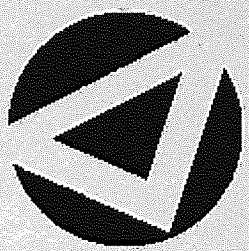
A Novel Entropy Measure for Analysis of the Electrocardiogram

DAN WOODCOCK

Doctor of Philosophy



ASTON UNIVERSITY

February 2007

ASTON UNIVERSITY

A Novel Entropy Measure for Analysis of the Electrocardiogram

DAN WOODCOCK

Doctor of Philosophy (February 2007)

Thesis Summary

There has been much recent research into extracting useful diagnostic features from the electrocardiogram with numerous studies claiming impressive results. However, the robustness and consistency of the methods employed in these studies is rarely, if ever, mentioned. Hence, we propose two new methods; a biologically motivated time series derived from consecutive P-wave durations, and a mathematically motivated regularity measure. We investigate the robustness of these two methods when compared with current corresponding methods. We find that the new time series performs admirably as a compliment to the current method and the new regularity measure consistently outperforms the current measure in numerous tests on real and synthetic data.

Keywords: P-wave, kernel entropy, chaos, time series, regularity, complexity, nonlinear dynamics

For
Adam Dauncey
and
Clive Woodcock

Acknowledgements

I am extremely fortunate that the following list contains so many people. My extreme gratitude to

- Ian Nabney for the initial project idea, astute advice, encouragement and extraordinary patience shown in the course of this research; as well as useful tips on piano technique. Any academic merit within this thesis is a result of his excellent supervision. Any mistakes are, of course, my own.
- my Mum and Dad for their encouragement, love and support in the last 7 years I've spent at university. The old cliché is literally true in this case; without their help, this undertaking would not have been remotely possible.
- Cardionetics Ltd. for allowing me to use their ECG processing program. As the project was intended to be mainly based on the creation of ECG signal processing techniques and not P-wave extraction, this gave me an invaluable head start.

I would also like to thank

- the members of the NCRG, particularly David Lowe, who was always available for advice and for the seminar which led to the idea for kernel entropy. Also many thanks to Laura Rebollo-Neira, Dan Cornford and Jort van Mourik for their support over the last three years.
- My undergraduate lecturers who helped me decide to take it further than a degree. Particularly Bill Cox, whose eloquent lecture on the Euler formula made such an impact on an indifferent undergraduate. Also many thanks to Sotos Generalis for his infectious enthusiasm for the subject.
- my fellow Ph.D students, particularly Dharmesh Maniyar, for putting up with relentless Linux questions and helping celebrate after Evel Knievel jumped four Mexican wrestlers during a rainy cricket match.
- the MSc PANN students of 2003-2004, especially Pierre Prévot, for enduring the weekly coursework marathon with me.
- the excellent NCRG support staff, Vicky Bond and Alex Brulo.
- my friends for keeping me sane. Particularly (in the order I think of them): Ste Weale, Ted Regan, Mark and Sarah Parsons, Matt Vickers, Kate Vernon, Mark and Claire Nichols, Stephen Foulger, Jon Copestake, Pete Phillpot, Richard and Karren Cavalot, Rob Butler, Paul Essex, James and Dave Curry, Ste Mattey, Dave Ward, Andrew Powell, Matt Parfey, Paul Brown and my brother Ben.

- the distinguished gentlemen from Aston Old Edwardians R.U.F.C. for allowing me to de-stress in a sporting context. Extra thanks to Paul Glenn, Tim Watson and Brendan Mulligan for the initiation.
- the rest of my family, with an apology for not seeing them as much as I should; Ness and Rick Woodcock, Puri and Mario Fernandez and Maria and Marcos Vieria.
- David Evans and Richard Everson for the invigorating discussion that was my viva!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Surface recordings of cardiac electrical activity were originally taken by Willem Einthoven in 1903 and have since become the primary diagnostic tool for cardiac disorders. Einthoven originally called this recording the *elektrokardiogramme*, which leads to the abbreviation *EKG* that is still used in some places, particularly the United States [Wagner, 2001]. However, we shall henceforth refer to it as the anglicised *electrocardiogram* or *ECG*.

Much effort has been put into determining the diagnostic features of various cardiac disorders from visual analysis of the electrocardiograph output with a good deal of success. In more recent times, the power of modern computers has led to a number of automated approaches being proposed, which have shown promising results. However, many of these automated approaches have not found their way into clinical use, despite innumerable potential benefits to clinicians. This is in the most part due to unknown reliability issues exacerbated by the cost and the difficulty in obtaining medical approval for the undertaking of such research. Therefore, it is paramount that when opportunities arise for clinical experiments to take place, the techniques to be trialled should be developed to the highest levels of robustness possible in a non-clinical environment.

The aim of this thesis is to introduce and investigate two novel approaches for use in the understanding and diagnosis of a range of cardiac disorders. The disorder that will be primarily under investigation (and hence motivated the development of the approaches) is paroxysmal atrial fibrillation, a condition causing the heart to beat in an irregular manner. It was noted, however, that for diagnostic use, any automated method should be able to distinguish between a number of cardiac conditions; if someone is complaining of heart problems, a method which distinguishes between a single condition and the normal heart beat is unlikely to be of much use. Unfortunately, tests on multiple conditions are comparatively rare in the

current research literature. Therefore, we also evaluate the effectiveness of the approaches in distinguishing between a number of cardiac conditions.

1.1 Summary of Chapters

The thesis is organised into chapters as follows

1.1.1 Chapter 2 - Literature Survey

This chapter is intended as a primer so that a reader with little or no prior knowledge of the main fields the thesis encompasses should be able to understand what follows. Hence, we review the relevant physiology of the heart, how the electrocardiogram is generated and how to understand its use in diagnostics. We then briefly summarise the pattern recognition techniques used in this thesis which are roughly split into classifiers, visualisation techniques and sampling methods. The final section gives a wider treatment of the methods used to extract information from cardiac time series. A number of methods are included here that are not used or mentioned again in the thesis. This is to give the reader a reasonable background of techniques to facilitate understanding of the place in the time series analysis catalogue in which the method introduced in Chapter 4 will occupy and how it relates to the other approaches.

1.1.2 Chapter 3 - Investigating the P-wave

This chapter is devoted to the preliminary investigation of the P-wave (one of the sub-waveforms of the ECG signal). The first section gives a short background of techniques currently in place and why we adopted the approach we used. The second section details our method for extraction of the P-wave from the rest of the ECG signal, with a pseudo-code version of the algorithm we used along with step-by-step instruction. The third section is concerned with the preliminary analysis of the P-wave data and uses some of the methods discussed in Chapter 2 to this end. This is followed by conclusions drawn from the analysis.

1.1.3 Chapter 4 - Development of Kernel Entropy

Here we introduce the *kernel entropy*, a novel method of quantifying the regularity of a time series. We start from the principle of entropy rate, with a theoretical definition, and then show how this has been used as the basis of previous regularity measures. We then show how

issues in these measures may be resolved by adopting a different density estimation scheme, with a mathematical property employed to ensure computation tractability of the method. Then we define and utilise a method for estimating the optimal parameter for the density estimation and then compare the new measure with the current standard measure in a variety of experiments using synthetic data sets. This is followed by a conclusion on the effectiveness of the new measure.

1.1.4 Chapter 5 - Application of the Methods

Now we apply a number of measures (including the kernel entropy) to both the series of consecutive P-wave durations, known as the P-wave length (PWL) series, and the more conventional RR-interval (duration between consecutive beats) series. The aim here is to evaluate the effectiveness of both the PWL series and the kernel entropy on real data in several situations. A number of experiments are proposed to this end which are carried out and conclusions drawn from the results.

1.1.5 Chapter 6 - Summary of Thesis

In this final chapter, we recapitulate the main points in this thesis, with consideration given to any outstanding matters. We then close with a discussion of possible extensions to this research and how they might be achieved.

Chapter 2

Literature Survey

As the scope of this research is quite broad, this chapter consists of an overview of the literature that has been studied in the compilation of this thesis. However, greater detail on important points may be found in the relevant chapters; the aim here is to provide insight into areas pertinent to the project as a whole since a certain amount of prior knowledge is needed to understand the range of topics covered and their validity to the project goals.

The chapter is split into three parts that correspond to the main areas investigated during the research. The first section describes the background of the cardiological processes and conditions relevant to the research, the second section introduces the classifiers, visualisation and sampling techniques used and the third section provides an overview of current complexity and disorder measures.

2.1 Cardiology

The aim of this project is to identify methods that can be used to extract features from the *electrocardiograph* (ECG) which can be used for diagnosis of a variety of cardiac disorders. The ECG is a recording of the electrical impulses generated by the heart and is the primary method of diagnosing cardiac problems, as well as being of use in the diagnosis of non-cardiac complaints [Wagner, 2001].

Before investigating any signal, it is important to understand how the signal is generated and the ECG is no exception. Many people are familiar with the general waveform of the ECG but few are able to explain the electrophysiological meaning of an ECG signal and how it relates to the physiological mechanisms of the heart. In this section, we shall see how the human heart works, how this relates to the generation of electrical signals and how this may

be of use in the analysis of the ECG.

2.1.1 Physiology

The heart is the organ that supplies blood to the rest of the body. It does this by rhythmically contracting certain blood-filled muscular chambers which forces blood around the body. The process whereby blood enters the heart, is pumped to the lungs and back, and is then pumped around the body is called the *cardiac cycle*. There is a substantial amount of literature that investigates the chemical and myocardial processes in the heart during the cardiac cycle (e.g. [Julian, 1978]) but, as we are only interested in the context of how the ECG is generated, such a detailed treatment is beyond the scope of this thesis.

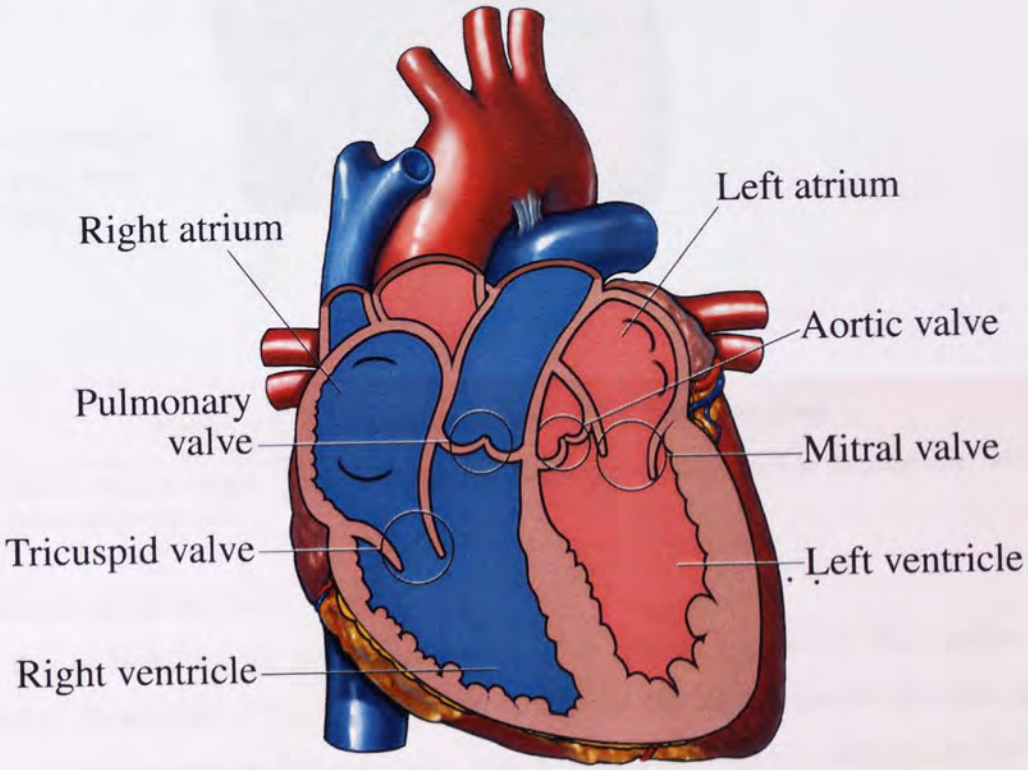


Figure 2.1: The heart with the chambers and valves identified.

The heart is split into four chambers as can be seen in Figure 2.1. These are further divided into the *atria*, which are at the top, and the *ventricles*, which are at the bottom. The atria and ventricles are further categorised as "left" and "right" which is taken from the patient's perspective; i.e.the "left" side of the heart is nearest to the left arm. The purpose of the atria is to fill with blood and then pump it to the corresponding ventricle which then

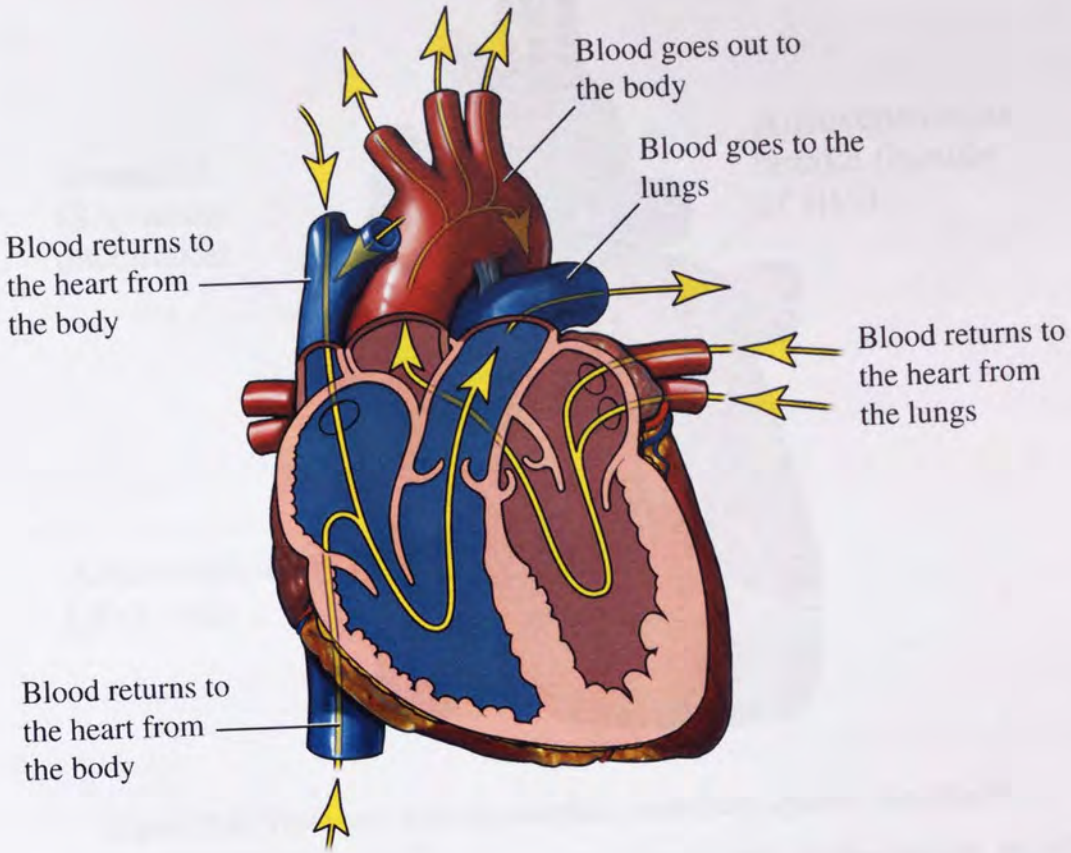


Figure 2.2: The heart with the blood flow identified.

pumps blood out of the heart.

The flow of blood during the cardiac cycle can be seen in Figure 2.2. The cardiac cycle begins when deoxygenated blood (indicated as blue on the figure) enters the right atrium via a blood vessel known as the vena cava. The atrium then contracts and pumps the blood into the right ventricle which, once filled, contracts and pumps blood through the pulmonary artery to the lungs where it absorbs oxygen. This oxygenated blood (indicated as red on the figure) then enters the left atrium via the pulmonary vein; the atrium then contracts, forcing the blood into the left ventricle which in turn contracts, pumping blood around the whole body. Although this appears to be a four stage process, the contraction of both atria occurs simultaneously, as does the contraction of the ventricles shortly afterwards. Therefore, the cardiac cycle is often thought of as a two stage process and it is this that gives rise to the characteristic rhythm of the heart.

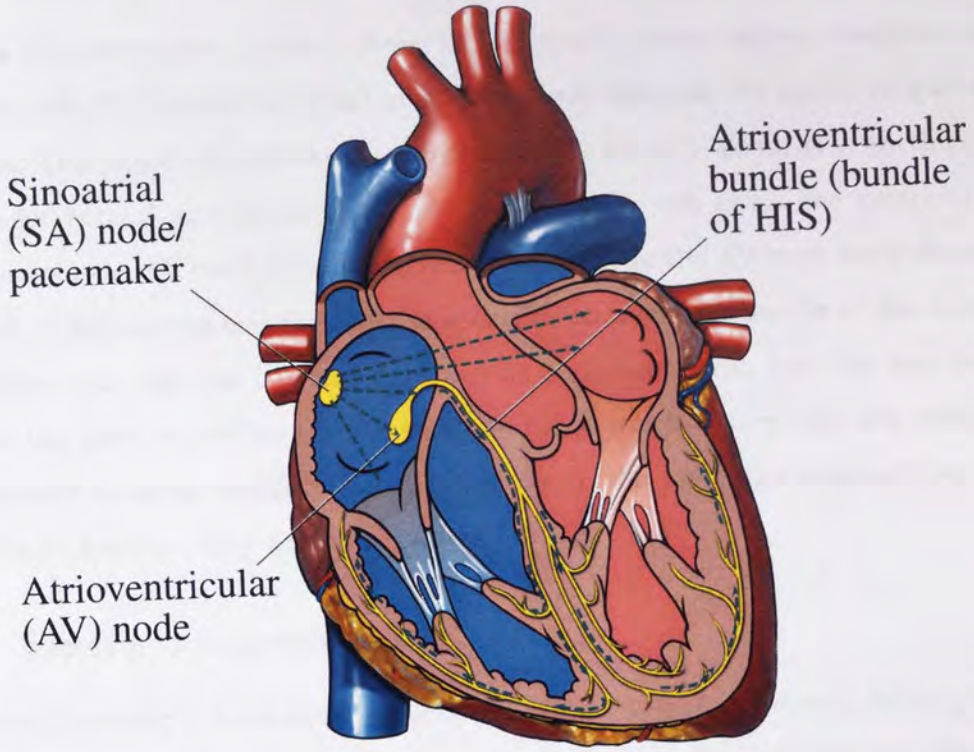


Figure 2.3: The heart with the cardiac conduction system identified.

2.1.2 Electrophysiology

When any muscle contracts there is an associated electrical depolarisation and the heart is no exception. In fact, the cardiac cycle is an ordered sequence of muscular contractions and therefore has a corresponding sequence of electrical depolarisations. The ECG detects these electrical impulses so it is important to be aware of the underlying electrical mechanisms that are present in the heart and their relevance in the cardiac cycle. Figure 2.3 shows the main areas involved in the electrical process of the heart which are collectively called the *cardiac conduction system* and are shown in yellow.

The electrical inception of the cardiac cycle is at the *sinoatrial* (SA) node which is a specialised bundle of nerve fibres that act as the body's natural pacemaker [Harel et al., 1998]. The SA node is located in the upper part of the right atrium; it starts the chain of electrical events which form the cardiac cycle. An electrical impulse is generated in the SA node which depolarises across the two atria (indicated by the green arrows in Figure 2.3), causing them to contract simultaneously. The impulse also reaches the *atrioventricular* (AV) node, which is also situated in the right atrium, nearer to the muscular partition that divides

the atria (the interatrial septum). Here, the electrical process pauses momentarily for the AV node (which generates potential at a slower rate than the SA node) to gain sufficient potential. This pause also allows the ventricles time to fill with blood. It is worth mentioning that, under normal circumstances, no electrical impulse can affect the ventricles without going through the AV node [Brembilla-Perrot, 2002]. Once the AV node has sufficient action potential, it depolarises and the impulse is carried through the bundle of His to where the nerve fibres split into the left and right bundle branches. From here the impulse spreads out into the lower ventricles via the Purkinje fibres which ensure that the contraction of the ventricles happens regularly from the bottom to the top. The muscles then relax and repolarise to become ready for the next beat.

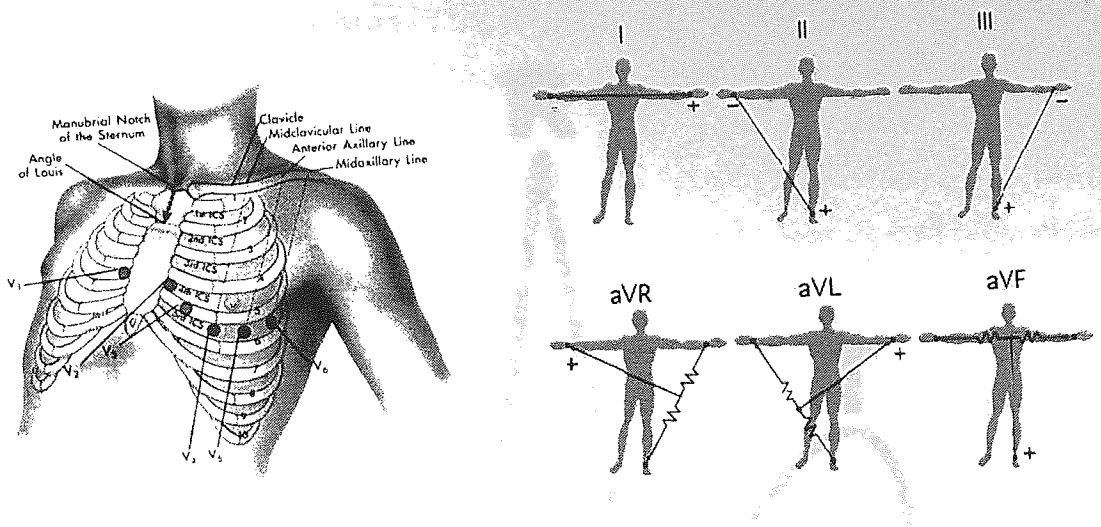
2.1.3 Electrocardiography

Electrocardiography is a non-invasive technique where electrodes (sensors) are attached to the outside of the body and recorded by a device known as an *electrocardiograph*. The recorded data is the *electrocardiogram*. There are a variety of ways to configure the position of the electrodes for an electrocardiogram but the standard is the 12-lead (12-channel) system. The basic principle is that if the depolarisation within the heart is moving toward a positive electrode, or repolarisation is moving away, then the ECG will register a positive (upward) deflection and vice-versa. It is clear from this that the position of any electrode is relevant to the interpretation of the output obtained. This is particularly evident in the orientation of the predominant features in the ECG; if these are generally positive then the depolarisation is moving toward the positive electrode. Further to this, it can also be seen that suitable positioning of an electrode would increase the focus on a particular area of the heart.

12-Lead ECG

In a twelve-lead ECG, there are 6 leads (known as the V-leads, V1 to V6) which monitor the heart on the horizontal plane and 6 leads that take measurements from the heart on the frontal plane which are known as the limb leads. The positioning of the lead attachments can be seen in Figure 2.4.

The positioning of the limb leads can be seen in Figure 2.4(b); electrodes are attached to both arms, and the left foot. The first three channels are known as leads I, II and III and measure the potential between the left and right arms, the left leg and the right arm, and the left leg and left arm respectively. The other three channels are called the augmented



(a) Positioning of leads V1-V6.

(b) Positioning and relative angles of the limb leads.

Figure 2.4: The positioning of the electrodes in the 12-lead system.

Figure (a) is adapted from `www.kauaicc.hawaii.edu`, (b) is adapted from `www.nobelprize.org`.

leads as they measure the potential between one of the limb leads, which is positive, and the other two negative leads. These are known as aVR (right arm is positive), aVL (left arm is positive) and aVF (left foot is positive).

The positioning of the 6 V-leads is shown in Figure 2.4(a). These are also known as the *precordial* leads as they are immediately in front of the heart. The V-leads are all positive with the negative pole being the central terminal formed by averaging the limb leads [Wagner, 2001].

As mentioned before, each one of the 12 leads measures the electrical activity from a different angle. If the depolarisation moves toward the positive electrode of any of these leads, it will register a positive deflection on the corresponding channel of the ECG. The principal direction of electrical depolarisation of the heart occurs along the interatrial septum which is normally orientated on an axis 30° anticlockwise from the vertical [Wagner, 2001]. For example, in the limb leads, this would mean a strong positive deflection in lead II which is approximately in the same direction as the interatrial septum but would register a smaller negative deflection in lead aVR. This shows that the choice of ECG channel will have an effect on the shape of the signal. It also causes a problem in the task of determining consistent diagnostic features using automatic computational methods due to inconsistent lead placement or variations in cardiac physiognomy or placement.

Identifying the ECG complexes

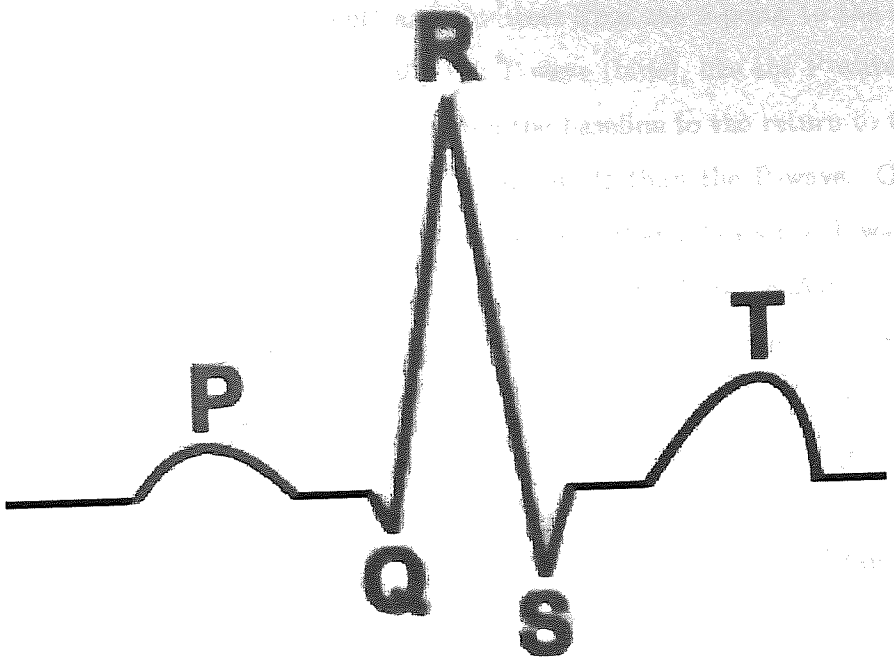


Figure 2.5: A representation of a typical ECG output with complexes identified.

Figure 2.5 shows a representation of an ECG signal for a normal heartbeat on a lead such as lead II with the main points of interest identified. Here we shall concentrate on the exact definitions of each as they appear in the signal.

The most recognisable part of the ECG signal is the QRS-complex (shown in green), which is often used as a reference point to find other ECG features as it is the most distinctive structure in the ECG [Dotsinsky and Stoyanov, 2004]. The P-wave (red) precedes this and is defined as the segment of the signal from the first upward deflection from the baseline to the point when the signal returns to the baseline (although, with all ECG structures, this may be inverted depending on the lead). There is then sometimes a period of rest known as the PR segment which is the time from the beginning of the P-wave to the start of the QRS-complex. The components of the QRS-complex are labelled depending on its morphology and often look quite different depending on the location of the sensors used to measure the ECG. However, the definitions of the points remain the same. If the first deflection in the QRS-complex is downwards then it is known as the Q-wave. Any upward deflection in the QRS-complex is called an R-wave, whether it is preceded by a Q-wave or not, and a deflection

below the baseline after an R-wave is always known as an S-wave. The Q, R and S-points are the turning points of the waves. The point where the S-wave returns to the baseline is known as the junction point (or J-point) and the time from the J-point to the start of the T-wave is known as the ST-segment. Finally the T-wave (blue), like the P-wave, is defined as the signal from the first upward deflection from the baseline to the return to the baseline after the QRS-complex. This is often larger in amplitude than the P-wave. Occasionally, there is a further wave known as the U-wave (not shown) that follows the T-wave; however the U-wave has no known clinical significance and therefore is often ignored.

As described above, on some leads the QRS-complex will be predominantly negative and also may not start with a Q-wave. In this case, the first upward deflection is still called an R-wave but is denoted by lower case `r` as it is of small amplitude. The downward deflection after the `r`-wave is still known as the S-wave but the subsequent deflection above the baseline is known as `r'` (r prime) so as not to get confused with the initial `r`-wave [Hampton, 1986].

Electrophysiological Meanings of ECG Structures

The depolarisations and repolarisations that cause the heart to beat correspond with the various structures of the ECG output. The depolarisation of the SA node and subsequent transmission of the impulse around the atria correspond to the P-wave. The PR-segment is the pause for the AV node to generate the action potential needed to continue the cardiac cycle. The QRS-complex corresponds to the depolarisation of the AV node and the impulse moving down the bundle branches and into the Purkinje fibres. The T-wave is due to the repolarisation of the ventricles which occurs in the opposite direction to the depolarisation (giving it the same orientation on the ECG). It is worthy of note that there is a lack of a defined structure that corresponds to the repolarisation of the atria. This is because it happens concurrently with the depolarisation of the AV node and the small amplitude of the atrial repolarisation is hidden by the far greater amplitude of the QRS-complex.

Artifacts

The ECG signal is prone to interference as the electrodes are placed on the outside of the body and can be susceptible to other depolarisations in the body caused by other muscles contracting, movement by the patient, poor electrode contact and alternating current artifacts. Removal of these artifacts has obvious benefits in signal processing and much research has been put into this already [Liu and Kao, 2003]. In Chapter 3 we explain how we deal

with any artifacts.

2.1.4 ECG Intervals

Apart from the actual morphology of the ECG, information can be gained from the lengths and amplitudes of certain segments in an ECG recording. For example, an elevated ST-segment [American College of Cardiology/American Heart Association Task Force on Practice Guidelines, 2004] and slight changes in the T-wave [Hunt, 2002] can be indicators of a myocardial infarction or other disorders [Engel et al., 2004].

However, most automated studies are based on the *intervals* of, or between, certain ECG segments. By far the most researched are the interbeat intervals; the duration between successive beats (e.g. [Task force of the European Society of Cardiology and the Northern American Society of Pacing and Electrophysiology, 1996; Teich et al., 2001; Small et al., 2000; Maier et al., 2001b]). However, the use of other intervals has been investigated, such as the QT-interval [Doevendans, 2000] and the P-wave duration [Steinberg et al., 1993].

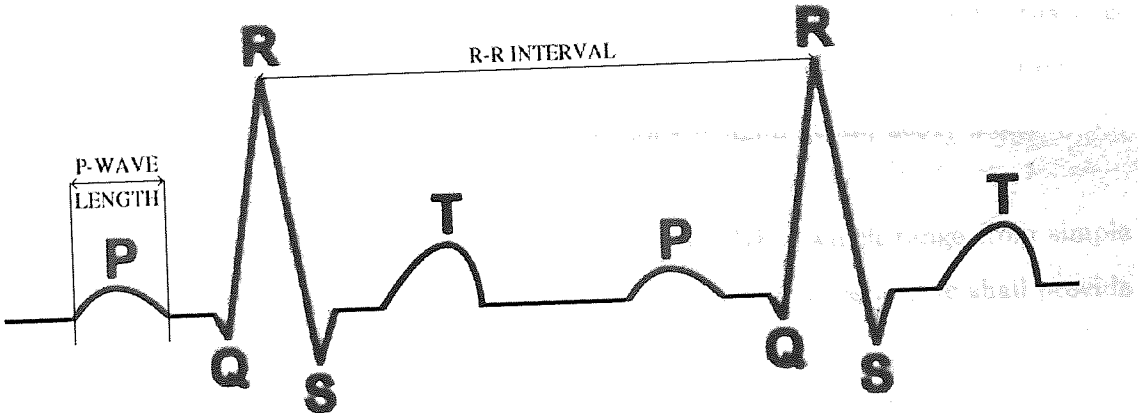


Figure 2.6: A representation of a typical ECG output with the RR-Interval and P-wave Length identified.

Interbeat Intervals

The timing of the sequence of interbeat intervals of the heart has been the subject of many studies and has been shown to be of significance in identifying various cardiac anomalies. To discern the time interval between beats in the ECG, the location of the R-point is often used as the reference point as it is the most significant and easily identifiable structure in the ECG. The duration between successive R-points is called the *RR-interval* and analysis of the

fluctuations in the sequence of RR-intervals is known as heart rate variability (HRV) analysis. Most techniques utilise what are known as NN (Normal to Normal) intervals which are the time intervals between adjacent QRS-complexes resulting from sinus node depolarisations so anomalous occurrences such as ectopic beats are excluded. A representation of these can be seen in Figure 2.6.

