UNIVERSITÉ CATHOLIQUE DE LOUVAIN

INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES, ELECTRONICS AND APPLIED MATHEMATICS

MACHINE LEARNING GROUP

Automated Modeling and Processing of Long-Term Electrocardiogram Signals

Gaël de Lannoy

Thesis presented for the Ph.D. degree
in Engineering Sciences

Thesis Jury:
Prof. **M. Verleysen**, Advisor (UCL - ICTEAM Institute, Machine Learning Group)
Dr. **J. Delbeke**, Advisor (UCL - Neuroscience Institute)
Dr. **D. François** (UCL - ICTEAM Institute, Machine Learning Group)
Prof. **J.-P. Peters** (FUNDP - Departement de Physique)
Prof. **F. Melgani** (University of Trento, Italy)
Prof. **C. De Vleeschouwer**, President (UCL - ICTEAM Institute)

April 2011

Contents

Notations and Vocabulary

We have tried to use consistent notation throughout the text. Scalars are represented by italic letters like x. Vectors are column vectors and represented by lowercase bold letters like $\mathbf{x}$. The transpose of $\mathbf{x}$ is $\mathbf{x}'$, a row vector. The elements of a vector are defined using square brackets and the dimensionality of the vector is defined with uppercase letters. For example, a particular $P-$dimensional observation $\mathbf{x}$ is defined as $\mathbf{x} = [x^p]_{p=1}^P = [x^1, x^2, \ldots, x^p, \ldots, x^P]' \in \Re^P$. The subscript is used to denote the observation index and the upperscript to denote the dimension index, for instance x_n^p is the pth coordinate of the nth observation. The set $\mathbb{D}$ containing the N observations $\mathbf{x}_n$ is defined using curly brackets as $\mathbb{D} = \{\mathbf{x}_n\}_{n=1}^N$. We have used lowercase letters for the index associated to uppercase letters such as the number of elements or the dimensionality. For example, the index n over observations ranges from 1 to N and the index p over dimensions ranges from 1 to P. Matrices are denoted using bold uppercase letters. For example, $\mathbf{X}$ could be the $N \times P$ matrix formed by concatenating all observation vectors defined above: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]'$.

Notations

Specific recurrent notations will be used throughout the text:

- a.u. stands for arbitrary units;
- t is the index of the sampled time of a time-series ($1 \leq t \leq T$);
- p is the index over the dimensionality of a vector ($1 \leq p \leq P$);
- n is the index over the number of observations ($1 \leq n \leq N$);
- k is an index over class labels ($1 \leq k \leq K$);
- j is another index over class labels ($1 \leq j \leq K$);

- x is used to represent the values of a time-series;
- y is used to represent the class labels associated to the values of a time-series;
- $\mathbf{x}$ is used to represent observations in a dataset;
- $\mathbf{y}$ is used to represent the class labels associated to observations.

Abbreviations

The following important abbreviations are defined in the text:

AAMI	American association for medical instrumentation;
ANS	autonomic nervous system;
AV	atrio-ventricular;
BSS	blind source separation;
CRFs	conditional random fields;
CWT	continuous wavelet transform;
DWT	discrete wavelet transform;
ECG	electrocardiogram;
EEG	electroencephalogram;
EMG	electromyogram;
ENG	electroneurogram;
HMMs	hidden Markov models;
HRV	heart rate variability;
HSMMs	hidden semi-Markov models;
ICA	independent component analysis;
LDA	linear discriminant analysis;
MI	mutual information;
PNS	parasympathetic nervous system;
SA	sino-atrial;
SNS	sympathetic nervous system;
SVMs	support vector machines.

Short glossary of medical terms

A short glossary of recurrent medical terms is provided:

- **Afferent nerve:** Carries impulses from the organs to the nervous system;

- **Arrhythmia:** A pathological variation of the heart rhythm characterized by beats resulting from a default in conductivity of the sino-atrial node;
- **Autonomic nervous system:** The nerves involved in the autonomic nervous system are of major importance to assure involuntary functions in the body such as the activity of the cardiac muscle. The ANS is classically divided into two complementary subsystems that typically function in opposition to each other: the parasympathetic nervous system (slows the heart rhythm) and the sympathetic nervous system (accelerates the heart rhythm);
- **Baroflex:** A reflex where the heart rhythm accelerates and achieves stronger contractions when blood pressure falls;
- **Blood pressure:** The pressure exerted by circulating blood upon the walls of blood vessel. BP varies between a maximum (systolic) and a minimum (diastolic) pressure;
- **Bradychardia:** Slowness of the rhythm of the heart;
- **Cardiomyopathy:** A weakness of the heart muscle caused by any kind of disorder, often leading to heart failure;
- **Efferent nerve:** Carries impulses from the nervous system to the organs;
- **Fibrillation:** Abnormal and sustained small contractions of a part of the heart which can be fatal;
- **Hypertension:** Persistent high blood pressure;
- **Myocardial infarction:** Blockage in the electrical conduction of the heart;
- **Parasympathetic nervous system:** The parasympathetic nervous system slows the heart rhythm via the vagus nerve, for example during food digestion, within a time range from 1 to 3 seconds. It is also responsible for such varied autonomic tasks as gastrointestinal peristalsis (stimulation of the intestine), sweating and quite a few muscle movements in the mouth, including speech;
- **Sympathetic nervous system:** The sympathetic nervous system acts as an heart rate accelerator, and is activated during stress or exercise in a time range from 12 to 17 seconds. It can also inhibits the intestine;
- **Tachycardia:** Acceleration of the rhythm of the heart;

- **Vagus nerve:** The vagus nerve is the supplier of efferent motor parasympathetic fibers to all the organs (except the suprarenal glands), from the neck down to the second segment of the transverse colon. Besides this output to the various organs in the body the vagus nerve also conveys sensory information about the state of the organs to the central nervous system. 80-90% of the nerve fibres in the vagus nerve are afferent (from the organs to the ANS) fibres communicating the state of the viscera to the brain.

Chapter 1

Introduction

House: “I had a heart attack this morning. I can’t do any more drugs until at least lunch.”

House MD, season 4 episode 2.

Contents

1.1 Foreword

With the growing complexity of clinical technology, the medical community is now facing large amounts of data. These data originate from such various sources as radiology imaging, microarray experiments, spectroscopy and physiological signals among others. Physiological signals consist in the recording of the electrical activity generated by the human body, for example in the muscles or in the brain. The analysis of these signals yields a great potential for such various tasks as brain-computer interfaces and the monitoring of body functions, including the diagnosis of disease conditions. For example, the recording of the brain electrical activity is an established technique for the diagnosis of epilepsy. Physiological signals are often obtained over a long-term period, sometimes up to several days. Then arises the difficulty of analyzing these long-term recordings. Nowadays, this analysis is still often performed visually by an expert. Needless to say, this task is a very tedious and time-consuming process which can lead to errors and misinterpretations, even when carried on by the best trained experts.

Machine learning is a discipline at the crossroad between mathematics, computer sciences and statistics that is concerned with the design of algorithms that automatically learn to recognize complex patterns from large amounts of empirical data. The learning task is very challenging in practice because the collected data are always finite (they do not cover the set of all possible instances) and noisy (they can contain errors). The goal is then to use that knowledge to come to a useful automated decision on new cases. Machine learning therefore appears as an appealing solution to overcome the difficulties faced by the manual analysis of long-term physiological signal recordings. In particular, the electrocardiogram (ECG) is a physiological signal representing the electrical activity produced by the heart. In the large majority of situations involving the recording of an ECG signal, a long-term monitoring is required not to miss any transient pattern.

This thesis focuses on the design and the assessment of machine learning algorithms to automatically process such long-term ECG recordings. In the following part of this introduction, the properties and the clinical uses of ECG signals are presented. Next, the four main ECG analysis topics investigated in this thesis are presented.

1.2 The electrocardiogram

In this section, the underlying physiological process that forms the electrocardiogram signal and the recording hardware are presented. The several sources of noise that contaminate electrocardiogram recordings are described. The clinical uses of the electrocardiogram signal are also shortly explained. Further details can be found in [1]

and in [20].

1.2.1 Physiology of the heart

Before attempting any signal processing of the ECG signal, it is important to first consider the basic anatomy and function of the heart, as shown in Fig. 1.1. The heart is a pump which drives the circulation of blood throughout the body. The heart cavity is vertically divided in two separated parts. Each part is then further horizontally divided in two chambers. The upper chamber contains the atrium, and the lower chamber contains the ventricle. The two chambers are linked by a valve that closes after its contraction to prevent any retrograde flow. The atria and the ventricles in the heart are composed of muscle cells. It is the rhythmical contractions of these cells that provide the driving force for the circulation of blood in the body. The role of the right atrium is to receive blood from everywhere in the body and to feed it into the right ventricle. The right ventricle then propels the blood to the lungs where it is oxygenated. The left atrium receives the blood from the lungs and the left ventricle ejects the oxygenated blood throughout the whole body.

The rhythmic contraction of the heart is triggered by a wave of electrical current called action potential. In normal situations, the action potential is initiated by cardiac cells in the sino-atrial (SA) node in the upper right atrium that have the ability to spontaneously depolarize. The depolarization turns into a mechanical contraction of the cell. Once initiated, the action potential is propagated by means of direct current spread (so without the need of electrochemical synapses) to adjacent cells in the atrio-ventricular (AV) node which is at the junction between the atria and the ventricles. The electrical impulse then penetrates into the ventricles via the His bundle. From the His bundle, the electrical impulse enters the two bundle branches (the right and the left). The right and left bundle branches send the electrical impulse to the right and left ventricle, respectively. When the bundle branches are functioning normally, the right and left ventricles contract nearly simultaneously.

As a result of the electrical activity of the cells, current flows within the body and potential differences are established at the surface of the skin. Recording these potential differences as a function of time produces the electrocardiogram. It is the propagation of the action potential through the atria and the ventricles during each heart beat that yields the characteristic waveforms of the ECG signal. Figure 1.2 shows two artificial heart beats and the corresponding characteristic waves. These waveforms represent either depolarization (electrical discharging) or repolarization (electrical recharging) of the heart muscle cells in the atria and the ventricles. The P wave corresponds to the depolarization of the atria and the QRS complex to the depolarization of the ventricles. Because the ventricles contain more muscle mass

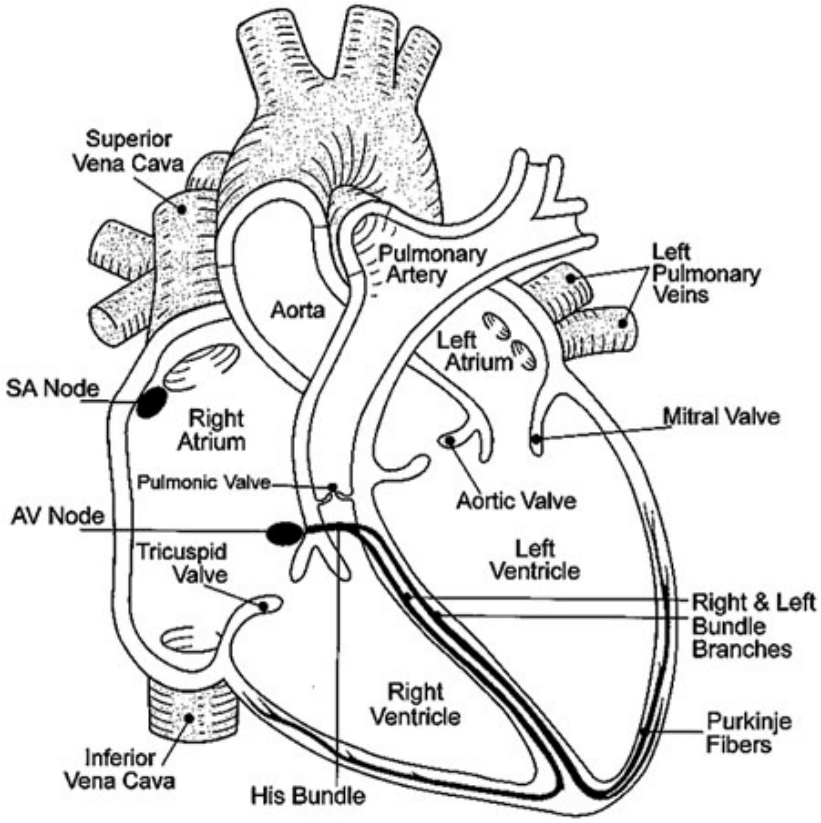

Figure 1.1: Basic anatomy of the heart, reproduced from [62]. The rhythmic contraction of the heart is triggered by a wave of action potentials that originates in the SA node in normal situations. The electrical wave propagates within the heart and initiates the coordinated contraction of the atria and of the ventricles.

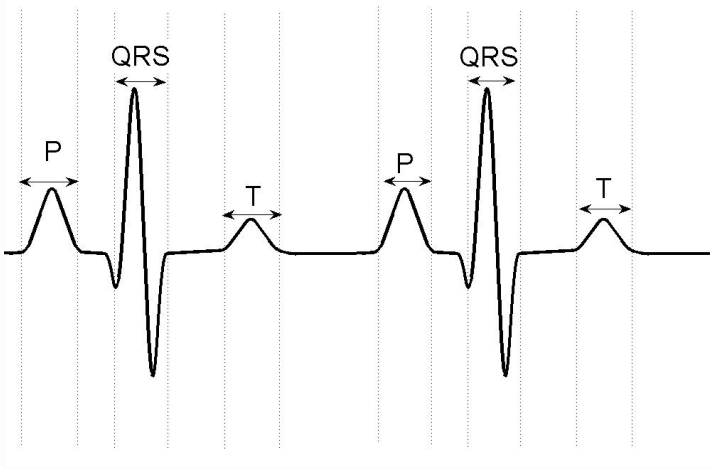

Figure 1.2: A clean artificial ECG signal with two annotated beats. The propagation of action potentials through the atria and the ventricles yields the characteristic P, QRS and T waves of the ECG signal. The P wave corresponds to the contraction of the atria, the QRS complex to the contraction of the ventricles and the T wave to their repolarization.

than the atria, the repolarization of the atria is masked by the depolarization of the ventricles. Finally, the T wave corresponds to the repolarization of the ventricles.

1.2.2 Recording systems

The electrocardiogram originated in the beginning of the 20th century from pioneer work of Willem Einthoven, a Dutch[1] doctor and physiologist who was the first to record the electrical activity of the heart at the surface of the skin using a string galvanometer. This work earned Einthoven the Nobel prize for medicine in 1924. The technique developed by Einthoven required that patients had both arms and left leg in separate buckets of saline solution which acted as electrodes to conduct the current. Needless to say, this was not a very convenient procedure.

Nowadays, the technique has evolved and the standard procedure for recording an ECG is the so-called 10 seconds 12-lead ECG. This type of recording makes use

[1] This is why the ECG is sometimes named EKG, after its Dutch origins.

of four limb electrodes and six chest electrodes carefully placed at strategic positions. The potential differences measured between pairs of electrodes are signals, often called *leads* in the medical literature, showing the activity of the heart from different viewpoints. In the standard 12-lead ECG, 12 pairs of electrodes are considered to represent distinct spatial perspectives of the heart's electrical activity. Modern 12-lead ECG hardware record the ECG at sampling frequencies between 250 and 500 Hz. The magnitude of the R-wave is between 1 and 2 mV and the frequency range of the ECG signal is between 0.1 and 250 Hz.

Nevertheless, this type of procedure is only used to record signals of very short duration, typically 10 seconds. Such short-term ECGs are only useful to observe permanent structural abnormalities of the heart, such as enlargement. In many practical situations such as exercise tests, clinical monitoring or pharmaceutical phase-one studies, an extended recording period is required to detect transient symptoms such as intermittent cardiac arrhythmias.

In these situations, long-term recordings can be obtained using the popular Holter[2] recorders. These systems are ambulatory heart activity recording units delivering signal storing capacity for at least 24 hours and up to eleven days with modern systems. Most Holter systems record 2 or 3 ECG signals obtained from between 3 and 8 electrodes, depending on the model. The electrodes are connected to a small recorder that is attached to the patient's belt or hung around the neck, and will log the heart's electrical activity throughout the recording period. Nevertheless, recordings from Holter monitors are of significantly lower resolution than those from a standard 12-lead ECG. The signals obtained are much noisier due to movements of the patient in his normal daily activities and the progressive degradation of electrode contact.

More recently, significant progresses in wireless monitoring and wearable sensors have enabled the development of so-called body sensor networks, a promising healthcare candidate solution. Such equipments replace the traditional electrodes attached to the skin with a sticker by miniature sensors embedded in textile or even implanted. Due to the miniature sized sensors and the ability to achieve wireless communication with processing and storing devices, the interference of this recent equipment with the wearer's daily life is very much reduced.

1.2.3 Sources of noise

Several sources of noise typically contaminate recorded ECG signals. This section provides a non-exhaustive list of the most important artifacts.

1. Power line: artifact corresponding to 50 Hz or 60 Hz interference, depending on the country. This noise can be filtered by notch filters.

[2]Named after its inventor Norman Holter.

2. Electrode contact: loss of contact between electrode surface and the skin, for example due to patient movements or conduction gel leakage. It is characterized by sharp changes with saturation of the signal.

3. Electromyogram: electrical activity due to muscle contractions near the ECG recording sites, for example pectorals or abdominals. This artifact overlaps the frequency spectrum of the ECG.

4. Respiration: baseline wanderings due to respiration that correspond to a sinus-like wave at frequencies below 1 Hz.

5. Movements: noise due to movements of the patient that yields rapid baseline jumps in the signal.

6. External noise: artifacts from any external hardware such as the recording hardware, mobile phone interference, etc.

1.2.4 Clinical uses

The ECG is a powerful non-invasive way to identify cardiac rhythm disorders as well as structural abnormalities of the heart. In this section, several pathologies that can be identified on the ECG signal are briefly described and illustrated in Fig. 1.3. Fibrillation and conduction blocks are two cases of transient unstable arrhythmias. Ischemia and the *torsades de pointes* syndrome are two cases of abnormal durations between ECG characteristic points which are often intermittent.

Fibrillation

When a heart beat originates from cells outside the SA node, it is called an ectopic or premature beat. Common causes of ectopic beats are drugs or damage of a part of the heart. If the cause of the ectopic beat is sustained, it can in some cases degenerate into a burst of smaller waves that propagates randomly. This phenomenon is called fibrillation. While atrial fibrillation is well tolerated in general, ventricular fibrillation can lead to a complete loss of coordination in muscle contractions and can be fatal within a few seconds. Atrial ectopic beats appear as abrupt normal beats (see Fig. 1.3, first plot) and ventricular ectopic beats appear as sharp spikes in the ECG (see Fig. 1.3, second plot). The last plot in Fig. 1.3 shows an episode with ventricular ectopic beats that degenerates into ventricular fibrillation.

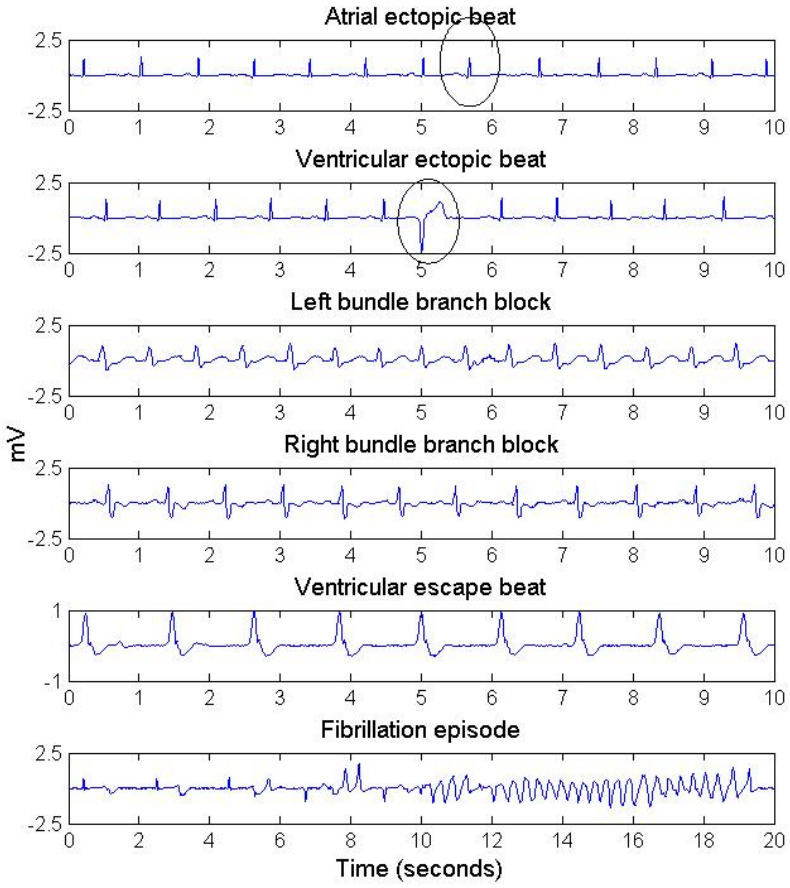

Figure 1.3: Several pathological conditions that can be identified on the ECG signal. The first two plots illustrate ectopic beats. Atrial ectopic beats appear as abrupt normal beats and ventricular ectopic beats appear as sharp spikes in the ECG. The third and fourth plots show a left and a right bundle branch block situation respectively. The fifth plot shows a ventricular escape beat. The last plot shows an episode with ventricular ectopic beats that degenerates into ventricular fibrillation.

Conduction blocks

In some cases, the conduction of the electrical wave can be blocked. For example, the AV node may fail to efficiently conduct action potentials from the atria to the ventricles. This phenomenon is called bundle branch block. The third and fourth plot in Fig. 1.3 show a left and a right bundle branch block situation respectively. In such situations, cells in the AV node who also have the ability to spontaneously depolarize will act as a backup and initiate the contraction of the ventricles to maintain blood flow. These fail-safe beats are called escape beats and are easily identified on the ECG signal, as shown in the fifth plot in Fig. 1.3.

Ischemia

Ischemia is a shortage in blood supply. The heart can suffer from ischemia, which will have functional consequences. This condition can lead to loss of consciousness and in worst cases to myocardium infarction. Ischemia can be diagnosed from the ECG signal by monitoring duration abnormalities in the S-T interval (the period when the ventricles are depolarized) and changes in appearance of the T wave.

Torsade de pointes

The administration of certain drugs which are not expected to affect the heart have been shown to lengthen the duration of ventricular repolarization [84]. It is a serious condition that can lead to very fast abnormal heart rhythm known as *torsades de pointes*[3] which can itself quickly turn into cardiac death. This phenomenon can be observed on the ECG by monitoring the duration of the Q-T interval. Pharmaceutical phase-one studies for the evaluation of new drugs always include an ECG monitoring for this reason.

1.3 Objectives of this thesis

Long-term ECG recordings, for example recorded using a Holter device, typically contains hundreds to thousands of heart beats to evaluate. The analysis of the recording is therefore a very time-consuming process and the task is difficult since the diagnosis may rely on just a few transient patterns. For this reason, computer-aided analysis of the ECG signal is of crucial help to cardiologists. In this thesis, four main ECG analysis topics are investigated using machine learning tools. These four topics are now presented and will be investigated in details in the next four chapters.

[3] A French term that literally means “twisting of the peaks”.

1.3.1 Segmentation of ECG characteristic points

The first topic is the accurate automatic segmentation of the ECG characteristic points, corresponding to the onset and ending of the P, QRS and T waves. Measurements pertaining to these characteristic points are of major importance for evaluating the cardiac function of patients. Moreover, the annotated ECG characteristic points are the foundation on which the next three ECG analysis topics rely. In the past few years, variants of the hidden Markov models (HMMs) have successfully been applied to automate the ECG segmentation. HMMs offer significant improvements over standard heuristic segmentation approaches. Recently, another probabilistic model called conditional random fields (CRFs) is gaining popularity. CRFs outperform HMMs on a wide number of sequence labeling tasks. In Chapter 2, a methodology using sparse conditional random fields and the continuous wavelet transform is proposed and compared to previously reported models. Experiments are conducted on both normal and pathological Holter ambulatory recordings from the Physiobank database.

1.3.2 Heart rate variability metrics

The second topic concerns the analysis of fluctuations in heart rate. In normal situations, the heart rate is determined by the autonomic nervous system. The variation in autonomic activity is typically quantified by heart rate variability (HRV) metrics which are computed on the timing intervals between successive R points. In Chapter 3, the possible use of heart rhythm variations as a marker of seizures and seizure onset is investigated. This topic is rapidly gaining interest in the epileptology community but it has confronted the potential clinical users with a large number of relatively complex HRV metrics to choose from. For this reason, standards for the measurement, interpretation and use of heart rate variability metrics are presented in a form directed at clinicians. This critical overview discloses the state-of-the art including more recent non-linear metrics and describes why some previous methods for processing the ECG signal are inadequate. Experiments are conducted on real ECG data from epileptic patients.

1.3.3 Classification of heart beats

The third topic focuses on the automatic and supervised classification of heart beats, i.e. the labeling of beats in a recorded ECG signal as either a normal beat, an ectopic beat, a bundle branch block beat, etc. Given a labeled set of beats, a model is created and later used to label new beats. This is primordial in many applications requiring long-term monitoring of the cardiac function where thousands of beats have to be diagnosed. The main difficulties are the strong unbalance in the number of beats

of each type and the extraction of discriminative features from the heart beat time-series. In Chapter 4, the problems associated to learning with unbalanced datasets are described and solutions are proposed. The relevance of feature sets previously proposed in the literature is also investigated using feature selection techniques. Good practice rules for the establishment of a reliable heart beat classification methodology are emphasized, such as the inter-patient classification paradigm. Experiments are conducted on real Holter recordings from the Physiobank arrhythmia database.

1.3.4 ECG as an artifact

Due to the proximity of the recording sites with the heart, the electrical current associated to each heart beat appears in several other physiological signals as an artifact. This is for example the case with EMG signals obtained from the trunk musculature or with fetal ECG recordings. Vagus nerve recordings have recently been achieved to gain a better understanding of its mechanisms of interaction in several pathologies such as refractory epilepsy and severe depression. It has been observed that the vagus recordings also suffer from the cardiac artifact contamination. These artifacts can significantly prevent the extraction of accurate information from the nerve signal. The fourth topic, investigated in Chapter 5, concerns the removal of cardiac artifacts in neural recordings using filtering and blind source separation techniques. Experiments are conducted on real vagus nerve recordings in rats.

Foreman: “You want to bet on the patient’s health?”
House: “You think that’s bad luck? You think that God will smite him because of our insensitivity? Look, if God does, you make a quick fifty.”

House MD, season 1 episode 3.

Chapter 2

Automatic Segmentation of the ECG Signal

Contents

2.1 Introduction

The ECG signal is characterized by a time-variant cyclic occurrence of patterns with different frequency contents (QRS complexes, P and T waves, remember Fig. 1.2). The P wave corresponds to the contraction of the atria, the QRS complex to the contraction of the ventricles and the T wave to their repolarization. The ECG characteristic points are the onset and ending of these three ECG waveforms.

During clinical monitoring or phase-one evaluation studies of new drugs, measurements pertaining to the ECG characteristic points are used to assess the state of a patient's heart. In particular, the time between the Q peak and the end of the T wave (Q-T interval) is currently the gold standard for evaluating the cardiac safety of new drugs in clinical trials [56]. The R-R time intervals also serve for the quantification of the heart rate variability metrics used in the diagnosis of cardiac arrhythmia and of specific cases of neuropathies [1].

Manually annotating long-term ECG signals is a tedious and time-consuming process which can lead to errors and misinterpretations. This difficulty has for example been experienced during the clinical trials of the antihistamine terfenadine, a drug that helps against allergies by blocking the action of histamine[1]. This drug has the side effect of significantly prolonging the Q-T interval in patients who have a rare abnormality of the heart muscle or who are under heart disease medications. Unfortunately, this side effect was not observed on time during the clinical trials and only came to light after a large number of patients unexpectedly died while taking the drug [84, 20].

Automatic computer-aided segmentation of the ECG signal is thus a milestone for ECG analysis and can greatly help physicians in their diagnosis. However, it is a difficult task in real situations. First, because of the physiological variations resulting from patient activities and disease instabilities, the ECG is a non-stationary signal. Second, the time interval between successive waveforms is normally varying. Third, many sources of noise pollute the ECG signal, such as power line interferences, muscular artifacts, poor electrode contacts and baseline wanderings due to respiration. These problems highly compromise the effective segmentation of the signal, especially with real-life signals such as ambulatory recordings. Figure 2.1 shows a 20 seconds extract of an ambulatory recording including artifacts. Clearly, the identification of the characteristic points is not straightforward.

Standard computer-aided methods for segmentation of the ECG signals attempt to find the characteristic points in a number of successive steps by using sophisticated thresholding methods and heuristic rules such as tangent slope criterion [20]. The detection is then often followed by a post-processing step to remove the aber-

[1] Histamine is an organic nitrogen compound involved in local immune responses, it is present in virtually all body cells.

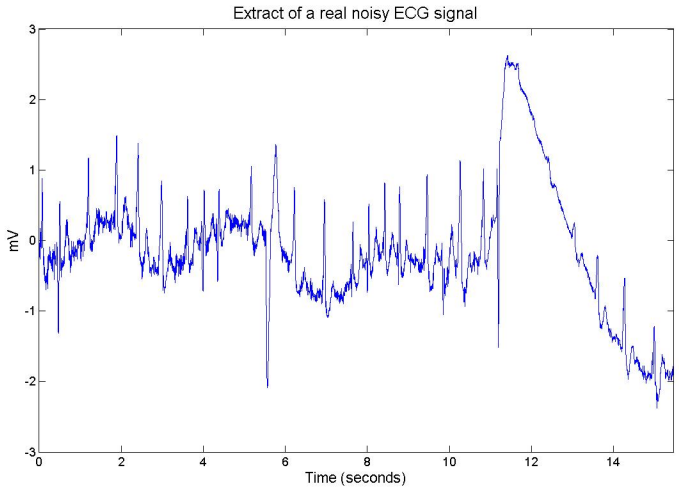

Figure 2.1: Extract from a real ambulatory ECG recording. The presence of many sources of noise such as muscular artifacts and baseline wanderings compromise the effective segmentation of the signal.

rant segmentations made by the algorithm. We refer to these algorithms as heuristic algorithms. The most famous heuristic algorithm performing QRS points detection is the Pan and Tompkins algorithm [92]. Although a large number of other heuristic algorithms have since been published (see for example [3] for a review), very few significant improvements have actually been achieved and the Pam and Tompkins algorithm remains the gold standard in clinical applications. When the segmentation of all the characteristic points is required, heuristic algorithms of reference include the algorithm of Li [66] performing complete beat segmentation using thresholding of wavelet coefficients, and the improvement of the Pan and Tompkins algorithm by Laguna [60] to perform a complete segmentation of all the characteristic points. Nevertheless, these heuristic algorithms all require the setting of many empirical parameters and have been found very sensitive to noisy waveforms, arrhythmias and changes in wave morphology [20, 72].

In order to circumvent these issues, probabilistic modeling approaches and especially hidden Markov models (HMMs) have more recently been introduced in the frame of automatic ECG segmentation. Probabilistic models offer significant improvements over heuristic methods such as the ability to learn their parameters from expert annotated data and the ability to include prior knowledge about the problem, like the statistical properties of waveforms [20]. This data-driven learning is of particular inter-

est since it enables the building of models that are optimized to detect specific patient waveform morphologies. This is inherently unreliable in heuristic approaches because of the wide range of ECG morphologies that can occur in practice. The use of probabilistic models such as HMMs for the analysis of ECG signals is not new [54], but it is only recently that the effect of the ECG representation and of the HMMs architecture on the segmentation reliability have been investigated [20, 49].

