1587 –1597, 2009.

[106] P. Warrick, E. Hamilton, D. Precup, and R. Kearney. Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography. *Biomedical Engineering, IEEE Transactions on*, 57(4):771 –779, 2010.

[107] P. A. Warrick, E. F. Hamilton, R. E. Kearney, and D. Precup. A machine learning approach to the detection of fetal hypoxia during labor and delivery. In *Proc. of the Twenty-Second Innovative Applications of Artificial Intelligence Conf*, pages 1865–1870, 2010.

[108] G. Xu, X. Lu, H. Kong, X. Shi, X. Zhao, J. Tian, and G. Lu. [Applications and progresses of expert system on chromatography]. *Se pu= Chinese journal of chromatography/Zhongguo hua xue hui*, 23(5):449, 2005.