Whether the RR-interval series arises from a deterministic dynamical system has long been a matter of uncertainty amongst researchers with papers claiming both that it is [Wagner and Persson, 1998] and that it is not [Teich et al., 2001].

There is also some indication that the cardiac behaviour of normal patients varies throughout the day, indicating that any underlying chaotic disorder is related to a periodic circadian cycle [Cugini et al., 1999] which would complicate matters further. Also, the heart is not an isolated system, impulses arrive from the autonomous nervous system at regular intervals which affect the variability. The presence of a cardiac condition could interfere with these natural sources of variability which may be able to be detected by suitable techniques.

The study of HRV is of undetermined physiological importance as only limited conclusions can be drawn regarding the behaviour of the system from solely the heart rate fluctuations [Harel et al., 1998]. However, it has proved beneficial in studies of cardiac disfunction and has been of use in distinguishing a number of conditions [Cugini et al., 2001; Maier et al., 2001a; Teich et al., 2001].

There are a wide variety of techniques used to measure HRV which range from simple average interval duration to non-linear analysis of the resulting time series; we shall provide an overview of these in Section 2.4.

2.2 Cardiac Disorders

There are a very large number of cardiac disorders, with a wide variety of causes and treatments. As a consequence, we shall provide only an overview of the conditions pertinent to this research; atrial fibrillation, sleep apnoea and congestive heart failure. The overview consists of review of the cardiac mechanisms that related to the condition, how the condition may be detected from the ECG and a review of the data sets that we use for each condition. All the data sets were from the PhysioNet website (www.physionet.org) Goldberger et al. [2000].

2.2.1 Paroxysmal Atrial Fibrillation

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia with an estimated 0.4% of the population counted as sufferers which accounted for 34.5% of patients hospitalised with a cardiac rhythm disturbance. It is also increasingly prevalent amongst the elderly [Committee to Develop Guidelines for the Management of Patients With Atrial Fibrillation, 2001].

Its physical symptoms are a rapid and irregular heart rate [Martin, 2003] often with severe discomfort. Paroxysmal atrial fibrillation (PAF) is self-terminating and may be symptomatic of underlying problems in a variety of areas especially as this uncoordinated atrial activation can lead to deterioration of atrial function. The causes are not clearly known; sometimes there can be structural abnormalities in the atria, where the juxtaposition of normal and diseased atrial fibres could account for the fractioned electrical activity. It has also been suggested that fatty infiltration or inflammation of the atria could be a cause [Committee to Develop Guidelines for the Management of Patients With Atrial Fibrillation, 2001].

Complications arising from paroxysmal AF are often devastating because of the sporadic dramatic changes of heart rate and regularity. Strokes can often be attributed to atrial fibrillation and in some cases, the AV node over-stimulation can lead to ventricular fibrillation and death. However, many patients with atrial fibrillation have no symptoms and are unaware of the abnormal heart rhythm [Lip and Li Saw Hee, 2001].

Cardiac Mechanisms

The mechanisms of AF are not fully understood but there are certain processes that are believed to occur during an AF episode. One is that several foci other than the SA node depolarise in an accelerated and irregular manner. These may be located in the pulmonary vein or other blood vessels [Committee to Develop Guidelines for the Management of Patients With Atrial Fibrillation, 2001]. Another suggestion is that the wave fronts become fractioned as they propagate through the atrial mass leading to multiple wavelets propagating in different directions [Poli et al., 2003].

ECG Indicators

The ECG of a person in AF will show the absence of defined P-waves with them being replaced by rapid oscillations which vary in size, shape and timing. The erratic timing of the activation of the AV node due to the disordered depolarisation causes the ventricles to

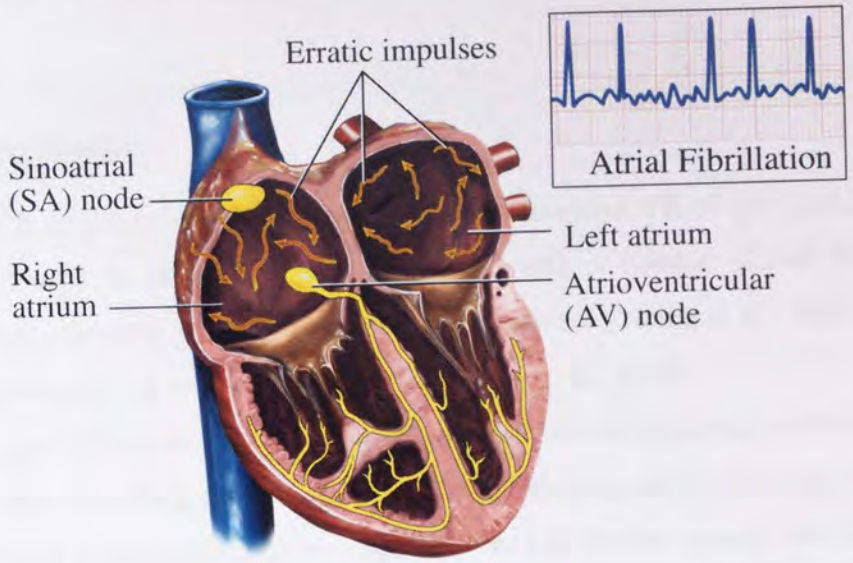


Figure 2.7: A diagram showing the heart during atrial fibrillation.

contract in an irregular manner. The QRS-complex and T-wave, however, have a normal shape as the process of ventricular contraction is not affected by AF itself. This does mean that the time in between heart beats will vary from beat to beat despite the ventricular component of the ECG being unaffected *per se*.

This indicates that the primary detection method for atrial fibrillation would be predicting it from changes in the P-wave since the QRS-complex will remain largely unchanged due to normal ventricular operation. Also, the variability of interbeat intervals due to irregular AV node excitation may be a useful indicator of an impending PAF episode and analytic methods based on this idea have been applied with reasonable success [de Chazal and Henegan, 2001; Maier et al., 2001a].

Data

The paroxysmal atrial fibrillation data set that we used in this thesis was from the *PAF Prediction Challenge Database*. The data consists of 100 pairs of half-hour ECG recordings equally split into groups N and P. Each pair of recordings is obtained from a single 24-hour ECG. Subjects in group P experienced PAF; for these subjects, one recording ends just before the onset of PAF (group P_c), and the other recording (group P_d) is distant in time from any PAF (there is no PAF within 45 minutes before or after the excerpt). Subjects in group N do not have PAF; in these, the times of the recordings have been chosen at random Goldberger

automated diagnostic method is more difficult.

Data

The data set is the *Detecting Sleep Apnea from the ECG* data set. It consists of 25 continuous recordings of approximately 8 hours in length. For the purposes of this thesis, a continuous 30 minute section was randomly drawn from the larger sample for fair comparison with the atrial fibrillation data set. Twenty of these records contain recordings with at least 100 minutes of apnoea in the recording. Five of the recordings contain between 5 and 99 minutes of apnoea.

2.2.3 Congestive Heart Failure

Congestive heart failure (CHF) is a term used to describe any condition that impairs the heart's ability to pump a sufficient supply of blood around the body due to a structural or functional abnormality [American College of Cardiology/American Heart Association Task Force on Performance Measures, 2005] leading to the impairment of cellular respiration [Shamsham and Mitchell, 2000]. An estimated 0.4% to 2% of the population suffer from CHF with most sufferers aged 65 and over with no apparent predeliction for a specific gender [Task Force for the Diagnosis and Treatment of Chronic Heart Failure, European Society of Cardiology, 2001; Senni et al., 1998]. There does, however, appear to be an increasing number of sufferers from the western world which has been attributed to the rise in obesity [Hubert et al., 1983].

There are a large number of conditions that may cause the heart to function in this detrimental fashion. They include hereditary and congenital structural causes, myocardial dysfunction, arrythmias, valve abnormalities or rhythm disturbances and the condition can be exacerbated by smoking, obesity and alcohol and drug abuse [Task Force for the Diagnosis and Treatment of Chronic Heart Failure, European Society of Cardiology, 2001].

The usual method of diagnosis is by assessment of the left ventricular systolic (contraction) function which shows a reduced flow in diseased hearts. This is often done using an ultrasound of the heart (echocardiography) to obtain an image which a clinician can then use to assess the ventricular flow (ejection fraction) [American College of Cardiology/American Heart Association Task Force on Practice Guidelines, 2001]. The diagnosis by echocardiogram provides the clinician with a lot of information on the specific anomalies within the heart as the resolution is such that abnormal valve and myocardial function can be determined [Task Force for the Diagnosis and Treatment of Chronic Heart Failure, European Society of

Cardiology, 2001]. However, echocardiography is often only used after referral to a hospital by a general practitioner, based on cardiac history and symptoms, and then only after an ECG and other methods have failed to preclude the possibility of CHF. The ECG is of limited diagnostic value as, although it can detect anomalous cardiac behaviour easily and cheaply, it is considered insensitive and non-specific. Its clinical use is to provide useful information but only as a guideline toward a positive diagnosis [American College of Cardiology/American Heart Association Task Force on Practice Guidelines, 2001].

Cardiac Mechanisms

Apart from reduced blood flow, there is no singular cardiac malfunction that characterises CHF. The presence of any anomalous cardiac defect can cause CHF which is the reason for the unusual meticulousness in obtaining a rigorous diagnosis as the more precise the diagnosis, the more effective the treatment and subsequent prognosis will be.

ECG Indicators

As previously mentioned, the ECG should not be used alone to diagnose a sufferer but a number of signs can be picked up on the ECG that may warrant further investigation. A number of sufferers exhibit tachycardia as the heart speeds up to counteract the insufficient blood flow. This can happen intermittently or in a sustained fashion. Apart from this, the ECG can be examined for evidence of ventricular hypertrophy (which can be determined by a enlarged QRS-complex), atrial enlargement (abnormally large P-waves), conduction abnormalities (unusual ECG structure morphology/orientation dependent on location of abnormality), myocardial infarction (ST depression or T-wave inversion), and active ischemia (ST depression) [Bales and Sorrentino, 1997].

Data

This data set is from the *BIDMC Congestive Heart Failure Database*. It consists of 15 recordings with severe congestive heart failure of about 20 hours duration. For the purposes of this thesis, a continuous 30 minute section was randomly drawn from the larger sample for fair comparison with the other data sets.

2.3 Pattern Recognition Techniques

We have used a number of pattern recognition techniques throughout this project and so it is necessary to provide the background of the methods to aid understanding of the results. The algorithms used can be categorised into visualisation techniques, classifiers, and sampling methods so we will review those categories here.

2.3.1 Classifiers

In this thesis, we use neural networks for classification. The goal in classification is to take an input vector $\mathbf{x}$ and assign it to one of K discrete classes $\mathcal{C}_k$ where $k = 1, 2, \ldots, K$. The neural network divides the input space into *decision regions* whose boundaries are called *decision boundaries*. The nature of these decision boundaries depends on the type of neural network model. A linear model will produce straight decision boundaries whereas nonlinear models will produced curved decision boundaries.

The classifiers used in this thesis are purely for evaluation of the feature extraction and so only a basic understanding of them is necessary and details on peripheral matters such as standard training algorithms are omitted. They are all types of neural network ranging from the simplest single-layer network to multi-layer networks implemented in the NETLAB [Nabney, 1999] toolbox.

Neural networks are inspired by, and loosely modelled on, the way the central nervous system works. They consist of a number of interconnected elements known as *neurons* or *units* which work in parallel in a similar fashion to the brain. When a number of neurons are combined, it forms the neural network. This is then 'trained' on a *training set* which is a data set that is of a similar type to the intended real data set, often a subset. The training process assigns values to certain parameters with reference to the known outputs which are called *target* variables. The target variable is dependent on the application of the neural network and the number of classes. In this thesis we use *0-1 encoding* which we now describe.

For a two class problem, the target value t is 0 for $\mathcal{C}_1$ and 1 for $\mathcal{C}_2$. For a multiclass problem, we use a vector $\mathbf{t}$ of length K where an element t_j is 1, and all other elements are 0 if the input belongs to class $\mathcal{C}_j$. After training, the neural network is now functional and can be used by inputting a previously unused data value which is propagated through the network and an output y obtained. In a two class problem, the value of y is a number between 0 and 1 that can be interpreted as the probability that the new input belongs to class $\mathcal{C}_2$. In the multiclass setting we have a vector $\mathbf{y}$, of which the elements y_j correspond

to the probability that the new data value belongs to class $\mathcal{C}_j$.

Generalised Linear Model

The simplest linear model is basically a linear combination of the parameters $\mathbf{w} = (w_0, w_1, \ldots, w_d)^T$ and the d-dimensional input values $\mathbf{x}$ so that

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_d x_d, \tag{2.1}$$

$$= \mathbf{w}^T \mathbf{x} + w_0. \tag{2.2}$$

This model will not output values consistent with the 0-1 encoding so we transform the linear function using a nonlinear *activation function* $f(\cdot)$ thus

$$y(\mathbf{x}, \mathbf{w}) = f(\mathbf{w}^T \mathbf{x} + w_0). \tag{2.3}$$

We can extend this further by considering linear combinations of fixed nonlinear basis functions $\phi_j(\mathbf{x})$. This will take the form

$$y(\mathbf{x}, \mathbf{w}) = f(\mathbf{w}^T \phi(\mathbf{x})), \tag{2.4}$$

or

$$y(\mathbf{x}, \mathbf{w}) = f\left(\sum_{j=0}^{c} w_j \phi_j(\mathbf{x})\right), \tag{2.5}$$

where $\phi_0(\mathbf{x}) = 1$ and c is the number of outputs. This is called a *generalised linear model* [McCullagh and Nelder, 1989].

Multilayer Perceptron

A nonlinear neural network is organised into *layers* which are categorised into three types.

- Input layer -
  This is the input data.
- Hidden layers -
  There can be a number of hidden layers. These take the outputs of the input layer or another hidden layer as their inputs.
- Output layer -
  These are the outputs of the network. They take the outputs of the last hidden layer as its inputs.

The number of outputs c and units in the hidden layer M is determined by the user. The number of inputs is the dimensionality of the data d. Each unit in one layer has a connection to each unit in the next layer, with no other connections permitted. Hence, this type of structure is known as a *feed-forward* neural network [Bishop, 1995].

The neural network model is an extension of the model given in Equation 2.5. The model is extended by making the basis functions $\phi_j(\mathbf{x})$ depend on parameters and then allow these parameters to be adjusted, along with the coefficients w_j during training. The multilayer perceptron (MLP) used in this thesis has a single hidden layer and uses basis functions of the same form as Equation 2.5; so each basis function itself is a nonlinear function of a linear combination of the inputs, and the coefficients in the linear combination are adaptive parameters [Bishop, 2006, p226].

The first step to constructing an MLP is to create M linear combinations of the input variables in the same manner as in Equation 2.5. The difference is that the outputs are the inputs to the hidden layer $z(\mathbf{x}, \mathbf{w})$, so the vector $\mathbf{z} = (z_1, z_2, \ldots, z_M)$. Also, the activation function used in this thesis is the '*tanh*' function [Nabney, 1999] which we shall denote $h(\cdot)$. So the function for the inputs to the each of the hidden units z_j is

$$z_j(\mathbf{x}, \mathbf{w}^{(1)}) = h\left(\sum_{i=0}^{D} w_j i^{(1)} x_i\right), \tag{2.6}$$

where the superscript (1) indicates the layer of the network and $j = 1, 2, \ldots, M$.

Consequently, the function from the hidden layer to the output layer is

$$y_k(\mathbf{z}, \mathbf{w}^{(2)}) = \sigma\left(\sum_{j=0}^{M} w_k j^{(2)} z_j\right), \tag{2.7}$$

where σ is chosen due to the nature of the data and the application. In this thesis, we use the logistic sigmoid activation function for two class problems and the softmax activation function for multiclass problems.

So, to see how this has extended the model formulated in Equation 2.5, we can write the whole process to decide an output value as

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=0}^{M} w_k j^{(2)} h\left(\sum_{i=0}^{D} w_j i^{(1)} x_i\right)\right). \tag{2.8}$$

It is easy to see from this how $h(\cdot)$ replaces $\phi(\cdot)$, and $\sigma(\cdot)$ replaces $f(\cdot)$ in Equation 2.5 [Bishop, 2006, p228].

This architecture has been shown to be able to model any smooth function (with suitable weights and bias) and is therefore known as a *universal approximator* [Hecht-Nielsen, 1987].

Radial Basis Function

As an alternative for the choice of $\phi_j(\mathbf{x})$ in Equation 2.5, *radial basis functions* are often used which means that the basis function depends only on a distance measure from a centre μ_j, so

$$\phi_j(\mathbf{x}) = h(\|\mathbf{x} - \mu_j\|). \tag{2.9}$$

The RBF also includes a hidden unit layer but differs from the MLP in that the activation functions of the hidden units are non-linear functions between the input vector and the hidden units with a linear function of the hidden units giving the output layer. The non-linear activation functions are radial basis functions, that is, the function is constructed using a distance metric with respect to the function 'centre', relative to a variance parameter. Other than more efficient training, the performance is similar to that of the MLP.

Network Training

Often, the way of determining the network parameters is to maximise the likelihood of the parameter values. This is equivalent to minimising a sum-of-squares error function of the form

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N} \|\mathbf{y}(\mathbf{x}_i, \mathbf{w}) - \mathbf{t}_i\|^2, \tag{2.10}$$

where N is the number of input vectors in the training set $\{\mathbf{x}_i\}$ and $\{\mathbf{t}_i\}$ is the corresponding target values.

Most training algorithms involve an iterative procedure to minimise such an error function, with adjustment of the weights being carried out at each step.

In this thesis we use the MLP for classification tasks. The method we use to train the MLPs used in this thesis is known as the Bayesian evidence procedure [MacKay, 1992]. This method regularises the weights of the network by introducing hyperparameters, the choice of which can be incorporated into the learning process. Full treatment of the regulation of weights via the Bayesian evidence procedure is outside the scope of this thesis; full details can be found in [MacKay, 1992].

2.3.2 Visualisation and Dimensionality Reduction

These techniques are primarily concerned with representing the data in such a manner that any underlying structure that may be exploited will become apparent. The basic aim of

visualisation methods is to map the data to a lower dimensional space that can be understood by a human observer (i.e. a 2 or 3 dimensional space).

Dimensionality reduction often uses similar techniques, but the aim is not to represent the data in a fashion that can be understood by a human. Instead, the aim is to reduce the dimensionality of the data to retain as much information as possible to reduce the computational burden and to avoid the curse of dimensionality (wherein a overly high dimensional data space compared to the number of data points will lead to a poor representation of the structure of the data [Bishop, 1995]). There are two main methods of dimensionality reduction known as *feature selection* and *feature extraction*. Feature selection involves reducing the dimensionality by choosing a subset of the original data that retains a suitable amount of information. Feature extraction methods transform the data in such a fashion that optimal information is retained in fewer dimensions. In this thesis, as our data is derived from the P-wave and are continuous recordings of different lengths and each value is likely to be highly correlated with the others, taking subsets of the data values at specific times would not be suitable. Hence, feature extraction methods are more applicable and it is these that are reviewed below. This means that the data points are projected onto a lower dimensional manifold defined by a combination of the original variables. We now review the two methods used.

Principal Component Analysis

Principal component analysis (PCA) is a *linear projection* method which can be used for both visualisation and feature extraction [Jolliffe, 1986]. It is a common technique that combines the inputs using a linear transformation in such a manner that the maximal variance is retained. The aim is to map a dataset of N vectors

$$\mathbf{x}_i \text{ of a } d\text{-dimensional space,}$$

$$\text{onto}$$

$$\mathbf{y}_i \text{ in an } Q\text{-dimensional space,}$$

where $i = 1, 2, \ldots, n$.

If consider the case where $Q = 1$, and $\mathbf{u}_1$ is a d dimensional vector of unit length ($\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$). Each data point is then projected onto a scalar value $\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_i$.

If the sample mean is given by

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i, \tag{2.11}$$

then the variance of the projected data is given by

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_i - \mathbf{u}_1^{\mathrm{T}}\bar{\mathbf{x}}_i\right\}^2 = \mathbf{u}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_1, \tag{2.12}$$

where

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}. \tag{2.13}$$

Therefore, as we want to maximise the variance of the projected data, we need to maximise $\mathbf{u}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_1$. As this is constrained by the normalisation condition $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$, we can construct a Lagrangian and maximise that. The Lagrangian in this case is

$$\mathbf{u}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_1 - \lambda_1(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1). \tag{2.14}$$

Taking the derivative with respect to $\mathbf{u}_1$ equal to zero, we can see this quantity will have a stationary point when

$$\boldsymbol{\Sigma}\mathbf{u}_1 - \lambda_1\mathbf{u}_1. \tag{2.15}$$

So multiplication of $\mathbf{u}_1$ and $\boldsymbol{\Sigma}$ results in a scaling of the vector $\mathbf{u}_1$ by a value λ_1. This is the definition of an *eigenvector* so $\mathbf{u}_1$ must be an eigenvector of $\boldsymbol{\Sigma}$. As the magnitude of an eigenvector is related to its corresponding eigenvalue, the maximal variance is retained when we use the eigenvector that corresponds to the largest eigenvalue. This eigenvector is known as the first *principal component*. For $Q > 1$, additional principal components can be chosen in an incremental fashion in orthogonal directions to those already chosen. This means that the optimal linear projection for which the variance of the projected data is maximised is defined by the Q eigenvectors having the Q largest eigenvalues [Bishop, 2006, p561].

For visualisation we choose $Q = 2$, but for feature extraction, we need to take the dimensions that contain the most information and discard those that contain noise. A heuristic rule is to plot the eigenvalues in descending order and see if there is a point at which the values level off. We then take the corresponding eigenvectors before this point.

However, in an automated approach, there is no steadfast rule for selecting the number of dimensions for use in PCA. The normal method is to take the first Q Eigenvectors which incorporate 95% of the relative value of the Eigenvalues. However, such thresholding is not mathematically robust and we employ a method which uses Bayesian model selection on the probabilistic PCA (PPCA) [Tipping and Bishop, 1999] model which is a generative model for PCA. Bayesian model selection generally selects simpler models that can be applied to

a wider range of data sets. The probability of the data given the model is calculated by integrating over all the unknown parameters in that model:

$$p(D|\mathcal{M}) = \int_{\theta} p(D|\theta)p(\theta|\mathcal{M})\mathrm{d}\theta. \tag{2.16}$$

This is known as the *evidence* for model $\mathcal{M}$.

As the Bayesian evidence of the PPCA model is not computationally tractable, it is approximated in Minka [2000] using Laplace's method, simplifying the resulting equation. However, testing of this method on synthetic data did not estimate the intrinsic dimensionality correctly and so we use a further simplification of Laplace's method, known as *Bayesian information criterion* (BIC) which performs well.

Minka [2000] gives the BIC approximation of the Bayesian evidence of the PPCA model is as

$$p(D|k) \approx \left(\prod_{j=1}^{k} \lambda_j\right)^{-N/2} \hat{v}^{-N(d-k)/2} N^{-(m+k)/2}, \tag{2.17}$$

where k is the dimension being tested, d is the original dimensionality, N is the total number of data vectors, λ_j is the j^{th} eigenvalue, $\hat{v} = \frac{\sum_{j=k+1}^{d} \lambda_j}{d-k}$, and $m = k(d-1)(k+1)/2$.

This calculates the probability that the data D has an intrinsic dimensionality k. The value of k that corresponds to the largest probability is the correct dimensionality by this method.

NeuroScale

A drawback of PCA is that it is a linear technique which means that any non-linear correlations in the data would be overlooked. NeuroScale [Lowe and Tipping, 1997] is a non-linear projection method that uses a radial basis function neural network to perform the mapping to the feature space.

The aim of NeuroScale is to retain the *topological* structure of the data upon projection to a lower dimensional space. In essence, the projection mapping is constructed so that the Euclidean distances between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the data space $d^*_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ should correspond as closely as possible to the distances in feature space $d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$ [Lowe and Tipping, 1996]. This is achieved by using a special error function is known as the *stress* which is given as

$$E = \frac{1}{\sum_{ij} d_{ij}} \sum_{i<j} \frac{(d^*_{ij} - d_{ij})}{d_i j}. \tag{2.18}$$

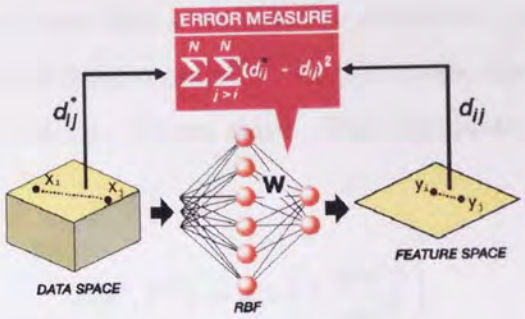


Figure 2.8: Schematic representation of the NeuroScale model, reproduced with permission.

Figure adapted from [Tipping and Lowe, 1997]

This is differentiable and can therefore be trained using conventional techniques.

This error function is used as the basis function in a radial basis function network. The original data points are used as the inputs to the radial basis function which has the projection space points as its outputs. The error function given in Equation 2.18 is differentiable and can therefore the network can be trained using conventional techniques. However, as the computational demands to train such a neural network with such techniques grow in the order of the square of the number of data points, an efficient training algorithm is used, known as *shadow targets* [Tipping and Lowe, 1997]. This makes use of the special form of the error function and the linear dependence of the network outputs on the output layer weights [Nabney, 1999]. It is this efficiency combined with the topological preservation inherent in the NeuroScale method that makes it suitable for effectively mapping a large number of data points to a suitable visualisation space.

2.3.3 Sampling Methods

The basic principle of sampling methods is that we select a statistical sample to approximate a posterior [Denison et al., 2002]. Often, precisely determining the posterior is not plausible, hence we draw an approximation by sampling from an appropriate distribution. Many sampling methods exist, but the one used in this thesis is known as Markov chain Monte Carlo (MCMC); specifically the Metropolis-Hastings algorithm [Hastings, 1970], details on which we give below.

The principle of MCMC is that we aim to sample the distribution by maintaining a record of the current state $p(\mathbf{z}^{(t)})$ at iteration t, and a proposal distribution for the next state dependent on the current state, $q(\mathbf{z}|\mathbf{z}^{(t)})$. Thus, the sequence of samples $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(T)}$ is a Markov chain [Hastings, 1970; Denison et al., 2002; Bishop, 2006].

The Metropolis-Hastings algorithm is an iterative procedure where a candidate state $\mathbf{z}^*$ is drawn from a proposal distribution $q(\mathbf{z}|\mathbf{z}^{(t)})$. In this thesis, the proposal distribution is simply a Gaussian centred on the current state. This candidate state is then assigned a probability

$$A(\mathbf{z}^*, \mathbf{z}^{(t)}) = \min\left(1, \frac{p(\mathbf{z}^*)}{p\mathbf{z}^{(t)}}\right). \tag{2.19}$$

We then choose a uniform random number u over the unit interval $(0, 1)$ and accept this new state if $A(\mathbf{z}^*, \mathbf{z}^{(t)}) > u$. The use of this acceptance criterion allows the algorithm to sample the full distribution space as $t \to \infty$. If the candidate state is accepted, then $\mathbf{z}^{(t+1)} = \mathbf{z}^*$, otherwise $\mathbf{z}^{(t+1)} = \mathbf{z}^{(t)}$ [Hastings, 1970; Bishop, 2006].

The choice of the variance of the proposal distribution is of some importance; if the variance is too high then many candidate states are likely to be accepted but it will take a significant number of iterations for the algorithm to converge to the correct distribution. Also, if the variance is too small then the algorithm may not be able to explore the full space as it may get stuck in a local minima.

2.4 Time Series Analysis Methods

There are a number of methods that have been used to analyse the RR-interval time series. It would be impossible to analyse each method in depth and so we provide a brief overview of the methods here.

2.4.1 Standard HRV Measures

There are a variety of standard techniques used in HRV analysis which utilise both the frequency and time domain. Although they are not complex enough to model the full dynamics of HRV, they can often be used for comparative purposes when developing new methods. Detailed descriptions can be found in many sources [Task force of the European Society of Cardiology and the Northern American Society of Pacing and Electrophysiology, 1996; Teich et al., 2001].

Time Domain

There are many commonly used time-domain measures in HRV analysis which use NN-intervals. They are:

- **NN50** Number of pairs of adjacent NN-intervals differing by more than 50 milliseconds (ms).
- **pNN50** The proportion of consecutive NN-intervals with differences that exceed 50 ms.
- **SDNN** The standard deviation of all NN-intervals.
- **SDNN** ($\sigma_{\mathbf{int}}$) The mean of the standard deviations of the NN-interval set for all five minute segments.
- **SDANN** The standard deviation of the averages of NN-interval over five minute segments
- **SDSD** Standard deviation of the differences between adjacent NN-intervals
- **RMSSD** The square root of the mean squared differences between adjacent NN-intervals

These measures are relatively simple but they have been widely applied in HRV analysis with some good results and have been shown to be clinically viable.

Frequency Domain

Transforming to the frequency domain allows analysis of how the power distributes as a function of the frequency [Task force of the European Society of Cardiology and the Northern American Society of Pacing and Electrophysiology, 1996]. The usual method of transformation from the time domain to the frequency domain is by non-parametric techniques such as the fast Fourier transform (FFT). Parametric methods such as autoregressive modelling have also been used.

The analysis of the power spectral density (PSD) of the data can take place over the short term (normally 5 minute intervals) or long term recordings (the entire period of the recording, typically 24 hours) which give quite different PSDs and therefore different measures. Long term recordings are difficult to interpret as they suffer from nonstationarity [Furlan et al., 1990]. Moreover, features based on long term data are highly correlated with that of the time domain due to mathematical and physiological relationships. As such, time domain analysis is often preferred as it is easier to perform [Task force of the European Society of Cardiology and the Northern American Society of Pacing and Electrophysiology, 1996].

The standard clinical frequency bands are Ultra Low Frequency (ULF), Very Low Frequency (VLF), Low Frequency (LF) and High Frequency (HF). For short term recordings, ULF is not used as a measure as is not meaningful in such a short time window. Indeed, the physiological explanation of VLF (the 0.003-0.04 Hz range) assessed from short term recordings is not well defined and it is normally not included in interpreting the data. The LF and HF boundaries are given as 0.04-0.15 Hz and 0.15-0.4 Hz respectively and it is these that are the basis for most of the spectral analysis. LF and HF may also be measured in normalised units (n.u.) which is the power of each frequency band divided by the total power minus the VLF component. Also the LF/HF ratio is often used as it is often more useful than the individual frequency information on its own.

The analysis of the frequency domain, particularly in short term recordings, is widely used and is regarded as being better understood than the equivalent time domain analysis.

2.4.2 Standard P-wave Measures

There are two simple measures which have been applied to P-waves for investigative purposes,

1. P-wave duration,

2. P-wave dispersion.

The P-wave duration is normally reported as the average P-wave duration plus or minus the standard deviation. The P-wave dispersion is the longest P-wave duration minus the shortest. Both have been used in a number of tests [Steinberg et al., 1993; Wong et al., 2004] which show that these statistics can be of use in determining the presence of certain cardiac conditions.

2.4.3 Information Theoretic Measures

There are a variety of information theoretic measures that are in use for cardiac time series analysis, most of which are termed as 'entropy' which is a quantitative measurement of disorder in a system. The measures that we use in this thesis are classed as entropy measures, with one notable exception being the Fisher information. We also look at measures that do not feature again in this thesis for comparative purposes.

The calculation of information theoretic measures in HRV analysis can potentially overcome some of the pitfalls of other techniques such as the limitations of linear statistical measures and the complexity of non-linear techniques [Wessel et al., 2000]. A variety of

methods have been presented [Fusheng et al., 2001; Cugini et al., 1999; Quiroga et al., 2000] which show that this field merits further study.

Information Entropy

Information entropy is a method which directly estimates the amount of unpredictable information in a discrete system [Shannon, 1948]. Conversely, it can also be regarded as a measurement of information *contained* in a system. For a data series, $\mathbf{x}$, with N possible outcomes, the information entropy is

$$H_p(\mathbf{x}) = -\sum_{i=1}^{N} p(x_i) \log p(x_i), \tag{2.20}$$

where p is the probability of value x_i being observed. The logarithm in Shannon's original paper is base 2 as it refers to binary bits but we use natural logarithms of base e in this equation and the rest of the thesis.

It has been used in some HRV studies [Cugini et al., 2001, 1999] but it does not hold any information on the sequencing of the values, only on their level of disorder.