Nevertheless, an important weakness of HMMs is the independence assumption made on observations. More precisely, in a HMM, an observation at a given time step must be independent from the other observations in the sequence. This is an appropriate assumption for a few simple data sets, but most real-world observation sequences such as physiological signals are best represented in terms of long-range dependencies between observations. Tricky features that voluntarily violate the independence assumption such as the wavelet transform [49] or autoregressive emissions [20] have then been designed to achieve good ECG segmentation performances with HMMs. More recently, another probabilistic model for labeling sequential data called conditional random fields (CRFs) is expanding outside of its original natural language processing context [58]. CRFs relieve the strong independence assumption over observations and outperform HMMs on a wide number of sequence labeling tasks such as image processing, text processing, activity recognition and bioinformatics [113].

The methodology for using CRFs in the context of ECG segmentation is the focus of this chapter. The performances of CRFs are compared to those of state-of-the-art hidden Markov models in both normal and pathological situations. Moreover, the advantages offered by the regularization of the CRF objective function by the L_1-norm of the parameter vector are investigated. Such regularization of the objective function encourages sparsity over the parameters of the model and therefore achieves instrinsic feature selection. Personal publications about the content of this chapter are [30, 31, 32, 41, 42, 25].

2.2 Graphical models

Graphical models are an elegant framework that combines uncertainty (probabilities) and logical structure (independence constraints) to compactly represent complex real-world phenomena such as non-independently distributed sequential data. The framework is quite general in that many of the commonly proposed statistical models such as HMMs and CRFs can be described as specific instances of graphical models.

More specifically, graphical models represent a probability distribution over some set $\mathbb{V}$ of i random variables. Even in the simplest cases where these variables are binary-valued, a joint distribution requires the specification of the probabilities of the 2^i distinct assignments. However, it is often the case that there is some structure in

the distribution that allows us to factor its representation into modular components. The structure that graphical models exploit is the independence properties that exist in many real-world phenomena to represent such high-dimensional distributions much more compactly.

Let us define such a set $\mathbb{V} = \{\mathbf{x}_t, y_t\}_{t=1}^{T}$ corresponding to a $P-$dimensional observation sequence $\mathbf{x}'_t = [x_t^p]_{p=1}^{P} \in \mathbb{R}^P$ and the associated labels, called states, $y_t \in \{1,2,\ldots,k,\ldots,K\}$. Directed graphical models, also known as Bayesian networks or finite-state automa, represent the family of distributions that factorize as

$$p(\mathbb{V}) = \prod_{\upsilon \in \mathbb{V}} p(\upsilon|\pi(\upsilon)) \tag{2.1}$$

where υ are the nodes in the graph and $\pi(\upsilon)$ are the parents of node υ.

In undirected graphical models, also known as random fields, there is no such notion of parents and children. Instead, the probability is proportional to the product of a series of non-negative potential functions ψ, with one potential function for each clique (a fully and directly connected subgraph) c in the graph:

$$p(\mathbb{V}) = \frac{1}{Z} \prod_{c \in \mathit{cliques}(\mathbb{V})} \psi(c) \tag{2.2}$$

where Z is a normalization term which guarantees that the probabilities sum to one. In undirected graphs, the cliques represent the dependencies between variables.

Figure 2.2 shows a directed (left) and an undirected (right) graphical model representing a probability distribution over $\mathbb{V}$. In the left graph, the shaded nodes correspond to observations $\mathbf{x}_t$ and the white nodes to states y_t. In the right graph, the single black node is the entire observation sequence $\mathbf{x}$. The graphical representation illustrates the independence assumptions and the way both models factorize probabilities. In the directed graph, each observation node only depends on its father state node and not any other node. On the other hand, in the undirected graph, each clique contains the whole observation sequence hence no independence assumption is made over observations. About state nodes, we can see in both graphs that each of state node only depends on the preceding state node. As we will now detail in the next sections, these two particular graph structures actually correspond to the specific cases of HMMs (the directed graph) and CRFs (the undirected graph).

2.3 Hidden Markov models

Hidden Markov models are a form of directed graphical model represented in Fig. 2.2 (left). As the graphical representation illustrates, the HMM relies on two strong assumptions. First, in order to keep inference in this joint model tractable, the HMM

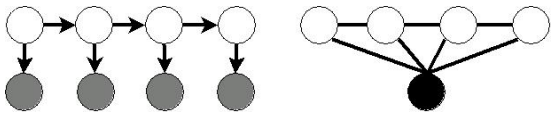

Figure 2.2: Graphical representation of a directed (left) and an undirected (right) graphical model. The graphs here represent the special case of a HMM (left) and of a CRF (right). The shaded nodes correspond to observations $\mathbf{x}_t$ and the clear nodes to labels y_t. In the CRF model, the single black node is the entire observation sequence $\mathbf{x}$. The graphical representation illustrates the independence assumptions and the way both models factorize probabilities.

relies on the naive Bayes assumption and treats the observations as conditionally independent given the state labels. More precisely, the observation element at any given time step only directly depends on the state at that time and not on any other observation. This is a very severe limitation, since most real-world applications are best represented in terms of multiple interacting features and long-term dependencies between observation elements. Second, the first-order Markov assumption defines that a future state, at any given moment, depends only on the present state, and not on any past states.

According to Eq. (2.1) and the underlying graphical representation which defines the independence constraints, the joint probability defined by HMMs is

$$p(\mathbf{x},\mathbf{y}) = \prod_{t=1}^{T} \overbrace{p(y_t|y_{t-1})}^{\text{First-order Markov}} \overbrace{p(\mathbf{x}_t|y_t)}^{\text{Naive Bayes}} . \tag{2.3}$$

2.3.1 Training

The HMM requires the learning of the three following items :

1. The $K \times K$ transition matrix $\mathbf{A}$ with

$$a_{kj} = p(y_{t-1} = k, y_t = j). \tag{2.4}$$

2. An observation probability distribution b_k for each state k, with

$$b_k(\mathbf{x}_t) = p(\mathbf{x}_t|y_t = k), \tag{2.5}$$

   which can for example be a parametric continuous probability distribution, i.e. a Gaussian (mixture) distribution.

3. The vector π of initial state probabilities with

$$\pi_k = p(y_0 = k). \tag{2.6}$$

In a supervised setting, the a_{kj}, π_k and the parameters of the emission distributions can be estimated from the available training data [97].

2.3.2 Inference

In a segmentation task, the final objective is to find the best label sequence $\mathbf{y}^*$ for an unlabeled observation sequence $\mathbf{x}$, given a particular HMM model. This sequence can be inferred efficiently by the forward-backward algorithm [97]. Consider the forward variable

$$\alpha_t(k) = p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, y_t = k), \tag{2.7}$$

the probability of the partial observation sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ until time t and state k at time t. We can solve $\alpha_t(k)$ recursively as follows:

1. Initialization:

$$\alpha_1(k) = \pi_k b_k(\mathbf{x}_1), \quad 1 \leq k \leq K. \tag{2.8}$$

2. Induction:

$$\alpha_t(k) = \left(\sum_{j=1}^{K} \alpha_{t-1}(j) a_{jk} \right) b_k(\mathbf{x}_t), \quad t = 2, 3, \ldots, T \quad 1 \leq k \leq K. \tag{2.9}$$

3. Termination:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_T(k). \tag{2.10}$$

Let us also consider the backward variable $\beta_t(k) = p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T | y_t = k)$, the probability of the partial observation sequence from $t+1$ to the end given state k at time t. We solve $\beta_t(k)$ recursively in a similar manner:

1. Initialisation:

$$\beta_T(k) = 1, \quad 1 \leq k \leq K. \tag{2.11}$$

2. Induction and termination:

$$\beta_t(k) = \sum_{j=1}^{K} a_{kj} \beta_{t+1}(j) b_j(x_{t+1}), \quad t = T-1, T-2, \ldots, 1 \quad 1 \leq k \leq K. \tag{2.12}$$

Since $\alpha_t(k)$ accounts for the partial observation sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ and state k at time t, and $\beta_t(k)$ accounts for the remainder of the observation sequence $\{\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T\}$ given state k at time t, the forward and backward terms can be combined to obtain

$$\gamma_t(k) = p(y_t = k|\mathbf{x}) = \frac{\alpha_t(k)\beta_t(k)}{\sum_{k=1}^{K} \alpha_t(k)\beta_t(k)}. \tag{2.13}$$

The individually most likely state at every instant $\mathbf{y}^*$ can then be solved for all t as

$$y_t^* = \max_k \gamma_t(k), \quad 1 \leq t \leq T. \tag{2.14}$$

2.3.3 Hidden semi-Markov models

One of the major weaknesses of the standard HMM depicted in Eq. (2.3) is the modeling of state durations. The inherent duration probability $p_k(d)$, the probability of staying d time steps in state k, is

$$p_k(d) = (a_{kk})^{d-1}(1 - a_{kk}). \tag{2.15}$$

This exponential state duration density is inappropriate for most physical signals. The alternative proposed in hidden semi-Markov models (HSMMs) [65] is to use a parametric family of continuous probability density functions to provide duration probabilities, such as the Poisson or gamma distributions. An HSMM differs from an HMM in the following two ways: the transition matrix is constrained to have a leading diagonal of zeros, $a_{jj} = 0 \; \forall j$, and an explicit parametric duration probability distribution $p_k(d)$, the probability of staying d steps in state k, is specified for the duration of each state. This distribution can for example be the gamma or Poisson distribution. A transition is therefore made only after the appropriate number of observations has occurred in a given state, as specified by the duration distribution, and no transition back to the same state can occur.

The forward induction step in Eq. (2.9) becomes:

$$\alpha_t(k) = \sum_{j=1}^{K} \sum_{d=1}^{\min(t,D)} \left(\alpha_{t-d}(j) a_{jk} p_k(d) \prod_{t'=t-d+1}^{t} b_k(\mathbf{x}_{t'}) \right) \tag{2.16}$$

where D is the maximum duration within any state. The backward induction step in Eq. (2.12) can also be modified as follows:

$$\beta_t(k) = \sum_{j=1}^{K} \sum_{d=1}^{\min(t,D)} \left(\beta_{t+d}(j) a_{kj} d_j(d) \prod_{t'=t+1}^{t+d} b_j(\mathbf{x}_{t'}) \right). \tag{2.17}$$

The most likely sequence of states given a sequence of observations can then be obtained using Eq. (2.13) and Eq. (2.14) left unchanged.

2.3.4 Limitations of HMMs

The main drawback of HMMs (and HSMMs) is their generative nature. This means that they require the modeling of $p(\mathbf{x})$, which is not needed for classification anyway [86]. With non-independent data like time-series, the difficulty in modeling $p(\mathbf{x})$ lies in the fact that $\mathbf{x}$ often contains many highly dependent features that are difficult to model. To model $p(\mathbf{x})$, all possible observation sequences should be enumerated, a task which is hard if observations have long-distance dependencies [113]. For this reason, generative models must make strict independence constraints among observations to retain tractability, like the naive Bayes assumption. We illustrate this by expanding the HMM model without the naive Bayes and first-order Markov assumptions:

$$\begin{aligned} p(\mathbf{x},\mathbf{y}) &= p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) && (2.18)\\ &= p(\mathbf{x}_T,\mathbf{x}_{T-1},\mathbf{x}_{T-2},\ldots,\mathbf{x}_1|\mathbf{y})p(\mathbf{y}) && (2.19)\\ &= p(\mathbf{x}_T|\mathbf{y})p(\mathbf{x}_{T-1},\mathbf{x}_{T-2},\ldots,\mathbf{x}_1|\mathbf{y},\mathbf{x}_T)p(\mathbf{y}) && (2.20)\\ &= p(\mathbf{x}_T|\mathbf{y})p(\mathbf{x}_{T-1}|\mathbf{y},\mathbf{x}_T)p(\mathbf{x}_{T-2},\mathbf{x}_{T-3},\ldots,\mathbf{x}_1|\mathbf{y},\mathbf{x}_T,\mathbf{x}_{T-1})p(\mathbf{y}) && (2.21)\\ &= \ldots && (2.22) \end{aligned}$$

and so on using repeated applications of the definition of conditional probability.

Clearly, such a model is intractable in practice even for small values of T. This is where the naive Bayes assumption comes into play by assuming $p(\mathbf{x}_t|\mathbf{y},\mathbf{x}_{t'}) = p(\mathbf{x}_t|y_t)$. The joint model can then be written as

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{y})\prod_{t=1}^{T} p(\mathbf{x}_t|y_t). \tag{2.23}$$

From the first-order Markov assumption, $p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t|y_t-1)$ and Eq. (2.23) renders the classical HMM formulation depicted in Eq. (2.3). Nevertheless, most sequential data contain long-distance dependencies between observation elements and benefit from being represented by a discriminative model that allows such dependencies and enables observation sequences to be represented by non-independent overlapping features [86, 113].

Table 2.1 shows existing models for sequential data and their properties. Maximum entropy Markov models (MEMMs) are directed graphical models directly modeling the conditional probability $p(\mathbf{y}|\mathbf{x})$ to overcome the limitations of generative models like HMMs [78]. The problem with discriminative directed models is that they suffer from the so-called label bias issue [58]. Briefly said, because of the per-state normalization factor in directed models which guarantees that the outgoing transitions of any state sums to one, states with fewer transitions give more probability mass to their successors. This causes a bias towards states with fewer outgoing transitions during inference and it can have serious performance decrease in real applications [58].

	Directed	**Undirected**
Generative	Hidden Markov models (HMMs)	Markov random fields (MRFs)
Discriminative	Maximum entropy Markov models (MEMMs)	Conditional random fields (CRFs)

Table 2.1: Probabilistic models for sequential (non independent) data can be either discriminative or generative, and have a graphical structure being either directed or undirected. The combinations of these criterion leads to four distinct types of models which are presented in this table.

The solution to this label bias problem is to use transition scores rather than transition probabilities and to use a global normalization factor. This is the idea followed in conditional random fields, described in the next section.

2.4 Conditional random fields

Conditional random fields (CRFs) [58] are a form of undirected graphical model represented in Figure 2.2 (right), which also relies on the first-order Markov assumption over labels. In the specific case of CRFs, as illustrated by the graph, the cliques consist of an edge between y_t, y_{t-1} and the full sequence of observations $\mathbf{x}$.

Following from the fact that CRFs are discriminative and their associated undirected graphical structure, the probability distribution defined by CRFs is represented from Eq. (2.2) as

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \psi(y_{t-1}, y_t, \mathbf{x}) = \frac{\prod_{t=1}^{T} \psi(y_{t-1}, y_t, \mathbf{x})}{\sum_{\mathbf{y}} \prod_{t=1}^{T} \psi(y_{t-1}, y_t, \mathbf{x})} \tag{2.24}$$

where $\sum_{\mathbf{y}}$ is the sum over all possible $\mathbf{y}$ sequences. In the original definition of the CRF model, $\psi(y_{t-1}, y_t, \mathbf{x})$ is a parametric logistic function:

$$\psi(y_{t-1}, y_t, \mathbf{x}) = \exp\left(\sum_{kj} \chi_{kj} f_{kj}(y_{t-1}, y_t, \mathbf{x}) + \sum_{kp} \omega_{kp} g_{kp}(y_t, \mathbf{x}) \right) \tag{2.25}$$

where $1 \leq k \leq K$ and $1 \leq j \leq K$ are indexes ranging over the number of labels, $1 \leq p \leq P$ is an index over the number of features, $\chi = \{\chi_{11}, \chi_{12}, \ldots, \chi_{kj}, \ldots, \chi_{KK}\}$ are transition weights and $\omega = \{\omega_{11}, \omega_{12}, \ldots, \omega_{kp}, \ldots, \omega_{KP}\}$ are emission weights. The f_{kj} are called transition feature functions and the g_{kp} are called emission feature functions.

These feature functions compute the features of the observations and have to be user-specified. As it can been deducted by the argument $\mathbf{x}$ of these features functions, there is no independence assumption over the observations since the features can be constructed by using the whole observation sequence. The other two arguments are y_t and y_{t-1}, and the CRF therefore maintains the first-order Markov assumption over labels.

To illustrate how to set feature functions, we can for instance define an HMM-like CRF by using one emission feature function for each state-feature pair and one transition feature function for each state-state pair:

$$\begin{aligned} f_{kj} &= \mathrm{I}(y_t = k)\mathrm{I}(y_{t-1} = j) \\ g_{kp} &= \mathrm{I}(y_t = k)x_t^p \end{aligned}$$

where $\mathrm{I}()$ is the boolean indicator function evaluating to 1 if its argument is true and to zero otherwise.

2.4.1 Training

We now discuss how to estimate the parameters $\{\chi, \omega\}$ of the linear-chain CRF model depicted in Eq. (2.24). CRFs are trained by maximizing the conditional log-likelihood $\mathcal{L}(\chi, \omega)$:

$$\max_{\chi,\omega} \mathcal{L}(\chi, \omega) \tag{2.26}$$

$$= \max_{\chi,\omega} \log(p(\mathbf{y}|\mathbf{x})) \tag{2.27}$$

$$= \max_{\chi,\omega} \sum_{t=1}^{T} \sum_{kj} \chi_{kj} f_{kj}(y_{t-1}, y_t, \mathbf{x}) + \sum_{t=1}^{T} \sum_{kp} \omega_{kp} g_{kp}(y_t, \mathbf{x}) - \log(Z(\mathbf{x})). \tag{2.28}$$

This function cannot be maximized in closed form, so numerical optimization is used. Since it is a convex function, quasi-Newton methods or conjugate gradient optimization methods using only first-order derivatives are directly applicable. A review over methods for training CRF models can be found in [121].

The derivative of $\mathcal{L}(\chi, \omega)$ with respect to a transition parameter χ_{kj} is:

$$\frac{\partial \mathcal{L}}{\partial \chi_{kj}} = \sum_{t=1}^{T} f_{kj}(y_{t-1}, y_t, \mathbf{x}) - \sum_{t=1}^{T} \sum_{y,y' \in \mathbf{q}} f_{kj}(y, y', \mathbf{x}) \times p(y_t = y, y_{t-1} = y'|\mathbf{x}) \tag{2.29}$$

and with respect to an emission parameter ω_{kp}:

$$\frac{\partial \mathcal{L}}{\partial \omega_{kp}} = \sum_{t=1}^{T} g_{kp}(y_t, \mathbf{x}) - \sum_{t=1}^{T} \sum_{y} g_{kp}(y, \mathbf{x}) \times p(y_t = y|\mathbf{x}). \tag{2.30}$$

The first term in both derivatives is the expectation of f_{kj} or of g_{kp} under the data distribution (a feature count over training instances) and the second term, which arises from the derivative of $\log(Z(\mathbf{x}))$, is the expectation of f_{kj} or of g_{kp} under the model distribution. At the maximum likelihood solution, the derivatives are equal to zero and the two terms are equal.

The computation of the normalizer $\log(Z(\mathbf{x}))$ requires a summation over all possible label sequences and the number of possible sequences grows exponentially with sequence length. Nevertheless, the forward-backward algorithm can be used to reduce the computational cost from $O(K^T)$ to $O(TK^2)$, thanks to the first-order Markov assumption. In more details, $\log(Z(\mathbf{x}))$ is computed using using Eq. (2.10) and both $p(y_t = y, y_{t-1} = y'|\mathbf{x})$ and $p(y_t = y|\mathbf{x})$ are computed using Eq. (2.13) with

$$a_{kj} = \exp(\chi_{kj}) \tag{2.31}$$

$$b_k(\mathbf{x}_t) = \exp\left(\sum_p \omega_{kp} g_{kp}(y_t, \mathbf{x})\right) \tag{2.32}$$

$$p(\mathbf{x}) = Z(\mathbf{x}). \tag{2.33}$$

The total training cost is $O(TK^2G)$ where G is the number of iterations in the gradient ascent optimization.

2.4.2 Regularization

Conventional approaches to regularizing CRFs, and logit models in general, focus on adding (a constant times) the L_2 norm of the model parameters to the objective function as a penalty to avoid overfitting [111]. This is equivalent to performing maximum a posteriori estimation of the parameters assuming that each model parameter is drawn independently from a Gaussian prior [113]. Ignoring the terms that do not affect the parameters, the optimization of the CRF log-likelihood regularized with a Gaussian prior becomes:

$$\max_{\mathbf{w}} \mathcal{L}(\mathbf{w}) - \frac{1}{2}\sum_k \left(\frac{w_k - \mu_k}{\sigma_k}\right)^2 \tag{2.34}$$

where $\mathbf{w} = \{\chi, \omega\} = [w_i]_{i=1}^{K\times K + K\times P}$ is the vector of all parameters, μ_k and σ_k are the mean and the variance of parameter w_k. For simplicity, the μ_k are assumed zero and σ_k is held constant across all parameters:

$$\max_{\mathbf{w}} \mathcal{L}(\mathbf{w}) - \frac{1}{2}\sum_k \left(\frac{w_k - \mu_k}{\sigma_k^2}\right)^2 \tag{2.35}$$

$$= \max_{\mathbf{w}} \mathcal{L}(\mathbf{w}) - \frac{1}{2\sigma^2}\sum_k w_k^2 \tag{2.36}$$

$$= \max_{\mathbf{w}} \mathcal{L}(\mathbf{w}) - \lambda \|\mathbf{w}\|_2 \tag{2.37}$$

where λ is a hyper-parameter controlling the amount of regularization. In practice, the L_2 norm regularization penalizes solutions with large parameter values and therefore helps against overfitting.

In recent years, there has been a growing interest in the L_1-norm regularization, which is equivalent to a Laplacian prior on parameters [111]. This type of regularization enforces sparsity in the parameters and yields models that are more easily interpreted [87]. In particular, the L_1-regularized logistic regression model has proven to be very efficient [63]. CRFs can actually be cast as a multiclass logistic regression model with extra parameters for the first-order Markov dependencies between labels. For this reason, the L_1-regularization of the CRF model yields the same benefits and has been investigated with success [111]. Sparsity is especially useful in sequence models having two sets of parameters: transition parameters and emission parameters. The L_1 penalty indeed achieves feature selection by encouraging sparsity in the emission parameters and in addition leads to a sparse transition matrix, which is of particular importance to yield good performances. In particular, so-called left-to-right models consist in a special case where a given state only has one successor beside itself. In such situations, many coefficients of the transition matrix should be zero. However, the training of the standard CRF model can lead to non-zero coefficients for non-existent transitions. The addition of the L_1-regularization penalty enforces sparsity in the transition matrix and avoids such transitions in the inference process.

Ignoring the terms that do not affect the parameters, the optimization of the CRF log-likelihood regularized with a Laplacian prior becomes:

$$\max_{\mathbf{w}} \mathcal{L}(\mathbf{w}) - \frac{1}{2} \sum_k \frac{|w_k - \mu_k|}{\sigma_k} \tag{2.38}$$

where here μ_k is a location parameter and $\sigma_k > 0$ is a scale parameter. Assuming once again that the μ_k are zero and that σ_k is held constant across all parameters, we obtain:

$$\begin{aligned} & \max_{\mathbf{w}} \mathcal{L}(\mathbf{w}) - \frac{1}{2} \sum_k \frac{|w_k - \mu_k|}{\sigma_k} && (2.39) \\ = \quad & \max_{\mathbf{w}} \mathcal{L}(\mathbf{w}) - \frac{1}{2\sigma_k} \sum_k |w_k| && (2.40) \\ = \quad & \max_{\mathbf{w}} \mathcal{L}(\mathbf{w}) - \lambda \|\mathbf{w}\|_1 . && (2.41) \\ & && (2.42) \end{aligned}$$

It is more convenient in practice to minimize the negative regularized log-likelihood defined as

$$\min_{\mathbf{w}} f(\mathbf{w}) = \min_{\mathbf{w}} -\mathcal{L}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 . \tag{2.43}$$

The drawback is that the objective function $f(\mathbf{w})$ in Eq. (2.43) is no longer continuously differentiable for $w_i = 0$. Three techniques can be considered to overcome this problem [107]:

1. The first technique is to use a smooth approximation of the L_1 penalty term, the most famous being the *epsL1* function:

$$\text{epsL1}(w_i, \varepsilon) = \sqrt{w_i^2 + \varepsilon}. \tag{2.44}$$

   The difficulty arises in the choice of the additional meta-parameter ε. Moreover the obtained solution does not have a strict zero sparsity but rather very small values for the useless parameters. A threshold has thus also to be defined.

2. The second technique is to cast Eq. (2.43) as a constrained optimization problem by replacing λ with a variable ρ proportional to $1/\lambda$ and solve the constrained problem

$$\min_{\mathbf{w}} -\mathcal{L}(\mathbf{w}) \;\; \text{s.t.} \, \|\mathbf{w}\|_1 \leq \rho. \tag{2.45}$$

   An efficient method to solve this constrained problem for the binary logistic regression is the IRLS-LARS algorithm [63]. Nevertheless, this algorithm is only applicable to objective functions that yield an iterative reweighted least-square update. Despite its similarity with the logit, there is to our knowledge no established IRLS formulation for the objective function of the CRF model.

3. The third technique is the use of sub-gradients to extricate the task of dealing with the non-differentiable gradients [107]. The sub-gradient at a point of non-differentiability is defined as the interval by the derivatives at the limit of each side of that point [15]. At a local minimizer $\tilde{\mathbf{w}}$ of Eq. (2.43), we have the following optimality conditions:

$$\begin{cases} \nabla_i \mathcal{L}(\tilde{\mathbf{w}}) + \lambda \text{sign}(\tilde{w}_i) = 0, & |\tilde{w}_i| > 0 \\ \nabla_i \mathcal{L}(\tilde{\mathbf{w}}) + \lambda \{-1, 1\} = 0, & \tilde{w}_i = 0 \end{cases} \tag{2.46}$$

   with $\nabla_i \mathcal{L}(\mathbf{w}) = \frac{\partial \mathcal{L}(\mathbf{w})}{\partial w_i}$. The second optimality condition comes from the non-differentiability of the absolute value function when its argument is zero. In this case, the sub-gradient is used. Note that Eq. (2.46) is equivalent to

$$\begin{cases} \nabla_i \mathcal{L}(\tilde{\mathbf{w}}) + \lambda \text{sign}(\tilde{w}_i) = 0, & |\tilde{w}_i| > 0 \\ -\lambda \leq \nabla_i \mathcal{L}(\tilde{\mathbf{w}}) \leq \lambda, & \tilde{w}_i = 0. \end{cases} \tag{2.47}$$

From these conditions, the gradient for each w_i computed during the optimization process is

$$\nabla_i f(\mathbf{w}) = \begin{cases} \nabla_i \mathcal{L}(\mathbf{w}) + \lambda \text{sign}(w_i), & |w_i| > 0 \\ \nabla_i \mathcal{L}(\mathbf{w}) + \lambda, & w_i = 0, \nabla_i \mathcal{L}(\mathbf{w}) < -\lambda \\ \nabla_i \mathcal{L}(\mathbf{w}) - \lambda, & w_i = 0, \nabla_i \mathcal{L}(\mathbf{w}) > \lambda \\ 0, & w_i = 0, -\lambda \leq \nabla_i \mathcal{L}(\mathbf{w}) \leq \lambda. \end{cases} \tag{2.48}$$

In this work, we consider the sub-gradient strategy since it has the advantage of using standard optimization tools and does not require the setting of any additional meta-parameters.

2.4.3 Inference

CRFs maintain the first-order Markov assumption over labels, and the predicted label sequence $\mathbf{y}^*$ can thus still be obtained by the forward-backward algorithm [113], with

$$a_{kj} = \exp(\chi_{kj}) \tag{2.49}$$

$$b_k(\mathbf{x}_t) = \exp\left(\sum_p \omega_{kp} g_{kp}(y_t, \mathbf{x})\right). \tag{2.50}$$

On the other hand, the independence assumption between the observations is not required for tractable inference like in HMMs (as detailed in Sec. 2.3.4). As a result, CRFs offer complete freedom over the choice of the features of the input, while the inference can still be achieved with the same time complexity as in HMMs.

2.5 Continuous wavelet transform

In this section, the continuous wavelet transform (CWT) is introduced. The CWT will later be used to build features from the ECG signal. The CWT of a signal x_t produces a time-frequency decomposition of this signal by the convolution with a so-called *wavelet function* ψ_t [73].

From a chosen wavelet function, one can obtain a family of time-scale waveforms by translation and scaling

$$\psi_t^{a,b} = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \tag{2.51}$$

where $a > 0$ is the scale (or dilatation) factor and is b the translation factor. When $a = 1$ and $b = 0$, the wavelet is called the *mother wavelet*.

The wavelet transform of a signal x_t is a projection of this signal on the wavelet $\psi_t^{a,b}$:

$$T(a,b) = \int_{-\infty}^{+\infty} x_t \psi_t^{a,b} dt. \tag{2.52}$$

For each a, the wavelet coefficients $T(a,b)$ are signals (that depend on b) which represent the matching degree between the wavelet $\psi_t^{a,b}$ and the analyzed signal x_t. Despite its name, the continuous transform is applied to discrete signals. The discretizations of the continuous wavelet transform involve a discrete approximation of the transform integral (i.e. a summation) computed on a discrete grid of a scales and b locations. In practice, the location parameter is usually discretized at the sampling interval and the scale parameter is discretized logarithmically. What is continuous about the CWT, and what distinguishes it from the discrete wavelet transform (DWT) which will later be introduced in Sec. 5.3.3, is the set of scales and positions at which it operates. Unlike the discrete wavelet transform which requires dyadic scales and locations, the CWT can operate at every scale, from that of the original signal up to some maximum scale (defined by the sampling frequency). The CWT is also continuous in terms of shifting: during computation, the analyzing wavelet is shifted smoothly over the full analyzed signal.

The CWT is a particularly useful tool to represent the ECG signal [3]. First, the multi-scale feature of the CWT allows to separate the different characteristic ECG waveforms over noise, baseline drift, and artifacts. Next, the time-frequency aspect preserves the important time course of the non-stationary ECG signal. Finally, efficient implementations of the algorithm exist and a low computational complexity is required, allowing real-time analysis. Figure 2.3 shows the CWT coefficients computed on a 10 second ECG signal for a large number of scales. The darker the color, the smaller the coefficient. As the figure illustrates, the CWT at small scales matches high-frequency patterns such as QRS waves. To the opposite, the higher scales, corresponding to a higher dilation of the mother wavelet, start to match lower frequency contents.

2.6 Application to ECG signals

In this section, the framework for segmentation of the ECG signal using probabilistic models such as HMMs and CRFs is introduced. Next, two experiments are conducted on real Holter ECG data and the results are presented.

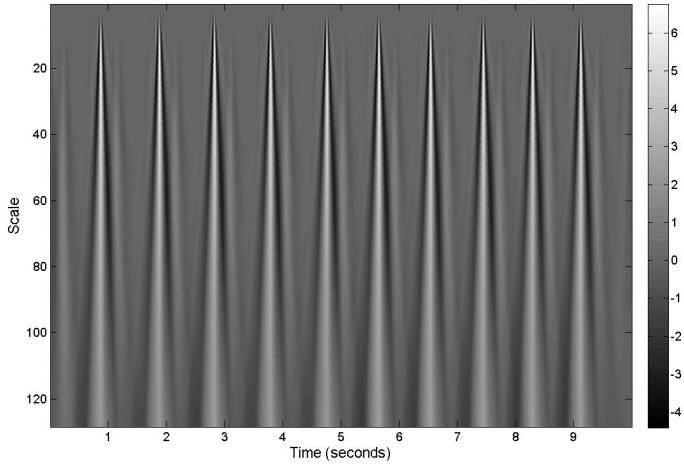

Figure 2.3: Coefficients of the CWT of a 10 second ECG signal for scales between 1 and 128. The darker the color, the smaller the coefficient. The CWT at small scales matches high-frequency patterns such as QRS waves. The higher scales, corresponding to a higher dilation of the mother wavelet, start to match lower frequency contents.

