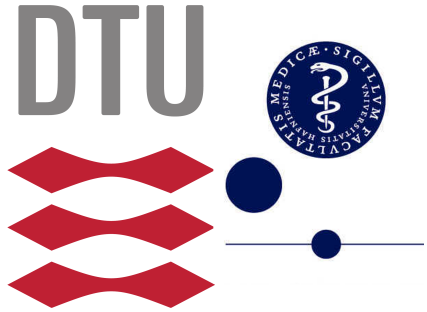
Characterization and Discrimination of Pathological Electrocardiograms using Advanced Machine Learning Methods

Andreas Seliger
&
Lasse Bergenholz Hansen



Kongens Lyngby 2013
IMM-M.Sc.-2013-14

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-M.Sc.-2013-14

Summary

Cardiac arrhythmia and other heart related conditions are potentially life-threatening, making fast and accurate diagnosis vital. This thesis describes an approach to characterize and discriminate ECGs by applying machine learning methods. The investigation concerns the discrimination of subjects suffering from the inherited genetic disorder Long QT type 2 (LQT2) from a normal population. Applying 10-second raw ECGs as input, various hidden Markov models are trained for each group. The generative properties of the models are assessed and the log-likelihoods of the test ECGs are applied in an initial classification scheme. Further, the Support Vector Machine is included to improve the classification using the log-likelihoods of multiple hidden Markov models.
ECG simulations from the trained hidden Markov models produced recognizable waveforms and some of the expected morphological changes, seen in LQT2 subjects, were observable in the simulated ECGs. The best classification result observed was a classification accuracy of 78.1% with a corresponding specificity of 78.2% and a sensitivity of 78.2%. Experience showed, however, that biological noise and power line interference in the ECG affected the classification, but it appears that the application of hidden Markov models using raw ECG data is well suited for the purpose of ECG characterization and discrimination.

Resume

Hjertearytmier og andre hjerterelaterede lidelser er potentielt livstruende, hvorfor en hurtig og præcis diagnose er altafgørende. Denne afhandling beskriver en tilgang til at karakterisere og diskriminere EKG'er ved anvendelse af maskinlæringsmetoder. Undersøgelsen handler om diskriminationen af genetisk nedarvet lang QT type 2 testpersoner fra en normal population. Ved at anvende 10-sekunders rå EKG'er som input, trænes forskellige skjulte Markov modeller for hver gruppe. Modellernes generative egenskaber undersøges, og log-sandsynligheden for test EKG'erne anvendes i en tidlig klassifikationsfase. Herudover inkluderes Support Vector Machine for at forbedre klassifikationen ved at anvende log-sandsynlighederne fra flere skjulte Markov modeller. EKG simulationer fra de trænede skjulte Markov modeller viste genkendelige bølgeformer, og nogle af de forventede morfologiske forandringer, der ses hos LQT2 patienter, kunne observeres i de simulerede EKG'er. Den bedst fundne klassifikationsnøjagtighed var 78,1 % med en tilsvarende specificitet på 78,2 % og en sensitivitet på 78,2 %. Det viste sig dog at biologisk og 50 Hz støj i EKG'erne påvirkede klassifikationen, men det fremgår alligevel, at modellering af rå EKG data ved anvendelse af skjulte Markov modeller, er velegnet til karakterisering og diskriminering af EKG'er.

Preface

This thesis is written by Andreas Seliger and Lasse Bergenholz Hansen and was prepared at the department of Informatics and Mathematical Modeling at the Technical University of Denmark in collaboration with the Department of Biomedical Sciences, Heart and Circulatory Research Section, University of Copenhagen, in fulfillment of the requirements for acquiring a M.Sc. in Biomedical Engineering. The thesis was produced between 1th of September 2012 and the 1th of March 2013 and corresponds to a workload of 30 ECTS.

Supervisors:

Associate Professor, Ph.D., Ole Winther

Associate Professor, M.D., Jørgen K. Kanters

M.Sc. Esben Vedel-Larsen

Lyngby, 01-March-2013

Andreas Seliger

Lasse Bergenholz Hansen

Acknowledgements

We would like to show our gratitude to Associate Professor, Ph.D., Ole Winther, Associate Professor, M.D., Jørgen K. Kanters and M.Sc. Esben Vedel-Larsen whose encouragement, guidance and support from the initial to the final level enabled us to develop an understanding of the subject. A special thanks to the Heart and Circulatory Research Section at the University of Copenhagen for providing the data.

Contents

Abbreviations

Abbreviation	Description
I,II,V1,V2,V3,V4,V5,V6	ECG leads
AUC	Area Under (ROC) Curve
AV node	AtrioVentricular node
BMI	Department of Biomedical Sciences
CVDHMM	Continuous Variable Duration HMM
ECG	Electrocardiogram
EM	Electrode Motion
EMG	Electromyogram
FULL	Full transition
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
Inf	Infinite
KKT	Karush-Kuhn-Tucker
LSE	Log-Sum-Exp trick
NaN	Not a Number
LR	Left-Right
LR1	Left-Right, one forward degree of freedom
LR2	Left-Right, two forward degrees of freedom
LQT2	Long QT type 2 syndrome
MA	Muscle Artifacts
P-wave	ECG waveform
PCA	Principle Component Analysis
PLI	Power Line Interference
PV	Premature Ventricular
QRS-complex	ECG waveform
RMS	Root Mean Square
ROC	Receiver Operator Characteristic
RR-interval	heartbeat duration
SAN	Sinoatrial Node
SD	Standard Deviation
SNR	Signal to Noise Ratio
SVM	Support Vector Machine
T-wave	ECG waveform
U-wave	ECG waveform
WCT	Wilsons Terminal Central
WGN	White Gaussian Noise

Table 1: Abbreviations

Chapter 1

Introduction

Cardiac arrhythmia, myocardial infarction and other heart related conditions are potentially life threating, making fast and accurate diagnosis vital. The heart conditions are either inherited, induced by drugs or related to lifestyle. The electrocardiogram (ECG) is one of the most widely used non-invasive diagnostic tools for monitoring cardiac disease. It enables the clinician to register the electrical activity of the heart in an inexpensive way. In Denmark the leading experts in the field of inherited and drug induced arrhythmias reside at the Department of Biomedical Sciences (BMI), Heart and Circulatory Research Section, University of Copenhagen. A general approach used at BMI, when examining ECGs, is to explore different stationary features, such as amplitude and duration measures, derived from median heart beats formed from 10-second ECGs. Participating in research at BMI, the potential of creating a method able to capture the temporal variation of a 10-second ECG was identified by the authors.

In this thesis we aim to develop a general modeling and classification method able to characterize and discriminate normal and pathological ECGs. The aim is the construction of a model that could aid in the diagnosis of cardiac disease or as a tool used in ECG-based heart research. The model should be able to capture temporal variations in the ECG, variations between ECG leads and should be independent of the currently applied software algorithms used at BMI. In short, the method should be able to provide both characterization and discrim-

ination of ECGs.

To investigate the performance of the model, a population of normal ECGs and a population of Long QT Type 2 syndrome (LQT2) subjects were applied. With the LQT2 syndrome being a highly researched, inherited genetic disorder at BMI, it could be validated whether the model is able to capture some of the expected morphological changes in the ECGs.
Furthermore, cardiac arrhythmias are one of the most feared adverse reactions to drugs, in which most cases occur due to a block of cardiac ion channels inhibiting certain potassium currents. The same currents are also inhibited in LQT2 patients. Being able to discriminate normal and LQT2 ECGs, the model could possibly also be applied in the evaluation of drug safety.

The thesis commences with an introduction (Chapter 2) to the heart, its electrical conduction system and relates this to the pathophysiology of LQT2 in order to clarify the clinical motivation and give an understanding of the signals obtained with the ECG technique. In Chapter 3, the ECG method is explained, giving an understanding of how the physiological processes are expressed in the ECG signal, how the ECG signals are obtained and how biological and machine generated noise can degrade the information content in the ECGs. Chapter 4 presents selected works within the field of ECG characterization and classification providing insight into the state-of-the-art machine learning methods applied. The acquired knowledge is applied in determining the choice of methods to be implemented in this work. The concepts behind machine learning and the theory of the chosen machine learning models are elucidated in Chapter 5, while the applied ECG data, the model training and the classification setup are presented in Chapter 6. The generative properties of models are also illustrated here. The verification of the implementation is performed and a test setup for the classification tolerance with regards to noise is outlined. The generative capability of the model is explored in Chapter 7 together with the results of classifying the normal and LQT2 ECGs. The effect of noise on the classification accuracy is also treated therein. Chapter 8 discusses the different aspects of the models and the general setup of the method and rounds off discussing interesting work to be undertaken in future endeavors. The conclusion of the project is given in Chapter 9.

Sometimes the heart sees what is invisible to the eye.
H. Jackson Brown, Jr.

Chapter 2

Physiological Background

The electrophysiological processes that are captured by the ECG and the pathophysiology of the LQT2 syndrome are addressed in the following. A brief description of the anatomy of the heart is given in section 2.1 and the electrophysiology of the cardiac action potential at the cellular level is presented in section 2.2. Section 2.3 describes the electrical conduction system of the heart, which in part explains the appearance of the measured ECG. Finally, the pathophysiology of Long QT syndrome is explained in section 2.4.

2.1 General Anatomy of the Heart

The human heart is roughly situated in the middle of the thorax. It consists of four chambers with the right and left atria situated superiorly and the right and left ventricles situated inferiorly. The left ventricle and atrium are separated from the right ventricle and atrium by the septum and as such the heart can be viewed as two separate pumps. The left ventricle and atrium are larger and have thicker walls than their right counterparts, thus the heart appears as unsymmetrical. A frontal plane section of the heart is presented in Figure 2.2. The differences in chamber size and wall thickness are due to the physiological function of the heart, where the left part supplies systemic circulation through the

aorta and the right part supplies pulmonary circulation through the pulmonary artery. The positive pressure difference between the systemic and pulmonary circulation requires a stronger, and therefore, larger left side of the heart.
Besides asymmetry, relative to the thorax and due to the structure of the heart itself, the heart is also rotated around three anatomical axes. That is, rotation with respect to the frontal plane, rotation with respect to the transverse plane and longitudinal rotation (base-apex axis) [29] [48].

2.2 Cardiac Action Potential

The cell membrane is polarized due differences in charge between the immediate inside and immediate outside of the cell membrane. An action potential is a transient change in the membrane potential. The semi-permeable cell membrane facilitates the existence and maintenance of the resting membrane potential, which occurs due to an electrochemical equilibrium. The main ions involved in the membrane potential are Na^+, K^+, Ca^{2+}, Cl^- and negative proteins. Two forces act on these; a chemical force and an electrical force collectively called electrochemical forces. The cell membrane permeability of K^+ and Cl^- are far larger than for the other ions [9] hence playing the main role in the formation of the resting potential. The cell membrane is impermeable to the negative proteins within the cell. The concentration of K^+ is largest within the cell and the concentration of Na^+ and Ca^{2+} is largest outside the cell. These concentration gradients are sustained by energy driven transport over the cell membrane. K^+ tend to diffuse out of the cell, down its concentration gradient, leaving the inside of the cell more negative. The electric force of the negative proteins inside the cell attracts the K^+ back to the cell membrane and into the cell, resulting in an accumulation of positive charges outside the cell. When the chemical forces acting on K^+ to move out of the cell are in equilibrium with the electrical forces acting on K^+ to move into the cell, a negative resting membrane potential of around -90 mV is established. The reason Cl^- does not influence the resting membrane potential, despite its high membrane permeability, is due do the fact that its equilibrium potential is close to the resting membrane potential [46].
When the cell membrane is sufficiently stimulated an action potential may occur that involves Na^+, K^+ and Ca^{2+}. The four phases of the cardiac action potential are presented in Figure 2.1. In phase 0 (depolarization phase) Na^+ channels are activated resulting in Na^+ influx and depolarization but they are inactivated shortly thereafter. In phase 1 (early repolarization) the cell is briefly repolarized due to an efflux of K^+ through K^+ channels and a closing of Na^+ channels. In phase 2 (plateau phase) Ca^{2+} channels open and counteract the effect of the K^+ efflux, creating the plateau phase that distinguishes the cardiac

action potential from the skeletal muscle action potential. Phase 3 (repolarization) begins when the increasing efflux of K^+ exceeds the decreasing influx of Ca^{2+} through the closing of Ca^{2+} channels. Furthermore, the Na^+ channels begin to open, resulting in repolarization of the cell. In phase 4 (resting potential) the cell returns to resting conditions. The steady influx of Na^+ is counteracted by the energy driven Na^+-K^+ pump. Ca^{2+} concentrations are restored by the $3Ca^{2+}$-Na^- energy driven pumps [9]. The involvement of the Ca^{2+} channels and the resulting plateau phase (phase 2) makes the cardiac action potential duration larger by a factor of 100 than actions potentials in skeletal muscle [30]. The action potential will cease to exist at one location with time but it can activate neighboring regions or cells since the cardiac tissue is electrically connected (see section 2.3). Hence, the activation can propagate in any direction in a large number of cells creating complex wavefronts on larger scale [30].

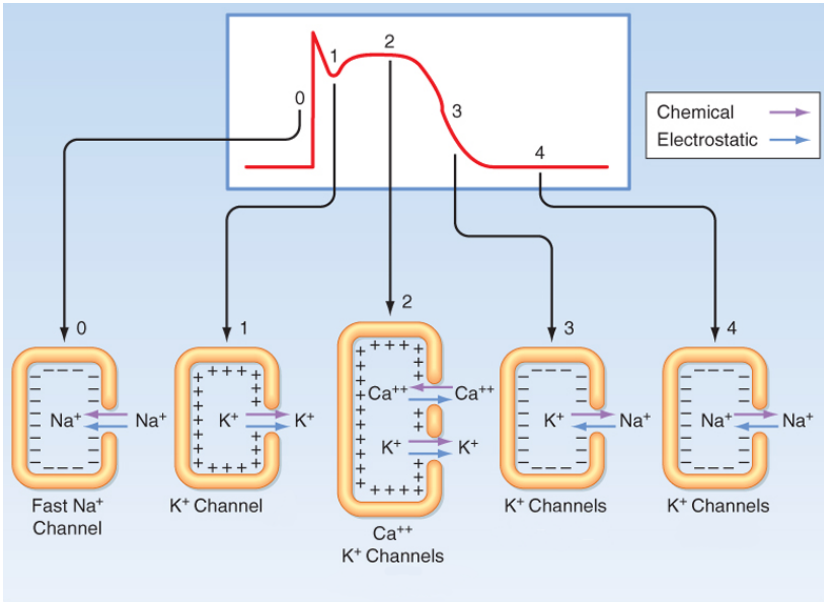


Figure 2.1: The four phases of the cardiac action potential. In phase 0 the cell membrane depolarizes. Phase 1 is the early repolarization which is counteracted in phase 2, called the plateau phase. Phase 3 is the repolarization phase that ends with reestablishment of resting conditions in phase 4, see text for details. Modified from [9].

2.3 The Electrical Conduction System of the Heart

In the normal heart the action potentials initiating the contraction of the heart occur in the sinoatrial node (SAN), located in the right atrium as shown in Figure 2.2. The atria are electrically insulated from the ventricles with only the atrioventricular node (AV node) as an electrical passage way. The AV node delays the propagation of activation such that the atria contract before the ventricles. When the action potentials have passed the AV node they first propagate through the bundle of His. Subsequently the propagation continues along the right and left bundle branches that extends through the septum before the action potentials reach the Purkinje fibers which extend through the inner ventricular walls. The propagation speed in the conduction system after the passage of the AV node is several times that of the surrounding cardiac tissue [30]. Besides the conduction system of the heart the myocardial cells are further coupled by GAP junctions, which provide direct connection of the cytoplasm of the cell. These cellular junctions provide a low resistance passage way for ionic currents, and therefore the cellular activation will spread (intracellularly) through the myocardium. Thus, the heart effectively behaves as an syncytium [30].

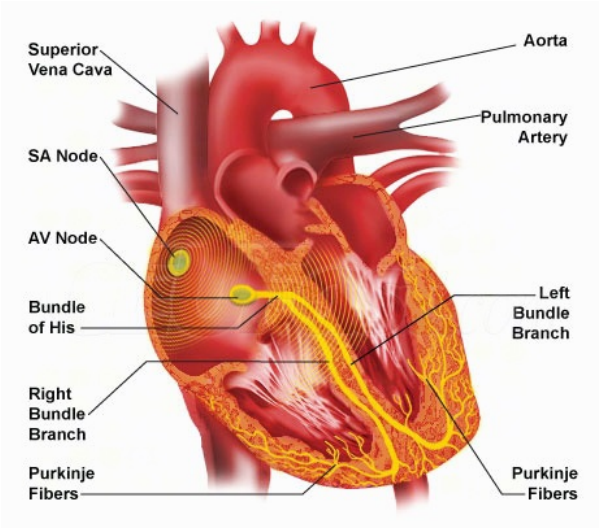


Figure 2.2: Illustration showing the general anatomy of the heart and its electrical conduction system. Adopted from [13].

2.4 Pathophysiology of Long QT Syndrome

Long QT syndrome can be either congenital or acquired (drug induced). Congenital long QT syndrome is characterized by an abnormal cardiac repolarization observed in the ECG as a prolonged QT interval and changes in the T-wave morphology. The prevalence is estimated to be 1: 5,000-10,000 [34] and the majority remain asymptomatic [20]. The phenotype[1] is extremely varied however, including syncope, ventricular arrhythmias and sudden cardiac death. The most typical ventricular arrhythmia is Torsades de Pointes which is described by an observation in the ECG where the QRS complex is twisted around the baseline. Prognosis in symptomatic cases is poor and if not treated 20% die within one year and 50% die within 10 years [20]. Symptoms are related to the cardiac system when the inheritance pattern is autosomal[2] dominant (Romano-Ward Syndrome). In the autosomal recessive case, however (Jervell and Lange Nielsen), a further clinical manifestation is deafness [20]. Diagnosis is usually based on QT prolongation (corrected for heart rate) in the ECG although other T-wave morphology parameters have been investigated recently and some found clinically relevant [11, 45, 10, 35]. Further, genetic tests, epinephrine tests and exercise tests are applied in the diagnosis.
Considering the autosomal dominant case, 12 gene mutations all related to cardiac ion channels are known. Both potassium (K^+), sodium (Na^+) and calcium (Ca^{2+}) are involved; long QT1, QT2, QT5, QT6 and QT7 are potassium current or potassium current related. Long QT3, QT10, QT9, QT12 includes sodium current or sodium related current. Finally long QT8 and QT4 includes calcium current or calcium related current [20]. The most common types of long QT syndrome are LQT1 and LQT2 covering 90% of LQTS [34]. This work is based on two gender and age matched populations of normal subjects compared with LQT2 subjects, and therefore emphasis will be put on the LQT2 type in the following.

In section 2.2 the action potential was described at the cellular level. The importance of potassium (K^+) flux and channels were presented without describing the channels in detail. Several types of K^+ channels are known to exist, all of which are involved in repolarization as they facilitate outward flux of potassium. One of such channels is the rapid delayed rectifying I_{Kr} channel, also known as an HERG channel. In LQT2 subjects mutations in the KCNH2 gene that codes for channel-proteins results in abnormal function of I_{Kr} channel (HERG) expressed as abnormal repolarization due to loss of potassium current [20]. Clinical findings in the ECG are related to the abnormal repolarization with prolonged QT interval, notched T-waves and T-wave alternans and ar-

[1] Expressed heredity.
[2] Other than sex chromosomes.

rhythmias [11, 34, 54].
Treatment involves β-adrenergic blocking agents, pacemakers, implantable cardioverter defibrillators and others. β-adrenergic blockers are effective with LQT2 [31]; besides the pace making abilities of the SAN and the cardiac muscle, the heart is innervated by parasympathetic and sympathetic nerve fibers. Parasympathetic innervation generally lowers heart rate whereas sympathetic innervation increases heart rate. As clinical manifestations of LQT2 often occur in stressful situations [20] the β-blockers are effective in that they block the receptors of the sympathetic neurotransmitters norepinephrine and epinephrine.
In acquired LQT2 certain drugs can affect the I_{Kr} channel mimicking the abnormalities caused by gene mutation. Graff et al. [11] showed that distinct patterns in the T-wave morphology seen in congenital LQT2 could quantify drug induced ECG changes in normal subjects.

Chapter 3

The Electrocardiogram

The ECG is a non-invasive diagnostic tool that measures the electrical activity of the heart via electrodes placed on the skin. The 12-lead ECG technique is over 70 years old and the most widely used cardiac diagnostic tool in clinical practice [27]. With the physiological processes underlying the ECG having been presented in the previous chapter, the formation of the ECG and related issues are presented in the following; section 3.1 describes the formation of the ECG, the placement of the leads and relates the observed signal to the underlying physiological process, section 3.2 addresses lead redundancy and explains the significance of the individual leads, section 3.3 compares a normal and a LQT2 ECG and finally sections 3.4 and 3.5 introduce ECG noise and noise filtering, respectively.

3.1 The ECG signal

The limb leads; lead I, II and III are obtained by placing the skin electrodes on the right arm (R), left arm (L) and left foot (F). Differences in the measured potentials yields these leads; $I = \Phi_L - \Phi_R$, $II = \Phi_F - \Phi_R$ and $III = \Phi_F - \Phi_L$, where Φ denotes the potential. Further, three augmented leads can be obtained; aVR, aVL and aVF, by subtracting the augmented average of the limb

potentials from each of the limb potentials, respectively. The average of the limb potentials is found with a setup called Wilson's Central Terminal (WCT), where the sum of the three potentials is measured after a $5k\Omega$ resistor connected to each. The augmentation is performed by omitting one of the resistances of the WCT; namely the resistance that is connected to the measurement electrode [30]. Finally there are the six precordial leads, $V1$-$V6$, and a reference electrode placed on the right leg. Hence the 12-lead system is comprised of 10 physical "leads". The precordial leads are placed in accordance with specific anatomical indicators and their potentials are measured with respect to the average of the limb potentials (WCT) without augmentation. Figure 3.1 presents the placement of the precordial leads. In the following the propagation of action

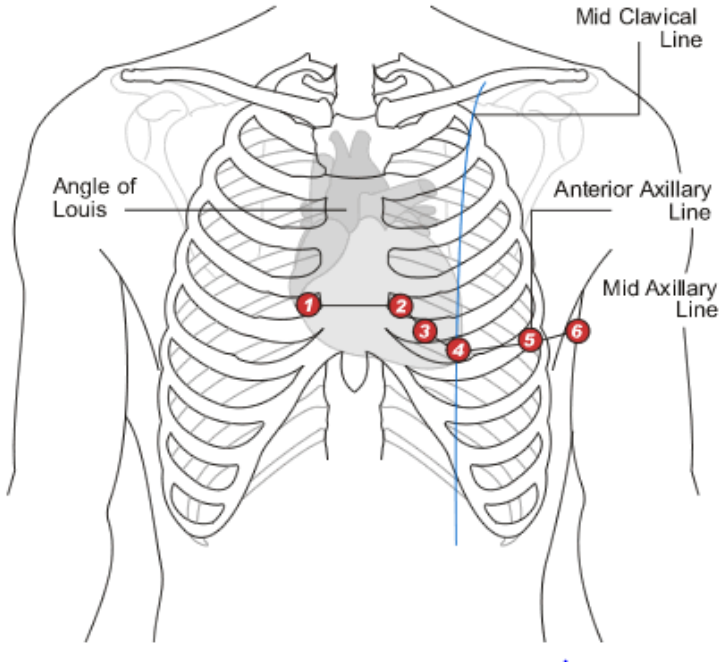


Figure 3.1: Placement of precordial leads ($V1$-$V6$) on the basis of anatomical indicators. Adopted from [49].

potentials through the conduction system of the heart is described with respect to the appearance of the typical ECG presented in Figure 3.2. To aid this description the term resultant vector is introduced. At any instant of time during the cardiac depolarization the propagation will occur in a number of directions. Assigning a potential vector to wavefronts traveling in these directions, a resultant vector can be calculated at any instant of time (dipole source assumption)

[59]. In the following it is assumed that the resultant vector from depolarization will produce a positive signal when the wavefront is propagating towards a positive electrode. Similarly it will produce a negative signal when the wavefront is propagating away from the (positive) electrode as the resultant vector is pointing away. In repolarization, the situation is opposite in that the resultant vector representing a depolarizing wavefront traveling towards a positive electrode will result in a negative signal and vice versa. When the atria are activated from the SAN the (depolarizing) action potentials spread from the right atrium to the left resulting in a vector that is fairly aligned with the septum. Considering the transverse plane, lead $V5$ is placed close to this direction and is considered in the following. This atrial depolarization appears as a positive P-wave in the ECG, shown in Figure 3.2. The action potentials propagate through the AV node to the septum where the left part of the septum depolarizes first, giving rise to a resultant vector pointing to the right. This is observed as the negative Q-wave. Subsequent apical depolarization results in a resultant vector aligned with the septum, initiating an increase in the ECG amplitude which eventually gives rise to a peak called the R-wave. As the left ventricular wall is thicker than the right the depolarization continues longer on the left side, resulting in a resultant vector oriented to the left. This orientation contributes to the continued rise in the ECG forming the peak of the R-wave. The resultant vector shifts upwards, but maintains its leftward orientation throughout the rest of the depolarization phase. It then decreases in magnitude until a minimum is reached, termed the S-wave, which finalizes the QRS complex. The onset of atrial repolarization is not visible in the ECG due to the contraction of the ventricles. Finally the ventricular repolarization begins in a transmural fashion from the epicardium to the endocardium resulting in a vector still oriented to the left, since direction and sign of the repolarizing wavefronts are opposite the depolarizing wavefronts. The repolarization of the ventricles is observed as the T-wave in the ECG and is strongly dependent on the heart rate, in that it becomes narrower and occurs closer to the QRS complex at high heart rates. Following the T-wave, a U-wave can appear under some conditions (not presented in Figure 3.2). The origin is not well explained, but it is probably due to delayed repolarization [59].

3.2 Polarity and Redundancy of ECG Leads

Traditionally ECG leads are divided into unipolar and bipolar leads reflecting measuring variation in voltage of a single electrode or between electrodes, respectively [59]. A true unipolar signal is measured with respect to an infinitely remote reference. Traditionally the limb leads, I, II and III are viewed as bipolar

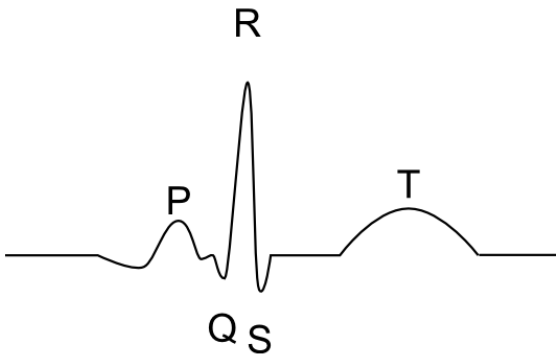


Figure 3.2: Idealized example of an ECG showing how the atrial depolarization is observed as the P-wave, the ventricular depolarization is observed as the QRS complex and finally the ventricular repolarisation is observed as the T-wave. Adopted from [62].

leads as they measure the electrical activity of the heart from a distance using one positive and one negative electrode. Considering the unipolar electrode the concept of an infinitely remote reference point, or an indifferent electrode, is not feasible in the human body as it constitutes a volume conductor. The WTC is an attempt to produce an indifferent electrode that approximates the potential at infinity [30]. The WTC, however, does not approximate zero potential [32] but rather an average of the limb potentials as mentioned earlier. Even so, the WTC still serves as a satisfactory reference [30]. Despite this limitation the augmented leads and precordial leads are termed unipolar and leads I, II and II are termed bipolar.

As a consequence of Kirschoffs law it must hold that lead I + II = III. In fact any two of leads I, II, III, aVR, aVL and aVF contain the same information as the rest as they are all derived from the same three measuring points [30]. Due to the placement of the precordial leads close to the heart, with respect to the WCT, they detect unipolar components of diagnostic value due to the proximity to the frontal part of the heart [30]. In other words, when measuring a complex source from a distance (e.g. the limb leads) the dipole assumption makes sense, but it deteriorates when the measuring electrodes are placed close to the heart [17]. The redundancy explains that the 12 lead ECG is represented by only 8 leads; I, II and V1-V6.

Several other systems for recording the electromyographic signals of the heart have been suggested. These methods include systems with a smaller or larger number of leads or different lead placement. Also, the technique of body surface potential mapping where 200 electrodes may be applied has been introduced. Donnely et. el. [47] provides a retrospective review of different systems in terms

of signal content and diagnostic value suggesting both limitations and improvements over the 12 lead system. Despite promising results with some systems the 12 lead ECG system remains the most widely accepted cardiac diagnostic tool in clinical practice.
An example of the 8-lead ECG from a normal subject is presented in Figure 3.3.

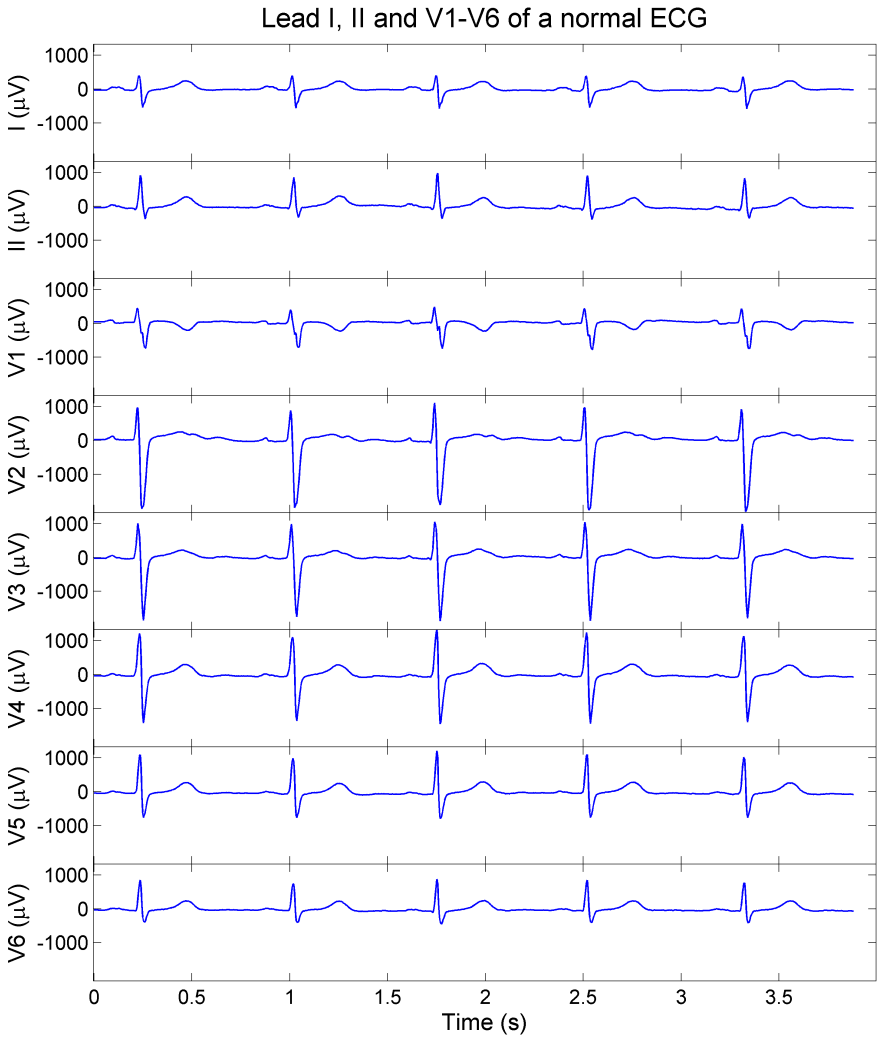


Figure 3.3: Normal ECG: 8 leads of a 12-lead ECG from a typical normal subject. Leads I and II represent the electrical activity of the entire heart whereas the precordial leads represent more localized variations in the electrical activity of the heart in the transversal plane. See text for details.

Comparing with Figure 3.2 it is evident that the signal represents five consecutive heartbeats. Further it is noticed that the leads vary in amplitude and shape despite sharing the same general excursions. Factors like the skin-electrode impedance and other noise sources can influence the measured ECG greatly as described in section 3.4.
The limb leads I and II reflect the electrical activity of the entire heart in the frontal plane. The precordial leads reflect the electrical activity of the heart in the transversal plane and are considered to capture more localized variations as indicated in Figure 3.4. Leads V1-V2 primarily reflect the right ventricle and septal wall, while leads V3-V4 reflect the anterior wall of the left ventricle and leads V5-V6 reflect the lateral wall of the left ventricle [6].