Approximate and Sample Entropy

These are similar techniques based on the concept of entropy rate which is used as a method of determining the chaotic nature of a system. Approximate Entropy (ApEn) is a promising characterising parameter which measures the irregularity or complexity of a signal [Fusheng et al., 2001] whilst incorporating information on the sequential properties of the data. Sample entropy (SampEn) is an improvement of approximate entropy which addresses some issues inherent in the method. For an extensive overview of these techniques refer to Chapter 4.

Fisher Information

Fisher information is another information theoretic measure that predates the information entropy by about 25 years but never achieved its popularity [Frieden, 2004].

Fisher information is defined as

$$I = \int \frac{(p'(x))^2}{p(x)}\, dx, \tag{2.21}$$

with the discrete form

$$I = \sum_{i=1}^{N} \frac{[p(x_{i+1}) - p(x_i)]^2}{p(x_i)}. \tag{2.22}$$

As can be seen from Equation 2.21, this statistic is a function involving the first derivative which means that it incorporates information on the local gradient of the function. Hence the Fisher information is sensitive to the local rearrangement of points, a property known as *locality*. This contrasts with the information entropy given in Equation 2.20 which is a *global* measure [Frieden, 2004], i.e. the ordering of the values in the series will have no effect on the result of the statistic. This is used as a useful indicator of the effect that the local arrangement of points can have on the discriminative potential of a series. This is of importance as the analysis done in this thesis is based on the hypothesis that the values of consecutive points will provide useful discriminative features. Hence a simple statistic that incorporates sequential information is of great use as a benchmark for comparison of other sequential measures.

2.5 Nonlinear Dynamics

The problem of identifying nonlinearities is non-trivial since both chaotic and random processes have similar broadband spectra. We shall overview the main approaches to determining if a series displays chaotic behaviour. These techniques can also be regarded as feature extraction methods. Firstly, we show how a time series may be represented in a more natural fashion using Takens' embedding theorem [Takens, 1981].

2.5.1 Phase-Space of a Dynamical System

Time series such as those derived from cardiac data can be regarded as a projection of the d-dimensional state of a system to a one-dimensional space. Reconstruction of the underlying function space is one method of exploring the dynamical nature of the system. Takens' embedding theorem asserts that if a time series is one component of an attractor that can be represented by a smooth d-dimensional manifold then the topographical properties of the attractor are equivalent to the topological properties of the reconstruction [Henry et al., 2001].

The method for reconstruction of the time series in phase space is the time delay method. The process involves calculating a set of delay vectors from the observation space of the form

$$\mathbf{y}_n = \big(x(n), x(n+\tau), x(n+2\tau), \ldots, x(n+(m-1)\tau)\big)^T, \tag{2.23}$$

where n is the initial time, τ is known as the lag (delay time) and m is the embedding dimension. A condition of Takens' embedding theorem is $m \geq 2d+1$ for complete reconstruction

of the attractor. Because of this, and to avoid the curse of dimensionality, accurate calculation of the lag and the embedding dimension are of paramount importance in time delay embedding.

2.5.2 Delay Time

Any value of the delay time or lag is theoretically viable but the shape of the embedded time series will change depending on the value of τ. Therefore it is prudent to choose a value for τ that will separate the data as much as possible [Packard et al., 1980]. Therefore a lag time that yields a reasonably small amount of correlation between points in the reconstructed series is preferential.

A widely used method to estimate the lag is to take the first zero (or close to zero) in the autocorrelation function of the signal $f(t)$ with itself

$$R_{ff}(\tau) = \int_{-\infty}^{\infty} f(t)f(t-\tau)\mathrm{d}t. \tag{2.24}$$

The value of τ that this corresponds to is then used as the lag. As the autocorrelation is a linear function, this will only show where linear correlations in the time series are negligible [Henry et al., 2001; Teich et al., 2001].

Another well used method in determining an appropriate value for the lag is by taking the first minimum of the mutual information [Fraser and Swinney, 1986] which is a measure of how much information about a time series point can be predicted given full information about another [Henry et al., 2001].

There are also composite methods which predict the lag and the embedding dimension at the same time as presented in [Gautama et al., 2003] which we shall return to later.

2.5.3 Embedding Dimension

There are many methods for calculating the value for the embedding dimension which have been implemented. As knowledge of these techniques is only needed for contextual understanding, we supply only a brief overview of the false nearest neighbour, singular spectrum analysis and differential entropy methods.

(i) False Nearest Neighbour

If the embedding dimension is too low, two points may be close when projected onto a low dimensional embedding space when in the correct attractor dimension they are quite far apart

[Davey et al., 2001; Signorini et al., 2001]. False nearest neighbour (FNN) [Kennel et al., 1992] exploits this property by determining the embedding dimension using an incremental search starting at $m = 1$.

For each of the set of time-lagged vectors $\mathbf{y}_m$, the algorithm determines the nearest neighbour $\tilde{\mathbf{y}}_m$ and calculates the distance between them. The value of m is then incremented by one and the distance between $\mathbf{y}_{m+1}$ and its nearest neighbour $\tilde{\mathbf{y}}_{m+1}$ is calculated in the higher dimensional space. The relative additional separation is then calculated

$$\left| \frac{d(\mathbf{y}_m, \tilde{\mathbf{y}}_m) - d(\mathbf{y}_{m+1}, \tilde{\mathbf{y}}_{m+1})}{d(\mathbf{y}_m, \tilde{\mathbf{y}}_m)} \right|. \tag{2.25}$$

If this number is greater than an absolute value R_{TOL}, then $\mathbf{y}_m$ and $\tilde{\mathbf{y}}_m$ are classed as false nearest neighbours. A suitable value of $R_{TOL} \geq 10$ was empirically shown to lead to the clear identification of the false neighbours [Kennel et al., 1992]. The percentage of false nearest neighbours is then calculated for the dimension m. The correct dimension is determined when the number of false nearest neighbours is low and further increase in m does not lower it significantly further.

(ii) Singular Spectrum Analysis

The first step of this method is to create a matrix out of the lagged delay vectors with a temporary value for the delay that is large enough so that there is redundancy in the embedding results [Broomhead and King, 1986; Roberts et al., 1998]. The matrix is of the form

$$\mathbf{Y} = \left(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{N-(n-1)}\right)^T, \tag{2.26}$$

where $\mathbf{y}$ is defined in Equation 2.23. This matrix is then decomposed into three matrices using singular value decomposition thus:

$$\mathbf{Y} = \mathbf{USV}^T. \tag{2.27}$$

The redundancy in the embedding manifests itself in rank deficiency in the matrix $\mathbf{S}$ which is a diagonal matrix with the elements $s_{ii} = \sigma_i$ where σ_i^2 are the eigenvalues of $\mathbf{YY}^T$. Rank deficiency can be determined where the values of s_{ii} become 0. In the presence of noise, as often found in biomedical systems, the magnitude of the singular values will 'level out' at a noise-floor. The first singular value s_{ii} on this noise floor can be used for the value of the embedding dimension $m = i$. The rank deficiency can also be investigated in other ways (see [Ko et al., 1999]).

(iii) Differential Entropy Method

The differential entropy based method [Gautama et al., 2003] aims to calculate the lag and the embedding dimension at the same time. The method produces favourable results using the Hénon map [Henon, 1976] with variable time delay d when compared to using the combination of the mutual information method and the false nearest neighbour technique.

The method proposed employs the Kozachenko-Leonenko estimate [Kozachenko and Leonenko, 1987] of the differential entropy

$$H(\mathbf{x}) = \sum_{j=1}^{N} \ln(Np_j) + \ln 2 + C_E. \tag{2.28}$$

N is the number of samples in the data set, p_j is the Euclidean distance of the j^{th} delay vector to its nearest neighbour and $C_E(\approx 0.5772)$ is the Euler constant. As further discussion of the procedure involves variable embedding dimension and lag, the function of the entropy, embedding dimension and lag is henceforth denoted $H(\mathbf{x}, m, \tau)$.

As the Kozachenko-Leonenko estimate is not robust to dimensionality, $H(\mathbf{x}, m, \tau)$ is standardised by dividing the entropy of the original time series by the average of a set of surrogates $H(\mathbf{x}_{s,i}, m, \tau)$ thus

$$I(m, t) = \frac{H(\mathbf{x}, m, \tau)}{\mathrm{E}[H(\mathbf{x}_{s,i}, m, \tau)]}. \tag{2.29}$$

Also, to penalise for higher embedding dimensions the minimum description length (MDL) method is superimposed, yielding the entropy ratio

$$R_{ent}(m, \tau) = I(m, \tau) + \frac{m \ln N}{N}, \tag{2.30}$$

which is the statistic that is tested for a range of values for m and τ. The minimum of the function is the correct value for m and τ. Although it has been found to be of use in [Gautama et al., 2003], preliminary testing on heart data led to inconsistent results.

2.5.4 Detecting Deterministic Behaviour

Correlation Dimension Estimation

The correlation dimension is a measure of the spatial complexity of a system [Çelebi et al., 2001]. The approach uses the principle that while a chaotic time series will have a finite dimensional attractor, a series generated by a stochastic generator will have an infinite dimensional attractor [Maier et al., 2001b].

The estimation of the dimensionality of the attractor is therefore a seemingly simple test for determinism. This can be done by calculating the *correlation dimension* d_c, traditionally

by using the algorithm introduced by Grassberger and Procaccia [Grassberger and Procaccia, 1983a,b]. A modified version has recently been proposed by Judd [Judd, 1992] which yields better results, especially in high dimensions [Galka et al., 1998; Small et al., 2000].

Judd proposed that the correlation dimension be estimated by

$$d_c = \lim_{\epsilon \to 0} \lim_{N \to \infty} \frac{\log\left(\frac{P(\epsilon))}{p(\epsilon)}\right)}{\log \epsilon}, \tag{2.31}$$

which differs from the Grassberger-Proccacia method by the inclusion of a polynomial correction term $p(\epsilon)$ [Small et al., 2000] which is an approximation of the distribution of interpoint distances and is of degree equal to the dimension of the attractor. $P(\epsilon)$ is known as the correlation function and is given by

$$P(\epsilon) = \lim_{N \to \infty} \binom{N - W}{2}^{-1} \sum_{|i-j|>W} H(\epsilon - \|x_i - x_j\|), \tag{2.32}$$

where W is equivalent to the lag τ as it is a windowing constant to prevent close temporal correlations in the series. $H(x)$ is the Heaviside function which is zero for a negative argument and one otherwise. Therefore its sum gives the number of pairs of points whose distance is less than ϵ and $P(\epsilon)$ reflects the probability that the distance between two randomly chosen points is closer than ϵ [Guerrero and Smith, 2003]. In practice, d_c is estimated by assigning an arbitrary value to ϵ or a range of values for ϵ.

Lyapunov Exponent

The Lyapunov exponent method is considered to be a quantitative test of the sensitivity of the system to initial conditions [Beckers et al., 2003] as they measure the rate of convergence or divergence of the trajectories of a dynamical system in the phase space [Bračič and Stefanovska, 1998; Çelebi et al., 2001].

Lyapunov exponents are defined as the long time average exponential rates of divergence of nearby states [Owis et al., 2002]. The mathematical formulation is given as

$$\lambda(\delta \mathbf{y}_0) = \lim_{t \to \infty} \frac{1}{t} \log\left(\frac{\|\delta \mathbf{y}(t)|_{\delta \mathbf{y}_0}\|}{\|\delta \mathbf{y}_0\|}\right), \tag{2.33}$$

where $\delta \mathbf{y}$ is a small perturbation to the orbit of a dynamical system

$$\dot{\mathbf{y}}(t) = f(\mathbf{y}(t)), \qquad \mathbf{y}(0) = \mathbf{y}_0 \qquad \mathbf{y} \in \Re^d. \tag{2.34}$$

It is normally the largest Lyapunov exponent that is studied as if this is positive it is a strong indicator of chaos: the larger the exponent, the more chaotic the time series. It

is therefore conventional to order the Lyapunov exponents by their magnitude. Calculation of the largest Lyapunov exponent is also done by algorithms such as [Rosenstein et al., 1993]. The analytic calculation of Lyapunov exponents is non-trivial and is only applicable to systems with known differential equations [Ott, 1993]. More often, the Lyapunov exponents are calculated using suitable algorithms such as the one developed by [Wolf et al., 1985] which is used in this thesis. However, Wolf's algorithm still relies on knowledge of the equations (and the Jacobian must be known). For raw time series data where the equations governing the system are unknown, more advanced algorithms must be used such as the one in [Brown et al., 1991].

The Lyapunov exponents are linked to other chaotic measures. For instance, Pesin's formula [Pesin, 1977] states

$$h = \sum_i \text{ positive } \lambda_i, \tag{2.35}$$

where h is the entropy rate. Also, the Kaplan-Yorke conjecture [Kaplan and Yorke, 1979] states that

$$d_I = k + \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{|\lambda_{k+1}|}, \tag{2.36}$$

where d_I is the information dimension, and k is the maximum value of j such that $\lambda_1 + \lambda_2 + \ldots + \lambda_j > 0$. It is also known that d_I provides an upper bound for the correlation dimension d_c.

Surrogate Data

The principle of surrogate data analysis involves methods that work under the assumption that the data comes from a particular class or system and then tests that assumption by generating *surrogate data* from this system [Small and Judd, 1998; Schreiber and Schmitz, 2000]. The surrogate data can be generated in a variety of ways which destroy some aspect of the underlying structure of the data. e.g. randomly sampling from the same distribution as the data to form a new, uncorrelated, time series. A number of surrogates are normally generated to increase confidence in the results.

Statistics are then generated from the surrogates and the series being tested. These are then compared, normally using conventional hypothesis testing, and the statistical confidence that the original data set does or does not belong to the same system as the surrogate is calculated. The hypothesis testing is normally hierarchical; the simplest explanation is

normally tested first due to Occam's razor (i.e. test for linearity) and then further, more stringent tests can take place.

There are many methods that use surrogate data for comparative purposes, for example, the differential entropy based method discussed earlier and the delay vector variance method [Gautama et al., 2004] which compares sets of pairs of delay vectors whose distance is less than some span r_d with sets derived from surrogate time series.

Surrogate data is a widely used and important method of establishing determinism within a system. This is because it is a powerful approach as many hypotheses can be tested and the surrogate data itself is easy to generate. However, the whole field of hypothesis testing relies on arbitrary confidence levels which, although it can be related to the number of surrogates tested, is not a principled approach and cannot be considered robust.

For a review of the methods used to create the surrogate series used in this thesis, refer to Appendix B.

Chapter 3

Investigating the P-wave

The motivation for investigating the P-wave is due to paroxysmal atrial fibrillation being the primary cardiac condition we are aiming to diagnose. As such, we are investigating the hypothesis that as paroxysmal atrial fibrillation is a condition affecting the atria, then any information pertaining to the onset of an AF episode may be extracted from the corresponding atrial ECG component; the P-wave.

The ECG recordings we shall investigate are taken on a Holter monitor over a long duration and therefore contain a large number of beats. Therefore, isolating the P-waves by hand is not viable for most applications.

Due to its low amplitude compared to the other ECG complexes, the P-wave is usually the most difficult to extract reliably. Indeed, the task of automatically extracting the P-wave from the ECG is non-trivial in its own right and numerous studies have been devoted to it [Anant et al., 2000; Hughes et al., 2003; Lepage et al., 2001; Rieta et al., 2003]. In Section 3.1 we shall look at some of the previous methods that have been employed to extract the P-wave. Section 3.2 outlines the method that we have used in this thesis including a pseudo-code algorithm and a discussion of its effectiveness.

Once the P-wave has been extracted, we can begin investigating the P-wave data with a view to identifying reliable discriminative features in earnest. Section 3.3 details the preliminary investigation of the P-wave data with a discussion of the results obtained.

3.1 Previous Research

Much of the previous research has involved the use of wavelet transforms, both as a pre-processing step and as an actual analytic tool. As such, this section is further divided into

methods using wavelet transforms and heuristic methods. However, as we do not use wavelet transforms, the discussion of wavelets is limited to a brief review of previous methods and results.

3.1.1 Wavelet Transforms

Of the papers that utilise wavelet transforms, there are those that extract the P-wave as part of a larger ECG processing task and those that are solely aimed at P-wave extraction.

One example of the former is to be found in [Lepage et al., 2001] where wavelets were used to isolate the P-wave in stationary ECG recordings by combining the analysis with a hidden Markov model (HMM). There were three steps to the analysis:

- apply a Haar transform with four levels of resolution to the ECG signal;
- locate the QRS-complexes by thresholding the wavelet coefficients;
- segment the signal by applying an HMM to each resolution level.

The results for the P-wave extraction of this automated method were compared to the results obtained by the average of two cardiologists examining the ECG by hand. This process achieved 77% accuracy which is not really suitable for robust P-wave analysis.

HMMs were used again, both in conjunction with multi-resolution wavelet analysis and without in [Hughes et al., 2003]. Again, the wave onset and offset were classified by experts and then a HMM was trained on the data in a supervised fashion. The probability densities for the start of each P-wave were estimated using Gaussian mixture models; although results for the other complexes were reasonable, the method only correctly extracted the P-waves 5.5% of the time. This was then compared to an HMM trained using data that was preprocessed with an wavelet transform in a similar fashion to [Lepage et al., 2001]. This yielded much better P-wave segmentation (74.2%). Despite this pronounced improvement, this accuracy is still unsuitable for robust P-wave analysis.

Another study that compared the benefits of preprocessing using wavelets to using the raw data used a neural network in place of the HMM [Anant et al., 2000]. The aim in this case was to detect the highest point on the P-wave. A wavelet decomposition of the ECG data was obtained and five different decompositions at different scales performed; the one which had the most information about the P-wave was retained. A neural network (multi-layer perceptron) was trained on the processed data and compared with a neural network that was trained on the raw data. The preprocessed neural network detected the peak 70%

of the time compared to 50% of the time for the raw data. Again, this shows the potential benefits of wavelet decomposition but it currently remains unsuitable for P-wave analysis.

3.1.2 Heuristic Methods

Heuristic approaches are often used to process ECG data, and often accurate detection of the R-point is a prerequisite for the detection of other ECG features. This is the most prominent feature of the ECG due to its size and shape. It is also essential to determine the baseline of a signal (usually taken as a line between the points just before the onset of the Q-wave of two consecutive beats). This baseline can then be subtracted from the signal to give a level representation of the ECG. These steps are important as they facilitates further feature extraction from the ECG.

These approaches adopt the following steps:

Filtering the ECG signal

As mentioned before, the signal can be susceptible to alternating current interference or low signal/noise ratio due to a poor electrode contact. This can be counteracted to a degree by filtering the data using finite impulse response (FIR) filters and infinite impulse response (IIR) filters. These can be targeted to filter certain frequencies and so, for example, could be targeted to remove alternating current noise specifically.

R-Point Detection

Reliable detection of a reference point is a pre-requisite for further ECG structure analysis. Most ECG processing packages work by detecting the QRS-complex, normally the R-wave, as that is of the highest amplitude and is the most distinctive structure in an ECG signal. This can be done in a variety of ways, such as analysis of the signal in the frequency domain or an amplitude triggering approach.

Q-Point and Isoelectric Point Detection

Having detected the R-point, the next step is to detect the Q-point and subsequently the isoelectric point (the point on the ECG baseline immediately preceding the Q-wave where there is no electrical potential). If a Q-point does not exist then the isoelectric point still needs to be identified. This can be achieved by using a simple gradient based approach to detect the turning point (at the Q-point) and then detect where the gradient levels out.

Baseline Correction

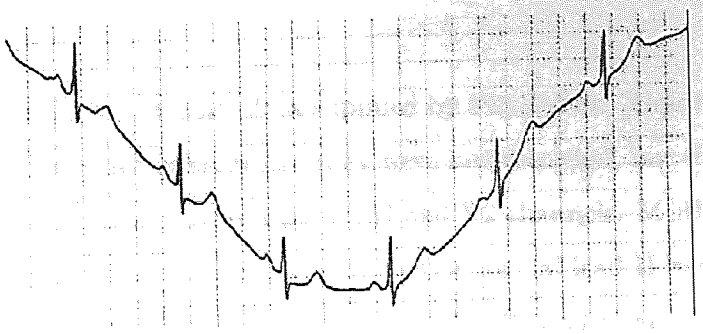


Figure 3.1: An ECG output displaying baseline wandering.

Figure adapted from `library.med.utah.edu`

ECG recordings are prone to what is known as *baseline wandering*. This is when the baseline is not flat and can cause problems with automatic complex detection and importantly, if the baseline is not corrected properly, any comparison of extracted complexes will be largely meaningless due to the fact that they will be of inconsistent orientations. There are several methods that have been implemented to correct the baseline, mainly by subtracting a baseline approximation created using cubic splines [Meyer and Keiser, 1977] or wavelet approximations [Cuesta-Frau et al., 2001]. The cubic spline method uses the isoelectric point as the knots but the wavelet analysis can be carried out without the previous steps.

P-Wave Detection and Extraction

The P-wave is the lowest amplitude normal complex in the ECG waveform. This means that in some situations it is difficult to determine the exact beginning and end points although this can be made considerably easier if the earlier stages are carried out to a high accuracy. Again, there are many methods that have been applied to this such as signal averaging [Carlson et al., 2005; Langley et al., 2001], Markov models [Lepage et al., 2001; Hughes et al., 2003] and wavelet analysis [Anant et al., 2000].

3.2 Method

3.2.1 Data

The data that we used is two channel data supplied by PhysioNet[1] [Goldberger et al., 2000]. The identities of the leads are unknown as they were not recorded, but it is known that the usual practice for their data is to record the MLII and V2 channels. MLII is an abbreviation of modified lead II as it records the similar data as the normal lead II but the electrodes are placed near the right shoulder and above the left hip [Goldberger et al., 2000] so as to be less intrusive to the patient. The other electrode is in position V2 (see Section 2.1.3). However, not all the recordings in the data set were done at PhysioNet and they could not guarantee that all the data was recorded on these two channels. This means that our algorithms must be potentially able to find the complexes on any one of the twelve leads and extract the P-wave effectively. We only used a single channel of the data set corresponding to MLII wherever possible.

3.2.2 Cardionetics Software

We were fortunate to have a working R-point detection program supplied by Cardionetics Ltd. The Cardionetics software works in a similar way to the heuristic method in Section 3.1.2 and was arranged as a series of functions with different purposes. The Cardionetics functions were applied to the PhysioNet data with varying levels of success.

The Cardionetics filtering function worked well on the noisy data. As Figure 3.2 shows, the signal after the filtering is a lot cleaner. The amount of noise was reduced considerably; however, the data seemed unnecessarily distorted by some of the filtering operations, as seen in Figure 3.3. If the 50 and 60Hz notch filters were not used, the distortion of the data was less and the performance on noisy data was not significantly impaired.

R-point detection was done in two functions. The first function detected the R-points and worked well, approximately finding their positions reliably. It picked up the location of the QRS-complex accurately enough but did not distinguish between the R and S-points and was occasionally misled by anomalous data. However, this is not a problem as the function is only intended to provide starting points for the next function.

This function did not work well on the data as it did not differentiate between the R and the S-point and sometimes crashed with anomalous data. This is probably due to the fact that the PhysioNet data set is considerably more diverse than the data produced by the

[1] www.physionet.org

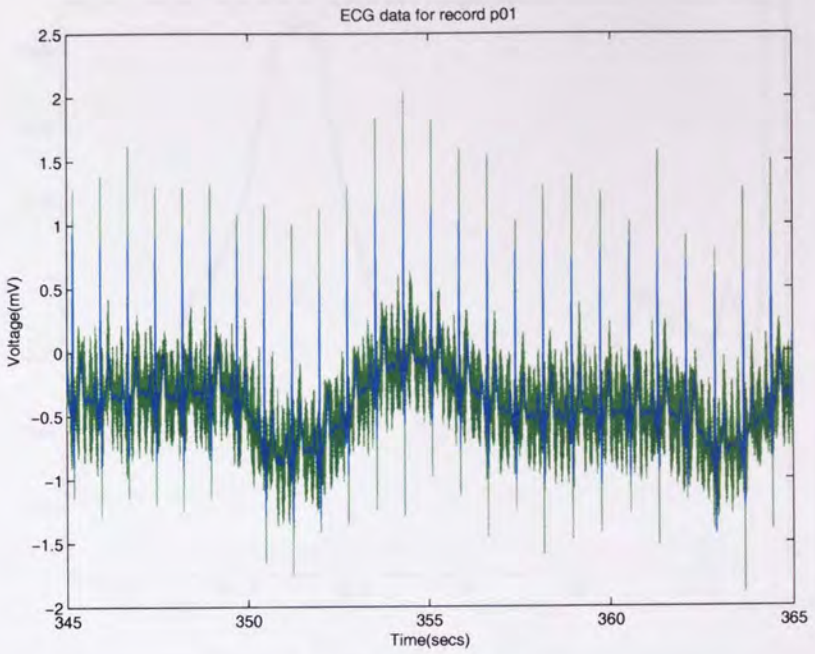


Figure 3.2: Record P01 before (green) and after (blue) filtering.

single channel Cardionetics product. This is because the recordings in the PhysioNet data do not all follow the morphology typical of the MLII lead; often the S-point is more pronounced than the R-point. As the data is squared before R-point detection, a pronounced S-point will be detected instead of the R-point.

The location of the isoelectric point also did not perform very well, in no small part due to the inaccurate distinction of the R and S-points in the previous function. Even with accurate R-point detection, the function inconsistently estimated the position of the isoelectric point. This, in fairness, could be due to the sampling rate of the Cardionetics product being different to the sampling rate of the PhysioNet data so the isoelectric point detector cannot be expected to perform optimally.

Finally, the baseline correction using a spline function worked very effectively and significantly reduced the baseline wandering, even in extreme situations as in Figure 3.4. The intermediate baseline in between the knots is not perfect but it would be difficult to get better results at this stage.

When appraising the Cardionetics software, it is important to be mindful that it was most likely never designed for an input set as diverse as the PhysioNet data. Considering this, the performance of the whole Cardionetics package provides a solid basis for further expansion and adaptation for our particular problem.

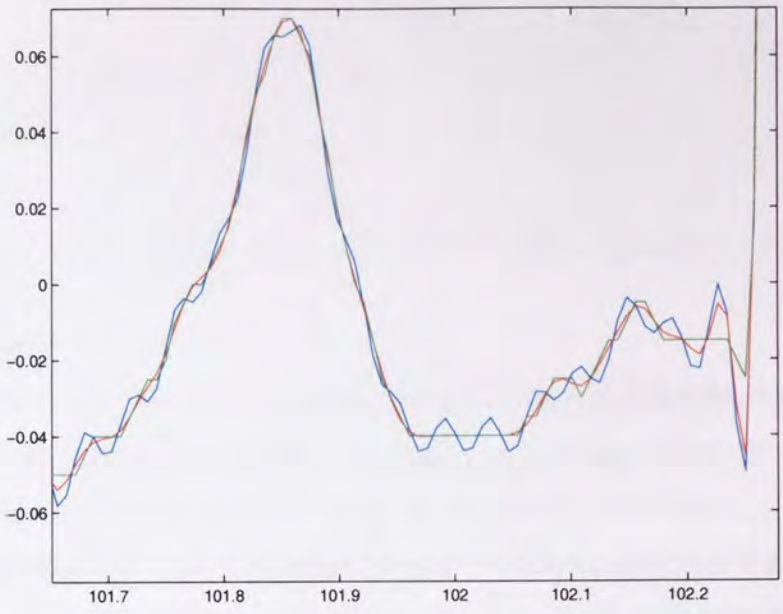


Figure 3.3: A filtered ECG output showing the data after the Cardionetics filters have been applied (blue), the data after just a 50Hz IIR filter (red) has been applied and the data after a 50Hz IIR and a morphological filter has been applied (green).

3.2.3 P-wave Detection

After the Cardionetics software showed some flaws while testing on the PhysioNet data it was apparent that we would have to modify the previous algorithms or create entirely new ones.

Our method follows the heuristic framework in Section 3.1.2 and is shown in Algorithm 1. The low-pass filters are applied in line 3 of the algorithm. We found we achieved best results using a combination of Butterworth and morphological filters (see Appendix A). The Butterworth filter was a fourth order low pass filter with a cutoff frequency of 0.4 multiplied by the Nyquist frequency of the data. This is applied first and the resulting series filtered further using the morphological filter. The morphological filter uses the method in [Seedhagi, 1998] where a flat structure element is chosen (in our case the structure element was of a length one fortieth of the sampling rate) and then the data was both closed then opened, and opened then closed. The filtered signal was the average of the morphological operations and the output of the application of the Butterworth filter. The results of the filtering can be seen in Figure 3.5. The Butterworth filter reduces the main source of noise in the channel (Figure 3.5(b)) and the morphological filter (Figure 3.5(c)) reduces the oscillations on the baseline and smoothes the turning points on the P-wave and the other ECG structures. The amplitude of the R and S points is reduced but this has no adverse effect on their detection.

```
1  begin
2      read in signal;
3      apply Butterworth filter with frequency cut-off of 0.4 multiplied by the Nyquist
       frequency (one half of the sampling rate) and morphological filter with a flat
       structure element of length 1/40^th of the series length to the signal;
4      detect high amplitude turning points on squared data as potential R-points. This
       was done using the rdetector.m function from the Cardionetics package, but it
       can be done with other R-point detection algorithms, such as the one in Pan and
       Tompkins [1985];
5      for i=1:number of potential R-points do
6          confirmed R-point vector(i)=Rverify(potential R-points,data)
7      end
8      for j=1:number of confirmed R-points do
9          iteratively search backward from the R-point(j) for a minima that corresponds
           to the Q-point;
10         store Q-point location in vector Qvec;
11     end
12     create a cubic spline using the elements of Qvec as knots;
13     subtract spline from the original signal to correct baseline;
14     for j=1:number of R-points do
15         temporarily store 300ms of data preceding the Qvec(i) in tempPvec;
16         Pwave(i)=findPwave(tempPvec);
17     end
18 end
```

Algorithm 1: Pseudo-code for processing the ECG data and extracting the P-wave.

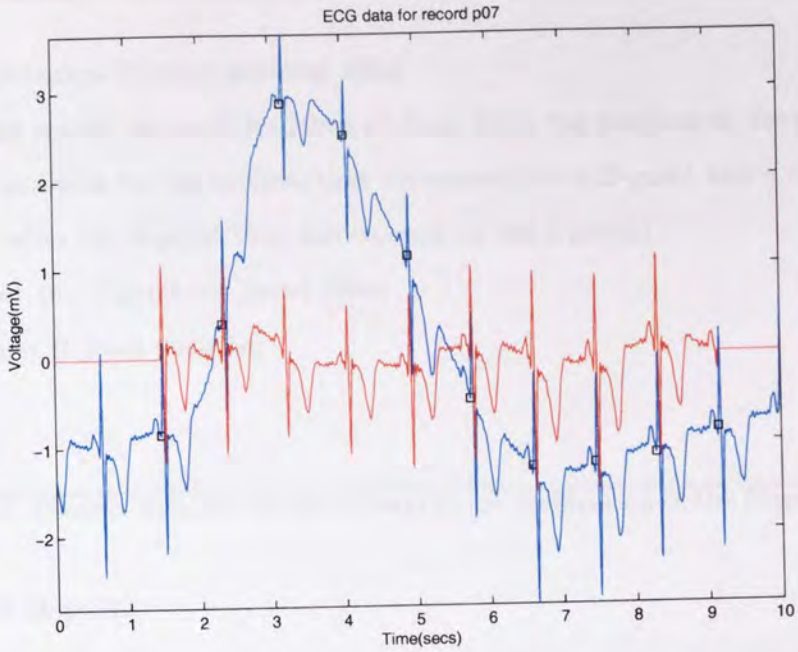


Figure 3.4: Record P07 before (blue) and after (red) baseline correction. The knots of the spline are shown by the black squares.

After filtering, we then square the data and differentiate to detect the turning points. The squared data ensures that the high amplitude R-points are easier to pick out. Actual detection of the R-points is started in line 4 and completed in line 5 where any fine tuning and verification is done. The process of correcting the baseline is carried out from line 6 to 11. Firstly, the start of the Q-wave is established for each R point in the series (lines 6-9). In line 10, these points are used as knots for a cubic spline which fits the underlying baseline drift which is corrected by simply subtracting the spline from the signal in line 11. The actual P-wave detection is carried out in lines 12-18. Firstly, the section of the signal immediately preceding the Q-wave is isolated in line 13. This section is then further baseline corrected using lines of best fit which align the P-wave to facilitate more robust and precise extraction which is carried out in line 14. The P-wave is detected in line 15 as the longest sequence where the signal leaves the baseline in this section, and it was checked to make sure the T-wave or other artifact had not been detected by mistake. In lines 16 and 17, the start and finish of the P-wave are detected as the points where the signal first deviates from, and then rejoins, the baseline. These are then used to extract the P-wave in line 18.

```
1 begin
    | input: potential R-point position, data
2   | iteratively search forward (for 50ms of data) from the position of the potential
    | R-point position for the minima that corresponds to a S-point and a corresponding
    | maxima after the S-point that corresponds to the J-point.;
3   | if S-point and J-point are found then
4   |   | return R-point position;
5   | end
6 end
```

Algorithm 2: Pseudo-code for function `Rverify` for verification of the R-point location

Detecting the R-point

The Cardionetics R-point detection consisted of two functions, one to determine the rough location of the R-points the second to pinpoint the R-point precisely. The preliminary R-point detector function performed satisfactorily in detecting the approximate locations and so was left untouched. After several attempts to modify the second R-point detector function failed to increase its robustness adequately, it was decided to start afresh.