2.6.1 Methodology

The ECG signal can be viewed as the result of a generative process, where each waveform is generated by a particular state of the heart (the atria, the ventricles, etc), this state being hidden to the observer. The cardiological process is sequential, and each state is solely dependent on the previous state. Probabilistic models for sequential data such as HMMs, HSMMs and CRFs relying on the first order Markov assumptions are thus naturally appropriate for modeling this kind of process.

In order to represent the dynamic of the ECG sequence, five different states are defined in the model: (1) P wave, (2) baseline 1, (3) QRS or N wave, (4) T wave and (5) baseline 2. Figure 2.4 shows the graphical structure of the CRF and HMM in the frame of ECG segmentation and Fig. 2.5 shows a real ECG beat and the associated state sequence.

Four models are compared: HMMs, HSMMs, CRFs and L_1-regularized CRFs (CRF+L_1). For this purpose, two experiments are conducted:

1. In the first experiment, the performances of HMMs, HSMMs and CRFs are compared using the raw unfiltered ECG values as features.

2. It has been shown in [49] that the use of features from the CWT of the ECG

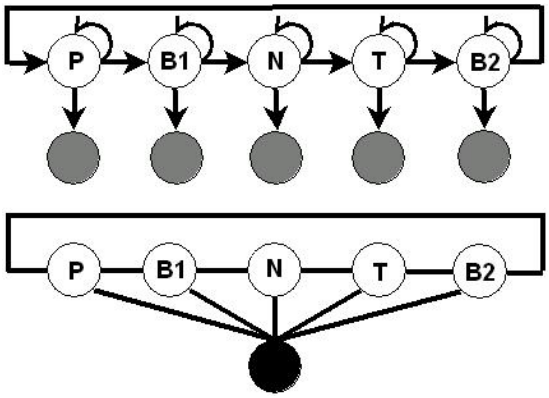

Figure 2.4: Graphical representation of a hidden Markov model (top) and a conditional random field (bottom) for ECG segmentation. Five different states corresponding to the ECG characteristic points are defined: (1) P wave, (2) baseline 1, (3) QRS or N wave, (4) T wave and (5) baseline 2.

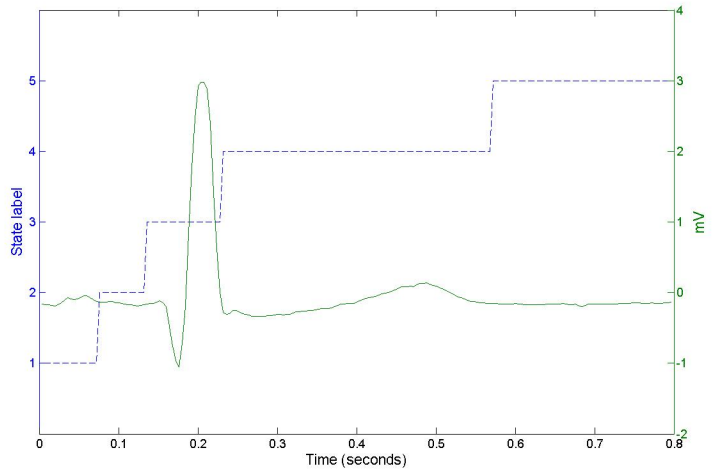

Figure 2.5: A real ECG beat (straight line) and the associated state sequence (sharped line). Five different states corresponding to the ECG characteristic points are defined: (1) P wave, (2) baseline 1, (3) QRS or N wave, (4) T wave and (5) baseline 2.

signal computed with an order two Coiflet mother wavelet and the first seven dyadic decomposition scales significantly improve the performances of HMMs. For this reason, in the second experiment, the performances of the four models are be compared using such features. More formally, the features included in the CRF model are

$$f_{kj}(y_{t-1}, y_t, \mathbf{x}) = \begin{cases} 1 \text{ if } y_{t-1} = k \text{ and } y_t = j, & 1 \leq k, j \leq K \\ 0 \text{ otherwise} \end{cases} \tag{2.53}$$

$$g_{kp}(y_t, \mathbf{x}) = \begin{cases} T(2^p, t) \text{ if } y_t = k, & 1 \leq k \leq K, 1 \leq p \leq P \\ 0 \text{ otherwise} \end{cases} \tag{2.54}$$

and in the HMM

$$x_t^p = T(2^p, t), \quad 1 \leq p \leq P \tag{2.55}$$

where P is the number of wavelet scales ($P = 7$) and K is the number of states in the model ($K = 5$). Remember from Eq. (2.52) that $T(a,b)$ represents the CWT coefficients for a given scale a and translation b.

The use of CWT features has two main practical advantages. First, it acts as a filter which separates noise and useful information within the signal. The performances of each model should therefore be improved. Second, the CWT creates time-dependent features. The CWT at a given location and scale can indeed be seen as a correlation measure between the chosen wavelet, dilated to a custom length (the scale parameter), and a portion of the signal of the same length centered at the time step corresponding to the location parameter. The CWT of the ECG signal at a given observation hence requires the use of surrounding observations and therefore theoretically violates the naive Bayes independence assumptions of the HMM and HSMM. On the other hand, the features in CRFs can be constructed by using the whole observation sequence, hence no independence assumption is made over observations. Combining these two benefits from the CWT, it is expected that each model provides better results than on the raw signal values, but CRFs should yield better results than HMMs since none of their assumptions has been violated.

2.6.2 Experiments

The two experiments are conducted on the public MIT-BIH QT database from Physiobank [44], designed for the evaluation of algorithms that detect waveform boundaries in the ECG. Over 100 excerpts of two-leads Holter ECG recordings have manually been annotated by cardiologists with waveform boundaries for 30 to 50 selected beats

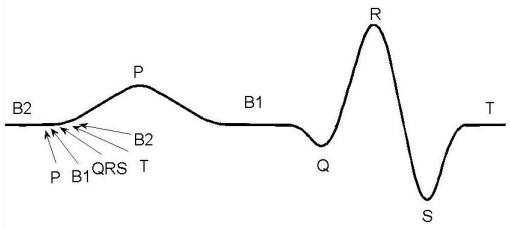

Figure 2.6: Example of a double beat inserted into a P wave. Wrong additional waves strongly alter Q-T, R-R or other statistics over beat labels which may lead to incorrect diagnosis.

in each recording. A broad variety of patients, pathologies and ECG signal morphologies are represented. All recordings are sampled at 250 Hz and have 11-bit resolution over a 10 mV range.

Each ECG recording has between 30 and 50 annotated beats. For each recording, 80% of the annotated beats are used as a training set to estimate the model parameters and the remaining 20% of the annotated beats are used as a test set for performance evaluation. In the case of the HSMM and HMM models, a Gaussian distribution is used for the state emission probabilities. A gamma distribution is used for the state durations of the HSMM. The transition matrix and the parameters of the gamma and Gaussian distributions are estimated from the available annotated training samples. The training of the CRF parameters is achieved on the same annotated samples using a subspace trust-region optimization method based on the interior-reflective Newton method described in [23] and [22]. In the case of the L_1-regularized CRFs, the regularization parameter is estimated using cross-validation on the training set. The forward-backward algorithm is used to compute the most likely sequence of states on the test set, with the initial state probability vector defined as $\pi = [1,0,0,0,0]'$, so that the model always starts in the first state (the P wave).

The performances of the models for the segmentation of normal and pathological signals are separately investigated. For this reason, two distinct datasets are extracted from the wide variety of recordings in the Physiobank QT database. The first dataset contains 10 recordings each from a distinct patient with normal sinus rhythm and no significant arrhythmias. The second dataset also contains 10 recordings each from a distinct patient but with pathologies such as arrhythmia, S-T changes and supraventricular arrhythmia. The performances of the models can therefore be evaluated in both normal and pathological situations.

Three performance measures are reported:

1. The first performance measure is the wave segmentation accuracy. It is defined

as the percentage of points being classified in the correct state.

2. The second performance measure is the wave apparition rate (later referred to as double wave rate). This measure quantifies the percentage of waves that the model incorrectly infers. The wave apparition rate is of particular importance because statistics over the time intervals between characteristic ECG points such as Q-T interval, R-R interval or S-T intervals are often used as a marker of pathologies [56, 20]. Degenerative additional waves may only slightly decrease the accuracy, but will strongly alter these statistics. For example, a complete double beat inserted into a P wave is illustrated in Fig. 2.6.

3. The third performance measure concerns the CRF model. The percentage of parameter values below 10^{-5} is reported. This percentage reveals the number of parameters, hence of features, which are useless to achieve high performances. Spurious features can harm the classifier. It can indeed be expected that the segmentation performances degrades when wavelet scales corresponding to noise and baseline drift are not rejected. Moreover, including the set of all consecutive dyadic scales from one to seven in the model may lead to redundancy. It is therefore expected that better performances are achieved by the CRF+L_1 than the unregularized CRF since it encourages sparsity in the parameters.

2.6.3 Results

The results of the four models using the raw ECG signal values and the wavelet coefficients as features are reported in Table 2.2 and in Table 2.3 respectively. We first analyze the results on the raw ECG values. The performances of the four models are unsurprisingly worst on arrhythmic recordings than on sinus rhythm recordings. Nevertheless, the two CRFs seem to be less impacted by arrhythmic events than the HMM and the HSMM. The loss in accuracy is indeed around 5% for the two CRFs and up to 10% for the HMM and the HSMM. It can be observed that the two CRF models significantly outperform the HMM and the HSMM in terms of accuracy, for both types of recordings.

In terms of double wave rate, the HSMM seems to benefit from the modeling of durations and obtains a better double wave rate than both CRF models and the HMM. In particular, the HMM clearly suffers from the independence assumption and achieves a double wave rate of around 60% for both types of recordings, which is clearly unacceptable. It is also worth mentionning that the L_1-regularization of the CRF objective function increases the accuracy and decreases the double wave rate on both types of recordings. Moreover, around 70% of the parameters are pruned. This pruning reveals that several parameters are useless to achieve high segmentation

Model	Dataset	Accuracy	Double waves	Pruning
HMM	Sinus	66.65%	62.00%	0.00%
HSMM	Sinus	67.16%	9.33%	0.00%
CRF	Sinus	76.89%	17.34%	37.00%
CRF+L_1	Sinus	77.67%	14.67%	72.00%
HMM	Arrhythmic	56.42%	63.45%	0.00%
HSMM	Arrhythmic	57.22%	16.55%	0.00%
CRF	Arrhythmic	71.14%	26.67%	38.00%
CRF+L_1	Arrhythmic	72.77%	22.82%	71.00%

Table 2.2: Segmentation results using unfiltered ECG values as features on sinus rhythm recordings and arrhythmic recordings.

performances and actually decrease the performances if left not rejected.

We now analyze the results of the models using the CWT coefficients as features. Here again, each model yields worst results on arrhythmic recordings than on sinus rhythm recordings, as can be expected. Compared to previous results using raw ECG values, the CWT clearly increases the performances of each model and especially those of the HMM, as previously reported in the literature [49]. On sinus rhythm recordings, the HMM and the HSMM outperform the CRF both in terms of accuracy and double wave rate. Nevertheless, when the L_1-regularization is added to the CRF, the performances increase significantly and the regularized CRF achieves the best overall performances with an accuracy of 95% and 0% of double wave rate. The same conclusions can be observed in arrhythmic recordings where the regularized CRF yields the best accuracy and a double wave rate equal to the one of the HSMM. On the other hand, the HMM once again obtains a unsatisfactory double wave rate of 36%. In both types of recordings, more than 75% of the parameters are pruned by the L_1-regularization.

2.7 Discussion

In this chapter, the methodology for automatically segmenting ECGs with the CRF model is presented and its theoretical advantages over standard HMMs approaches are emphasized. In particular, due to its discriminative form and the undirected nature of its graphical representation, the CRF model relieves the strong naive Bayes assumption made in HMMs. Moreover, the CRF model can achieve feature selection in an embedded manner by adding a L_1-norm regularization to its objective function. In sequence models such as CRFs, the L_1-regularization is especially interesting since

Model	Dataset	Accuracy	Double waves	Pruning
HMM	Sinus	94.33%	4.00%	0.00%
HSMM	Sinus	93.09%	0.00%	0.00%
CRF	Sinus	89.74%	4.60%	50.50%
CRF+L_1	Sinus	95.29%	0.00%	77.00%
HMM	Arrhythmic	89.33%	35.85%	0.00%
HSMM	Arrhythmic	88.47%	4.00%	0.00%
CRF	Arrhythmic	87.57%	9.67%	51.00%
CRF+L_1	Arrhythmic	89.99%	4.04%	75.25%

Table 2.3: Segmentation results using CWT coefficients as features on both sinus rhythm recordings and arrhythmic recordings.

it enforces sparsity in the transition parameters in addition to the feature parameters. Not existing transitions between states are therefore much more likely to be avoided in the inference process.

The performances of the CRF model and of the regularized CRF model are compared to those of the state-of-the-art segmentation model with HMMs and the HSMMs variant. For this purpose, two experiments are conducted on both sinus rhythm and arrhythmic Holter recordings. The results show that the CRF seems to be much less impacted by arrhythmic events than the HMM and the HSMM. The results also confirm the hypothesis that the naive Bayes assumption made in HMMs and HSMMs leads to a significant loss in performance for the segmentation of ECG signals. The proposed L_1-regularized CRF model does not require such assumptions and outperforms state-of-the-art HMMs and HSMMs on both sinus rhythm and arrhythmic recordings.

In our experiments, the regularization of the CRF by the L_1-norm increases the accuracy and decreases the double wave rate in all situations. Furthermore, a much larger number of model parameters are pruned out. These results confirm the benefit of the L_1-norm regularization in CRFs and reveal that several features and transition parameters may actually decrease the performances if left not rejected.

The accuracy and the double wave rate on pathological recordings can probably be further improved by considering a semi-Markov variant of the CRF model. Semi-Markov variants of the CRF model have been investigated previously [105] for speech processing, but their ability to model the ECG sequence remains to be investigated. In this work, we have assumed that the expert segmentations were perfect and reliable. However in practice, the wave boundary measurements in the training set as defined by the experts are not always located in the correct positions. These errors in the training set may adversely influence the resulting parameters of standard supervised learning models like CRFs and HMMs. Previous work have investigated the use of a

semi-supervised learning algorithm for HMM, based upon the Baum-Welch algorithm, which can be used to adjust the HMM parameters from subjectively labeled ECG data [48]. Although semi-supervised learning does not truly address the label noise issue itself, the term semi-supervised was used by the authors because labels at the boundary of two waves are ignored and estimated during the expectation step of the Baum-Welch algorithm. Their experiments demonstrate that the proposed algorithm improves the estimation of the Q-T interval measurements. It is therefore expected that including such uncertainties over the labels in the CRF model would also improve the results achieved in our experiments.

Taub: “The deodorant has a high proportion of propylene glycol. Same stuff made a kid in Singapore develop a heart condition, and get seizures. Our patient may have never needed split-brain surgery.”
House: “I’m sure he’ll half appreciate the irony.”

House MD, season 5, episode 24

Chapter 3

Heart Rate Variability Metrics for Epileptic State Identification

Contents

3.1 Introduction

The variation over time of the heart rate is conventionally described as the heart rate variability (HRV). In normal situations, the heart rate is determined by the sino-atrial (SA) node which acts as the heart's primary pacemaker. The SA node has an intrinsic firing rate, but it can be modulated by the central nervous system, and specifically by the autonomic nervous system (ANS) with central nuclei in the brain stem and the hypothalamus. The ANS is subdivided in two mutually opposing systems connected to the periphery through the sympathetic (SNS) and parasympathetic (PNS) nerves respectively. The balancing action of these two systems regulates the activity of the SA node. The SNS acts on the SA node's firing rate as an accelerator, and is activated during stress or exercise in a time range from 12 to 17 seconds. Conversely, the PNS slows the firing rate via the vagus nerve, for example during food digestion within a time range from 1 to 3 seconds. The ANS thus controls the basic heart rate but the corresponding neural activity is even better reflected in the beat-to-beat variations of the heart rate.

A vast literature has been published about HRV metrics for quantifying beat-to-beat variations together with their varying success for discerning specific disease conditions. In particular, the clinical use of HRV is now established for the risk assessment after acute myocardial infarction and for the assessment of diabetic neuropathy [1]. The reliability of the suggested methods highly relies on a correct beat identification from the digitalized ECG signal. Because the ventricles contain more muscle mass than the atria, the QRS complex is the most visible ECG pattern and is therefore conventionally used as a beat marker. Several methods exist for computer-aided automatic segmentation of the ECG and have been presented in Chapter 2. Once the R spikes have been marked using such algorithms, the successive time values of the intervals between them (R-R, sometimes called N-N for normal to normal) form the data series used for estimation of the HRV metrics.

In 1996, a Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology published standards for the measurement, interpretation and use of HRV metrics [1]. New experiments have since raised doubts over the interpretation of some of the reference metrics. Moreover, the computation of several reference metrics have been proven unreliable but these metrics still remain commonly used without precautions. More powerful metrics have also been proposed more recently for new applications with varying success. In this chapter, standards for the measurement, interpretation and use of heart rate variability are presented in a form directed to clinicians. This critical overview discloses the pros and cons of state of the art metrics and describes why some previous methods for processing the ECG signal are inadequate.

Epileptic seizures are associated with several changes in autonomic functions [12]. It is also now demonstrated that functional electrical stimulation of the parasympathetic vagus nerve is a successful treatment for refractory epilepsy [43]. Nevertheless, although the detection and prediction of epileptic seizures using EEG recordings is a widely studied research topic, very few authors investigated the possible correlation of the variations in heart rate and epileptic seizures. In this chapter, the possible use of HRV metrics in the frame of epileptic state identification is also investigated, including the controversial pre-ictal state. The materials in this chapter have been submitted for publication in [29].

3.2 HRV metrics

Assume the time indexes t corresponding to R spike positions are detected in a sampled ECG signal x_t containing N identified beats. We define $\mathbf{r} = [r_n]_{n=1}^{N}$ as the vector containing the time positions of the R spikes. The R-R interval series $\mathbf{rr} = [rr_n]_{n=1}^{N-1}$ is then formed as $rr_n = r_{n+1} - r_n$. Figure 3.1 illustrates the extraction of the **rr** series. There are thus $N-1$ R-R intervals that are computed from N beats. For convenience, in the remaining of this chapter we use N to refer to the number of beat intervals rather than to the number of beats. HRV metrics can be grouped in three categories: linear time domain metrics, non-linear time domain metrics and frequency domain metrics.

3.2.1 Linear time domain metrics

Time domain metrics can be subdivided into two categories: statistical measures and geometrical measures. Statistical measures are mostly based on the computation of the standard deviation of the R-R interval series (SDNN - standard deviation of the R-R intervals) or of the difference between consecutive R-R intervals (SDSD - standard deviation of the differences between successive R-R intervals).

$$SDNN = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(rr_n - \mu)^2} \tag{3.1}$$

$$SDSD = \sqrt{\frac{1}{N-1}\sum_{n=2}^{N}(rr_n - rr_{n-1} - \theta)^2} \tag{3.2}$$

with

$$\mu = \frac{1}{N}\sum_{n=1}^{N} rr_n \tag{3.3}$$

$$\theta = \frac{1}{N-1}\sum_{n=2}^{N} rr_n - rr_{n-1}. \tag{3.4}$$

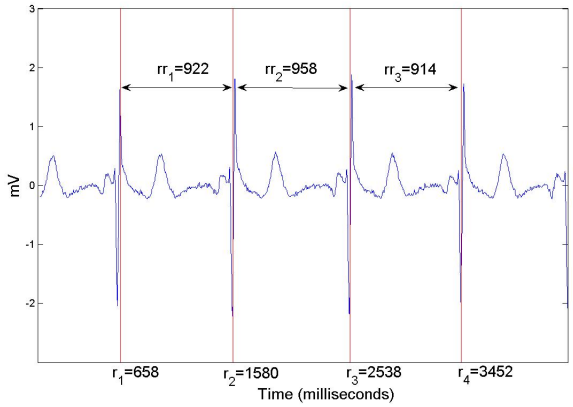

Figure 3.1: Because the ventricles contain more muscle mass than the atria, the QRS complex is the most visible pattern on the ECG signal and therefore conventionally serves as beat marker. Once the timings r_n of the N identified R spikes have been extracted from the sampled ECG signal, the time durations between them form the **rr** series used for estimation of the HRV metrics.

Another popular time domain metric counts the number of interval differences that are larger than 50ms and expresses the result as a percentage over the total number of beats analyzed (pNN50 - percentage of R-R intervals greater than 50 msec).

Geometrical methods roughly estimate the shape of the R-R interval distribution histogram built from the data series. The shape of the histogram is estimated by dividing the total number of R-R intervals with the number of R-R intervals in the modal bin (TI - triangular index) [1]. In essence, geometrical metrics are simply an attempt to estimate higher-order statistics of the R-R interval distribution such as the skewness and the kurtosis. With the more recent introduction of entropy-based methods [100], the geometrical metrics have faded out of the literature.

Time domain metrics prove to be powerful indicators of pathologies such as premature ventricular contraction, sick sinus syndrome, atrial fibrillation, complete heart block, left bundle branch block and ischemic cardiomyopathy, among others [2].

3.2.2 Non-linear time domain metrics

It has been suggested that in normal physiological situations the heart is not a periodic oscillator [94]. Moment statistics used in linear metrics may therefore not reveal all of the variability of the heart rate series. For this reason, non-linear metrics have more recently been introduced in the field of HRV. Non-linear methods are based on the

theory of chaos and fractal dimensions in order to quantify the entropy or complexity in a time-series.

Entropy relates to dynamical systems and represents the rate of information production. Usual methods for estimating the entropy of a system are not well suited to analyze short-term and noisy data such as R-R interval series [94, 95]. For this reason, the approximate entropy (ApEN) [95] and sample entropy (SampEN) [100] methods have been proposed for the assessment of HRV. ApEN and SampEN are regularity statistics quantifying the unpredictability and complexity of the heart rate time-series. These methods are based on the reasoning that the presence of repetitive patterns of fluctuation (i.e. a sinus wave) in a time-series renders it more predictable than a time-series in which such patterns are absent (i.e. a random time-series).

Both ApEN and SampEN rely on the correlation dimension. Given two fixed positive integer parameters m and d, we embed $\mathbf{rr}$ in a m-dimensional phase space[1] by constructing $N-m+1$ series $\mathbf{u}_j = [rr_j, rr_{j+1}, \ldots, rr_{j+m-1}] \in \mathbb{R}^m$. We define $b_j(d)$ as the number of vectors $\mathbf{u}_g$ such that $\text{dist}(\mathbf{u}_j, \mathbf{u}_g) \leq d$, dist() being the Euclidean distance function. The quantity $C_j(d,m)$ is computed as

$$C_j(d,m) = \frac{b_j(d)}{N-m+1}, \quad 1 \leq j \leq N-m+1 \tag{3.5}$$

and corresponds to the probability that any vector $\mathbf{u}_g$ is within d of $\mathbf{u}_j$. The quantity

$$\Phi(d,m) = \frac{\sum_{j=1}^{N-m+1} \log C_j(d,m)}{N-m+1} \tag{3.6}$$

is the average of the logarithm of the $C_j(d,m)$. The entropy of the underlying process can be approximated by the correlation dimension defined as

$$\lim_{d\to 0} \lim_{m\to\infty} \lim_{N\to\infty} [\Phi(d,m) - \Phi(d,m+1)]. \tag{3.7}$$

Because of the limits, this definition is not well suited to the analysis of noisy and finite time-series such as R-R intervals. Nevertheless, Pincus [95] observed that the calculation of $\Phi(d,m) - \Phi(d,m-1)$ had intrinsic interest as a measure of regularity and complexity in numerous situations and proposed the ApEN metric for quantifying HRV as

$$ApEN(m,d) = \Phi(d,m) - \Phi(d,m-1). \tag{3.8}$$

It has later been demonstrated that ApEN is a biased statistic [100]. For this reason, Richman [100] later proposed the SampEN metric which is closely related to ApEN as

$$SampEN(m,d) = -\log(\Phi(d,m)/\Phi(d,m+1)). \tag{3.9}$$

[1]The phase space embedding is an established technique for visualizing the dynamical behavior of a process over time, and relates to the notion of *attractors* in physics (see [114] for details).

There are two differences between ApEN and SampEN. First, during the computation of $C_j^m(r)$, the comparison with $\mathbf{u}_j$ itself is removed in SampEN. Second, SampEN decreases monotonically when d increases. In practice, this means that ApEN is a biased statistic and estimates values below those predicted by the theory; it has indeed been shown that SampEN agrees much more closely with the expected result than ApEN does, and this, over a broad range of conditions [100].

Another method for quantifying the fractal scaling property of R-R interval signals is Detrended Fluctuation Analysis (DFA) [94]. This technique is a modification of the root-mean-square analysis of random walks applied to non-stationary signals. The **rr** series is first integrated to form a new series $\mathbf{z} = [z_n]_{n=1}^N$ as

$$z_n = \sum_{c=1}^{n} rr_c - \mu. \tag{3.10}$$

For a given window size q, the integrated series $\mathbf{z}$ is then divided into windows of equal length q and a least-square line representing the local trend is fit to each window. The root-mean-square fluctuation $F(q)$ of the integrated and detrended series is computed as

$$F(q) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}[z_n - z'_n)]^2} \tag{3.11}$$

where z'_n corresponds to the value of z_n on the associated fitted straight line segments. This computation is repeated over all possible window lengths to provide a relationship between $F(q)$, the average fluctuations, and the window length q. Typically, $F(q)$ will increase when the window size also increases. The DFA metric value is then estimated by the slope of the line relating the logarithm of $F(q)$ to the logarithm of q. The DFA metric for healthy subjects is close to one, and this value is significantly different for various types of cardiac abnormalities such as severe heart failure or also in children with disordered breathing during sleep [94].

3.2.3 Frequency domain metrics

Since its first introduction in 1981 [6], power spectral density (PSD) has become one of the most popular method for HRV analysis. Spectral power of the R-R series is computed using conventional PSD estimation methods such as parametric models (AR - autoregressive) or non parametric models (FFT - fast Fourier transform). However, for a correct evaluation, these PSD estimation methods require a constant sampling rate, which is not the case with the R-R interval series (each R-R interval is of different duration). A suggested solution is to resample the series at frequencies between 2 and

10 Hz by linear or cubic interpolation as if the interval values were taken from a continuous signal [6, 1]. This interpolation has however been shown to yield incorrect results in the power spectrum estimation and the resulting error can affect the results quite significantly [61, 21]. The non-parametric Lomb-Scargle [71, 106] method for spectral analysis of unequally spaced samples, which was originally designed for the analysis of astronomical data, has therefore more recently been introduced in the field of HRV analysis [61].

More formally, the generalized N-point discrete Fourier transform of the series **rr** with non constant associated time indexes **r** leads to the following expression for the normalized periodogram:

$$p(\omega) = \frac{1}{\sigma^2}\left(\frac{[\sum_n (rr_n - \mu)cos(\omega(r_n - \tau))]^2}{\sum_n cos^2(\omega(r_n - \tau))} + \frac{[\sum_n (rr_n - \mu)sin(\omega(r_n - \tau))]^2}{\sum_n sin^2(\omega(r_n - \tau))}\right) \quad (3.12)$$

$$\tau = tan^{-1}\left(\frac{\sum_n sin(2\omega r_n)}{2\omega\sum_n cos(2\omega r_n)}\right) \quad (3.13)$$

where ω is the frequency of interest and σ^2 is the variance of the **rr** series. The τ value is an offset that makes $p(\omega)$ independent of shifting of the time indexes by any constant. See [61] or [21] for details on mathematical derivations. For even sampling, i.e. $rr_n = rr_m, \forall m,n$, Eq. (3.13) reduces to the classical discrete Fourier transform. Unfortunately, and probably in order to comply with the Task Force Standard described above, most papers still ignore this issue and use resampled PSD estimators instead of the Lomb-Scargle method.

The standard procedure then goes on splitting the frequency spectrum into four consecutive bands:

1. Ultra low frequency (ULF): from 0.0001 Hz to 0.003 Hz;
2. Very low frequency (VLF): from 0.003 Hz to 0.04 Hz;
3. Low frequency (LF): from 0.04 Hz to 0.15 Hz;
4. High frequency (HF): from 0.15 Hz to 0.4 Hz.

The Task Force recommends normalizing the energy in the ULF, LF and HF bands by subtracting the energy in the VLF band and dividing by the energy in the complete spectrum. The result is then multiplied by 100 to obtain the normalized metric value. The values obtained in the VLF and ULF bands have been associated with a wide range of physiological factors. Their interpretation is very controversial and these bands are therefore most often not taken into account for the analysis [1].

Conversely, several experiments using either passive head-up tilt tests in humans [83] or chronic destruction of sympathetic fibres in rats [17] tend to demonstrate that the HF band reflects PNS activity only and that it shows synchrony with the respiration rate. The interpretation of the LF band is more controversial. Some authors found that it is reflecting SNS activity alone [74] or link it to a mix of PNS and SNS activity [83]. Others conclude that the LF band does not reflect any ANS activity but rather baroflex function [81]. On the whole, the sympatho-vagal balance is commonly accepted to be estimated by the value of the energy ratio between the LF and HF bands.

Frequency domain metrics are successful in the diagnostic of cardiac pathologies and non-myelinated fiber neuropathies, for example in renal failure and diabetes [2]. They have furthermore been found successful in some clinical applications where time domain metrics failed such as chronic heart failure [2].

3.3 HRV and epilepsy

Epileptic seizures are associated with several changes in autonomic functions, which lead to cardiovascular, respiratory, gastrointestinal, cutaneous and urinary or sexual manifestations during or soon after the seizure [12]. Some are secondary to seizure manifestations (tachycardia related to motor activity by instance) when others are ictal manifestations directly triggered by the epileptic discharge itself. In particular, the analysis of cardiovascular changes has recently gained interest. Indeed, several studies suggest the possible use of ECG processing algorithms for the detection [46] and perhaps for the prediction [53] of seizures with promising results.

Stimulation of the vagus nerve has become a recognized adjunctive treatment for refractory epilepsy [43]. However, the exact mechanisms involved in this therapeutic action remain unknown. One of the main functions of the vagus nerve is the parasympathetic innervations of the sino-atrial node. The variability in heart rate contains information about the autonomic activity and a study of this parameter appears to be an appropriate non invasive method to answer questions such as the possible correlation of the autonomic activity and epileptic symptoms. For these reasons, a few studies have evaluated heart rate variability (HRV) in epileptic patients and differences in HRV metrics were observed when compared to a healthy reference group [40, 117].

Another interesting application of HRV metrics is the observation of autonomic variations between the different states of an epileptic seizure. Such variations may reveal information about the possible role of the autonomic system in epileptic seizures. An epileptic patient can find him or herself in one of the following three states: (1) the 'healthy' inter-ictal state, where no epileptic activity is observed clinically, (2) the ictal state, corresponding to a seizure and (3) the post-ictal state, where the brain is progressively recovering from an epileptic event.

The hypothetical existence of a fourth state, the pre-ictal state, is very controversial. It is defined as a transitional state between the inter-ictal and the ictal state. From the early 1970s on, numerous approaches have been attempted to detect such a state using advanced EEG signal processing methods [39]. While the automatic detection of ictal events in the EEG signals has reached good yield and accuracy, the detection of the pre-ictal state remains elusive or poorly convincing [85].

While comparing the pre-ictal, ictal, and the post-ictal states, time and frequency domain HRV changes were observed in [120]. Since then, a Task Force published recommendations for the assessment of HRV and new HRV metrics such as non-linear metrics gained popularity [1, 2]. Furthermore, the way frequency domain metrics were computed has been found unreliable and a superior method has since been introduced [61, 21]. The pertinence of the state-of-the-art HRV metrics in the frame of epileptic state identification (including the controversial pre-ictal state) therefore remains to be evaluated.