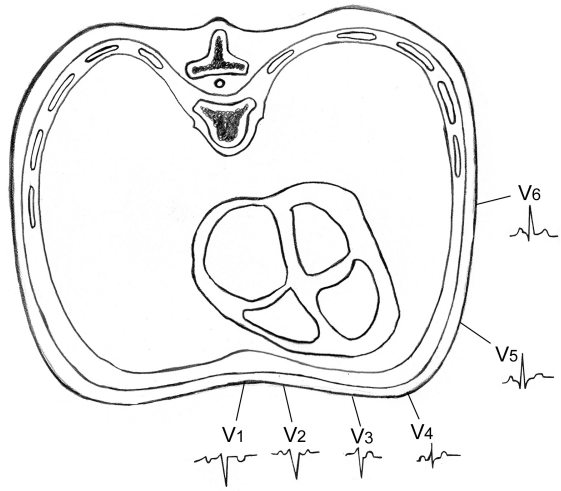


Figure 3.4: Transversal cross section of the heart showing which localized regions of the heart each of the precordial leads reflect. Leads V1-V2 primarily reflect the right ventricle and septal wall, leads V3-V4 reflect the anterior wall of the left ventricle and leads V5-V6 reflect the lateral wall of the left ventricle. Adopted from [21].

3.3 A Normal and a LQT2 ECG

Figure 3.3 shows 8 leads of a normal ECG. It graphically demonstrates that the leads show the same excursions, to a large extent, but still exhibit inter lead variation. Presenting results, it is sometimes desirable to show a single lead

rather showing all 8 leads as in Figure 3.3 as this becomes excessive. When considering LQT2 ECGs the T-wave is of interest as explained in section 2.4. Struijk et. al. [33] argues that when considering the T-wave, a good choice of a single lead would be lead V5 due to its physical position with respect to the principle direction of the T-wave loop.

The classification of ECGs performed in this work is based on all (8) leads and as such the rationale above has no impact in that context. However, when evaluating the generative properties of the models it is desirable to attempt to rediscover known morphological differences between normal and LQT2 ECGs (T-wave morphology). However these would not necessarily be the main foundation of the discriminative properties of the model. In summary, it is sometimes convenient to show only one lead when presenting data, and a reasonable candidate in the context of this work is lead V5.
To further evaluate lead V5 a principle component analysis (PCA) was per-

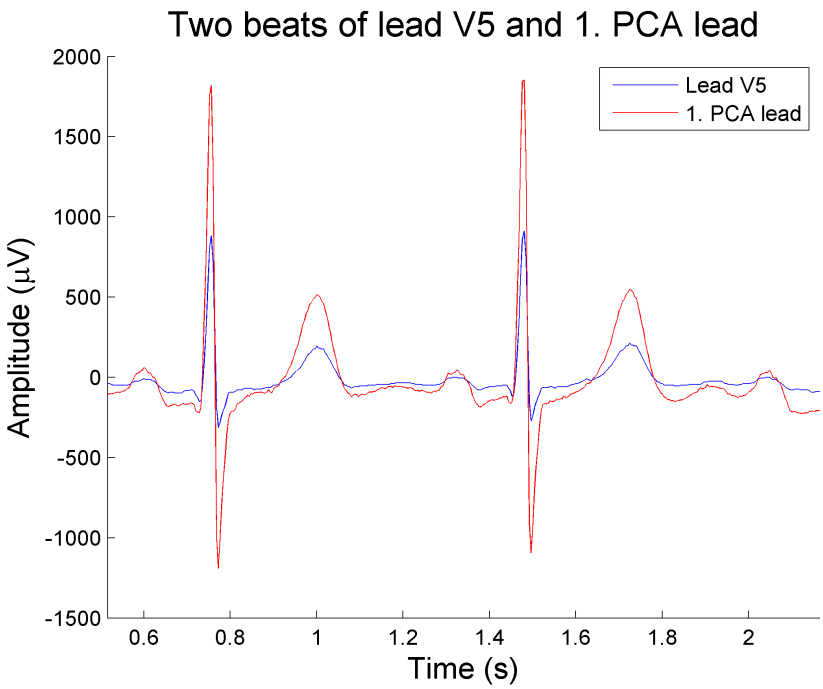


Figure 3.5: ECG from a normal subject corresponding to that of Figure 3.3. The blue graph presents lead V5 and the red graph presents the 8-lead ECG data projected on to the first principle direction found using principle component analysis.

formed where the 8-lead ECG data was projected on to the first principle direction, denoted as the first PCA lead. In the entire study population of normal and LQT2 ECGs the first PCA lead on average explains 70.1±9.7% of the variation. Lead V5 and the first PCA lead are presented in Figure 3.5 for the same normal subject as in Figure 3.3. The figure indicates that lead V5 captures the general excursions of the first PCA lead but that the P-wave, QRS-complex and the T-wave are of lower amplitude. The P-wave is less well-defined and the U-wave in particular is difficult to distinguish in lead V5, but the comparison still supports applying lead V5 when presenting data. In Figure 3.6 lead V5 of the ECG from the same normal subject is compared with a typical LQT2 ECG. The description in section 2.4 suggests morphological changes in the T-wave as

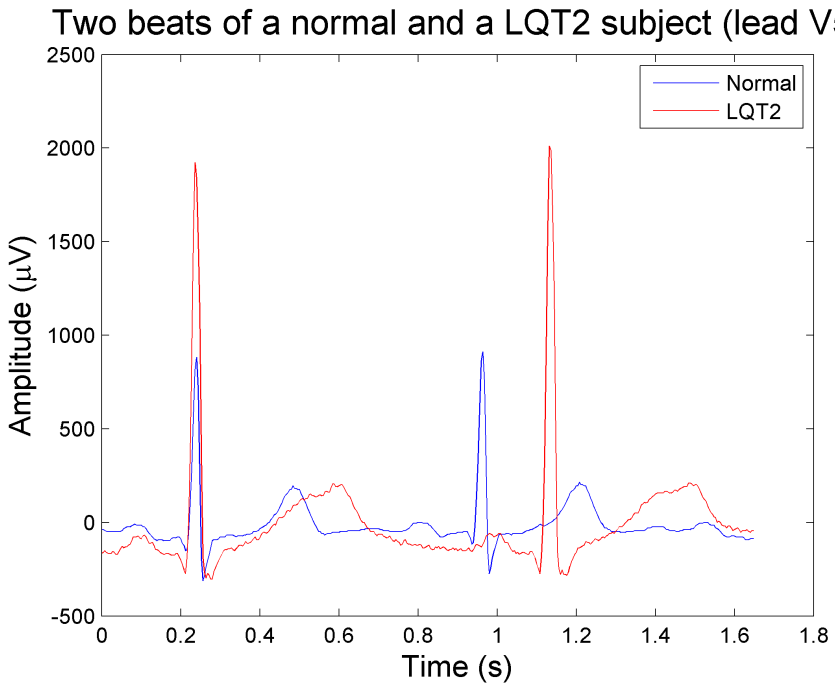


Figure 3.6: Comparison of a normal and an LQT2 ECG. The blue graph represents Lead V5 from a normal subject corresponding to that of Figure 3.3 and the red graph represents a typical LQT2 subject.

well an obviously longer QT interval. Before evaluating the appearance of the T-wave, it is noted that the ECGs presented correspond to different heart rates. As mentioned earlier, the T-wave is strongly dependent on the heart rate, in that it becomes narrower and occurs more closely to the QRS complex at high heart rates. Hence part of the difference in the appearance of the T-wave and transi-

tion to next beat may be contributed to heart rate. However, the difference in shape and duration of the T-wave is still distinct beyond heart rate differences. Besides the longer duration of the T-wave in the LQT2 ECG a notch appears before the maximum of the T-wave, which is not uncommon. The amplitude and baseline difference is probably due to measurement conditions and subject variations not related to LQT2.

3.4 Noise Sources in the ECG Signal

Even within normal ECGs the biological variation is large. Further, the ECG quality is very dependent on the clinician performing the measurement as well as the subject itself. In this work it is desirable that the discriminative properties of the models are able to capture a general trend in the ECGs within each group. The variation in each group can be thought to consist of inter-subject variations as well as various noise types. In regard to the latter there is an undesirable situation in which one of the groups to be classified contains a higher amount of noise, e.g. a bias or very low frequency noise, that may contribute substantially to the classification. Capturing which group of ECGs are most noisy in the classification would diminish the classification abilities as this is specific to the study population. In section 3.4.1 the generally accepted types of noise in ECGs are presented. Section 3.4.2 visualizes the effect of noise by adding various noise sources to a normal ECG and finally a brief overview of ECG filtration is provided in section 3.5.

3.4.1 Five types of ECG Noise

Electromyographic signals (EMGs) arising from extremities, can produce noise of a bandwidth that overlaps or exceeds the ECG bandwidth [25]. The interface between skin and electrode is described by the skin-electrode impedance. The preparation and condition of skin leads to differences in skin-electrode impedance, which contributes to the ECG noise [63]. Changes in skin-electrode impedance, due to electrode movement caused by e.g. skin stretch or perspiration, can produce low frequency noise that is observed as baseline wander. Further, depending on the nature of the electrode movements, the noise can mimic the elements of the ECG and have a wider bandwidth than baseline wander. This behavior are referred to as electrode motion. This type of noise is usually caused by intermittent mechanical forces acting on the electrodes [25]. PLI (Power Line Interference) is also a well-known contributor. The mentioned noise sources are common in that that they are controllable in some sense.

However differing clinical environments, operators and subjects leave questions about the degree to which this control is achieved.
To visualize the effect of noise on the ECG, five types of noise are added to a normal ECG; 1: baseline wander (BW), 2: muscle artifacts (MA), 3: electrode motion (EM), 4: white Gaussian noise (WGN) and 5: power line interference (PLI). Noise types 1-3 were obtained from the The Massachusetts Institute of Technology-Beth Israel Hospital Noise Stress Test Database (MIT-BIH NST Database) [1, 25]. The database consists of 3 half-hour noise recordings with two channels each and are described at `physionet.org`: The noise recordings were made using physically active volunteers and standard ECG recorders, leads, and electrodes; the electrodes were placed on the limbs in positions in which the subjects' ECGs were not visible. The three noise records were assembled from the recordings by selecting intervals that contained predominantly baseline wander, muscle (EMG) artifact, and electrode motion artifact.
As described in [25] the selection procedure was based on visual inspection and the half hour signals were formed by concatenating segments of similar noise and amplitude. Due to the selection procedure, the noise signals do not exclusively contain one of the three types of noise. Thus, some overlap between the noise signals is present [44]. Muscle artifacts and electrode motion were especially hard to isolate from baseline wander [25]. Both channels of the three half-hour signals were applied in this work.
Noise types 4 and 5 were implemented in `MATLAB`® as described in the following. WGN are essentially physically unrealizable since the bandwidth will be limited by a finite sampling frequency. Further, the random numbers generated on a computer are in actuality pseudo random. In order to simulate a totally uncorrelated and normally distributed signal, the function `randn.m` in `MATLAB`® was used. This function generates pseudo independent, pseudo random numbers whereby WGN is simulated. The PLI was defined as a sinusoidal oscillation consisting of a natural frequency and three overtones, with a random phase and an amplitude inversely proportional to the overtone number.

3.4.2 Applying Noise Sources Individually to Visualize the Effect

The recorded ECGs were sampled at 500 Hz for the LQT2 patients and at 250 Hz for the normal subjects. Noise types 1-3 from the MIT-BIH NST Database were sampled at 360 Hz. The LQT2 ECGs and the noise recordings were downsampled to 250 Hz using the `MATLAB`® function `resample.m`. The PLI sinusoidal was also defined at this frequency.
An ECG from a normal subject, appearing free of noise, was chosen. Lead I, II and V1-V6 were corrupted with different randomly sampled noise signals in

each subject, thus creating a set of noise samplings corresponding to the number of leads. When adjusting the noise level, this set of noise signals was fixed such that only the magnitude was adjusted at the different noise levels. For noise types 1-3, the 10 s noise signal required for each baseline lead, was sampled by randomly selecting a starting point from the half-hour noise signal and then sampling 10 s of consecutive data. Both channels of each of the two half-hour signals were sampled. The WGN and PLI were simulated after the same principle, i.e. individual realizations for each lead were fixed when increasing the noise by adjusting the magnitude.
For each subject the magnitude of a given noise signal was identified by merging the leads and the noise signals, respectively, to two long signals which facilitates the calculation of an overall SNR. Thus, the merged signals provided means of calculating which magnitude of the merged noise signal corresponded to a given overall SNR. Subsequently each of the individual noise signals were adjusted with this calculated magnitude. As a consequence, the SNR's stated in the following correspond to the overall SNR. Figure 3.7 shows lead V5 of the normal ECG with noise applied, following the procedure described above. All types of noise are shown for three levels of noise; SNR: 10 dB, SNR: 0 dB and SNR: -4 dB (the corresponding root mean square amplitude ratios are 3.2, 1 and 0.6). As these SNR's corresponds to the overall SNR described above, the SNR of lead V5 depicted in Figure 3.7 may deviate from the stated levels.

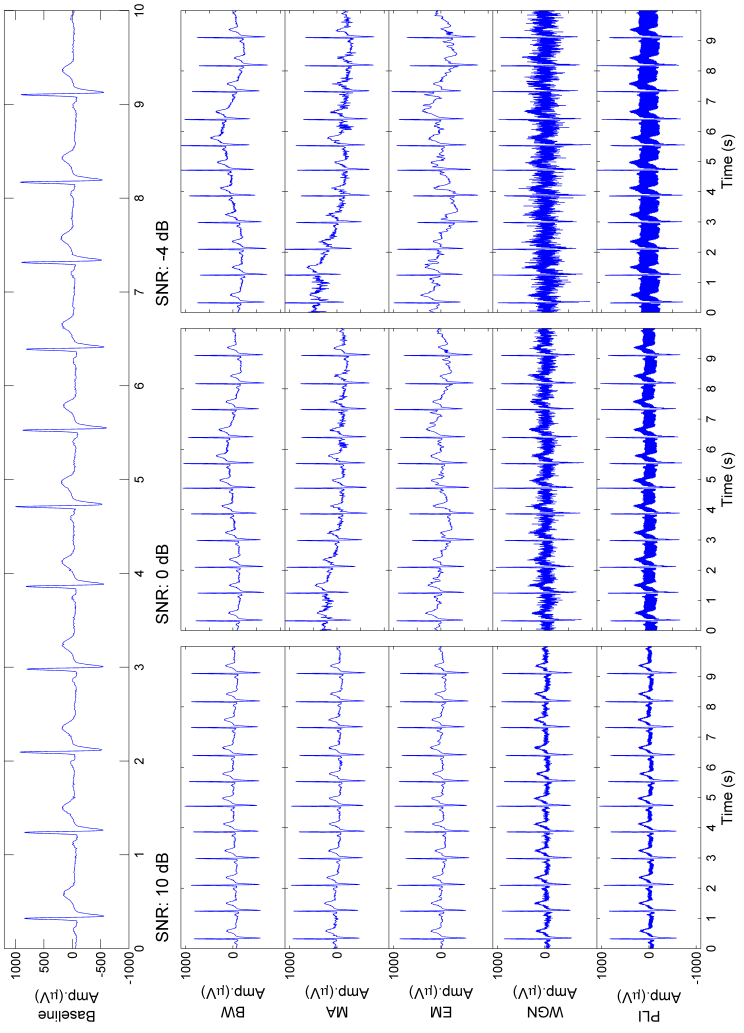


Figure 3.7: The top plot shows lead V5 of a normal baseline ECG with no noise applied. Rows 2-6 corresponds to the 5 types of noise applied; baseline wander, muscle artifacts, electrode motion, white Gaussian noise and power line interference. The columns represent three levels of noise; SNR: 10 dB, SNR: 0 dB and SNR: -4 dB. The corresponding root mean square amplitude ratios are 3.2, 1 and 0.6. Note that the stated SNR's are not uniquely calculated from lead V5. See section 3.4.2 for details.

3.5 Filtering ECGs to Remove Noise

Section 3.4.1 presented five examples of ECG noise. White Gaussian noise is, by the nature of its theoretically infinite sampling frequency, not treatable with regards to lowering the SNR. It can be thought of as unexplained variation, measurement errors and the like. It was included in the presentation of ECG noise to visualize the effect of a noise source with an equally distributed spectrum on the ECG. It is highly desirable that the differences between the two study populations is founded in a physiological process related to the heart and not in artifacts of the measurement process, biological or otherwise. In order to plot the amplitude spectrum of the noise signals, the noise application procedure described in section 3.4.2 was followed; a random starting point in the noise recordings is chosen and a noise signal of the same length as the ECG is sampled. If the full length (non sampled) root mean square (rms) values of the 3 biological noise sources are added, EM, MA and BW correspond to 49%, 18% and 33% of the total rms, respectively. As BW was hard to isolate from the remaining during noise recording ([25]) it is expected that EM and MA overlaps BW in the low frequency range (below 1 Hz). Figure 3.8 presents the amplitude spectra of lead V5 of an ECG and the three biological noise sources. To ease the comparison

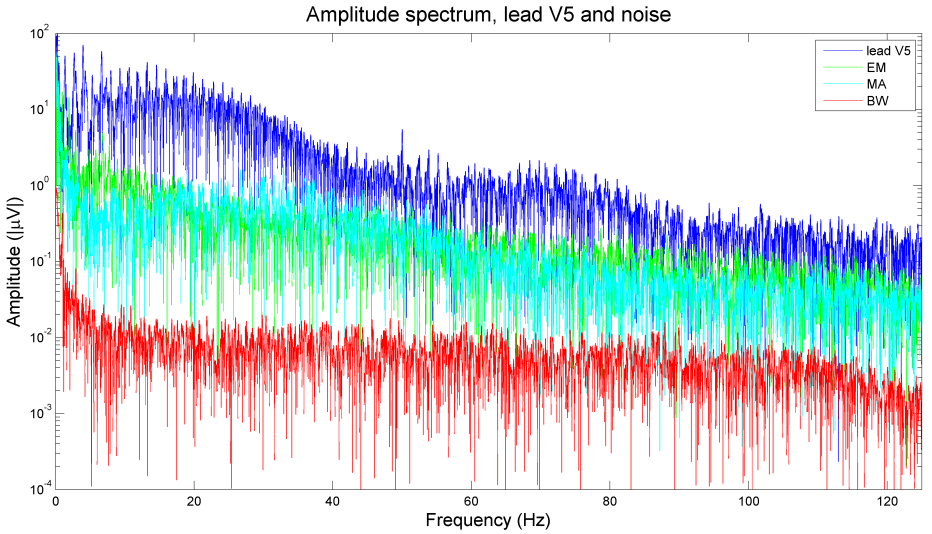


Figure 3.8: Amplitude spectrum lead of V5 of ECG (blue), electrode motion noise (green), muscle artifact noise (cyan) and baseline wander noise (red). The original noise amplitude is adjusted such that SNR is 14 dB for the three types.

the amplitude of the original noise signals was adjusted such that the SNR was

14 dB (rms amplitude ratio close to 5) for all three. The fundamental frequency of the QRS complex is around 10 Hz while it is 1-2 Hz for the T-wave. Also, most diagnostic information is contained below 100 Hz in adults [50]. Higher frequency components could be notches within the QRS complex or the T-wave which, for the latter, is observable in LQT2 subjects. The frequencies depend on the heart rate, which sets a lower bound for the frequency content [50]. Bradycardic subjects (<40 beats per minute) corresponds to a lower bound of 0.667 Hz and are uncommon in the clinic. Further, the study population does not include any subjects with a heart beat in that region. Since the study population ECGs are sampled at 250 Hz the highest frequency content in the sampled ECGs are 125 Hz.

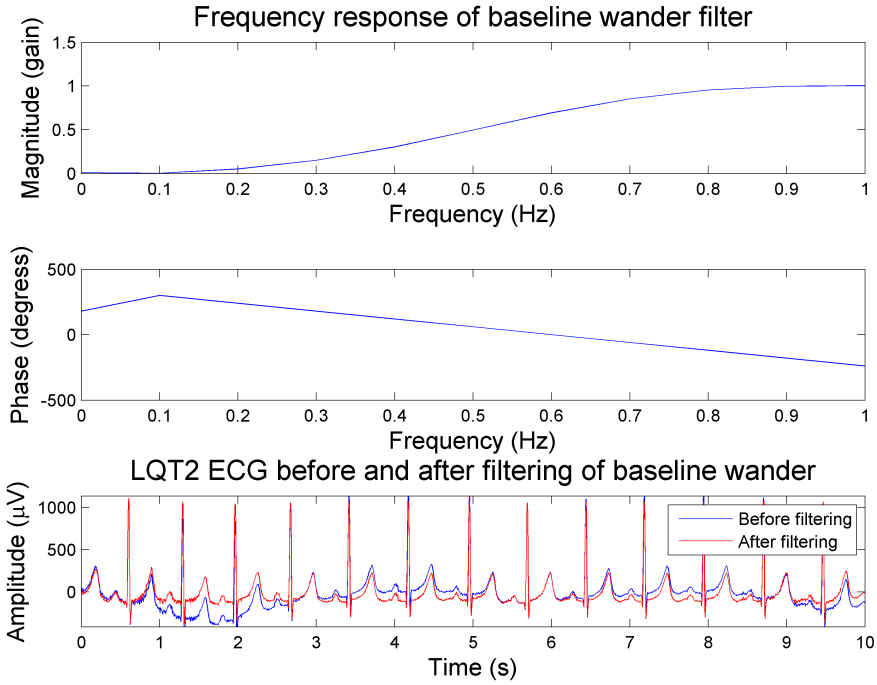


Figure 3.9: Frequency response of filter and example of signal filtering. Top panel shows the gain in the frequency range 0-1 Hz, middle panel shows the phase in the frequency range 0-1 Hz and the bottom panel shows an ECG from study population before and after filtering.

Figure 3.8 shows that at an SNR of 14 dB both EM and MA have high frequency components (100-125 Hz), with an amplitude in the range of the ECG. However, in order to preserve information and prevent introducing differences in the study population by inappropriate filtering, focus is maintained on the low frequency range. The main components of BW is typically said to be found below 0.5 Hz and BW can be greatly reduced by high pass filtering. The cutoff frequency has been the subject of some concern as a cutoff of 0.667 Hz can result in distortion of repolarization and ST-segment changes. However, bidirectional digital filters eliminate phase shift and so high pass filtering of this kind, with a cutoff frequency of up two 0.667 Hz, is in compliance with AHA recommendations, *Recommendations for the Standardization and Interpretation of the Electrocardiogram. Part I: ...* [50]. Hence it was chosen to apply a high pass filter to the data to remove baseline wander and other noise sources having spectral components in this region. A bidirectional digital high pass Kaiser Window FIR filter with a cutoff frequency of 0.5 Hz was implemented. Figure 3.9 presents the frequency response in terms of gain (top panel) and phase (middle panel), within the frequency range 0-1 Hz. Furthermore, an ECG from the study population is shown before and after filtering. The example ECG was chosen by visual inspection and shows the beneficial effect of removing baseline wander.

Chapter 4

Previous Work

This chapter presents selected works within the field of ECG characterization and discrimination. A reflection on the methods, relevant to the current thesis, are provided at the end of this Chapter.
The literature indicates that a large amount of work has been performed in the field of ECG segmentation, i.e. wave labeling and the like. A relevant example could be that of identifying abnormal beats in a 24 hour Holter ECG recording, which is a very time consuming task. Computerizing the process, ECG beats, as defined by their segmentation, can be identified and characterized automatically. The features extracted from the ECG and the methods applied are numerous. Experience shows that hidden Markov models in various forms have been applied extensively in ECG segmentation and discrimination in different contexts. The selection of works presented below is chosen as representative of the methods that are typically encountered in the field, but a strong emphasis is put on the application of hidden Markov models (HMMs).

4.1 Signal Recognition and ECG Modeling

Title: *The Application of Pattern Recognition Technology in the Diagnosis and Analysis on the Heart Disease: Current Status and Future* (**2012**)

In this review article Jin et. al [22] motivates the increasing importance of automatic detection and analysis methods applied in ECG pattern recognition in the diagnosis of cardiovascular disease. Pattern recognition methods are usually either statistical or structural. The process is broken down in to 1) feature extraction and 2) classification and prominent methods within both are described. It is stated that detection and location of ECG waveforms form the basis of an automatic ECG diagnosis system. Current basic methods of feature extraction include the adaptive threshold method, the syntax method, the wavelet analysis method, morphology operation, hidden Markov models, linear prediction and correlation method among others. Methods are briefly presented and their advantages and disadvantages are described. With regards to classification it is stated that classification of QRS waves are mainly dependent on the effectiveness of the feature extraction besides, of course, the classification method. Pattern recognition oriented classification mainly applies linear classification, Bayes classification, K-adjacent rules, support vector machine classification, clustering methodology and neural network methods, among others. Methods of combining classifiers are also presented.
Perspective: It is stated that ECG denoising and specific ECG pattern recognition (P wave, QRS wave etc.) have currently been performed with good results whereas automatic ECG classification has not shown satisfactory results. The need of a global ECG classifier is motivated and it is pointed out that existing automated analysis systems are based on short term observational data.

Title: *Machine Learning in Electrocardiogram Diagnosis* (**2009**)

Salem et. al [41] provides a review of machine learning applications in ECG classification. The classification process is split in feature extraction and classification and comparative tables of classification accuracies within each machine learning scheme are presented. **Support Vector Machine Methods:** feature extraction covers symptoms, PCA, direct cosine and wavelet transform and the raw 8 lead ECG, amongst others. Classification accuracy ranges from 88% using the raw ECG as feature to a 100% using symptoms obtained from patients. **Fuzzy Methods:** direct wavelet transform and other ECG parameters as feature. Classification accuracy ranges from 98.1% to 100%. **Artificial Neural Network Methods:** Feature extraction methods are the discrete wavelet transform, eigenvector methods, rate of heartbeat and waveform characteristics such as amplitudes and duration, amongst others. Various types of ANN methods are presented and the classification accuracy ranges from 79% to 100%. **Rough Set Theory:** various non-time series features are extracted and applied in ECG classification resulting in an accuracy of 87% to 93%. **Hidden Markov**

Models: as feature extraction method the wavelet analysis method is applied resulting in a classification with 70% sensitivity. In terms of classification Salem et. al describe hybrid methods where multiple classifiers are fused. Results show 80% to 99.9% accuracy.
Perspective: Generally classification accuracies fell within the range of 70% to 100% but it must be noted that the referenced works were performed with different data and with different objectives for classification. Salem et. al further mentiones the importance of considering sensitivity and specificity issues given the possible diagnostic context.

Title: *Kernel based Hidden Markov Model with applications to EEG signal classification* (**2005**)

In this study [66] Xu et. al introduces a kernel based HMM, where they combine the hidden Markov and support vector machine (SVM) framework (KHMM) to be applied in signal classification. They state that the hidden Markov model is an elegant statistical model particularly suitable for modeling temporal signals such as speech and biosignals, giving a good generalized representation. The support vector machine is also a discriminative model capable of maximizing the margin between classes, and thus considering the error rate during classification. By combining them they explore the temporal dynamics of the signals while maximizing the margins between classes, thus taking the misclassification margin into account while training the HMM.
Perspective: The model performance was evaluated using features from 100 training examples of 28-channel EEG signals using 20 fold cross validation and comparison to SVM and HMM. The accuracy obtained was 78% for SVM, 84% for HMM and 88% for KHMM.

Title: *Support Vector Machine for Assistant Clinical Diagnosis of Cardiac Disease* (**2009**)

Wei et. al [24] evaluated SVM methods in classification of normal and abnormal (not specified) ECGs in cardiovascular disease. Then it is stated that automatic ECG classification relies on an initial effective ECG segmentation followed by analysis and classification of the extracted waves. The input data in this work are already-segmented full beats. Wei argues that these differ in length and so they are transformed in some non specified manner (presumably some form of time-warping). Channels of ECG data are applied in the classification either in series or in parallel and in both cases no form feature extraction (besides

segmentation) is performed prior to the classification; the raw ECG is used as input. Various types of radial basis functions are applied in the SVM classification and classification precision varies from 58%-89.25% when ECG data input is applied in series to 87%-91.25% in the parallel input case.
Perspective: Raw ECG beats are applied with good result, particularly in the parallel case. Classification is not based on derived ECG features, however, as the beats are extracted and transformed in to the same length some feature extraction has effectively been performed. Results also show that the choice of kernel parameters greatly influences the classification accuracy.

Title: *Cardiac arrhythmia classification using wavelets and hidden Markov models–a comparative approach* (**2009**)

The study [26] applies a HMM to model features derived from linear segmentation or wavelets of ECG beats in order to classify beats involved in cardiac arrhythmia. The study uses a left-to-right HMM with six states and five Gaussian components per state to model the features.
Perspective: It concludes that features from the wavelet transform outperform linear segmentation in beat classification.

4.1.1 HMM Methods Applied in ECG Recognition

Title: *Myocardial infarction classification with multi-lead ECG using hidden Markov models and Gaussian mixture models* (**2012**)

The general scope of this study by Chang et al. [51] is the separation of normal ECGs from ECGs containing changes related to myocardial infarction. In summary, the methods cover segmentation of the ECG using hidden Markov models, evaluating the likelihood of extracted segments with the HMM and finally classifying on the basis of the HMM features using both GMM and SVM. Chang et al. stress the need for an automatic classification system and state that previous work is mostly comprised of pattern recognition (segmentation), noise removal and ischemia detection. In this context HMM has mostly been applied in delineation, segmentation or component detection (seemingly covering the same concept, namely that of defining segments of the ECG as corresponding ECG

waves). Further, it is stated that the study is the first to identify each (presumably full) beat by its waveform and apply it in classifying myocardial infarction. The study applies HMM in both segmentation and log-likelihood calculation. Whole beats are extracted and sample sizes are fixed because time-warping is not applied. Four relevant leads of each ECG are evaluated with regards to log-likelihood applying the HMM models. A left-to-right transition matrix assumes time-series input, which is beneficial. Full transition is also evaluated to test the time-series assumption (full type seems to capture most of the left-right properties). Applying 6 and 16 state transition matrices, a different number of components in the GMM and an RBF kernel in the SVM, the classification accuracies were; 71%-83% for GMM and 71%-75% for SVM.
Perspective: HMM were used in both segmentation and log-likelihood evaluation of whole beats as feature for classification. However, the extracted beat sample lengths were truncated because otherwise the probability value would be "unfair" with regards to classification. Illustrations of the matter are vague and presumably new tachycardic subjects would pose a problem. Best results were seen with the 16 state HMM and GMM outperformed SVM. With SVM the key issue was found to be the selection of kernel function.

Title: *Modelling ECG Signals With Hidden Markov Models* (**1996**)

In this study Koski [37] uses a continuous probability hidden Markov model to model segmented ECG signals. The ECG signals are approximated with broken lines, providing two features; the duration of the line segment and the amplitude of the line's starting point. Subsequently features are modeled using a hidden Markov model. To validate the trained model ECG simulations are performed. Koski found that a small model using 15 states was not able to capture the dynamics of the ECG, since it wrongly mixes the QRS complexes with the T-waves. A 25 state model was found to be sufficient in modeling an entire heart beat cycle. However, he argues that an increased number of states might be required to model different ECG variations while simultaneously constraining the number of states due to the potential of overfitting the training data and the loss of generalization capability. To investigate the classification property of the HMM, Koski used a 30 state HMM to model four normal ECG signals and four ECG signals containing premature ventricular (PV) beats. Subsequently, the models were tested using two normal ECG signals and two containing PV beats. Using the maximum probability of the signals given the models, all test signals were correctly classified.
Perspective: The study concludes that HMM is a very suitable method for modeling ECG signals and further it can be used to classify new unseen ECG signals. Koski states that the strength of the HMM is that it can be used with-

out expert knowledge, can model the signal directly, and it produces probability values instead of simply yes/no decisions. The disadvantages are that the HMM must be analyzed in order to be trusted. However, a simulated ECG generated from the model is an excellent way to visually inspect the result of the learning.

Title: *Heart Signal Recognition by Hidden Markov Models: The ECG Case* (**1994**)

This work [39] covers to ECG segmentation applying a specialized form of the continuous variable duration hidden Markov model (CVDHMM). In a segmentation context (i.e. labeling P, QRS and T-waves) Thoraval explains the application of HMMs in ECG segmentation and points out some weaknesses of the HMM approach; in a segmentation context the wave is associated with a state who's emission density is considered to be stationary with time and forms the basis of the segmentation. It is further stated that the non-stationary properties of ECG waves degrade the robustness of a segmentation model based solely on the stationary statistical properties of the ECG waves, though marginal stationarity is observed in the ECGs. Furthermore, the stationary assumption might eliminate important shape descriptors characterizing the ECG waves. To overcome this issue a modification of the CVDHMM is proposed; one state is partitioned in to two subsets where one subset models the wave and the other an "interwave" corresponding to intermediate observations. Intermediate observations need not be present, and so the one-to-one registration of ECG samples and observations is not necessary, effectively decoupling the simultaneous segmentation-identification process as in the normal CVDHMM. Preprocessing amounts to a non-linear transform and wavelet analysis producing the required features.
Perspective: Without quantifying the applicability further than presenting two examples of segmentation of noisy ECGs it is implied that the lacks of the normal CVDHMM were confirmed during simultaneous segmentation applying the new and regular method, respectively.