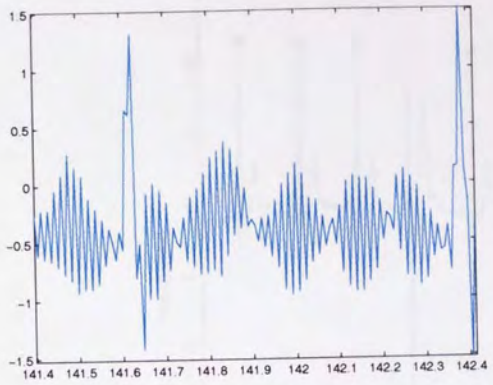
The new function would have to detect the R-point accurately every time and not the S-point. The program should detect the R-point as per the definition given in Section 2.1.3; as the point immediately preceding the downward deflection of the S-wave. This meant that the R-wave would always be a positive deflection, which was an immediate refinement of the Cardionetics software which detected the first upward or downward deflection in the window centred on the candidate R-point, which was not necessarily the R-point as given in the previous definition. It is also worthy of note that in the case of having two R-points, `r` and `r'`, the program should detect the first one, `r`.

For robustness, the function finds a suitable S-point as well as a suitable R-point. This was done by taking the highest and lowest points of a data range centred around the approximate R-point. If the lowest point preceded the highest point (i.e. the Q-point was lower than the S-point), then the S-point needed to be properly located by only searching the area after the highest point.

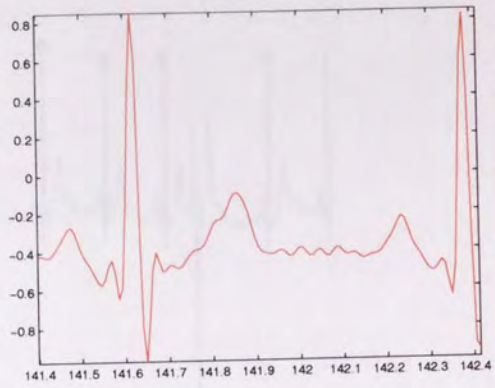
It is likely that the R-point would correspond to the highest point in this data range. However, as this is not always true, for example, if the start of the T-wave encroaches into the data range; the program would examine all the maxima, discount any that occur after

```
for k=1:2 do
    begin
        input: tempPvec
        calculate the turning points of tempPvec and calculate line of best fit of these;
        subtract the line of best fit from the data;
        use the result to update tempPvec;
    end
    set a threshold as a 20% of the largest value in tempPvec;
    identify P-wave as the section of the signal that remains above this threshold for
    the most consecutive time points;
    verify this signal section is not the T-wave by iteratively searching backward from
    the signal section maxima for first maxima that lies above the threshold (which
    corresponds to the T-wave) if this is found to be equal to the previous R-point,
    then we have located the T-wave;
    if T-wave located then
        finish=locate the first minima below the threshold after the T-wave peak where
        the signal rejoins the baseline;
        start=locate the first minima below the threshold after the T-wave peak where
        the signal rejoins the baseline;
        Twave=vector of zeros the same size as tempPvec;
        Twave(start:finish)=signal(start:finish);
        tempPvec=tempPvec-Twave;
        set a threshold as a 20% of the largest value in tempPvec;
        identify P-wave as the section of the signal that remains above this threshold
        for the most consecutive time points;
    end
    finish=locate the first minima below the threshold after the P-wave peak where the
    signal rejoins the baseline;
    start=locate the first minima below the threshold after the P-wave peak where the
    signal rejoins the baseline;
    Pwave=vector of zeros the same size as tempPvec;
    Pwave(1:finish-start)=tempPvec(start:finish);
    return Pwave;
end
```

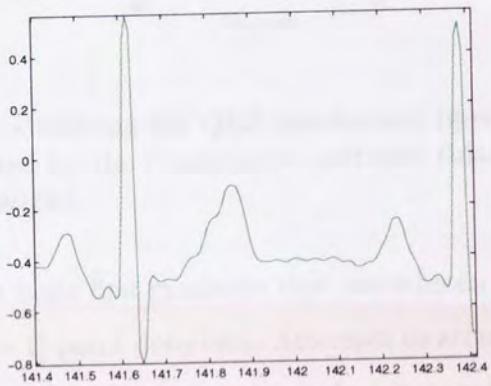
Algorithm 3: Pseudo-code for function `findPwave` for P-wave detection



(a) Unfiltered Data



(b) Butterworth Filter



(c) Butterworth and Morphological Filters

Figure 3.5: An ECG recording of a heartbeat from record P01 at the different stages of filtering.

the S-point and take the maximum of those that remain. Also, for computational efficiency, if the approximate R-point was the same as the maximum value, it was accepted that this was the accurate R-point and no further examination was carried out on that R-point.

The duration between subsequent R-points for each record was stored in a vector for future analysis.

Q-Point and Isoelectric Point Detection

A useful precursor to finding the isoelectric point is accurate location of the Q-point which is defined as "a negative downward deflection preceding an R-wave" [Julian, 1978]. In effect,

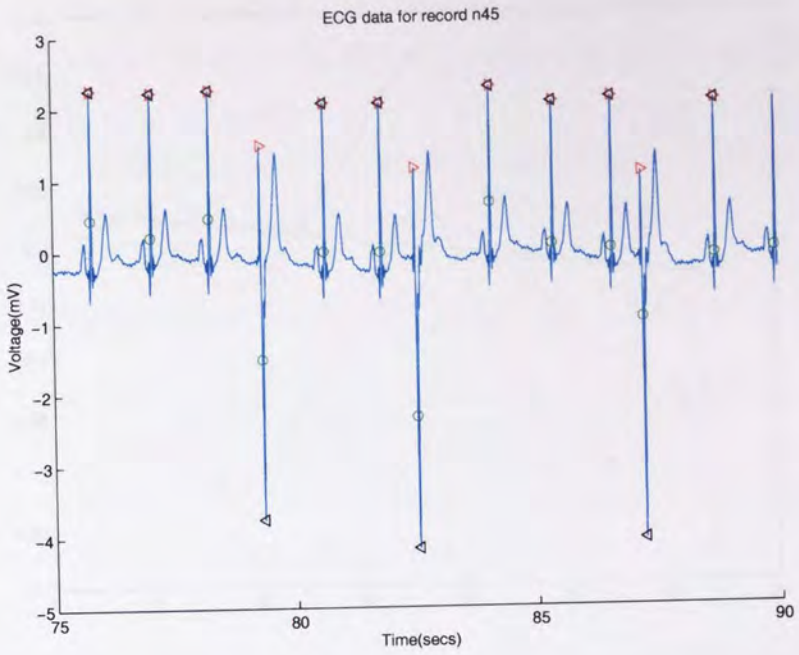


Figure 3.6: The filtered data showing the QRS annotations (green circle) supplied by PhysioNet, the R-points detected by the Cardionetics software (black triangle) and the newly developed software (red triangle).

this means that the Q-point is the first minimum that precedes an R-wave so simply detecting this is all that is required for Q-point detection. Attempts to accurately detect the isoelectric point yielded inaccurate and inconsistent results and so it was decided to use the Q-points as the knots for the correcting spline as they could be detected more consistently.

Baseline Correction

This was implemented in the same way as the Cardionetics software but with the Q-points being used as the knots instead of the isoelectric points. Correcting the data with the cubic spline aligned the Q-points correctly as well as significantly improving the baseline wandering but the baseline was still not completely flat in between knots. This problem needs to be addressed as accurate ECG segmentation, particularly the P-wave, is dependant on a consistent baseline. Although the baseline was not totally flat in most cases, the improvement is sufficient for P-wave extraction to begin.

3.2.4 P-wave Extraction

A pseudo-code representation of this function can be found in Algorithm 3. To extract the P-wave, we needed to examine the segment before the Q-point. An interval of 300ms was

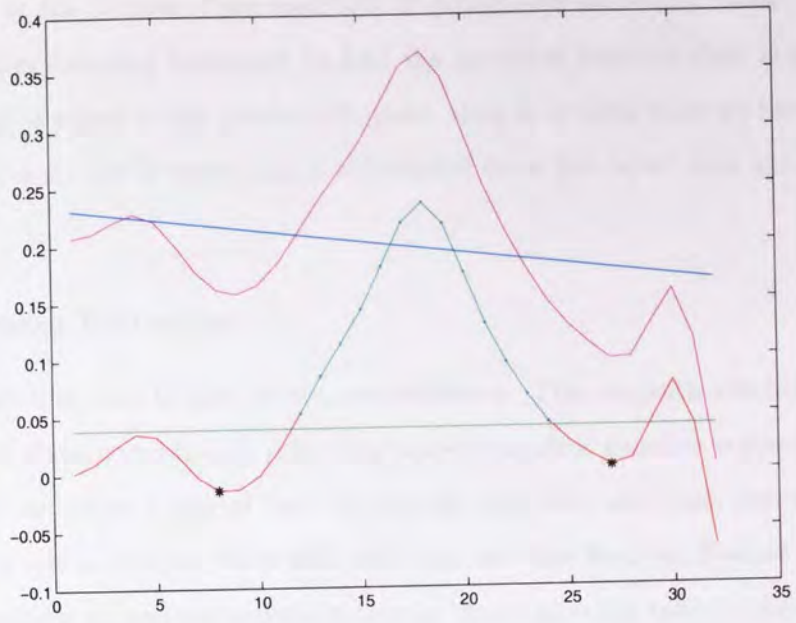


Figure 3.7: The stages of the P-wave extraction process. The original data (magenta) is baseline corrected (red), the P-wave located (cyan) and the beginning and end points found (black stars).

chosen as this is more than the maximum duration of the P-wave.

As the baseline was still fairly inconsistent, accurately extracting the P-wave at this stage would have been particularly difficult. To address this problem, we corrected the baseline further. Figure 3.7 shows how this was done. The raw data 300ms before the Q-point is shown in magenta. The turning points of the data are found (minima and maxima) and a line of best fit is calculated that fits those turning points. The line of best fit is then subtracted from the data to correct the orientation of the P-wave. This was then repeated to correct the baseline further (not shown on the figure for clarity), which gives the P-wave data shown in red and cyan. This process consistently flattened the baseline around the P-waves.

After several tests, a threshold was set at 20% of the maximum value of the corrected data as that would be able to rule out a substantial level of baseline wandering. The threshold is shown in green in Figure 3.7. Any consecutive points above this threshold were noted and the longest sequence of consecutive points was taken to be the potential P-wave which is the cyan part of the corrected data in Figure 3.7. The maximum of this was taken to be the peak of the P-wave and the closest minimum prior to the peak that was less than the threshold value was taken to be the start of the P-wave. Likewise, the minimum succeeding the peak of the P-wave fitting these criteria was take to be the end of the P-wave. The beginning and end points are shown as the black stars in Figure 3.7. The algorithm sometimes returned the

T-wave instead of the P-wave if the interbeat duration was unusually small. To counteract this the algorithm searches backward to find the previous maxima that is also above the threshold. If this is equal to the previous R-point, then it is likely that we have detected the T-wave. In this case, the T-wave data is subtracted from the input data and the process is repeated.

Results of P-wave Extraction

Figure 3.7 shows how this is done in the new software. The magenta waveform is a typical example of ECG data immediately preceding a Q-point after baseline correction. As this is not totally flat, we make a line of best fit through the data and then correct the baseline again giving the red waveform. It is with this that we then find the P-wave (highlighted in cyan) and the beginning and end points (shown in black) are then taken to be the the upward deflection preceding the P-wave and the minimum where the wave rejoins the baseline.

Figure 3.8 shows the P-waves extracted from a typical 20 second interval. Many of the P-waves are of different amplitudes and lengths but the software detects the beginning and end points with high accuracy.

Comparison of five hundred of the extracted P-waves showed that the average distance our own visual estimates of the P-wave fiducial points were away from those generated by the software was just over a 200th of a second. This seemed acceptable enough to begin the analysis in earnest.

The P-waves were then all aligned to the beginning of a vector and any space left between the end of the P-wave and the end of the vector was filled with zeros. This was done to aid comparison of the P-waves whilst keeping the vectors the same size so they could be manipulated easily.

A very small amount ($< 1\%$) of the data was so distorted that the P-wave was not distinguishable, even by visual means. In this case, the algorithm reported that no P-wave was present and the corresponding output was the same as the previous P-wave vector. This was done for future analysis so that the RR-interval data and the P-wave data was of the same length.

3.2.5 Conclusions

We have created an algorithm based on software supplied by Cardionetics to extract RR-intervals and P-waves from ECG data.

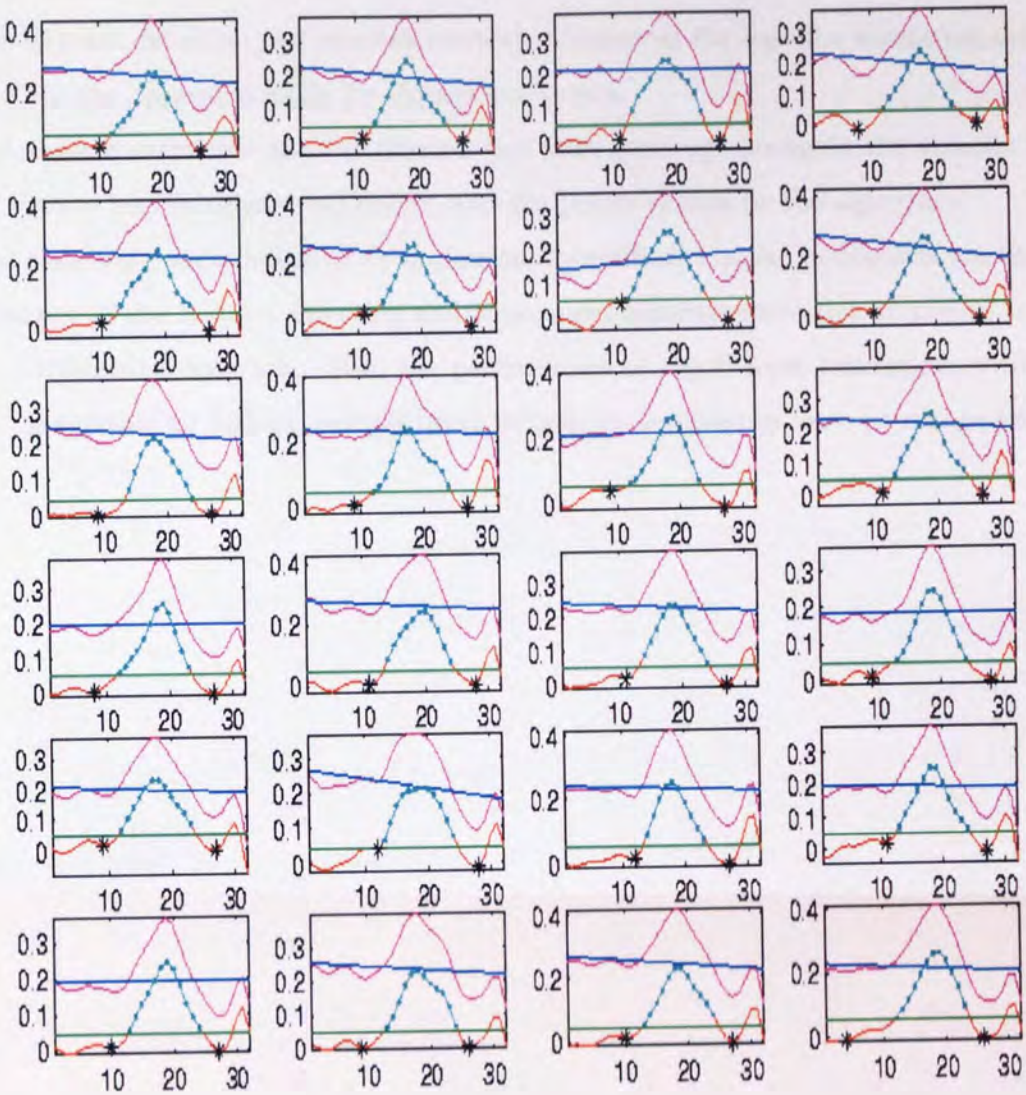


Figure 3.8: A set of P-waves extracted from a typical 20 second segment of ECG.

The filtering in the new software is similar to that in the Cardionetics version in that it uses a low-pass Butterworth filter. We have also used a morphological filter as it reduces the noise further. The results of the filtering can be seen in Figure 3.5.

The R-point detection showed an improvement on the QRS annotations supplied by PhysioNet and was more robust to anomalous beats than the Cardionetics software as seen in Figure 3.6. This is obviously of great benefit when trying to locate other structures in the ECG as robust detection of the R-points facilitates subsequent tasks. Of one thousand R-points that were visually inspected from several different records, each was correctly identified by the software. This is very important for heart rate analysis as the location of the R-points needs to be determined precisely.

The Q-point detection and baseline correction improved the baseline wandering substantially. This was deemed suitable for P-wave extraction.

The P-wave extraction itself performed well with good agreement in the visually determined P-wave beginning and end points, and the points chosen by the algorithm.

The tests were not exhaustive, or professionally verified, but the indications are that the performance of the R-point detection and subsequent baseline correction improved on that of the Cardionetics software. Also, the performance of the P-wave extraction, whilst not perfect, is suitable for further analysis using techniques designed to work on real world data.

3.3 Analysis

In this section we will look at the techniques used to extract features from the P-wave data sets generated from the raw ECG data. We shall examine the results of simple P-wave statistics on the data and investigation of the data using visualisation techniques.

For this investigation, we shall be using the paroxysmal atrial fibrillation prediction challenge data set. The classes we shall be trying to distinguish between are

1. Recordings from patients with no medical history of PAF (group N) from patients who suffer from PAF episodes (group P).
2. Of those recordings from those who suffer from PAF, which are immediately prior to a PAF episode onset (group P_c), and which are distant (group P_d).

3.3.1 P-wave Statistics

A set of P-wave statistics were generated for each class. These were:

1. P-wave duration.
2. P-wave mean duration.
3. P-wave variance.
4. P-wave skew.
5. P-wave kurtosis.

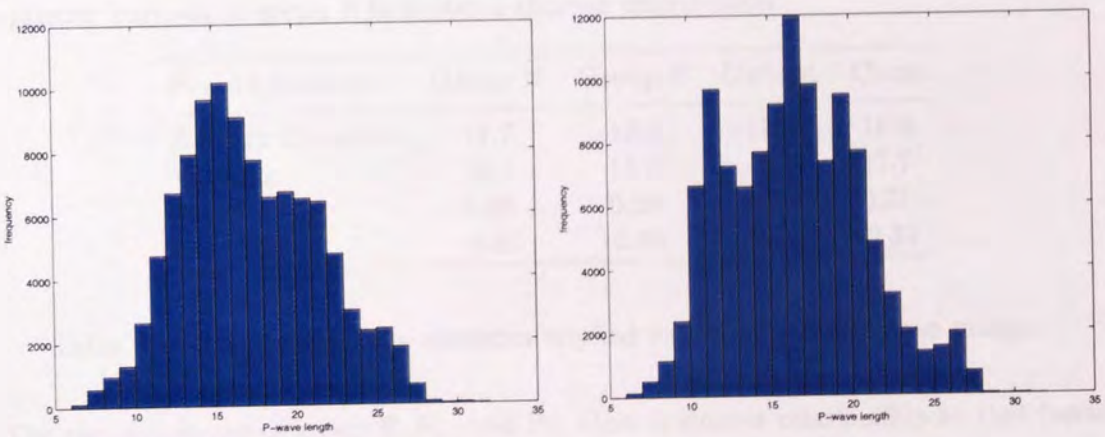


Figure 3.9: Histograms of the P-wave durations of the N (left) and P (right) groups.

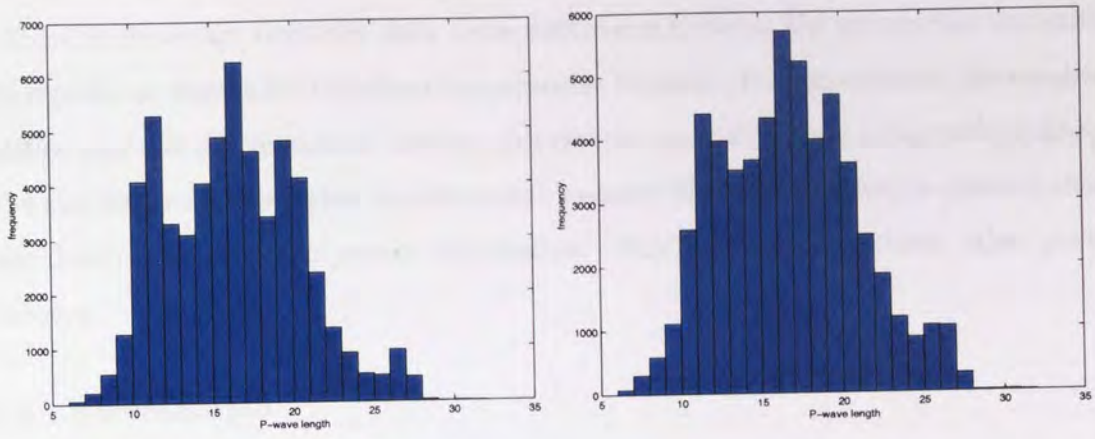


Figure 3.10: Histograms of the P-wave durations of the P_c (left) and the P_d (right) groups.

The duration histograms of the groups N and P can be seen in Figure 3.9. The distributions appear similar but the group P group has a double peak and also seems 'sharper'. The histograms of the recordings from group P_c and group P_d are shown in 3.10. The distributions of these are very similar, with both groups exhibiting the bimodal appearance of group P. These properties can be quantitatively assessed by calculation of the P-wave statistics, as shown in Table 3.1.

The statistics for groups N and P are quite similar. The average P-wave duration is slightly longer in group N. The variance is also very similar in both groups showing that the overall spread is approximately the same. The slight positive skew for both sets indicates a mild asymmetry; the distribution mode is less than the mean in both cases. This can be seen in the two histograms. The normalised kurtosis for both sets is less than zero implying that they are *platykurtic* (or sub-Gaussian) which means they have short tails and flattened tops. The greater kurtosis in group P indicates a sharper distribution.

P-wave Statistic	Group N	Group P	Distant	Close
Average Duration	17.7	17.0	17.2	16.8
Variance	18.3	18.0	18.3	17.7
Skew	0.26	0.28	0.24	0.31
Kurtosis	-0.46	-0.38	-0.41	-0.34

Table 3.1: Results of simple statistics applied to the different P-wave groups.

The two subclasses of group P, P_c and P_d, show a similar relationship to that between groups N and P. This indicates that there may not be suitable information in the basic P-wave statistics to distinguish between recordings.

These rudimentary statistics show some differences between the groups but are unlikely to be significant enough for classification purposes. However, the expansion of the number of statistics may aid differentiation between the classes especially when using such techniques as P-wave dispersion and other statistics that measure the temporal and sequential change in the data rather than the overall distribution. This led us to investigate other possible techniques.

3.3.2 Visualisation

Principal Component Analysis

The first step of the initial analysis was to apply linear feature extraction to the extracted P-wave data by computing the principal components. To do this, the data was split into a learning set and a test set. The learning set consisted of 20 records from each of the groups N and P, and the test set consisted of 30 records from each group.

To visualise, we computed the three principal components corresponding to the largest eigenvalues of the learning set. We then projected the rest of the data onto these PCs and plotted them pairwise against each other. The results are shown in 3.11.

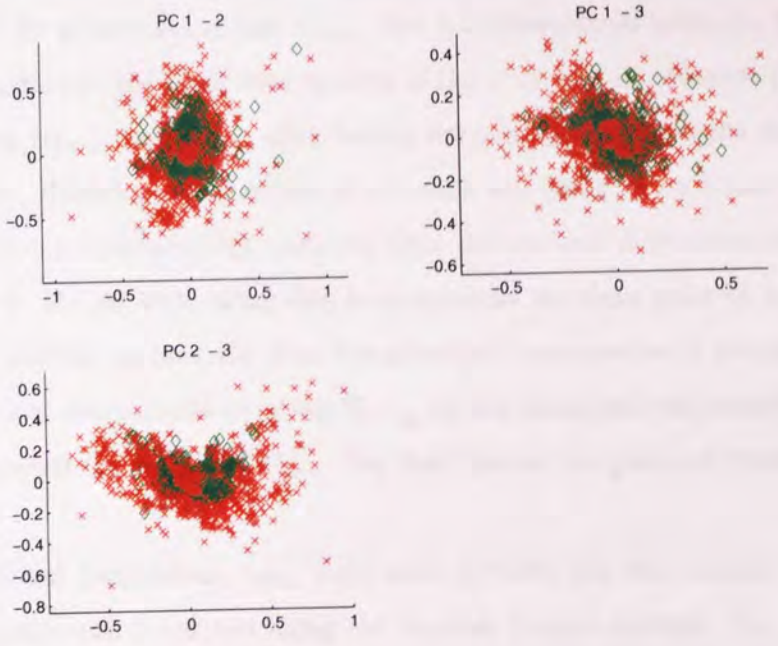


Figure 3.11: The projection of the training data onto the first three principal components defined by BIC analysis on the learning set. The green diamonds are data from group N and the red crosses represent data from group P.

The classes are not separated at all so we decided to use a non-linear visualisation ap-

proach. A problem became immediately apparent in the size of the data; too large and the computational demands become very intensive, too small and it would not be a valid representation of the data set. This implied that a method that can be trained on a small subset of the data would be needed; a criterion that the NeuroScale method fulfils. PCA would be used to preprocess the data for dimensionality reduction before using NeuroScale.

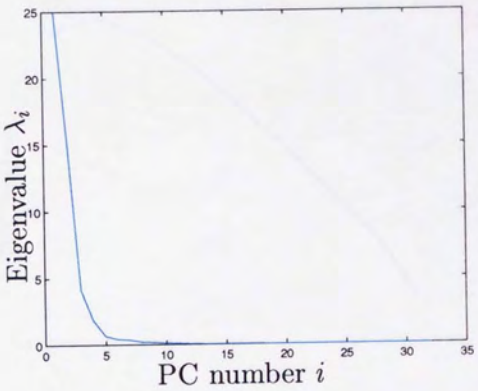
NeuroScale

Using PCA to preprocess two-class data leads to the problem of whether the principal components should come from a combined learning set using PCs from both the N and P groups or taking the PCs from the N and P groups separately and then combining them before the projection. We took 100 P-waves prior to the end of each recording from each record and divided the recordings into the learning sets and the training set as follows. The learning set L_{PCA} for the principal components was 20 records from group N ($N_{L_{PCA}}$) and 20 from group P ($P_{L_{PCA}}$). The learning set for the NeuroScale network was the next 15 records from each group ($N_{L_{NS}}$ and $P_{L_{NS}}$). This left the test set 15 records from each group (N_T and P_T).

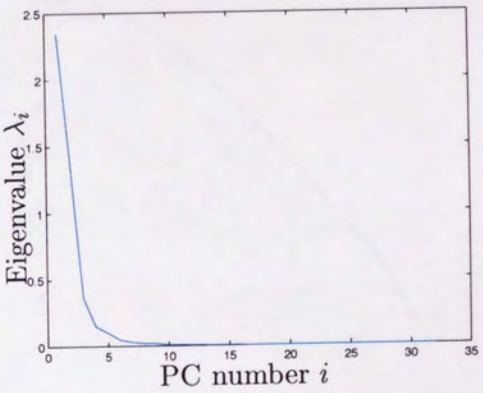
Figure 3.13 shows the results of the application of the BIC method to the PCA learning groups (on the 100 P-waves prior to the end of each recording). The analysis suggested a 5-dimensional space for groups L_{PCA} and $N_{L_{PCA}}$ but a 3-dimensional space for $P_{L_{PCA}}$. This was confirmed by looking at the eigenvalue spectra of the principal components (Figure 3.12. Therefore, combining $N_{L_{PCA}}$ and $P_{L_{PCA}}$ after taking the principal components would give an 8 dimensional space. However, some of the dimensions are likely to be noise and there is also no guarantee of non-orthogonality, meaning that the optimal dimensionality reduction may not be achieved. As we were using this to preprocess the data prior to implementing NeuroScale, it was decided to take the first five principal components of group $N_{L_{PCA}}$ and the first three principal components of group $N_{L_{PCA}}$ as the principal component basis B_{PCA} upon which we projected the $N_{L_{NS}}$ and $P_{L_{NS}}$. We shall denote the group of these projections as L_{NS}.

These 8-dimensional projections, L_{NS}, were used to train the NeuroScale RBF with 8 inputs, 25 hidden units and 2 outputs using the shadow targets method. So, in Equation 2.18, d_{ij}^* is the distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ in 8-dimensional space and d_{ij} is the distance between points $\mathbf{y}_i$ and $\mathbf{y}_j$ in the final projection space.

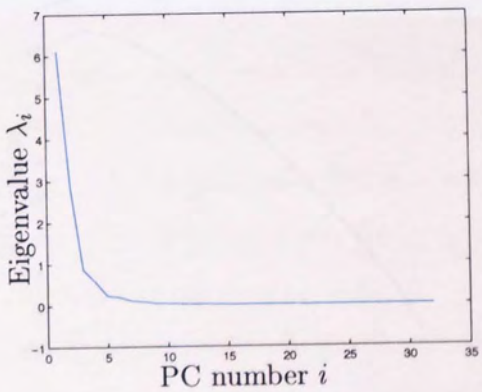
The results are shown in Figure 3.14. There are discernible patterns in the data although the interclass separation is not very good. However, this is an improvement on the principal



(a) $\mathrm{L}_{\mathrm{PCA}}$

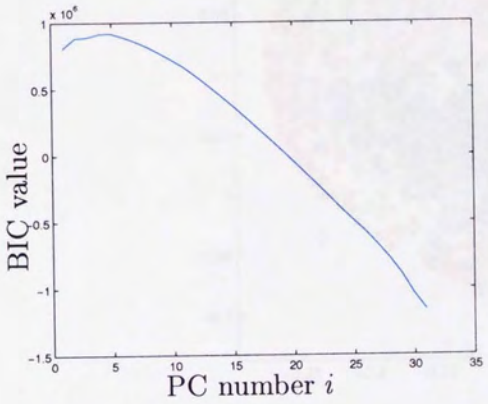


(b) $\mathrm{N}_{\mathrm{L_{PCA}}}$

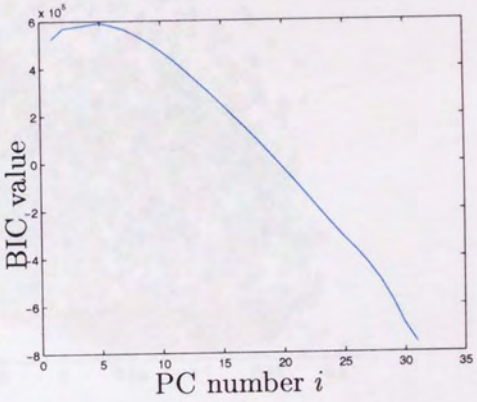


(c) $\mathrm{P}_{\mathrm{L_{PCA}}}$

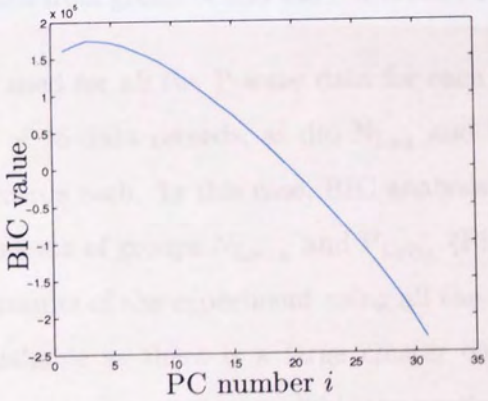
Figure 3.12: The eigenvalue spectra of the last 100 P-waves of groups $\mathrm{L}_{\mathrm{PCA}}$, $\mathrm{N}_{\mathrm{L_{PCA}}}$ and $\mathrm{P}_{\mathrm{L_{PCA}}}$.