In the past, and until recent years, some studies have suggested that cardiac sympathetic and parasympathetic functions may be lateralized. Sympathetic representation would be lateralized to the right hemisphere, and parasympathetic, to the left. These studies have suggested that ictal bradycardia is most commonly a manifestation of the activation of the left temporal and insular cortex [89, 99, 101] meanwhile ictal tachycardia is more often seen in right temporal epilepsy [14, 126]. More recent studies performed by means of depth EEG recording have shown that ictal bradycardia occurs when both temporal lobes are involved, and starts at the time of spreading to the second temporal lobe [35, 103, 16]. A clear-cut lateralization pattern cannot therefore be identified between the different cardiac arrhythmia patterns. Nevertheless, left and right hemisphere do not seem equivalent in cardiac rhythm regulation. For this reason, HRV experiments should consider left and right lateralized seizures separately.

In the next section, an evaluation and a comparison of the state-of-the-art HRV metrics in both left and right lateralized seizures is investigated in the frame of epileptic state identification, including the pre-ictale state.

3.4 Experiments and results

The primary data used in our experiments are polygraphic recordings (19 EEG scalp electrodes - 1 ECG chest derivation - Twin software on Grass Telefactor system) associated with video monitoring performed on patients suffering from temporal lobe epilepsy (see Table 3.2 for patient clinical details). The analysis of the patient data has been approved by the Committee for biomedical ethics of the Faculty of Medicine of the Université catholique de Louvain. The beginning and the end of the seizures are defined by an expert using EEG signals and the video images. R spike locations are

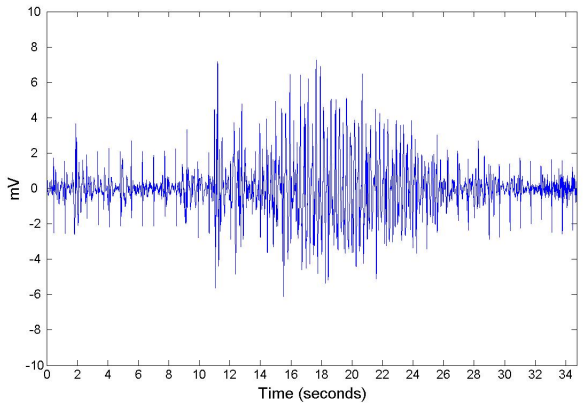

Figure 3.2: Noisy segment in an ECG recording corresponding to movement artifacts. The detection of R spikes becomes impossible in the noise burst.

detected from the ECG signal using an R spike detection algorithm such as the ones described in Chapter 2. The R spike markings are then reviewed manually to ensure no false or missing detection alters the results. The heart rate signal is thus reduced to a data series containing successive R-R interval durations (often referred to as R-R tachogram).

In order to obtain a reliable estimate of the HRV metrics, the Task Force recommends a five minutes tachogram duration. Representative data segments for the four seizure states described above are therefore defined as follows: (1) inter-ictal state: a five minute segment starting 1 hour before the seizure starts (2) pre-ictal state: the 5 last minutes before the seizure starts (3) ictal state: 5 minutes from the seizure onset on and (4) post-ictal state: 5 minutes immediately after the end of the seizure.

The preparation of the data is a very tedious procedure. It starts with the rejection of segments with noisy ECG signals by eliminating those segments where R spike detection was impossible, especially in the ictal states with convulsive movements. Figure 3.2 shows a noisy ECG segment corresponding to movement artifacts; the detection of R spikes becomes impossible in the noise burst. At the end of the data preparation step, a total of 13 satisfactory seizures samples were obtained from our 6 different patients. In the 13 seizures, seven have a right lateralization and six have a left lateralization.

Table 3.1 holds a reminder of the metrics presented in Section 3.2 and the corresponding abbreviations. The ten HRV metrics are evaluated in the four seizures states. The parameters for the TI, ApEN and SampEN methods are defined according to the

Abbrev.	HRV Metric
SDNN	Standard deviation of R-R intervals
SDSD	Standard deviation of the differences between successive R-R intervals
pNN50	Percentage of R-R intervals greater than 50 msec
TI	Triangular index
ApEN	Approximate entropy
SampEN	Sample Entropy
DFA	Detrended fluctuation analysis
LF	Low frequency
HF	High frequency
LF/HF	LF/HF ratio

Table 3.1: List of the HRV metrics investigated in this study and the corresponding abbreviations.

Task Force and the original papers. The Lomb method for frequency analysis of unequally spaced data is used for the frequency domain metrics. An ANOVA test is applied next in order to determine whether the four seizure states defined actually exhibit differences in one of the metrics. Briefly, in the ANOVA test the null hypothesis is that all the states do not exhibit significant differences in the given characteristic and the alternative hypothesis is that at least one state is different from the others. Tukey-Kramer's multiple comparison procedure is also applied to specifically detect which pairs of states are significantly different.

The p-values of the ANOVA test are shown in the first column of Table 3.3 for the left lateralized seizures and in Table 3.4 for the right lateralized seizures. A low p-value (i.e. < 0.01) indicates a significant difference in the seizure states according to the metric. The results of the Tukey-Kramer's multiple comparison test between the four states are shown in columns 3 to 8 of Table 3.3 and Table 3.4. Label 1 designates the inter-ictal state, 2 pre-ictal state, 3 ictal state and 4 post-ictal state. Each column reports whether the metric value between two states is significantly different, for example the value in column 1-2 indicates whether state 1 is different from state 2 according to the test.

3.5 Discussion

In this study of 13 seizures, we evaluated 10 metrics for the quantification of heart rate variability in the frame of epileptic state identification. Four of these metrics are time

Patient	Sex	Handedness	Age	Seizure origin
1	F	R	24yo	R temporal
2	M	L	25yo	R mesiotemporal
3	F	L	31yo	R temporal
4	F	-	48yo	L mesiotemporal
5	F	-	70yo	L mesiotemporal
6	F	R	34yo	L mesiotemporal

Table 3.2: Clinical details of patients involved in this study; R is right, L is left, F is female and M is male.

Metric	p-value	1-2	1-3	1-4	2-3	2-4	3-4
SDSD	0.27	No	No	No	No	No	No
SampEN	0.55	No	No	No	No	No	No
ApEN	0.57	No	No	No	No	No	No
LF/HF	0.69	No	No	No	No	No	No
SDNN	0.82	No	No	No	No	No	No
TI	0.86	No	No	No	No	No	No
DFA	0.89	No	No	No	No	No	No
NHF	0.94	No	No	No	No	No	No
NLF	0.96	No	No	No	No	No	No
pNN50	0.98	No	No	No	No	No	No

Table 3.3: Significance of HRV metrics for discriminating between the four states (1 inter-ictal, 2 pre-ictal, 3 ictal and 4 post-ictal) of patients with left seizure lateralization. The p-value is the result of the ANOVA. A low p-value indicates a significant difference in the seizure states according to the metric. The binary comparisons are the result of the Tukey-Kramer test and reveal which pairs of groups actually differ in the given metric.

Metric	p-value	1-2	1-3	1-4	2-3	2-4	3-4
SDSD	< 0.01	No	Yes	No	Yes	No	Yes
ApEN	< 0.01	No	Yes	Yes	Yes	No	Yes
SampEN	< 0.01	No	Yes	Yes	Yes	Yes	Yes
pNN50	< 0.01	No	Yes	Yes	Yes	Yes	No
SDNN	< 0.01	No	Yes	No	Yes	No	Yes
TI	0.2	No	No	No	No	No	No
NLF	0.69	No	No	No	No	No	No
LF/HF	0.74	No	No	No	No	No	No
NHF	0.79	No	No	No	No	No	No
DFA	0.92	No	No	No	No	No	No

Table 3.4: Significance of HRV metrics for discriminating between the four states (1 inter-ictal, 2 pre-ictal, 3 ictal and 4 post-ictal) of patients with right seizure lateralization. The p-value is the result of the ANOVA. A low p-value indicates a significant difference in the seizure states according to the metric. The binary comparisons are the result of the Tukey-Kramer test and reveal which pairs of groups actually differ in the given metric.

domain metrics (SDSD, SDNN, TI, pNN50), three are frequency domain metrics (normalized LF, normalized HF, LF to HF ratio) and the last three are non-linear metrics (SampEN, ApEN, DFA). We observe that the lateralization of the seizures is a crucial indicator of whether the ECG signal will be correlated with the seizures. Left lateralized seizures did not show any change in HRV metrics in the four seizure states. To the opposite, in right lateralized seizures we observe that time domain metrics based on the standard deviation of the R-R series or the pNN50 metric and the non-linear metrics based on the entropy of the R-R distribution show significant differences between the seizure states.

Frequency metrics did not reveal any significant differences where time domain metrics succeeded. These findings are in contradiction with the results from [120] where multiple patterns of changes in the frequency bands were observed at the onset of seizures. This difference may be explained by the fact that the frequency transform used in this previous study relies on a resampling step, and this has been shown to lead to significant errors in some frequency domain metrics [61, 21]. In particular, the resampling is known to add LF components and reduce the HF components of the spectrum [21]. However, it is important to note that R spike labels are sometimes hard to detect during seizures, because of the motion artifacts in the ECG signal. The resampling effect is therefore expected to be much more significant during seizures where fewer R-R intervals are available. Furthermore, parameters of the resampling

such as the choice of the interpolation method and the sampling frequency can have influence on the power spectrum. The difference between our results and the results obtained in [120] may be explained by this phenomenon.

When observing pair to pair comparisons for the significant metrics in right lateralized seizures, we find a significant difference in heart rate variability during the ictal state of seizures. These findings confirm the suggestion that the ECG signal could be used in seizure detection algorithms along with EEG signals [46]. In particular, time domain metrics based on the standard deviation of the R-R series and the non-linear metrics based on the entropy of the R-R distribution can be useful tools for the detection of the seizures. Nevertheless, this change in variability does not seem to happen before the actual start of the seizure. These findings confirm the doubts raised in recent studies about the existence of a preictal state that could be used in order to predict the onset of seizures [39, 85]. Results also validate the pertinence of the more recent non-linear metrics based on the entropy of the R-R series.

The experiments conducted in this work involved retrospective collection of data. Several important factors have therefore been left uncontrolled. For instance, the medications taken by the involved patients or the unequal representation of males and females may alter the results. The interesting findings obtained in our preliminary experiments should therefore be validated on a bigger sample size specifically recruited for this particular HRV study.

House: “Occam’s Razor. The simplest explanation is almost always somebody screwed up.”

House MD, season 1 episode 3.

Chapter 4

Supervised Classification of Heart Beats

Contents

4.1 Introduction

The analysis of electrocardiogram (ECG) signals provides critical information on the cardiac function of patients. Cardiac disease conditions can be diagnosed by identifying abnormal heart beats in the ECG signal. In such applications as clinical monitoring or pharmaceutical phase-one studies, long-term recordings of the ECG signal are required to this end. These long-term recordings are typically obtained using the popular Holter recorders. Holter ambulatory systems record at least 24 hours of heart activity, resulting in data that contain thousands of heart beats. The analysis is usually performed off-line by cardiologists, whose diagnosis may rely on just a few transient patterns. Because of the high number of beats to evaluate, this task is very expensive and reliable visual inspection is difficult.

Computer-aided classification of pathological beats is therefore of great importance to help physicians perform correct diagnosis. Before considering the automatic classification of the beats, each beat in the recorded signal must first be identified and extracted from the full sampled recording. Chapter 2 focused on automatic methods to annotate the characteristic points of the beats. Because the ventricles contain more muscle mass than the atria, the QRS complex is the most representative feature of the heart beat. The detected R spikes therefore typically serve as beat identifiers. The automatic classification of the identified beats can then be achieved subsequently by machine learning classifiers. In real situations, designing such automatic classifiers is however difficult for two main reasons.

First, the vast majority of the heart beats are normal healthy beats while just only a small number of beats are pathological, and of course those are of uttermost importance. In such situations, standard machine learning algorithms generally perform poorly because they are designed to generalize from training data and to output the simplest hypothesis that best fits the data, based on Occam's razor. As a result, machine learning algorithms tend to treat the pathological beats as noise and the learning process often leads to a dummy classifier always predicting the healthy class.

Second, artificial intelligence methods used to automate classification require the extraction of discriminative features from the heart beat signals. Unfortunately, very little information is available to decide how to extract and build those features. Spurious features can harm the classifier, especially in the presence of unbalanced classes [47, 88]. As a consequence, the classification results can be suboptimal. Feature selection techniques can solve this issue by identifying the discriminative features and can also improve the interpretation of the classifier. This property is especially useful in medical applications where the selected features may help to understand the causes and the origin of the pathologies.

In this chapter, the pertinence of a large number of features previously proposed for

heart beat classification is investigated using feature selection techniques. Moreover, several classification models able to cope with the class unbalance are considered. Finally, best practice rules for constructing a reliable heart beat classification system are also presented. Experiments are conducted on real ambulatory signals from the Physiobank arrhythmia database. Personal publications over the materials in this chapter are [34, 24, 37, 26, 27].

4.2 Learning with unbalanced datasets

The class unbalance problem arises when the number of observations in a particular class is sorely higher than the number of observations in the other classes. This difference can either arise from a lack of collected data or be an inherent property of the underlying application. In a credit card fraud detection problem, the data is for instance intrinsically unbalanced since only a very few portion of the transactions are illegal compared the total amount of transactions in a given time range. It is also the same situation in the heart beat classification problem where pathological beats are transient scarce observations.

More formally, two difficulties arise from learning with unbalanced datasets. First, the choice of the classification performance metric should carefully be made to take the minority class into account. Second, the decision boundary established by standard accuracy-driven algorithms tends to be biased towards the majority class; hence the minority class instances are more likely to be misclassified. Several approaches have been designed to improve classification algorithms when faced with unbalanced datasets. In the following of this section, the performance metrics and the methodological approaches to learning from unbalanced data are presented.

4.2.1 Performance metrics

As far as unbalanced data are concerned, the metric should value the minority class. To formulate metrics for measuring the performance of a classifier, let us consider the confusion matrix in Table 4.1. The commonly used metric is the overall classification rate, or accuracy, defined as $\frac{tp+tn}{tp+tn+fp+fn}$. However, in the case of unbalanced classes, it is not a suitable metric since the small class has less effect on accuracy than the majority class. The poor performance obtained for the minority class with the accuracy metric has for example been experimentally assessed on 23 datasets in [123].

As illustrated in Table 4.1, we define N_+ and N_- as respectively the number of positive and negative examples and we assume that the positive class is the minority

		Predicted		
		-	+	total
True	-	tn	fp	N_-
	+	fn	tp	N_+

Table 4.1: Confusion matrix: the columns represent the classes assigned by the classifier and the rows represent the true classes. Value tn stands for the true negatives count, tp true positives count, fn false negatives count and fp false positives count.

class. The sensitivity (se) and the specificity (sp) are defined as

$$\text{se} = tp/(fn+tp) = tp/N_+ \tag{4.1}$$

$$\text{sp} = tn/(tn+fp) = tn/N_-. \tag{4.2}$$

Note that the sensitivity and the specificity represent the accuracy of the positive and negative classes respectively. Any metric that values both classes should be a trade-off between these two parameters.

The receiver operating characteristic (ROC) curve is a graph showing the relationship between the sensitivity and $1-$specificity as a function of the decision threshold on the output of the classifier. The area under the curve (AUC) is employed to summarize the performance of the classifier into a single metric without reference to a given threshold. In a binary problem, the ROC curve can thus be used to adapt the decision threshold to yield satisfactory performances in both classes. Although generalizations of the ROC curve for multiclass problems have been investigated [57], the construction of ROC curves remains however difficult in multiclass problems.

Another metric that values both classes equally is the balanced classification rate (BCR). In a two class problem, the BCR is defined as the mean between the sensitivity and the specificity. It is naturally generalizable to multiclass problems by simply computing the mean of all class accuracies. In practice, the geometric mean (GM) is often used instead of the arithmetic mean [55]. In a K class problem, the geometric mean is mathematically defined as follows:

$$\text{Geometric mean (GM)} = \sqrt[K]{\prod_{k=1}^{K} \frac{tp_k}{N_k}}. \tag{4.3}$$

4.2.2 Approaches to unbalanced classification

In this section, the methodological approaches to learning from unbalanced data are described. The pros and cons of each technique are outlined.

Sampling

One of the most common approaches to the class unbalance problem is to pre-process the training data to obtain a more balanced class prior distribution. The two basic sampling methods are oversampling and undersampling. Oversampling increases the number of minority observations by randomly duplicating them. The main issue associated to oversampling is the increase in computational cost that can be very dramatic in highly unbalanced situations. Moreover, for a given dataset, any oversampling method will only result in duplicating already known information. For this reason, advanced oversampling algorithms such as SMOTE (synthetic minority oversampling technique) and its variants create new observations in the minority class by modeling their probability distribution and generating new samples from the estimated distribution [69].

In contrast, undersampling randomly removes observations from the majority class while keeping the minority class intact. The major drawback with undersampling is that discarding observations may lead to loss of potentially useful information. For this reason, more advanced undersampling methods attempting to discard only redundant information in the majority class have been investigated. This is typically achieved by clustering techniques aiming to group similar observations together. It has experimentally been shown that clustering methods outperform blind undersampling methods in a wide range of applications [124]. In the case of ECG signals, heart beat clustering techniques include the work of D. Cuesta-Frau (see for instance [102]) and the work of D. Novak (see for instance [80]) who investigated the use of several algorithms, features and distance measures to reduce the number of beats in Holter ECG signals.

It is important to mention that an active field of machine learning called active learning has the focus of developing classifiers with an embedded requirement of sparsity over the training observations. The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns. Active learning enabled methods include support vector machine classifiers where the observations valuable in learning the separating hyperplane (called support vectors) are identified. These support vectors identified in the learning process can then be used as a starting point for dataset subsampling purposes [122]. Nevertheless, these methods are yet based on accuracy-driven classifiers that still suffer from the class unbalance when establishing the decision boundary and are therefore not applicable in unbalanced situations.

One-class learning

From the point of view of the classifier, in unbalanced datasets the minority class is simply considered as noise with respect to the majority class. Therefore, another way

to tackle the problem is to use methods designed for outliers detection or so-called novelty detection. One-class classifiers have been developed for this purpose. One-class classifiers produce a probability of being in the learned class for each test observation. A threshold on the probability is then used to detect non class members. One of the early one-class classifiers is the one-class neural network proposed in [51]. The concept has later been applied to support vector machine classifiers [108]. The drawback of such methods is the difficulty in choosing an appropriate decision threshold.

Ensemble learning

Ensemble learning is a field of machine learning in which the prediction of multiple classifiers is combined to classify test observations. The two most widely known ensemble learning approaches are boosting and bagging. The idea behind boosting algorithms is to improve the performances of so-called weak classifiers that individually only perform slightly better than a random classifier into a weighted combination that classifies with high accuracy. For example, the popular Adaboost (adaptative boosting) algorithm iteratively builds subsequent classifiers which are giving more weight to observations misclassified by previous classifiers. The final prediction is then obtained by a linear combination of all the subsequent classifier outputs. The question of how to adapt boosting to handle unbalanced datasets was considered in [52] and [64] where the AdaUboost (adaptative uneven boosting) and LPUBoost (linear programming uneven boosting) algorithms have respectively been introduced. The simple idea behind these algorithms is to iteratively bias the weighting of observations in favor of examples of the minority class. In general, boosting algorithms have been shown to exhibit a strong resistance to overfitting but also to be very sensitive to outliers and noise.

On the other hand, bagging methods train multiple classifiers on distinct random subsampling of the full dataset and their outputs are then combined to yield the final prediction. These methods are however rather poor in strong unbalanced situations or in multiclass situations where is it difficult to randomly obtain a representative subsampling.

Cost-sensitive learning

Traditional classification algorithms assume that misclassification errors cost equally. However, in many applications such as medical diagnosis, fraud detection and risk management, the small classes are of primary interest. For example, in a heart beat classification task where the objective is to detect heart diseases, false negatives can have dramatic consequences while false positives are of course undesired but still not life-threatening. In weighted learning, the idea is to design cost-enabled classifiers

	Discard observations	Computational burden	Meta Parameters	Model dependent
Oversampling	no	yes	yes	no
Undersampling	yes	no	yes	no
One-class	no	no	yes	yes
Ensemble	no	yes	yes	no
Cost-sensitive	no	no	no	yes

Table 4.2: Comparison of learning approaches for unbalanced datasets. The first column of the table shows whether a given approach discards observations. The second column considers the computational burden. The third column shows whether additional meta-parameters have to be fixed. The last column shows whether the method is model dependent.

that include distinct misclassification costs for each class in their objective function. Higher costs are then given to minority classes so as to guide the training process to solutions which favor these classes. For example, the weighted LDA model proposed in [18] directly integrates costs into the training of the linear discriminants.

4.2.3 Discussion

Table 4.2 summarizes the pros and cons of each approach. The first column of the table considers the fact that a given approach makes use of all the sampled observations or discards some of them. The second column considers the computational cost, for example the need to train a large number of subsequent models in boosting approaches or to duplicate many observations in the oversampling approach. The third column shows whether meta-parameters have to be a priori fixed, such as the outliers threshold in one-class learning. The last column represents the fact that the approach can be applied to any kind of model or if it is limited to a specific form such as one-class learning where very few one-class versions of traditional learning models have been designed.

The only drawback of cost-sensitive learning methods for unbalanced learning is the model dependency. In the next section, we will see that any model whose optimization can be written in a specific form can in fact be modified to achieve embedded cost-sensitive learning.

4.3 Cost-sensitive models

Define $\mathbb{D} = \{(\mathbf{x}_n, \mathrm{y}_n)\}_{n=1}^{N}$, a dataset containing N pairs of $P-$dimensional observations $\mathbf{x}_n = [x_n^p]_{p=1}^{P} \in \mathbb{R}^P$ and associated class labels $\mathrm{y}_n \in \{1, 2, \ldots, k, \ldots, K\}$. We define N_k as the number of observations in class k. In this section, we consider the general class of models whose optimization takes the following form:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \mathrm{L}(\mathrm{y}_n, f(\mathbf{x}_n, \mathbf{w})) + \lambda \|\mathbf{w}\| \tag{4.4}$$

where $\mathbf{w}$ are the parameters of the model, $\mathrm{L}(\mathrm{y}_n, f(\mathbf{x}_n))$ is a loss function measuring the discrepancy between the true label vector y_n and the model output $f(\mathbf{x}_n)$ for each training instance n and the right-hand side of the sum is a regularization term. It seems natural to choose a loss function counting the number of misclassified points (the accuracy). Unfortunately, the optimization of such a function combined with the regularization penalty is known to be NP hard [109]. In practice, an approximation of the accuracy is thus commonly used as loss function. Examples of loss functions are the hinge loss in support vector machines and the logit loss in binary logistic regression:

$$\mathrm{L}_{\mathrm{logit}}(\mathrm{y}_n, f(\mathbf{x}_n, \mathbf{w})) = \log(1 + e^{-\mathrm{y}_n f(\mathbf{x}_n, \mathbf{w})}) \tag{4.5}$$

$$\mathrm{L}_{\mathrm{SVM}}(\mathrm{y}_n, f(\mathbf{x}_n, \mathbf{w})) = \max(0, 1 - \mathrm{y}_n f(\mathbf{x}_n, \mathbf{w})). \tag{4.6}$$

Any model that can be cast as an optimization of the form in Eq. (4.4) can be modified to achieve cost-sensitive learning. First, the sum over all observations in Eq. (4.4) is reorganized into two sums over the classes and the observations of each class:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \mathrm{L}(\mathrm{y}_n, f(\mathbf{x}_n, \mathbf{w})) + \lambda \|\mathbf{w}\| \tag{4.7}$$

$$= \min_{\mathbf{w}} \sum_{k=1}^{K} \sum_{\{n|\mathrm{y}_n=k\}} \mathrm{L}(\mathrm{y}_n, f(\mathbf{x}_n, \mathbf{w})) + \lambda \|\mathbf{w}\|. \tag{4.8}$$

Next, K cost parameters are added to weight the terms associated to each class by a factor c_k:

$$\min_{\mathbf{w}} \sum_{k=1}^{K} c_k \sum_{\{n|\mathrm{y}_n=k\}} \mathrm{L}(\mathrm{y}_n, f(\mathbf{x}_n, \mathbf{w})) + \lambda \|\mathbf{w}\|. \tag{4.9}$$

The main drawback is that cost-sensitive methods assume that costs are known in advance; it is rarely the case in real situations. Nevertheless, when the goal is to give equal importance to every class, one can set the costs to the inverse of the class priors:

$$c_k = \frac{N}{N_k}. \tag{4.10}$$

Using such cost values gives the same cost, or weight, to errors in both majority and minority classes during the training of the model. The accuracy criterion in the objective function of a given model is then replaced by a balanced criterion. Nevertheless, precautions must be taken when faced with small datasets having few labeled samples or samples not representative of the real population priors. In such situations, the class prior estimations can be biased and unreliable, as discussed in [125]. In the remaining of this section, we will show how specific classifiers like linear discriminants, support vector machines and conditional random fields are modified to integrate class specific misclassification costs in their learning process following the same reasoning.

4.3.1 Weighted SVM

A support vector machine (SVM) is a supervised learning method introduced by Vapnik [119]. The two-class case is described here, so $y_n \in \{-1,+1\}\ \forall n$, because its extension to multiple classes is straightforward by applying the one-against-all or one-against-one methods.

SVMs are linear machines that rely on a preprocessing to represent the features in a higher dimension, typically much higher than the original feature space. With an appropriate non-linear mapping $\varphi(\mathbf{x})$ to a sufficiently high-dimensional space, finite data from two categories can always be separated by a hyperplane. In SVMs, the distance from this hyperplane to the nearest data point on each side, referred to as the margin, is maximized. Assume each observation $\mathbf{x}_n$ has been transformed to $\mathbf{z}_n = \varphi(\mathbf{x}_n)$. The soft-margin formulation of the SVM allows examples to be misclassified or to lie inside the margin by the introduction of slack variables ξ_n in the objective constraints:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \xi_n + \lambda \|\mathbf{w}\|_2 \tag{4.11}$$

$$\text{s.t.} \begin{cases} y_n(\langle \mathbf{w}, \mathbf{z_n} \rangle) \geq 1 - \xi_n, & \forall n = 1 \ldots N \\ \xi_n \geq 0 & \forall n = 1 \ldots N \end{cases} \tag{4.12}$$

where $\mathbf{w}$ are the parameters of the hyperplane. For any feasible solution, misclassified examples have an associated slack value ξ_n greater than 1. We can see from Eq. (4.11) that minimizing the first term minimizes the classification error while minimizing the second term is equivalent to maximizing the classification margin. This soft-margin formulation of the SVM can be rewritten in the form of Eq. (4.4) by rewriting the soft-margin constraint as $\xi_n \geq 1 - y_n(\langle \mathbf{w}, \mathbf{z_n} \rangle)$. Since each ξ_n is minimized and $\xi_n \geq 0$, at the optimum we have $\xi_n = \max(0, 1 - y_n(\langle \mathbf{w}, \mathbf{z_n} \rangle))$. Substituting this result in Eq.

(4.11) yields the hinge loss formulation. This classical SVM formulation has been shown to suffer from class unbalance and in worst unbalanced cases to yield a classifier biased towards the majority class [5]. The reason is that classifying everything in the majority class is what makes the margin the largest, with zero cumulative loss on the abundant majority examples. The only trade-off is the small amount of cumulative loss on the few minority examples which do not count for much.

To overcome this problem, different penalties for each class can be included in the loss function term as in Eq. (4.9) and Eq. (4.10). The optimization problem then becomes:

$$\min_{\mathbf{w}} \left(\frac{N}{N_+} \sum_{\{n|\mathrm{y}=1\}} \xi_n + \frac{N}{N_-} \sum_{\{n|\mathrm{y}=-1\}} \xi_n \right) + \lambda \|\mathbf{w}\|_2 \tag{4.13}$$

$$\text{s.t.} \begin{cases} \mathrm{y}_n(\langle \mathbf{w}, \mathbf{z_i} \rangle) \geq 1 - \xi_n, & \forall n = 1 \ldots N \\ \xi_n \geq 0, & \forall n = 1 \ldots N. \end{cases} \tag{4.14}$$

Using this formulation, the model now optimizes a convex approximation of the BCR rather than the accuracy. By introducing the Lagrangian multipliers α_n, this *primal* formulation can be rewritten in a so-called *dual* form. The optimization is then typically achieved by solving the system using quadratic programming. For this type of optimization, there exist many effective learning algorithms. A common method is Platt's Sequential Minimal Optimization (SMO) algorithm [96], which breaks the problem down into 2-dimensional sub-problems that may be solved analytically, eliminating the need for a numerical optimization algorithm such as conjugate gradient methods.

In the dual form, the explicit form of the mapping function φ must not be known as long as the kernel function $K(\mathbf{x}_n, \mathbf{x}_j) = \varphi(\mathbf{x}_n)\varphi(\mathbf{x}_j)$ is defined. The kernel can for example be the linear kernel $K(\mathbf{x}_n, \mathbf{x}_j) = \mathbf{x}_n^t \mathbf{x}_j$ or the radial basis function kernel $K(\mathbf{x}_n, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_n - \mathbf{x}_j||^2)$ where γ is a kernel parameter to be tuned. The sign of the following decision function is then used to determine the predicted class value y_n^* for a new unlabeled observation:

$$\mathrm{y}_i^* = \mathrm{sign}(f(\mathbf{x}_i)) \tag{4.15}$$

$$f(\mathbf{x}_i) = \mathbf{w}'\varphi(\mathbf{x}_i) = \sum_{n=1}^{N} \alpha_n \mathrm{y}_n K(\mathbf{x}_n, \mathbf{x}_i). \tag{4.16}$$

4.3.2 Weighted LDA

The weighted LDA model was introduced for heart beat classification in [18]. Linear discriminant analysis (LDA) approaches the classification problem by assuming that the conditional probability density functions $p(\mathbf{x}_n|\mathrm{y} = k)$ are normally distributed with

the simplifying homoscedastic assumption that the class covariances are identical. All the parameters $\mathbf{w}$ of the model are thus summarized by the mean class vectors μ_k and the unique covariance matrix Σ. These parameters are identified by maximizing the log-likelihood function defined as

$$\max_{\mu_1,\mu_2,\ldots,\mu_K,\Sigma} \sum_{k=1}^{K} \sum_{\{n|\mathsf{y}_n=k\}} \log(f_k(\mathbf{x}_n,\mu_k,\Sigma)) \tag{4.17}$$

where $f_k(\mathbf{x}_n,\mu_k,\Sigma)$ are the value of a Gaussian distribution with mean μ_k and covariance Σ. The similarity with Eq. (4.4) is obvious by minimizing the negative likelihood and choosing $\lambda = 0$. The optimization can be done in closed form and yields the following solution:

$$\mu_k = \frac{\sum_{\{n|\mathsf{y}_n=k\}} \mathbf{x}_n}{N_k} \tag{4.18}$$

$$\Sigma = \frac{1}{N} \sum_{k=1}^{K} \sum_{\{n|\mathsf{y}_n=k\}} (\mathbf{x}_n-\mu_k)(\mathbf{x}_n-\mu_k)^T. \tag{4.19}$$

To add costs in the model, we apply the same reasoning from Eq. (4.9) to Eq. (4.17):

$$\max_{\mu_1,\mu_2,\ldots,\mu_K,\Sigma} \sum_{k=1}^{K} c_k \sum_{\{n|\mathsf{y}_n=k\}} \log(f_k(\mathbf{x}_n,\mu_k,\Sigma)) \tag{4.20}$$

is now solved with

$$\mu_k = \frac{\sum_{\{n|\mathsf{y}_n=k\}} \mathbf{x}_n}{N_k} \tag{4.21}$$

$$\Sigma = \frac{\sum_{k=1}^{K} c_k \sum_{\{n|\mathsf{y}_n=k\}} (\mathbf{x}_n-\mu_k)(\mathbf{x}_n-\mu_k)^T}{\sum_{k=1}^{K} c_k N_k}. \tag{4.22}$$

Inference is then achieved using

$$\mathsf{y}_n^* = \max_k f_k(\mathbf{x}_n) \tag{4.23}$$

$$f_k(\mathbf{x}_n) = -(1/2)\mu_k^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mathbf{x}_n \tag{4.24}$$

which corresponds to assigning $\mathbf{x}_n$ to the class having the smallest Mahalanobis distance between the class mean and $\mathbf{x}_n$.