Title: *ECG Signal Analysis Through Hidden Markov Models* (**2006**)

In this work Andreão et. al [5] applies hidden Markov models in both ECG segmentation and classification of premature ventricular beats and ventricular beats. The relevance of automated ECG analysis is stressed and it is pointed out that the ECG segmentation prior to the actual classification is crucial for accurate results. Also, most works apply heuristic rules in the segmentation

process. A large number of classification methods exist but the advantages of HMMs are pointed out; these include that a waveform sequence can be modeled, intra-individual variability can be incorporated in to the model state transitions, and that the HMMs can be applied in both beat detection, segmentation and classification. The approach is a two-step process where the ECG data is first segmented using the HMM and then premature ventricular beats and ventricular beats are classified using a heuristic and a statistical approach, respectively. The heuristic approach applies segmentation results whereas the statistical approach applies the likelihood of the QRS complex as given by the HMM model (which is essentially the first step). The method covers both single and double channel ECG data and a continuous wavelet transform that is performed prior to the segmentation. The HMM model is comprised of several sub models for each waveform such that it is effectively waveform modeling and not beat modeling. This elementary waveform model consists of 4 HMMs for the QRS complex, 2 HMMs for the P-wave, PQ-segment, ST-segment and T-wave, respectively, and one HMM for the baseline. A single Gaussian is applied and summing the HMM states for one set of waveform sub models, 19 states are applied (i.e. plus remaining sub models). In the segmentation a generic model is adapted to each individual. Considering the non-heuristic approach the QRS complexes are labeled as abnormal (ventricular beats) by considering the dominant QRS sub HMM in each individual. The labeling is performed by comparing with the remaining part of the individuals' QRS complexes while holding the log-likelihood against an adaptive threshold, meaning that it is therefore unsupervised.
Perspective: Hidden Markov models are suitable for ECG modeling, beat detection, segmentation and classification. Classification of ventricular beats, based on the QRS log-likelihood is performed with 99.79% sensitivity. Premature ventricular beat detection is performed with 87% sensitivity.

Reflections on Methods: As mentioned in the introduction the motivation of this work was the possibility of characterizing ECGs without the use of stationary features extracted from MUSE®. It seems, however, that most works adopt this approach in that ECGs are most often segmented before any form of discrimination of the waveform or ECG types is performed. HMMs in automated ECG analysis are often applied in the segmentation process by using the hidden state sequence. However, the HMM approach also provides log-likelihood which can be used to discriminate the ECGs. Chang et al. [51] claims to be the first to both identify and classify full beats using the HMMs. Koski [37] states that the strength of the HMM is that it can be used without expert knowledge, it can model the signal directly, and it produces probability values instead of simply yes/no decisions. The disadvantages is that the HMM must be analyzed in order to trust them, but a simulated ECG generated from the model is an excellent

way to visually inspect the result of the learning. Thoraval [39] observes that a weakness of the HMM approach is, in a segmentation context, that a given wave is associated with a state who's emission density is considered to be stationary with time which forms the basis of the segmentation. It is further stated that stationarity of ECGs (e.g. wave mean) is not an appropriate assumption, although marginal stationarity is observed in ECGs. In a classification context however, using a method that forces stationarity in some ways, could be beneficial because the aim is to capture general trends in each group. Thus, the HMMs should provide a good generalized representation of the ECGs.
Besides the actual classification, emphasis in this work is also put on characterization of the ECGs. Preferably the applied machine learning methods should also maintain some generative capabilities that could potentially lead to the identification of the general ECG trends captured by the models. Perhaps these observations could even be related to the underlying physiological process. Finally, to improve the classification results while applying HMMs, the literature suggests that SVM poses a good candidate. Also, SVM appears to have been used extensively in the field of ECG characterization and discrimination.

Chapter 5

Machine Learning Methods

In following chapter the different machine learning methods applied in this work are explained. First a brief introduction to machine learning is given. The basic concepts of training models are described in section 5.1 and their validation is described in section 5.2. In section 5.3 the reasoning behind the choice of machine learning models is presented with emphasis on the knowledge acquired in the literature review in section 4.
The Hidden Markov Model and its framework are explained in the next four sections. In section 5.4 the discrete Markov Model is introduced followed by a description of the Gaussian Mixture model in section 5.5. Section 5.6 describes the fusion of the Markov model and the Gaussian Mixture model to form the Hidden Markov Model. Issues regarding the implementation of Hidden Markov models are discussed in section 5.7, addressing problems such as underflow, singularity issues and speed. Finally an explanation of the discriminative model Support Vector Machine is provided.

5.1 Basic Concepts of Machine Learning

Machine learning is a cross field between statistics, data mining and pattern recognition. The basic idea of machine learning is to construct a system that

can adapt or learn from data in order to build a model capable of performing descriptive and/or predictive tasks. A descriptive task is concerned with finding interpretable patterns in data, while predictive tasks strive to predict unknown or future values of variables given some input features. The latter could be the classification of some unknown object based on its features or a regression where values could be predicted based on the learned functional relationship between input and output.
Machine learning models are separated into *supervised* (where the class labels of input data are known) and *unsupervised* (where they are not).
In a supervised classification scheme a model is built using a *training set* containing information/features of the data objects including the class labels. Based on the training data, a learning algorithm is used to construct a model with a good generalization capability, i.e. it can accurately classify new unknown data.

5.2 Evaluating Model Performance

To measure the performance of a given model the training data is often split in a training set and a *test set*. The training set is used to build the model while the test set is utilized as unknown data, to be classified by the model. A measure for how well the model fits the training data is termed the *training error*, whilst the performance on the test set is termed the *test error* or *generalization error*. Models that fit the training data well, but have a high test error, are termed *overfitted*. Various measures are taken to avoid overfitting, based on the model setup, but often the generalization error provides a means of determining the complexity of the best models [61].
A way to obtain a good estimate of the generalization error, is to use k-fold cross-validation. Using this approach data is segmented into k equal size partitions. In each k run one of the partitions is chosen as test set and the remaining are used for training. Each data object is used $k-1$ times for training and one time for testing resulting in k so-called cross-folds. For each cross-fold the accuracy is determined, as

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{5.1}$$

By comparing the accuracy between the different models, one can determine the better model. Having multiple estimates of the accuracy for each model, it can be determined whether the differences are statistically significant.
Two other important model evaluation metrics are the *sensitivity* and *specificity*. The sensitivity or true positive rate is the fraction of positive examples predicted correctly by the model and is in a healthy care setting given as:

$$\text{Sensitivity} = \frac{\text{Number of subjects correctly diagnosed as sick}}{\text{Total number subjects diagnosed as sick}} \tag{5.2}$$

The specificity or true negative rate is the fraction of negative examples predicted correctly by the model and is, in a health care setting, given as:

$$\text{Specificity} = \frac{\text{Number of subjects correctly diagnosed as healthy}}{\text{Total number subjects diagnosed as healthy}} \tag{5.3}$$

Other ways to examine the performance of a model is to calculate the *confusion matrix*. The confusion matrix is set up as presented below.

	Predicted Class 1	Predicted Class 2
Class 1	True Positive	False Negative
Class 2	False Positive	True Negative

Table 5.1: Illustration of a confusion matrix for a binary classification problem.

The confusion matrix simply summarizes the number of objects predicted correctly or incorrectly. This matrix may reveal if, for example, the test objects are more often incorrectly classified as class 2, but never incorrectly classified as class 1.
A graphical approach, showing the trade-off between the true positive rate and the false positive rate, is the receiver operating characteristics curve (ROC). In the ROC the true positive rate (sensitivity) is plotted on the y axis while the x axis shows the false positive rate (specificity). A good classification model should have an ROC curve that is located in the upper left corner. A random classification model would reside along the diagonal.
A way to quantify the ROC curve is to calculate the area under the curve (AUC). This evaluates which model is better on average. An random classification model would have an AUC close to 0.5 while a perfect classification model would have an AUC of 1. The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen sick subject higher than a randomly chosen healthy subject, where a high ranking indicates sickness [23]. It should be noted that it is possible for a classifier with a high AUC to perform worse in specific regions than one with a low AUC.
Different types of ROC curves and their corresponding AUC are presented in Figure 5.1.

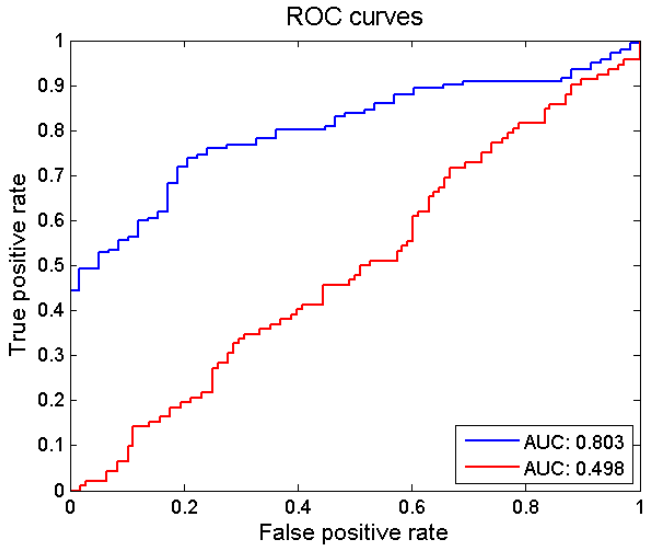


Figure 5.1: Illustration of two ROC curves for two different classification models. The model represented by red curve is a random classifier, while the model represented by the blue curve is superior to that of the red.

5.3 Choice of Machine Learning Models

Given the previous work in the field of ECG modeling and classification the hidden Markov model appears to be a good candidate for the intended purpose, since

- It is particularly suitable for modeling temporal signals such as bio-signals [66] and is used in a range of applications in bioinformatics [18].
- It has been used to model ECG signals and has proven able to capture features of the ECG's [37], [24] as well as ECG derived features [26], [41].
- It has been used for ECG classification providing good results [37], [22], [14], [26], [15].
- It has the capability of providing output to more discriminative machine learning models [51].
- It is a generative model, which means it can be used to simulate what has been learned, thus providing insight into the captured features from training data [37].
- Numerous well described methods of modification exist to improve the classification or modeling capabilities [55], [66], [4], [38], [58], [67] .
- Extensive use of HMMs have seen many ECG applications such as feature extraction and delineation [40].

Many studies have tried to model single ECG beats [37], [24], with a few modeling multiple leads, but often only with individual HMM's [51]. In this thesis an 8-dimensional continuous emission HMM is chosen to model 10 second ECG's, to able to model covariance between leads, to simulate 8-lead ECGs and to capture some relation between beats. To have complete control of the model and since such a model is not available in the MATLAB® statistical toolbox [42] the model is implemented from scratch in MATLAB® version R2012a (7.14.0.739) using both 32-bit and 64-bit versions. As in the work of Chang et al. [51] a Support Vector Machine using the output probabilities from the HMMs as features, is applied in order to investigate possible improvement of the classification. The SVM is chosen since, as will be described in section 5.8, it is able to produce a higher feature space and has good discriminative capabilities. The SVM implementation from the Bioinformatics Toolbox Version 4.1 is used only applying default kernel options, hence black-boxing the choice of kernel.

5.4 Markov Models

Markov models are stochastic processes often used to model sequential data [3]. They are used to create mathematical models for the temporal evolution of given phenomena with probability as an important factor. They have been applied in a very broad range of disciplines spanning financial modeling [28], operational research [2], environmental predictions [60], biomedical signal analysis [19] and especially in speech recognition [36], [68], [56]. A simple Markov model is also called an observable Markov model, since the output of the process is observable events equal to the states of the model [55].
In a first-order Markov chain model the state at time $t+1$ only depends on the previous state at time t. This memoryless property of the process is also referred to as the Markov property. A simple first-order Markov chain is illustrated in Figure 5.2.

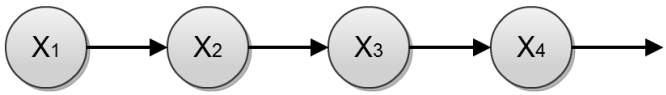


Figure 5.2: Illustration of a first-order Markov chain in which the current state only depends on the previous states. The observable values x_n are equal to states of a simple Markov model.

A Markov model can be formalized with the following elements:

1. N: Number of states in the model
   $\mathrm{S} = \{S_1, S_2, ..., S_N\}$
2. State transition probabilities:
   $\mathbf{A} = [a_{ij}]$ where $a_{ij} \equiv P(q_{t+1} = S_j | q_t = S_j)$
3. Initial state probabilities:
   $\Pi = [\pi_i]$ where $\pi_i \equiv P(q_t = S_i)$

where

- t is the discrete time at a given state:
  $\mathrm{t} = \{1, 2, ..., T\}$
- Q is the state sequence:
  $Q = \{q_1, q_2, ..., q_T\}$

With a_{ij} being probabilities, it must satisfy

$$a_{ij} \geq 0 \text{ and } \sum_{j=1}^{N} a_{ij} = 1 \tag{5.4}$$

which also applies for π_i

$$\pi_i \geq 0 \text{ and } \sum_{i=1}^{N} \pi_i = 1 \tag{5.5}$$

An example of a Markov model with two states is shown in Figure 5.3.

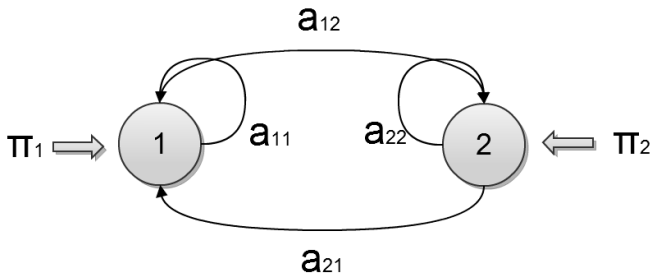


Figure 5.3: Graphical representation of a simple Markov model with two states. Π_i is the probability that the system starts in state S_i while a_{ij} is the probability of a transition from state S_i to S_j.

Weather predictions are an often used as an intuitive example of a sequence, where a simple Markov could be used as a model. For simplification assume that the weather can be either sunny (☼) or rainy (|||) on a specific day. The states of the Markov model are directly observable and the states will therefore directly describe the weather. An example of some parameter values for a Markov model describing this scenario could be:

$$\mathbf{A} = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$$
$$\Pi = \begin{bmatrix} 0.2 & 0.8 \end{bmatrix}^T$$

The top diagonal element represents self-transitions to the sunny state whereas the lower diagonal element corresponds to the rainy state. Hence, the probability that the next day should be sunny given that the current day is sunny, is

0.6. On the other hand, the probability of rain next day is 0.4. The probability of a five day weather prediction, e.g. {Sunny,Sunny,Rain,Rain,Sunny} would be:

$$\pi_{☼} a_{☼,☼} a_{☼,☼} a_{☼,☁} a_{☁,☁} a_{☁,☼} \Rightarrow$$
$$0.2 \cdot 0.6 \cdot 0.4 \cdot 0.7 \cdot 0.3 = 1.01\%$$

In order create an applicable model, the parameters must be estimated from already observed sequences. Given K sequences of length T, the parameter Π and **A** is trained using the following estimation formulas [3]:

$$\hat{\pi}_i = \frac{\#\text{ of sequences starting with } S_i}{\#\text{ of sequences}} = \frac{\sum_k 1(q_1^k = S_i)}{K}$$

with 1(x) being 1 if x is true or 0 if x is false.

$$\hat{a}_{ij} = \frac{\#\text{ of transitions from } S_i \text{ to } S_j}{\#\text{of transitions from} S_i} = \frac{\sum_k \sum_{t=1}^{T-1} 1(q_t^k = S_i, q_{t+1}^k = S_j)}{\sum_k \sum_{t=1}^{T-1} 1(q_t^k = S_i)}$$

5.5 Gaussian Mixture Models

The Gaussian mixture model (GMM) is not as much a model as it is a probability distribution consisting of multiple Gaussian distributions. Each Gaussian can be viewed as being a hidden or latent process from which visible data values can be drawn. The purpose of the GMM is to estimate an optimal (local or global) set of parameters of a predefined number of Gaussians in order to maximize the conditional probability as if the specific data were being generated from the GMM. A number of techniques for maximizing the likelihood exists [8], and one of them is the elegant expectation-maximization (EM) algorithm.
The probability of a data point $\boldsymbol{x}$ given a GMM, is defined as

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} W_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5.6}$$

with K being the predefined number of Gaussians, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ being the mean values and covariance structure of the k'th Gaussian component and W_k the weighting or mixing coefficient. These parameters must satisfy

$$0 \leq W_k \leq 1 \tag{5.7}$$

and

$$\sum_{k=1}^{K} W_k = 1 \tag{5.8}$$

An important quantity to be estimated is the conditional probability of a specific Gaussian component given a data point. This can evaluated by introducing γ_{kx} as the responsibility that the k'th component takes in explaining the observation x [8] and it can be calculated by

$$\gamma_{kx} = \frac{W_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} W_j N(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{5.9}$$

In order to model data in the best possible way from a probabilistic standpoint, the following log-likelihood function should be maximized

$$\ln(p(\boldsymbol{x}|W, \boldsymbol{\mu}, \boldsymbol{\Sigma})) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} W_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \tag{5.10}$$

with N being the number of data points. The maximization can be performed by applying the EM algorithm.

5.5.1 EM algorithm in GMM

The expectation-maximization or EM algorithm is a general framework for identifying maximum likelihood solutions for models having hidden or latent variables. The EM algorithm is an iterative process that optimizes the log likelihood until either a convergence or maximum iteration criterion is satisfied [8]. The steps of an EM algorithm in a GMM framework are presented below and a visual illustration in 2D is presented in Figure. 5.4:

1. Initialize $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and weights W_k and calculate the log-likelihood. In Figure 5.4 (a) the green data points will be modeled using two Gaussians. The two Gaussians (a red and a blue) are randomly initialized with different means, equal variance and with zero covariance.

2. Expectation step: Evaluate γ_{kx} using equation 5.9. This step is illustrated in Figure 5.4 (b). Given the means and covariance structure of the two Gaussians each data point is given a color in the spectrum between red and blue, illustrating the probability of being generated from a particular Gaussian. Hence, a data point having a high probability of being generated from the blue Gaussian is colored blue, while a data point is red if it has a high probability of being generated from the red Gaussian. Data points having almost the same probability of being generated from either of the Gaussians are colored purple.

3. Maximization step: The parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and W_k are re-estimated using the responsibilities from the expectation step, by setting the derivative of

equation 5.10 to zero with respect to each of the parameters. The re-estimation formulas are:

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^{N} \gamma_{kx} \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_{kx}} \tag{5.11}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{\sum_{n=1}^{N} \gamma_{kx} (\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})^T}{\sum_{n=1}^{N} \gamma_{kx}} \tag{5.12}$$

$$W_k^{new} = \frac{\sum_{n=1}^{N} \gamma_{kx}}{N} \tag{5.13}$$

In Figure 5.4 (c) the means and covariance structure of the blue and red Gaussians are updated based on the given color/responsibility of the data points.

4. Evaluate the log-likelihood.

5. Terminate if convergence criteria are met, otherwise repeat steps 2-5. In Figure 5.4 (d)-(f) the Gaussians repeat the EM algorithm until they satisfactorily model the two data clusters. At this point the log-likelihood will have converged indicating a maximum in the log-likelihood function.

If performed correctly the log-likelihood of the EM algorithm will always increase or stay the same if convergence is achieved. However, the log-likelihood function will in general have multiple local maxima and so the identification of a global optimal solution cannot be expected. The problem regarding presence of singularities is discussed in section 5.7.

A frequently used method of determining a suitable initialization of the means and covariances is the K-means algorithm [8]. Using the K-means algorithm could help finding consistently good solutions [16]. The K-means algorithm is an unsupervised cluster method that resembles the GMM. It aims to construct k-clusters for each observation belonging to the nearest mean and requires a smaller number of computations than the GMM.

5.6 Hidden Markov Models

Having introduced the Markov chain and the Gaussian mixture model, the extension to the Hidden Markov model is made easier.
In the Hidden Markov Model (HMM) the states are not directly observable [3].

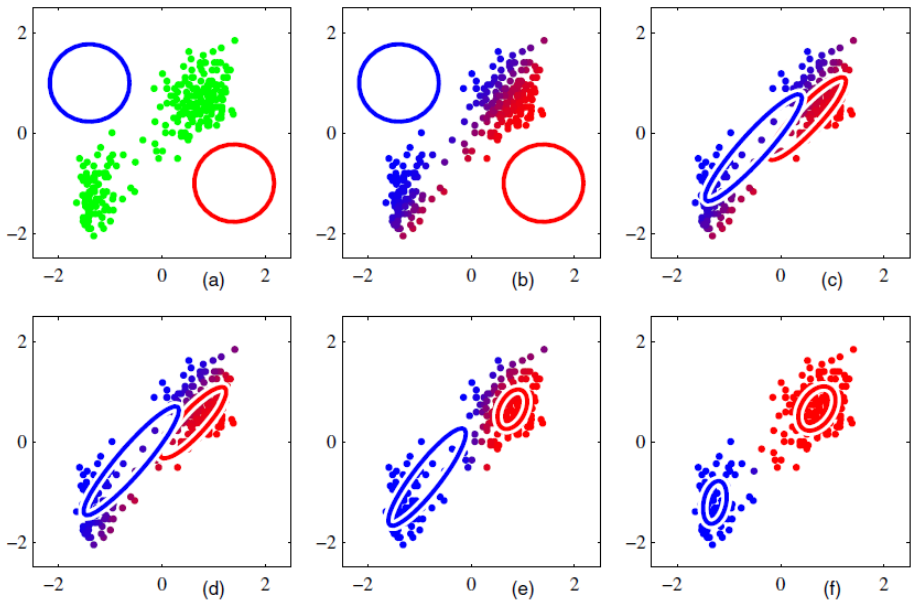


Figure 5.4: A illustration of the EM algorithm in a Gaussian mixture model framework. In (a) two Gaussians are randomly initialized with zero covariance. In (b) the responsibilities for each data point are calculated (expectation step). In (c) the parameters are updated (maximization step). In (d) through (f) the EM algorithm is repeated until convergence is achieved. Modified from [8].

In a given signal the observable values are thought of as being emissions from these latent states.
The discrete HMM can be formalized with the following elements [55]:

1. N: Number of states in the model
   $\mathrm{S} = \{S_1, S_2, ..., S_N\}$
2. M: Number of distinct emissions in the signal
   $\mathrm{V} = \{v_1, v_2, ..., v_M\}$
3. State transition probabilities:
   $\mathbf{A} = [a_{ij}]$ where $a_{ij} \equiv \mathrm{P}(q_{t+1} = S_j | q_t = S_i)$
4. Emission probabilities:
   $\mathbf{B} = [b_{jm}]$ where $b_{jm} \equiv \mathrm{P}(O_t = v_m | q_t = S_j)$
5. Initial state probabilities:
   $\Pi = [\pi_i]$ where $\pi_i \equiv \mathrm{P}(q_t = S_i)$

where

- t is the discrete time at a given emission:
  $\mathrm{t} = \{1, 2, ..., T\}$
- O is a given d-dimensional emission sequence:
  $O = \{O_1, O_2, ..., O_T\}$
- Q is the state sequence:
  $Q = \{q_1, q_2, ..., q_T\}$

The HMM model can be defined by the parameters $\lambda = (\mathbf{A},\mathbf{B},\Pi)$, since N and M are implicitly defined by $\mathbf{A}$ and $\mathbf{B}$.
The discrete HMM can easily be extended to a continuous observation HMM by modeling the emission probabilities as probability densities. A popular choice is to model the emissions as Gaussian probability densities specified only by a mean and a covariance parameter [55]. Applying a mixture of Gaussians to model the emission probability for each state S_j, the probability of a multi-dimensional observation O at time t given a state S_j, is calculated using the multivariate Gaussian distribution and weight W_k:

$$\boldsymbol{b}_{jO_t} = \sum_{k=1}^{K} W_k \mathcal{N}(\boldsymbol{O}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5.14}$$

with K being the number of Gaussian components for the emission probability of each state S_j, $1 \leq j \leq N$. The multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{O}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is defined as:

$$\mathcal{N}(\mathbf{O}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{2\pi^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{O}_t - \boldsymbol{\mu}_{jk})^T \boldsymbol{\Sigma}^{-1}(\mathbf{O}_t - \boldsymbol{\mu}_{jk})\right) \quad (5.15)$$

with n being the dimension of the input sample $\boldsymbol{x}$.
An illustration of a Hidden Markov model using a Markov chain representing hidden variables and an emission model with single a Gaussian is presented in Figure 5.5.

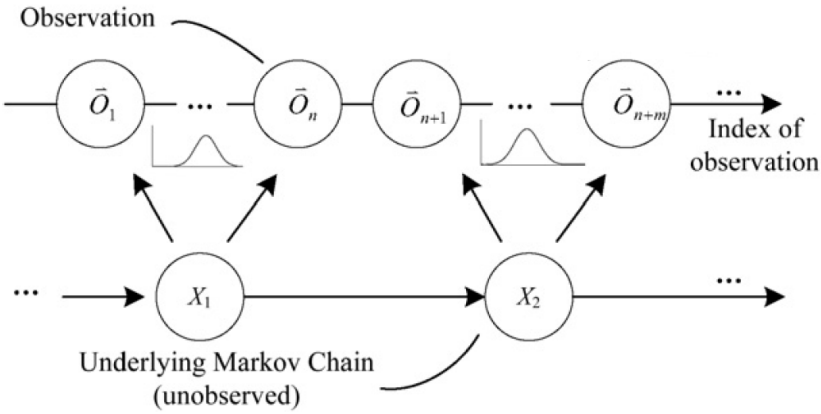


Figure 5.5: Structure of a Hidden Markov model using a Markov chain representing hidden variables and an emissions model with a single Gaussian. Modified from [65].

The three basic problems of HMMs are:

1. To calculate the probability of an observation sequence given the model λ.
2. To find the state sequence, that has the highest probability, given an observation sequence O and an HMM λ.
3. Given a number of observation sequences, train an HMM such that it maximizes the probability of generating the sequences.

The solution to these three questions will be answered in the following sections.

5.6.1 Probability of an Observation Sequence

In order to calculate the probability of an observation sequence given an HMM model λ the joint probability of the observation sequence O and state sequence Q must be calculated followed by marginalization over the joint probability, by summing all possible Q [3].

$$P(O|\lambda) = \sum_{\text{All possible Q}} P(O, Q|\lambda) \tag{5.16}$$

with the joint probability given by:

$$P(O, Q|\lambda) = P(q_1) \prod_{t=2}^{T} P(q_t|q_{t-1}) \prod_{t=1}^{T} P(O_t|q_t) \tag{5.17}$$

$$= \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) ... a_{q_{T-1} q_T} b_{q_T}(O_T) \tag{5.18}$$

where $B = [b_{q_x}(O_t)]$ is the observation probabilities in state x at time t.
Eq. 5.18 states that the joint probability of an observation sequence and the corresponding hidden states are calculated as the probability of a transition from one hidden state at time $t-1$ (q_1) to another hidden state at time t (q_2), multiplied by the probability of emitting the observation O at time t, for all T where the initial hidden state is assumed to be known.
However, this calculation would require $2TN^T$ operations, which for a model with 4 states and a length of an observation sequence of 500 would amount to $2 \cdot 500 \cdot 4^{500} = 1.0715 \cdot 10^{304}$ operations. This would take today's computers much longer than the estimated age of the earth to calculate [53]. To solve this problem the forward-backward procedure is applied.

5.6.1.1 Forward-Backward Algorithm

The general idea behind the forward-backward algorithm is to divide the observation sequence in two parts, having a *forward* variable explaining the sequence from observation 1 to t and a *backward* variable explaining from $t+1$ to end of the sequence. It is a technique based on dynamic programming, which is a way to break a complex problems into sub problems, solving these and storing the results, such that it does not need to be recomputed. The forward-backward algorithm is used to calculate the responsibility variable $\gamma_t(i)$, which is the probability of being in state S_i at time t given an observation sequence O and a model λ. This variable is equal the responsibility variable introduced in section 5.5, which is used in the expectation step of the EM algorithm.
The forward variable $\alpha_t(i)$ is defined as the probability of observing the partial sequence $\{O_1 ... O_t\}$ and being in state S_i at time t given model λ, with

$1 \leq t \leq T$:

$$\alpha_t(i) = P(O_1, O_2, ..., O_t, q_t = S_i|\lambda) \tag{5.19}$$

It can be solved recursively.
Initialization:

$$\alpha_1(i) = P(O_1, q_1 = S_i|\lambda) = \pi_i b_i(O_1) \tag{5.20}$$

with $1 \leq i \leq N$. It is the probability of starting in state S_i multiplied with the probability that the given state has emitted the first observation.
Recursion:

$$\alpha_{t+1}(j) = P(O_1..O_{t+1}, q_{t+1} = S_j|\lambda) \tag{5.21}$$

$$= \Big[\sum_{i=1}^{N} \alpha_t(i)a_{ij}\Big] b_j(O_{t+1}) \tag{5.22}$$

for $1 \leq t \leq T-1$ and $1 \leq j \leq N$.
The forward variable $\alpha_t(i)$ represents the probability of ending in state S_i at time t. This is multiplied with the probability of a transition from all possible previous states to state S_j $(a_{1j}, a_{2j}, ..., a_{Nj})$ at time $t+1$ and summed. Finally, we multiply with the probability of the given observation O_{t+1} in state S_j at time $t+1$, which yields $\alpha_{t+1}(j)$.
By summing over the terminal forward variable the probability of an observation sequence given a model λ is found:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{5.23}$$

Solving the probability of an observation sequence using the forward variable only uses N^2T computations, which is only $4^2 \cdot 500 = 8000$ computations for 4 states and an observation sequence of 500 [3].

To be able to calculate $\gamma_t(i)$, which is applied in the last two problems, the backward variable $\beta_t(i)$ is introduced. It is defined as:

$$\beta_t(i) = P(O_{t+1}, ..., O_T|q_t = S_i, \lambda) \tag{5.24}$$

where $\beta_t(i)$ is the probability of being in state S_i at time t and observing the partial sequence $\{O_{t+1}, ..., O_T\}$.
$\beta_t(i)$ is recursively calculated with the time going in a backward direction:
Initialization:

$$\beta_T(i) = 1 \tag{5.25}$$

Recursion:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{5.26}$$

for $t = T-1, T-2, ..., 1$ and $1 \leq j \leq N$.
Being in state S_i there are N possible state transitions to state S_j with the probability of a_{ij}. When in state S_i the $t+1$ observation is generated and $\beta_{t+1}(j)$ explains all observations after $t+1$ [3].
The probability of being in state S_i at time t can be calculated using the forward and backward variables. This is defined as $\gamma_t(i)$, which is derived in the following:

$$\gamma_t(i) = P(q_t = S_i|O, \lambda) \tag{5.27}$$

Using Bayes second Theorem, which states

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{n=1}^{N} P(A|B_n)P(B_n)}$$

we have

$$\gamma_t(i) = \frac{P(O|q_t = S_i, \lambda)P(q_t = S_i|\lambda)}{\sum_{j=1}^{N} P(O|q_t = S_j, \lambda)P(q_t = S_j|\lambda)} \tag{5.28}$$

By applying the following conditional independence property [8]

$$P(B, A_n) = P(b_1, ..., b_n|A_n)P(b_{n+1}, ..., b_N|A_N)$$

we have

$$\gamma_t(i) = \frac{P(O_1, .., O_t|q_t = S_i, \lambda)P(O_{t+1,...,O_T}|q_t = S_i, \lambda)P(q_t = S_i|\lambda)}{\sum_{j=1}^{N} P(O_1, .., O_t|q_t = S_j, \lambda)P(O_{t+1,...,O_T}|q_t = S_j, \lambda)P(q_t = S_j|\lambda)} \tag{5.29}$$

Using the product rule

$$P(AB) = P(A|B)P(B)$$

we have

$$\gamma_t(i) = \frac{P(O_1, .., O_t, q_t = S_i|\lambda)P(O_{t+1,...,O_T}|q_t = S_i, \lambda)}{\sum_{j=1}^{N} P(O_1, .., O_t, q_t = S_j|\lambda)P(O_{t+1,...,O_T}|q_t = S_j, \lambda)} \tag{5.30}$$

which can be rewritten using the definitions of α and β

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{5.31}$$

The product of $\alpha_t(i)$ and $\beta_t(i)$ explains the whole observation sequence given that the system is in state S_i at time t. This is normalized by all possible intermediate states and guarantees that

$$\sum_i \gamma_t(i) = 1 \tag{5.32}$$

Modeling each state with k-multiple Gaussians the responsibility variable is calculated by [55]

$$\gamma_t(j,k) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)} \frac{W_{jk}b_{jk}(O_t)}{\sum_{k=1}^{K}W_{jk}b_{jk}(O_t)} \tag{5.33}$$

where W_{jk} is the mixing weights for the k'th Gaussian component of S_j, with $\sum_{k=1}^{K} W_{jk} = 1$ and $W_{jk} \geq 0$ for $1 \leq j \leq N$ and $1 \leq k \leq K$.