(a) L_{PCA}



(b) $N_{L_{PCA}}$



(c) $P_{L_{PCA}}$

Figure 3.13: The results of the BIC analysis applied to the last 100 P-waves from groups L_{PCA}, $N_{L_{PCA}}$ and $P_{L_{PCA}}$.

component visualisation on its own.

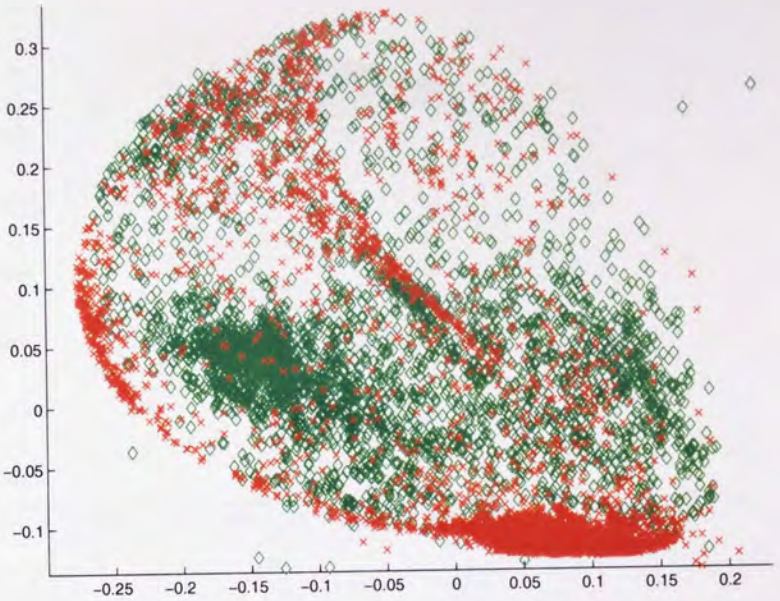


Figure 3.14: The NeuroScale visualisation using the last 100 P-waves of each record in the test set projected onto the principal components defined by BIC analysis of the learning set. The green diamonds are data from group N and the red crosses represent data from group P.

The same method was used for all the P-wave data for each record. In this case, $N_{L_{PCA}}$ and $P_{L_{PCA}}$ each consisted of 15 data records, as did $N_{L_{NS}}$ and $P_{L_{NS}}$, and the test sets, N_T and P_T, consisted of 20 records each. In this case, BIC analysis suggested that we keep the first three principal components of groups $N_{L_{PCA}}$ and $P_{L_{PCA}}$ (Figure 3.18).

Figure 3.15 shows the results of the experiment using all the data for each record. Again, the classes are indistinguishable as there is a large cluster containing both classes. The clustering is similar to the previous experiments. This can particularly be seen in Figure 3.16 which shows two close ups of the plot in Figure 3.15.

The experiment was repeated for the groups P_c and P_d. In this case, there were 5 recordings of each class in the learning sets ($P_{cL_{PCA}}$, $P_{dL_{PCA}}$,$P_{cL_{NS}}$, $P_{dL_{NS}}$) and the remaining 30 recordings were used as the test sets P_{cT} and P_{cT}. As can be seen from Figure 3.20, BIC analysis suggested 3 principal components for group $P_{cL_{PCA}}$ and four for group $P_{dL_{PCA}}$.

The results, plotted in Figure 3.21, also show very little separation between the data classes.

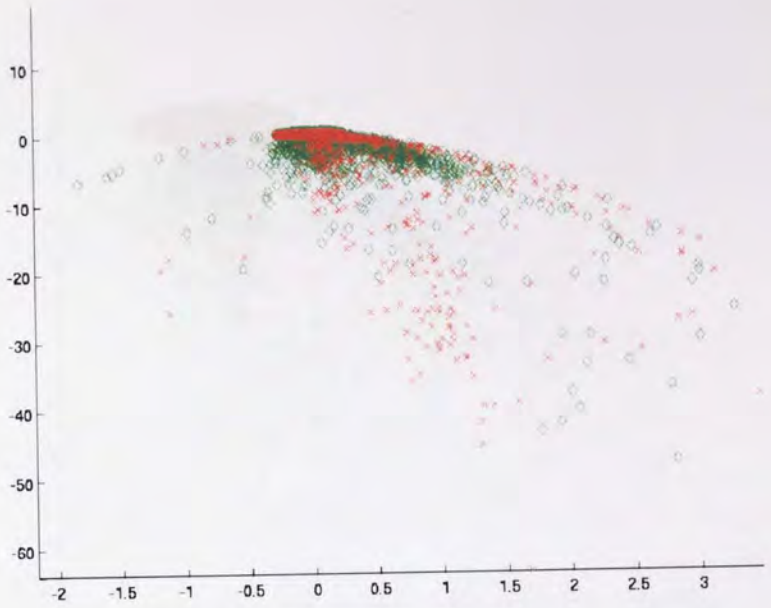


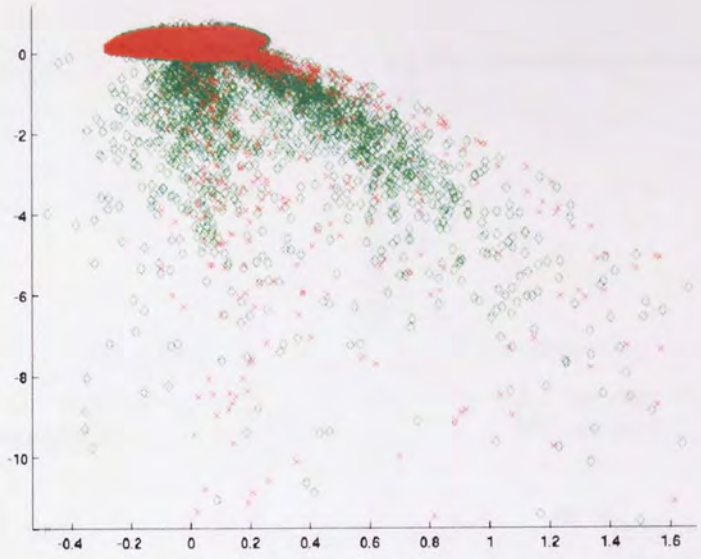
Figure 3.15: The NeuroScale visualisation using all the test data projected onto the principal components defined by BIC analysis of the learning set. The green diamonds are data from group N and the red crosses represent data from group P.

3.3.3 Neural Network Classifier

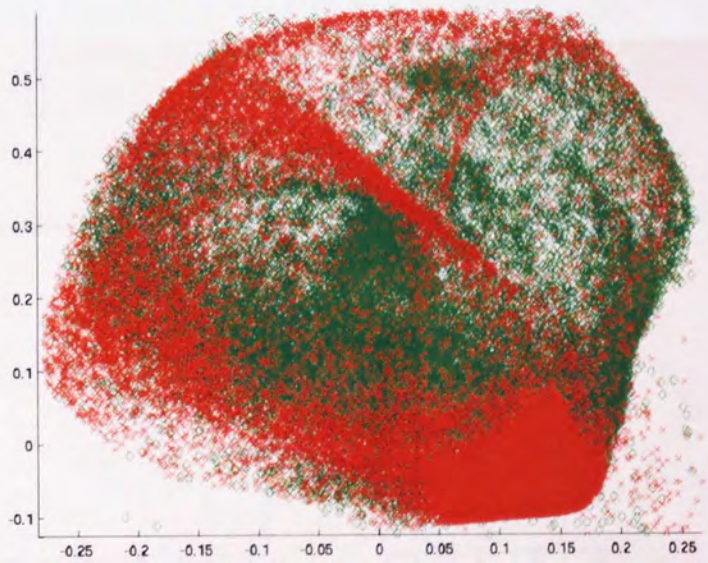
A quantitative assessment of the success of the method of dimensionality reduction using principal component analysis followed by neural network implementation was carried out using a similar method as above but with the NeuroScale RBF replaced by a multi-layer perceptron (MLP) with logistic sigmoidal activation functions [Nabney, 1999] trained using the Bayesian evidence procedure [MacKay, 1992]. This was again done using the last 200 P-waves of the each record.

Leave-one-out cross validation [Weiss and Kulikowski, 1991] was performed on the data set. This involves taking one record out of the data and performing the whole training process using the rest of the data then classifying the omitted record. This is then repeated, omitting each record in turn and training on the rest, until all records have been classified. This negates the need for a test set due to the repetition over all the records. In our implementation we left out one record from each class for each iteration and trained on the rest.

The MLP was trained with d inputs, ten hidden units and one output; the targets for the output were defined as 0 for group N and 1 for group P. The inputs were the P-waves of the training set projected onto the first eight principal components of the P-wave learning

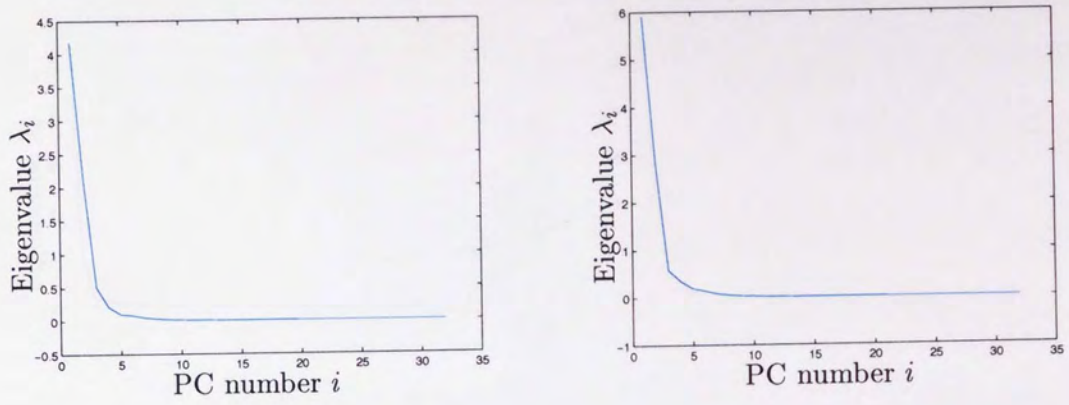


(a) Medium Magnification



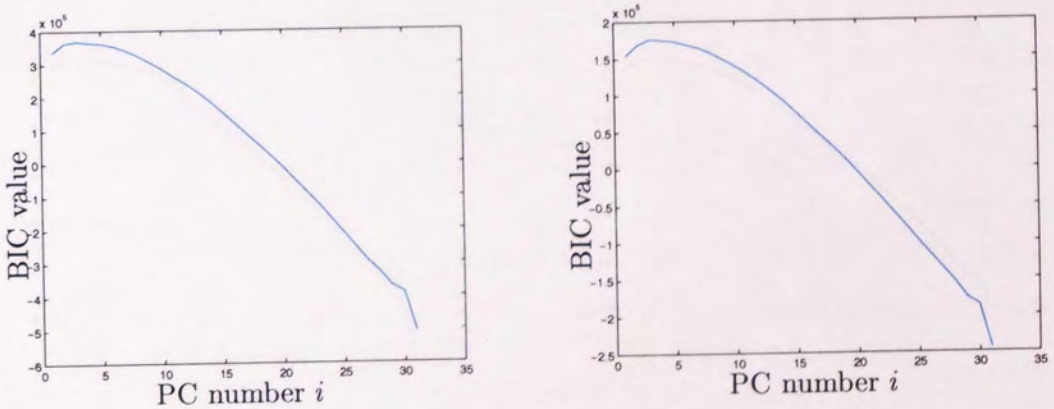
(b) High Magnification

Figure 3.16: A closer look at the cluster in Figure 3.15 for medium magnification (a) and high magnification (b). The green diamonds are data from group N and the red crosses represent data from group P.



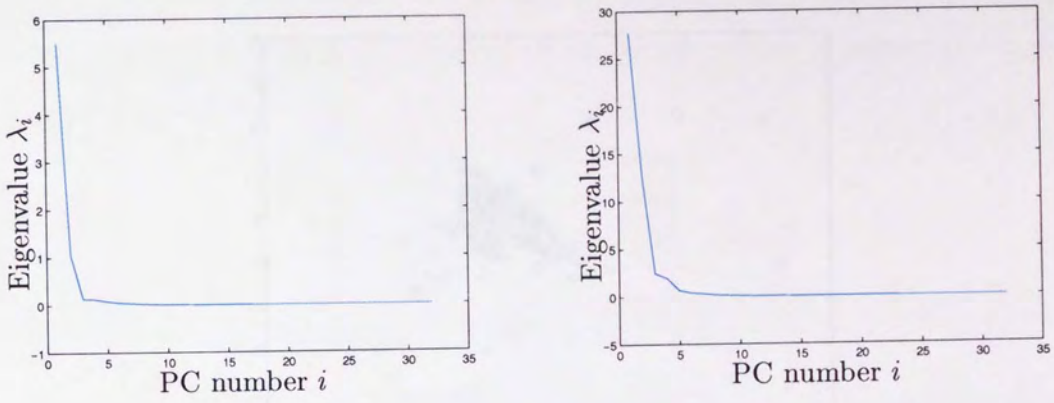
(a) $\mathrm{L_{PCA}}$ (b) $\mathrm{N_{L_{PCA}}}$

Figure 3.17: The eigenvalue spectra of groups $\mathrm{N_{L_{PCA}}}$ and $\mathrm{P_{L_{PCA}}}$ for every P-wave in these groups.



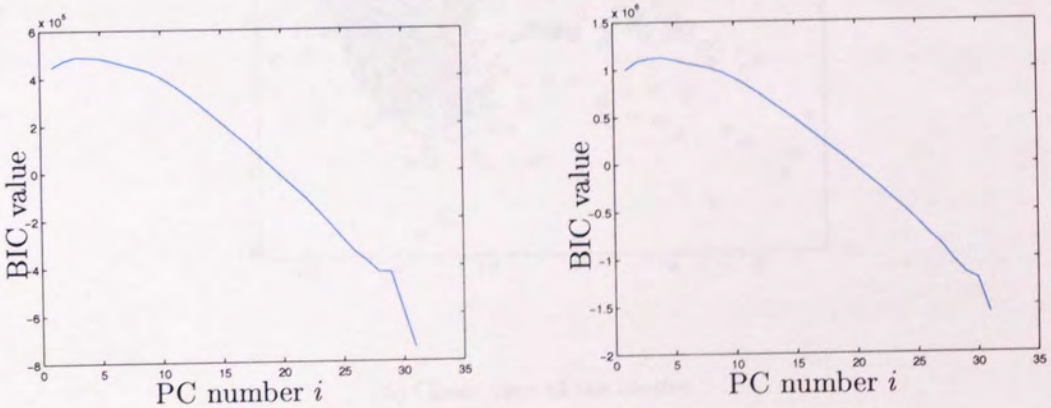
(a) $\mathrm{N_{L_{PCA}}}$ (b) $\mathrm{P_{L_{PCA}}}$[

Figure 3.18: The results of the BIC analysis applied to groups $\mathrm{N_{L_{PCA}}}$ and $\mathrm{P_{L_{PCA}}}$ for every P-wave in these groups.



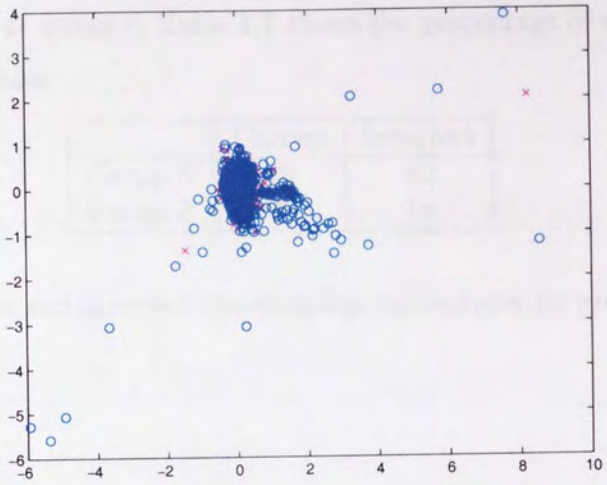
(a) $P_{cL_{PCA}}$ (b) $P_{dL_{PCA}}$[

Figure 3.19: The eigenvalue spectra of groups $N_{L_{PCA}}$ and $P_{L_{PCA}}$ for every P-wave in these groups.

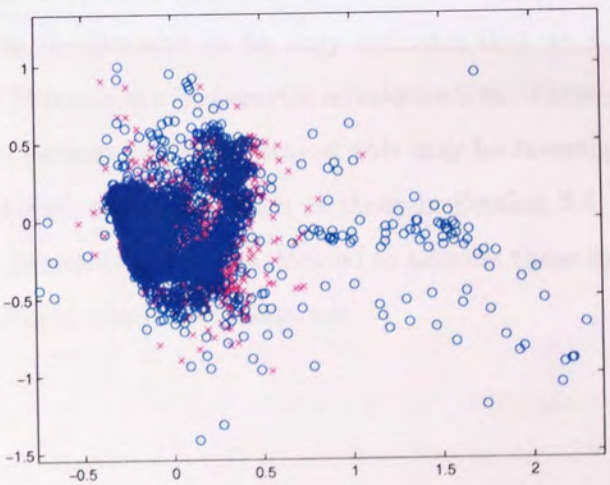


(a) $P_{cL_{PCA}}$ (b) $P_{dL_{PCA}}$

Figure 3.20: The results of the BIC analysis applied to groups $N_{L_{PCA}}$ and $P_{L_{PCA}}$ for every P-wave in these groups.



(a) View of all results



(b) Closer view of the cluster

Figure 3.21: The NeuroScale visualisation of the test data projected onto the principal components defined by BIC analysis of the learning set. Plot (a) shows the data with no magnification and plot (b) shows the magnification of the central cluster. The blue circles are data from those distant to AF and the magenta crosses represent those close to an AF episode.

set as decided by the BIC method. The P-waves of the absent records were then propagated through the MLP and which output a number between 0 and 1 for each P-wave (200 for each record). A classification level, at 0.5, was set and each P-wave below that classified as group N and those above that as group P. Table 3.2 shows the percentage of correct and incorrect classifications for each class.

	Correct	Incorrect
Group N	18	82
Group P	90	10

Table 3.2: Correct and incorrect classification percentages for groups N and P.

3.3.4 Conclusions

We can see from all of the initial analysis that it is difficult to extract meaningful information from the P-wave data; at this point we cannot conclude that the P-wave contains information pertinent to the diagnosis of atrial fibrillation. However, this is not surprising as in any cardiac condition the P-waves are likely to be very similar, especially over relatively long time periods. That this is evident from the research so far only indicates that we need to use a different set of features, not that P-waves are necessarily uninformative. These results motivated the subsequent research into *variation* in the P-waves; this may be investigated in a similar way to RR-interval variation with techniques such as those in Section 2.4. However, flaws were apparent in the current measures and it was decided to address these to create a more robust measure before application to the P-wave data set.

Chapter 4

Development of Kernel Entropy

In this chapter we further explore the concept of *entropy rate*, first mentioned in Chapter 2, and investigate how measures based on this can be used to quantify the level of regularity in a time series.

We shall, in Section 4.1, review in detail the dynamical systems theory behind entropy rate and how this was used to create the regularity measures that are currently in use. We expand on this to incorporate signal processing and probability density estimation theory in a novel approach in Section 4.2.

Section 4.3 focusses on the validation of the new measure. It is compared to the best current measure on synthetic data and subjected to a variety of tests to gauge its robustness and consistency when compared with the previous approach.

4.1 Theory

The entropy rate (also known as metric entropy or the Kolmogorov-Sinai invariant) is the *mean rate of creation of information* [Eckmann and Ruelle, 1985]. This is the average information gain that can be obtained from each observation; a measure of the time rate of creation of information as a chaotic orbit evolves [Ott, 1993]. It is, therefore, 0 for non-chaotic motion as no information is gained by additional observations and greater than 0 for chaotic motion where, on average, more observations will yield more information about the system.

The entropy rate is of great use in the mathematical theory of chaos when the equations governing a system are known, but is of much less use in practice as the quantity is difficult to determine numerically for a sequence of measurements $\mathbf{x}$, with elements x_i, such that

$$\mathbf{x} = [x_1, x_2, \ldots, x_N]. \tag{4.1}$$

As our aim is to develop a measure to investigate real world data sets, the theory given here is presented with a view to provide understanding of the origin of the regularity measures which form the latter part of this chapter. There are a number of ways to define the entropy rate; we shall explore two of them as understanding both facilitates the comprehension of how the subsequent regularity measures work.

4.1.1 Partitions

Typically in older papers in dynamical system theory, the entropy rate is defined in terms of partitions as this gives an intuitive representation of how a chaotic system returns a positive entropy rate.

We define a probability space $(\Omega, \mathcal{B}, \rho)$, where Ω is the *sample space* and $\mathcal{B}$ is a set of subsets of Ω, where each subset is known as an *event*. An element of Ω is denoted by ω. ρ is an ergodic probability measure assigned to each event.

Hence, a set of α disjoint subsets, Q_i, of Ω forms a partition $\mathcal{Q}$ if

$$\mathcal{Q} = Q_1 \cup Q_2 \cup \cdots \cup Q_\alpha, \tag{4.2}$$

and $\mathcal{Q}$ is non-empty.

If $A = \{a_1, a_2, \ldots, a_\alpha\}$ is the finite alphabet of possible outcomes of a mapping $M : \Omega \to A$, we can consider the partition $\mathcal{Q} = \{Q_i; i = 1, 2, \ldots, \alpha\}$ defined by $Q_i = \{\omega : M(\omega) = a_i\} = M^{-1}(\{a_i\})$ [Gray, 1990]. We can then write the entropy as a function of the partition defined by the disjoint sets of the preimages under M on the alphabet A as

$$H_\rho(\mathcal{Q}) = -\sum_{i=1}^{\alpha} \rho(Q_i) \log \rho(Q_i), \tag{4.3}$$

where we also define $u \log u = 0$ if $u = 0$. So $H_\rho(\mathcal{Q})$ is the information content of the partition with respect to the probability measure which is denoted ρ, as opposed to p, to indicate that it is ergodic.

Define M^τ to be the mapping M applied τ times. The inverse images $M^{-\tau}$ can be partitioned in a similar way to above, denoted $\mathcal{Q}^\tau$, and so we can describe the evolution of the system in time as

$$\mathcal{Q}^\tau = \mathcal{Q}^0 \cup \mathcal{Q}^1 \cup \ldots \cup \mathcal{Q}^{\tau-1}. \tag{4.4}$$

Q^τ is a partition of the phase space created by applying the mapping τ times. We can write the components of this partition as

$$P_k^\tau = Q_{i_1}^0 \cap Q_{i_2}^1 \cap \ldots \cap Q_{i_\tau}^{\tau-1}, \tag{4.5}$$

where $i_j \in \{1, 2, \ldots, \alpha\}$ and $k \in \{1, 2, \ldots, \alpha^\tau\}$. Therefore, the information entropy of an interval of time period of length τ with respect to the state ρ can be written as

$$H_\rho(\mathcal{Q}^\tau) = -\sum_{k=1}^{\alpha^\tau} \rho(P_k^\tau) \log \rho(P_k^\tau). \tag{4.6}$$

The *rate* of information creation is defined as the limit

$$h_\rho = \lim_{\tau \to \infty} [H(\mathcal{Q}^{\tau+1}) - H(\mathcal{Q}^\tau)]. \tag{4.7}$$

4.1.2 Probabilistic Representation

Moving away from partitions, an alternative derivation of the entropy rate can be from a more probabilistic perspective. The whole approach can be seen as being based on the conditional probability on the sequence of observations, $\mathbf{x}$, from Equation (4.1),

$$p(x_{i+m} | x_{i+m-1}, x_{i+m-2}, \ldots, x_i), \tag{4.8}$$

where x_k is an event in the series at time k, and i is an index variable that denotes the start of the sequence. In other words, the probability of the next event in the sequence occurring given that the previous m events have occurred in sequence; m is analogous to τ above. p is used instead of ρ as the condition of ergodicity has been relaxed.

If we set $\mathbf{x}_i^m = \{x_{i+m-1}, x_{i+m-2}, \ldots, x_i\}$, and we write $\mathbf{x}^m = \{\mathbf{x}_i^m\}$ as the set of all $\mathbf{x}_i^m$, then we can express the entropy rate in Equation (4.7) in probabilistic terms;

$$h = \lim_{m \to \infty} \big[H(x_{m+1}, \mathbf{x}^m) - H(\mathbf{x}^m)\big], \tag{4.9}$$

where $H(x_{m+1}, \mathbf{x}_i^m)$ is the joint entropy of x_{m+1} and $\mathbf{x}^m$. From this we can write

$$
\begin{aligned}
h &= \lim_{m\to\infty}\left[\sum_{x\mathbf{x}^m} p(x_{m+1},\mathbf{x}^m)\log\frac{1}{p(x_{m+1},\mathbf{x}^m)} - \sum_{\mathbf{x}^m} p(\mathbf{x}^m)\log\frac{1}{p(\mathbf{x}^m)}\right], \\
&= \lim_{m\to\infty}\left[\sum_{x\mathbf{x}^m} p(\mathbf{x}^m)p(x_{m+1}|\mathbf{x}^m)\left[\log\frac{1}{p(\mathbf{x}^m)} + \log\frac{1}{p(x_{m+1}|\mathbf{x}^m)}\right] - \sum_{\mathbf{x}^m} p(\mathbf{x}^m)\log\frac{1}{p(\mathbf{x}^m)}\right], \\
&= \lim_{m\to\infty}\left[\sum_{x\mathbf{x}^m} p(\mathbf{x}^m)\log\frac{1}{p(\mathbf{x}^m)} + \sum_{x\mathbf{x}^m} p(\mathbf{x}^m)p(x_{m+1}|\mathbf{x}^m)\log\frac{1}{p(x_{m+1}|\mathbf{x}^m)} - \sum_{x\mathbf{x}^m} p(\mathbf{x}^m)\log\frac{1}{p(\mathbf{x}^m)}\right], \\
&= \lim_{m\to\infty}\left[\sum_{x\mathbf{x}^m} p(\mathbf{x}^m)p(x_{m+1}|\mathbf{x}^m)\log\frac{1}{p(x_{m+1}|\mathbf{x}^m)}\right], \qquad (4.10)
\end{aligned}
$$

which gives us

$$h = \lim_{m\to\infty} H(x_{m+1}|\mathbf{x}^m), \qquad (4.11)$$

which is analogous to Equation 4.7.

4.1.3 Previous Measures

Equation (4.7) cannot be applied to real world applications as the data is always of finite length and so the limit $\tau \to \infty$ cannot be calculated when the dynamical equations governing a system are unknown. Pincus noted that even an approximation of this may have intrinsic interest in determining the nature of a dynamical system and developed the approximate entropy measure (ApEn) to investigate this [Pincus, 1991]. Despite being called 'entropy', it is better to think of them as measures of *regularity* [Fusheng et al., 2001] rather than measures of disorder. Here we review ApEn and another measure based on the entropy rate, known as sample entropy.

Approximate Entropy

If we approach the problem as in Section 4.1.1, we can partition the dynamical space by introducing a ball centred at $\mathbf{x}_i^m$ with radius r, denoted $B(\mathbf{x}_i^m, r)$.

As we have a signal as in Equation (4.1), then we can define a distance measure to be the maximum Euclidean distance between a window vector $\mathbf{x}_i^m$, consisting of m consecutive values starting at value x_i, and a corresponding window $\mathbf{x}_j^m$ starting at the value x_j,

$$d_E[\mathbf{x}_i^m, \mathbf{x}_j^m] = \max\{|x_{i+k} - x_{j+k}| : 0 \le k \le m-1\}. \qquad (4.12)$$

As such, if a vector $\mathbf{x}_j^m$ lies within the ball $B(\mathbf{x}_i^m, r)$, then $d_E[\mathbf{x}_i^m, \mathbf{x}_j^m] \le r$. Hence, we can find the number of $\mathbf{x}_j^m, j = \{0, 1, \ldots, N-m+1\}$ that lie within the ball $B(\mathbf{x}_i^m, r)$,

$$N_i^m(r) = \#\{\mathbf{x}_j^m \in B(\mathbf{x}_i^m, r), j = \{0, 1, \ldots, N-m+1\}\}. \qquad (4.13)$$

We can then calculate the probability that a sequence of size m will occur (within the tolerance value) thus

$$C_i^m(r) = \frac{N_i^m(r)}{N-m+1}. \tag{4.14}$$

It is now possible to obtain h_ρ directly [Eckmann and Ruelle, 1985]. If we define

$$\phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log C_i^m(r), \tag{4.15}$$

then, using Equation (4.7), we can say

$$h_\rho = \lim_{r\to 0} \lim_{m\to\infty} \lim_{N\to\infty} [\phi^m(r) - \phi^{m+1}(r)]. \tag{4.16}$$

As this is still not suitable for finite data sets, further adjustments need to be made. Approximate entropy (ApEn) is essentially Equation (4.16) but with fixed m and r, and N data points. It is defined as

$$ApEn(m, r, N) = \phi^m(r) - \phi^{m+1}(r). \tag{4.17}$$

Although ApEn is derived from and resembles Equation (4.7), it is worth noting that it is not intended to represent it precisely and should be considered as a separate statistic in its own right [Pincus, 1991]. It is worth mentioning again that this formulation, as well as the subsequent ones, cannot robustly determine chaotic behaviour and so should not be considered a quantitative test for chaos. As they do supply information on the regularity of the signal evolution in time they are termed *regularity measures* [Richman and Moorman, 2000].

Sample Entropy

It has been shown that ApEn is inherently biased [Richman and Moorman, 2000], and therefore sample entropy was developed to address this bias and provide a more rigorous regularity measure.

One source of this bias is the necessity of ensuring a non-zero value for Equation (4.13). This is so that the logarithm taken in Equation (4.14) is well defined. The method of assuring that this constraint is fulfilled in approximate entropy is by allowing $i = j$ (known as 'self-matching' [Richman and Moorman, 2000]) in Equation (4.13). This is equivalent to saying that $\mathbf{x}_i^m$ is always in $B(\mathbf{x}_i^m, r)$ and so $N_i^m(r)$ is always positive.

We can see how this causes bias in the statistic by considering $N_i^m(r)$ and $N_i^{m+1}(r)$. If we express approximate entropy in a similar fashion to Equation (4.11), it can be considered as the log of the conditional probability that $N_i^m(r)$ and $N_i^{m+1}(r)$ stay the same over time:

$$ApEn(m,r,N) \approx \frac{1}{N-m}\sum_{i=1}^{N-m}\log\left(\frac{N_i^m(r)}{N_i^{m+1}(r)}\right). \tag{4.18}$$

Now, as neither $N_i^m(r)$ or $N_i^{m+1}(r)$ can be 0 for this to be defined, the self-matching biases the statistic to give a positive result for the statistic which causes bias, especially in small series [Richman and Moorman, 2000].

SampEn removes this bias by removing all self matches. The formulation is also slightly different. Only the first $N-m$ values of the series are used when calculating $N_i^m(r)$ and $N_i^{m+1}(r)$ ensuring an equal series length for each value. Also, two new variables were defined, based on the *correlation integral*. The correlation integral is simply the average of the series of $C_i^m(r)$ (defined in Equation (4.14)),

$$C^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} C_i^m(r). \tag{4.19}$$

Sample entropy is defined in a similar fashion to the approximate entropy above. Firstly, $N_i^{'m}(r)$ is defined to discount the self matches,

$$N_i^{'m}(r) = \#\{\mathbf{x}_j^m \in B(\mathbf{x}_i^m, r),\ j = \{0,1,\ldots,N-m\}, j \neq i\}, \tag{4.20}$$

We now define

$$U_i^m(r) = \frac{1}{N-m-1}N_i^{'m}(r), \tag{4.21}$$

and, following from Equation (4.19),

$$U^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} U_i^m(r). \tag{4.22}$$

$U^{m+1}(r)$ is similarly defined for $m+1$

$$N_i^{'m+1}(r) = \#\{\mathbf{x}_j^{m+1} \in B(\mathbf{x}_i^{m+1}, r),\ j = \{0,1,\ldots,N-m\}, j \neq i\}, \tag{4.23}$$

for $j = 1,2,\ldots,m, j \neq i$. We now proceed as before

$$U_i^{m+1}(r) = \frac{1}{N-m-1}N_i^{'m+1}(r), \tag{4.24}$$

$$U^{m+1}(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} U_i^{m+1}(r). \tag{4.25}$$

Sample entropy is the negative logarithm of the ratio of these probabilities

$$SampEn(m,r,N) = -\log\left(\frac{U^{m+1}(r)}{U^m(r)}\right) \tag{4.26}$$

We can see how this compares to the approximate entropy given in Equation (4.18) by noting that the $1/(N-m)$ and $1/(N-m-1)$ terms cancel so we can write it as

$$SampEn(m,r,N) = \log\left(\frac{\sum_{i=1}^{N-m} N_i^{'m}}{\sum_{i=1}^{N-m} N_i^{'m+1}}\right). \tag{4.27}$$

This is the log of the sum of the conditional probability that if two sequences are classified as *similar* within a tolerance of r for m points, the next points of each sequence will also be within r of each other [Richman and Moorman, 2000]. It can be seen from this that if there are no matches for either $N_i^{'m}$ or $N_i^{'m+1}$ then the measure is undefined.