4.3.3 Weighted CRF

If the dataset $\{(\mathbf{x}_n,\mathsf{y}_n)\}_{n=1}^N$ is not independently distributed, there is some dependency between subsequent observations. Assume for instance the observations $\mathbf{x}_n$ correspond

to heart beats. Clearly, if beat n is a healthy beat, there are more chances that the subsequent beat $n+1$ will also be a healthy beat. To the opposite, if a pathological beat has occurred, there are more chances that another pathological beat will also occur in the future. The observations are therefore not truly independent and sequential models like CRFs can use this dependence in the classification process. Remember the objective function of the L_1-regularized CRF model from Section 2.4:

$$\max_{\lambda,\mu} \sum_{n=1}^{N} \sum_{kk'} \lambda_{kk'} f_{kk'}(y_{n-1}, \mathrm{y}_n, \mathbf{x}) + \sum_{n=1}^{N} \sum_{kp} \mu_{kp} g_{kp}(\mathrm{y}_n, \mathbf{x}) - \log(Z(\mathbf{x})) - \lambda \|\lambda, \mu\|_1 \tag{4.25}$$

$$= \max_{\lambda,\mu} \sum_{n=1}^{N} \left(\sum_{kk'} \lambda_{kk'} f_{kk'}(y_{n-1}, \mathrm{y}_n, \mathbf{x}) + \sum_{kp} \mu_{kp} g_{kp}(\mathrm{y}_n, \mathbf{x}) \right) - \log(Z(\mathbf{x})) - \lambda \|\lambda, \mu\|_1 \tag{4.26}$$

$$= \min_{\lambda,\mu} - \sum_{n=1}^{N} \left(\sum_{kk'} \lambda_{kk'} f_{kk'}(y_{n-1}, \mathrm{y}_n, \mathbf{x}) + \sum_{kp} \mu_{kp} g_{kp}(\mathrm{y}_n, \mathbf{x}) \right) + \log(Z(\mathbf{x})) + \lambda \|\lambda, \mu\|_1 . \tag{4.27}$$

The difficulty to cast Eq. (4.27) in a form similar to Eq. (4.4) lies in writing the $\log(Z(\mathbf{x}))$ term as a sum over observations. We now show that this can be achieved by using the scaling trick for the computation of the forward-backward variables. In practical implementations, the values of the forward $\alpha_k(n)$ and backward $\beta_k(n)$ variables head exponentially to zero. For sufficiently large N (i.e. 10 or more), the dynamic range of both α and β will exceed the precision range of any machine. Hence the only reasonable way of performing the computation is either to work in the log domain or to incorporate a scaling procedure [97]. In the scaling procedure, at each time step, the forward variables are normalized to sum to one as follows:

$$z_n = \sum_{k=1}^{K} \alpha_n(k) \tag{4.28}$$

$$\widehat{\alpha}_n(k) = \frac{\alpha_n(k)}{z_n}. \tag{4.29}$$

Next, we use the same scaling factors z_n for scaling the $\beta_k(n)$ variables:

$$\widehat{\beta}_n(k) = \frac{\beta_n(k)}{z_n}. \tag{4.30}$$

Since each scale factor effectively restores the magnitude of the α terms to one and since the magnitudes of the α and β variables are comparable, using the same scaling factor is an efficient way for keeping the computation within reasonable bounds. When

computing the forward-backward terms in Eq. (2.13), since the scaling factors are common to both variables, we obtain:

$$p(\mathrm{y}=k|\mathbf{x}) = \gamma_n(k) = \widehat{\alpha}_n(k)\widehat{\beta}_n(k)z_n. \tag{4.31}$$

The only drawback is that we cannot merely sum up the $\widehat{\alpha}_N(k)$ terms for computing $\mathrm{Z}(\mathbf{x})$ since these are scaled already. Nevertheless, we now show that the computation of $\log(\mathrm{Z}(\mathbf{x}))$ can be rewritten as a product over observations thanks to the scaling factors z_n. Let us first define $\widetilde{\alpha}_n$ as the forward variable computed from $\widehat{\alpha}_{n-1}$ before scaling. We have

$$\begin{aligned}
\widetilde{\alpha}_0(i) &= \alpha_0(i) && (4.32)\\
\widehat{\alpha}_0(i) &= z_0\widetilde{\alpha}_0(i) = \frac{\widetilde{\alpha}_0(i)}{\sum_j \widetilde{\alpha}_0(j)} && (4.33)\\
\widetilde{\alpha}_1(i) &= \sum_j \widehat{\alpha}_0(j)a_{ji}b_i(\mathrm{x}_1) && (4.34)
\end{aligned}$$

and by induction

$$\begin{aligned}
\widehat{\alpha}_1(i) &= z_1\widetilde{\alpha}_1(i) && (4.35)\\
&= z_1\sum_j \widehat{\alpha}_0(j)a_{ji}b_i(\mathrm{x}_1) && (4.36)\\
&= z_1\sum_j z_0\widetilde{\alpha}_0(j)a_{ji}b_i(\mathrm{x}_1) && (4.37)\\
&= z_1z_0\sum_j \alpha_0(j)a_{ji}b_i(\mathrm{x}_1) && (4.38)\\
&= z_1z_0\alpha_1(i). && (4.39)
\end{aligned}$$

And the general case is

$$\widehat{\alpha}_n(i) = (\prod_{r=1}^{n} z_r)\alpha_n(i). \tag{4.40}$$

Using this result, we can write $\log(Z(\mathbf{x}))$ as

$$\begin{aligned}
\sum_j \widehat{\alpha}_N(j) = (\prod_{n=1}^{N} z_n)\sum_j \alpha_N(j) &= 1 && (4.41)\\
(\prod_{n=1}^{N} z_n)Z(\mathbf{x}) &= 1 && (4.42)\\
Z(\mathbf{x}) &= \frac{1}{\prod_{n=1}^{N} z_n} && (4.43)\\
\log(Z(\mathbf{x})) &= -\sum_{n=1}^{N} z_n. && (4.44)
\end{aligned}$$

It is thus only feasible to compute the logarithm of $Z(\mathbf{x})$ but not $Z(\mathbf{x})$ since it would be out of the dynamic range of the machine anyway. Substituting Eq. (4.44) into Eq. (4.27) yields the desired formulation:

$$\min_{\lambda,\mu} \sum_{n=1}^{N} \left(\sum_{kk'} \lambda_{kk'} f_{kk'}(y_{n-1}, \mathsf{y}_n, \mathbf{x}) + \sum_{kp} \mu_{kp} g_{kp}(\mathsf{y}_n, \mathbf{x}) - \log(z_n) \right) + \lambda \|\lambda, \mu\|_1 . \quad (4.45)$$

The weighted CRF objective function then becomes:

$$\min_{\lambda,\mu} \sum_{k=1}^{K} c_k \sum_{\{n|\mathsf{y}=k\}} \left(\sum_{kk'} \lambda_{kk'} f_{kk'}(y_{n-1}, \mathsf{y}_n, \mathbf{x}) + \sum_{kp} \mu_{kp} g_{kp}(\mathsf{y}_n, \mathbf{x}) - \log(z_n) \right) + \lambda \|\lambda, \mu\|_1 . \quad (4.46)$$

4.4 Time-series feature extraction

In this section, two methods previously proposed in the literature to extract features from the ECG signal are introduced.

4.4.1 Hermite basis functions

The representation of the heart beat signal via Hermite basis functions (HBF) was first introduced by [59] for a clustering application and later by [91] for classification. This approach exploits similarities between the shapes of HBF and typical ECG waveforms. The expansion of x_t into a Hermite series of order Q is written as

$$x_t = \sum_{q=0}^{Q-1} e_q \phi_q(t, \sigma) \quad (4.47)$$

where e_q are the expansion coefficients and σ is a width parameter. The $\phi_q(t, \sigma)$ functions are the Hermite basis functions of order q defined as follows:

$$\phi_q(t, \sigma) = \frac{1}{\sqrt{\sigma 2^q q! \sqrt{\pi}}} e^{-t^2/2\sigma^2} H_q(t/\sigma) \quad (4.48)$$

where $H_q(t/\sigma)$ is the Hermite polynomial of the qth order. The Hermite polynomials satisfy the following recurrence relation:

$$H_q(x) = 2xH_{q-1}(x) - 2(q-1)H_{q-2}(x) \quad (4.49)$$

with $H_0(x) = 1$ and $H_1(x) = 2x$.

The higher the order of the Hermite polynomial, the higher its frequency in the time domain, and the better the capability of the expansion in Eq. (4.47) to reconstruct

the signal. The width parameter σ can be tuned to provide a good representation of signals with large differences in durations. The coefficients e_q of the HBF expansion are typically estimated by minimizing the sum of squared errors using singular value decomposition and the pseudo-inverse technique. These coefficients summarize the shape of signal and can be treated as the features used in the classification process.

4.4.2 Higher order cumulants

The statistical properties of the heart beat signal can be represented by its higher order cumulants. For convenience with the heart beat classification literature, we refer to these higher order cumulants as higher order statistics (HOS). The cumulants of order two, three and four are usually considered [90]. Assuming the signal x_t has a zero mean, its cumulant C_x^i of order i can be computed as follows:

$$\begin{aligned} C_x^2(\tau_1) &= E\{x_t x_{t+\tau_1}\} \\ C_x^3(\tau_1,\tau_2) &= E\{x_t x_{t+\tau_1} x_{t+\tau_2}\} \\ C_x^4(\tau_1,\tau_2,\tau_3) &= E\{x_t x_{t+\tau_1} x_{t+\tau_2} x_{t+\tau_3}\} \\ &\quad -C_x^2(\tau_1)C_x^2(\tau_3-\tau_2) \\ &\quad -C_x^2(\tau_2)C_x^2(\tau_3-\tau_1) \\ &\quad -C_x^2(\tau_3)C_x^2(\tau_2-\tau_1) \end{aligned}$$

where E is the expectation operator, and τ_1, τ_2, τ_3 are the time lags. Note that the order 2 cumulant is the classical autocorrelation function.

4.5 Guidelines for heart beat classification

In this section, best practice recommendations for constructing reliable ECG classification methodologies are presented. The standards defined by the American Association for Medical Instrumentation (AAMI) are presented and the inter-patient classification paradigm is introduced. The importance of these two paradigms for the design of heart beat classification systems and for the assessment of their relative merits is emphasized. Next, the pros and cons of feature selection techniques are evaluated in the specific context of heart beat classification.

4.5.1 AAMI standards

A large variety of automatic ECG classification methods have been investigated in the literature. However, very few reported experiments follow the beat classification standards defined by the AAMI [9], which makes it very difficult to assess the relative

merits of the methods and of the proposed extracted features. The AAMI defines the four clinically relevant heart beat classes:

N-class includes beats originating in the sinus node: normal beats, bundle branch block beat types, atrial and nodal escape beats;

S-class includes supraventricular ectopic beats: (aberrant) atrial, nodal and supraventricular premature beats;

V-class includes ventricular ectopic beats: premature ventricular contraction and ventricular ectopic beats;

F-class includes beats that result from fusing normal and ventricular ectopic beats.

For a given classification algorithm, the AAMI outlines the necessity to report the classification performances for each of these four classes.

4.5.2 Inter-patient classification

Another important aspect of heart beat classification is the way the training and test sets are composed. Most published methods for classifying the heart beats of a patient require access to previous data from that particular patient. In other words, data for each patient must be present both in training and test sets. We refer to this as "intra-patient" classification. On the other hand, "inter-patient" classification consists in classifying the beats of a new tested patient according to a reference database and a model built from data from other patients. This process thus implies generalization from one patient to another.

The results that can be achieved with intra-patient methods are naturally better than when inter-patient classification is performed, but the patient labeled beats are usually not timely available in real situations. Furthermore, because pathological beats can be very rare, there is no guarantee that the few training beats that would be labeled for this patient would contain representatives for each class; and the classifier could possibly fail in predicting something it has not learned.

Despite these major drawbacks, the large majority of previously reported work focuses on intra-patient classification. A comprehensive review of intra-patient classification methods can be found in [20] and in [79] for recent results. In this chapter, we focus on inter-patient classification of heart beats following the AAMI guidelines. The "inter-patient" classification is a much harder task of generalization but it is also much more useful since labeled beats from a new patient are usually not timely available in real clinical situations.

The first study to establish a reliable inter-patient classification methodology following AAMI guidelines is [18]. Since then, two other algorithms using the same methodology have been proposed in [93] and [70] but did not yield better results.

4.5.3 Feature selection

Previously reported inter-patient heart beat classification algorithms apply feature selection using either domain knowledge [93] or by evaluating a few combinations of several feature groups [18] without evaluating the relevance of each individual feature included in the classifier. Furthermore, distinct features groups are considered in each study which makes it difficult to assess their discriminative power on a fair basis. Spurious features can deteriorate the performances of the classifier and this effect is even more pronounced with unbalanced classes [47, 88]. Hence, the results obtained by these previously reported models can be suboptimal. Moreover, feature selection does not only serve the classification performances; it also provides insight in the classification process. This property is especially useful in medical applications where the selected features may help to understand the causes and the origin of the pathologies.

In this section, techniques for the selection of discriminative factors from a large set of computed features are investigated in the specific frame of heart beat classification. Inter-patient classification of heart beats following AAMI standards involves a large number of observations and of features, several distinct patients and multiclass labels. These characteristics must be taken into account when choosing the feature selection technique associated with a given model. The selection of discriminative features can typically be achieved either by wrapper, filter or embedded approaches. Wrappers utilize the learning machine of interest as a black box to score subsets of variables according to their predictive power. Filters make use of a criterion independent of the prediction model to select the variables like a pre-processing step. Embedded methods perform variable selection in the process of training and are usually specific to given learning machines.

Wrapper approaches

The exhaustive wrapper approach consists in feeding a model with the $2^P - 1$ possible feature subsets (P being the total number of features) and to choose the one for which the model performs best. The exhaustive wrapper approach is therefore the optimal feature selection technique for a given model. Such an exhaustive search is however intractable in practice since it would require the training (including the time-consuming optimization of potential hyper-parameters by a so-called "leave-one-patient-out" cross-validation procedure to ensure inter-patient validation) of $2^P - 1$ distinct models. When simple and fast (e.g. linear) models are considered, one can

nevertheless circumvent this issue by using an incremental wrapper approach. One of the most common incremental search procedures is the forward selection algorithm. Its principle is to select at each step the feature whose addition to the current subset yields the highest increase in prediction performances.

More precisely, the procedure usually begins with the empty set of features. The first selected feature is the one which individually maximizes the performances of the model. The second step consists in finding the feature from the feature set which leads to the best increase in performance when combined to the previously selected feature. The procedure is then repeated until no feature can increase the performance anymore. Although this incremental search is not guaranteed to converge to the selection of the optimal subset of features, it has proven to be very efficient in practice and it reduces the required number of models to train from $2^P - 1$ to $O(P)$. Because they are designed for a specific model, wrapper approaches are expected to produce better results than filter approaches.

In an inter-patient heart beat classification task, the dataset contains tens of patients, several thousands of observations, a large number of features and four classes. If a multi-class SVM classifier is considered, $K \times (K-1)/2$ models must be trained for each feature set candidate and each model itself requires the tuning of two hyper-parameters by "leave-one-patient-out" cross-validation. Needless to say, a wrapper feature selection strategy for such a classification model is very time-consuming and can take several months of computer time. In such situations, filter approaches are to be preferred over wrapper approaches.

Filter approaches

Filter approaches can be subdivided in two categories: ranking and multivariate filters. Both methods rely on a criterion that is independent of the prediction model in order to select the features as a pre-processing step. The difference lies in the fact that in the univariate selection, a subset of features that are individually discriminative is selected. In the multivariate selection, a subset of features that are discriminant when taken as a group is selected.

The drawback of the ranking filter approach is that a variable will be rejected if completely useless by itself although it may provide a significant performance improvement when taken with others. This is not the case with multivariate filters since they are able to grasp covariate discriminative information. Nevertheless, the major computational advantage of filter approaches over wrapper approaches is lost with multivariate filters since they also require some kind of incremental search strategy.

One of the most popular criterion for both kinds of filter approaches is the mutual information (MI) value [110]. Mutual information is able to detect non-linear rela-

tionships between variables and it is naturally suitable for multiclass problems. For ranking filters, the mutual information value between each feature and the class labels can efficiently be computed by a histogram-based estimator [82]. For multivariate filters, the MI between a group of features and the class labels is typically estimated using a nearest neighbor rule based algorithm [45]. This algorithm requires the computation of the $N \times N$ distance matrix for each candidate feature subset and the tuning of the number of neighbors considered. In a heart beat classification task with tens of thousands of beats, the multivariate filter approach is thus also computationally intractable even with an incremental forward algorithm.

For this reason, the ranking procedure may be preferable because of its computational scalability. Ranking is indeed very efficient since it requires only the computation and sorting of P scores. In our heart beat classification task, the ranking approach is thus well suited in conjunction with more sophisticated models that do not have a closed form solution.

Embedded approaches

Embedded approaches consist in letting the learning process decide which variables to use. This is of particular interest for multiclass classifiers because it allows the learning process to consider distinct sets of variables for each class. Embedded methods are not new: decision trees such as CART, for instance, have a built-in mechanism to perform variable selection. The L_1-norm regularization is recently gaining popularity to embed feature selection in traditional models. The regularization of an objective function by the L_1-norm of its parameter vector encourages sparse solutions during the optimization process [107]. The parameter vector itself therefore reveals the discriminative features. The drawback of the L_1-norm regularization is the additional meta-parameter controlling the amount of regularization.

4.6 Application to ECG signals

In this section, several experiments concerning the supervised classification of heart beats are conducted. The database used in the experiments is first presented together with the features that are extracted from the heart beat time-series. Next, the methodology followed by the experiments is described. Finally, the results are presented.

4.6.1 ECG database and preprocessing

Data from the MIT-BIH arrhythmia database [44] are used in the experiments. The database contains 48 half-hour long ambulatory recordings obtained from 48 patients,

for a total of approximately 110'000 heart beats manually labeled into 15 distinct types. Following the AAMI recommendations, the four recordings with paced beats are rejected.

The inter-patient dataset configuration defined in [18], which has since been used in each inter-patient classification systems [70, 93], is used in this work. The 44 available recordings are divided in two independent datasets of 22 recordings each with approximately the same ratio of heart beats classes. The first dataset is the training set, and is used to build the model. The second dataset is the test set, and is used to obtain an independent measure of the performances of the classifier.

The sampled ECG signals are first filtered to remove unwanted artifacts using the filtering procedure proposed in [18]. Two median filters are designed for this purpose. The first median filter is of 200 msec width and removes the QRS complexes and the P waves. The resulting signal is then processed with a second median filter of 600 msec width to remove the T waves. The signal resulting from the second filter operation contains the baseline wanderings and can be subtracted from the original signal. Powerline artifacts are then removed from the baseline corrected signal with a 60 hz band-stop filter.

The location of R spikes and the associated beat types are provided with the database. These R locations serve as beat identifiers and the heart beats are recognized in the signals accordingly. The MIT-BIH heart beat labels are then grouped in the four classes defined by the AAMI recommendations (see Sec. 4.5.1). Table 4.3 shows the number of beats in each class and their frequencies in the two datasets. The class unbalance is obvious. Beats having a R-R interval smaller than 150 msec or higher than 2 seconds most probably involve segmentation errors and are discarded.

	N	S	V	F	Total
Training	45809	942	3784	413	50948
	89.91%	1.85%	7.43%	0.81%	100%
Test	44099	1836	3219	388	49542
	89.01%	3.71%	6.50%	0.78%	100%

Table 4.3: Distribution of heart beat classes in the two independent datasets. The class unbalance is obvious.

4.6.2 Feature extraction

A large variety of popular feature groups previously proposed for heart beat classification are extracted from the heart beat time-series. The feature groups involved in this work are R-R intervals (used in almost all previous works), segmentation intervals

[19, 18], morphological features [18, 79], Hermite basis function expansion coefficients (HBF) [59, 91, 93] and higher order statistics [90, 93]. We also introduce two additional features groups, corresponding to the normalized R-R intervals and the normalized segmentation intervals. Each group is populated with the following individual features:

1. Segmentation intervals (24 features): ECG characteristic points, corresponding to the onset and ending of P, QRS and T waves, are annotated in each beat using the unsupervised algorithm in [77]. A large variety of 24 features are then computed from the annotated characteristic points:

   (a) QRS wave: boolean flag indicating whether both Q and S points have been annotated, area, maximum, minimum, positive area, negative area, standard deviation, skewness, kurtosis, length, QR length, RS length;

   (b) P wave: boolean flag indicating whether its onset and ending have been annotated, area, maximum, minimum, length;

   (c) T wave: boolean flag indicating whether its onset and ending have been annotated, area, maximum, minimum, length, QT length, ST length.

   When the characteristic points needed to compute a feature failed to be detected in the heart beat segmentation step, the feature value is set to the patient's mean feature value.

2. R-R intervals (8 features): This group consists of four features built from the original R spike segmentations provided with the MIT-BIH database; the previous R-R interval, the next R-R interval, the average R-R interval in a window of 10 surrounding R spikes and the patient's mean R-R interval. The same four features are also computed using the R spikes detected by the segmentation algorithm.

3. Morphological features (19 features): Ten values are measured by uniformly sampling the ECG amplitude in a window defined by the onset and ending of the QRS complex, and nine other features in a window defined by the QRS ending and the T-wave ending. As the ECG signals are already sampled, linear interpolation is used to estimate the intermediate values of the ECG amplitude. Here again, when the onset or ending points needed to compute a feature were not detected, the feature value is set to the patient's mean feature value.

4. HBF coefficients (20 features): The parameters for the HBF expansion coefficients are chosen as in [93]: the order of the Hermite polynomial is set to 20 and the width parameter σ is estimated so as to minimize the reconstruction error for each beat.

5. High order statistics (30 features): The 2nd, 3rd and 4th order cumulant functions are computed. The parameters as defined in [91] are used: the lag parameters range from -250 msec to 250 msec centered on the R spike and 10 equally spaced sample points of each cumulant function are used as features, for a total of 30 features.

6. Normalized R-R intervals (6 features): These features correspond to the same features as in the R-R interval group except that they are normalized by their mean value for each patient. These features are thus independent from the mean normal behavior of the heart of patients, which can naturally be very different between individuals, possibly misleading the classifier. The normalization is obviously not applied to the R-R feature corresponding to the patient's mean itself, for a total of 6 features.

7. Normalized segmentation intervals (21 features): This group contains the same features as in the segmentation group, except that they are normalized by their mean value for each patient. The normalization is obviously not applied to boolean segmentation features. Here again, the objective is to make each feature independent from the mean behavior of the heart of a patient, because it can naturally be very different between individuals.

These features are illustrated in Figures 4.1 to 4.4. Several studies have shown that using the information from both leads can increase the classification performances [18, 70]; all features are therefore computed independently on both leads (except the four R-R intervals and the three normalized reference R-R intervals computed from the original segmentations since they are common to both leads), for a total of 249 individual features.

4.6.3 Methodology

Two distinct experiments are conducted. In the first experiment, it is observed whether the cost-sensitive learning paradigm actually improves the results of SVMs, CRFs and L_1-regularized CRFs in the context of heart beat classification. The weighted variant of these algorithms will later be referred to as weighted SVM (wSVM), weighted CRF (wCRF) and L_1-regularized weighted CRF model (wCRF+L_1). In the second experiment, these three proposed weighted models are compared to the state-of-the-art weighted LDA model (wLDA) introduced by [18]. We now detail these two experiments:

1. In the first experiment, the effect of the class unbalance on the performances of the standard SVM, CRF and CRF+L_1 models is evaluated. To this end, the

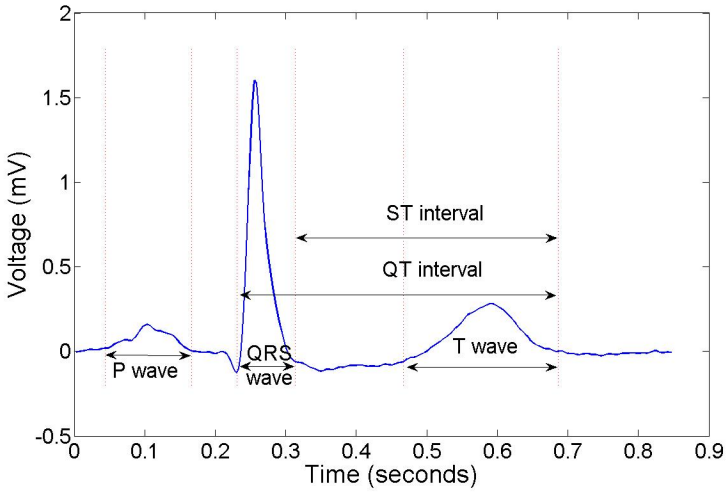

Figure 4.1: One normal heart beat and segmentation interval features: showing P length, QRS length, T length, ST interval and QT interval features.

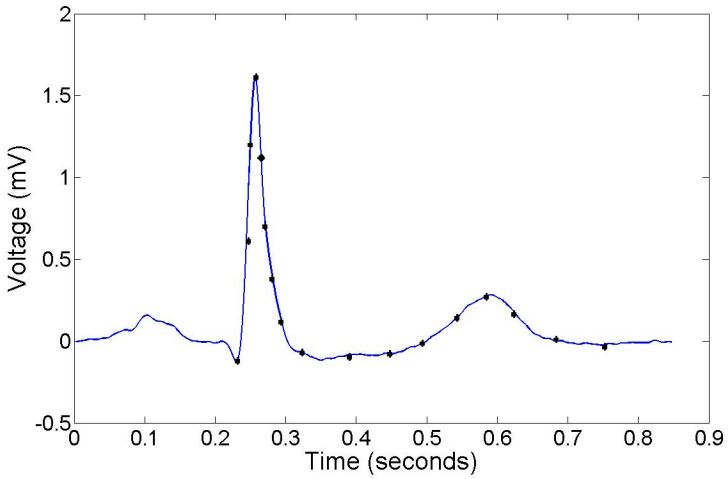

Figure 4.2: Morphological features: 19 features corresponding to 10 equally sampled points in the QRS complex and 9 equally sampled points in the S-T interval.

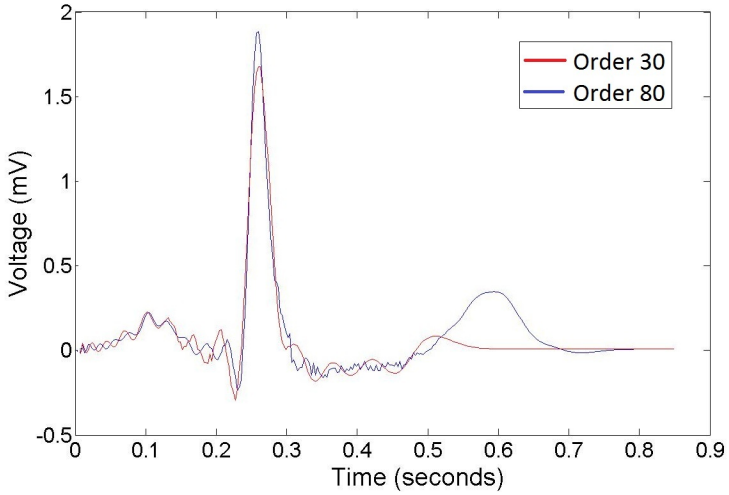

Figure 4.3: Beat reconstruction using 30 and 80 HBF coefficients.

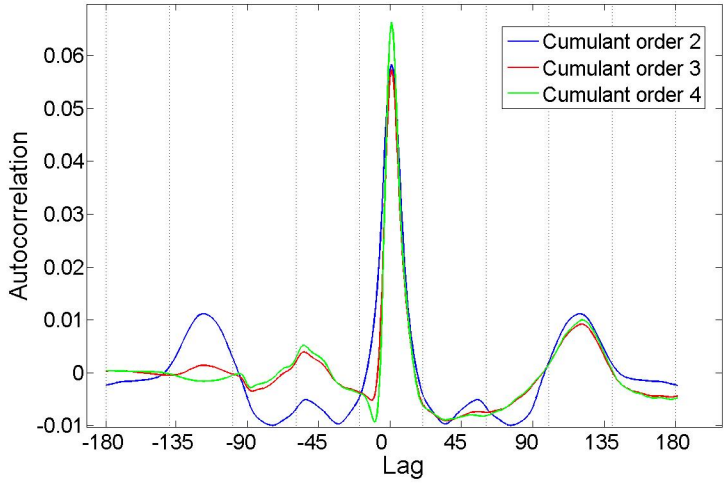

Figure 4.4: High-order statistics features: 10 equally sampled points, illustrated by the vertical lines, are taken from the three cumulants.

performances of these three classifiers are evaluated (1) when no weights are included in the model, (2) when weights corresponding to the inverse of the class priors are included in the model and (3) on an undersampled dataset where a subsample of each class is randomly selected to have equal priors. The number of elements that are randomly selected in each class corresponds to the number of observations in the minority class (the F class, 413 observations). In this experiment, the features are selected by a filter ranking procedure with the MI criterion. The histogram-based estimator [82] is used to estimate the mutual information value between each feature and the class labels. For the randomly undersampled dataset, an average of the performances over ten random selections is reported.

Since the objective of this experiment is not to report optimal generalization performances but rather to evaluate the impact of the class unbalance in equal configurations, the top 5 ranked features are empirically selected. The models are trained on the training set and the performances on the test set are reported. The best performances in terms of BCR that can be achieved on the test set are reported for the SVM and the CRF+L_1 classifiers, by scanning a large range of hyper-parameter values. In other words, in this experiment, the test set acts as a validation set. The results of this experiment therefore do not allow us to report the real performance of the models but only allow us to illustrate the importance of addressing the class unbalance in the classifiers.

2. In the second experiment, the performances of the three weighted models are compared to the state-of-the-art LDA classifier introduced by Chazal [18]. It is expected that the performances are increased by the proposed wSVM classifier with a non-linear kernel. The benefit of the first-order Markov dependencies integrated in the wCRF classifier is also investigated. Finally, the advantages offered by the L_1-norm regularization of the wCRF classifier are evaluated. In this experiment, the hyper-parameters of the wSVM and of the wCRF+L_1 classifiers are estimated by a "leave-one-patient-out" cross-validation procedure on the training set. The BCR, estimated by the geometrical mean of the class accuracies, is used as performance measure.

   The state-of-the-art wLDA-based algorithm performs feature selection by evaluating very few combinations of three feature groups (R-R intervals, morphological features and segmentation features). As described in details in Sec. 4.5.3, the results can be suboptimal. The training of the wLDA classifier does not require the estimation of any hyper-parameter and has a closed-form solution. It is therefore evaluated whether the performances of the wLDA can be increased with a forward wrapper feature selection strategy on the whole feature set. More

formally, the criterion used in the forward search strategy is the BCR obtained by the wLDA model with a "leave-one-patient-out" cross-validation procedure on the training set. The forward search is halted when no supplementary feature can increase the BCR.

On the other hand, the other models do not have a closed-form solution and require the estimation of their hyper-parameters by a "leave-one-patient-out" cross-validation. A forward search strategy is thus clearly computationally intractable (in the order of several months of computer time on recent computers). For this reason, a filter approach with a mutual information ranking criterion is used in conjunction with the two CRF models and the SVM model. According to expert knowledge [19] and to preliminary experiments, the number of discriminative features is known to be typically much smaller than the number of features in our dataset. Therefore, and mostly in order to ensure computational tractability, only the top ten ranked features are considered in this experiment. The optimal number of features between one and ten is chosen by the "leave-one-patient-out" cross-validation on the training set, as an additional hyper-parameter.