5.6.2 Finding the Optimal State Sequence

To find the single best state sequence Q for an observation sequence O, given a model λ, the Viterbi algorithm is applied. The Viterbi algorithm is, just as forward-backward algorithm, based on dynamic programming. The structure of the algorithm is outlined below.
Initialization:

$$\delta_1(i) = \pi_i b_i(O_1) \tag{5.34}$$
$$\psi_1(j) = 0 \tag{5.35}$$

Recursion:

$$\delta_t(j) = \max_i \left[\delta_{t-1}a_{ij}\right] b_j(O_t) \tag{5.36}$$
$$\psi_t(j) = \text{argmax}_i \delta_{t-1}a_{ij} \tag{5.37}$$

Termination:

$$p^* = \max_i \delta_T(i) \tag{5.38}$$
$$q_T^* = \text{argmax}_i \delta_T(i) \tag{5.39}$$

Backtracking:

$$q_T^* = \psi_{t+1}(q_{t+1}^*),\ \text{t = T-1,T-2,...,1} \tag{5.40}$$

where $\delta_t(i)$ is the highest probability along a single path at time t which ends in state S_i. The variable $\psi_t(j)$ contains the information of the state that maximizes $\delta_t(j)$ at time $t-1$. By backtracking through $\psi_t(j)$ the optimal state sequence can be found.

5.6.3 Training Model Parameters

When training the model parameters, $\lambda^* = \{A, \mu, \Sigma, \pi\}$, the goal is to maximize the likelihood of the training data, $P(O|\lambda^*)$. Applying the Baum-Welch algorithm, which is an iterative procedure based on the EM algorithm introduced in section 5.5.1, the model parameters can be trained.

5.6.3.1 Baum-Welch Algorithm

Before the different steps of the Baum-Welch algorithm are outlined, the variable $\xi_t(i,j)$ is defined as the probability of being in state S_i at time t and in state S_j at time $t+1$, given the observation sequence O and model λ [3]:

$$\begin{aligned}
\xi_t(i,j) &= P(q_t = S_i, q_{t+1} = S_j|O,\lambda) \\
&= \frac{P(O|q_t = S_i, q_{t+1} = S_j,\lambda)P(q_t = S_i, q_{t+1} = S_j|\lambda)}{P(O|\lambda)} \\
&= \frac{P(O|q_t = S_i, q_{t+1} = S_j,\lambda)P(q_{t+1} = S_j|q_t = S_i,\lambda)P(q_t = S_i|\lambda)}{P(O|\lambda)} \\
&= \frac{1}{P(O|\lambda)}P(O_1,...,O_t, q_t = S_i|\lambda)P(O_{t+1}|q_{t+1} = S_j,\lambda)... \\
&P(O_{t+2},...,O_T, q_{t+1} = S_j,\lambda)a_{ij} \\
&= \frac{1}{P(O|\lambda)}P(O_1,...,O_t, q_t = S_i|\lambda)P(O_{t+1}|q_{t+1} = S_j,\lambda)... \\
&P(O_{t+2},...,O_T, q_{t+1} = S_j,\lambda)a_{ij}P(q_t = S_i|\lambda)
\end{aligned}$$

Using that

$$\begin{aligned}
\alpha_t(i) &= P(O_1,...,O_t, q_t = S_i|\lambda) \\
\beta_{t+1}(i) &= P(O_{t+2},...,O_T, q_{t+1} = S_j,\lambda) \\
b_j(O_{t+1}) &= P(O_{t+1}|q_{t+1} = S_j,\lambda) \\
a_{ij} &= P(q_{t+1} = S_j|q_t = S_i,\lambda)
\end{aligned}$$

the expression for $\xi_t(i,j)$ is:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{\text{all i}}\sum_{\text{all j}} P(q_t = S_i, q_{t+1} = S_j|O,\lambda)} \tag{5.41}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{\text{all i}}\sum_{\text{all j}} \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \tag{5.42}$$

Calculating $\sum_t \xi_t(i,j)$ yields the expected number of transitions from state S_i to S_j, while calculating $\sum_t \gamma_t(i)$ gives the total number of transitions from S_i. Calculating the ratio of these two gives the probability of transition from S_i to S_j, which provide us with an estimate of A as suggested by equation 5.43. To estimate μ and Σ for each k-Gaussian the observation sequence is weighted with the responsibility γ.
The Baum-Welch algorithm for a continuous Gaussian emission HMM is outlined below:

1. Initialize the HMM parameters A, μ, Σ, W and π. Often A, W and π are randomly initialized considering the relevant restrictions, while μ and Σ can be initialized using the k-means algorithm [8]. Following initialization the log-likelihood is calculated.

2. Expectation step:
   Estimate $\alpha_t(i)$ and $\beta_t(i)$ using the forward-backward algorithm from section 5.6.1.1.
   Evaluate $\xi_t(i,j)$ and $\gamma_t(i,k)$ using equations 5.42 and 5.33, respectively.

3. Maximization step:
   Calculate the model parameters for λ:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^{K} \gamma_t(i,k)} \tag{5.43}$$

$$\hat{\mu}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(i,k) O_t}{\sum_{k=1}^{K} \gamma_1(i,k)} \tag{5.44}$$

$$\hat{\Sigma}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(i,k)(O_t - \mu_{jk})(O_t - \mu_{jk})^T}{\sum_{k=1}^{K} \gamma_1(i,k)} \tag{5.45}$$

$$\hat{\pi}_i = \frac{\sum_{k=1}^{K} \gamma_1(i,k)}{\sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_1(i,k)} \tag{5.46}$$

$$\hat{W}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(i,k)}{\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(i,k)} \tag{5.47}$$

4. Evaluate the log likelihood.

5. Stop if convergence criteria are met or repeat step 2-5.

Using multiple sequences all parameters are averaged over all observations for all sequences.

5.6.4 Types of Transition Structures

A way to force the HMM to have a specific transition structure is to initialize A with zeros for state transition coefficients where a transition should not be able to occur. Since the numerator of equation 5.43 is always going to be zero for a_{ij} initialized as zero, the transition element will never be updated. An ergodic or fully connected model is initialized with values different from zero for all elements of the transition matrix, while a strict left-right model is restricted to making only transitions to higher states. Illustrations of different transition structures in HMM models are presented in Figure 5.6.

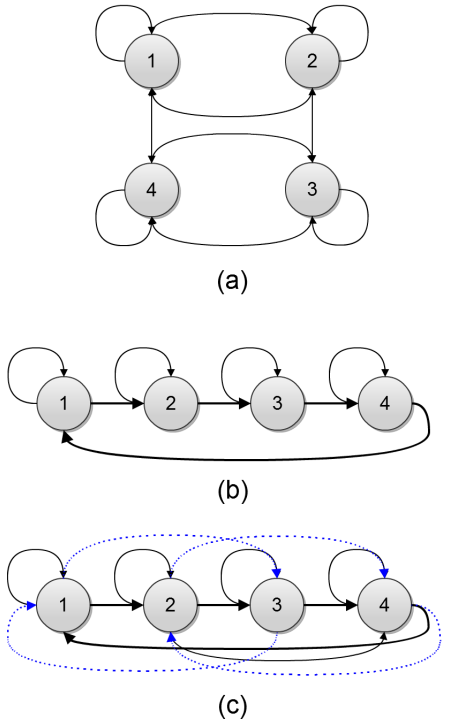


Figure 5.6: Simple illustrations of three types of transition HMM's. a) A fully connected 4 state ergodic model. b) A 4 states left-right model with a recirculation loop. c) A 4 state left-right model with a recirculation loop able to make transition two states ahead. The blue dotted arrows indicate a transition where a state is skipped. The models in this thesis are given the notation, FULL, LR1 and LR2, respectively.

5.7 Implementation Issues

Implementing the entire Hidden Markov Model from scratch leads to different issues some of which are described in the following.

5.7.1 Underflow Problems

Working with HMM's frequently requires computations using small probability values. While in theory this poses no problem, working with finite memory machines like computers, it may cause implementations to crash, due to *underflow* [8]. Underflow occurs when the result of a computation is smaller than the computer is able to represent in its memory and it then sets the parameter to zero. At best this can cause a degraded solution and in the worst case, the program crashes, when NaN's are introduced. Two initiatives were taken in order to overcome these problems. The first is the scaling of the forward-backward algorithm, which is a well-documented underflow solution in HMM literature [8], [3] and [55]. However, working with 8-lead covariance structures caused further problems while trying to evaluate probabilities in an 8-dimensional space. Hence, all related calculations were reworked to be represented in the log domain.

5.7.1.1 Scaling of the forward-backward algorithm

When calculating the forward-backward algorithm both α and β are multiplied with small probability values, which will cause underflow for longer sequences. In order to avoid this, $\alpha_t(i)$ is normalized by multiplying it with the scaling factor c_t

$$c_t = \frac{1}{\sum_j \alpha_t(j)} \tag{5.48}$$

Since the magnitude of $\alpha_t(i)$ and $\beta_t(i)$ is comparable [55] the same scaling factor c_t can be used to scale $\beta_t(i)$

$$\beta_t(i) = c_t \beta_t(i) \tag{5.49}$$

Having scaled $\alpha_t(i)$ the probability of a sequence O given a model λ can no longer be calculated using equation 5.23. Instead the following expression is applied

$$\ln(P(O|\lambda)) = -\sum_t \log c_t \tag{5.50}$$

5.7.1.2 Log space calculations

To avoid the emission probabilities to numerically underflow, the multivariate Gaussian distribution given by equation 5.15 is calculated in the log domain:

$$\ln\left(\mathcal{N}(\mathbf{O}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) = \frac{n}{2}\ln\left(2\pi\right) - \frac{1}{2}\ln\left(|\boldsymbol{\Sigma}|\right)\sum\left(\mathbf{O}_t - \boldsymbol{\mu}_{jk}\right)^T\boldsymbol{\Sigma}^{-1}(\mathbf{O}_t - \boldsymbol{\mu}_{jk}) \tag{5.51}$$

The emission probability given in equation 5.14 is then rewritten in log domain:

$$\ln(\boldsymbol{b}_{jO_t}) = \sum_{k=1}^{K}\ln(W_k) + \ln(\mathcal{N}(\boldsymbol{O}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))) \tag{5.52}$$

While products are easily replaced by addition in log domain, the sum cannot as easily be replace by another operator. Summation has to be performed in the linear domain, but simply transforming the values to the linear domain using $\exp(-)$ would instantly cause numerical underflow and results would be lost. To solve this the *Log-Sum-Exp trick* (LSE) is applied [57]. This is a computational trick where the maximum of the log values is found and subtracted from all of the log values. This operation shifts all the large magnitude negative values towards zero leaving the previously maximal value zero. Having done this, taking the $\exp(-)$ of the values would not cause underflow [1]. Then the values are summed in the linear domain and returned to the log domain, by simply taking the log. Finally the maximum value that was subtracted at first is added, to return the values to their proper magnitude. It can be written as

$$LSE[\ln(x)] \equiv \ln(\sum_{n=1}^{N} x_n) \qquad = \ln(x_m) + \ln[\sum_{n=1}^{N}\exp^{\ln(x_n)-\ln(x_m)}] \tag{5.53}$$

with x_m being the largest term. The cost of avoiding numerical underflow is a high increase in computations which can effectively decrease the speed of the program.

5.7.2 Singularity of Covariance

A significant problem while applying the EM algorithm in log-likelihood maximization of the Gaussian components for each state, is the presence of singularities. A Gaussian component could be fitting a single point or points of similar value. This would cause the variance to go towards zero and the log-likelihood

[1] With the exception of values both containing relatively high and extremely small numbers. In this case the LSE trick will not be able to shift values away from potential underflow.

to go towards infinity. This is often thought of, as severe over-fitting [8] and can cause the program to crash due to the introduction of Inf values when taking the inverse of Σ in equation 5.51. Inf values can cause NaN values in MATLAB® if an Inf value is divided by another Inf value or if a positive Inf value is added with a negative Inf value. A common way to avoid this singularity, is to reset the mean and covariance to some random values [8] when detecting a collapsing Gaussian component. If one wants to be able to model such singularities, one could simply "freeze" the variance at some specified small number when the covariance fall below this. However, this can cause the log-likelihood to increase at some point, but would ensure stability of the program. Illustration of an arising singularity is presented in Figure 5.7.

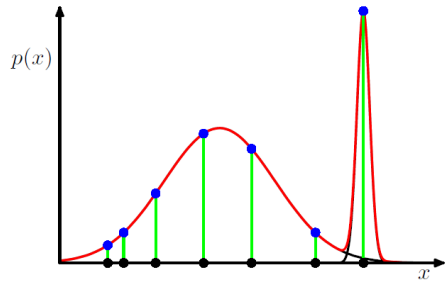


Figure 5.7: Illustration of an arising singularity for a Gaussian component. Modified from [8].

Another form of singularity that can arise is when the covariance matrix Σ is close to singular or badly scaled. This happens if the data modeled by the Gaussian is highly correlated. Consider multiple two-dimensional data points having the same value in the first dimension and different values in the second dimension. This would cause variance in the first dimension to be zero along with the covariance. Taking the inverse to such covariance matrix would cause Inf values caused by division by zero. To overcome this problem the covariance matrix Σ must be conditioned to secure numerical stability. Pekka Paalanen [52] have proposed following conditioning algorithm.

Covariance matrix conditioning

1. While Σ is not positive definite Do
2. Extract the diagonal d
3. If d is greater than some predefined value min_{limit} the diagonal is increased with 1% Else
4. 1% of max of the diagonal is stored in m.

5. If m is less than min_{limit} it is set equal to min_{limit}
6. The diagonal of Σ is increased by m

5.7.3 Speed

MATLAB® is a high level programming language where programming and testing of algorithms are fairly easy compared to low level language such as C++. However, the speed of lower level language is still superior. When performing data heavy machine learning in MATLAB®, one should put a great deal of thought into the implementation in order to obtain acceptable execution times. The best ways to optimize MATLAB® code is to consider [43]

1. Code vectorization.
2. Variable preallocation.
3. Memory access optimization.

Doing the above vastly increase the speed of the MATLAB® program. However, being forced to use multiple loops or applying the LSE trick, as in the forward-backward recursion in log domain, one can greatly improve the speed by building the functions as MEX files. A MEX file or Matlab EXecutable is an interface between MATLAB® and a C/C++ function that can vastly increase speed of functions that use computational heavy loops.

5.8 Support Vector Machine

Support Vector Machine or SVM is one of the most widely used classification algorithms and have shown very good empirical results [61]. The basic idea of the SVM is to create an optimal linear decision boundary (hyperplane) in a given feature space for two or more classes using a subset of training examples. The selected subset of training examples represents the decision boundary and are called *support vectors*. As opposed to the HMM, which is a generative model, the SVM is a discriminative model.
Since only two class classification is of interest, multiclass SVM classification will not be discussed in the following.

Consider two classes each having a feature vector $\mathbf{x}$ containing two features for each training example. Given that the classes are separable by a linear hyperplane, there exists an infinite number of decision boundaries that could divide the two classes. An illustration of this scenario with N = 8 training examples for each class, is illustrated in Figure 5.8.
In order to choose an optimal hyperplane the SVM uses the Maximum Margin method.

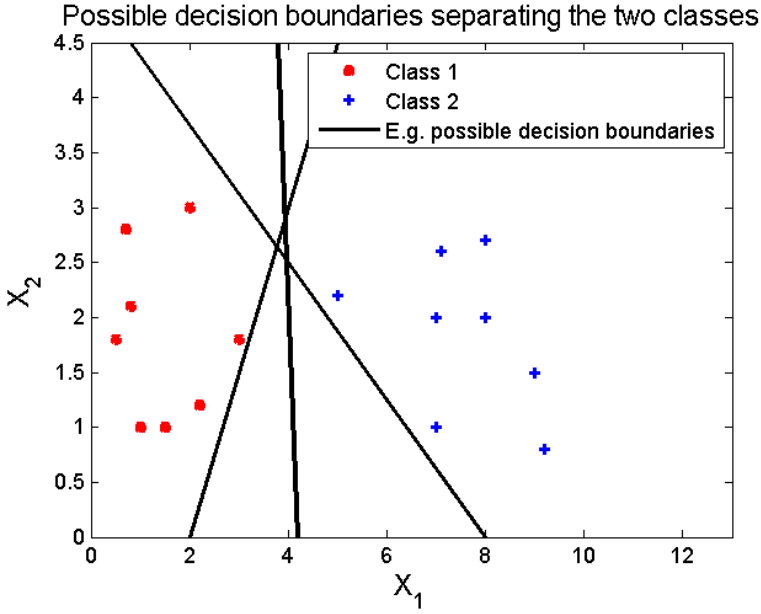


Figure 5.8: Illustration of two classes, red and blue, and different possible decision boundaries that could separate the two classes perfectly.

5.8.1 Maximum Margin Hyperplane

To determine the margin of a given hyperplane one shifts parallel hyperplanes in all dimensions until one of them coincides with a training example. The distance between the parallel hyperplanes, shifted until they to coincide with the training data, is called the margin. The decision boundary with the largest margin is called the maximum margin hyperplane. An illustration of the maximum margin hyperplane and its margin is presented in Figure 5.9. The maximum margin

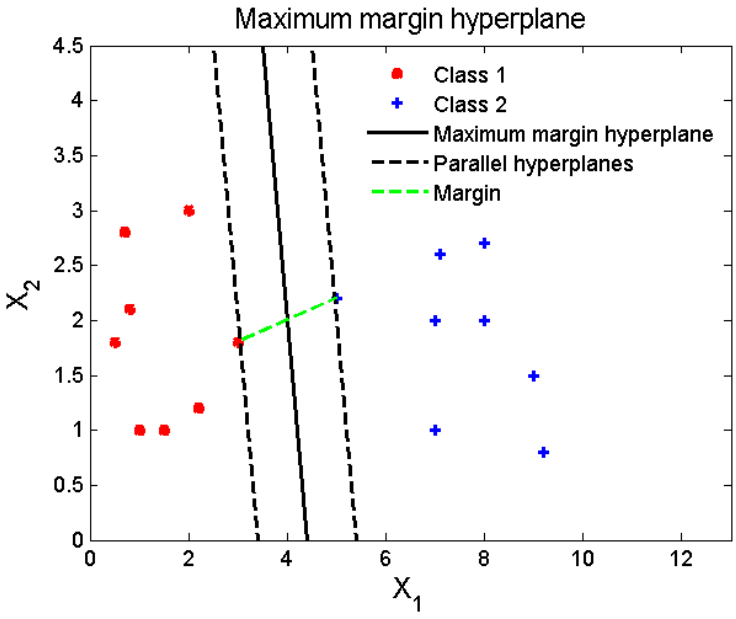


Figure 5.9: Maximum margin hyperplane (solid black line) and the illustration of the margin (dotted green line).

hyperplane is chosen as the decision boundary since it theoretically has a better generalization error which can potentially lead to a better classification of unseen test examples [8].
A linear classifier model can be written as:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{5.54}$$

where $\mathbf{w}$ are the weights, b the bias of model and $\mathbf{x}$ are features of the training examples. Each training example is labeled with $y_i \in \{-1, 1\}$ with i = 1,2,..,N where N is the number of training examples. The two parallel hyperplanes b_{i1}

and b_{i2} can be expressed as:

$$b_{i1} : \mathbf{w} \cdot \mathbf{x} + b = 1 \tag{5.55}$$

$$b_{i2} : \mathbf{w} \cdot \mathbf{x} + b = -1 \tag{5.56}$$

if the $\mathbf{w}$ and b are rescaled [61].
Any test example $\mathbf{z}$ would be classified using

$$y = \begin{cases} 1, & \text{if } \mathbf{w} \cdot \mathbf{z} + b > 0 \\ -1, & \text{if } \mathbf{w} \cdot \mathbf{z} + b < 0 \end{cases} \tag{5.57}$$

To find the model parameters given in equation 5.54 one must maximize the margin, which is equivalent to minimizing following objective function [61]:

$$\min \frac{||w||^2}{2} \tag{5.58}$$

$$\text{subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \tag{5.59}$$

for i = 1,2,..,N.
This is a convex optimization problem, since the objective function is quadratic and the constraints are linear [61]. A known method for solving constrained optimization problems is the Lagrange multiplier method (see Appendix A.1). The Lagrangian for the optimization problem is:

$$L_P = \frac{||\mathbf{w}||^2}{2} - \sum_{i=1}^{N} \lambda_i (y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1) \tag{5.60}$$

with λ_i being the Lagrange multipliers. The first term is the original objective function, while the second term captures the inequality constraints. To minimize the Lagrangian the first-derivative of equation 5.60 with respect to $\mathbf{w}$ and b are taken and set equal to zero:

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \tag{5.61}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0 \tag{5.62}$$

To handle the inequality constraints, they are transformed into equality constraints using the Karush-Kuhn-Tucker (KKT) conditions (see Appendix A.1).

$$\lambda_i \geq 0 \tag{5.63}$$

$$\lambda_i [y_i(\mathbf{w} \cdot x_i + b) - 1] = 0 \tag{5.64}$$

From equation 5.64 it is seen that only training examples where $y_i(\mathbf{w}\cdot x_i+b)=1$ can have Lagrange multipliers different from zero. These training examples are the support vectors.
A way to simplify the Lagrangian is to transform it into it's dual problem by substituting equation 5.61 and 5.62 into equation 5.60 giving:

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i \mathbf{x}_i \cdot \mathbf{x}_j \tag{5.65}$$

It is seen that the dual problem Lagrangian only depends on the Lagrange multipliers and the training examples. Because of the negative quadratic term the problem is now a maximization problem. The equation can be solved using numerical techniques such as the Sequential Minimal Optimization method or Quadratic programming [61]. When the Lagrange multipliers have been determined equation 5.61 and 5.64 is used to obtain $\mathbf{w}$ and b. Then the decision boundary can be expressed as:

$$\left(\sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \cdot \mathbf{x} \right) + b = 0 \tag{5.66}$$

5.8.2 Linear Nonseparable Classification

Until now it has been assumed that the two classes could be perfectly separated by a linear hyperplane. However, this is not always possible. This could occur if noise is present or if the features does not represent a strong difference between the two classes. To overcome the issue the *soft margin* method is applied where some amount of training misclassification is allowed. Thus, the learning of the SVM decision boundary is a trade-off between training misclassification and the width of the margin. To model this the constraints are relaxed by introducing a positive valued slack variable ξ and by modifying the objective function [61] such that increasing slack is penalized. The slack is an estimate of the error introduced by the decision boundary in the training examples. The modified objective function is written as [8]:

$$f(w) = \frac{||w||^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{5.67}$$

where C is a user-specified parameter regulating the penalty of misclassification. The Lagrangian for this is:

$$L_P = \frac{||\mathbf{w}||^2}{2} + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \lambda_i (y_i(\mathbf{w} \cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^{N} \mu_i \xi_i \tag{5.68}$$

where both μ_i and λ_i are Lagrangian multipliers. The first two terms are from the objective function, the third represents the inequality constraints and the last term is a non-negativity requirement of the slack variable.
The new KKT conditions are [61]:

$$\lambda_i \geq 0, \xi_i \geq 0, \mu_i \geq 0 \tag{5.69}$$

$$\lambda_i[y_i(\mathbf{w} \cdot x_i + b) - 1 + \xi_i] = 0 \tag{5.70}$$

$$\mu_i \xi_i = 0 \tag{5.71}$$

The new first-order derivative of L with respect to the parameters $\mathbf{w}$, b and ξ are:

$$\frac{\partial L}{\partial \mathbf{w}_j} = 0 \Rightarrow \mathbf{w}_j = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_{ij} \tag{5.72}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0 \tag{5.73}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \lambda_i + \mu_i = C \tag{5.74}$$

To find the dual formulation the three above equations are substituted into 5.68 and it is found that it is identical with equation 5.65, but with the $0 \leq \lambda \leq C$. The solution can be found in the same manner as in the separable case.

5.8.3 Nonlinear SVM

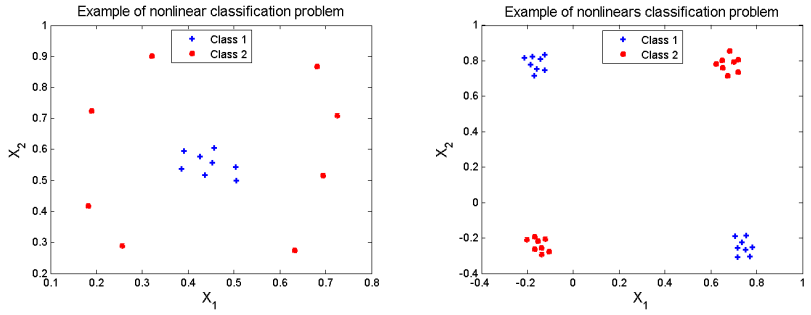
Sometimes is a linear decision boundary is not the best choice when searching for the optimal classification hyperplane. Two illustrations of such situations are presented in Figure 5.10.
The trick to solve this is to transform the data from the original feature space into a new feature space $\phi(x)$ (often of higher dimensions), such that a linear hyperplane can separate the training examples satisfactorily. As an example the following transformation could be applied to the nonlinear problem in Figure. 5.10(a):

$$\phi(x_1, x_2) \longrightarrow (x_1^2 - x_1, x_2^2 - x_2) \tag{5.75}$$

and the problem in Figure 5.10(b) could be applied the transformation:

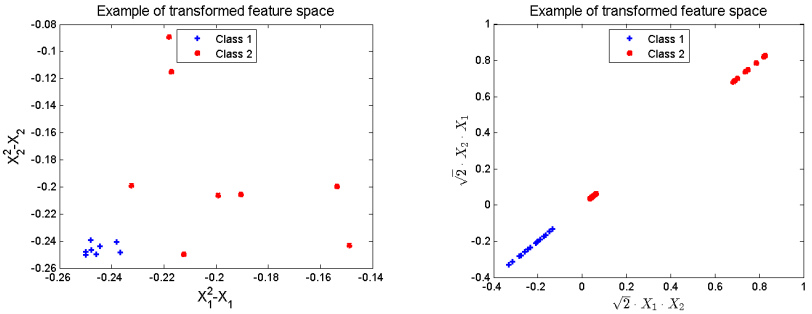
$$\phi(x_1, x_2) \longrightarrow (\sqrt{2}x_1 x_2, \sqrt{2}x_2 x_1). \tag{5.76}$$

The results are shown in Figure 5.11. The example clearly shows how a linear hyperplane could perfectly separate the two classes in the new feature space.



(a) (b)

Figure 5.10: The two examples in Figure 5.10(a) and 5.10(b) illustrate scenarios where a nonlinear decision boundary would be preferable.



(a) Problem from Figure 5.10(a) transformed into a new feature space.

(b) Problem from Figure 5.10(b) transformed into a new feature space

Figure 5.11: Figures 5.11(a) and 5.11(b) show how the features of the classifications problems depicted in Figures 5.10(a) and 5.10(b) are transformed using $\phi(x_1, x_2) \longrightarrow (x_1^2 - x_1, x_2^2 - x_2)$ and $\phi(x_1, x_2) \longrightarrow (\sqrt{2}x_1x_2, \sqrt{2}x_2x_1)$, respectively. In the new feature spaces the problems can easily be separated by a linear hyperplane.

A nonlinear SVM can be learned by simply replacing x with the transformed $\phi(x)$ in the equations stated in section 5.8.2. The curse of dimensionality and the multiple dot products that must be calculated in equation 5.65 present a potential problem. This is avoided using the *kernel trick*, which is a method for computing the dot products in the transformed space using only the original features. The function that computes the dot product using the original feature

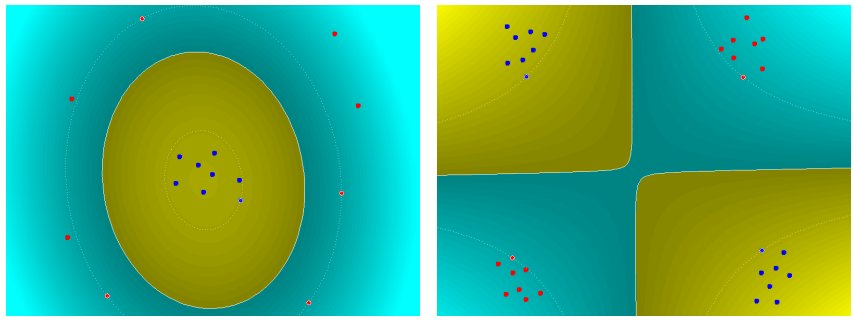
space is called the *kernel function* [61]. Kernels used in a nonlinear SVM must satisfy Mercers' theorem, which ensures that the kernel can always be expressed as the dot product of two vectors in some high-dimensional space. An example of a kernel is the quadratic polynomial:

$$K(u,v) = (\mathbf{u}\cdot\mathbf{v}+1)^2 \tag{5.77}$$
$$= u_1^2v_1^2 + u_1^2v_1^2 + 2u_1v_1 + 2u_2v_2 + 1 \tag{5.78}$$
$$= (u_1^2, u_2^2, \sqrt{2}u_1, \sqrt{2}u_2, 1)\cdot(v_1^2, v_2^2, \sqrt{2}v_1, \sqrt{2}v_2, 1) \tag{5.79}$$
$$\tag{5.80}$$

which leads to the following transformation

$$\phi(x_1, x_2) \longrightarrow (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1) \tag{5.81}$$

Thus the two dimensional feature space is lifted to a five dimensional feature space, but the dot products can still be calculated in the original feature space due to the kernel trick.
The nonlinear decision boundaries in the original features space of the problems presented in Figure 5.10 are presented in Figure 5.12. These were created using a SVM with second order polynomial kernel.



(a) (b)

Figure 5.12: Illustration of the nonlinear decision boundaries for the two examples in Figure 5.10(a) and 5.10(b) in the original feature space using an SVM with a second order polynomial kernel. The cyan area illustrates where a test example would be classified as red, while the yellow area would label the test example as blue.

Chapter 6

Model Identification

6.1 ECG Acquisition and Study Population

The data were collected from the Marquette Electronics (Milwaukee, WI) ECG database and the ECG acquisition for the normal ECGs were performed as stated in [12]. Ten-second ECGs were recorded with a standard 12-lead system on a MAC15 digital ECG recorder (GE Medical Systems, Milwaukee, WI). The subjects were in a supine position and the ECGs were obtained after the subjects had 5 minutes of rest. The sampling rate was 250 Hz and the amplitude resolution was 1.22 μV. The ten-second ECGs from the LQT2 patients were acquired on a MAC5000 digital ECG recorder (GE Medical Systems, Milwaukee, WI) with the subjects being in a supine position. The duration of the ECGs was 10 s, and the sampling rate was 500 Hz (Graff et al., 2009). The LQT2 ECGs were down sampled to 250 Hz using the `MATLAB`® function `resample.m`. The data consisted of ECGs from 64 LQT2 patients and 64 normal subjects that were gender and age matched by extraction from a database of 1109 normal ECGs, giving a total sample size of 128. The normal subjects consisted of 29 men of mean age 36±18 and 35 women of mean age 39±17. There were 29 men of mean age 36±18 and 35 women of mean age 38±18 in the LQT2 patient group.

Unfortunately PLI filtering was assumed to be performed during the ECG

recording but experience showed that the data applied in this work is completely raw. By visual inspection of the ECGs and their spectra none of the normal ECGs contain visible PLI whereas a few of the LQT2 ECGs do. However, those LQT2 ECGs containing PLI also seems to contain more biological noise which masks the PLI in the time domain. Due to the time limit of the project it was chosen to filter only BW and instead focus the effect of noise on classification to assess whether the potentially inadequate filtration impairs the classification method.

6.2 Model Training Setup and Implementation of HMM

A significant amount of the project duration was spent on the satisfactory implementation of the multi-lead continuous emission Hidden Markov model. It turned out that the limited time of the project and the speed of the program limited the extent to which different methods could be applied in the characterization and discrimination of ECGs in this work. However, within the limits of those methods, a thorough approach was adopted, which is described in the following.

Transition matrices: Three transition types were chosen. Due to the periodic nature of ECGs the left-right (LR) type transition was adopted in two forms. First, each non self-transition could occur only to the immediately following state, termed one forward degree of freedom (LR1). Correspondingly a two forward degree of freedom LR type transition was also applied to investigate the impact of the extra flexibility (LR2). Transition from the "last" state to the "first" were allowed to maintain periodicity. Also the full type transition (FULL) was applied, in that results could either confirm the choice of the LR types or could potentially capture other trends in the data. The full and LR type transition models were applied, since these were found to be the most commonly used in studies covering HMM and ECG modeling.