If defined, an upper bound for the value of SampEn can also be calculated, minimising $U^{m+1}(r)$ and maximising $U^m(r)$ in Equation (4.26).

Unless there are no matches, $U^{m+1}(r)$ is minimised when only 1 pair of vectors (say, v and w) of length $m+1$ match. In that case $N_i^{'m+1(r)} = 2$ so

$$U_v^{m+1}(r) = \frac{1}{N-m-1},$$

and

$$U_w^{m+1}(r) = \frac{1}{N-m-1},$$

then

$$U^{m+1}(r) = \frac{2}{[N-m][N-m-1]}.$$

Similarly, $U^m(r)$ is maximised when all $N-m-1$ vectors are within a tolerance r of each other so that $N_i^{'m+1(r)} = N-m-1$ which means

$$U_i^m(r) = \frac{N-m-1}{N-m-1} = 1,$$

for each value of i so

$$U^m(r) = \frac{N-m}{N-m} = 1.$$

This means the maximum probability that the equation can achieve is $2/[N-m][N-m-1]$ and therefore the upper bound is

$$\begin{aligned} SampEn(m,r,N) &\leq -\log\frac{2}{[N-m][N-m-1]}, & (4.28)\\ &\leq \log(N-m)+\log(N-m-1)-\log(2). & (4.29) \end{aligned}$$

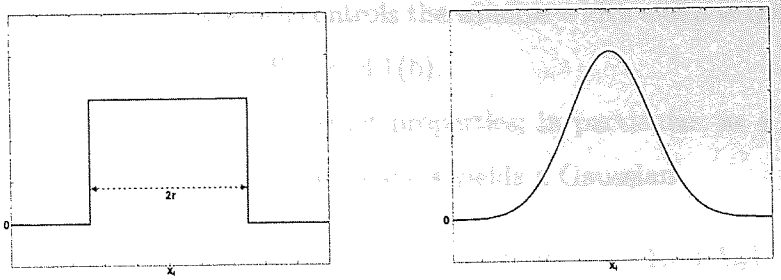
Therefore, the upper bound and therefore the values of sample entropy scales as approximately $2\log(N)$ as the value of m should be negligible compared to N.

4.2 Kernel-Based Entropy Measure

The entropy measures introduced so far can be seen as *phase space reconstruction* methods, as $\mathbf{x}_i^m$ is a delay vector of size m. The set of these, for $i = 1, 2, \ldots, N - m + 1$, is the phase space representation of the signal for dimension m. We also need to estimate the probability that this path in phase space repeats itself. In the previous measures, calculation of this probability is based on a binary classification of whether two delay vectors are *similar* to each other or not, the degree of similarity allowed being within a tolerance of r. However, although this is conventional in dynamical systems theory, arguably the application of a regularity measure such as this to a time series also falls in the signal processing and pattern recognition domain where it is quite unusual; it is equivalent to estimating a probability density using a mixture of uniform distributions and does not consider distances greater than r. In probability density estimation terms, this is a square kernel Parzen window around each point $\mathbf{x}_i^m$. The common noise model assumption is that of *additive white noise* [Bishop, 1995]. One method to use for probability density function estimation under this assumption is a Gaussian kernel Parzen window.

Using Gaussian kernels instead of square kernels would have obvious benefits; for instance, a higher probability would be assigned to points closer to the central point. Also, it is easy to avoid the pitfalls associated with $\log 0$ because every point has a non-zero density. There is a computational issue if an outlier is so distant that the associated probability falls below machine number representation. This can easily be dealt with if we are aware of it by incorporating a failsafe in the program where the probability cannot fall below machine precision.

There are some obvious drawbacks with using Gaussian kernels too. The main one is the computational cost; one of the greatest benefits of the square kernel method is that it is very computationally efficient. However, using some mathematical properties of Gaussians, we can show how the use of Gaussian kernels in an entropy formulation can be reconciled with computational efficiency whilst still retaining a sound analytical justification.



(a) Square Kernel (b) Gaussian Kernel

Figure 4.1: A graphical representation of the two kernel types for Parzen window probability density estimation.

4.2.1 Parzen Window

A Parzen window is a type of probability density estimation scheme that utilises kernels [Wand and Jones, 1995]. A kernel is a parametric density model such as a Gaussian which is placed on top of each data point and the full density is evaluated as the average of the kernels. In our application, we wish to evaluate the density function at a point $\mathbf{x}_i^m$. The density estimation model can be written as

$$f_P(\mathbf{x}_i^m) = \frac{1}{N}\sum_{j=1}^{N} K(\mathbf{x}_i^m - \mathbf{x}_j^m, \eta), \tag{4.30}$$

where η is the window width parameter and K is the kernel function. As with any density function, it is positive and satisfies the constraint

$$\int K(\mathbf{x}, \eta)\ \mathrm{d}\mathbf{x} = 1. \tag{4.31}$$

We can see parallels with the methods employed in the entropy measures outlined above. The kernel is the function $d[\mathbf{x}_i^m, \mathbf{x}_j^m] \leq r$, which in density estimation notation would be written as

$$K(\mathbf{x}_i^m, r) = \begin{cases} 1 & \text{if } \max\{|x_{j+k}| : 0 \leq k \leq N\} \leq r \\ 0 & \text{otherwise} \end{cases} \tag{4.32}$$

with r corresponding to the window width. This can be seen in Figure 4.1(a). With a Gaussian kernel, the functional form is given by

$$G(\mathbf{x}, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right), \tag{4.33}$$

where Σ is the covariance matrix which controls the window width and d is the dimensionality of the data. This kernel is shown in Figure 4.1(b).

The Gaussian kernel has some important properties; in particular, as a property of the Fourier transform, a convolution of two Gaussians yields a Gaussian

$$\int G(\mathbf{x}_i - \mathbf{x}_j, \Sigma_1) G(\mathbf{x}_i - \mathbf{x}_k, \Sigma_2)\ \mathrm{d}\mathbf{x} = G(\mathbf{x}_j - \mathbf{x}_k, \Sigma_1 + \Sigma_2). \tag{4.34}$$

We shall now show how this property can be used to improve the computational efficiency of an entropy formulation whilst retaining analytical justification.

4.2.2 Renyi Entropy

The family of Renyi entropies are defined by

$$H_{R_\alpha} = \frac{1}{1-\alpha} \log \int p(\mathbf{x})^\alpha\ \mathrm{d}\mathbf{x},\ \alpha \neq 1, \tag{4.35}$$

where α denotes the order of the entropy, $\alpha > 0$ [Rényi, 1961]. In the limit $\alpha \to 1$, this is equivalent to the information entropy given in Equation (4.3).

As mentioned previously, the use of the term 'entropy' has always been rather loosely used in the approximate entropy family of regularity measures. When ϕ^m is calculated in Equation (4.15), the measure is simply the logarithm of the probabilities rather than the information entropy or any other standard entropy measure. However, recently it has been noted that the approximate entropy, given in Equation (4.17), approximates the Renyi entropy of order 1 (the information entropy) and the sample entropy, given in Equation (4.26), approximates the Renyi entropy of order 2 [Costa and Healey, 2003].

We directly use the Renyi entropy of order 2, which is termed the *quadratic entropy* as it uses the second power of the probabilities [Xu and Príncipe, 1998]. Numerically calculating the integral of a squared probability function would not normally be feasible for many real world data sets due to the computational expense. However, if we use Gaussian kernels in the quadratic entropy, we can use the property from Equation (4.34) to provide a much more computationally tractable approach. For notational simplicity, we say $n = N - m + 1$. If we also assume that the Gaussians are spherical ($\Sigma = \sigma^2 \mathbf{I}^m$, where $\mathbf{I}^m$ is the identity matrix of

dimension m), from [Príncipe et al., 1999] we have

$$\begin{aligned} H_{R_2} &= -\log \int p(\mathbf{x})^2 \, \mathrm{d}\mathbf{x} \\ &= -\log \int \frac{1}{n^2} \left(\sum_{j=1}^{n} \sum_{k=1}^{n} G(\mathbf{x}^m - \mathbf{x}_j^m, \sigma^2 \mathbf{I}^m) G(\mathbf{x}^m - \mathbf{x}_k^m, \sigma^2 \mathbf{I}^m) \right) \mathrm{d}\mathbf{x} \\ &= -\log \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} G(\mathbf{x}_j^m - \mathbf{x}_k^m, 2\sigma^2 \mathbf{I}^m). \end{aligned} \tag{4.36}$$

In its full form, it is expressed as

$$H_{R_2}^m = -\log \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{1}{(2\pi)^{\frac{m}{2}} |2\sigma^2 \mathbf{I}^m|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x}_j^m - \mathbf{x}_k^m)^T (2\sigma^2 \mathbf{I}^m)^{-1} (\mathbf{x}_j^m - \mathbf{x}_k^m) \right). \tag{4.37}$$

This means that we can precisely calculate the quadratic Renyi entropy for a probability density estimated using Gaussian kernels with pairwise sums. This has significant computational benefits as the computation is of $O(N^2)$, which is significantly more efficient than numerical integration whilst remaining theoretically sound.

Kernel Entropy

The quadratic Renyi entropy can easily be incorporated into the entropy rate framework by incorporating it in Equation (4.7),

$$h_{R_2\rho} = \lim_{\tau \to \infty} [H_{R_2}(Q^{\tau+1}) - H_{R_2}(Q^{\tau})]. \tag{4.38}$$

For calculating the statistic from finite data, we need to determine the time scale, m, as before, and the width of the Gaussian distribution σ. We can then define an approximation of the Renyi entropy rate as

$$KernEn(m, \sigma) = \lim_{\sigma \to 0} \lim_{m \to \infty} \lim_{N \to \infty} [H_{R_2}^{m+1}(\sigma) - H_{R_2}^m(\sigma)], \tag{4.39}$$

which, when estimated for finite data is defined as

$$KernEn(m, \sigma, N) = H_{R_2}^{m+1}(\sigma) - H_{R_2}^m(\sigma). \tag{4.40}$$

We term this the *Kernel Entropy* to distinguish it from other forms of entropy and to highlight the importance of the Parzen window model in its formulation.

Renyi entropy rate has been implemented in a very recent paper to quantify the Gaussianity present in heart rates under various conditions [Lake, 2006]. The approach to estimating the probabilities is based on the method used for the sample entropy in Equation (4.26), rather than utilising the properties of Gaussian kernels as we have. The paper does provide

an interesting insight into properties and applications of the Renyi entropy rate as opposed to the information entropy rate and suggests the use of Gaussian kernels would have beneficial properties. This is independent of Woodcock and Nabney [2006] which theoretically defines and implements the kernel entropy.

For comparison of this measure to the previous measures, we can calculate a bound in a manner similar to that calculated for sample entropy in Equation (4.29). In this case, we need to minimise $H_{R_2}^m$ and maximise $H_{R_2}^{m+1}$.

As $H_{R_2}^{m+1}$ is a negative function, to maximise it, we need to minimise

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{1}{(2\pi)^{\frac{m+1}{2}}|2\sigma^2\mathbf{I}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}_j^m-\mathbf{x}_k^m)^T(2\sigma^2\mathbf{I})^{-1}(\mathbf{x}_j^m-\mathbf{x}_k^m)\right). \tag{4.41}$$

The minimum occurs when all the $\mathbf{x}_j^m$ and $\mathbf{x}_k^m$ are sufficiently far apart that the associated probability tends to zero. If we attempt to approximate this, by saying $|\mathbf{x}_j - \mathbf{x}_k| \to \infty$, there will still be a single $\mathbf{x}_j = \mathbf{x}_k \ \forall\ j$ due to the function being derived from a square of itself. Hence, Equation 4.41 is minimised when

$$G(\mathbf{x}_j^m - \mathbf{x}_k^m, 2\sigma^2\mathbf{I}^{m+1}) = \begin{cases} 1 & \text{if } j=k, \\ 0 & \text{otherwise,} \end{cases} \tag{4.42}$$

Using this condition, along with the fact that $|2\sigma^2\mathbf{I}^{m+1}| = (2\sigma^2)^{m+1}$, we have

$$\begin{aligned}
\max H_{R_2}^{m+1} &= -\log\frac{1}{(n-1)^2}\sum_{j=1}^{n-1}\sum_{k=1}^{n-1}\frac{1}{(2\pi)^{\frac{m+1}{2}}(2\sigma^2)^{\frac{m+1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}_j-\mathbf{x}_k)^T(2\sigma^2\mathbf{I})^{-1}(\mathbf{x}_j-\mathbf{x}_k)\right), \\
&= -\log\frac{1}{(n-1)^2}\frac{n-1}{(2\pi)^{\frac{m+1}{2}}(2\sigma^2)^{\frac{m+1}{2}}}, \\
&= -\log\frac{1}{n-1} - \log\left[(2\pi)^{\frac{m+1}{2}}(2\sigma^2)^{\frac{m+1}{2}}\right]^{-1}, \\
&= \log(n-1) + \log(2\pi)^{\frac{m+1}{2}} + \log(2\sigma^2)^{\frac{m+1}{2}},
\end{aligned}$$

which gives us

$$\max H_{R_2}^{m+1} = \log(n-1) + \frac{m+1}{2}\log(2\pi) + \frac{m+1}{2}\log(2\sigma^2). \tag{4.43}$$

To minimise $H_{R_2}^m$ we need every vector of length m to be an exact match with every other vector in the system. That is to assume $\mathbf{y}_j = \mathbf{y}_k \ \forall\ j,\ k$. Therefore,

$$\begin{aligned}
\min H_{R_2}^m &= -\log \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{1}{(2\pi)^{\frac{m}{2}} |2\sigma^2 \mathbf{I}^m|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x}_j - \mathbf{x}_j)^T (2\sigma^2 \mathbf{I}^m)^{-1} (\mathbf{x}_j - \mathbf{x}_j) \right), \\
&= -\log \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{1}{(2\pi)^{\frac{m}{2}} |2\sigma^2 \mathbf{I}^m|^{\frac{1}{2}}} \exp(0), \\
&= -\log \frac{1}{n^2} \frac{n^2}{(2\pi)^{\frac{m}{2}} |2\sigma^2 \mathbf{I}^m|^{\frac{1}{2}}}, \\
&= -\log \left[(2\pi)^{\frac{m}{2}} |2\sigma^2 \mathbf{I}^m|^{\frac{1}{2}} \right]^{-1}, \\
&= \frac{m}{2} \log(2\pi) + \frac{1}{2} \log |2\sigma^2 \mathbf{I}^m|.
\end{aligned}$$

and as $|2\sigma \mathbf{I}^m| = (2\sigma)^m$, we have

$$\min H_{R_2}^m = \frac{m}{2} \log(2\pi) + \frac{m}{2} \log(2\sigma), \tag{4.44}$$

So, substituting Equations 4.43 and 4.44 in Equation 4.40, we have

$$\begin{aligned}
\max\ (KernEn) &= \max H_{R_2}^{m+1} - \min H_{R_2}^m, \\
&= \log(n-1) + \frac{m+1}{2} \log(2\pi) + \frac{m+1}{2} \log(2\sigma^2) - \frac{m}{2} \log(2\pi) - \frac{m}{2} \log(2\sigma^2), \\
&= \log(n-1) + \frac{1}{2} \log(2\pi) + \frac{1}{2} \log(2\sigma^2).
\end{aligned}$$

So the upper bound is

$$KernEn(m, \sigma, N) < \log(N-m) + \frac{1}{2} \left[\log(2\pi) + \log(2\sigma^2) \right]. \tag{4.45}$$

In comparison to the upper bound for the sample entropy in Equation 4.29, we can see that this scales as $\log(N)$ as opposed to $2\log(N)$. Also, this is different to the upper bound for the sample entropy as it is dependent on the standard deviation of the kernels, σ, as well as the size of the data set. It should also be noted that this highlights the fact that a small kernel variance may yield a negative kernel entropy value.

4.2.3 Selection of the Parameters

Of course, for use on real data, appropriate values of m and σ need to be found. For m, the problem is no different to that in the choice of the parameter for the other entropy measures; the standard approach is to use $m = 2$ [Acharaya et al., 2004; Fusheng et al., 2001; Pincus, 1991; Vikman et al., 1999]. However, as there may be benefits in working with different values of m, the method should be open to application to the widest possible range.

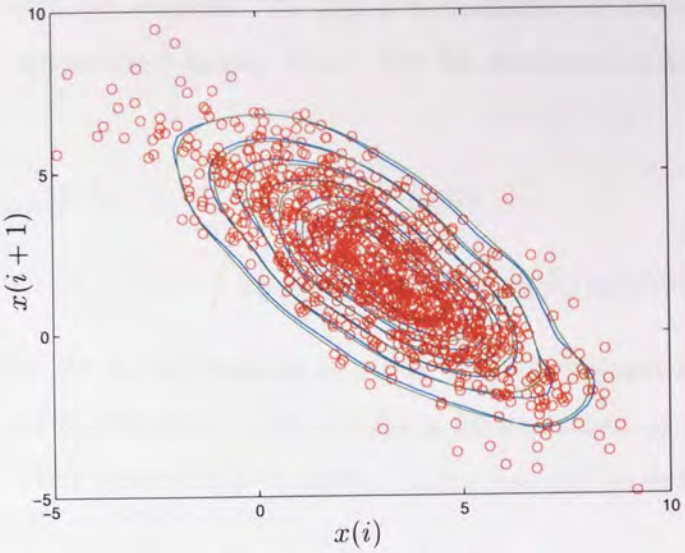


Figure 4.2: A plot showing the contour probabilities using a Gaussian kernel Parzen window (blue) and the normal reference rule (green) to choose the bandwidth. The 1000 data points (red) are sampled from a two dimensional Gaussian.

The same cannot be said for the window width parameter (often referred to as the *bandwidth*). The σ value is conceptually similar but not mathematically equivalent to the r threshold and so there is no reason to think that values that work well for r will do so for σ. As we are using a Gaussian kernel Parzen window, methods exist to estimate the optimal bandwidth from the data. However, although there are a number of these schemes available [Loader, 1999], most of them are inappropriate as the computation becomes increasingly prohibitive, especially for higher dimensional delay vectors. Because of this, we adopt a Bayesian approach using Markov chain Monte Carlo.

Bayesian Bandwidth Selection

The Bayesian approach in [Zhang et al., 2006] to bandwidth selection treats the components of Σ as parameters and aims to obtain the posterior density of the components of Σ by sampling with the Markov chain Monte Carlo (MCMC) method (see Section 2.3.3). As we are modelling the probability density using spherical Gaussians, this means that the bandwidth matrix is diagonal, and $\Sigma = \sigma^2\mathbf{I}$. Using MCMC is beneficial as we want our method to be flexible and the sampling algorithm can be applied to data of any dimension so we can determine reliable estimates for the bandwidth whatever the value of m is.

The method utilises the Kullback-Leibler (KL) divergence which is a non-symmetric dis-

tance measure between two densities. The aim is to minimise the distance from the target density $f(\mathbf{x})$ to the approximate density $\hat{f}_\Sigma(\mathbf{x})$. The KL divergence is defined as

$$D_{KL}(f, \hat{f}_\Sigma) = \int \log\left[\frac{f(\mathbf{x})}{\hat{f}_H(\mathbf{x})}\right] f(\mathbf{x})\mathrm{d}\mathbf{x} \tag{4.46}$$

$$= \int \log f(\mathbf{x}) f(\mathbf{x})\mathrm{d}\mathbf{x} - \int \log \hat{f}_\Sigma(\mathbf{x}) f(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{4.47}$$

which is non-negative. As the first term in Equation (4.47) is constant and we do not know the target density, the minimisation of $D_{KL}(f, \hat{f}_\Sigma)$ is the equivalent to the maximisation of $\int \log \hat{f}_\Sigma(\mathbf{x}) f(\mathbf{x})\mathrm{d}\mathbf{x}$. Using a kernel approximator, $K_\Sigma(\mathbf{y})$ this can be written as

$$\hat{E}\log[\hat{f}_\Sigma] = \sum_{i=1}^{N} \log \hat{f}_\Sigma(\mathbf{x}_i) = \sum_{i=1}^{N} \log\left[\frac{1}{N}\sum_{j=1}^{N} K_\Sigma(\mathbf{x}_i - \mathbf{x}_j, \Sigma)\right]. \tag{4.48}$$

As the maximisation of this directly leads to a bandwidth matrix of zeros, a leave-one-out cross validation estimator $\hat{f}_{\sigma,i}(\mathbf{x}_i)$ is used for the cost function in the MCMC method. We start by defining

$$\hat{f}_{\sigma,i}(\mathbf{x}_i) = \frac{1}{N-1}\sum_{\substack{j=1 \\ j\neq i}}^{N} |\sigma^2\mathbf{I}|^{-\frac{1}{2}} K\Big([\sigma^2\mathbf{I}]^{-\frac{1}{2}}(\mathbf{x}_i - \mathbf{x}_j)\Big), \tag{4.49}$$

and we use this to calculate the log likelihood,

$$L(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N|\sigma) = \frac{1}{N}\sum_{i=1}^{N} \log \hat{f}_{\sigma,i}(\mathbf{x}_i). \tag{4.50}$$

As we are using a Bayesian approach we need to fix a prior over σ. As the likelihood is flat when σ is large, a uniform prior will lead to a wide range of values accepted by the MCMC algorithm and therefore a reduction in robustness. Hence, we put a low prior probability on the areas of parameter space where the likelihood is very flat. A suitable prior for this purpose is a variant of the half-Cauchy prior [Bauwens and Lubrano, 1998]

$$\pi(\sigma, \lambda) \propto \frac{1}{1+\lambda\sigma^2}, \tag{4.51}$$

for $k = 1, 2, \ldots, m$ and where λ is a hyperparameter controlling the shape of the prior density. Zhang discusses the choice of λ in [Zhang et al., 2006], and concludes that the choice of λ does not affect the result to a noticeable extent. Hence, we use $\lambda = 1$.

From Bayes' theorem, the posterior for σ (up to a normalising constant) is given by

$$\pi(\sigma|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \propto \left[\prod_{k=1}^{m} \frac{1}{1+\lambda\sigma_k^2}\right] \prod_{i=1}^{n} \hat{f}_{\sigma,i}(\mathbf{x}_i). \tag{4.52}$$

We sample from this distribution using the Metropolis-Hastings algorithm implemented in NETLAB [Nabney, 1999]. The mean of these samples gives us the estimator for the optimal bandwidth.

Figure 4.2 shows a comparison of the Bayesian method with the Normal reference rule,

$$h = \sigma\left\{\frac{4}{(m+2)N}\right\}^{1/(m+4)}, \tag{4.53}$$

which is the standard method of approximating the optimal bandwidth for Gaussian target distributions [Zhang et al., 2006]. The Bayesian method is very close to the bandwidth suggested by the Normal reference rule and shows its usefulness in determining the bandwidth. For distributions that are non-Gaussian, the use of the Normal reference rule is not theoretically justified but the Bayesian method still determines a good approximation of the optimal bandwidth [Zhang et al., 2006]. Notice that at this stage, no assumption has been made about the nature of the kernel estimation method so it is applicable to any kernel method chosen, including the Gaussian or square kernels already mentioned.

4.3 Comparison with Previous Techniques

As ApEn has been shown to introduce bias, SampEn has recently superseded it as the current standard [Goldberger et al., 2000], and so we shall use it for comparison with KernEn. However, due to the inherent difference in the σ and r parameters, it is unsound to compare them for the same value. Therefore, the only option is to compare a range of parameter values and identify strengths and weaknesses in both statistics. Also, by testing σ without using the bandwidth selection enables us to investigate the optimal range for the new window width value. All of the experiments are carried out with $m = 2$ unless otherwise stated. This is because $m = 2$ is the standard value used for the previous measures in the literature [Richman and Moorman, 2000; Pincus, 1991; Acharaya et al., 2004].

Initially, we investigated how the measure behaves by directly varying σ without using the bandwidth selection procedure. As we want to be able to gauge the performance of kernel entropy as accurately as possible before we apply it to real world data, it is prudent to apply it to more than one synthetic data set. Here we shall apply it to two well known chaotic series, the Lorenz series and the Duffing-Van der Pol oscillator.

This first test was carried out by comparing the behaviour of the two measures under increasing noise for different r and σ. This is investigated on its own and then used to evaluate the behaviour of the measures for different series lengths N and different values of m.

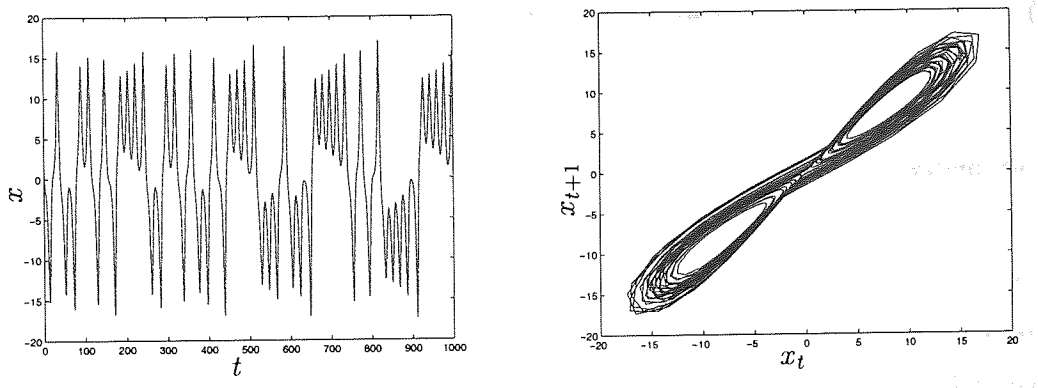
The second test we carried out was to determine how the measures performed in distinguishing a series from several surrogates sharing some statistical properties. This was again done for a range of r and σ values.

To investigate the effectiveness of the bandwidth selection procedure, we applied the method to a series and its surrogates to gauge the effectiveness at distinguishing a deterministic from a random series. Also, the behaviour of the bandwidth selection scheme was tested for a series with increasing noise.

Finally, we tested the ability of kernel entropy to quantify the level of chaos and disorder in a deterministic system. This was done in two related experiments. For the first, we used the Duffing-Van der Pol oscillator equations and altered the parameters to increase the level of disorder. We then compared the results of kernel entropy with the bandwidth selection procedure to the results of kernel and sample entropy for fixed parameter values as well as the information entropy. For the second experiment, the usefulness of the kernel entropy in determining chaos in a series is tested. This is done by comparing the value of the kernel

entropy with the bandwidth selection scheme with the value of the entropy rate for the Lorenz and the Duffing-Van der Pol oscillator series.

The Lorenz System



(a) Signal Space

(b) Phase Space

Figure 4.3: The x-value of the Lorenz series.

The behaviour of a dynamical system in time is mathematically represented by a coupled set of first order autonomous ordinary differential equations [Henry et al., 2001].

The Lorenz system [Lorenz, 1963] was originally conceived in the field of fluid dynamics [Ott, 1993] and is probably the most well known system in the field of non-linear dynamics and is almost ubiquitous in the literature. It is defined by the equations

$$\frac{dx}{dt} = a(y - x), \tag{4.54}$$

$$\frac{dy}{dt} = x(\rho - z) - y, \tag{4.55}$$

$$\frac{dz}{dt} = xy - \beta z, \tag{4.56}$$

with a, ρ and β as dimensionless parameters. In the generation of the time series for the subsequent experiments, we adopt the usual convention of setting $a = 10$, $\rho = 28$ and $\beta = 8/3$. The series was calculated using $x = 10$, $y = 0$ and $z = 10$ as the initial conditions and was sampled at unit time epochs with a burn in of 1000 iterations to make sure the phase path has forgotten the initial conditions and is following a normal path around the attractors. We shall only investigate the resulting series of x-values. The plots of the x-values and the phase space (x at time t plotted against x at time $t + 1$) are shown in Figure 4.3.

The Duffing-Van der Pol Oscillator

The Duffing-Van der Pol oscillator [van der Pol and van der Mark, 1927] is another well studied dynamical system. Its equations are

$$\frac{dx}{dt} = y, \tag{4.57}$$

$$\frac{dy}{dt} = \mu(1 - x^2)y - x^3 + f\cos z, \tag{4.58}$$

$$\frac{dz}{dt} = \omega, \tag{4.59}$$

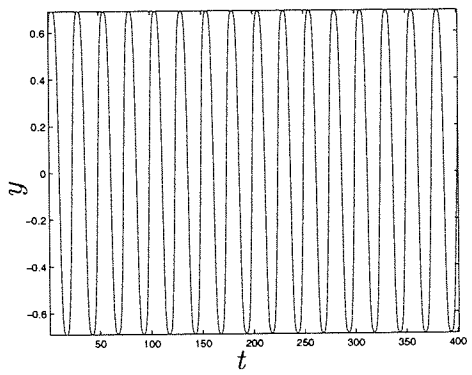
where μ, f and ω are dimensionless parameters. In contrast to the Lorenz system, we shall use a range of parameters to obtain several series with a varying amount of "disorder" but from the same generating functions for comparative purposes. The initial conditions were $x = 1$ and $y = z = 0$ and the series was sampled at a time epoch of 0.3 as these are the standard values [Henry et al., 2001]. The parameter values are shown in Table 4.1 along with the corresponding maximal Lyapunov exponent λ, calculated using the algorithm in [Wolf et al., 1985] run for 10000 iterations. The Lyapunov exponents in Table 4.1 indicate that it

System	μ	f	ω	λ
DVP1	0	0	0	0.0000
DVP2	0.2	0	0	0.0000
DVP3	0.2	1	0.9	-0.0002
DVP4	0.2	1	0.94	0.0235

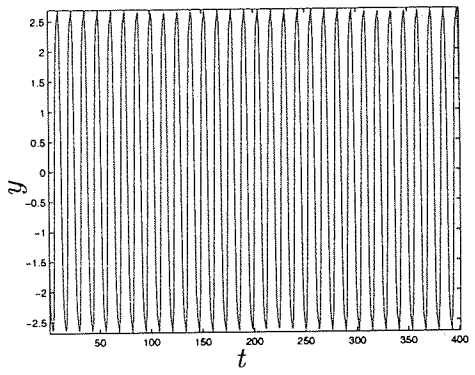
Table 4.1: Parameters used in the creation of four Duffing-Van der Pol oscillator systems.

is only DVP4 that exhibits any chaotic behaviour (albeit mildly).

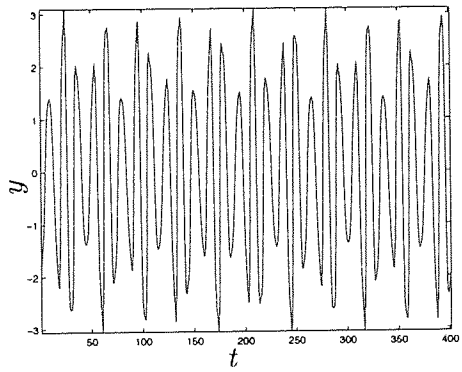
For the creation of the actual time series, a burn in of 1000 values was used so the phase orbit was settled in and we recorded the next 1000 values. In our case, we are only investigating the y-value. The plots of the y-value against time for all the systems are shown in Figure 4.4 and the phase space representations are shown in Figure 4.5.