The final models are obtained by training the four models on the complete training set with their selected hyper-parameters, including the selected feature subset. The final performances are then evaluated on the test set - which has never been involved in any computation before - as a fair measure of their real generalization capabilities on unseen data.

Since a sufficient number of data is available to obtain reliable class prior estimations, the cost parameters in the weighted methods are set to the inverse of the class priors in all the experiments, as previously detailed in Sec. 4.3.

4.6.4 Results

The results of the first experiment are presented in Table 4.4. Clearly, the performances of the three models without tackling the unbalance are unsatisfactory for the two smallest classes. In particular, each unweighted model yields an accuracy below 5% for the S class on the full dataset, which is clearly unacceptable. The subsampling of the dataset improves the performances of the SVM classifier, but it still achieves poor results for the minorities. The performances of both CRF classifiers dramatically collapse with the subsampling, probably because discarding observations alters the temporal dependencies between subsequent heart beats and their labels. Hence the first-order Markov transitions in the CRF model or the features built over consecutive beats (e.g. R-R intervals) do not make sense anymore.

Data	Model	BCR	Acc.	N	S	V	F
Full	SVM	53.22%	92.86%	98.23%	0.16%	79.18%	35.30%
Full	CRF	59.17%	89.16%	93.33%	3.37%	85.08%	54.89%
Full	CRF+L_1	59.26%	89.33%	93.52%	3.21%	85.18%	55.15%
Sub.	SVM	65.47%	80.63%	82.54%	38.12%	81.17%	60.05%
Sub.	CRF	18.40%	63.37%	71.06%	0.00%	1.80%	0.77%
Sub.	CRF+L_1	81.82%	24.76%	91.58%	6.53%	0.94%	0.00%
Full	wSVM	83.80%	78.57%	77.89%	81.32%	84.75%	91.24%
Full	wCRF	78.00%	76.62%	76.11%	72.22%	86.11%	77.58%
Full	wCRF+L_1	80.30%	76.90%	76.01%	81.15%	86.70%	77.32%

Table 4.4: Performances of the SVM, CRF and CRF+L_1 models when no weights are included in the cost functions on the full dataset, when an undersampled dataset with equal priors is considered, and when weights corresponding to the inverse of the class priors are included in the cost functions on the full dataset. N, S, V and F are the accuracies of the normal, supraventricular, ventricular and fusion class respectively. *Sub* is the subsampled dataset and *Full* is the complete dataset.

On the other hand, the class accuracies for the minority classes are significantly increased by the weighting of the objective functions. This class accuracy increases to around 80% when the weights are included in the model. Nevertheless, the increase of accuracy in small classes seems to be in disfavor of the accuracy of the majority class, which decreases of around 15% with the weighting. The results also reveal that the BCR is a much more suitable criterion than the global accuracy. For example, the SVM model with no weights leads to the best overall global accuracy although having a class accuracy below 50% for two classes. By contrast, this model yields the smallest BCR with only around 50%. The BCR criterion therefore adequately penalizes models with clearly unsuitable class accuracies.

The results of the second experiment are shown in Table 4.6. The state-of-the-art configuration based on the wLDA model and 26 features achieves a BCR of 71.39%. The forward feature selection strategy does not improve the performances of the state-of-the-art wLDA model. This can be explained by looking at the performances on the training set. The wLDA model with the forward selection obtains a BCR of 81.13% on the training set, and the BCR falls down to 68.76% on the test set. Hence, the forward selection with the LDA model actually seems to overfit the training data and to generalize poorly. In the heart beat classification task, errors in the pathological classes (i.e., missing a cardiac disease) can have dramatic consequences while errors in the normal class (i.e., incorrectly diagnosing a cardiac disease) are of course undesired but still not life-threatening. The pathological classes are therefore of uttermost importance.

Pos.	Description	Lead	wSVM	wCRF+L_1	wCRF
1	Previous R-R (normalized)	Ref.	•	•	•
2	T wave amplitude (normalized)	1	•	•	•
3	2nd-order statistic at -40msec	1	•	•	•
4	2nd-order statistic at +40msec	1	•	•	•
5	2nd-order statistic at -166msec	1	•	•	•
6	2nd-order statistic at 166msec	1		•	•
7	T wave interpolation at 50%	1			•
8	Previous R-R	Ref.			•
9	Next R-R (normalized)	Ref.			•
10	T wave interpolation at 60%	1			

Table 4.5: Top 10 features as ranked by the MI criterion. The features from this list that are selected by the models proposed in this work are also shown. For the 7th and 10th features, the percentages indicate the value of the interpolation at this particular percentage of the wave length.

The wLDA model, with any of the two feature choices, achieves unsatisfactory results in this aspect.

On the other hand, the non-linear polynomial wSVM model achieves a BCR of 82.45% with only five features. In particular, the wSVM model yields an accuracy over 80% for all the pathological classes. The wCRF model obtains results very similar to the ones of the wSVM model. When the L_1-regularization is added to the wCRF model, the best overall results are obtained with a BCR of 85.39% and an accuracy close to 85% for each pathological class. Moreover, the wCRF+L_1 model requires only 6 features from the 249 extracted features to achieve these results.

Table 4.5 holds the top 10 features, as ranked by the MI criterion, and reveals which of these features were selected by the models proposed in this work. As it can be observed from the table, the important features seem to be R-R intervals, the amplitude and length of the T wave and 2nd-order statistics (the autocorrelation function). The top 2 features are from patient-normalized feature sets. These results therefore validate the relevance of the normalization of the features. On the other hand, several popular feature sets do not seem to serve the classification performances. No features were indeed selected by the models from the HBF coefficients, the 3rd and 4th order statistics and the unnormalized segmentation intervals. Furthermore, it does not seem necessary to extract features on both leads since only features from the original annotations and from the first lead are selected.

Model	Features (#)	BCR	N	S	V	F
wLDA	Chazal [18] (26)	71.39%	88.63%	44.66%	80.58%	81.44%
wLDA	Wrapper (16)	61.76%	82.31%	58.12%	67.04%	45.36%
wSVM	Ranking (5)	82.45%	77.65%	84.48%	85.43%	82.47%
wCRF	Ranking (9)	81.29%	76.55%	80.72%	86.24%	81.96%
wCRF+L_1	Ranking (6)	85.39%	79.78%	92.59%	85.12%	84.54%

Table 4.6: Performances of the wSVM, wCRF and wCRF+L_1 models on the test set. The performances of the state-of-the-art wLDA-based algorithm in [18] are also first reproduced with the original feature choice and next evaluated with a wrapper selection strategy. The hyper-parameters and the number of features are selected on the training set by "leave-one-patient-out" cross-validation. N, S, V and F are the accuracies of the normal, supraventricular, ventricular and fusion class respectively.

4.7 Discussion

The classification of heart beats is of crucial importance for clinical applications involving the long-term monitoring of the cardiac function. In this chapter, the importance of the AAMI guidelines for the design of reliable automatic heart beat classifiers and for the evaluation of their relative merits is outlined. The inter-patient and intra-patient classification paradigms are also defined to this end. The inter-patient paradigm implies generalization from one patient to another and offers several advantages in practical situations compared to intra-patient classification.

The first main difficulty in the design of a heart beat classifier is the class unbalance. With temporal data such as ECG signals, any learning technique based on undersampling or oversampling is difficult to achieve. First, because the ratio between the size of the majority and the minority class is too high. For example, in our dataset, undersampling the majority class would imply to keep less than 1% of its observations, which is clearly a dramatic loss of information. Second, any artificial reduction or increase in the observations of a class must be done carefully to maintain first-order dependencies between labels of subsequent observations. For these reasons, this work has followed the cost-sensitive learning approach which does not require any change in the number of observations.

The cost-sensitive learning approach enables standard classifiers to handle the class unbalance by weighting the errors associated to each class with respect to their priors. In particular, the classical objective functions of the SVM classifier and of the (regularized) CRF model can be modified accordingly. Experiments on real Holter recordings show that the weighted models yield results significantly better than their non-weighted version both on the full dataset and on a subsampled dataset with equal

priors. Results also show that the balanced classification rate is a performance metric which is much more suitable than the usual accuracy in the case of unbalanced classes.

The second difficulty is the extraction and the selection of discriminative features from the heart beat time-series. In this chapter, a large variety of features proposed in the literature is reviewed. Two additional patient-normalized feature sets are also proposed. The selection of discriminative features is of great importance to help interpreting models and to increase the performances by removing spurious features. Inter-patient classification of heart beats following AAMI standards involves a large number of observations and of features, several distinct patients and multiclass labels. These characteristics must be taken into account when choosing the feature selection strategy associated to a given model. In particular, a wrapper approach can only be affordable for models where no hyper-parameters have to be tuned by "leave-one-patient-out" cross-validation (which is necessary for the inter-patient classification) or for models having a closed-form solution. Otherwise, the ranking approach using the mutual information criterion is a more tractable alternative.

Classification experiments are conducted on real Holter recordings to compare the proposed weighted SVM, weighted CRF and weighted L_1-regularized CRF to the state-of-the-art weighted LDA model. Results show that the weighted LDA model, even with a forward wrapper selection strategy, yields unsatisfactory results for the pathological classes. By contrast, best results are achieved by the L_1-regularized weighted CRF model with a BCR of 85.39% and an accuracy close to 85% for each pathological classes. These results show that the information contained in the first-order dependency in class labels increases the performances significantly. Moreover, the regularized weighted CRF model requires only 6 features from the 249 extracted features to achieve these results. These findings reveal that several previously reported feature sets do not serve the classification process and that a very small number of features are actually required to yield high performances. Moreover, the top two selected features are patient-normalized features, which validate the discriminative power of the proposed patient-normalized features.

It is worth mentioning that variants of SVMs have been proposed to integrate the dependence between observations in sequence classification. The task is not so trivial when dealing with multiclass problems solved by several binary classifiers, since each binary classifier is trained on a distinct subsample of the data and hence has no knowledge of observations from other labels. Nevertheless, multiclass-enabled SVMs have been proposed and extended to sequence classification. For example, hidden Markov support vector machines (HM-SVMs) [7] perform non-linear and maximum margin classification of sequences. Although these are promising theoretical advantages over CRFs, experiments have shown that HM-SVMs do not significantly outperform CRFs and suffer from a major computational burden [7].

The most important challenge that must be addressed in future works is the reduction of the dataset while keeping first-order dependencies between labels. Adjacent beats are indeed very similar and a large portion of the beats in the training set could probably be pruned out without any loss in performance. The reduction of the dataset would enable the use of more powerful feature selection techniques than the ranking filter technique which misses covariate information. Two approaches could be investigated to reduce the dataset. First, on-the-fly techniques iteratively estimate a similarity value between a new heart beat and the previously extracted beat(s). The new beat is included in the dataset if the similarity with the previous beat(s) is sufficiently small. Second, clustering techniques involve the extraction of all the beats in the signal and subsequent reduction by computing and processing the full $N \times N$ similarity matrix. The advantage of the first technique is that it does not require the computation and storage of the similarity matrix which can be quite problematic in practical situations. By contrast, the second technique allows non adjacent beats to be compared. Preliminary experiments tend to show that there is a dependency between beats separated by a long duration and the second approach is therefore much more likely to provide better results.

House: “The body does crazy things.”
Foreman: “Well, that explains everything.”

House MD, season 2 episode 9.

Chapter 5

Elimination of ECG Contamination from Vagus Nerve Recordings

Contents

5.1 Introduction

Due to the proximity of the recording sites to the heart and beating arteries, the recordings of several physiological signals are typically contaminated with noise related to the cardiac activity, including the ECG. These artifacts can significantly distort the frequency spectrum and the amplitude of the sought signal, thus presenting a major challenge to the subsequent extraction of accurate and useful information.

This problem has been identified in the electromyogram (EMG) collected from the trunk muscles, where the proximity to the heart and the volume conduction characteristics induce a cross talk from the ECG through the torso. The same situation appears in the non invasive recordings of fetal ECGs from electrodes on the maternal abdomen; the signal to noise ratio of these recordings is relatively low and the main undesired signal here is the maternal ECG.

In the ADVENS project, vagus nerve electroneurograms (ENG) have been obtained in the hope to get better insight in the nerve interaction with several pathologies such as refractory epilepsy and severe depression. It has been observed that these recordings suffer from ECG contamination. These neural signals are typically characterized by higher frequency contents than ECG or EMG signals, but the heart related contamination is not only electrocardiographic in nature: it can contain plethysmographic and movement artifacts as well.

In this chapter, methods for the removal of cardiac artifacts in vagus ENG recordings are investigated. This is, to our knowledge, the first study investigating the filtering of ECG activity in neurograms. In small animals, the design of small cuffs carrying multiple recording contacts can be challenging. For this reason, it is not always possible to include many recording contacts in the cuff. In this chapter, both single and multiple channel filtering methods are considered separately. The experiments are conducted on real ENG recordings from the vagus nerve in rats. These experiments examine whether the use of multiple contacts improves the filtering performances. Personal publications over the materials in this chapter are [28, 76, 75, 33].

5.2 Vagus nerve recordings

In this section, the properties of the nerve signals and the motivations for recording vagus nerve ENG are described. A nerve can be seen as a cordlike structure that contains many axons. Axons are the long slender projections of neurons, referred to as fibers to designate the axon and the associated myelin sheet. Very schematically, the axon can be compared to an electrical cable linking neurons. Axons transport information under the form of propagated regenerative electrical impulses that result from changes in the resting transmembrane potential known as action potentials (AP). The

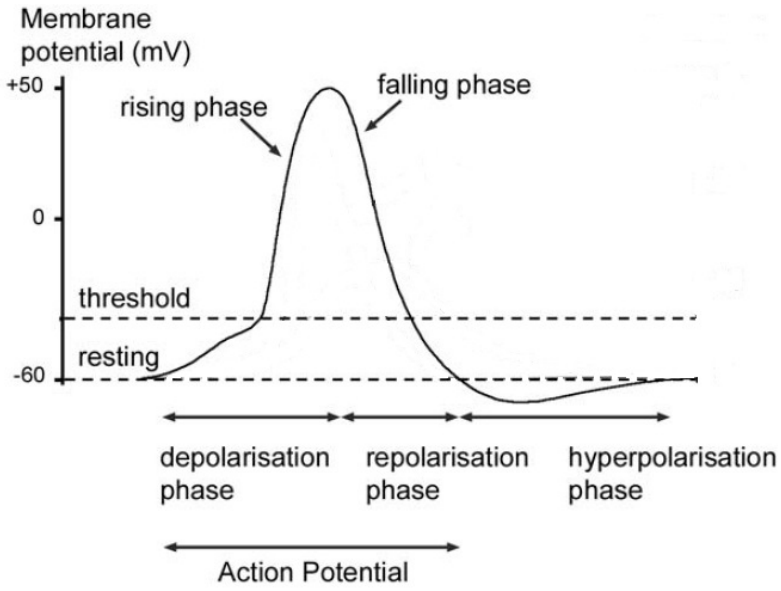

Figure 5.1: An action potential, reproduced from [115]. The membrane has a resting potential of -60 mV. The *all-or-none* law states that only a depolarization higher than a threshold value can initiate the action potential.

all or none law requires a sufficient depolarization to generate and propagate an AP. During an AP, recordings of the membrane currents show a fast, transient depolarizing followed by a delayed and sustained hyperpolarizing current as illustrated in Fig. 5.1 [4, 115]. The signal generated across the neuronal membrane is then transported over a relatively long distance at high velocity without distortion or decrement. The conduction velocity varies in different nerve fibers, depending on the axon diameter, the myelin thickness and the internodal length [112].

The nerves of the autonomic nervous system (ANS) are of major importance to ensure involuntary functions in the body. The ANS is classically divided into two complementary subsystems that typically function in opposition to each other: the parasympathetic nervous system (PNS) and the sympathetic nervous system (SNS). For an analogy, one may think of the sympathetic division as the accelerator and the parasympathetic division as the brake. The vagus nerve is the supplier of efferent (from the ANS to the organs) motor parasympathetic fibers to all the organs (except the suprarenal glands), from the neck down to the second segment of the transverse colon. This means that the vagus nerve is responsible for such varied tasks as heart

rate, gastrointestinal peristalsis and some muscle movements in the neck area, including speech. Besides this output to the various organs in the body the vagus nerve also conveys sensory information about the state of the organs to the central nervous system. Actually 80-90% of the nerve fibres in the vagus nerve are afferent (from the organs to the ANS) fibres communicating the state of the viscera to the brain.

The treatment of several refractory neurological or psychiatric diseases involves invasive electrical stimulation of the vagus nerve through an implanted a pacemaker-like stimulation device. Vagus nerve stimulation (VNS) has for example become a recognized therapy for intractable epilepsy and severe depression [8]. Nevertheless, the exact mechanisms involved with this treatment remain largely unknown. In the ADVENS project, experimental recordings of the vagus nerve activity aiming at a better understanding of those mechanisms have recently been achieved.

For this purpose, nerve cuff electrodes offer many advantages over standard electrodes such as silver hook electrodes. They make it possible to include several stimulation and recording contacts in a single cuff and to obtain long-term recordings. The cuff electrode is a self-sizing spiral cuff made of two silicon rubber sheets carrying platinum contacts between them. They are designed to tightly fit the nerve while allowing flexibility in the case of swelling.

Despite the many advantages offered, since the cuff electrode contacts are at the surface of the nerve, they can only record multiple action potentials and have a rather poor recording signal-to-noise ratio compared with nerve penetrating alternatives, the latter being able to record single fiber action potentials such as the one in Fig. 5.1. This represents a major challenge for the subsequent extraction of accurate and useful information from signals obtained with cuff electrodes. In particular, cardiac activities induce a major noise component in the signal.

5.3 Filtering in single channel recordings

In this section, three techniques for the removal of ECG contamination in singe channel recordings are presented: high-pass filtering [98], template subtraction [38] and the discrete wavelet transform [10].

5.3.1 High-pass filtering

High-pass frequency filtering has been applied as a simple and potentially efficient method for the removal of artifacts in EMG signals [98]. The major advantages of this method are its simplicity and its popularity. On the other hand, the order of the filter must be chosen empirically while this parameter can significantly affect the results. Furthermore, this method is only effective if the artifact spectral content is limited and

does not overlap with the signal spectrum. The high-pass filter considered in this work is the classical symmetric finite impulse response filter (FIR) with a Hamming-window based linear-phase transfer function [36].

5.3.2 Template subtraction

The template subtraction algorithm has shown a great potential for filtering EMG signals when the ECG signal can be recorded as supplemental data [38]. The general idea is that a template is subtracted from the nerve signal at each occurrence of a heart beat. The occurrence of the heart beats are measured on the recorded ECG signal, for example by using a QRS detection algorithm (remember Chapter 2).

Once the contamination times have been identified on the ECG signal, the template subtraction algorithm involves two steps. First, a template is created from the nerve signal at the contamination times. It is estimated by averaging neural recording samples corresponding to each QRS occurrence. In the second step, the template is subtracted from the neural signal at each contamination time.

The main advantage of this method is that it is not blind: it makes use of the information from the ECG signal. Moreover, because the template is estimated from the signal itself, any delay between the heart ECG signal and the ECG contamination is taken into account. However, the size of the template must be chosen carefully. It must be long enough to capture P and T waves but it must not overlap between successive heart beats. Since the heart is not a periodic oscillator, the choice of the template size can be very challenging. Furthermore, the method is very dependent on the reliability of the QRS detection algorithm.

5.3.3 Discrete wavelet transform

The discrete wavelet transform (DWT) [73, 3] is a time-frequency representation of a signal that was introduced to overcome the limitations in time resolution of classical frequency transforms such as the Fourier transform. The DWT can be summarized as follows: the signal x_t is first passed through two filters: a lowpass to extract the approximations coefficients and a highpass to extract the detail coefficients. Next, in the decomposition tree, only the approximation details are passed again through the filters until the last level of the decomposition. At each level, the frequency band of the signal and the sampling frequency are halved. The DWT algorithm can be implemented by a bank of bandpass filters each having a frequency band and a central frequency half the previous one.

More formally, the DWT of a signal x_t can be written as

$$T^{m,n} = \int_{-\infty}^{\infty} x_t \phi_t^{m,n} dt, \tag{5.1}$$

with

$$\phi_t^{m,n} = \frac{1}{\sqrt{a}}\phi\left(\frac{t - nba^m}{a^m}\right) \tag{5.2}$$

where $T^{m,n}$ is known as the detail coefficient at scale m and location n, $a > 1$ is a fixed dilatation step parameter and $b > 0$ is the location parameter. In its most common form, the DWT employs a dyadic grid and orthonormal wavelet basis functions and exhibits zero redundancy. A common choice for these parameters are $a = 2$ and $b = 1$. This power of two logarithmic scaling of both dilatation and translation steps is known as the dyadic grid arrangement. Substituting these values in Eq. (5.2), the dyadic grid wavelet can be written compactly as

$$\phi_t^{m,n} = 2^{-m/2}\phi(2^{-m}t - n). \tag{5.3}$$

Orthonormal dyadic discrete wavelet are associated with scaling functions $\varphi_t^{m,n}$ having a form similar to wavelet functions:

$$\varphi_t^{m,n} = 2^{-m/2}\varphi(2^{-m}t - n). \tag{5.4}$$

The scaling function can be convolved with the signal to produce approximation coefficients as follows:

$$A^{m,n} = \int_{-\infty}^{\infty} x_t \varphi_t^{m,n} dt. \tag{5.5}$$

The signal x_t can then be represented using a combined series expansion using both the approximation coefficients at a given level m_0 and all successive detail coefficients:

$$x_t = \sum_{n=-\infty}^{\infty} A^{m_0,n}\varphi_t^{m_0,n} + \sum_{m=-\infty}^{m_0}\sum_{n=-\infty}^{\infty} T^{m,n}\phi_t^{m,n}. \tag{5.6}$$

The signal detail coefficients at scale m is defined as

$$D_t^m = \sum_{n=-\infty}^{\infty} T^{m,n}\phi_t^{m,n}. \tag{5.7}$$

A large bank of predefined wavelet exists, for example the Mexican hat wavelet is defined as

$$\phi_t = (1 - t^2)e^{\frac{-t^2}{2}} \tag{5.8}$$

and is shown in Fig. 5.2.

The DWT has been used with success for artifact removal in EMG signals [10]. Its application is based on the assumption that spectral separation between the original signal and the artifact is possible. Artifact removal can then be performed by decomposing the signal until the level at which approximation coefficients mainly correspond to ECG activity, as judged visually. The reconstructed and summed detail coefficients at previous levels then correspond to signal activity. The main drawback of the DWT is that the wavelet function and the decomposition order must be empirically chosen.

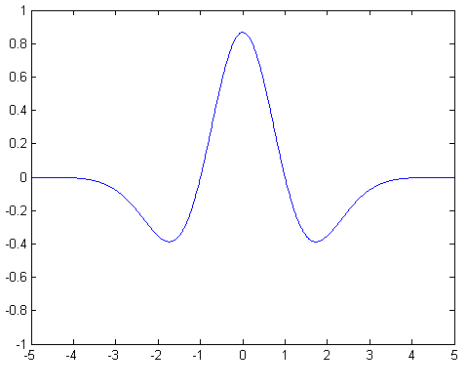

Figure 5.2: The Mexican hat wavelet.

5.4 Filtering in multiple channel recordings

One of the main advantages of the cuff electrode is the possibility to include multiple recording contacts in the cuff. This opens the path to the use of blind source separation (BSS) methods for the extraction of cardiac artifacts. The BSS problem consists in separating a multivariate signal (called observations or mixtures) into additive subcomponents (called sources) without the aid of information about the source signals or the mixing process. The heart related artifacts can thus be isolated as one of the estimated sources and the clean vagus nerve signals can then be reconstructed by discarding it.

In this section, we first introduce traditional methods for solving the BSS problem and show why these methods are not suitable for the extraction of ECG artifacts in vagus neurograms. Semi-blind source separation methods exploiting the time structure of the data are then introduced as an alternative.

5.4.1 Blind source separation

Let us define $\mathbf{x}$, a multivariate sequence of temporal observations $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T}$ with $\mathbf{x}'_t = [x_t^p]_{p=1}^{P} \in \mathbb{R}^P$ and the multivariate sequence of source observations $\mathbf{s} = \{\mathbf{s}_t\}_{t=1}^{T}$ with $\mathbf{s}'_t = [s_t^p]_{p=1}^{P} \in \mathbb{R}^P$. The BSS model is a linear transformation of the form:

$$\mathbf{s}_t = \mathbf{U}'\mathbf{x}_t \tag{5.9}$$

$$\mathbf{U} = \mathbf{Q}\mathbf{H}^{-1/2}\mathbf{W} \tag{5.10}$$

where $\mathbf{Q}$ and $\mathbf{H}$ are respectively the eigenvector and eigenvalue matrix of the covariance matrix $\mathbf{C} = E_t\{\mathbf{x}_t\mathbf{x}'_t\}$ and $\mathbf{W}$ is a rotation matrix left undefined for now. The

transformation by **Q** and **H** performs whitening (decorrelation and sphering) of the data, and **W** yields a stronger separation than only decorrelation. The estimation of **W** is the focus of source separation methods.

The underlying assumptions are first that the mixing matrix is invertible and square. In other words, the maximum number of sources that can be estimated is equal to the number of observed mixtures. Second, the mixing is instantaneous: there is no delay in the mixing process. The latter assumption is rather strong in physiological signals where volume conduction velocity may vary and convolutive BSS algorithms have been developed for that purpose, although not considered in this work (see [50], Chapter 19 for further details).

BSS algorithms thus rely on prewhitening of the data, typically achieved using principal component analysis, leaving only the rotation matrix **W** to be estimated. One of the most popular techniques for solving the BSS problem is independent component analysis (ICA) [50] where the matrix **W** is estimated by assuming that the source signals are statistically independent and that a maximum of one of the sources has a Gaussian distribution. The sources are estimated by searching for the transformation that maximizes the nongaussianity of the whitened observations, typically estimated by kurtosis or negentropy criteria [50].

However, this nongaussianity assumption does not always hold. In the case of a neural signal obtained by a cuff electrode, remember that the recorded signal is a multiple action potential: the sum of the activity of numerous fibers. By the central-limit theorem, the resulting signal distribution is expected to be approximately Gaussian. Other assumptions should therefore be made in the BSS model.

5.4.2 Semi-blind source separation

In recent work over fetal ECG extraction, the temporal structure of the sources has been exploited in source separation algorithms. For example, the nongaussianity assumption has been replaced by a non-white assumption (non-zero autocorrelation). These models are often referred to as semi-blind source separation (semi-BSS) models, because they require some kind of a priori knowledge about the temporal structure of the sources.

The time structure of the sources is summarized by the $P \times P$ autocovariance matrix $\mathbf{C}_\tau$ defined as

$$\mathbf{C}_\tau^x = E\{\mathbf{x}_t \mathbf{x}'_{t-\tau}\} \tag{5.11}$$

for a given lag τ. The key point in semi-blind algorithms is that the information in this autocovariance matrix can be used to estimate **W** in Eq. (5.10) instead of the higher-order statistics used in ICA methods [118]. In addition to make the instantaneous covariances of $\mathbf{s}_t$ go to zero by the whitening process, the lagged autocovariances are

made zero as well:

$$E\{s_t^p s_{t-\tau}^q\} = 0, \quad \forall\, p,q,\tau. \tag{5.12}$$

AMUSE

The first semi-BSS algorithm exploiting the autocovariance matrix is the AMUSE algorithm [118], which was later redefined in [11]. The AMUSE algorithm jointly whitens the data and diagonalizes the autocovariance matrix for a given lag τ. Assume the desired source signal $\mathbf{s}^p$ is temporally correlated, satisfying the following relations for a specific time delay τ:

$$E\{s_t^p s_{t-\tau}^p\} > 0 \tag{5.13}$$

$$E\{s_t^p s_{t-\tau}^q\} = 0 \tag{5.14}$$

$$E\{s_t^q s_{t-\tau}^r\} = 0,\ \forall q \neq p, r \neq p. \tag{5.15}$$

Let us define the prewhitened observations $\mathbf{z}_t = \mathbf{Q}\mathbf{H}^{-1/2}\mathbf{x}_t$. Under the constraint $\|\mathbf{w}\| = 1$, finding

$$\max_{\mathbf{w}} \mathbf{w}' E\{\mathbf{z}_t \mathbf{z}_{t-\tau}'\}\mathbf{w} \tag{5.16}$$

where $\mathbf{w}$ is the first row of the matrix $\mathbf{W}$, leads to the first desired source signal. It can be shown that maximizing Eq. (5.16) is similar to solving the following eigenvalue-eigenvector decomposition [118]:

$$\mathbf{W} = \mathrm{EIG}(\tilde{\mathbf{C}}_\tau^z) \tag{5.17}$$

where $\mathrm{EIG}(\mathbf{A})$ is an operator that returns the matrix whose rows are formed by the eigenvectors of $\mathbf{A}$, sorted in decreasing order of the corresponding eigenvalues. The transformation $\tilde{\mathbf{C}}_\tau^z = \frac{1}{2}(\mathbf{C}_\tau^z + \mathbf{C}_\tau'^z)$ ensures the symmetry of $\mathbf{C}_\tau^z$. In PCA, the eigenvalues from the decomposition of the correlation matrix measure the amount of variance explained by the associated component. Here, the eigenvalues of the autocorrelation matrix measure the amount of periodicity in τ of each component. In other words, the component associated to the largest eigenvalue has the most periodicity in τ, and the component associated to the smallest eigenvalue has the least periodicity in τ.

For two $P \times P$ symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, the problem of generalized eigenvalue decomposition (GEVD) of the matrix pair $\{\mathbf{A},\mathbf{B}\}$ consists in finding the matrices $\mathbf{V}$ and $\mathbf{D}$ such that $\mathbf{V}'\mathbf{A}\mathbf{V} = \mathbf{D}$ and $\mathbf{V}'\mathbf{B}\mathbf{V} = \mathbf{I}$ where $\mathbf{D}$ is the diagonal generalized eigenvalue matrix corresponding to the eigenvector matrix $\mathbf{V}$. Therefore, $\mathbf{V}$ is a transformation that simultaneously diagonalizes $\mathbf{A}$ and $\mathbf{B}$. It follows that the complete unmixing matrix $\mathbf{U}$ can directly be estimated by computing the GEVD of the covariance and autocovariance matrices of $\mathbf{x}_t$ together:

$$\mathbf{U} = \mathbf{Q}H^{-1/2}\mathbf{W} = \mathrm{GEVD}(\tilde{\mathbf{C}}_\tau^x, \mathbf{C}_0^x). \tag{5.18}$$

The AMUSE algorithm for the extraction of all the P sources can then be summarized as follows:

1. Find τ, the most significant autocorrelation lag in the desired signal.
2. Compute $\mathbf{C}_\tau^x$ and $\mathbf{C}_0^x$ using Eq. (5.11);
3. Compute $\tilde{\mathbf{C}}_\tau^z = \frac{1}{2}(\mathbf{C}_\tau^z + \mathbf{C}_\tau'^z)$;
4. Compute GEVD$(\tilde{\mathbf{C}}_\tau^x, \mathbf{C}_0^x)$;
5. The rows of the matrix $\mathbf{U}$ are given by the eigenvectors, ranked in descending order of their corresponding eigenvalues;
6. The components of $\mathbf{s}_t$ are given using Eq. (5.9) and are sorted according to their periodicity with the R peaks of the ECG.