Number of states: Consulting the relevant literature shows that the chosen number of states varies between studies, but it seems to be in the range of 5 to 35 depending on the application. To be able to reproduce a single beat such that medical staff would recognize it as an ECG, the number of states should be more than 20. Based on this it was decided to train models defined with 5 to 50 states at increments of 5 [5,10,15,..,50]. Covering this large range of states could potentially aid in determining the trade off between generative and discriminative properties, if any.

Stopping criteria: The choice of stopping criteria will often be a compromise between finding the optimal parameters and time consumption during training. Through empirical experiments, tweaking the tolerance for the minimum acceptable change in log-likelihood, it was decided to use 10^{-2} as the stopping criterion. To secure an upper bound for model training a maximum of a 1,000 iterations was set as limit. However, not a single time during the training of the models did the number of iterations reach the maximum allowable.

Emission distribution: It was chosen to model the emission distributions using Gaussian mixtures. The method is well understood and employs easily interpretable distributions and was found to be the most commonly used emission distribution in studies concerning HMM. Models using one or two Gaussians per state were applied in this project.

Parameter initialization: The parameters were initialized differently based on the state transition type. Using the full transition model the parameters $\mathbf{A}$ and Π were initialized randomly using the built-in random `MATLAB`® function. The means μ were initialized with 100 iterations of k-means. The k-means function from the Statistics Toolbox Version 8.0 in `MATLAB`® was applied, using a squared Euclidean distance measure, a uniform initial start of centroids and three replicates. One centroid was used for each state. Using mixed Gaussians for each state, the calculated k-means centroids were shifted a small random amount for each Gaussian component. The variances were initialized using the data samples assigned to the cluster with the covariances initially set to zero. Considering the LR models, the state transitions were restricted to a LR cyclic sequence. Through empirical experiments it was found to be undesirable to use the k-means as initialization. It was observed that the model could lock on undesirable local maxima. To establish a more flexible initialization the means were set as the mean of the total data shifted a small amount using a random function. The variance was set to be that of the entire data set. All transition probabilities were initialized equally and the initial state probability was set with the first state to be more likely than the others.

Handling singularity issues: To avoid the issues of covariance matrices becoming singular due to very correlated data, the covariance fixer routine described in section 5.7.2, was applied. To avoid collapsing Gaussians, e.g. when fitting a single cluster of points having the same value, a minimum variance limit was defined. Should the variance fall beneath this limit it would be locked at that specific value. This was done to be able to model a straight line segment, which could be the iso-line in an ECG. Should the covariance fixer or the limitation of the variance due to potential collapse be used during training, the program displays a warning. Using LR state transition models, these precautions were never applied at any combination of states or number of Gaussian components in the data. The precautions were, however, necessary to be able

to model data with a high number of states in the full transition model.

Underflow: Underflow was a considerable issue while implementing the HMM. Calculating probabilities in a 8-dimensional space using the ECG data yielded extremely small values, some of which could not be represented in the regular domain why the probabilities were calculated in the log domain. The down stream effect was that the forward-backward algorithm, the estimation step of the EM-algorithm and the update of the transition matrix had to be implemented in log domain as well. The consequence of this was that the well-structured MATLAB® code that was vectorized for speed needed rewriting. The necessary application of the log-sum-exp trick multiple times caused a large increase in computation time.

Speed: To be able to investigate multiple model setups within a reasonable time frame some effort was put into code optimization. The built-in MATLAB® profiler was used to locate the bottlenecks in the code. Experience showed, that in a significant portion of the iterations, the backward-algorithm and the updating of the transition matrix could be performed without the log-sum-exp trick without underflowing, making vectorization in linear domain possible. Should underflow occur, recalculation was carried out with all steps of the parameter estimation in the log domain. No difference in the resulting models was observed using this speed improvement trick.
Furthermore, the forward-algorithm, the backward-algorithm and the transition matrix update were reprogrammed as MEX files. The effect of the code optimization is presented in Figure 6.1 for a specific training example, and indicates a reduction in computation time by a factor of 10. As a result of the optimization the model training setup could be performed in a week using four average laptop computers instead of a month.

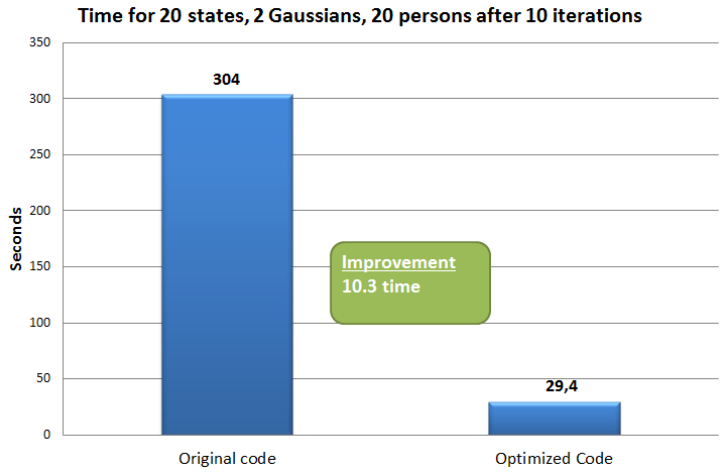


Figure 6.1: Illustration of the effect of optimizing the program code in 10 iterations modeling 20 ECG's using a full transition matrix with 20 states each having two Gaussian components. The test was performed on a 64-bit laptop using 2.53GHz Intel core i5 processor with 8 GB of RAM

6.3 Classification Setup

Different schemes were attempted in order to find the model with the best classification accuracy. For each different model setup the HMMs were trained using either normal or LQT2 training ECG's. Subsequently the probability of the test ECG's given the normal and LQT2 model, respectively, were calculated. The test ECG's were classified according to the model having the highest probability of generating the test ECG in question.
To utilize the multiple HMM's, two different classification schemes were developed.

1. **Biggest difference**: For each test ECG the difference in log-likelihood evaluated is evaluated for each of the included HMM models. The model with the biggest difference classifies the test ECG.

2. **Majority voting**: Each of the included models votes whether a given test ECG is normal or LQT2. The ECG is classified according to the class receiving the majority of the votes. If the number of votes are equal, the model with the biggest difference in log-likelihood decides.

To investigate whether a decision boundary in a higher feature space could improve classification accuracy an SVM using seven different kernels was applied using several combinations of HMM differences between normal and LQT2 log likelihood as the input feature. The seven kernels applied were linear, second order polynomial, third order polynomial, fourth order polynomial, quadratic kernel, radial basis function and multi-layer perceptron. All kernels were used with `MATLAB`®'s default parameter settings.
The following two different schemes were used to select HMM models to be combined:

1. Combining models based on their accuracy for each transition type and number of Gaussian components.
2. Combining the best six models based on their accuracy disregarding their transition type and number of Gaussian components.

A simple flow chart of the entire program structure is outlined in Figure 6.2. The model or combination of models with the classification scheme yielding the highest accuracy is selected to be the best model.

As explained in section 2.4 the LQT2 syndrome is usually diagnosed based on the heart rate corrected QT interval (QTc) of the ECG. The QTcB interval ($\frac{QT}{\sqrt{RR}}$) for each of the 128 subjects were calculated using the commercial ECG software `MUSE`® in order to compare the best model accuracy with that of the QTc and to investigate if the model could contribute to the accuracy, were they combined. To do this the QTc was used as input to the SVM in a 4-crossfold train/test classification scheme using the different kernels. Subsequently, both the log-likelihood difference from the HMMs of the best model and QTc were applied in the SVM classification scheme using the kernel achieving the highest mean accuracy found by the best model.

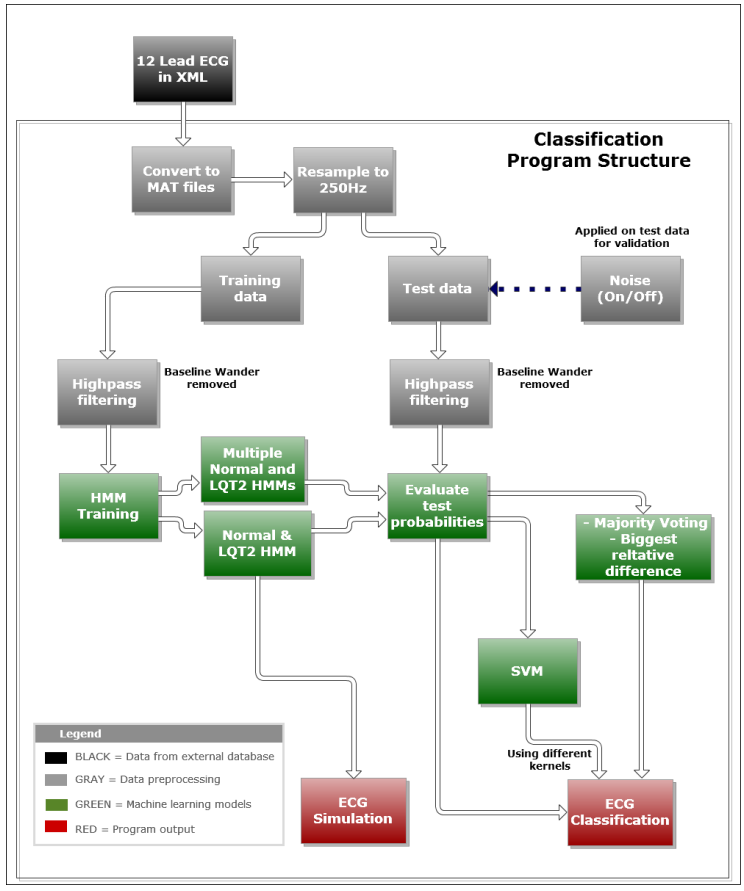


Figure 6.2: Program structure for the classification setup. The ECG's are extracted from the ECG database MUSE® as XML files. The ECG's are converted to mat-files, resampled to 250 Hz and high pass filtered using a bidirectional digital high pass Kaiser window FIR filter with a cutoff frequency of 0.5 Hz, before feeding the signals to the Hidden Markov model. The 8-dimensional continuous Hidden Markov model train different models, which are evaluated doing 4-crossfold validation. ECG's simulated from the HMM are output. The probabilities from a single or multiple HMM's are used as input in a SVM, using different kernels, to classify the ECG's. The multiple HMMs are also used in different classification schemes, such as majority voting and decision based on biggest log-likelihood difference. ECG noise can be applied to the test data used in classification.

6.4 Generative Properties

The discriminative properties of the models can not necessarily be related to the physiological process underlying the ECG. Since the choice of model parameters is based on the classification accuracy, there is a risk of capturing population specific properties of the ECG such as e.g. overall amplitude or noise. As both examples are likely to be specific to the data used in this work and not to normal and LQT2 ECGs in general, it is of interest to evaluate the generative properties of the models.

6.4.1 ECG Simulation

To evaluate the generative properties of the models a simulation can be performed where the best models, according to the classification accuracy, are applied. Considering the HMM as a simulator, the transition matrix creates the sequence of hidden states, where each hidden state has an emission distribution with a mean and covariance. As such, there are two sources of randomness in a simulation. First, the generated sequence of hidden states will not be strictly periodic and second the covariance matrix associated with the mean emission values introduces a further source of randomness. Preferably, an ECG simulation of some fixed time length should be generated iteratively and these realizations should be used to find an average ECG simulation of the model in question. However, this would require alignment of the different realizations such that each full ECG cycle or "heartbeat" would be aligned with that of the next sequence. This is not feasible in an automated fashion as the simulations vary considerably. To overcome the problem two measures were taken; the mean emission values were considered and an expected number of self-transitions were applied in the simulation. An exponential state duration is a characteristic of the Markov chain [55] and so the expected number of self-transitions, or duration, can then be defined as:

$$E[d_i] = \sum_{d=1}^{\infty} dp(d_i) = d(a_{ii})^{d-1}(1 - a_{ii}) = \frac{1}{1 - a_{ii}} \tag{6.1}$$

where, d_i is the duration of state i and a_{ii} is the probability of self-transition. This expectation can be applied when simulating the hidden state sequence; when the sequence starts the expectation is calculated and rounded down to nearest integer and a sequence of this length is simulated. When the duration is complete, a new state is drawn. In the case of LR transitions with only one forward degree of freedom the next state drawn will always be the immediately following state. In the case where there are two forward degrees of freedom

there will still be a source of variability, because when the duration is complete, the next state drawn has two outcomes. To accommodate that situation (when drawing a new state) the probability of self-transition is removed and the remaining probabilities are normalized such that they sum to one. Thus, in the case of two forward degrees of freedom the next state is drawn according to a Bernoulli distribution. In the case of the full transition matrix this generalizes to the categorical distribution. In summary, taking expectation of the duration and considering only the mean emissions eliminates randomness for the LR type transition with one forward degree of freedom but leaves randomness in the non-self-transitions for the LR type transition with two forward degrees of freedom and the full type transition. Figure 6.3 illustrates a 35 state (1 Gaussian) ECG simulation of lead V5 for the one degree freedom LR type transition matrix. In the left column the hidden state sequence is generated by random according to the transition matrix. The right column corresponds to the hidden state sequence generated by calculating the expected number of self-transitions. In the top row the emissions are simulated while applying the covariance matrix. In the second row the variance of the mean emissions is shown by plotting the standard deviation. Row three shows the mean emissions and row four shows the corresponding state sequence.

6.4.2 Period of a Transition Matrix

The state sequence in the bottom row of Figure 6.3 suggests that a full pass through the transition matrix corresponds to a single heartbeat. Summing the number of self-transitions and inter state transitions yields the total number of transitions. If the number of transitions is considered to be a number of samples at a specified sampling frequency the "heart rate" of a transition matrix can be calculated. For the one forward degree of freedom LR type transition matrix it is straight forward to collect the total number of transitions as the result is independent of the initial state as long as the return to the initial state is monitored. In the two forward degrees of freedom LR transition type however, there is a source of randomness as mentioned before. Thus, it is necessary to simulate a number of realizations and find an average "heart rate" in this manner. The LR with two degrees of freedom can jump two states at a time, which will increase the modeled "heart rate". Also, experience shows that some states will be close to absorbing. It was chosen to limit number of self-transitions to $1e6$ corresponding to $a_{ii} = 0.999999$ and run the LR with the two degrees of freedom calculation 100 times to find an average "heart rate". In the case of the full transition matrix the bookkeeping with regards to when the process returns to the initial state is more ambiguous in that it may, in theory, return at any time. Thus, the average "heart rate" of the transitions matrices are only

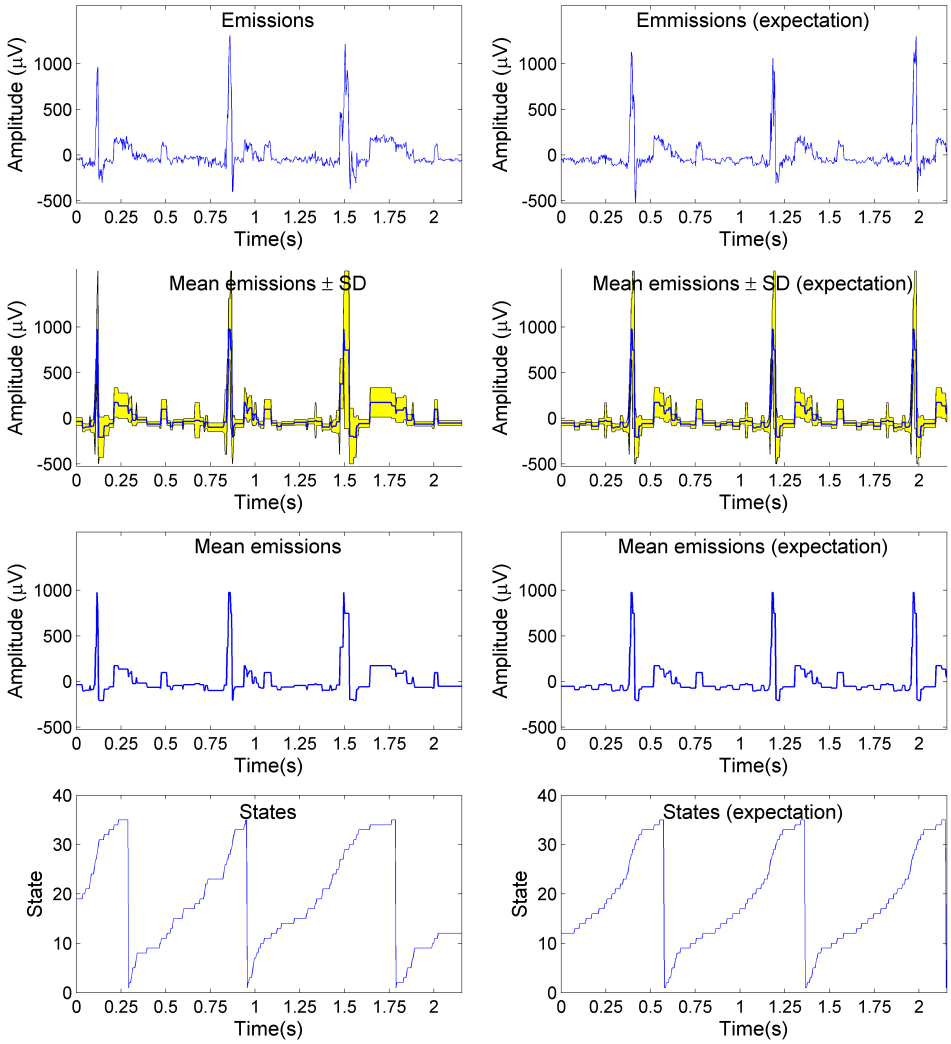


Figure 6.3: ECG simulation. In the left column the hidden state sequence is generated by random according to the transition matrix. The right column corresponds to the hidden state sequence generated by calculating the expected number of self-transitions. In the top row the emissions are simulated while applying the covariance matrix. In the second row the variance of the mean emissions are shown by plotting the standard deviation. Row three shows the mean emissions and row four shows the corresponding state sequence.

calculated for the one and two forward degrees of freedom LR types.
Obviously this "heart rate" will be dependent on the number of states in the model, and so it is of interest to calculate the "heart rate" of all the differently sized transitions matrices for each crossfold. The optimal number of states with regards to classification accuracy then yields a corresponding "heart rate" that can be compared with the true heart rate of the two study populations.

6.5 Most Probable ECG According to Model

The other means of expressing which dynamics the models are capturing are through inspection of the population ECGs creating the most extreme log-likelihoods; the most likely subjects evaluated with the normal model and the LQT2 model and the subjects yielding the largest difference between the normal and LQT2 models. In the latter case, the difference is evaluated as a ratio in the linear domain (subtraction of log-likelihoods) implying that it is the relative size of the likelihoods that are of interest and not the numerically largest difference. Figure 6.4 shows the likelihood for a test set given a normal model plotted against the likelihood for the test set given a LQT2 model. The line through the diagonal represents equality in likelihoods. Blue asterisks denote normal subjects whereas red asterisks denote LQT2 subjects. The green triangles indicate the most likely subject with each model and the black circles represent the maximal probability ratios for the case where the normal model is more likely than the LQT2 model and vice versa. These ECGs will be plotted and inspected in section 7.1.

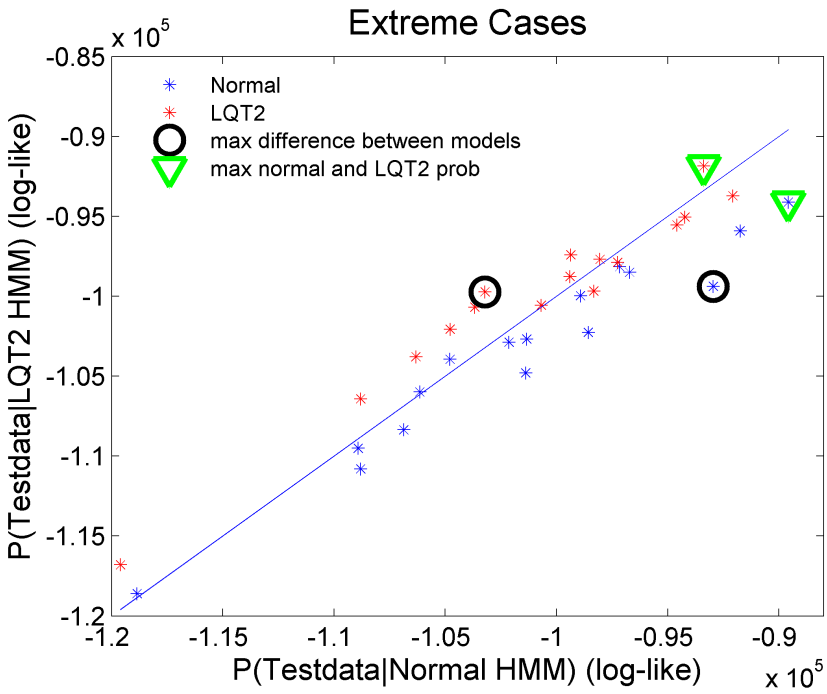


Figure 6.4: Likelihood for a test set given a normal model plotted against the likelihood for the test set given a LQT2 model. The line through the diagonal represents equality in likelihoods. Blue asterisks denote normal subjects whereas red asterisks denote LQT2 subjects. The green triangles indicate the most likely subject with each model and the black circles represent the maximal probability ratios for the case where the normal model is more likely than the LQT2 model and vice versa.

6.6 Verification of Implementations

Different measures are taken to verify that the implementation of the HMMs and related functions behave as expected. The four following tests are performed:

- Modeling 1D artificial signal:
  Constructing a simple signal with a ground truth, the HMM's ability to estimate the parameters can be compared.
- Modeling 2D artificial signal with random component:
  Modeling a clearly sectioned two dimensional signal allows a visual verification of the HMM's ability to capture means and covariance.
- Capture the dynamics of a ECG using optimal state path:
  With regards to ECG modeling, it of interest to evaluate if the HMM can capture the periodic structure of an ECG bin in the hidden state sequence.
- Approximate the HMM of a process by modeling the simulation of a teacher model:
  With the constructed simulator function a simulation is generated applying a predefined HMM which is subsequently approximated by training a new HMM.

All models correspond to the full transition type and a tolerance of 10^{-2} is applied as convergence criterion.

6.6.1 Modeling 1D Artificial Signal

A signal is defined as having three underlying processes, each having two sub processes. Each process and sub process are equal in the number of samples. The signal is presented in Figure 6.5. The mean of the signal processes are:

$$\mu_{Process1} = 5,\ \mu_{Subprocess11} = 2, \mu_{Subprocess12} = 8$$
$$\mu_{Process2} = 24,\ \mu_{Subprocess21} = 21, \mu_{Subprocess22} = 27$$
$$\mu_{Process3} = 48,\ \mu_{Subprocess31} = 45, \mu_{Subprocess32} = 51$$

The standard deviations are equal within the processes and within the sub processes.

$$\sigma_{Process} = 3.1623,\ \sigma_{Subprocess} = 1.0$$

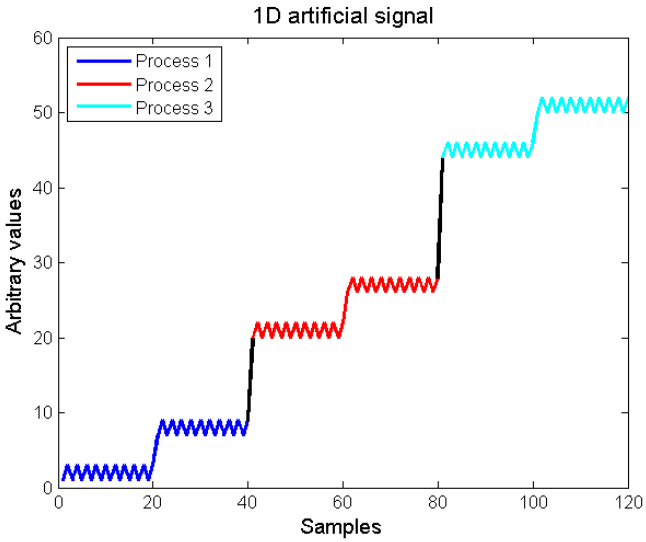


Figure 6.5: An artificial signal with "known" parameters. The signal is thought to be made of three hidden processes each having two sub processes. Each color represents a hidden process, from which the observable values are emitted.

Modeling this clearly sectioned signal with a 3 state 1-Gaussian component HMM, it is expected that the model captures the $\mu_{Process}$ and $\sigma_{Process}$. The sub processes should be discovered when performing the modeling with a 3 state 2-Gaussian component HMM. Considering the single Gaussian variant, the transition probability for a_{12} and a_{23} should be $\frac{1}{40} = 0.025$, since the transition between processes in the signal occurs at every 40'th sample. The self-transition probability of a_{33} should be 1, since a transition to another state should not occur as the constructed signal is not periodic.
A 3 state 1-Gaussian component HMM is used to model the signal. The model converges after 4 iterations. The estimated model parameters are presented below:

$$\mathbf{A} = \begin{bmatrix} 0.975 & 0.025 & 0.000 \\ 0.000 & 0.975 & 0.025 \\ 0.000 & 0.000 & 1.000 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 5 & 24 & 48 \end{bmatrix}$$

$$\sigma = \begin{bmatrix} 3.1623 & 3.1623 & 3.1623 \end{bmatrix}$$

$$\Pi = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

with the following differences from the *expected* parameters:

$$\begin{aligned}
\mathbf{A}_\Delta &= \begin{bmatrix} -.0009 \cdot 10^{-12} & 0.0009 \cdot 10^{-12} & 0.000 \cdot 10^{-12} \\ 0.000 \cdot 10^{-12} & -.1434 \cdot 10^{-12} & 0.1434 \cdot 10^{-12} \\ 0.000 \cdot 10^{-12} & 0.000 \cdot 10^{-12} & 0.000 \cdot 10^{-12} \end{bmatrix} \\
\mu_\Delta &= \begin{bmatrix} -0.7961 \cdot 10^{-5} & 0.7959 \cdot 10^{-5} & 0.0001 \cdot 10^{-5} \end{bmatrix} \\
\sigma_\Delta &= \begin{bmatrix} 0.2391 \cdot 10^{-4} & 0.2392 \cdot 10^{-4} & 0.0001 \cdot 10^{-4} \end{bmatrix} \\
\Pi_\Delta &= \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}
\end{aligned}$$

The estimated parameters correspond exactly to the expected values with only a small difference in numerical precision between the true and estimated parameters.

6.6.2 Modeling 2D Artificial Signal with a Random Component

A clearly sectioned two dimensional signal with some randomness in the respective processes is created. Like the one dimensional artificial signal, the two dimensional artificial signal is constructed with three underlying processes each having two sub processes. Each dimension of the signal is presented in Figure 6.6.

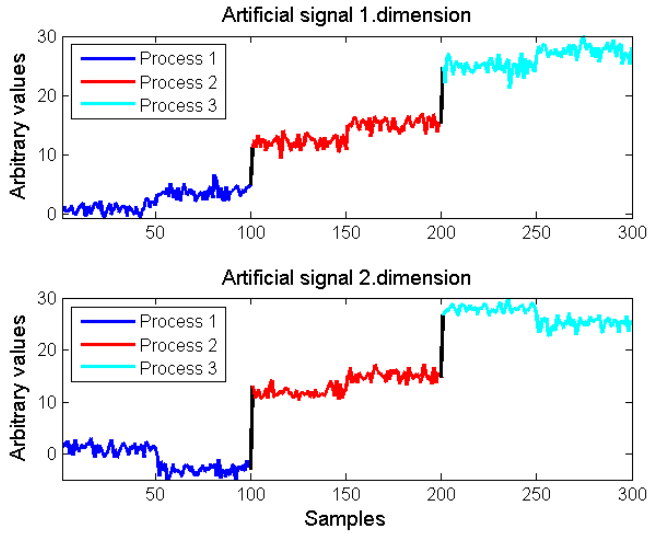


Figure 6.6: Two dimensional artificial signal with random component. The three colors indicate three underlying processes.

The two dimensions are plotted against each other and presented in Figure 6.7.

To visually verify that HMM captures the latent processes of the two dimensional artificial signal, it is modeled both with a 3 state 1-Gaussian component HMM and a 3 state 2-Gaussian component HMM. The results are shown as contour plots. The result for the 3 state 1-Gaussian component HMM is visualized in Figure 6.8 and the result of the 3 state 2-Gaussian component HMM is shown in Figure 6.9. These plots can be compared against the actual data presented in Figure 6.7.

Figure 6.8 indicates how the 3 state 1-Gaussian component HMM very satisfactorily captures the three processes when comparing with Figure 6.7. The 3 state 2-Gaussian component HMM uses it's extra flexibility to capture the sub processes, which is evident when comparing Figure 6.7 and Figure 6.9.

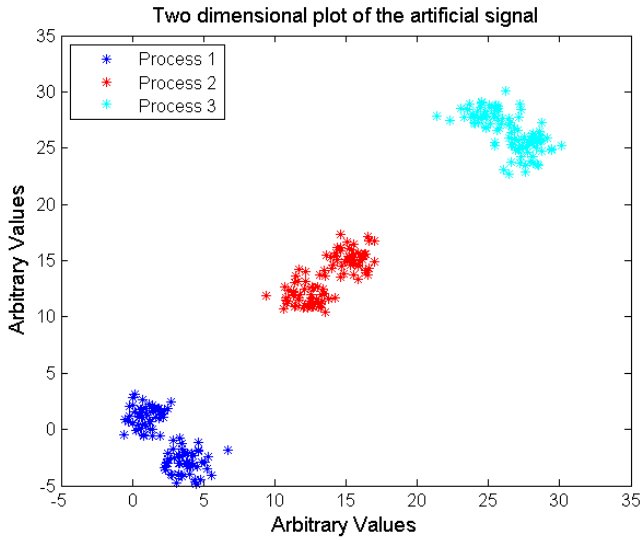


Figure 6.7: Values of the two dimensional artificial signal plotted against each other. The clearly defined clusters are color coded in order to indicate the underlying processes specified in Figure 6.6. Visual identification of the subclusters, corresponding to the sub processes, is straightforward.

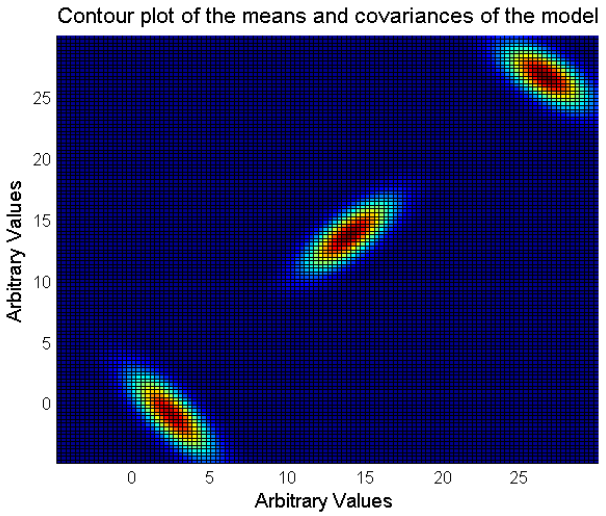


Figure 6.8: Contour plot of the means and covariances of the 3 state 1-Gaussian component HMM of the two dimensional artificial signal shown in Figure 6.6 and Figure 6.7.

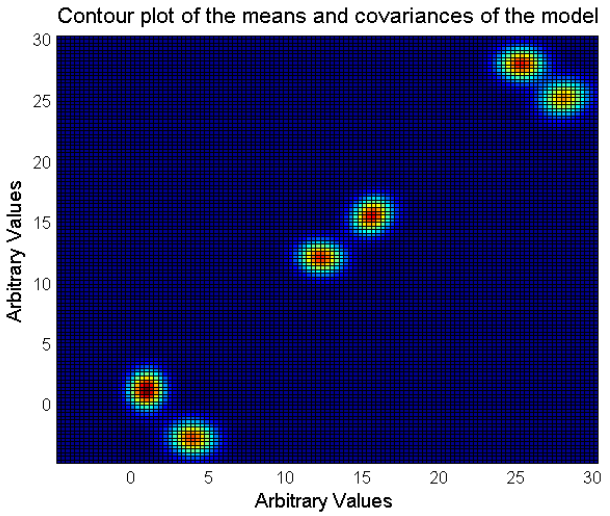


Figure 6.9: Contour plot of the means and covariances of the 3 state 2-Gaussian component HMM of the two dimensional artificial signal shown in Figure 6.6 and Figure 6.7.

6.6.3 Capturing the Dynamics of an ECG using Optimal State Path

The previous sections established that the HMM is capable of modeling a clearly sectioned signal. To verify that the model has the ability to capture the characteristics of an ECG signal, a normal ECG is modeled with different numbers of states and a 1-Gaussian component. Using the Viterbi algorithm to calculate the optimal state path given the trained HMM, it should be possible to recreate the modeled ECG. The results of applying 8, 25 and 50 states are presented in Figure 6.10 where the means of the states have been plotted. Not surprisingly the model with 50 states recreates the ECG better than the 8 state model. However, the basic ECG structure is recognizable in all of the recreations.