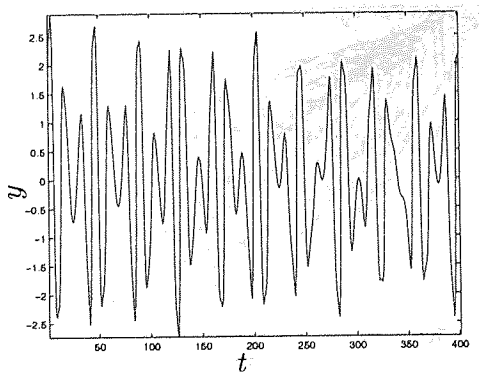
(a) DVP1



(b) DVP2

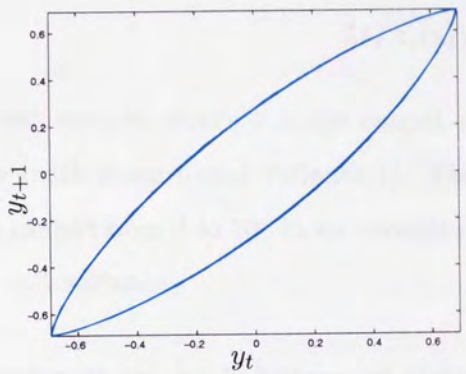


(c) DVP3

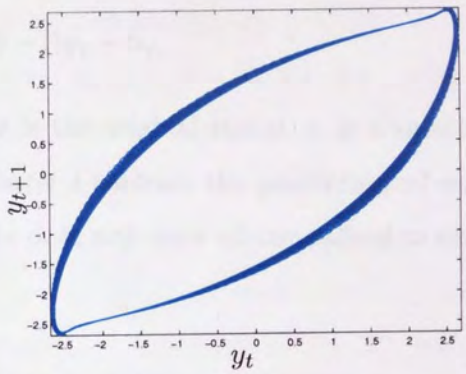


(d) DVP4

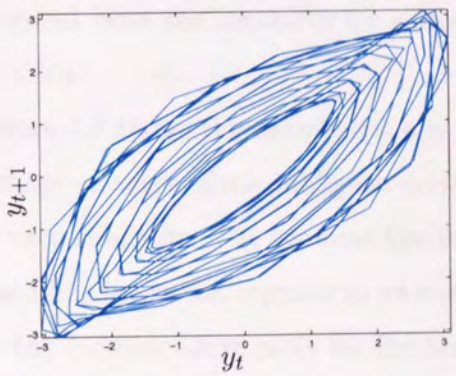
Figure 4.4: The y values for all four Duffing-Van der Pol oscillator systems.



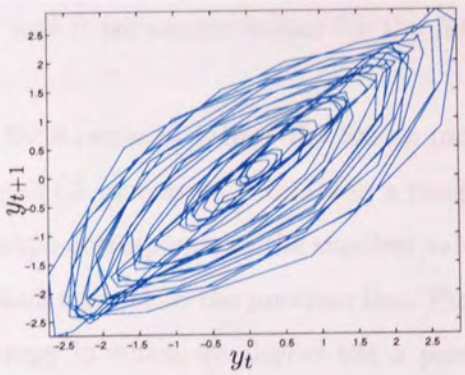
(a) DVP1



(b) DVP2



(c) DVP3



(d) DVP4

Figure 4.5: Phase space representation for all four Duffing-Van der Pol oscillator systems.

4.3.1 Robustness to Noise

To compare the measures, we need to gauge how robust they are under several conditions. Firstly, we need to know how the measures perform in the presence of noise.

For these tests we used both the Lorenz x-value series, and the DVP3 series. To investigate the effect of increasing noise on the system, Gaussian noise was added as a percentage of the overall signal. This can be formalised as

$$MIX_l(\mathbf{x}_t) = (100 - l)\mathbf{y}_t + l\epsilon_t, \tag{4.60}$$

for each sample, where x is the output signal, $\mathbf{y}$ is the original signal, ϵ_t is Gaussian white noise (with mean 0 and variance 1). The parameter l controls the percentage of noise; the tests ranged from 0 to 100 in increments of 1. The data sets were all normalised to zero mean and unit variance.

Robustness to the Influence of Noise for Different Values of r and σ

What we are looking for when we investigate the robustness of the entropy measures when applied with different values of r and σ is unusual or inconsistent behaviour. Firstly, we investigated both the measures for a range of r and σ parameter values for the full series ($N = 1000$).

Figure 4.6 shows the sample entropy results for a range of values of r for an increasing percentage of noise. Each subfigure shows a range of 5 increasing r values in a range given under each subfigure. The topmost line in each graph corresponds to the smallest values of r and each line below corresponds to an r value of 0.02 more than the previous line. Figure 4.7 shows the corresponding plots for the kernel entropy in which we altered the σ parameter. Also, it is different in that the lower line on each graph corresponds to the lower σ value and the higher lines correspond to higher values. This is possibly due to the dependence on σ in the upper bound (Equation 4.45).

Both sets of entropy values show the trend of increasing to an asymptotic upper limit as the noise increases which is as we would expect as the random element begins to dominate the signal.

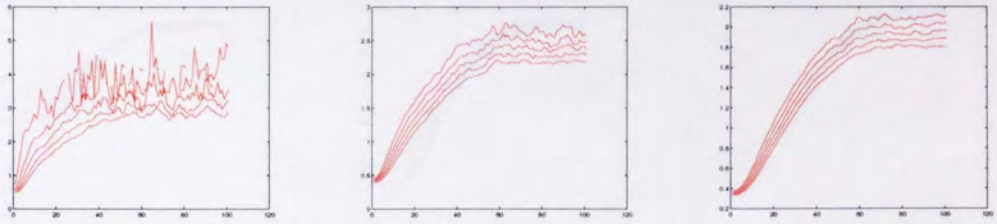
In Figure 4.6(a), we can see that for smaller values of r, the sample entropy values are erratic. This continues from $r = 0.02$ where it is very erratic (and discontinuous) though to $r = 0.3$. This is of particular interest as this covers the parameter values that Pincus recommends (r = [0.18 - 0.25] for normalised data). These aberrations are due to the small

number of matches when the tolerance is low so as the noise increases, the overall percentage of matches can change considerably. From Figure 4.7, we can see that kernel entropy does not share this trait due to the smooth kernel function; the values are continuous throughout the noise range for all values of σ.

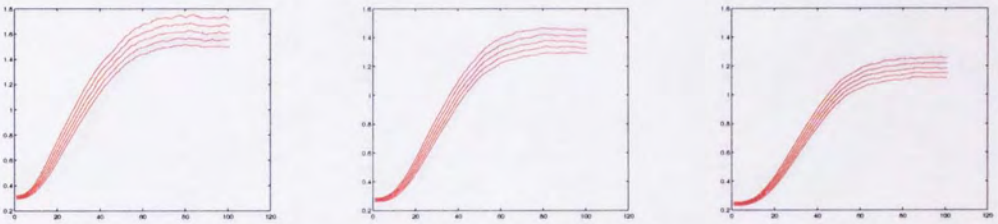
Another difference between the two measures is that the entropy values are closer together in the sample entropy than kernel entropy for low noise, and further apart for high noise. This is due to the smooth Gaussian kernel giving points further away a smaller probability compared to the square kernel assigning an equal probability to any points within it. In a very ordered system, all the points in the system will follow a simple path around the attractor(s). Therefore, as long as the square kernel is of such a size that it can encompass most of the noise variance, it will give similar results at low noise levels. However, at higher noise levels, the size of the kernel will make more of a difference in encompassing more of phase space, hence a small increase in the kernel size will affect the result disproportionately.

With kernel entropy, the opposite happens. All the points are assigned a probability density, with those within the same phase path receiving a higher probability. However, although the time domain epochs are equally spaced, the phase space ones are not, leading to a lower value of kernel entropy as the phase path inevitably deviates away from the one under consideration due to the distance between points in discrete phase space. Thus, an increase in the size of the encompassing Gaussian kernel results in more points in the phase being assigned a similar value as the distribution around the point becomes less sharply 'peaked'. In an ordered system, the flatter distribution will result in a more even probability assignment to the phase space points that are on the same phase path which is why a change in σ has more effect than when the noise increases and most of the points will be assigned a small probability regardless of the kernel size.

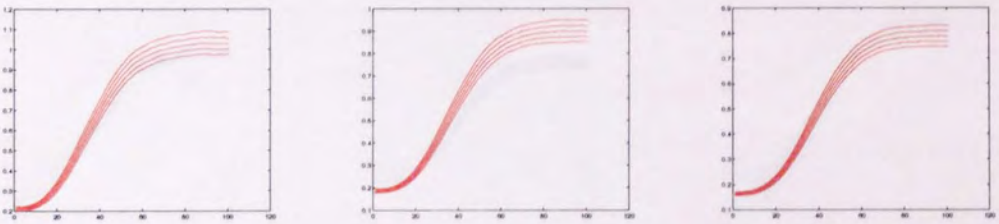
For the the Duffing-Van der Pol oscillator y-value series the results are similar. In Figure 4.8, we can see that the results for sample entropy are still erratic and not defined for smaller values of r. Kernel entropy, in Figure 4.9, also behaves slightly differently; whilst still not as erratic as sample entropy, for some of the smaller values of σ (Figure 4.9(a)), there is some inconsistent behaviour as the kernel entropy value appears to approach a plateau and then wavers slightly. This indicates that, although kernel entropy is robust to small σ values, some care should be taken in their choice. However, in comparison to sample entropy, which in this case has inconsistent behaviour at all values of r, it is much more consistent.



(a) $r = \{0.02 - 0.1\}$ (b) $r = \{0.12 - 0.2\}$ (c) $r = \{0.22 - 0.3\}$

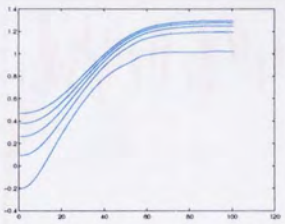


(d) $r = \{0.32 - 0.4\}$ (e) $r = \{0.42 - 0.5\}$ (f) $r = \{0.52 - 0.6\}$

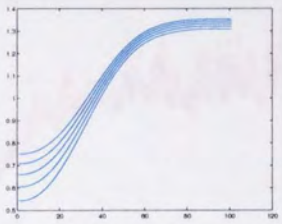


(g) $r = \{0.62 - 0.7\}$ (h) $r = \{0.72 - 0.8\}$ (i) $r = \{0.82 - 0.9\}$

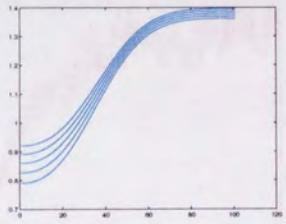
Figure 4.6: The sample entropy values (y-axis) for the x value of the Lorenz series with noise term l (x-axis) for a range of r values.



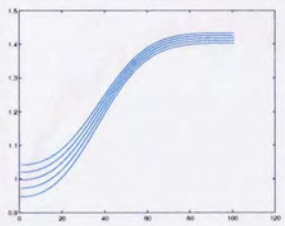
(a) $\sigma = \{0.02 - 0.1\}$



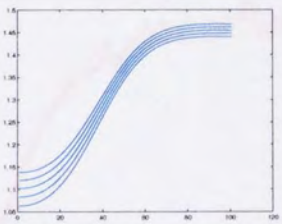
(b) $\sigma = \{0.12 - 0.2\}$



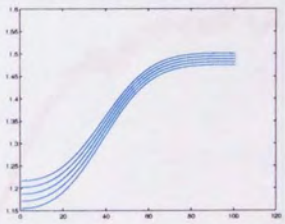
(c) $\sigma = \{0.22 - 0.3\}$



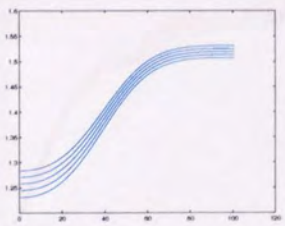
(d) $\sigma = \{0.32 - 0.4\}$



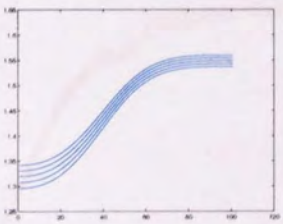
(e) $\sigma = \{0.42 - 0.5\}$



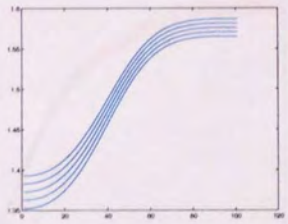
(f) $\sigma = \{0.52 - 0.6\}$



(g) $\sigma = \{0.62 - 0.7\}$

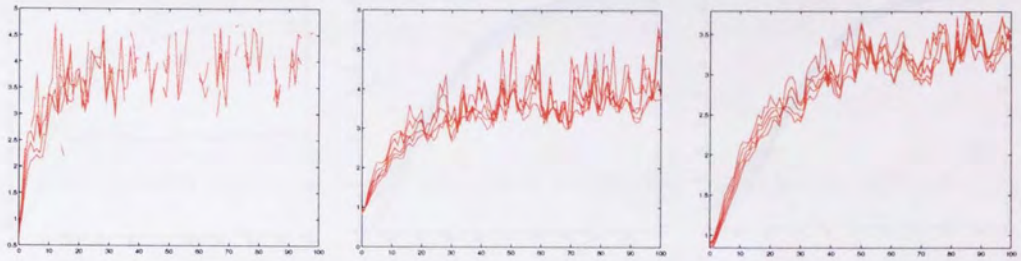


(h) $\sigma = \{0.72 - 0.8\}$

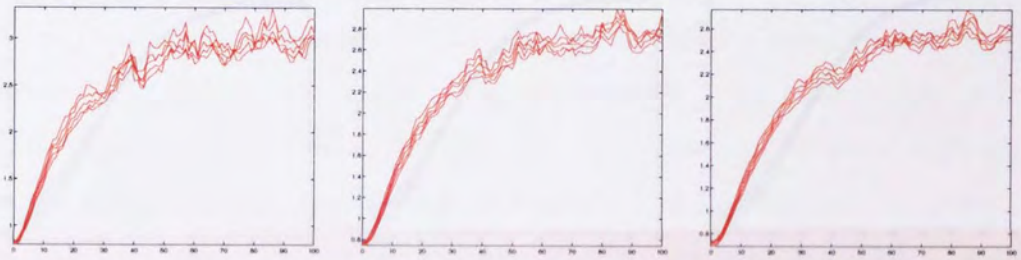


(i) $\sigma = \{0.82 - 0.9\}$

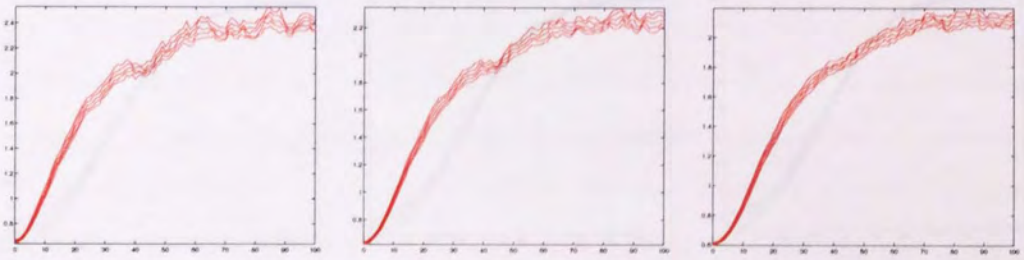
Figure 4.7: The kernel entropy values (y-axis) for the x value of the Lorenz series with noise term l (x-axis) for a range of σ values. Bayesian bandwidth selection suggests $\sigma = 0.1244$.



(a) $r = \{0.02 - 0.1\}$ (b) $r = \{0.12 - 0.2\}$ (c) $r = \{0.22 - 0.3\}$

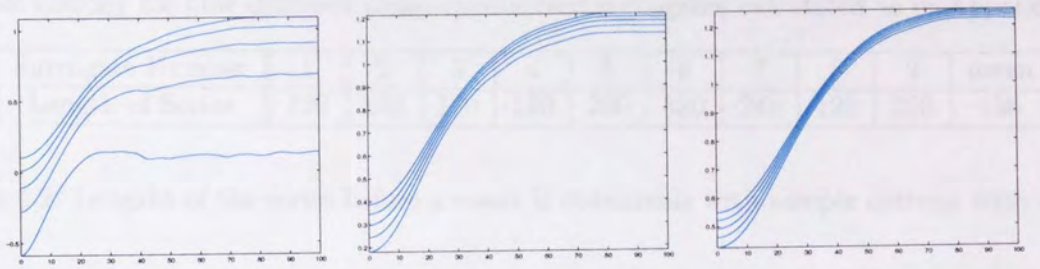


(d) $r = \{0.32 - 0.4\}$ (e) $r = \{0.42 - 0.5\}$ (f) $r = \{0.52 - 0.6\}$

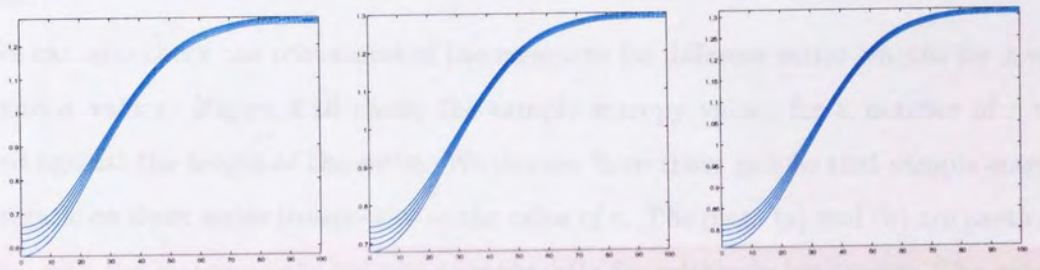


(g) $r = \{0.62 - 0.7\}$ (h) $r = \{0.72 - 0.8\}$ (i) $r = \{0.82 - 0.9\}$

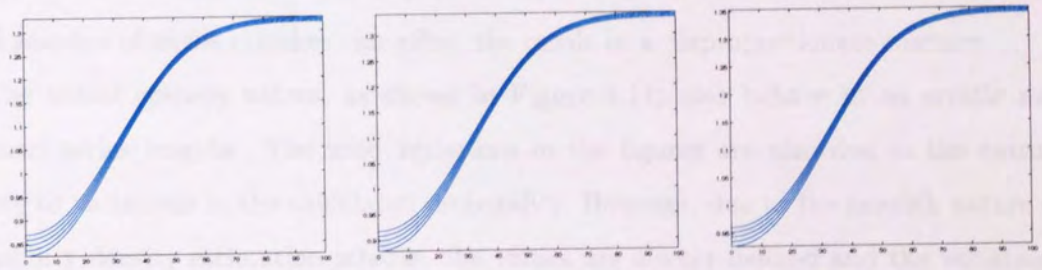
Figure 4.8: The sample entropy values (y-axis) for the x value of the series DVP3 with noise term l (x-axis) for a range of r values.



(a) $\sigma = \{0.02 - 0.1\}$ (b) $\sigma = \{0.12 - 0.2\}$ (c) $\sigma = \{0.22 - 0.3\}$



(d) $\sigma = \{0.32 - 0.4\}$ (e) $\sigma = \{0.42 - 0.5\}$ (f) $\sigma = \{0.52 - 0.6\}$



(g) $\sigma = \{0.62 - 0.7\}$ (h) $\sigma = \{0.72 - 0.8\}$ (i) $\sigma = \{0.82 - 0.9\}$

Figure 4.9: The kernel entropy values (y-axis) for the x value of the series DVP3 with noise term l (x-axis) for a range of σ values. Bayesian bandwidth selection suggests $\sigma = 0.0691$.

Robustness to the Influence of Noise for Different Series Lengths N

The first thing we need to examine regarding the series length is where the sample entropy is defined. As sample entropy works on matches, if there are none, then the statistic is undefined. This is particularly pertinent for short data sets. To test this, we calculated sample entropy for nine different phase-randomised surrogates calculated as in Appendix B.

Surrogate Number	1	2	3	4	5	6	7	8	9	mean
Length of Series	120	140	170	150	260	320	240	120	250	196

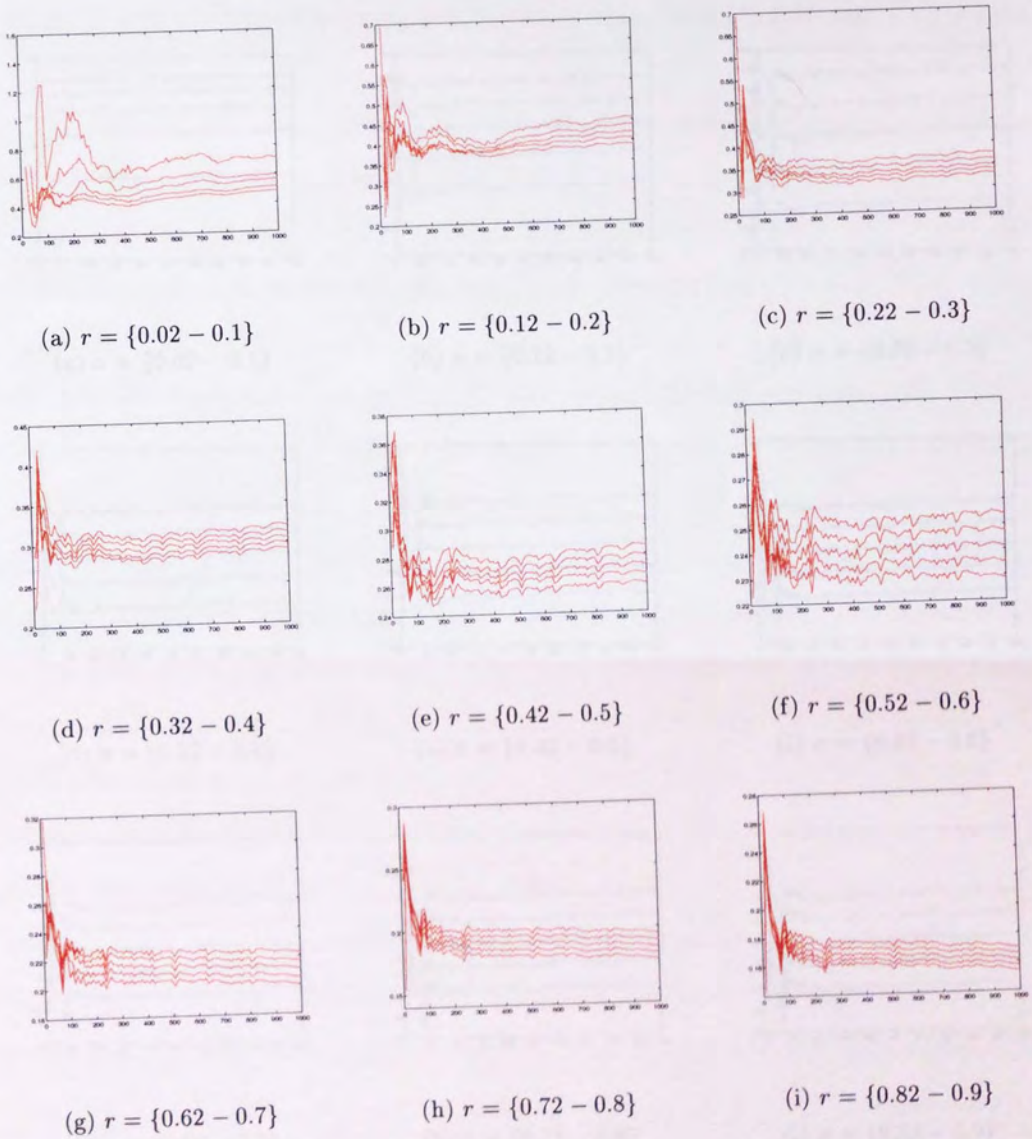
Table 4.2: Lengths of the series before a result is obtainable with sample entropy with r=0.1

Table 4.2 shows that sample entropy is often not defined for short data sets when N is relatively small. This is not a problem with kernel entropy as it is defined for all series lengths.

We can also check the robustness of the measures for different series lengths for a variety of r and σ values. Figure 4.10 shows the sample entropy values for a number of r values plotted against the length of the series. We can see from these graphs that sample entropy is very erratic on short series irrespective of the value of r. The plots (a) and (b) are particularly erratic, with the measure only behaving consistently for relatively long series. The values do level out as the series lengths increase although there are a number of small fluctuations for all values of r. This is to be expected, however, as the nature of the tolerance means that a small number of extra matches can affect the result in a disproportionate manner.

The kernel entropy values, as shown in Figure 4.11, also behave in an erratic manner for short series lengths. The mild variations in the figures are also due to the extra data leading to variations in the calculated probability. However, due to the smooth nature of the probability density estimation scheme, the values are always defined and the variations are smaller than those of the square kernel sample entropy technique.

This means that it is more consistent for varying series lengths and is important when the series lengths are not constant, say if you are comparing two series of different lengths. Also, the kernel entropy values for small σ are remarkably stable at low series lengths. This may indicate that the optimal value for σ is smaller than the optimal value of r in sample entropy.

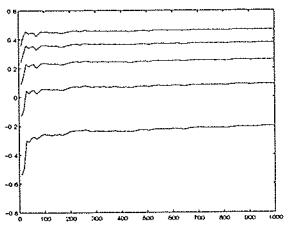


(a) $r = \{0.02 - 0.1\}$ (b) $r = \{0.12 - 0.2\}$ (c) $r = \{0.22 - 0.3\}$

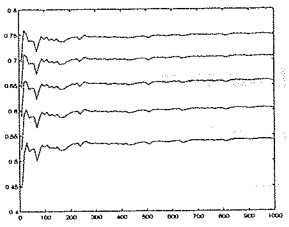
(d) $r = \{0.32 - 0.4\}$ (e) $r = \{0.42 - 0.5\}$ (f) $r = \{0.52 - 0.6\}$

(g) $r = \{0.62 - 0.7\}$ (h) $r = \{0.72 - 0.8\}$ (i) $r = \{0.82 - 0.9\}$

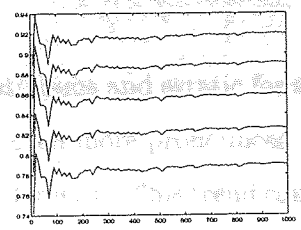
Figure 4.10: The sample entropy values (y-axis) for the x value of the Lorenz series with increasing series length (x-axis) for a range of r values.



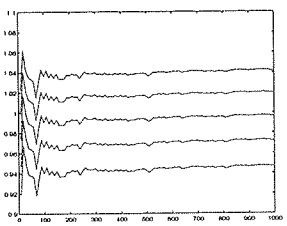
(a) $\sigma = \{0.02 - 0.1\}$



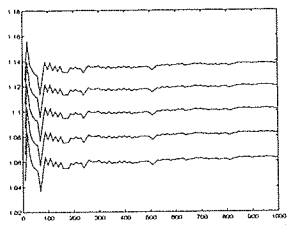
(b) $\sigma = \{0.12 - 0.2\}$



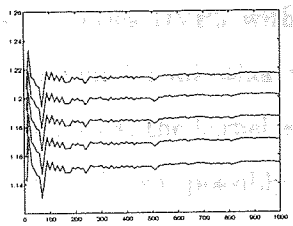
(c) $\sigma = \{0.22 - 0.3\}$



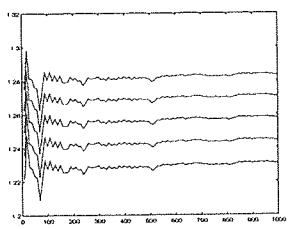
(d) $\sigma = \{0.32 - 0.4\}$



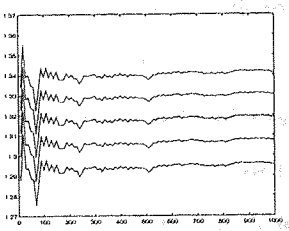
(e) $\sigma = \{0.42 - 0.5\}$



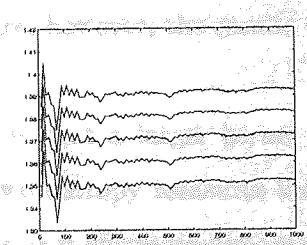
(f) $\sigma = \{0.52 - 0.6\}$



(g) $\sigma = \{0.62 - 0.7\}$



(h) $\sigma = \{0.72 - 0.8\}$



(i) $\sigma = \{0.82 - 0.9\}$

Figure 4.11: The kernel entropy values (y-axis) for the x value of the Lorenz series with increasing series length (x-axis) for a range of σ values.

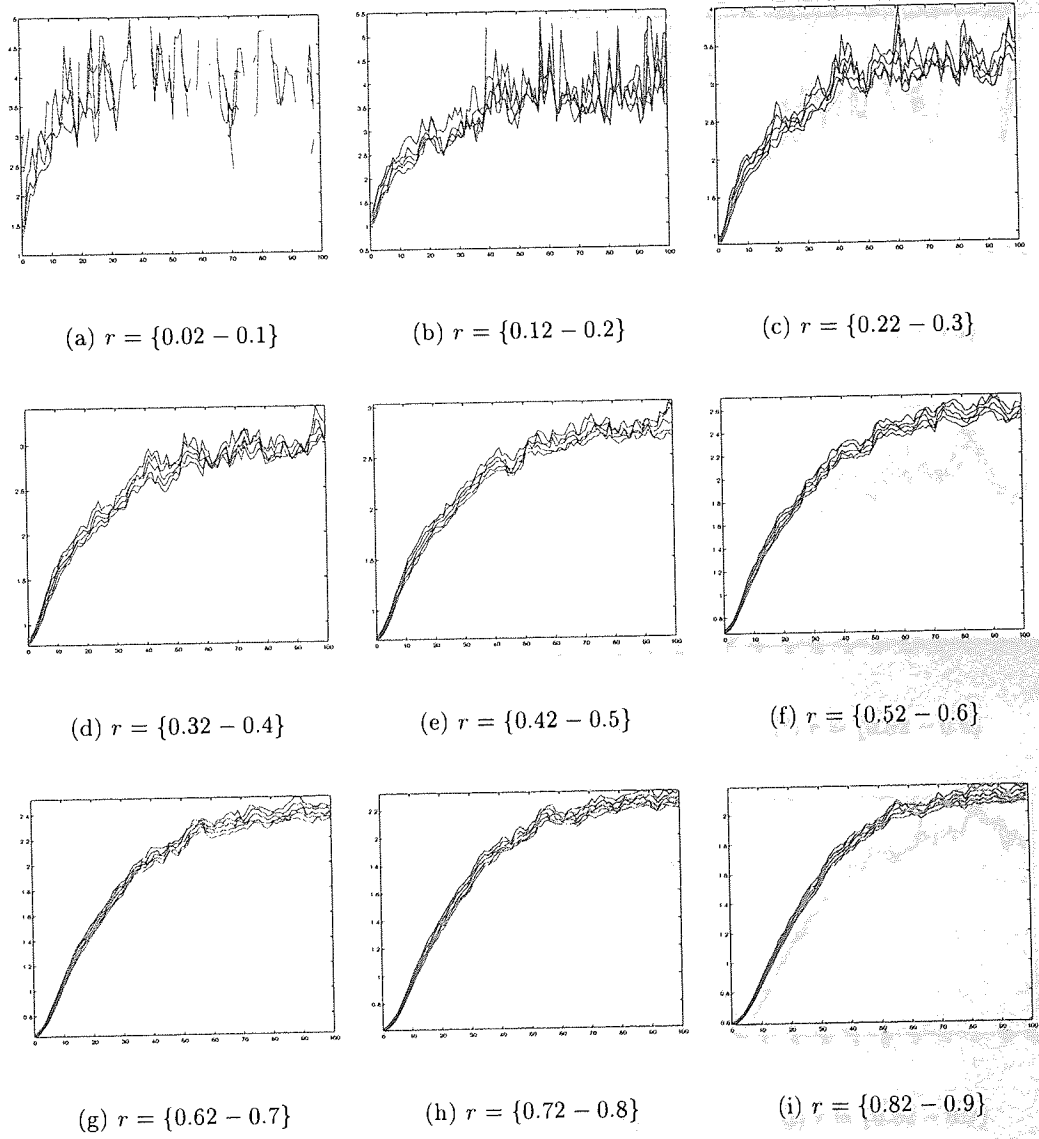
Robustness to the Influence of Noise for Different Window Sizes

We can investigate how the series behave with different m values by comparing plots for ranges of r and σ with different values of m. For this we use series DVP4, and plot the ranges for $m = 2, 3, 4$. The results for the sample entropy are shown in Figures 4.12, 4.13 and 4.14 and the results of kernel entropy are shown in Figures 4.15, 4.16 and 4.17 for $m = 2, 3$ and 4 respectively.

As with series DVP3 above, at $m = 2$, the results are discontinuous and erratic for sample entropy, especially at lower r values. With $m = 3$, this effect is even more pronounced with a larger number of discontinuities and the signal is erratic even for larger r. This trend continues at $m = 4$, where at small values of r the results are next to useless for any reasonable noise value. At this m value, the series is very erratic; although at higher r values, the general trend of the results are more consistent and well defined for low noise values.

At $m = 2$, kernel entropy behaves similarly to the results for series DVP3 with some slight inconsistent behaviour at small σ but with consistent behaviour for all other values. At $m = 3$, the small σ values behave inconsistently, as the noise increases, the kernel entropy values actually decrease which is contrary to what would be expected. This is possibly due to the sharply peaked kernels being unable to accurate encompass the higher dimensional data. However, at the range $\sigma = 0.12 - 0.2$ the behaviour is more consistent and by $\sigma = 0.32 - 0.4$ it is behaving completely consistently. Again, a similar pattern occurs at $m = 4$, with the small σ values displaying inconsistent behaviour as before. Here, however, the results do not start behaving consistently until $\sigma = 0.52 - 0.6$.

The first thing we can notice about both sets of results is that a larger kernel size is needed for larger values of m. Comparing the results of the two entropy measures indicates that kernel entropy is far more consistent for differing m values as it shows no discontinuities and is less erratic. It would also appear that kernel entropy is able to encompass a higher percentage of noise as the result becomes unreliable with sample entropy whilst kernel entropy remains consistent for almost all noise values as long as the σ value is chosen appropriately. In fact, from these results, it would be difficult to rely on any results produced by sample entropy at higher values of m.