Robust AMUSE

In [67], a more robust estimation of the autocovariance matrix is used in the GEVD decomposition in Eq. (5.18):

$$\mathbf{C}_\tau^x = \frac{1}{D}\sum_{d=1}^{D}(\mathbf{C}_{d\tau}^x + \mathbf{C}'^x_{d\tau}) \tag{5.19}$$

where D corresponds to the number of multiplicative factors of τ and must be chosen by the user. A signal having a significant autocorrelation at lag τ will indeed also have significant autocorrelations at lags $d\tau$ for any choice of d, and the idea is to include this information in the diagonalization process. The authors [67] show that the estimation of this matrix is more robust to the number of available samples and to small lag estimation errors.

Is it worth mentioning that if the objective is the extraction of only one source of interest, semi-blind source extraction (semi-BSE) can be performed rather than semi-BSS. The algorithm in [68] performs robust AMUSE iteratively by extracting only one source of interest at a time via iterative approximate joint diagonalization of autocorrelation matrices. Semi-BSE has some advantages over semi-BSS such as increased flexibility, lower computational complexity and extraction of only the sources of interest. Nevertheless, such BSE methods are not well suited for filtering of the extracted source since the full mixing matrix is not estimated.

SOBI

In situations where no value for the τ parameter can empirically be chosen, the SOBI (Second-Order Blind Identification) algorithm [13] performs a joint diagonalization of an arbitrary set of v autocovariance matrices with consecutive time delays $\tau = 1$ to v. Since no parameter has to be a priori chosen (beside the maximal number of considered time delays v), this method rather belongs to blind methods but still permits the separation of Gaussian sources. The difficulty arises in the joint diagonalization of more than 2 matrices which can be solved by extending the Jacobi technique for diagonalizing a unique matrix to the joint approximate diagonalization of a set of matrices [13].

Periodic Component Analysis

Although ECG signals have a periodic structure that is repeated in every cycle of the heart beat, normal ECGs can have a variation in R-R periods up to 20% [44]. Hence, the τ parameter does not fully describe the periodicity of the ECG signal. The periodic component analysis (πCA) algorithm [104] is merely a restatement of the AMUSE algorithm specifically customized for quasi-periodic signals such as the ECG signal. The difference arises in the integration of a time-varying delay in the estimation of the autocovariance matrix.

First, the R spikes are detected using an ECG segmentation algorithm such as the ones introduced in Chapter 2. A linear phase Φ_t ranging from $-\pi$ to $+\pi$ is then assigned to each ECG samples on a beat-to-beat basis with the R peaks being fixed at $\Phi_t = 0$, as illustrated in Fig. 5.3. Next, the constant time lag τ in Eq. (5.18) is replaced by a variable τ_t that is computed from Φ_t. Therefore, in each beat, the sample at time t is compared with the sample $t + \tau_t$ which corresponds to the sample with the same phase in the successive beat. More formally, the autocovariance matrix in (5.18) is now computed as

$$\mathbf{C}^x = E_t\{\mathbf{x}_{t+\tau_t}\mathbf{x}'_t\} \tag{5.20}$$

with

$$\tau_t = min\{\tau|\Phi(t+\tau) = \Phi(t), \tau > 0\}. \tag{5.21}$$

The matrix $\mathbf{U}$ is then computed using the GEVD solution of the $(\tilde{\mathbf{C}}^x, \mathbf{C}^x_O)$ pair with the eigenvectors ranked in descending order of their corresponding eigenvalues. The components of $\mathbf{s}_t$ are sorted according to their amount of periodicity with the R peaks of the ECG. The algorithm not longer requires the a priori fixing of a τ parameter but instead requires the detection of the R peaks in the ECG signal.

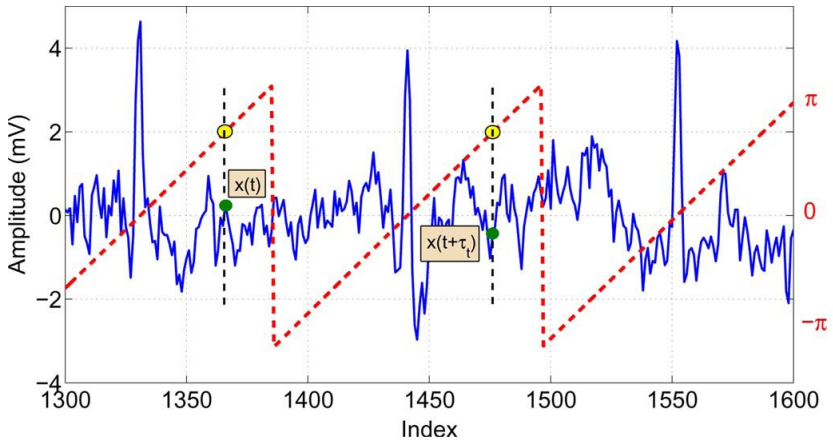

Figure 5.3: Computation of the autocovariance matrix with a varying delay, reproduced from [104]. A linear phase Φ_t ranging from $-\pi$ to $+\pi$ is assigned to each ECG sample on a beat-to-beat basis, with the R peaks being fixed at $\Phi_t = 0$. The constant time lag is then replaced by a variable τ_t that is computed from Φ_t. Finally, The autocovariance matrix is obtained by comparing the sample at time t with the sample $t + \tau_t$.

5.5 Experiments

In this section, we investigate the filtering of ECG artifacts from vagus nerve recordings. Real data obtained from acute recordings of the vagus nerve in rats are used. The experimental protocol for recording neural activity from the vagus nerve in rats has been approved by the Committee for Ethical use of animals of the Faculty of Medicine of the Université catholique de Louvain. Adult albino rats (Wistar) were used for these experiments. Animals were anesthetized with intraperitoneal Xylazine (10 mg/kg body weight) and Ketamine (50 mg/kg body weight).

A 0.9 mm diameter self-sizing spiral cuff made of two silicon rubber sheets carrying 9 pieces of platinum foils of 1 by 1 mm contacts between them is used for recording [75]. Eight internal contacts are made by opening in front of each piece of platinum a 500 μm diameter circular window in the inner silicone sheet. These are used for recording and one external (opening in the external silicone sheet) contact is the reference. The cuff is 17 mm long and the distance between the centers of each contact is 2 mm. The sampling rate is 16384 Hz. The impedance between each contact and the reference is verified prior to recording as suggested in [116] (below 7 kOhms). The ECG signal is recorded from a silver wire electrode inserted subcutaneously.

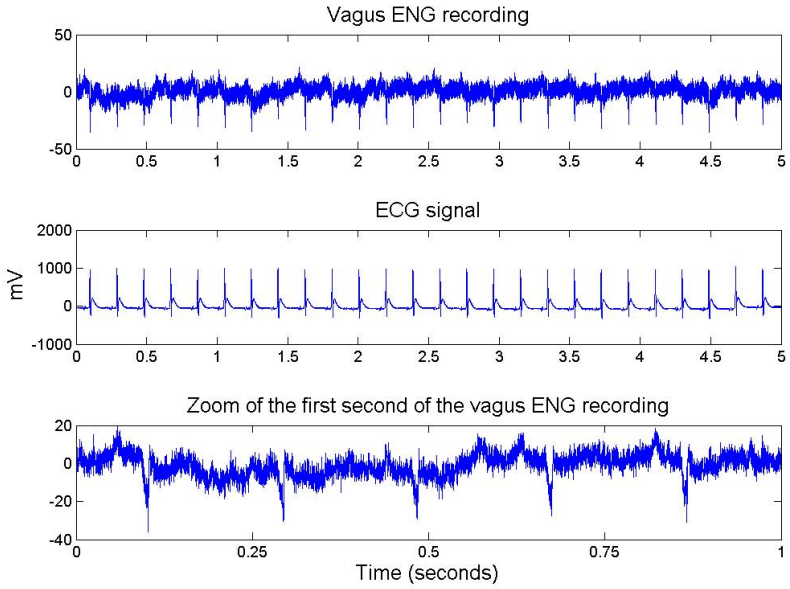

Figure 5.4: ECG contamination in a single rat vagus ENG recording. From top to bottom: the first plot shows 5 seconds of the recorded ENG signal; the second plot is the ECG signal that is recorded apart; the third plot is a zoom on the first sedcond of the recorded ENG signal. Cardiac contamination is obvious.

5.5.1 Single channel filtering

The three single channel ECG removal methods introduced in Sec. 5.3 are evaluated on a 5 seconds neural signal (the signal recorded from the first contact). The raw nerve signal is shown in Fig. 5.4. The lower trace is a zoomed extract corresponding to the first second of the signal in the upper trace. Cardiac contamination is obvious in either case. This signal is the input to each of the three filtering algorithms: FIR filtering, template subtraction and DWT filtering.

Since the frequency band of the ECG signal is mainly localized between 3 and 50 Hz, a safe cutoff frequency of 100 Hz is used for the FIR filtering method. For the template subtraction algorithm, a template size of 2000 points (approximately 1/8th seconds, or 125 milliseconds) is empirically chosen. This value is chosen so as to have a template which is in accordance with the theoretical mean heart rhythm of the rat which is on average approximately 5 beats per second (270-350 beats/min). For the DWT method, an order 6 Daubechies wavelet has been shown to be efficient for ECG

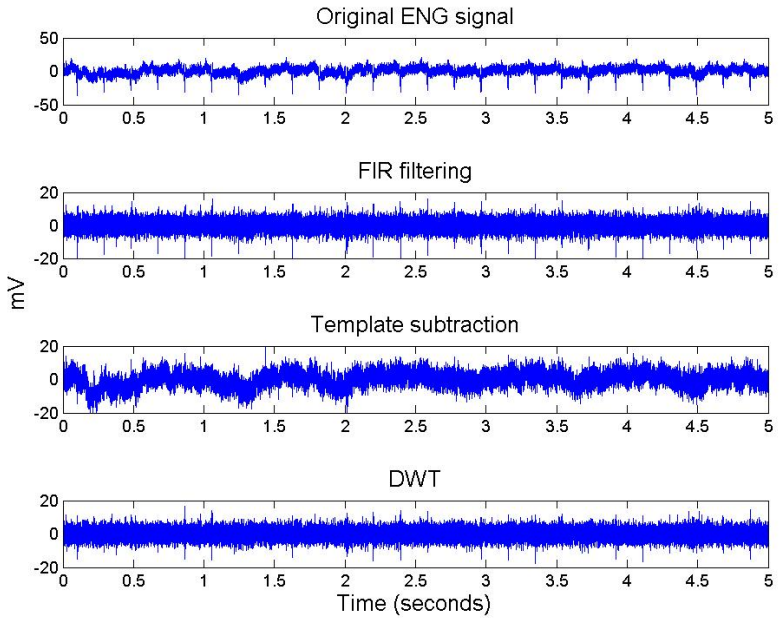

Figure 5.5: ECG filtering in a single rat vagus ENG recording. From top to bottom: the original signal, the signal filtered by the FIR method, the template subtraction method and the DWT method respectively.

filtering in EMG signals [4], and it has therefore been selected for this work as well. The best decomposition level is visually chosen. The detail coefficients at previous levels correspond to nerve activity and are summed to obtain the neural signal. The approximation coefficients correspond to the ECG artifacts.

Figure 5.5 shows the results on the full 5 second signal length and Fig 5.6 zooms on the first second to better judge the filtering. As revealed by these pictures, the FIR filter is not able to remove the QRS complex completely. The template subtraction method performs better but still leaves some low frequency baseline wanderings probably corresponding to respiration artifacts. The DWT outperforms the other two methods, with an almost perfect removal of the ECG contamination and of the baseline wanderings as can be judged visually.

When looking at the approximation coefficients of the DWT, the signal appears to be highly correlated with a time-delayed version of the original ECG signal. Figure 5.7 illustrates the delay between the estimated ECG noise and the real ECG signal that was

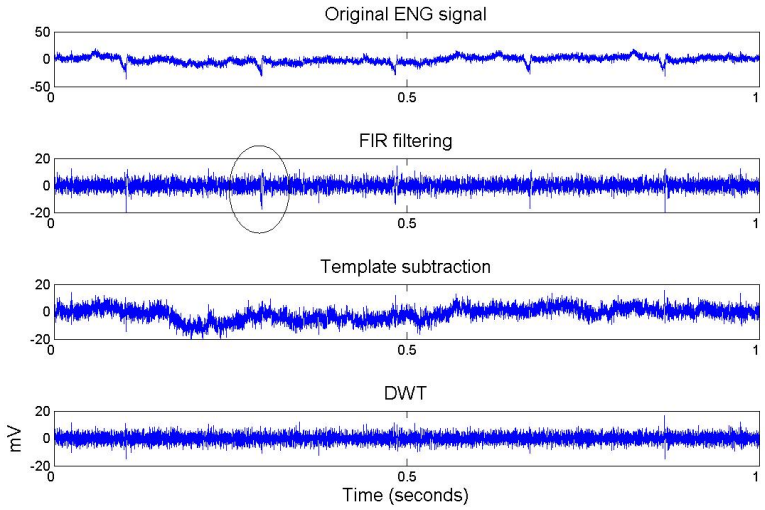

Figure 5.6: Zoom on the first second of Fig. 5.5. The FIR method is unable to completely remove the QRS complex and the template subtraction method leaves artifacts corresponding to the respiration.

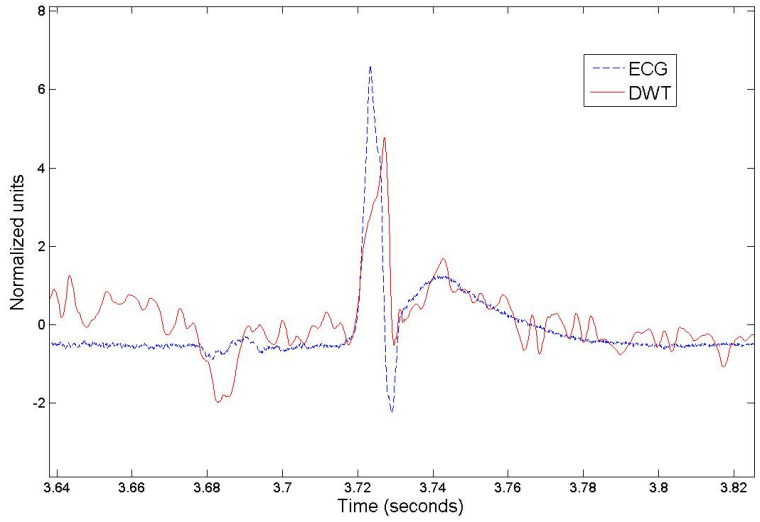

Figure 5.7: Delay between the estimated ECG artifact (straight line) and the real ECG signal (dashed line). This delay is estimated by maximal cross-correlation to be around 2.5 msec.

recorded alongside. The maximal cross-correlation value of this delay is estimated to 2.5 msec. The similarity between this rejected signal and the ECG signal suggests that the artifact mainly results from volume conducted heart electrical activity. The delay can be explained by cardiac conduction times.

It is difficult to obtain a quantitative criterion of the quality of the filtering since it is not possible to obtain a clean benchmark nerve signal. One possible quantitative measure is the correlation between the result of the algorithms and the ECG signal. A correlation value close to zero would indicate that the filtering was efficient. This smallest correlation criterion is however dangerous since a totally random signal would obtain a value of zero. For this reason, this correlation value is only reported in conjunction with visual results as an additional performance measure. Before measuring the correlation, the delay corresponding to the cardiac conduction time is first corrected in the filtered signal.

Table 5.1 shows the correlation values between the original ECG and the signal filtered by the three methods after correction of the delay. The correlation with the raw nerve signal after correction of the delay is also shown for comparison. The three filtering methods significantly reduce the correlation value. Nevertheless, the DWT

Filtering method	Correlation with the ECG
None	-0.5822
FIR	-0.0623
Temp. Sub.	0.0222
DWT	0.0195

Table 5.1: Correlation between the original ECG and the results of the three methods after correction of the delay. The table is sorted in decreasing order of the absolute value of the correlation. The correlation with the raw nerve signal after correction of the delay is also shown for comparison ("None" method).

outperforms the other two methods, which confirms the visual interpretation of Fig. 5.6.

5.5.2 Multiple channel filtering

Four BSS algorithms using the time structure of the signals are applied to the eight recorded signals: the AMUSE algorithm, the robust AMUSE algorithm, the SOBI algorithm and the πCA algorithm. One second of the recorded signals is shown in Fig. 5.8; cardiac contamination is obvious in all channels. The algorithms are applied on the same five seconds as in the single channel experiments.

From the autocorrelation function of the ECG signal (see Fig. 5.9), a delay of $\tau = 3127$ points (approximately equal to 190 milliseconds, a value in accordance with the average heart rhythm in rats) seems to be a good choice for the τ parameter in AMUSE and robust AMUSE. For the SOBI algorithm, which is a blind method, the authors mention that a very limited number of subsequent autocovariance matrices are often sufficient [13]. In our experiments, 10 autocovariance matrices are considered, which is a fairly sufficient number. Figure 5.10 shows the source associated to the largest eigenvalue for each of the four algorithms. As can be judged visually, all methods correctly identify and extract the ECG signal.

The SOBI algorithm (with $v = 2$ to ensure fair comparison) has a running time in the order of 1000 times higher than methods based on the AMUSE algorithm. This is the price to pay for the blindness of the method, and can be problematic for long-term signals required by many applications. To observe the influence of errors in the choice of the delay, an error of 10 milliseconds is added to the τ parameter for AMUSE and robust AMUSE. As illustrated in Fig. 5.11, both methods are no longer able to identify the ECG signal.

The πCA method does not suffer from this problem since the delay is estimated

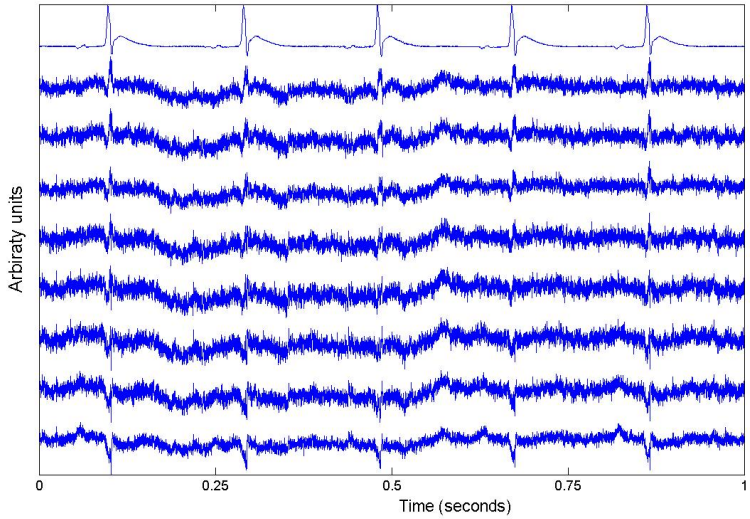

Figure 5.8: The ECG signal (top) and eight channels of vagus ENG recording, showing 1 second. Cardiac contamination is obvious in all channels.

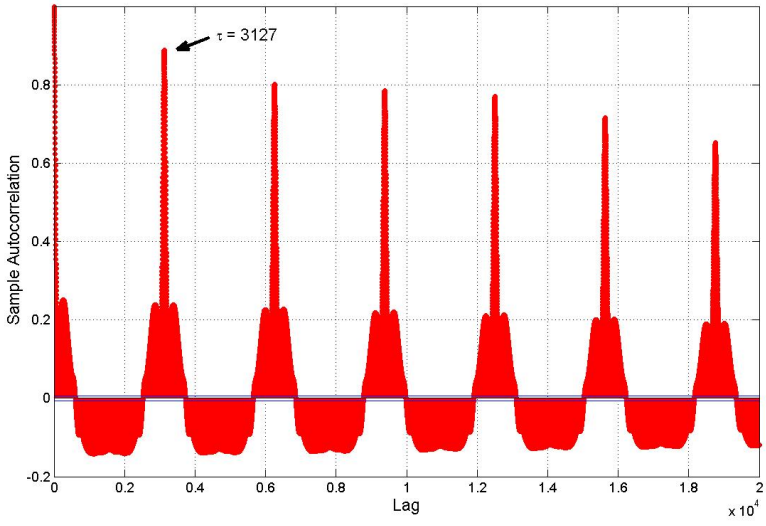

Figure 5.9: Autocorrelation function of the ECG signal. The lag axis is expressed in sampling units (16384 Hz). The most periodic delay τ is shown by the arrow.

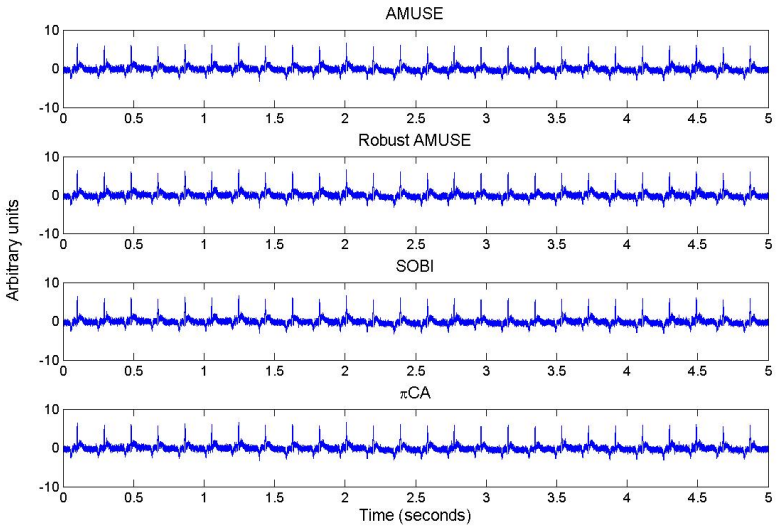

Figure 5.10: Source associated to the largest eigenvalue for each of the four algorithms. From top to bottom: The AMUSE algorithm, the robust AMUSE algorithm, the SOBI algorithm and the πCA algorithm. The identified sources correspond to cardiac artifacts.

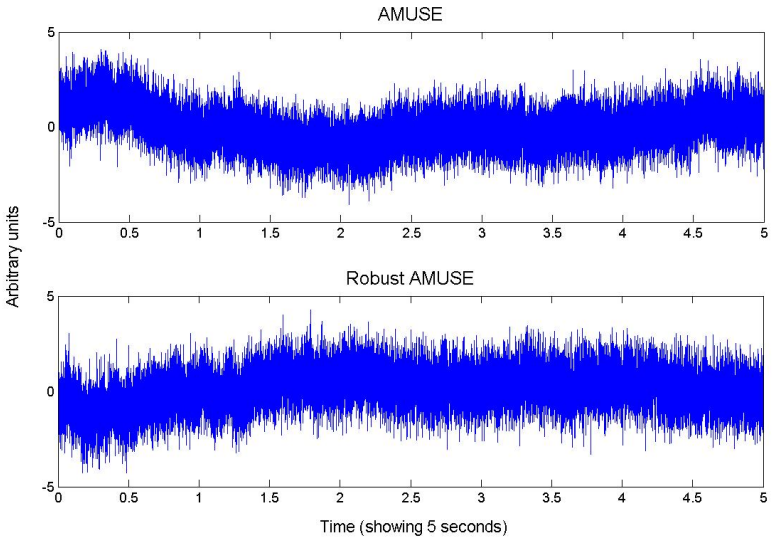

Figure 5.11: Source associated to the largest eigenvalue for the AMUSE and robust AMUSE algorithms, with a 10 millisecond error in the choice of the τ parameter. The estimated sources no longer correspond to cardiac artifacts.

on a beat to beat basis from the position of the R peaks. Furthermore, experiments have shown that the method is very robust to errors in R peak detections [104]. The reconstruction of the first signal after discarding the source associated to the largest eigenvalue is shown in Fig. 5.12. As can be judged visually, cardiac artifacts are successfully removed.

After correction of the cardiac conduction time delay, the correlation between the first reconstructed channel and the original ECG signal is computed to obtain a quantitative assessment of the filtering. Table 5.2 shows the correlation values for the four BSS methods. The four methods significantly reduce the cardiac contamination. Surprisingly, the robust AMUSE algorithm does not yield better results than the classical AMUSE formulation. The worst results are provided by the SOBI method, probably suffering from its blind delay parameter estimation. The three other semi-BSS methods provide better results than the single channel filtering methods, which confirms the interest to include all the available signals in the filtering process. The best overall results are obtained by the πCA algorithm.

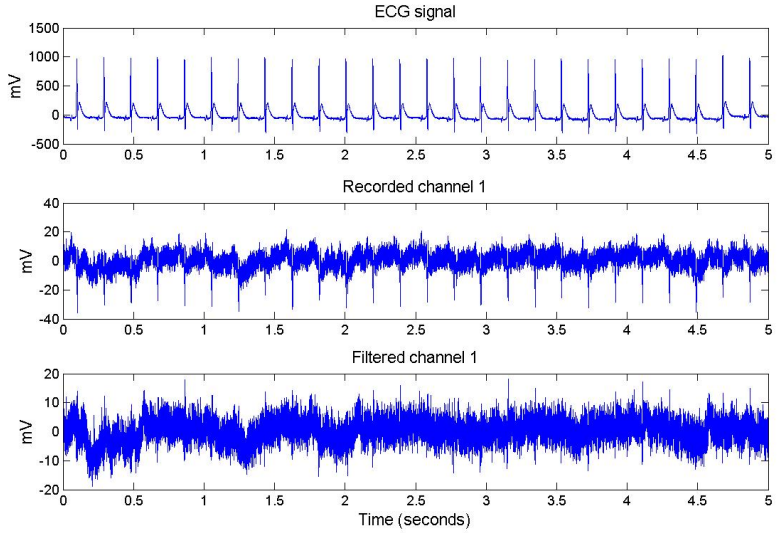

Figure 5.12: Reconstruction of the first neural signal by the πCA method after discarding the source related to ECG artifacts. The first plot shows the recorded ECG signal. The second plot shows the signal recorded by the first contact with cardiac contamination. The last plot shows the same signal after cardiac artifacts were filtered by the πCA method.

Semi-BSS algorithm	**Correlation with the ECG**
None	-0.5822
SOBI	0.0210
Robust AMUSE	-0.0192
AMUSE	-0.0171
πCA	0.0124

Table 5.2: Correlation between the original ECG and the reconstruction of the first channel after correction of the delay for the four BSS algorithms. The table is sorted in decreasing order of the absolute value of the correlation. The correlation with the raw nerve signal after correction of the delay is also shown for comparison ("None" method).

5.6 Discussion

Vagus nerve stimulation has become a recognized form of treatment for several conditions such as refractory epilepsy and depression. Vagus nerve signals have recently been recorded in rats in order to gain a better insight in the nerve involvement in these pathologies. For this purpose, nerve cuff electrodes offer many advantages but suffer from a low signal to noise ratio. In particular, cardiac artifacts represent a major burden for the subsequent extraction of information from the nerve signal. Because of the difficulty to record from a large number of recording contacts in some situations, the filtering in both single channel and multiple channel recordings are investigated separately. In the case of single recordings, cardiac contamination and baseline wanderings can be successfully removed using the DWT method. Other methods such as frequency filtering and template subtraction are unable to completely remove the ECG related noise in the signal. When multiple channels can be recorded, cardiac artifacts can be efficiently separated from the neural activity by blind source separation algorithms. These models are able to integrate all the recordings channels into the filtering process.

Because the recorded neural signal is a multiple action potential, it is expected to be approximately Gaussian distributed and classical BSS models such as ICA are not suitable. For this reason, semi-BSS models relying on assumptions about the temporal structure of the sources are rather considered. Furthermore, from a physiological point of view, the temporal structure criterion is a more suitable assumption for ECG signals which correspond to cyclic occurrences of patterns with different frequency content (the P, QRS and T waves). Moreover, in semi-BSS methods, the extracted components are ranked according to their degree of synchronization with the ECG signal, while in conventional ICA the order of the components cannot be predicted. The Periodic Component Analysis model was recently introduced for the semi-blind extraction of quasi-periodic signals such as the ECG. The method uses a phase-wrapping of the R-R intervals in order to extract the most periodic linear mixtures in a single step of generalized eigenvalue decomposition.

Results using this model show that the cardiac artifacts are automatically separated and that the clean neural signals can then be reconstructed accordingly. The correlation measured between the filtered signals and the recorded ECG, after correction of the cardiac conduction time delay, shows that semi-blind source separation methods outperform single channel filtering methods. These results confirm the interest to include the signals from all the available contacts in the filtering process. Moreover, the periodic component analysis algorithm which yields the best overall results does not require the empirical setting of any parameter, in contrast with single channel methods. The removal of ECG artifacts can therefore be made fully automatic.

The discarded source is highly correlated with a short time-delayed version of the ECG signal, which suggests that the artifact mainly results from volume conducted heart electrical activity. These findings reveal the interesting practical ability to obtain the ECG signal from only neural recordings. The recently developed nerve stimulation devices with recording capabilities are then also able to deliver cardiac and plethysmographic information without the need of additional hardware. Recordings of the vagus nerve in humans have recently been achieved. Further works should investigate whether the results from our experiments can be reproduced in these human vagus nerve signals.

Chapter 6

Conclusion

The ECG signal provides critical information for the diagnosis of cardiac conditions. Some of these diseases are potentially life-threatening if left undetected. Long-term recordings are required to this end, not to miss any transient pattern. Manual analysis of the ECG signal is therefore a very tedious and time-consuming process, and machine learning appears as an appealing solution enabling automatic processing of the ECG signals. The goal of this thesis was the investigation of view of four crucial topics about ECG recordings from a machine learning point. The findings and the results obtained around these four topics are summarized hereafter.

Automatic segmentation of the ECG signal

Standard approaches to annotate the ECG characteristic points rely on unstable heuristic rules and require the setting of many empirical parameters. Probabilistic modeling approaches, and especially hidden Markov models, have more recently been proposed to overcome these limitations. Nevertheless, these models suffer from strong independence assumptions. In this thesis, another class of probabilistic model for labeling sequential data called CRFs has been investigated for the segmentation task. Due to its discriminative form and the undirected nature of its graphical representation, the CRF model does not require the strong naive Bayes assumption as in HMMs and hence allows for complete free choice of the features.

We have shown that the CRF model can achieve feature selection in an embedded manner by regularization of its objective function with the L_1-norm. In sequence models such as CRFs, the L_1-regularization is especially interesting since it enforces sparsity not only in the feature parameters but also in the transition parameters. Not existing transitions between states are therefore much more likely to be avoided during

the inference process.

Results in this thesis have confirmed that the naive Bayes assumption made in HMMs and HSMMs leads to a significant loss in performance for the segmentation of ECG signals. It has nevertheless been observed that HSMMs are less impacted by this assumption since they benefit from the modeling of state durations. In particular, the HMM infers a large number of incorrect waves in the heart beats. This can be very problematic for applications requiring the computation of wave interval statistics like the Q-T interval. The results in this thesis have also shown that the proposed L_1-regularized CRF model outperforms state-of-the-art HMMs and HSMMs on both sinus rhythm and arrhythmic recordings. Moreover, the L_1-regularization significantly increases the performances of the CRF in all situations. These results therefore confirm the benefit from regularizing the CRF objective function with the L_1-norm of its parameter vector.

One of the main difficulties for the evaluation of segmentation performances lies in the errors in the labels provided by the experts. When standard supervised machine learning algorithms are used with such noisy measurements, the resulting model parameters can be adversely influenced. In such case, a given model providing the highest performances may only be overfitting the label errors and hence lead to bad performances in real situations. For this reason, further works should investigate the inclusion of the inherent uncertainty in the boundary measurements defined by the experts into the CRF model. Additional further works should be the evaluation of semi-Markov CRFs to better model the state durations, as in HSMMs.