6.6.4 Approximate a HMM of a Process by Modeling Simulation from a Teacher Model

In the following a 4 state 1-Gaussian component HMM is created. Then the simulation procedure is used to generate a signal of varying length. The model is called the *teacher*. The parameters of the teacher used are:

$$\mathbf{A} = \begin{bmatrix} 0.80 & 0.20 & 0.00 & 0.00 \\ 0.00 & 0.70 & 0.30 & 0.00 \\ 0.00 & 0.00 & 0.75 & 0.25 \\ 0.40 & 0.00 & 0.00 & 0.60 \end{bmatrix}$$
$$\mu = \begin{bmatrix} 10 & 20 & 30 & 40 \end{bmatrix}$$
$$\sigma = \begin{bmatrix} 1 & 2 & 1 & 2 \end{bmatrix}$$
$$\Pi = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

Defining the transition matrix in this way makes it a Left-Right model with periodicity. Using a Left-Right model starting in state one, avoids the problem with possible permuted states.
The teacher model is used to generate signals increasing in length. Starting at a length of 50 samples, a HMM is trained at signal lengths defined by increments of 10 from the initial length of 50 samples, i.e. the HMM is trained on generated signals of length 50,60,70 up to 2000 samples. For each model based on the generated signal, the summed difference between the various HMM parameters were calculated together with the total difference. The results are presented in Figure 6.11 and shows how the total sum of differences between the teacher model and learned model decrease inconsistently at first and later more steadily

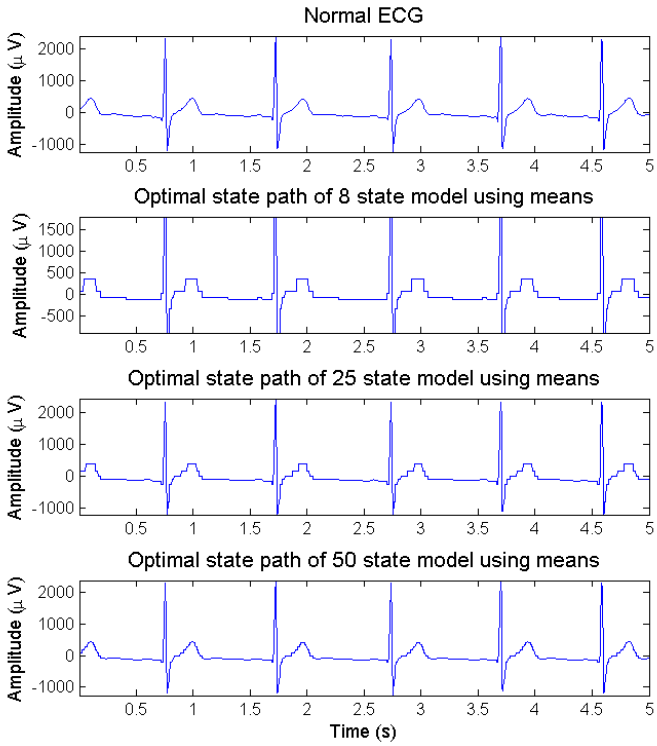


Figure 6.10: Top plot shows five seconds of a normal ECG. The remaining plot shows the optimal state sequence given the ECG of the top plot and a HMM trained on that ECG. Plots 2-4 from the top corresponds to a HMM trained with 8, 25 and 50 states, respectively. For comparison, the state means are added.

towards zero.

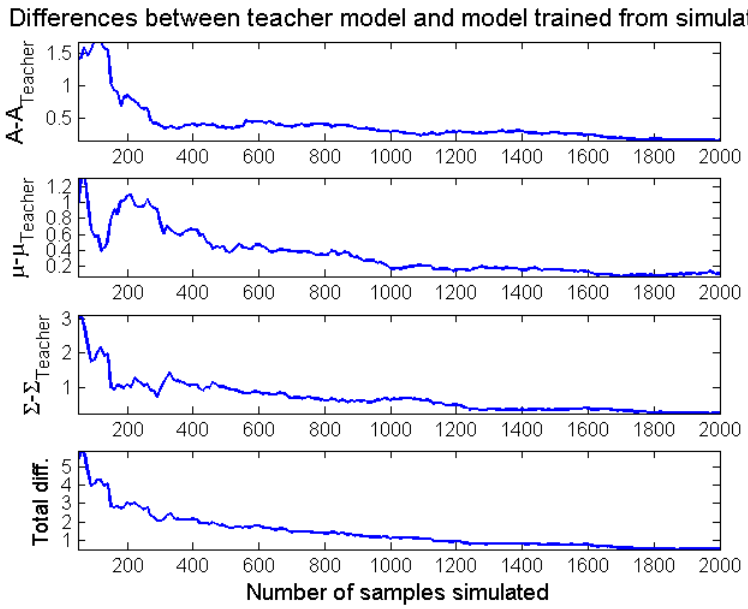


Figure 6.11: Four plots showing the summed difference between a teacher HMM and the learned HMM of a simulated signal at varying length. From the top: 1. Difference between transition matrices. 2. Difference between means. 3. Difference between covariance matrices. 4. Total sum of differences of all HMM parameters.

6.7 The Effect of ECG Noise on Classification

Section 3.4 presented five types of noise typically found in the ECG. Having a serious chronic illness might produce uncomfortable associations with the process of having an ECG recorded. As such, it is not unreasonable to think that the LQT2 ECGs might be more prone to a higher biological noise source content than the normal ECGs. The worst case scenario is that an artificial difference between the two study populations is introduced this way. This was the rationale behind the baseline wander filtration and preferably further filtration should have been performed.
It was chosen to assess the effect of noise on the classification by adding noise to

the test data used in the classification; i.e. noise was not added to the training data but only to the test data. A fused noise source were created by adding the 3 biological noise sources and PLI. The rms amplitude ratios between the three biological noise sources were preserved as they appear in the original noise data. As the impairment of the ECG caused by PLI, compared to e.g. BW at the same SNR, is far more devastating, the PLI rms amplitude in the fused noise source was considered. By visual inspection of the noisy ECGs it was chosen to define the rms amplitude of the PLI as half of the lowest rms amplitude of the three biological noise sources. This method resulted in a fused noise source whose component wise rms amplitude distribution was 30%, 16%, 46% and 8% for BW, MA, EM and PLI, respectively. Subsequently the fused signal was multiplied with a constant factor, calculated for each ECG to which noise was applied, in order to match a predefined level of noise in dB (SNR) as described in section 3.4.2. Besides the fusion of the noise signals, the individual and random noise samplings for each lead, were performed and applied as described in section 3.4.2. The BW filtration was performed after the addition of noise to the *unfiltered* test ECGs.

Chapter 7

Results of Model Identification and Classification Applied to ECGs

The current chapter presents the results of applying the methods described in chapter 6 to the normal and LQT2 ECGs. Section 7.1 presents the mean accuracies of the trained models and the most extreme ECGs identified by the best models, according to the mean classification accuracy. Section 7.2 presents the generative properties in terms of simulations and finally section 7.3 presents the results of classifying normal and LQT2 ECGs.

7.1 Basic Discriminative Properties

Following the procedure of model training described in section 6.2 the classification accuracy was evaluated for different parameter settings in the models. Results are summarized and presented in Figure 7.1, showing the mean classification accuracy of the four crossfolds for an increasing number of states. Results are presented for the three types of transition matrices all applying both one and two Gaussians, respectively. The blue line corresponds to the LR type transition

with one forward degree of freedom, the red line represents the corresponding two forward degrees of freedom and the magenta line shows the full type transition. Punctuation refers to the use of two Gaussians in the model. None of the models produce a steady course through the increasing number of states, however, there seems to be a common tendency of a decrease in classification accuracy from 40 to 45 states and a common increase between from 25 to 30 states. Comparing the mean accuracy curves representing the application of two Gaussians, with the corresponding curves for a single Gaussian, two Gaussians seem to produce a lower or similar mean classification accuracy except at states 40, 45 and 50 for the LR types. The mean accuracies corresponding to the full type transition is more ambiguous and shows no clear tendency with regards to the use of one or two Gaussians.

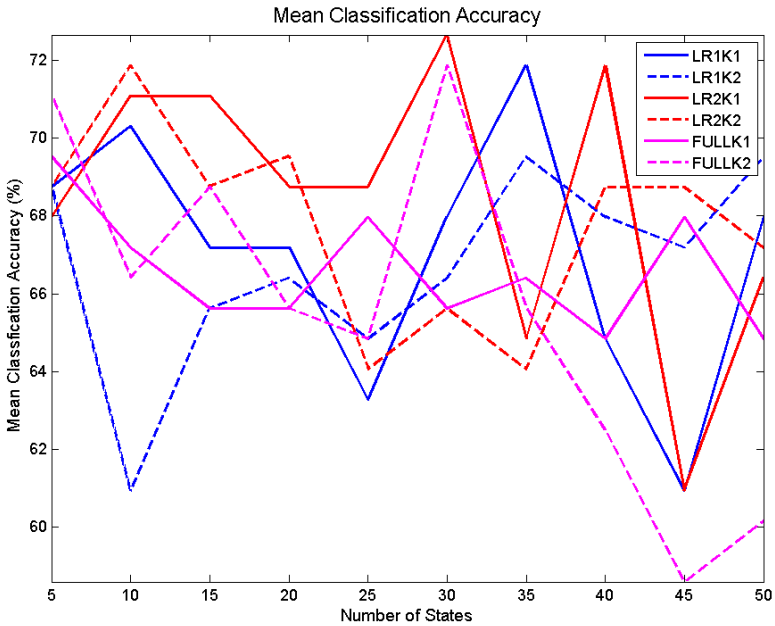


Figure 7.1: Mean classification accuracy for the models corresponding to the three types of transition matrices. The blue line corresponds to the LR type transition with one forward degree of freedom, the red line represents the corresponding two forward degrees of freedom and the magenta line shows the full type transition. Punctuation refers to the use of two Gaussians in the model.

Following the procedure introduced in section 6.5 the extreme case ECGs, identified by their likelihoods, are identified and plotted for inspection. The next six pages presents these extreme case ECGs for each type of model. First, the LR type transition with one forward degree of freedom, one Gaussian and 35 states are presented in Figure 7.2 for crossfold 1. The top part of the figure presents the likelihood for the test set given the normal model plotted against the likelihood for the test set given the LQT2 model. The line through the diagonal represents equality in likelihoods. Blue asterisks denote normal subjects whereas red asterisks denote LQT2 subjects. The green triangles indicate the most likely subject with each model and the black circles represent the maximal probability ratios for the case where the normal model is more likely than the LQT2 model and vice versa. The bottom part shows the most likely ECGs with both models in the top subfigure, corresponding to the green triangle. ECGs corresponding to the black circle are presented in the lower subfigure of the bottom part. The setup remains the same for the following figures and will not be addressed further.

Considering Figure 7.2 it appears that the most likely ECGs is the same normal subject for both models. The likelihood ratios capture a LQT2 ECG whose lead V5 is strongly corrupted with noise (remaining leads are more normal). Note that the 2D probability is based on all leads whereas only lead V5 is presented for inspection. The corresponding normal ECG seems to be similar to the most likely ECG. Figure 7.3 presents the LR type, one degree of freedom and two Gaussians (crossfold 4). Most likely ECGs are captured from each group and seem to have similar P-waves and QRS-complexes. The T-wave in the LQT2, however, seems to be wider, flatter and noisier. The maximum likelihood ratios seem to capture the same differences in two ECGs. Figure 7.4 (LR type, two degrees of freedom , one Gaussian and crossfold 3) captures the same normal ECGs as being the most likely with both models. The ECGs selected with the likelihood ratios show a normal looking LQT2 ECG and a normal ECG with a remarkably higher heart rate. Figure 7.5 presents the corresponding 2 Gaussian case, which also had crossfold 3 as the best, with regards to classification accuracy. Thus the same test data are plotted, using different models of course, but the selected ECGs are the same. Figure 7.6 presents the full type transition (crossfold 2). The most likely ECG is again the same normal subject. ECGs identified with the most extreme likelihood ratio show a smooth normal with a relatively wide T-wave. The LQT2 ECG is of lower amplitude, far more noisy and has a wider T-wave. The corresponding 2 Gaussian case (crossfold 2) is presented in Figure 7.7 where the most likely ECG is again the same normal subject for both models. ECGs identified with the most extreme likelihood ratio show a normal ECG of generally lower amplitude than the LQT2 ECG. The normal T-wave looks a bit noisy whereas the LQT2 T-wave is smooth and far wider than the normal T-wave.

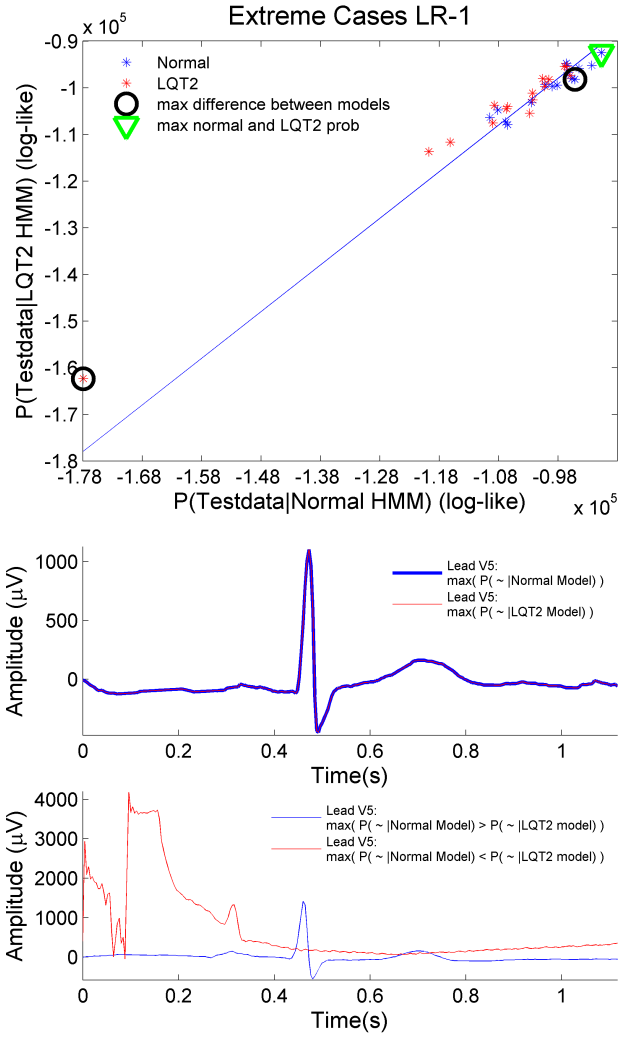


Figure 7.2: [35 states, 1K, crossfold 1] The top part of the figure presents the likelihood for the test given the normal and LQT2 model where the diagonal represents equality in likelihoods. Blue and red asterisks denote normal and LQT2 subjects, respectively. Green triangles indicate the most likely subject with each model and the black circle represents the maximal probability ratios for the case where the normal model is more likely than the LQT2 model and vice versa. The bottom part shows the ECGs corresponding to the green triangle(s) in the top subfigure and the black circles in the bottom subfigure.

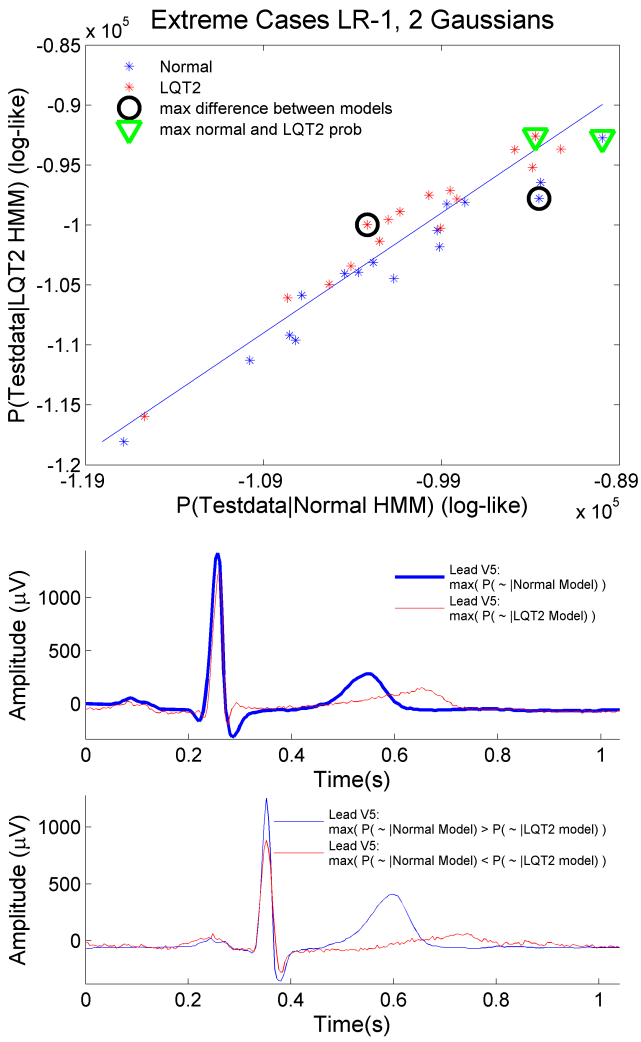


Figure 7.3: [35 states, 2K, crossfold 4] The top part of the figure presents the likelihood for the test given the normal and LQT2 model where the diagonal represents equality in likelihoods. Blue and red asterisks denote normal and LQT2 subjects, respectively. Green triangles indicate the most likely subject with each model and the black circle represents the maximal probability ratios for the case where the normal model is more likely than the LQT2 model and vice versa. The bottom part shows the ECGs corresponding to the green triangle(s) in the top subfigure and the black circles in the bottom subfigure.

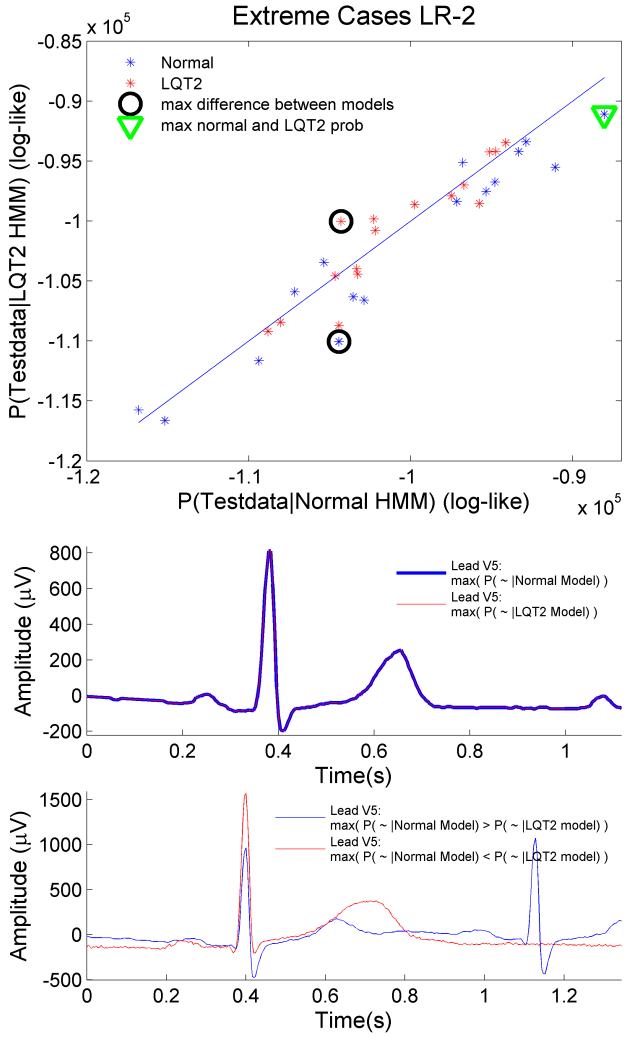


Figure 7.4: [30 states, 1K, crossfold 3] The top part of the figure presents the likelihood for the test given the normal and LQT2 model where the diagonal represents equality in likelihoods. Blue and red asterisks denote normal and LQT2 subjects, respectively. Green triangles indicate the most likely subject with each model and the black circle represents the maximal probability ratios for the case where the normal model is more likely than the LQT2 model and vice versa. The bottom part shows the ECGs corresponding to the green triangle(s) in the top subfigure and the black circles in the bottom subfigure.

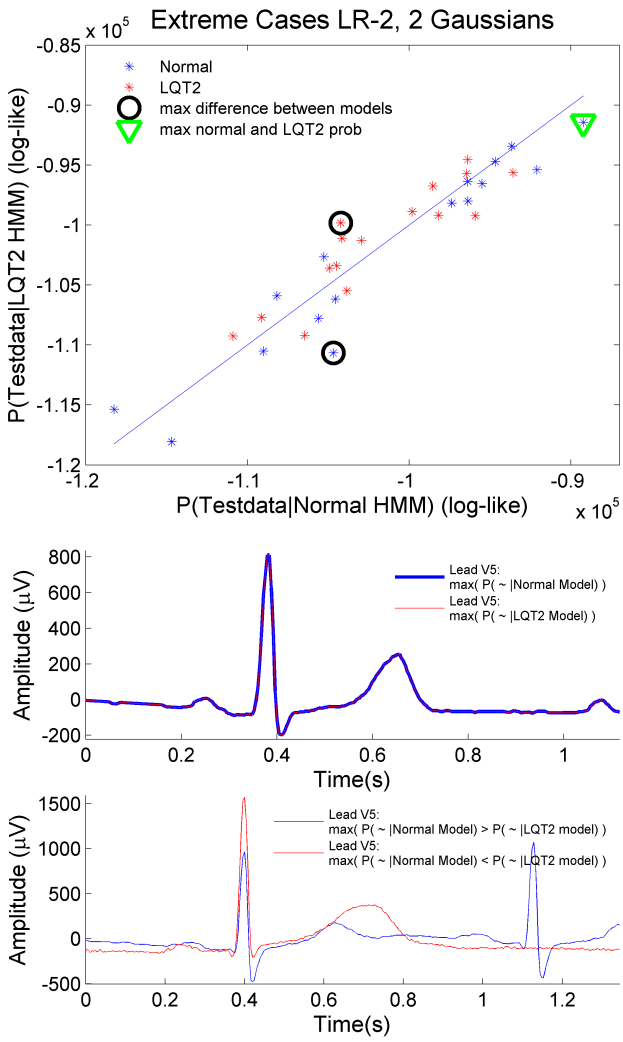


Figure 7.5: [10 states, 2K, crossfold 3] The top part of the figure presents the likelihood for the test given the normal and LQT2 model where the diagonal represents equality in likelihoods. Blue and red asterisks denote normal and LQT2 subjects, respectively. Green triangles indicate the most likely subject with each model and the black circle represents the maximal probability ratios for the case where the normal model is more likely than the LQT2 model and vice versa. The bottom part shows the ECGs corresponding to the green triangle(s) in the top subfigure and the black circles in the bottom subfigure.

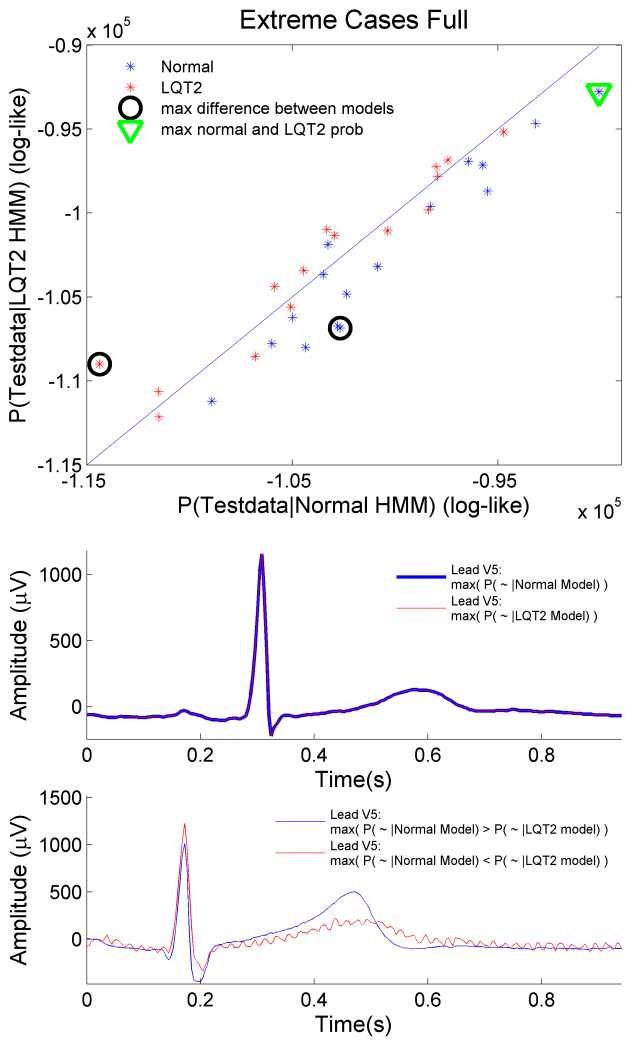


Figure 7.6: [5 states, 1K, crossfold 2] The top part of the figure presents the likelihood for the test given the normal and LQT2 model where the diagonal represents equality in likelihoods. Blue and red asterisks denote normal and LQT2 subjects, respectively. Green triangles indicate the most likely subject with each model and the black circle represents the maximal probability ratios for the case where the normal model is more likely than the LQT2 model and vice versa. The bottom part shows the ECGs corresponding to the green triangle(s) in the top subfigure and the black circles in the bottom subfigure.

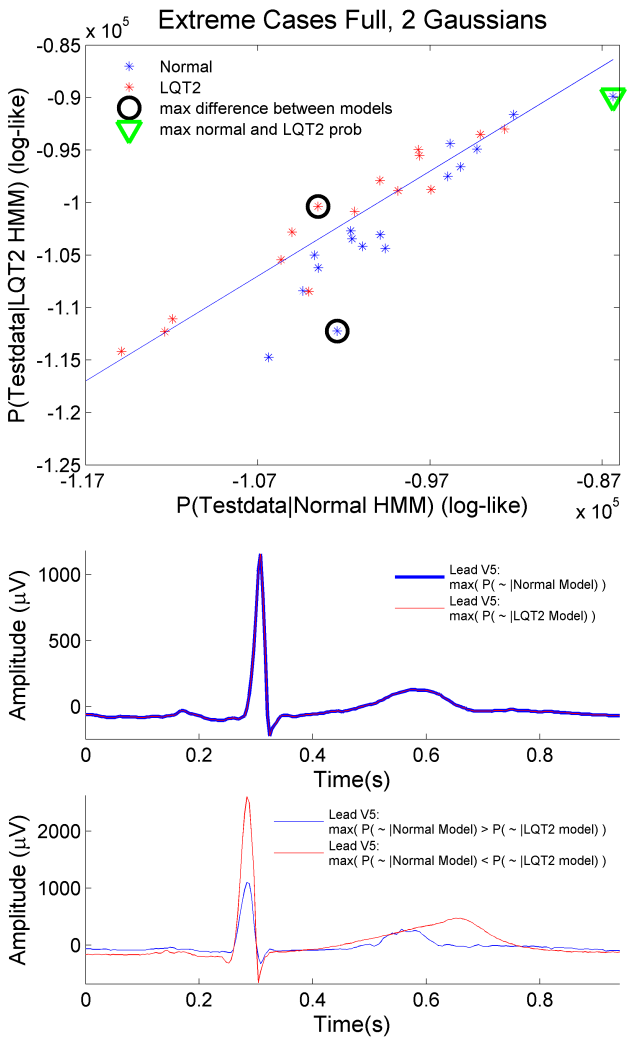


Figure 7.7: [30 states, 2K, crossfold 2] The top part of the figure presents the likelihood for the test given the normal and LQT2 model where the diagonal represents equality in likelihoods. Blue and red asterisks denote normal and LQT2 subjects, respectively. Green triangles indicate the most likely subject with each model and the black circle represents the maximal probability ratios for the case where the normal model is more likely than the LQT2 model and vice versa. The bottom part shows the ECGs corresponding to the green triangle(s) in the top subfigure and the black circles in the bottom subfigure.

7.2 Generative Properties

Following the procedure of simulation described in section 6.4.1 the mean emissions are simulated while constricting the number of self-transitions to expectation. Doing so, all 8 leads of an ECG are simulated with the best single Gaussians models, based on classification accuracy, for the normal and LQT2 group. Figure 7.8 presents the LR type transition with one forward degree of freedom. The best model consisted of 35 states and 1 Gaussian. One cycle, or "heartbeat", is shown for each lead with lead I and II in row one and leads V1, V2, V3, V4, V5 and V6 in row 2 through 4. All leads are presented in individual subfigures and for comparison the normal (blue line) and LQT2 (red line) ECG simulations are aligned by visual inspection. Comparing with the biological normal ECG presented in Figure 3.3 it seems that the P-wave, QRS-complex and T-wave are recognizable in some form although some additional excursions are present for both the normal and the LQT2 ECG simulations. The larger negative component of the biological QRS complex in some leads (e.g. lead V2 and V3 in Figure 3.3) also seem to be present in the simulation. Considering the excursions in the QRS-complex region in both groups they seem to be similar in their shape and amplitude except for lead V3 and lead V4 where the amplitude of the normal ECG simulations are larger. The baseline and P-wave section preceding the QRS-complex is stable with smaller excursions in the normal ECG simulation and none of the leads exhibit any distinct excursions that could be attributed to the model capturing the dynamics of the P-wave. The LQT2 ECG simulation however, has an excursion in the area of the P-wave that is distinct in most of the leads. The baseline section preceding these excursions, also seem to show a larger variability between mean values than in the normal simulation. The section of the simulation following the QRS-complex region where the T-wave is found in the biological ECGs shows distinct excursion in both groups. Comparing the groups the LQT2 excursions are the same or lower amplitude than in the normal ECG simulation. Furthermore, there is a tendency of the excursions to initiate earlier than with the normal simulation. Depending on the level at which the return to the baseline is defined, the LQT2 excursions in the T-wave region also seem to be wider in the LQT2 group.
Figure 7.9 presents an ECG simulation applying the LR type transition with two forward degrees of freedom. The best model, according to classification accuracy, consisted of 30 states and 1 Gaussian. Due to the randomness in the non self-transitions an interlead variability will be present, and so two consecutive beats are shown. Comparing with the biological normal ECG presented in Figure 3.3 again, the biological ECG features are not recognizable to the same degree as with the simulation in Figure 7.8. The most distinct excursions are seen at times 0.25 s and 0.55 s in both the normal and the LQT2 simulation and have similarities with the QRS-complex region of the biological ECGs. Except for lead V2 and lead V3 the LQT2 ECG simulations have larger amplitudes.

Considering the QRS-complex region further, the normal simulations shows flat areas in all the leads which is not observed in the LQT2 simulation. Baseline regions are stable in both cases and excursions related to the P-wave are not distinct in the LQT2 ECG simulation. In the normal case the flat part of the simulation in the QRS-complex region covers the expected area of the P-wave. The areas following the QRS-complex regions show distinct excursions immediately after the QRS-complex region. The differences in this region between the normal and LQT2 ECG simulation are not remarkable but the lower amplitude and wider excursions seen in Figure 7.8 with LQT2 in the T-wave region are observable on a far smaller scale in Figure 7.9 in lead I, II, V5 and V6.
Figure 7.10 presents an ECG simulation applying the full transition type. The best model, according to classification accuracy, consisted of 5 states and 1 Gaussian. The randomness in the non self-transitions has a far larger influence yielding the aperiodic appearance of the ECG simulation. Comparing the normal and LQT2 ECG simulation they both seem irregular and of lower amplitude than the LR types and the most distinct difference, if any, seems to be the larger amplitude seen in the LQT2 ECG simulations. In the LR type transition simulations the highest amplitude spikes appeared as being related to the QRS-complex region of the biological ECG. Considering these spikes in Figure 7.10 as related to the QRS-complex, no systematic excursions preceding or following this region, that could be related to the P-wave or T-wave, are seen in either the normal or the LQT2 ECG simulation.

To generalize, the trained HMMs assume stationarity in the ECG and map all observations corresponding to a given state into a density with mean and variance. As one group could potentially have a larger inter-subject variability in some waves, it is also of interest to inspect the SD of the means as presented in row two of Figure 6.3. As the clarity of these properties, by visual inspection, is poor, only lead V5 of the normal and LQT2 model are presented in Figure 7.11, for the three types of transitions matrices. Figure 7.11 shows the simulated mean emissions, while taking expectation of self-transitions, and their corresponding SDs. The left column represents the normal model and the right column represents the LQT2 model. The top row corresponds to the LR type transition with one forward degree of freedom, the second row presents the corresponding two forward degrees of freedom and row three presents the full type transition. Considering the top row (LR1) quantification of any differences is complex. However, the T-wave region in the normal simulation seems to have a larger part (longer duration) with higher SD than seen in the LQT2 case where the SD seems to decrease towards the end of the T-wave region. In the middle row (LR2) the oddly shaped P-Q region in the normal case obviously introduces a difference, however this beat type is not observed every second time as the figure might indicate. It suggests, however, a larger variability in this region than in the LQT2 case. In contradiction, within the QRS complex region, the LQT2 simulation seems to possess a slightly higher SD of the mean emission

which is present in a wider region. The full type transition in the bottom is difficult to quantify. The SDs of the normal and LQT2 simulations are comparable except the in the first 0.3 seconds where the LQT2 SD is larger. Due to the randomness of the simulations however, this might not be the general case.