Heart rate variability metrics for epileptic state identification

Epilepsy is associated with several changes in autonomic function. One of the most important roles of the ANS is the regulation of heart rate. HRV metrics therefore appear as an appealing solution to quantify the interactions between the ANS and epilepsy in a non-invasive manner. In this thesis, the pros and cons of state-of-the-art time domain, frequency domain and non-linear metrics have been reviewed. These HRV metrics have then been applied to seizures from epileptic patients with the objective to identify epileptic states. We have motivated the separation of the HRV analysis between left and right lateralized seizures. Results show that left lateralized seizures do not reveal any change in HRV metrics in the four seizure states. To the opposite, in right lateralized seizures we have observed that time domain metrics and non-linear metrics detect significant changes in HRV during the ictal state. These findings confirm the hypothesis that the lateralization of the seizures is a crucial indicator of

whether heart rhythm variations are likely to be observed during the seizures.

The results also confirm the suggestion that the ECG signal could be used in seizure detection algorithms along with EEG signals. Nevertheless, the change in heart variability does not seem to happen before the actual start of the seizure. These findings support the doubts raised in recent studies about the existence of a preictal state that could be used in order to predict the onset of seizures. Results also validate the pertinence of the more recent non-linear metrics based on the entropy of the R-R series.

On the other hand, frequency metrics did not reveal any significant difference where time domain metrics succeeded in doing so. These findings are in contradiction with previously reported results where multiple patterns of changes in the frequency bands were observed at the onset of seizures. This difference may be explained by the fact that the frequency transform used in previous studies relies on a resampling step, and this has been shown to induce significant errors in frequency domain metrics.

The results in this thesis were obtained from retrospectively collected data. Several important factors, such as medications taken by the patients, have therefore been left uncontrolled. The conclusions derived from our experiments should therefore be validated on a larger sample size specifically recruited for this particular HRV study.

Automatic classification of heart beats

In this thesis, guidelines for the design of reliable automatic heart beat classifiers and for the evaluation of their relative merits have been presented. In particular, we have shown that the AAMI standards and what we have defined as the inter-patient classification paradigm are of major importance for this purpose.

The main difficulty in the design of a heart beat classifier is the class unbalance. From the Occam's razor principle, standard classifiers perform poorly in such situations. In this thesis, several solutions to overcome the class unbalance in classifiers have been reviewed. In particular, we have focused on the cost-sensitive learning approach. The weighted SVM and the weighted L_1-regularized CRF models have been proposed as specific instances of the cost-sensitive learning approach.

In the literature, the features used in the model are chosen using domain knowledge or by evaluating non-exhaustive combinations of feature groups. The discriminative power of the individual features is thus left unknown and the results can be suboptimal. Moreover, each study involves distinct feature groups. In this thesis, a large number of feature groups previously proposed in the literature have been considered together with two new patient-normalized feature groups. Computationally affordable feature selection techniques have then been used to extract the discriminative features within the set of all computed features.

Results obtained on real ECG data have confirmed that the weighted models yield significantly better results in unbalanced situations than their non-weighted version. Moreover, we have shown that the balanced classification rate is a performance metric which is much more suitable than the usual accuracy in the case of unbalanced classes. Results have also shown that the L_1-regularized weighted CRF model achieves better results than the state-of-the-art inter-patient heart beat classification model. These findings confirm the hypothesis that the temporal information used in CRFs increases the classification performances. The feature selection results reveal that several previously reported features do not serve the classification process, and that a very small number of features are actually required to yield high performances. Moreover, the feature selection results have confirmed the pertinence of the proposed patient-normalized features.

Further works should include the use of more powerful feature selection techniques than the ranking selection strategy used in our experiments. Nevertheless, such feature selection techniques will only be affordable after a deep reduction of the redundancy in the dataset has been achieved. Clustering techniques appear as appealing solutions to this end.

Filtering of ECG contamination in vagus nerve recordings

Vagus nerve recordings are contaminated by cardiac artifacts. These artifacts can significantly prevent the extraction of accurate information from the neural signals. In this thesis, several filtering techniques have been reviewed and applied to real vagus nerve signals recorded in rats with a cuff electrode. Because of the difficulty to record from a large number of recording contacts in some situations, the filtering in both single channel and multiple channel recordings are explored separately.

In the case of single recordings, the FIR filtering, template subtraction and DWT method have been considered and applied to a five seconds neural signal with obvious cardiac contamination. Results in this thesis have shown that cardiac contamination and baseline wanderings can be successfully removed using the DWT method while other methods are unable to completely remove the ECG related noise in the signal. When multiple channels can be recorded, semi-BSS models relying on assumptions about the temporal structure of the sources are considered. Results have shown that the periodic component analysis model successfully extracts the source corresponding to the cardiac artifacts. The clean neural signals can then be reconstructed after discarding this source.

The interest to include multiple contacts in the cuff electrode, when possible, has

also been confirmed by our results. The filtering is indeed more efficient and can be made fully automatic when several signals can be obtained. These findings reveal the interesting practical ability to automatically obtain the ECG signal from only neural recordings. The recently developed nerve stimulation devices with recording capabilities are then also able to deliver cardiac and plethysmographic information without the need of additional hardware.

Recordings of the vagus nerve in humans have recently been achieved with the cuff electrode. Further works should investigate whether the results from our experiments can be reproduced in these human vagus nerve signals.

Take-home message: this thesis in one paragraph

Long-term recordings of physiological signals are often required to evaluate the state of the organs in the body. Machine learning tools are of particular interest to automatically extract and process useful information from these recordings. In this thesis, we have focused on the automatic analysis of the ECG signal, which records the activity of the heart. We have shown that traditional machine learning algorithms may fail to adequately model the ECG. In particular, the cyclic temporal structure of the ECG can better be represented by specific models that are able to grasp the temporal dependence in the signal and the structure of its cyclic patterns. In this thesis, we have shown that semi-blind source separation models and conditional random fields are successful candidates to this end. We have also shown that the ECG signal has an underestimated potential in the field of epilepsy where EEG signals are often the only reference.

From a Biomedical Point of View

The work presented here pertains to the fields of engineering as well as of biomedicine. Both disciplines bring in a different point of view. This section will summarize the results in a way that addresses more appropriately the concerns of the biomedical and the clinical community.

In Chapter 2, a new algorithm has been proposed to automatically detect the boundaries of P, QRS and T components of the ECG wave. This algorithm can be seen as a black box. The input is a sampled ECG signal and the output is the label of each component at any time-step of the recorded ECG signal. The content of the black box is optimized in order to yield the best results, that is to find the same labels as were given previously on the same signals by experts or any other 'golden standard' if available. The originality of the algorithm used here is the ability to use characteristics from the full recording length to predict the wave at a given time point. Furthermore, the algorithm is also able to automatically decide which of these characteristics are relevant. During experiments, the performances of the algorithm have been evaluated in real situations, on recordings from both normal and pathological patients, and were found superior to previously reported algorithms.

In Chapter 3, the question was explored as to whether the ECG signal could provide information about epileptic seizures. It is indeed known that epilepsy can induce sympathetic symptoms and vagus nerve stimulation is a recognized procedure for the treatment of refractory epilepsy. Changes in heart beat frequency could thus be expected on the basis of the parasympathetic influence of the vagus nerve on the sino-atrial node. Experiments in this thesis have indeed shown that epileptic seizures are associated with an abnormal variability in the heart rhythm. Recording and analyzing the ECG signal along with EEG thus appears to be potentially informative in epileptic patients. In the past twenty years, much work has been devoted to the development of EEG-based epileptic seizure prediction algorithms. Despite promising early results,

recent research raised serious doubts about the reliability and general applicability of such prediction algorithms. We have explored this avenue with ECG signals and have not been able to predict seizures on this basis. Early detection is a realistic objective but prediction might simply be impossible.

In Chapter 4, the design of algorithms to automatically diagnose heart beats as being either normal or pathological has been investigated. Here again, as in Chapter 2, these algorithms can be seen as black boxes. This time, the input is a heart beat and the output is the class label normal or one of the known pathological patterns. We point to the fact that the design of such models is not trivial because the number of beats is highly different in the distinct classes. We have also identified the characteristics of a given heart beat that are relevant to predict its label (For example, R-R the time interval is different in normal and in pathological situations). In experiments with real ECG recordings, the best performances are achieved using our own algorithm which has two advantages over previously reported beat classification programs. First, it is able to use the label of the previous beat to improve the decision about the current beat. Second, it is able to automatically select the characteristics of the beat that are relevant to come to that decision.

In Chapter 5, the removal of ECG related artifacts has been investigated in vagus nerve recordings. These artifacts can significantly distort the envelope of the neurogram and must therefore be filtered out as much as possible. The experimental vagus nerve recordings used here were obtained in rats using a cuff electrode with up to 8 embedded contacts. Ad-hoc blind source separation algorithms based on the temporal structure of the ECG recording rather than the more traditional non-gaussianity assumption was investigated. Results have shown that the automatic filtering achieved is more efficient when several signals can be recorded from the cuff electrode.

In conclusion, the results obtained from the experiments conducted in the frame of this thesis point to the importance of integrating the biological properties of the system at stake in the analysis algorithms. Although the designed algorithms can at first sight seem complex to medical practitioners, it often appears that the mathematical equations and their parameters actually only reflect the biological aspects of the underlying problem. For example, a CRF model can be summarized by its two sets of parameters: the first set of parameters captures the features of heart beats (i.e. their shape) and the second set of parameters captures their temporal sequence (the P, QRS and T waves). A deep understanding and efficient communication between the biomedical and the engineering worlds is therefore of major importance. Both sides directly benefit from the questions and knowledge of the other side. To cite the famous Disney character Hannah Montana: "*Mix it all together, it's so much better, because you know you've got the best of both worlds!*".

Bibliography

[1] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation and clinical use. *Circulation*, 93(5):1043–1065, Mar 1996.

[2] U. R. Acharya, P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri. Heart rate variability: a review. *Medical and Biological Engineering and Computing*, 44(12):1031–1051, 2006.

[3] P. S. Addison. Wavelet transforms and the ecg: a review. *Physiological Measurement*, 26(5):155–199, 2005.

[4] W. F. Agnew. *Neural Prostheses: Fundamental Studies (Spectrum Book)*. Prentice Hall, 1990.

[5] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *ECML 2004: Proceedings of the 15th European conference on Machine Learning*, pages 39–50. Springer Berlin / Heidelberg, 2004.

[6] S. Akselrod, D. Gordon, F. A. Ubel, D. C. Shannon, A. C. Berger, and R. J. Cohen. Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. *Science*, 213(4504):220–222, Jul 1981.

[7] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden markov support vector machines. In *ICML 2003: Proceedings of the 20th international conference on Machine learning*, pages 3–10. AAAI Press, January 2003.

[8] S. Ansari, K. Chaudhri, and K. Moutaery. Vagus nerve stimulation: indications and limitations. In *Operative Neuromodulation*, volume 97 of *Acta Neurochirurgica Supplementum*, pages 281–286. Springer Vienna, 2007.

[9] Association for the Advancement of Medical Instrumentation. Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms. *ANSI/AAMI EC38:1998*, 1998.

[10] B. Azzerboni, M. Carpentieri, F. La Foresta, and F. C. Morabito. Neural-ica and wavelet transform for artifacts removal in surface emg. In *IJCNN 2009: International Joint Conference on Neural Networks*, pages 3223–3228, 2004.

[11] A. K. Barros and A. Cichocki. Extraction of specific signals with temporal structure. *Neural Computation*, 13(9):1995–2003, 2001.

[12] C. Baumgartner, S. Lurger, and F. Leutmezer. Autonomic symptoms during epileptic seizures. *Epileptic Disorders*, 3(3):103–116, Sep 2001.

[13] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[14] L. D. Blumhardt, P. E. Smith, and L. Owen. Electrocardiographic accompaniments of temporal lobe epileptic seizures. *Lancet*, 1(8489):1051–1056, May 1986.

[15] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[16] J. W. Britton, G. R. Ghearing, E. E. Benarroch, and G. D. Cascino. The ictal bradycardia syndrome: localization and lateralization. *Epilepsia*, 47(4):737–744, Apr 2006.

[17] D. R. Brown, L. V. Brown, A. Patwardhan, and D. C. Randall. Sympathetic activity and blood pressure are tightly coupled at 0.4 Hz in conscious rats. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, 267(5):1378–1384, 1994.

[18] P. De Chazal, M. O'Dwyer, and R. B. Reilly. Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, 51:1196–1206, 2004.

[19] I. Christov, G. Gómez-Herrero, V. Krasteva, I. Jekova, A. Gotchev, and K. Egiazarian. Comparative study of morphological and time-frequency ecg descriptors for heartbeat classification. *Medical Engineering and Physics*, 28(9):876–887, 2006.

[20] G. D. Clifford, F. Azuaje, and P. McSharry. *Advanced Methods And Tools for ECG Data Analysis*. Artech House, Inc., Norwood, MA, USA, 2006.

[21] G. D. Clifford and L. Tarassenko. Quantifying errors in spectral estimates of hrv due to beat replacement and resampling. *IEEE Transactions on Biomedical Engineering*, 52(4):630–638, Apr 2005.

[22] T. F. Coleman and Y. Li. On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds. *Mathematical Programming*, 87(2):189–224, 1994.

[23] T. F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2):418–445, 1996.

[24] G. de Lannoy, D. François, J. Delbeke, and M. Verleysen. Feature relevance assessment in automatic inter-patient heart beat classification. In *BIOSIGNALS 2010: Proceedings of the 3rd International Conference on Bio-inspired Systems and Signal Processing*, pages 13–20, Valencia, Spain, January 20-23 2010.

[25] G. de Lannoy, D. François, J. Delbeke, and M. Verleysen. Supervised segmentation of ecg signals using l_1-regularized conditional random fields. In *Submitted to IWANN2011: Proceedings of the 11th International Work-Conference on Artificial Neural Networks (Advances on Computational Intelligence)*, 2011.

[26] G. de Lannoy, D. François, J. Delbeke, and M. Verleysen. Weighted svms and feature relevance assessment in supervised heart beat classification. *Communications in Computer and Information Science*, 127:212–225, 2011.

[27] G. de Lannoy, D. François, and M. Verleysen. Class-specific feature selection for one-against-all multiclass svms. In *ESANN 2011: Proceedings of the 19th European Symposium on Artificial Neural Networks (to be published)*, Bruges, Belgium, April 27-29 2011.

[28] G. de Lannoy, T. Costecalde, J. Marin, M. Verleysen, and J. Delbeke. Elimination of electrocardiogram contamination from vagus nerve recordings using ica. In *IFESS 2009: Proceedings of the 14th Annual International FES Society Conference*, pages 109–111, Seoul; Korea, September 13-17 2009.

[29] G. de Lannoy, M. de Tourchaninoff, M. Verleysen, and J. Delbeke. Evaluation of heart rate variability metrics between the different states of epileptic seizures. *Submitted to Epilepsia*, 2010.

[30] G. de Lannoy, A. De Decker, and M. Verleysen. A supervised learning approach based on the cwt for r spike detection in ecg. In *BIOSIGNALS 2008: Proceedings of the 1st International Conference on Bio-inspired Systems and Signal Processing*, pages 140–145, Madeira, Portgual, January 2008.

[31] G. de Lannoy, A. De Decker, and M. Verleysen. A supervised wavelet transform algorithm for r spike detection in noisy ecgs. *Communications in Computer and Information Science*, 25:256–264, 2008.

[32] G. de Lannoy, B. Frénay, M. Verleysen, and J. Delbeke. Supervised ecg delineation using the wavelet transform and hidden markov models. In *MBEC 2008: Proceedings of the fourth European Conference of the International Federation for Medical and Biomedical Engineering*, pages 22–25, Anvers, Belgium, November 23-27 2008. Springer.

[33] G. de Lannoy, J. Marin, M. Verleysen, and J. Delbeke. Filtering heart related activity from vagus nerve recordings in rats. In *Proceedings of IFESS'08, 13th Annual International FES Society Conference*, pages 16–18, Freiburg, Germany, September 21-25 2008.

[34] G. de Lannoy, M. Verleysen, and J. Delbeke. Assessment and comparison of time realignment methods for supervised heart beat classification. In *BIOSIGNALS 2009: Proceedings of the 2nd International Conference on Bio-inspired Systems and Signal Processing*, pages 140–145, Porto, Portugal, January 14-17 2009.

[35] O. Devinsky, S. Pacia, and G. Tatambhotla. Bradycardia and asystole induced by partial seizures: a case report and literature review. *Neurology*, 48(6):1712–1714, Jun 1997.

[36] Digital Signal Processing Committe. *Programs for Digital Signal Processing*. IEEE Press, Piscataway, NJ, USA, 1979.

[37] G. Doquire, G. de Lannoy, D. François, and M. Verleysen. Feature selection for supervised inter-patient heart beat classification. In *BIOSIGNALS 2011: Proceedings of the 4th International Conference on Bio-inspired Systems and Signal Processing*, pages 67–73, Roma, Italy, January 26-29 2011.

[38] J. D. M. Drake and J. P. Callaghan. Elimination of electrocardiogram contamination from electromyogram signals: An evaluation of currently used removal techniques. *Journal of Electromyography and Kinesiology*, 16(2):175 – 187, 2006.

[39] J. S. Ebersole. In search of seizure prediction: a critique. *Clinical Neurophysiology*, 116(3):489–492, Mar 2005.

[40] H. Evrengül, H. Tanriverdi, D. Dursunoglu, A. Kaftan, O. Kuru, U. Unlu, and M. Kilic. Time and frequency domain analyses of heart rate variability in patients with epilepsy. *Epilepsy Research*, 63(2-3):131–139, Feb 2005.

[41] B. Frénay, G. de Lannoy, and M. Verleysen. Emission modelling for supervised ecg segmentation using finite differences. In *MBEC 2008: Proceedings of the fourth European Conference of the International Federation for Medical and Biomedical Engineering*, pages 1212–1216, Anvers, Belgium, November 23-27 2008. Springer.

[42] B. Frénay, G. de Lannoy, and M. Verleysen. Improving the transition modelling in hidden markov models for ecg segmentation. In *ESANN 2009: Proceedings of the 17th European Symposium on Artificial Neural Networks*, pages 141–146, Bruges, Belgium, April 22-24 2009.

[43] M. S. George, Z. Nahas, D. E. Bohning, M. Lomarev, S. Denslow, R. Osenbach, and J. C. Ballenger. Vagus nerve stimulation: a new form of therapeutic brain stimulation. *CNS Spectrums*, 5(11):43–52, Nov 2000.

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

[45] V. Gomez-Verdejo, M. Verleysen, and J. Fleury. Information-theoretic feature selection for functional data classification. *Neurocomputing*, 72(16-18):3580–3589, Oct. 2009.

[46] B. R. Greene, P. de Chazal, G. B. Boylan, S. Connolly, and R. B. Reilly. Electrocardiogram based neonatal seizure detection. *IEEE Transactions on Biomedical Engineering*, 54(4):673–682, Apr 2007.

[47] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[48] N. P. Hughes, S. J. Roberts, and L. Tarassenko. Semi-supervised learning of probabilistic models for ecg segmentation. In *IEMBS 2004: Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 434–437, 2004.

[49] N. P. Hughes, L. Tarassenko, and S. J. Roberts. Markov models for automated ECG interval analysis. In *NIPS 2004: Proceedings of the 16th Conference on Advances in Neural Information Processing Systems*, pages 611–618, 2004.

[50] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, first edition, May 2001.

[51] N. Japkowicz, C. Myers, and M. Gluck. A novelty detection approach to classification. In *Proceedings of the Fourteenth Joint Conference on Artificial Intelligence*, pages 518–523, 1995.

[52] G. Karakoulas and J. Shawe-Taylor. Optimizing classifiers for imbalanced training sets. In *NIPS 1998: Proceedings of the 11th Conference on Advances in neural information processing systems*, pages 253–259, Denver, Colorado, USA, November 30 1999. The MIT Press.

[53] D. H. Kerem and A. B. Geva. Forecasting epilepsy from the heart rate signal. *Medical and Biological Engineering and Computing*, 43(2):230–239, Mar 2005.

[54] A. Koski. Modelling ecg signals with hidden markov models. *Artificial Intelligence in Medicine*, 8(5):453 – 471, 1996.

[55] M. Kubat, R. Holte, and S. Matwin. Learning when negative examples abound. In *ECML 1997: Proceedings of the 9th European Conference on Machine Learning*, pages 146–153, London, UK, 1997. Springer-Verlag.

[56] M. Laakso, A. Aberg, J. Savola, P. J. Pentikainen, and Pyorala K. Diseases and drugs causing prolongation of the qt interval. *American journal of Cardiology*, 59(8):862–5, 1987.

[57] N. Lachiche and P. A. Flach. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using roc curves. In *ICML 2003: Proceedings of the 20th International Conference on Machine Learning*, pages 416–423. AAAI Press, January 2003.

[58] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML 2001: Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[59] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo. Clustering ecg complexes using hermite functions and self-organizing maps. *IEEE Transactions on Biomedical Engineering*, 47(7):838–848, Jul 2000.

[60] P. Laguna. *New Electrocardiographic Signal Processing Techniques: Application to Long-term Records*. PhD thesis, Science Faculty, University of Zaragoza, 1990.

[61] P. Laguna, G. B. Moody, and R. G. Mark. Power spectral density of unevenly sampled data by least-square analysis: performance and application to heart rate signals. *IEEE Transactions on Biomedical Engineering*, 45(6):698–715, Jun 1998.

[62] M. R. Lauer. Heart function: Basic anatomy and function of the heart. http://www.permanente.net/homepage/kaiser/pages/c6166-63363.html.

[63] S. I. Lee, H. Lee, P. Abbeel, and A. Y. Ng. Efficient L1 Regularized Logistic Regression. In *AAAI 2006 : Proceedings of the Twenty-First National Conference on Artificial Intelligence*, page 401. 2006.

[64] J. Leskovec and J. Shawe-Taylor. Linear programming boosting for uneven datasets. In *ICML 2003: Proceedings of the 20th International Conference on Machine Learning*, pages 456–463. AAAI Press, 2003.

[65] S. E. Levinson. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech and Language*, 1(1):29–45, 1986.

[66] C. Li, C. Zheng, and C. Tai. Detection of ecg characteristic points using wavelet transform. *IEEE Transactions on Biomedical Engineering*, 42(1):21–28, 1995.

[67] X.-L. Li and X.-D. Zhang. Robust extraction of specific signals with temporal structure. *Neurocomputing*, 69(7-9):888–893, 2006.

[68] X.-L. Li and X.-D. Zhang. Sequential blind extraction adopting second-order statistics. *IEEE Signal Processing Letters*, 14(1):58–61, 2007.

[69] A. Liu, J. Ghosh, and C. Martin. Generative oversampling for mining imbalanced datasets. In *DMIN 2007: Proceedings of the 2007 International Conference on Data Mining*, pages 66–72, Las Vegas, Nevada, USA, June 2007.

[70] M. Llamedo-Soria and J. P. Martinez. An ecg classification model based on multilead wavelet transform features. In *Computers in Cardiology*, volume 35, pages 105–108, Sept. 2007.

[71] N. R. Lomb. Least-square frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39:447–462, 1976.

[72] M. Malik. Errors and misconceptions in ecg measurement used for the detection of drug induced qt interval prolongation. *Journal of Electrocardiology*, 37:25–33, 2004.

[73] S. Mallat. *A Wavelet Tour of Signal Processing, Second Edition (Wavelet Analysis & Its Applications)*. Academic Press, 2 edition, September 1999.

[74] A. Malliani, M. Pagani, F. Lombardi, and S. Cerutti. Cardiovascular neural regulation explored in the frequency domain. *Circulation*, 84(2):482–492, Aug 1991.

[75] J. Marin, T. Costecalde, G. de Lannoy, and J. Delbeke. Recording and stimulation contacts combined in a single spiral cuff electrode for use in rats. *Acta Physiologica*, 192 (Supp.661), 2008.

[76] J. Marin, G. de Lannoy, and J. Delbeke. When can we recover charges with a biphasic charge balanced stimulation pulse ? In *IFESS09: Proceedings of the 14th Annual International FES Society Conference*, pages 55–57, Seoul; Korea, September 13-17 2009.

[77] J. P. Martinez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna. A wavelet-based ECG delineator: evaluation on standard databases. *IEEE Transactions on Biomedical Engineering*, 51:570–81, 2004.

[78] A. McCallum, D. Freitag, and F. C. N. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML 2000: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 591–598, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[79] F. Melgani and Y. Bazi. Classification of electrocardiogram signals with support vector machines and particle swarm optimization. *IEEE Transactions on Information Technology in Biomedicine*, 12(5):667–677, Sept. 2008.

[80] P. Mico, D. Cuesta, and D. Novak. Preclustering of electrocardiographic signals using left-to-right hidden markov models. *Lecture Notes in Computer Science*, 3138:939–947, 2004.

[81] J. P. Moak, D. S. Goldstein, B. A. Eldadah, A. Saleem, C. Holmes, S. Pechnik, and Y. Sharabi. Supine low-frequency power of heart rate variability reflects baroreflex function, not cardiac sympathetic innervation. *Cleveland Clinic Journal of Medicine*, 76(Suppl 2):51–59, 2009.

[82] R. Moddemeijer. On estimation of entropy and mutual information of continuous distributions. *Signal Processing*, 16(3):233–246, 1989.

[83] N. Montano, T. G. Ruscone, A. Porta, F. Lombardi, M. Pagani, and A. Malliani. Power spectrum analysis of heart rate variability to assess the changes in sympathovagal balance during graded orthostatic tilt. *Circulation*, 90(4):1826–1831, Oct 1994.

[84] J. Morganroth and H. M. Pyper. The use of electrocardiograms in clinical drug development: Part 1. *Clinical Research Focus*, 12(5):17–23, 2001.

[85] F. Mormann, R. G. Andrzejak, C. E. Elger, and K. Lehnertz. Seizure prediction: the long and winding road. *Brain*, 130(2):314–333, 2007.

[86] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NIPS 2002: Proceedings of the 14th Conference on Advances in Neural Information Processing Systems*, pages 841–848. MIT Press, 2002.

[87] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *ICML 2004: Proceedings of the twenty-first international conference on Machine learning*, page 78, New York, NY, USA, 2004. ACM.

[88] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung. *Learning Pattern Classification Tasks with Imbalanced Data Sets*, volume Pattern Recognition. INTECH, 2009.

[89] S. M. Oppenheimer, A. Gelb, J. P. Girvin, and V. C. Hachinski. Cardiovascular effects of human insular cortex stimulation. *Neurology*, 42(9):1727–1732, Sep 1992.

[90] S. Osowski and L.T. Hoai. Ecg beat recognition using fuzzy hybrid neural network. *IEEE Transactions on Biomedical Engineering*, 48(11):1265–1271, Nov 2001.

[91] S. Osowski, L.T. Hoai, and T. Markiewicz. Support vector machine-based expert system for reliable heartbeat recognition. *IEEE Transactions on Biomedical Engineering*, 51(4):582–589, April 2004.

[92] J. Pan and W. J. Tompkins. A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3):230–236, March 1985.

[93] K. S. Park, B. H. Cho, D. H. Lee, S. H. Song, J. S. Lee, Y. J. Chee, I. Y. Kim, and S. I. Kim. Hierarchical support vector machine based heartbeat classification using higher order statistics and hermite basis function. In *Computers in Cardiology*, pages 229–232, Sept. 2008.

[94] C. K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos*, 5(1):82–87, 1995.

[95] S. M. Pincus. Approximate entropy (apen) as a complexity measure. *Chaos*, 5(1):110–117, 1995.

[96] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods*. The MIT Press, 1999.

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

[98] M. S. Redfern, R. E. Hughes, and D. B. Chaffin. High-pass filtering to remove electrocardio-graphic interference from torso emg recordings. *Clinical Biomechanics*, 8(1):44–48, 1993.

[99] A. L. Reeves, K. E. Nollet, D. W. Klass, F. W. Sharbrough, and E. L. So. The ictal bradycardia syndrome. *Epilepsia*, 37(10):983–987, Oct 1996.

[100] J. S. Richman and J. R. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology - Heart and Circulatory Physiology*, 278(6):2039–2049, June 2000.

[101] R. Rocamora, M. Kurthen, L. Lickfett, J. Von Oertzen, and C. E. Elger. Cardiac asystole in epilepsy: clinical and neurophysiologic features. *Epilepsia*, 44(2):179–185, Feb 2003.

[102] J. Rodriguez-Sotelo, D. Cuesta-Frau, and G. Castellanos-Dominguez. An improved method for unsupervised analysis of ecg beats based on wt features and j-means clustering. In *Computers in Cardiology*, pages 581–584, 30 Sept. 2007 2007.

[103] A. O. Rossetti, B. A. Dworetzky, J. R. Madsen, O. Golub, J. A. Beckman, and E. B. Bromfield. Ictal asystole with convulsive syncope mimicking secondary generalisation: a depth electrode study. *Journal of Neurology, Neurosurgery and Psychiatry*, 76(6):885–887, Jun 2005.

[104] R. Sameni, C. Jutten, and M. B. Shamsollahi. Multichannel electrocardiogram decomposition using periodic component analysis. *IEEE Transactions on Biomedical Engineering*, 55(8):1935–1940, 2008.

[105] S. Sarawagi and W. W. Cohen. Semi-markov conditional random fields for information extraction. In *NIPS 2004: Proceedings of the 17th conference on Advances in Neural Information Processing Systems*, pages 1185–1192, 2004.

[106] J. D. Scargle. Studies in astronomical time series analysis. ii. statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal*, 263:835–853, 1982.

[107] M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *ECML 2007: Proceedings of the 18th European conference on Machine Learning*, pages 286–297, Berlin, Heidelberg, 2007. Springer-Verlag.

[108] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[109] B. Schölkopf and A. Smola. *Learning with kernels*. MIT press, Cambridge, 2002.

[110] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423,623–656, 1948.

[111] A. Smith and M. Osborne. Regularisation techniques for conditional random fields: Parameterised versus parameter-free. In *IJCNLP 2005: Proceedings of the International Joint Conference on Natural Language Processing*, pages 896–907. Springer Berlin Heidelberg, 2005.

[112] S. Sunderland. *Nerves and nerve injuries*. Churchill Livingstone, 2nd edition, 1978.

[113] C. Sutton and A. McCallum. *An Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2007.

[114] F. Takens. Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence*, 898:366–381, 1981.

[115] M.-A. Thil. *Nerve reactions to self-sizing cuff electrode implantation and electrical stimulation: physiological and histological studies*. PhD thesis, Univeristé catholique de Louvain, 2006.

[116] M.-A. Thil, B. Gérard, J. C. Jarvis, and J. Delbeke. Two-way communication for programming and measurement in a miniature implantable stimulator. *Medical and Biological Engineering and Computing*, 43(4):528–534, 2005.

[117] T. Tomson, M. Ericson, C. Ihrman, and L. E. Lindblad. Heart rate variability in patients with epilepsy. *Epilepsy Research*, 30(1):77–83, Mar 1998.

[118] L. Tong, V. C. Soon, Y. F. Huang, and R. Liu. Amuse: a new blind identification algorithm. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 1784–1787, 1990.

[119] V. N. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, November 1999.

[120] B. V. Vaughn, S. R. Quint, M. B. Tennison, and J. A. Messenheimer. Monitoring heart period variability changes during seizures. ii. diversity and trends. *Journal of Epilepsy*, 9(1):27 – 34, 1996.

[121] H. Wallach. Efficient training of conditional random fields. Master's thesis, University of Edinburgh, 2002.

[122] D. Wang and L. Shi. Selecting valuable training samples for svms via data structure analysis. *Neurocomputing*, 71(13-15):2772–2781, 2008.

[123] G. M. Weiss and F. Provost. Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19(1):315–354, 2003.

[124] S.-J. Yen and Y.-S. Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, 2009.

[125] Z. Zhang and J. Zhou. Transfer estimation of evolving class priors in data stream classification. *Pattern Recognition*, 43(9):3151 – 3161, 2010.

[126] M. Zijlmans, D. Flanagan, and J. Gotman. Heart rate changes and ecg abnormalities during epileptic seizures: prevalence and definition of an objective clinical sign. *Epilepsia*, 43(8):847–854, Aug 2002.