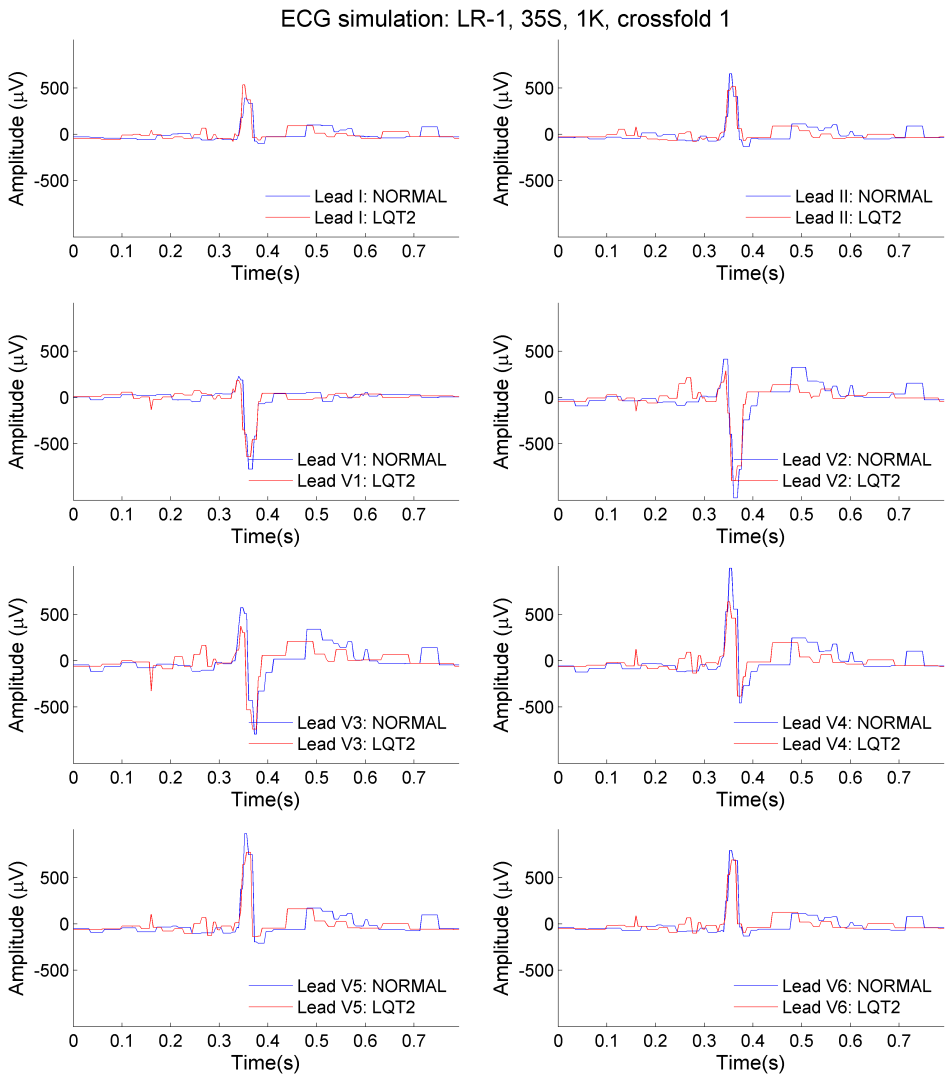


Figure 7.8: ECG simulation corresponding to LR type transition with one forward degree of freedom (35 states, 1 Gaussian). One cycle, or "heartbeat", is shown for each lead with lead I and II in row one and leads V1, V2, V3, V4, V5 and V6 in row 2 to 4. All leads are presented in individual subfigures and for comparison the normal (blue line) and LQT2 (red line) ECGs are aligned by visual inspection.

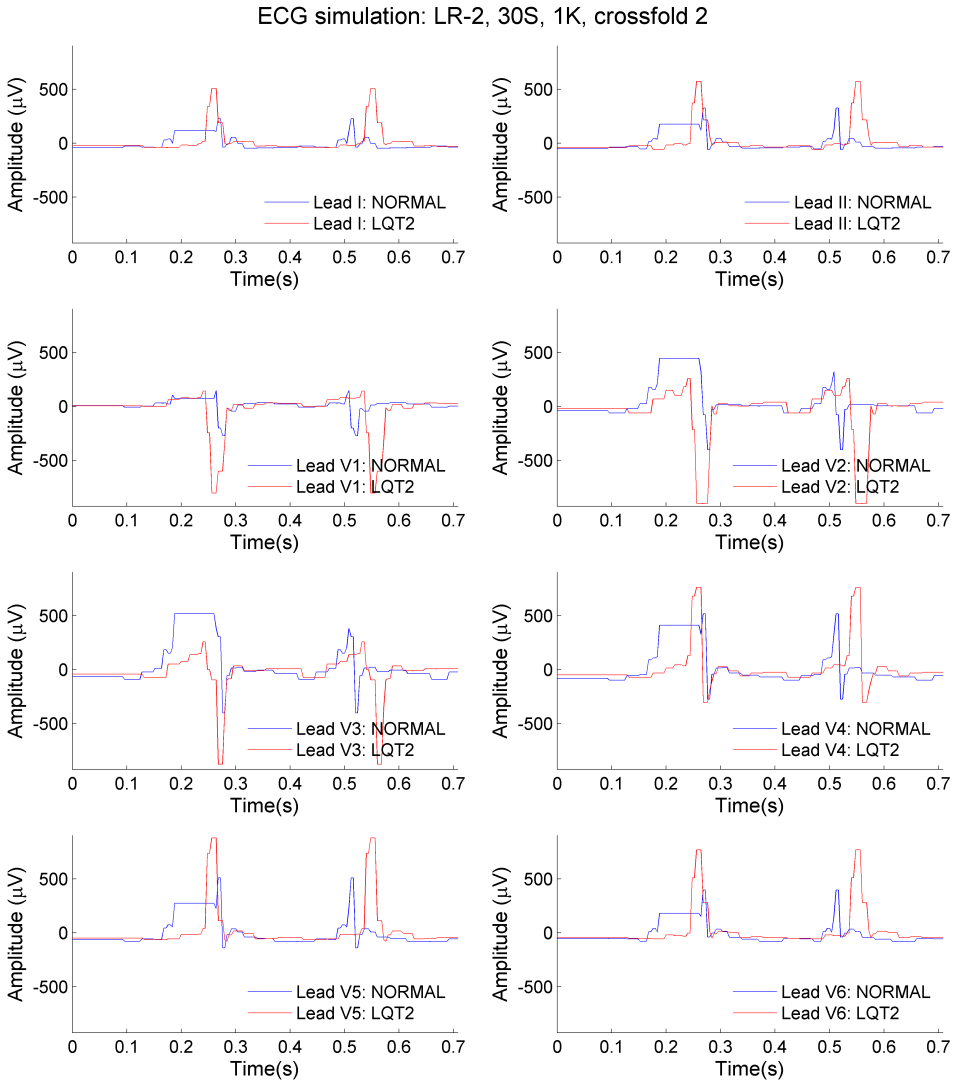


Figure 7.9: ECG simulation corresponding to LR type transition with two forward degree of freedom (30 states, 1 Gaussian). One cycle, or "heartbeat", is shown for each lead with lead I and II in row one and leads V1, V2, V3, V4, V5 and V6 in row 2 to 4. All leads are presented in individual subfigures and for comparison the normal (blue line) and LQT2 (red line) ECGs are aligned by visual inspection.

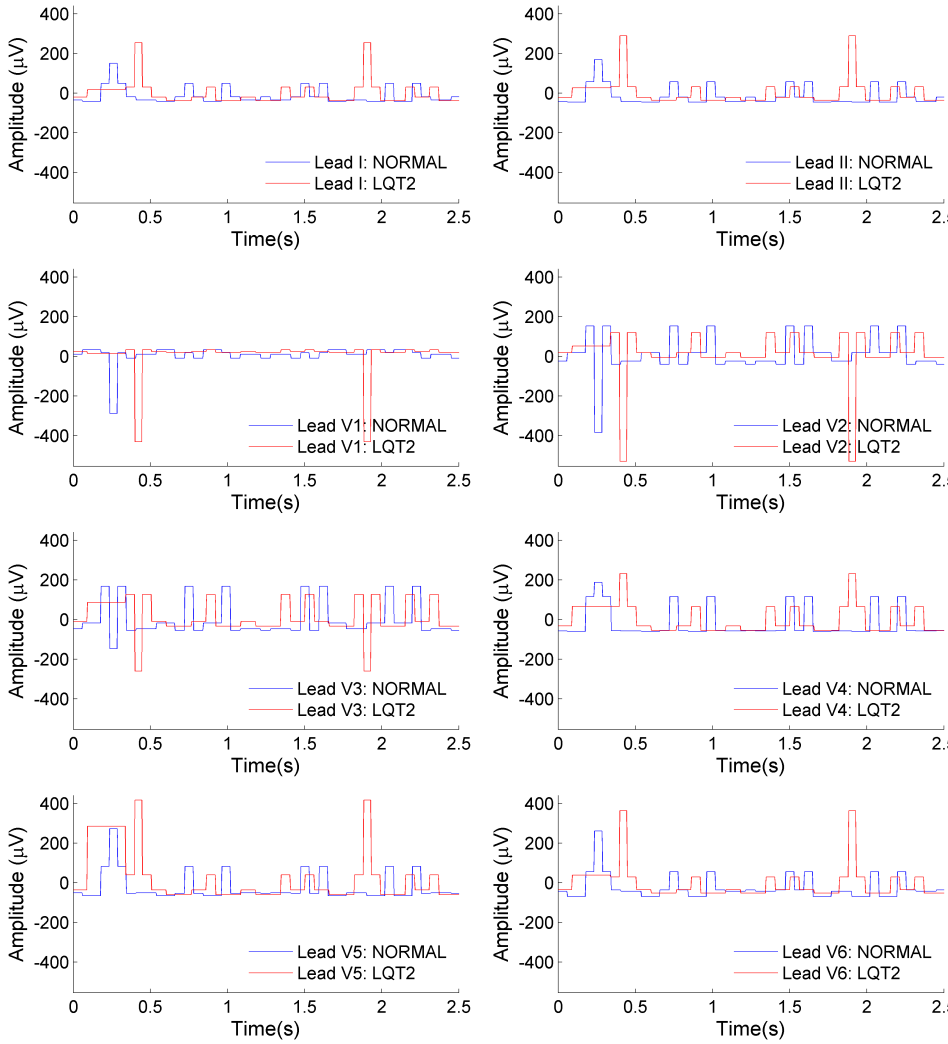


Figure 7.10: ECG simulation corresponding to full type transition (5 states, 1 Gaussian). One cycle, or "heartbeat", is shown for each lead with lead I and II in row one and leads V1, V2, V3, V4, V5 and V6 in row 2 to 4. All leads are presented in individual subfigures and for comparison the normal (blue line) and LQT2 (red line) ECGs are aligned by visual inspection.

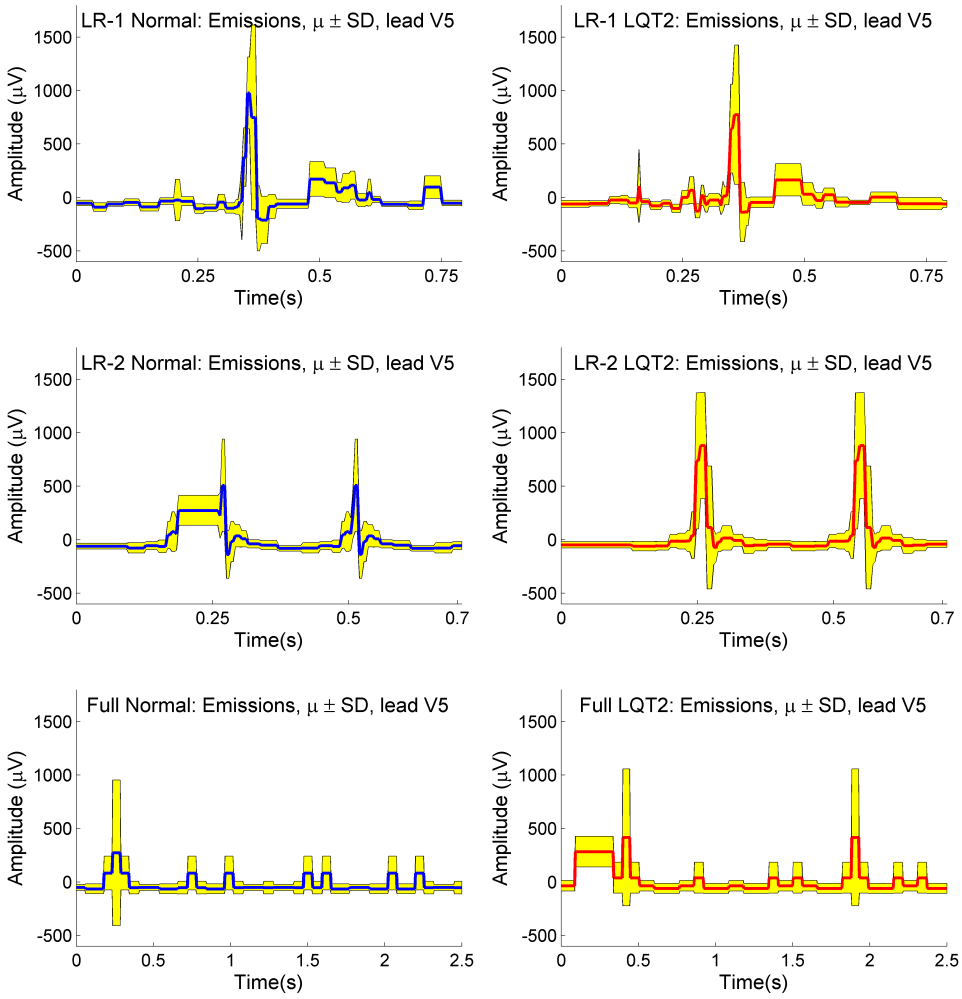


Figure 7.11: Simulated mean emissions, while taking expectation of self-transitions, and their corresponding SDs. The left column represents the normal model and the right column represents the LQT2 model. The top row corresponds to the LR type transition with one forward degree of freedom, second row presents the corresponding two forward degrees of freedom and row three presents the full type transition.

Comparing Figure 7.8 and 7.9 it appears as if the "heart rate" of the simulation doubles in the case with two forward degrees of freedom. Following the procedure described in section 6.4.2 the "heart rate" of the LR type transitions are calculated and presented in Figure 7.12. The top plot shows the "heart rate" of the LR type with one forward degree of freedom and the bottom plot shows the corresponding two forward degrees of freedom. The "heart rate" is calculated for all sizes of the transitions matrix (states) and for all crossfolds for both the normal (blue line) and LQT2 (red line) model. The best models, according to classification accuracy, are marked. Furthermore, the true heart rate of the biological ECGs are marked at the optimal number of states for comparison. Comparing the top and bottom plot, the "heart rate" of the one forward degree of freedom type is closer to the true heart rate of the data. The "heart rate" of the two forward degree of freedom model is high as suggested by Figure 7.9.

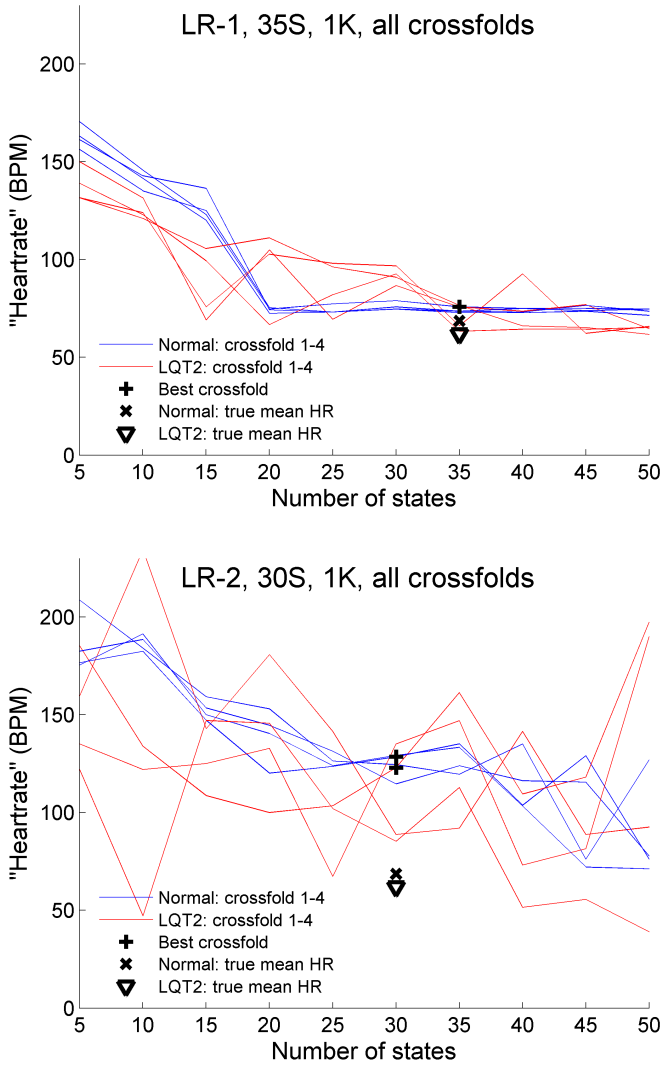


Figure 7.12: The top plot shows the "heart rate" of the LR type transition matrix with one forward degree of freedom and the bottom plot shows the corresponding two forward degrees of freedom calculated for all model sizes and crossfolds for the normal (blue line) and LQT2 (red line) model. The black plus marks the optimal models. The true heart rate of the data is marked with a cross (normal) and triangle (LQT2).

7.3 Classification

The of results using the different classification schemes described in section 6.3 are presented in the following.

The accuracy, specificity, sensitivity and area under the receiver operating characteristic curve (AUC), are calculated for each of the models with the highest accuracy and shown in Table 7.1. The models presented in this section is presented in a short notation form, where e.g. $LR1K1_{35}$, is a strict Left-Right model using 35 states and 1 Gaussian per state.

	Accuracy	*Specificity*	*Sensitivity*	*AUC*
$LR2K1_{30}$	72.6%	68.4%	79.9%	0.70
$FULLK2_{30}$	71.9%	68.4%	79.9%	0.72
$LR1K1_{35}$	71.9%	69.9%	79.4%	0.72
$LR2K2_{10}$	71.9%	71.5%	73.8%	0.72
$FULLK1_5$	70.3%	68,0%	76.5%	0.68
$LR1K2_{35}$	69.5%	67.7%	75.8%	0.70

Table 7.1: The optimal number of states for each of type of transition matrix models. The highest score for each measure is highlighted.

Due to the variance and small number of cross folds, no statistical significance can be established between the models using a significance level of 5%. Evaluating the table by simply looking at the magnitudes, the model using a 30 state Left-Right transition structure with an extra degree of freedom and states modeled with one Gaussian component seems to be the best choice. It has the highest accuracy and sensitivity yielding 72,6% and 79.9%, respectively. Weighting the AUC higher than the accuracy, the $FULLK2_{30}$ and $LR1K1_{35}$ would be good candidates, but in general is it difficult to distinguish between the models.

To visually compare the models the mean ROC curves are shown in Figure 7.13. The course of the mean ROC curves looks very much alike. Only $LR2K1_{30}$ and $LR1K1_{35}$ seem to positively stand out around a false positive rate of 0.3 and 0.6, respectively.

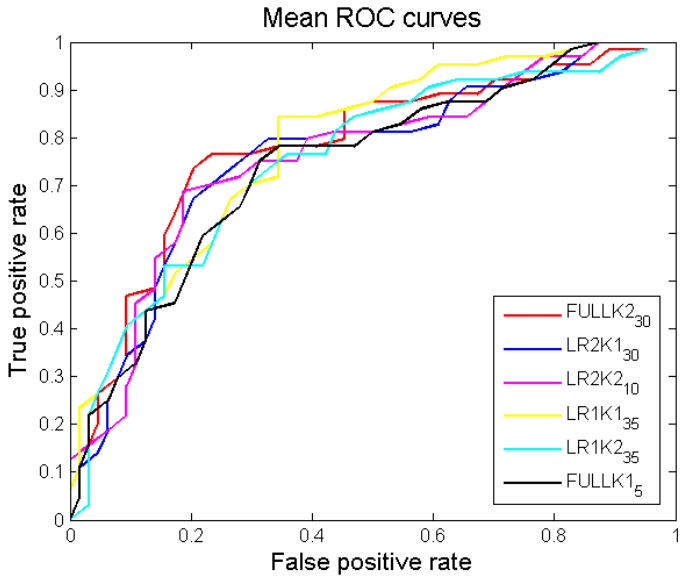


Figure 7.13: Mean ROC curves for the models with the highest accuracy for each combination of models from Table 7.1. The course of the mean ROC curves look very much alike. Only $LR2K1_{30}$ and $LR1K1_{35}$ seems to positively stand out around a false positive rate of 0.3 and 0.6, respectively

.

7.3.1 Combining Best Of Each Model Types

Classification using combinations of the six models from Table 7.1 with the different classification schemes is performed. The range of the resulting accuracies due to the different classification schemes are shown in Table 7.2.

	Range of Accuracy
$HMM_{Biggest\Delta}$	$71.1 - 71.9\%$
HMM_{Voting}	$71.9 - 75.0\%$
SVM_{Linear}	$72.7 - 75.8\%$
SVM_{poly2}	$71.1 - 74.2\%$
SVM_{poly3}	$72.7 - 73.4\%$
SVM_{poly4}	$71.1 - 75.0\%$
SVM_{rbf}	$71.9 - 74.2\%$
$SVM_{quadratic}$	$71.9 - 74.2\%$
SVM_{mlp}	$72.7 - 75.0\%$

Table 7.2: The Table shows the range of the resulting accuracy using the combination of models from Table 7.1 with different classification schemes.

The model with the highest accuracy from Table 7.2 is a combination of $LR2K1_{30}$ and $LR1K1_{35}$ using SVM without a kernel (linear kernel). This yields an accuracy of 75.8%.

7.3.2 Combining Models Having the Best Accuracies

The models having the six highest accuracies (see Figure 7.1) disregarding model type, were combined in different ways, using the different classification schemes. The results are seen in Table 7.3.

Based on the highest accuracy the best classification is performed doing the following: Calculate the log likelihood differences between the normal and LQT2 models corresponding to the LR2K1 models consisting of 10, 30 and 40 states, respectively, and apply them as features to an SVM using a multi-layer perceptron as kernel. From here on this procedure is named the $BestModel$. A mean confusion matrix for the $BestModel$ is seen in Table 7.4. The model is equally good at predicting LQT2 and Normal ECGs.

	Range of Accuracy
$HMM_{Biggest\Delta}$	70.3 – 73.4%
HMM_{Voting}	72.7 – 75.0%
SVM_{Linear}	73.4 – 77.3%
SVM_{poly2}	74.2 – 75.8%
SVM_{poly3}	73.4 – 76.6%
SVM_{poly4}	72.7 – 75.8%
SVM_{rbf}	71.9 – 75.8%
$SVM_{quadratic}$	74.2 – 76.8%
SVM_{mlp}	75.0 – 78.1%

Table 7.3: The table shows the classification accuracies of the models having the six highest accuracies (see Figure 7.1) disregarding model type while using the different classification schemes.

	Predicted as LQT2	Predicted as Normal
LQT2	12.5	3.5
Normal	3.5	12.5

Table 7.4: A mean confusion matrix for the *BestModel*.

The accuracy, specificity and sensitivity of the *BestModel* is presented in Table 7.5.

	Accuracy	*Specificity*	*Sensitivity*
BestModel	78.1%	78.2%	78.2%

Table 7.5: The accuracy, specificity and sensitivity of the model combination yielding the highest accuracy. The transition matrix is a Left-Right structure able to make a transition two states ahead. Each state is modeled using one Gaussian component.

Comparing the single best model from Table 7.1 and *BestModel* the accuracy increases with 5.5%, the sensitivity decreases 1.7% while specificity increases 9.8%.

7.3.3 Combining QTcB and Best Model

As described in section 2.4 the QTcB is the golden standard for classifying LQT2 patients. Applying the SVM with different kernels using QTcB yields the classification accuracy presented in Table 7.6.

	$QTcB$
SVM_{Linear}	81.3%
SVM_{poly2}	82.0%
SVM_{poly3}	82.0%
SVM_{poly4}	79.7%
SVM_{rbf}	81.3%
$SVM_{quadratic}$	82.0%
SVM_{mlp}	81.3%

Table 7.6: The classification accuracy using QTcB in an SVM with seven different kernels.

The accuracy found by combining the QTcB and the $BestModel$ is seen in Table 7.7.

	$QTcB + BestModel$
SVM_{mlp}	84.4%

Table 7.7: The accuracy using the QTcB feature combined with the best found combination of HMM's using an SVM with a multi-layer perceptron kernel.

Table 7.7 indicates that the $BestModel$ contributes to the classification accuracy with an increase of 2.4% compared to using only QTcB.

7.3.4 Effect of Noise on the Classification

To investigate the effect of noise, the noise application procedure described in section 6.7 is applied. The effect of noise on the accuracy, sensitivity and specificity of the *BestModel* using this procedure is presented in Figure 7.14. At SNR levels of 40 and 35 dB the noise does not influence accuracy, sensitivity nor specificity. Around 30 dB a small decrease in all three measures is observed. At 20 dB the accuracy reaches 71.9%. The sensitivity increases and the specificity decreases. At 15 and 10 dB the accuracy is quickly decreasing together with the specificity, while the sensitivity is increasing.

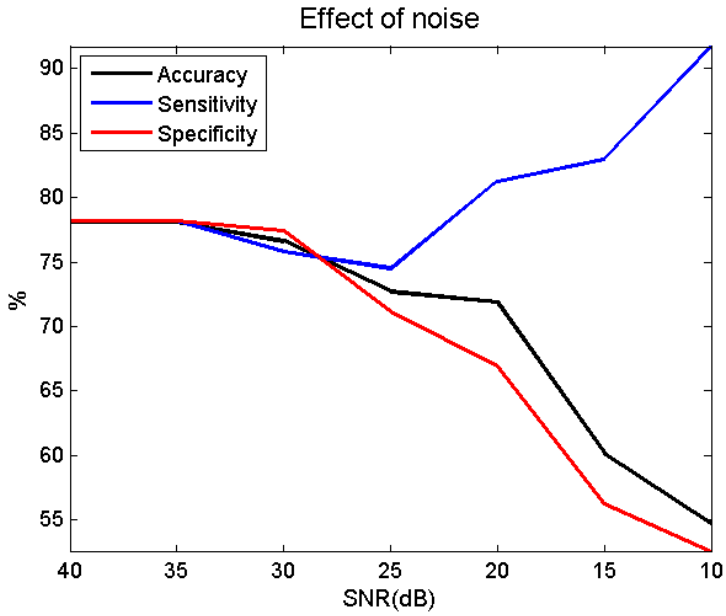


Figure 7.14: The effect of noise on the accuracy, sensitivity and specificity of the *BestModel*.

Chapter 8

Discussion

In the current work 8-lead ECGs from normal and LQT2 subjects have been modeled using a multi-dimensional continuous emission Hidden Markov model. Both the generative and discriminative properties of the models were investigated. The possibility of discriminating the data were further investigated through different classification schemes including the Support Vector Machines. The current chapter discusses the results obtained, starting with the choice of HMMs in section 8.1. The training and testing of the models are discussed in section 8.2 and the generative properties of the resulting models are discussed in section 8.3. The classification and the effect of noise on these results are discussed in sections 8.4 and 8.5, respectively.

8.1 Choice of Model

Motivation: As the introduction presented, the primary motivation was to create a method of characterizing and discriminating ECGs in general. Existing methods of quantifying ECGs seems to rely on single beat characterization. This could either be the median beat as in MUSE® (12SL® algortihm) or abnormal beat recognition in 24h Holter recordings. In the current work only LQT2 and normal ECGs were treated. Comparing the ECGs from the normal

and LQT2 subjects revealed relevant properties in terms of the appearances of the ECGs; the LQT2 ECGs show a large inter subject variability and more importantly a subset of the LQT2 ECGs resembles the normal ECGs, which is the typical situation in biosignal classification. It is well-known that the LQT2 is associated with abnormal T-waves, but with regards to the QRS-complex and P-wave no particular abnormalities are generally accepted. Applying gender and age matched study populations should provide a good test of a potentially more global method in that the general trends captured in each group should preferably show some meaningful differences. It was chosen to evaluate these differences both with regards to the classification itself but also with regards to the generative properties of the models. It was learned from the literature review that HMMs are often applied in ECG segmentation.

The choice of HMMs: Generally speaking the HMM is a stochastic model in which the hidden states correspond to the underlying stochastic process that cannot be observed; only the emissions associated with each hidden state are observable. The hidden Markov modeling approach hence combines both statistical and structural knowledge. However, it was learned that the structural properties that is characterized during model training assumes stationarity in the ECG; i.e. considering e.g. a state corresponding to some part of the T-wave, the estimated mean value of the emission density will be an expression of the general amplitude of this region, both inter-beat and inter-subject wise. In the study [39] by Thoraval et al. was it argued that the assumption of inter-beat stationarity is only marginally fulfilled. However, these arguments were proposed in a segmentation context. In the current work the aim was to capture general trends that characterizes the groups which could lay the basis of discrimination. In that context however, the assumption of stationarity might be appropriate. Information contained in the inter-beat variations might be lost, but considering that HMMs are not trained for each subject, these inter-beat variations would probably be hidden in the larger inter-subject variations anyway. In summary the HMM approach can model the signal directly, it can include inter-subject variability and considering the HMM as a classification method it provides probabilities rather than binary outputs.

Evaluation: In order to evaluate the HMM approach both the discriminative and generative abilities were evaluated. Depending on the criteria of success, the modeling approach does not need to fulfill both areas but preferably the models showing the best classification abilities should contain some generative properties making the classification method more reliable. As the discriminative and generative abilities are discussed in other sections only the segmentation and time warping issue is considered in the remaining. One of the selection criteria was that the modeling approach should be able to deal with unsegmented ECGs of varying heart rates. As the HMM approach was able to model 10 s ECG segments and capture the general trends in the semi periodic signals and,

at the same time, explain inter-subject variability, it seems that the application of HMMs using raw ECG data was well suited for the purpose of ECG characterization and discrimination.

8.2 Training and Testing the Hidden Markov Models

Hidden state transition structures: Having explored the literature covering ECG modeling with HMMs, it was found that the transition structures often applied were the fully connected and the strict Left-Right types (LR1). In this work a Left-Right transition matrix with an extra degree of freedom was also included (LR2). As the model achieving the highest accuracy exclusively used combinations of the models having this LR2 transition structure with a different number of states, it would be of interest to apply transition structures with increasing flexibility in the Left-Right setting. Having a strict Left-Right model with a sufficient number of states to represent a full heartbeat in one cycle (around 30-35 states) and one Gaussian per state, the model forces the specific parts of the heartbeat to be mapped into emissions of a single state, thus modeling the variations through the variance of a Gaussian. Consider a population with almost identical ECGs, with four sub populations each having a P-wave of different amplitude. A strict Left-Right model would find the mean and model the differences of the P-wave amplitude though a large variance. The mean value of the P-wave captured by the model could be a value that would actually not be present in that population. Providing a model with multiple degrees of freedom, the model would be able to capture the four specific P-wave amplitude subpopulations, by making it possible to skip states in the transition matrix when the states representing the top of the lower amplitude P-wave are encountered. Since the remaining parts of the ECGs are identical in this scenario, the following transitions of the model could adapt to a strict Left-Right structure.

Alternative transition structures: Having experienced that a number of 30 to 35 states seems to be able to model an entire heart beat realistically, it could be of interest to build a transition matrix incorporating two Left-Right structures with periodicity that were only linked in one or two states. This could potentially capture two different heartbeat variations present in a population of ECGs. An example could be normal heartbeats and premature ventricular beats. Furthermore, knowledge about the diseases that are modeled could be incorporated in the transition matrix through specific transition structures, which has shown to improve modeling in certain areas such as protein transportation. However, since the purpose was to evaluate HMMs in general ECG characteri-

zation and discrimination, this was not pursued.

The number of cross-folds: In order estimate the mean accuracy of the implemented models a 4-fold cross validation was performed. The number of folds were solely chosen based on an estimation of the training process computation time. Going through the machine learning literature the number of cross-folds applied ranges between 5 and 20. Had the number of cross-folds been increased, a more accurate estimate of the accuracy could perhaps have been found. Further this would facilitate the establishment of statistical significance between models (if any). However, given the limited time of the project and the speed of the implementation this was not feasible.

Covariance structure: In the current work the covariance between all 8 leads of the ECGs were modeled. It was included in order to capture any class related covariance between the leads which could potentially improve the discrimination between the normal and the LQT2 ECGs. However, estimating a full covariance implies estimating numerous parameters thus making it more prone to overfitting. One way to relax the model complexity would be to use only a diagonal covariance matrix, ignoring the covariance between leads. Doing so it could have been investigated whether the covariance actually contributed to the discriminative capabilities.

8.3 Generative Properties of the Hidden Markov Model

Simulations: The key issue with the HMM approach is the question of what general trends are actually captured in each group. As the trained models should be able to represent both inter-beat and inter-subject variations it is not straight forward to anticipate which ECG structures are captured. Consulting section 7.2 the three types of transition models, all applying only one Gaussian, were used to simulate ECGs. The best models with the LR types, according to the mean classification accuracy, both had a relatively high number of states (35 and 30). The best model with the full transition type, however, only had 5 states. Without considering the actual simulations it is not anticipated that the full type transition will be able to capture the same general trends as the LR types due to the low number of states. The LR1 type produced the most realistic looking simulation. Most of the anticipated ECG waveforms were recognizable and the amplitude and duration of the beat corresponded well to the actual data for both the normal and the LQT2 simulation. The LR2 case produced less rec-

ognizable waveforms and the duration of a beat seemed far shorter than in the data. Even though individual waveforms were not distinguishable to the same degree as in the LR1 case the notion of two beats in the presented time window seemed clear. As expected, the 5-state full transition model produced what looked like random patterns of which not even the peak amplitudes are comparable to that of the typical ECGs. Besides evaluating the fidelity of the simulations the differences between the groups was also of great interest. It is noted that it was chosen to consider the mean emissions and to perform the simulation using an expected number of self-transitions. In order to consider the variance of the mean emissions the SDs was also plotted (Figure 7.11). When comparing the captured trends between the normal and LQT2 group, both the course of the simulations in terms of the mean emissions as well as the captured variance should be considered. Generally speaking, the variances seem to be comparable whereas the amplitude of the excursions seem larger with the LQT2 simulations for the full transition and LR2 transition while being more comparable in the LR1 case. Hence, to generalize the visually most obvious difference in the simulations does not seem to be the captured variance or the general amplitude, but rather the structure of the simulation (excursions). Considering the LR1 case it seems reasonable to say that there appears to be a difference both in the P-wave area end T-wave area when comparing the normal and LQT2 model. However, observing abnormal T-waves in the LQT2 model is less clear. On the other hand, the wide and sometimes notched LQT2 T-waves in the data are often of low amplitude and maybe the simulation actually shows a wider T-wave. The most distinct difference though, would be the earlier onset of the T-wave in the LQT2 simulation. The LR2 case is more difficult to compare and also seems to exhibit some difference in the resulting "heart rate" between the normal and LQT2 simulations, besides generally being higher than in the LR1 case.

"Heart rate" of the simulations: As variable duration was not included explicitly in the HMM model, the inherent exponential duration of the states was effectively applied. This limitation might have affected the choice of model. Consulting the "heart rate" plots (Figure 7.12) it seems that only the LR1 type captured the true mean heart rate whereas the LR2 simulation seems to represent an almost doubled heart rate. It should also be noted that the best models as according to the classification accuracy seem to represent an intersection of the normal cross-folds' "heart rates" and the LQT2 cross-folds' "heart rates" (over the increasing states), which could imply that the best discrimination is achieved when the "heart rates" of the models representing each group are relatively similar, which corresponds well to the actual situation in the data. The "heart rate" plots shown for the LR types along with the simulations for all three types of transition suggest that the LR1 type seems to be the HMM type showing the best generative properties.

Extreme log-likelihood ECGs: By calculating the probability of the ECGs

given the HMMs the most likely and the most extreme ECGs were investigated. These ECGs does not present actual generative properties of the models, but they were included to aid in the evaluation of what trends that were actually captured. Following comments hence relates to Figures 7.2 - 7.7. To generalize, the most likely ECG was almost without exception the same (normal) ECG for both the normal and the LQT2 model. Depending on the crossfold and model in question this normal ECG changed but in all cases except with LR1, 2 Gaussians, a normal looking normal ECG was selected as the most likely for both models. This confirms the anticipation that some overlap occurs between the groups in that a normal looking normal ECG results in a high probability with both models. The largest differences in log-likelihoods however, were different from model to model. In most cases the LQT2 ECG would be of higher amplitude, have a wider T-wave and contain more noise. It should be noted however, that even though the mean heart rate of the populations are similar, the heart rate from ECG to ECG varies significantly. Take for example both LR2, 1 Gaussian and LR2, 2 Gaussians (Figures 7.4 and 7.5); the normal ECG seems to represent a higher heart rate than observed with the LQT2 ECG in both cases. The LR1 results seem to capture the expected typical normal and typical LQT2 better in the 2 Gaussian case but unfortunately the 1 Gaussian case captures an extremely noisy LQT2. The full transition seems to result in normal and LQT2 ECGs of similar heart rate and the T-wave is wider in the LQT2 ECG in both cases.

Evaluation: Considering the small number of test subjects and the fact that the differences in performance, as according to the mean classification accuracy, often differ with only a single subject, the difference in performance between LR1 and LR2 does not seem to be that large when consulting Figure 7.1. Hence, the type of model that provides the most realistic simulations, represents the true heart rate of the data well, results in reasonable extreme ECGs and has proven its basic discriminative capabilities is the LR type with one degree of freedom.

8.4 Discriminative Properties of the Hidden Markov Model

Classification techniques: The general approach when using the HMM as a classifier is building a HMM per class and subsequently evaluating the probability of a new test signal for each of the HMMs. The test signal is classified according to the HMM yielding the highest probability. The same approach was applied in this work. Furthermore, in utilizing multiple models two different classification schemes was applied. The biggest difference method, where

the models with the highest log-likelihood difference ruled and a voting scheme were the classification was based on the majority of votes. In general, utilizing multiple models increased the mean classification accuracy. It was found that the voting scheme was a better choice than relying on the biggest difference. However, none of these classification schemes achieved as high a mean accuracy as using the output from the HMMs in a SVM in the form of a log-likelihood difference between the normal and LQT2 models. Using a linear kernel and a multi-layer perceptron kernel yielded the highest mean accuracies of 77.3% and 78.1%, respectively. It was found that ECGs having a high probability of being generated from the normal HMM also had high probability being generated by the LQT2 HMM and vice versa. The same seemed to be the case for low probabilities. A possible explanation of why the SVM improved the accuracy is its ability to find the maximum margin decision boundary. It could be argued that the LR2 transition type using 1 Gaussian maybe is slightly better in capturing the important differences between the normal and LQT2 ECGs. Each of the models using different states could possible capture important features from sub groups in the normal and LQT2 population. Applying the SVM using multiple models, the possible enhanced separation of log-likelihood relationships between the models in the specific domains can be utilized. It must be stressed that no evidence for this hypothesis has been found and that the argumentation is solely based on speculation.

Best classification: The best model found in this project, based on the classification accuracy of normal and LQT2 ECGs, was found using three log-likelihood differences between the models of the normal and LQT2 classes, using a Left-Right transition structure (LR2) with 10, 30 and 40 states, respectively, and applying them as features to a SVM using a multi-layer perceptron kernel. The model yielded a mean accuracy of 78.1%. Comparing with literature the result seems satisfactory; a study using a HMM based diagnosis system for myocardial infarction classification applying multiple lead ECGs achieved accuracies ranging from 71%-83% [51].
Using the combined output of multiple HMMs as input to the SVM increase the model complexity and thereby the risk of overfitting. Considering the small test set and the statistically insignificant differences between some model combinations a less complex model combination could have been chosen. Applying the principle of Ockham's razor, the model using only the best LR1 and LR2 HMMs in combination with a SVM with a linear kernel, yielding a mean accuracy of 75.8%, could be a good candidate. The mean accuracy is still satisfactory, but the complexity is reduced and thereby possibly giving a better generalization. However, this thesis pursued the model yielding the highest mean accuracy.

Choice of the optimal number of states and classification: Due to the parameter learning of the HMM, the probability of the training data (say class 1) is maximized, without considering other classes. This is a way to accomplish

a general representation of training data, given that the HMM has a sufficient complexity. However, from a classification point of view, it is desirable to maximize the difference in probability between the models, given an ECG, to improve generalization. In this work, the best discriminative model was determined by exploring the mean accuracy of many different models while adjusting the number of states, the number of Gaussians per state and the transition structure in the model. Other ways of improving the classification accuracy of the HMM have been proposed in the literature; e.g. incorporating a more discriminative approach in the HMM training by combining the HMM and SVM framework [66]. In this work the training consisted of finding the best model representation of the data while minimizing the classification error.

Evaluation: An excellent way to establish a reliance of the accuracy of a machine learning system is to evaluate the chosen model, on a "hidden" test data set, which has never been used throughout the modeling process. However, biological and medical data are often sparse. As the general idea behind a learning system is that the data should be representative for the class in question, the removing of data from the training environment could degrade the system performance. In this work only 64 LQT2 subjects were available and it was decided to utilize all of the data in the modeling process.

8.5 The Effect of Noise on the Generative and Discriminative Properties

ECG noise: A thorough investigation of the most common types of ECG noise was performed in the current work. At the initial phase it was assumed that the ECG data, that were extracted from the XML files in the MUSE® database, were preprocessed similarly to the ECG shown on the physical cart during the ECG recording. It turned out, however, that the data were completely raw. Due to the time limits of the project the most straight forward task with respect to filtering, with the smallest risk of removing valuable information from the signals, seemed to be highpass filtering. Especially it seemed that with regards to the hidden states and their emissions, the low frequency and sometimes high amplitude BW, could be particularly troublesome. BW could easily be the initial part of any of the ECG waveforms, hence influencing the probability of an ECG given a model. EM and MA however, are more difficult to filter without removing valuable information in the signal. Unfortunately, also PLI was present in small amounts which should preferable have been filtered.

Noise vs. general trends: The general issue of concern is not as much the filtering as it is the difference in the signal quality that seems to characterize the two study populations. Very generally speaking the HMM will try to capture a general trend within each group as this structure will result in the highest probability. The model finds a structure that fit inter-beat and inter-subject variability well and here the noise might have a devastating effect. If one group is more noisy this would appear in the total group variability. As such, there is a risk that the general trend captured by the HMM modeling normal ECGs is more smooth and have a smaller variance than that of the LQT2 HMM. However, it is stressed that despite this negative effect, it only explains part of the differences in the general trends that were captured.

Influence on classification: Consulting the simulations the LR1 model simulated differently shaped T-waves in the normal and LQT2 group. Also, the most extreme ECGs chosen from data seemed to be differently shaped with regards to the T-wave all implying that some of the known physiological trends in the groups were actually captured. However, the extreme ECGs also seemed to show more noisy LQT2 ECGs suggesting that both the noise and the general trends were captured. Consulting the influence of noise on the best model this also reveals itself. Noise of an SNR between 30 and 25 can be added in which all three measures of the classification decrease slightly. This implies a generally negative effect, but apparently the general trends in the groups which are not related to noise are still distinct enough to provide means of classification. When the noise level increase beyond this point, however, it seems that all ECGs moves towards being classified as LQT2; the sensitivity increases at the cost of decreasing specificity and accuracy. From a patient safety point of view this is not the worst situation, since an LQT2 diagnosis would lead to more elaborate tests, revealing a higher number of LQT2 subjects. However, in the false positive case, the mental stress and inconvenience of being subject to further clinical examinations should not be taken lightly. From a health care economic perspective, the cost of unnecessary examinations would not be welcome.
In summary, experience showed that the LQT2 group is more noisy and that this, in part, is captured by the HMM model. The task of evaluating how much of the classification accuracy, can be attributed to difference in noise between the groups, is very complex. Certainly, less noisy data or better filtering prior to the model training would provide both a better idea of the discriminative and generative abilities of the suggested method as well as a final product that is less sensitive to noise.

8.6 Perspectives and Future Work

One of the key advantages of the HMM is its general capability of modeling temporal signals. It is not dependent on any segmentation algorithms and it can model any ECG even if the usual characteristics are not visible. Also, signals degraded by noise can be modeled, although it affects the resulting model.

With respect to the generative properties, the HMM should be improved by incorporating some sort of explicit state duration parameter. Besides improving the ECG simulation, multiple studies also report an increase in the classification accuracy [64], [7]. However, the cost of explicitly modeling the state duration is a considerable increase in computational load. However, we believe that there are many ways to improve the speed of the program, both by optimization the code further, by using parallel computing or by implementing the code in a lower level program language such as C++.

From a classification point of view, the training procedure of the HMM should also incorporate some sort of discriminative power in the optimization function. Having introduced the Support Vector Machine as an excellent discriminative model, it would be reasonable to try to incorporate a more discriminative approach in the HMM training by combining the HMM and SVM framework as proposed in [4] or [66].
Combining the output of the HMM with the currently applied stationary features at the Department of Biomedical Sciences in an optimized ensemble of machine learning methods could potentially also prove to be a strong classifier.

In this project the developed methods have been applied to ECGs from patients suffering from the well-studied LQT2 syndrome. Due to the general classification abilities of the method, it would be interesting to apply it to a variety of the different pathological ECGs where strong predictors have not yet been established.
In it's current form a clinical application could be the modeling of a large population of ECGs with the HMM; ECGs having very degraded, noisy or aberrant waveforms could be localized by evaluating their probability, given the model, and the ECGs with the lowest probability could be investigated.
From a research point of view, investigating the temporal variation and covariance of already recognized heart beat features such as T-wave amplitude and skewness, would be interesting and the developed HMM system could be an important tool in characterization as well as classification aspects in such a study.

Chapter 9

Conclusion

In this thesis six hidden Markov models using different transition structures and number of emission distributions were trained using raw ECG signals from normal and LQT2 subjects. The models were trained using a different number of hidden states and, based on the mean accuracy, the best number of states for each of the six HMMs were found. These hidden Markov models were able to capture the general trends in the ECGs and, at the same time, explain inter-subject variability. The strict Left-Right transition type showed promising generative properties which facilitated the observation of ECG waveforms that could be related to the ECG. Furthermore, the expected morphological changes in the T-wave were, to some degree, captured both in terms of the simulations and also in terms of the ECGs yielding the biggest difference in log-likelihood between the normal and LQT2 HMMs. However, some overlap between the groups resulted in several normal ECGs being most probable with both normal and LQT2 HMMs. Also, the Left-Right transition types were able to match the heart rate of the two study populations when simulating ECGs. The basic discriminative probabilities (applying only the log-likelihoods) resulted in classification accuracies ranging from 69.5% to 72.6% for the six HMMs. Applying the Support Vector Machine using different kernels and combining the models in several ways, improved the classification results. The best classification result observed was a classification accuracy of 78.1% with a corresponding specificity of 78.2% and a sensitivity of 78.2%.
Besides capturing general trends in the ECGs, making classification possible,

noise contained in the ECGs was also captured. Experience showed that the LQT2 group contained more noise, which affected the classification. Applying a substantial amount of noise to the test data prior to classification resulted in increasing sensitivity at the cost of drastically decreasing accuracy and specificity. Less noisy data or better filtering prior to the model training would definitely provide a better idea of the discriminative and generative abilities of the HMM approach. However, based on the results it seems that the application of HMMs using raw ECG data is well suited for the purpose of ECG characterization and discrimination.

Appendix A

Appendix

A.1 Lagrange Multiplier Method

The Lagrange multiplier method is used to solve constrained optimization problems [61]. The steps for solving an optimization problem where the goal is to find the minimum value of a function:

$$f(x_1, x_2.., x_d) \tag{A.1}$$

with an *equality constraint* of the form

$$g_i(x) = 0, \ i = 1, 2, ...p \tag{A.2}$$

is done analogously to the following three steps:

1. Construct the Lagrangian which is

$$L(x, \lambda) = f(x) + \sum_{i=1}^{p} \lambda_i g_i(x) \tag{A.3}$$

with λ_i being the Lagrange multipliers.

2. Set
$$\frac{\partial L}{\partial x_i} = 0 \tag{A.4}$$
for i = 1,2..d
and
$$\frac{\partial L}{\partial \lambda_i} = 0 \tag{A.5}$$
for i = 1,2..p.

3. Solve the above mentioned p+d equations to obtain stationary point x^* and λ_i.

Should the function have *inequality constraints* (e.g. $g_i(x) \leq 0$) the Lagrangian will have the following constraints known as the Karush-Kuhn-Tucker conditions:

$$\frac{\partial L}{\partial x_i} = 0, \qquad \forall_i = 1, 2.., d \tag{A.6}$$
$$g_i(x) \leq 0, \qquad \forall_i = 1, 2.., p \tag{A.7}$$
$$\lambda_i \geq 0, \qquad \forall_i = 1, 2.., p \tag{A.8}$$
$$\lambda_i g_i(x) = 0, \qquad \forall_i = 1, 2.., p \tag{A.9}$$

A.2 Flow Chart of Hidden Markov Model Implementation

A flow chart of the Hidden Markov Model implementation and initial classification setup, is shown in Figure A.1.

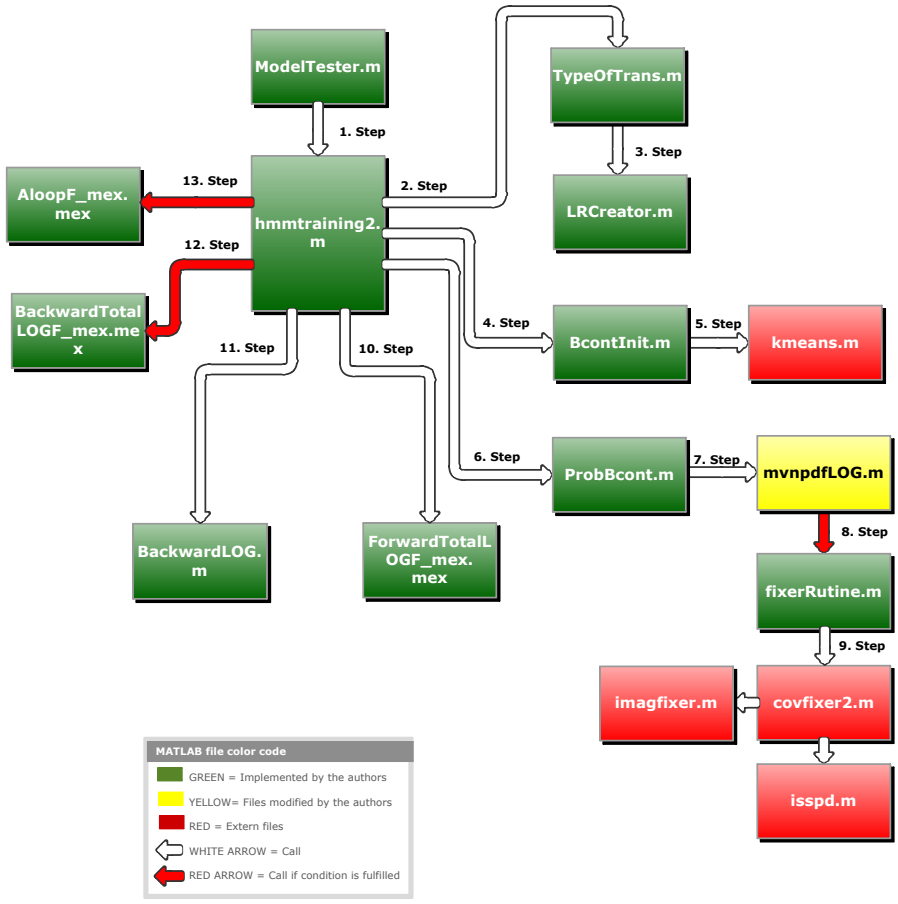


Figure A.1: Flow chart of Hidden Markov Model implementation. The green blocks are files implemented by the authors, the yellow are files modified by the authors and the red are files from a toolbox or implemented from others work.

Bibliography

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

[2] E. Almehdawe, B. Jewkes, and Q. He. A markovian queueing model for ambulance offload delays. *European Journal of Operational Research*, 2012.

[3] E. Alpaydin. *Introduction to machine learning*. MIT press, 2004.

[4] Yasemin Altun, Ioannis Tsochantaridis, Thomas Hofmann, et al. Hidden markov support vector machines. In *Machine Learning-International Workshop then Conference*, volume 20, page 3, 2003.

[5] Andre ao R.V., Dorizzi B., and Boudy J. Ecg signal analysis through hidden markov models. *IEEE Transactions on Biomedical Engineering*, 53(8):1541–1549, 2006.

[6] Jonson B., Westling H., White T., and Vollmer P. *Klinisk fysiologi*. Gads Forlag, 2002.

[7] Abdallah Benouareth, Abdel Ennaji, and Mokhtar Sellami. Semi-continuous hmms with explicit state duration for unconstrained arabic word modeling and recognition. *Pattern Recognition Letters*, 29(12):1742–1752, 2008.

[8] C.M. Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[9] Koeppen B.M. and Stanton B.A. *Berne and Levy Physiology*. MOSBY, Elsevier, 2010.

[10] Graff C., Struijk J.J. Matz J., Kanters J.K., Andersen M.P., Nielsen J., and Toft E. Covariate analysis of qtc and t-wave morphology: New possibilities in the evaluation of drugs that affect cardiac repolarization. *Clinical pharmacology & Therapeutics*, 88(1):88–94, 2010.

[11] Graff C., Andersen M.P., Xue J.Q., Hardahl T.B., Kanters J.K., Toft E., Christiansen M., Jensen H.K., and Struijk J.J. Identifying drug-induced repolarization abnormalities from distinct ecg patterns in congenital long qt syndrome. a study of sotalol effects on t-wave morphology. *Drug Safety*, 32(7):599–611, 2012.

[12] Haarmark C., Graff C.and Andersen M.P., Hardahl T., Struijk J.J., Toft Egon, Xue J., Rowlandson G.I., Hansen P.R, and Kanters J.K. Reference values of electrocardiogram repolarization variables in a healthy population. *Journal of Electrocardiology*, 43:31–39, 2010.

[13] CEUfast. 03-heart-inherent-electri.jpg. http://www.ceufast.com/courses/148/03-Heart-Inherent-Electri.jpg, 2013.

[14] WT Cheng and KL Chan. Classification of electrocardiogram using hidden markov models. In *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, pages 143–146. IEEE, 1998.

[15] Douglas A Coast, Richard M Stern, Gerald G Cano, and Stanley A Briller. An approach to cardiac arrhythmia analysis using hidden markov models. *Biomedical Engineering, IEEE Transactions on*, 37(9):826–836, 1990.

[16] A. Corduneanu and C.M. Bishop. Variational bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA, 2001.

[17] Geselowitz D.B. Dipole theory in electrocardiography. *The American Journal of Cardiography*, 14:301–306, 1964.

[18] Valeria De Fonzo, Filippo Aluffi-Pentini, and Valerio Parisi. Hidden markov models in bioinformatics. *Current Bioinformatics*, 2(1):49–61, 2007.

[19] B. Direito, C. Teixeira, B. Ribeiro, M. Castelo-Branco, F. Sales, and A. Dourado. Modeling epileptic brain states using eeg spectral analysis and topographic mapping. *Journal of Neuroscience Methods*, 2012.

[20] Brugada R. (Ed.). *Clinical Approach to Sudden Cardiac Death Syndromes*. Springer, 2010.

[21] Institute For Further Professional Education and Studies. Electrocardiogram (ecg/ekg). http://www.ifpes.net/2011/10/09/electrocardiogram-ecg-ekg/, 2013.

[22] Jin F., Liu J., and Hou W. The application of pattern recognition technology in the diagnosis and analysis on the heart disease: Current status and future. *IEEE*, Control and Decision Conference (CCDC), 2012 24th Chinese:1304–1307, 2012.

[23] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[24] Wei G. and Shoubin W. Support vector machine for assistant clinical diagnosis of cardiac disease. *IEEE Computer Society; Global Congress on Intelligent Systems*, 3:588–591, 2009.

[25] Moody G.B., Muldrow WE, and Mark R.G. A noise stress test for arrhythmia detectors. *Computers in Cardiology*, 11:381–384, 1984.

[26] Pedro R Gomes, Filomena O Soares, JH Correia, and CS Lima. Cardiac arrhythmia classification using wavelets and hidden markov models–a comparative approach. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 4727–4730. IEEE, 2009.

[27] Lines G.T., Buist M.L, Grøttum P., Pullan, A.J., Sundnes J., and Tveito A. Mathematical models and numerical methods for the forward problem in cardiac electrophysiology. *Computing and Visualization in Science*, 5:215–239, 2003.

[28] F. Guilbaud and H. Pham. Optimal high-frequency trading with limit and market orders. *Quantitative Finance*, 2012.

[29] Engblom H., Foster J.E., Martin T.N., Groenning B, Pahlm O., Dargie H.J, Wagner S.G., and Arheden H. The relationship between electrical axis by 12-lead electrocardiogram and anatomical axis of the heart by cardiac magnetic resonance in healthy subjects. *Am Heart J*, 12:150–507, 2005.

[30] Malmivuo J. and Plonsey R. *Bioelectromagnetism. Principles and Applications of Bioelectric and Biomagnetic Fields*. New York: Oxford University Press, 1995.

[31] Moss J.A. Management of patients with the hereditary long qt syndrome. *Journal of Cardiovascular Electrophysiology*, 9(6):668–674, 1998.

[32] Madias J.E. On recording the unipolar ecg limb leads via the wilson's vs the goldberger's terminals: avr, avl, and avf revisited. *Indian Pacing and Electrophysiology Journal*, 4:292–297, 2008.

[33] Struijk J.J., Kanters, J.K., Andersen M.P., Hardahl T., Graff C., Christiansen M., and Toft E. Classification of the long-qt syndrome based on discriminant analysis of t-wave morphology. *Medical and Biological Engineering*, 44:543–549, 2006.

[34] Kanters J.K., Graff C., Andersen MP, Hardahl T., Toft E., Christiansen M., Thomsen P.E.B, and Struijk J.J. Long qt syndrome genotyping by electrocardiography: fact, fiction, or something in between? *Journal of Electrocardiology*, 39:S119–S122, 2006.

[35] Kanters J.K., Fanoe S., Larsen L.A., Thomsen P.E.B, Toft E., and Christiansen M. T wave morphology analysis distinguishes between kvlqt1 and herg mutations in long qt syndrome. *Heart Rhythm*, 3:285–292, 2004.

[36] B.H. Juang and L.R. Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[37] Antti Koski. Modelling ecg signals with hidden markov models. *Artificial intelligence in medicine*, 8(5):453–471, 1996.

[38] Sven E Krüger, Martin Schafföner, Marcel Katz, Edin Andelic, and Andreas Wendemuth. Using support vector machines in a hmm based speech recognition system. In *Specom*, pages 329–331. Citeseer, 2005.

[39] Thoraval L. and Carrault G. Heart signal recognition by hidden markov models: Theecgcase. *Methods of Information in Medicine*, 33(1):10–14, 1994.

[40] G Lannoy, B Frenay, M Verleysen, and J Delbeke. Supervised ecg delineation using the wavelet transform and hidden markov models. In *4th European Conference of the International Federation for Medical and Biological Engineering*, pages 22–25. Springer, 2009.

[41] Salem A-B. M., Revett K., and El-Dahshan E-S. A. Machine learning in electrocardiogram diagnosis. *Proceedings of the International Multiconference on Computer Science and Information Technology*, 4:429–433, 2009.

[42] Mathworks. How can i use continuous sequence values with hmmestimate in the statistics toolbox 7.1 (r2009a)? http://www.mathworks.se/support/solutions/en/data/1-EFNRHP/index.html?product=ST&solution=1-EFNRHP, 2012.

[43] MathWorks. Speeding up matlab applications. http://www.mathworks.com/tagteam/70598_91991v00_MATLABApps_WhitePaper.pdf, 2013.

[44] Robert M.F. and Rowlandson G.I. The effects of noise on computerized electrocardiogram measurements. *Journal of Electrocardiology*, 39:S165–S173, 2006.

[45] Andersen M.P, Xue J.Q., Graff C., Kanters J.K., Toft E., and Struijk J.J. New descriptors of t-wave morphology are independent of heart rate. *Journal of Electrocardiology*, 41:557–561, 2008.

[46] Blaustein M.P., Kao J.P.Y., and Matteson D.R. *Cellular Physiology*. MOSBY, Elsevier, 2010.

[47] Donnelly M.P., Finlay D.D., Nugent C.D., and Black N.D. Lead selection: old and new methods for locating the most electrocardiogram information. *Journal of Electrocardiology*, 41:257–263, 2008.

[48] Fowler N.O and Braunstein J.R. Anatomic and electrocardiographic position of the heart. *Circulation*, 3:906–910, 1951.

[49] University of Nottingham. Practice learning resources. http://www.nottingham.ac.uk/nursing/practice/resources/cardiology/function/chest_leads.php, 2013.

[50] Kligfield P., Gettes L.S, Bailey J.J., Childers R, Deal B.J., Hancock E.W., van Herpen G., Kors J.A., Macfarlane P., and Mirvis D.M. AHA/ACC/HRS scientific statements recommendations for the standardization and interpretation of the electrocardiogram. part I: the electrocardiogram and its technology a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the american college of cardiology foundation; and the heart rhythm society. *Journal of the American College of Cardiology*, 49:1109–1127, 2007.

[51] Chang P-C., Lin J-J., Hsieh J-C., and Weng J. Myocardial infarction classification with multi-lead ecg using hidden markov models and gaussian mixture models. *Applied Soft Computing*, 12:3165–3175, 2012.

[52] Pekka Paalanen. covfixer2 - force matrix to be a valid covariance matrix. http://www2.it.lut.fi/project/gmmbayes/doc/gmmbayestb-v0.2/base2/covfixer2.html, 2003.

[53] V.A. Petrushin. Hidden markov models: Fundamentals and applications. In *Online Symposium for Electronics Engineer*, 2000.

[54] Hedley P.L, Jørgensen P., Schlamowitz S., Wangari R., Moolman-Smook J., Brink P.A., Kanters J.K., Corfield V.A., and Christiansen M. The genetic basis of long qt and short qt syndromes: A mutation update. *Human Mutation*, 30(11):1486–1511, 2009.

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

[56] LR Rabiner, B.H. Juang, SE Levinson, and MM Sondhi. Recognition of isolated digits using hidden markov models with continuous mixture densities. *AT&T technical journal*, 64(6):1211–1234, 1985.

[57] Ludwig Schwardt. Gaussian mixture models. University of Stellenbosch, Lecture Slides TW414/PR813, 2003.

[58] Alba Sloin and David Burshtein. Support vector machine training for improved hidden markov modeling. *Signal Processing, IEEE Transactions on*, 56(1):172–188, 2008.

[59] Sörnmo and Laguna. *Bioelectrical Signal Processing in Cardiac and Neurological Applications*. Elsevier, 2005.

[60] W. Sun, H. Zhang, A. Palazoglu, A. Singh, W. Zhang, and S. Liu. Prediction of 24-hour-average pm< sub> 2.5</sub> concentrations using a hidden markov model with different emission distributions in northern california. *Science of The Total Environment*, 443:93–103, 2013.

[61] P. Tan. *Introduction to Data Mining*. Pearson Education, 2006.

[62] School of Electrical Engineering Systems Ted Burke. Biomedical engineering - ecg assignment. http://eleceng.dit.ie/tburke/biomed/assignment1.html, 2013.

[63] Metting van Rijn A.C., Peper A., and Grimbergen C.A. High-quality recording of bioelectric events, part 2 low-noise, low-power multichannel amplifier design. *Medical and Biological Engineering and Computing*, 29:433–440, 1991.

[64] SV Vaseghi. State duration modelling in hidden markov models. *Signal processing*, 41(1):31–41, 1995.

[65] Yi Xie, S Tang, C Tang, and X Huang. An efficient algorithm for parameterizing hsmm with gaussian and gamma distributions. *Information Processing Letters*, 2012.

[66] Wenjie Xu, Jiankang Wu, Zhiyong Huang, and Cuntai Guan. Kernel based hidden markov model with applications to eeg signal classification. In *IASTED International Conference on Biomedical Engineering (BioMED), Austria*, pages 16–18, 2005.

[67] Oksana Yakhnenko, Adrian Silvescu, and Vasant Honavar. Discriminatively trained markov model for sequence classification. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[68] S.X. Zhang and M.J.F. Gales. Structured svms for automatic speech recognition. *Audio, Speech and Language Processing*, 2013.