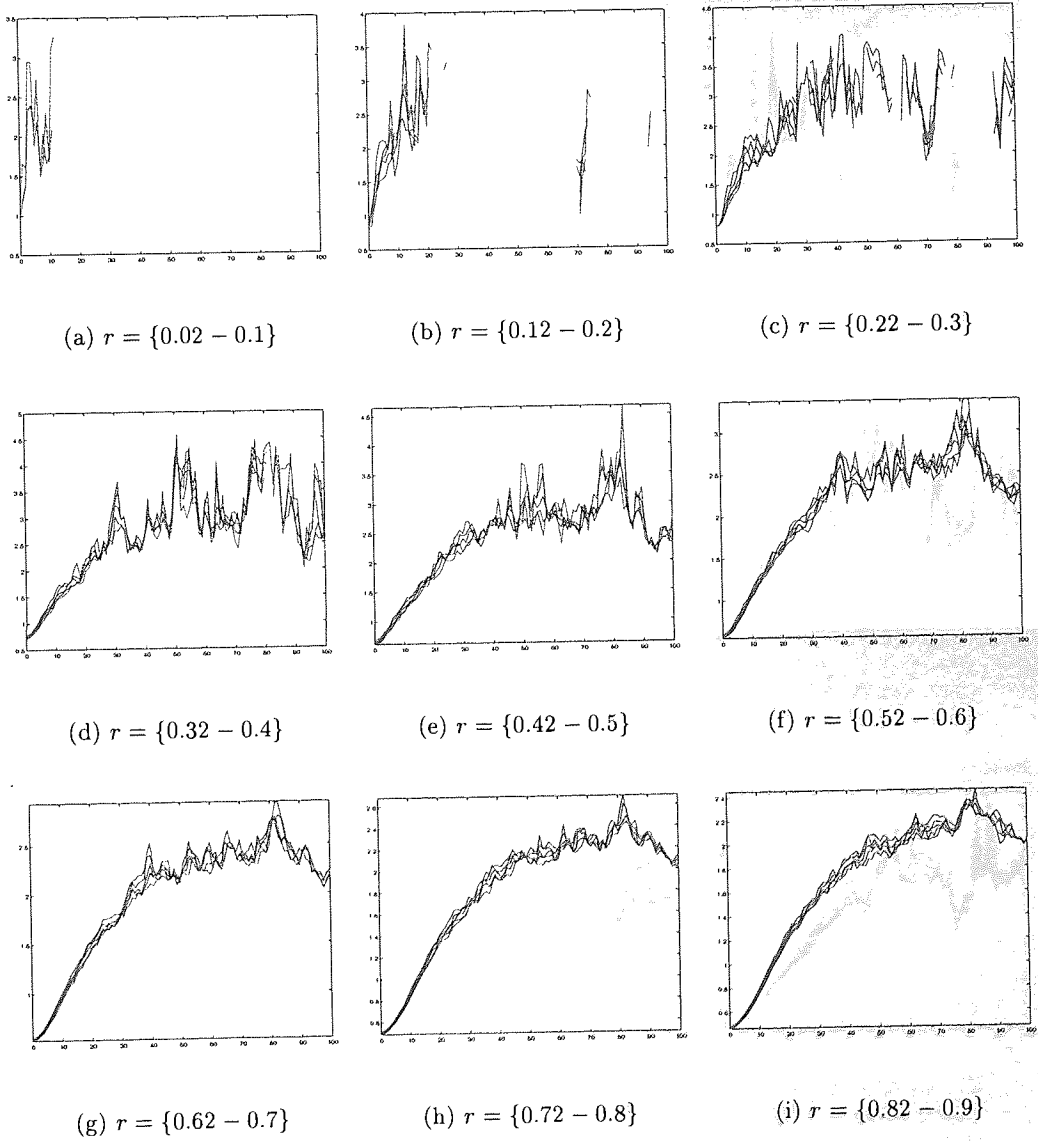


(a) $r = \{0.02 - 0.1\}$ (b) $r = \{0.12 - 0.2\}$ (c) $r = \{0.22 - 0.3\}$

(d) $r = \{0.32 - 0.4\}$ (e) $r = \{0.42 - 0.5\}$ (f) $r = \{0.52 - 0.6\}$

(g) $r = \{0.62 - 0.7\}$ (h) $r = \{0.72 - 0.8\}$ (i) $r = \{0.82 - 0.9\}$

Figure 4.12: The sample entropy values (y-axis) for the x value of series DVP4 with noise term l (x-axis) for a range of r values at $m = 2$.

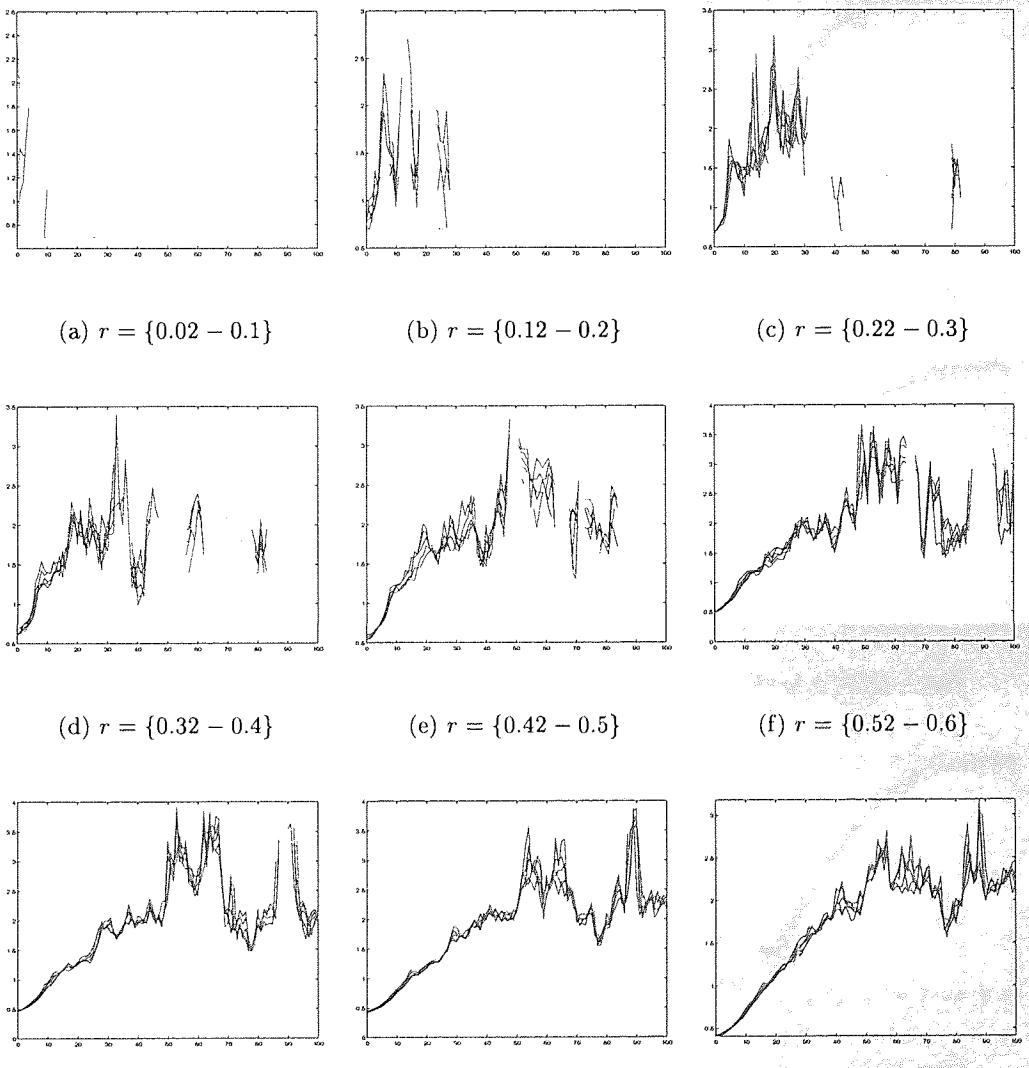


(a) $r = \{0.02 - 0.1\}$ (b) $r = \{0.12 - 0.2\}$ (c) $r = \{0.22 - 0.3\}$

(d) $r = \{0.32 - 0.4\}$ (e) $r = \{0.42 - 0.5\}$ (f) $r = \{0.52 - 0.6\}$

(g) $r = \{0.62 - 0.7\}$ (h) $r = \{0.72 - 0.8\}$ (i) $r = \{0.82 - 0.9\}$

Figure 4.13: The sample entropy values (y-axis) for the x value of series DVP4 with noise term l (x-axis) for a range of r values at $m = 3$.

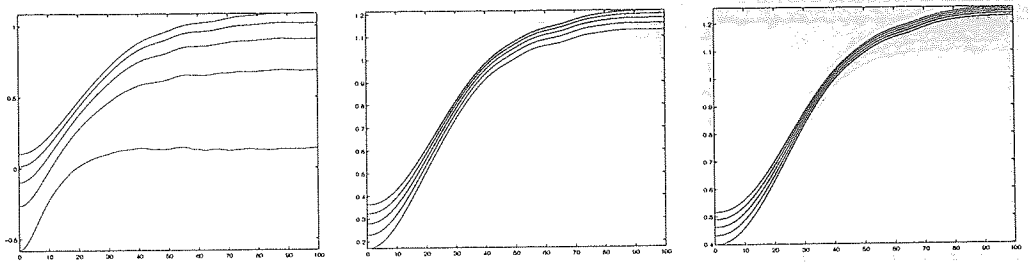


(a) $r = \{0.02 - 0.1\}$ (b) $r = \{0.12 - 0.2\}$ (c) $r = \{0.22 - 0.3\}$

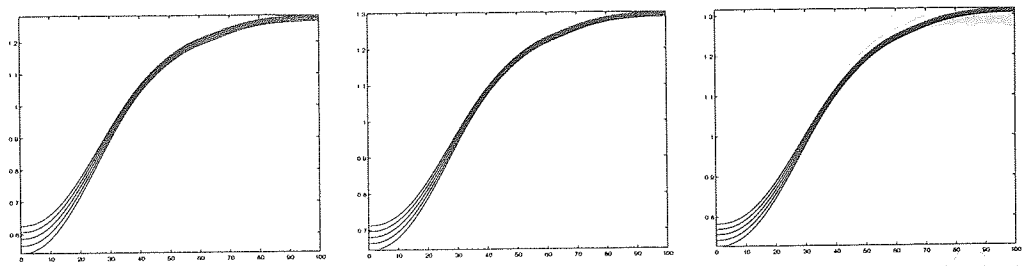
(d) $r = \{0.32 - 0.4\}$ (e) $r = \{0.42 - 0.5\}$ (f) $r = \{0.52 - 0.6\}$

(g) $r = \{0.62 - 0.7\}$ (h) $r = \{0.72 - 0.8\}$ (i) $r = \{0.82 - 0.9\}$

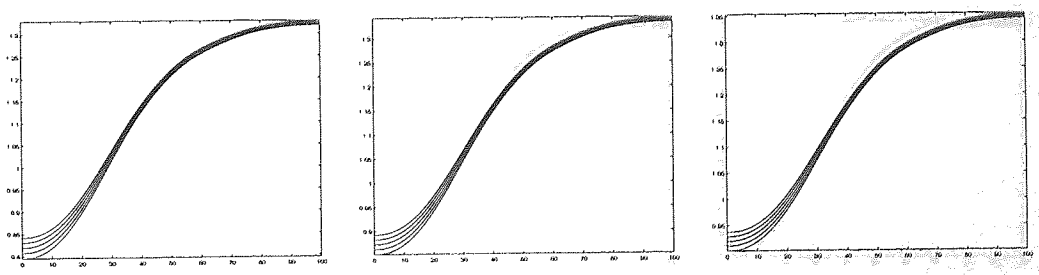
Figure 4.14: The sample entropy values (y-axis) for the x value of series DVP4 with noise term l (x-axis) for a range of r values at $m = 4$.



(a) $\sigma = \{0.02 - 0.1\}$ (b) $\sigma = \{0.12 - 0.2\}$ (c) $\sigma = \{0.22 - 0.3\}$

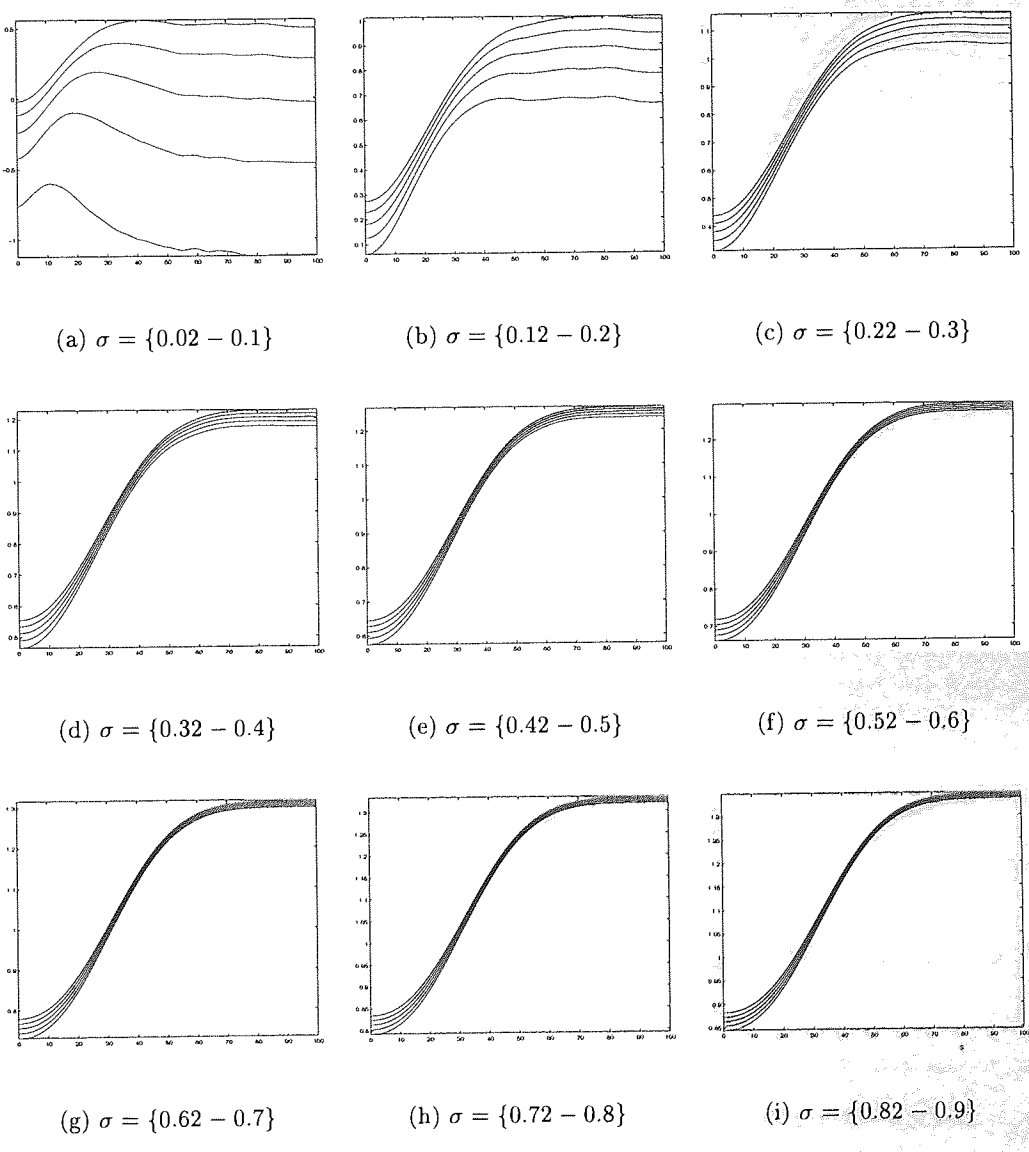


(d) $\sigma = \{0.32 - 0.4\}$ (e) $\sigma = \{0.42 - 0.5\}$ (f) $\sigma = \{0.52 - 0.6\}$



(g) $\sigma = \{0.62 - 0.7\}$ (h) $\sigma = \{0.72 - 0.8\}$ (i) $\sigma = \{0.82 - 0.9\}$

Figure 4.15: The kernel entropy values (y-axis) for the x value of series DVP4 with noise term l (x-axis) for a range of σ values for $m = 2$. Bayesian bandwidth selection suggests $\sigma = 0.1314$.

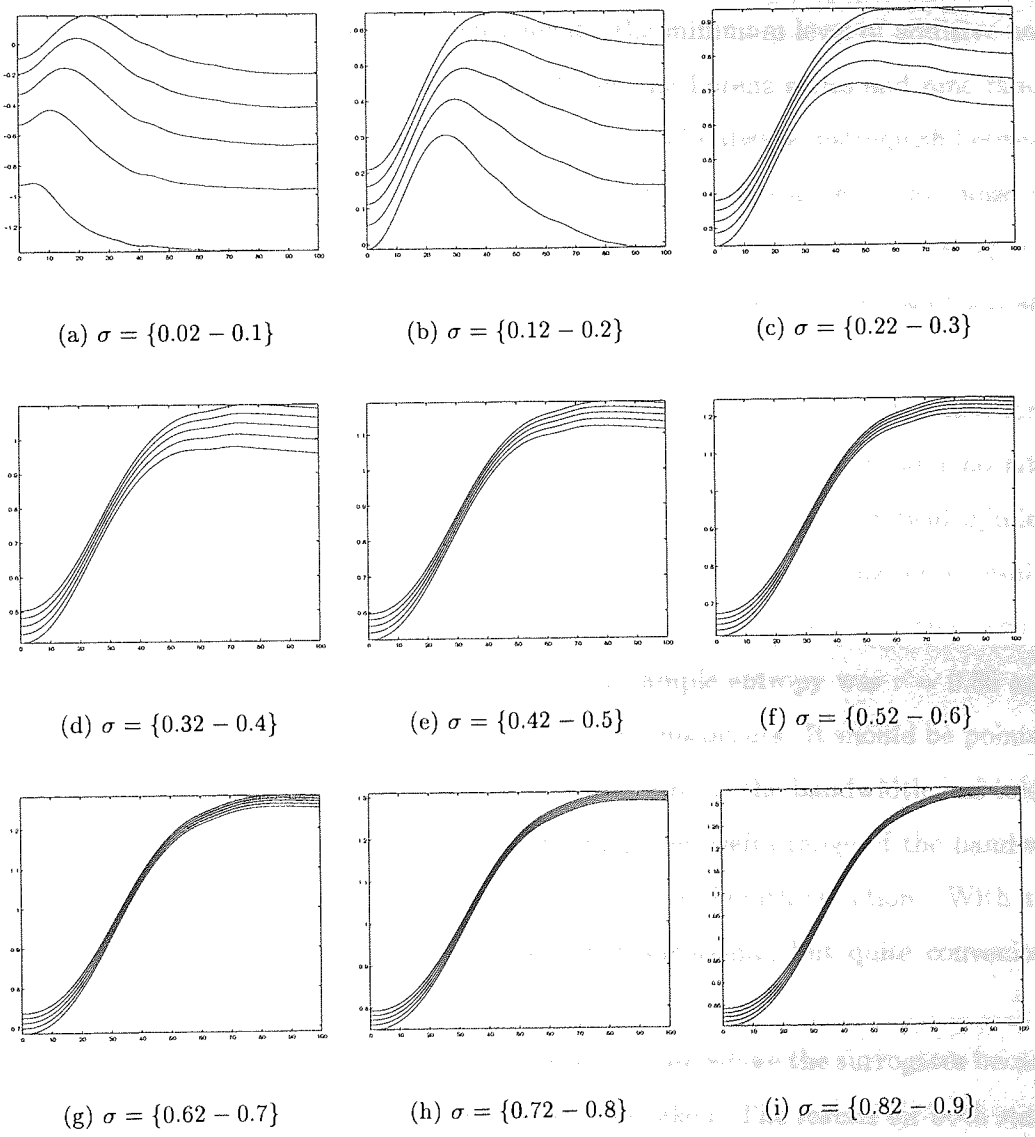


(a) $\sigma = \{0.02 - 0.1\}$ (b) $\sigma = \{0.12 - 0.2\}$ (c) $\sigma = \{0.22 - 0.3\}$

(d) $\sigma = \{0.32 - 0.4\}$ (e) $\sigma = \{0.42 - 0.5\}$ (f) $\sigma = \{0.52 - 0.6\}$

(g) $\sigma = \{0.62 - 0.7\}$ (h) $\sigma = \{0.72 - 0.8\}$ (i) $\sigma = \{0.82 - 0.9\}$

Figure 4.16: The kernel entropy values (y-axis) for the x value of series DVP4 with noise term l (x-axis) for a range of σ values for $m = 3$. Bayesian bandwidth selection suggests $\sigma = 0.1331$.



(a) $\sigma = \{0.02 - 0.1\}$ (b) $\sigma = \{0.12 - 0.2\}$ (c) $\sigma = \{0.22 - 0.3\}$

(d) $\sigma = \{0.32 - 0.4\}$ (e) $\sigma = \{0.42 - 0.5\}$ (f) $\sigma = \{0.52 - 0.6\}$

(g) $\sigma = \{0.62 - 0.7\}$ (h) $\sigma = \{0.72 - 0.8\}$ (i) $\sigma = \{0.82 - 0.9\}$

Figure 4.17: The kernel entropy values (y-axis) for the x value of series DVP4 with noise term l (x-axis) for a range of σ values for $m = 4$. Bayesian bandwidth selection suggests $\sigma = 0.1332$.

4.3.2 Distinguishing Between Ordered and Disordered Systems

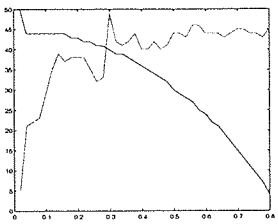
In this case, we are investigating how good the measures are at distinguishing an *ordered* (or deterministic) from a *disordered* (non-deterministic) system. For this, we investigate how well the measures discriminate between the types of system under increasing noise.

To evaluate this, we found, through experiments, the minimum level of additive noise at which the measures can no longer distinguish between the Lorenz series and nine randomly generated phase-randomised surrogates; both measures could always distinguish between the shuffled series and either the Lorenz or a phase-randomised surrogate for all noise values which shows that the an defined phase path (random or otherwise) is easily distinguishable from a completely disordered one by these measures. Hence, results and discussion are not given.

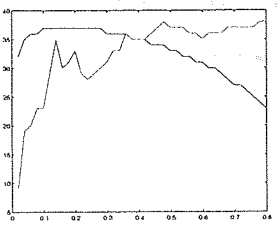
In the first instance, kernel entropy and sample entropy values were calculated for both series for $\sigma, r \in \{0.02, 0.04, \ldots, 2.5\}$ for the series with 1000 data points and no additive noise. This was done to find the sensitivity of the measures to the bandwidth/tolerance values by finding the smallest bandwidth/tolerance value when the Lorenz series resulted in a higher entropy value than each of the surrogates. The average value where this occurs was taken for comparison. The average tolerance value in sample entropy was $r = 0.86$ and the average bandwidth in kernel entropy was $\sigma = 0.88$ when this occurs. It should be pointed out that neither one should be considered "better" at this point as the bandwidth and tolerance are inherently different and we are merely determining the useful range of the bandwidths, although this does highlight the importance of proper bandwidth selection. With this in mind, the closeness of these two values is somewhat surprising, but quite convenient for future comparison.

Increasing levels of noise were added and the level of noise where the surrogates became indistinguishable from the deterministic Lorenz series was taken. The results for both statistics applied to each surrogate are plotted in Figure 4.18.

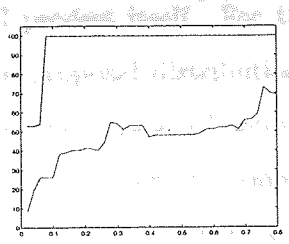
Looking at each measure individually, sample entropy, indicated by the red line, is generally erratic and inconsistent as the bandwidth increases. The kernel entropy behaves more consistently, often with a 'peak' which indicates the optimal bandwidth. Also, kernel entropy was able to discern the surrogate from the deterministic series more often than sample entropy. This gives it greater reliability and therefore more confidence can be attributed to the result.



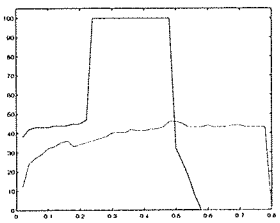
(a) Surrogate 1



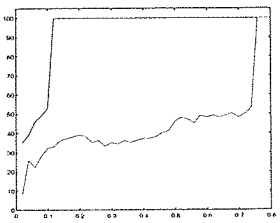
(b) Surrogate 2



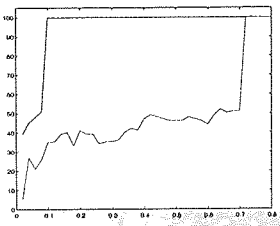
(c) Surrogate 3



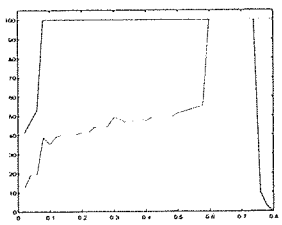
(d) Surrogate 4



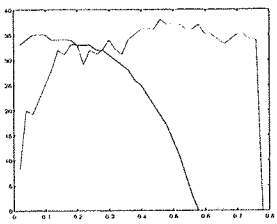
(e) Surrogate 5



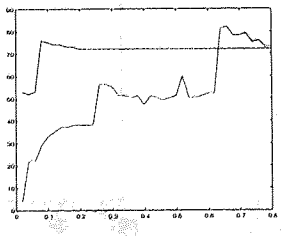
(f) Surrogate 6



(g) Surrogate 7



(h) Surrogate 8



(i) Surrogate 9

Figure 4.18: The noise level where the Lorentz series and a phase-randomised surrogate become indistinguishable by kernel entropy (blue) and sample entropy (red).

4.3.3 Effectiveness of the Bandwidth Selection Procedure

As there is no analogous method for the selection of r that can be used for sample entropy, the selection scheme can only really be evaluated for robustness and consistency. This means that to judge the robustness and consistency of the selection scheme we can only compare it with itself. This can be done in two ways.

First, we need to investigate the properties of the MCMC process itself. For this we started 5 different experiments, with different variances of the proposal distribution each with uniformly randomised start positions between 0 and 20, on the noiseless Lorenz series where we discarded a burn-in period of 500 iterations and kept the next 500 iterations. The variances of the proposal distributions were $\gamma = 0.01, 0.04, 0.09, 0.16, 0.25$. The σ values for the last 500 iterations for each of these proposal variances can be seen in Figure 4.20. Figure 4.19 shows the first 100 iterations of the burn-in period. It can be seen that regardless of start point or proposal variance, these chains converge to a value of $\sigma \approx 0.44$. Aside from

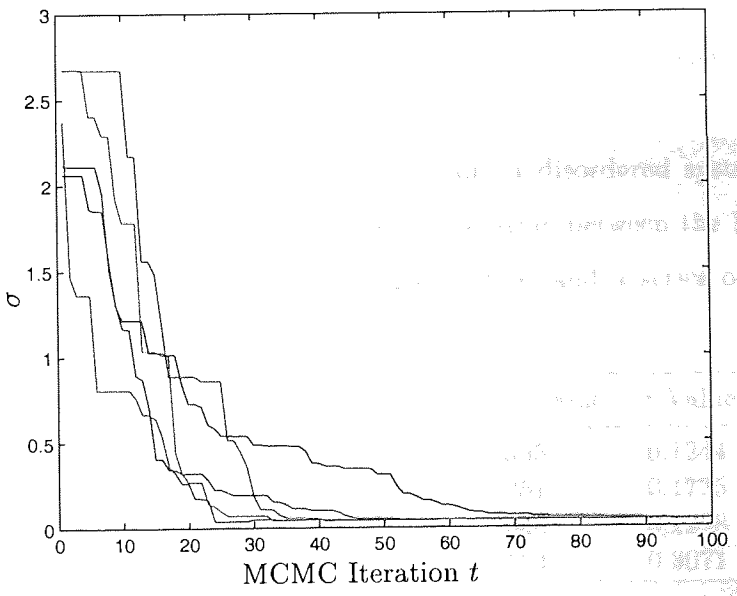


Figure 4.19: The MCMC burn-in period arising from a range of proposal distribution variances γ for the estimation of the bandwidth for the Lorenz series.

this, we can plot histograms of the resulting distributions of the MCMC process. These are shown in Figure 4.21. The resulting distributions for $\gamma = 0.01$ and $\gamma = 0.04$ do not exhibit any unusual behaviour like bimodality ($\gamma = 0.16$) and skew ($\gamma = 0.25$) so it was decided that using a proposal distribution variance of 0.04 and taking the mean of the last 500 σ values will give an acceptable approximation of the optimal bandwidth.

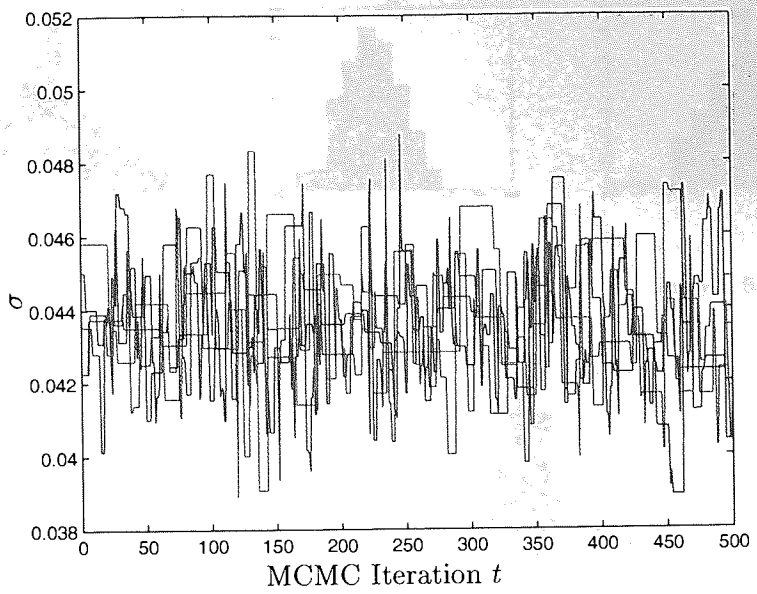


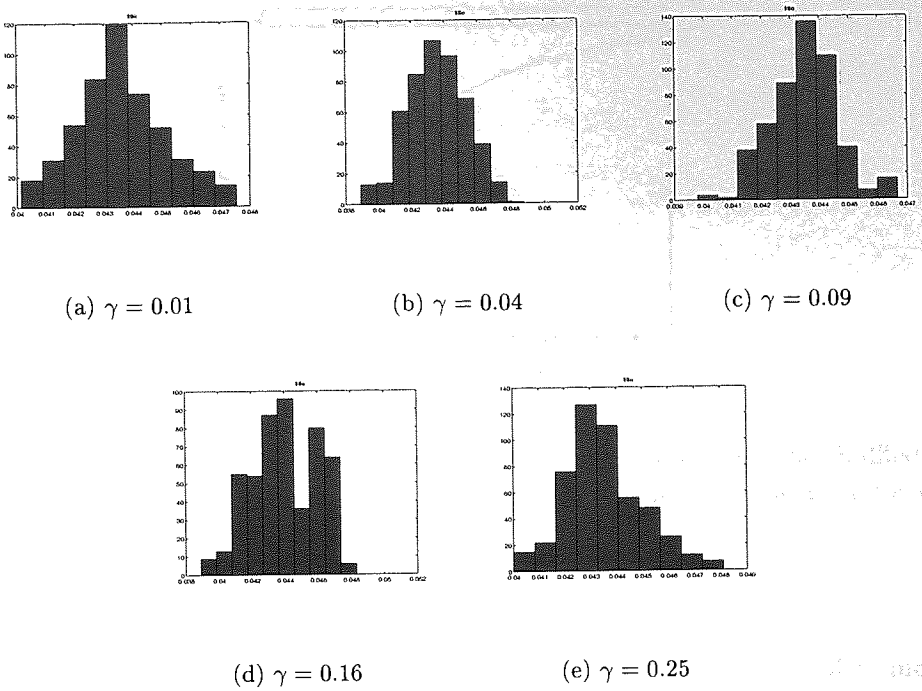
Figure 4.20: The MCMC chains arising from a range of proposal distribution variances γ for the estimation of the bandwidth for the Lorenz series.

Now that the question of the stability of the MCMC process is resolved, we applied kernel entropy with the bandwidth selection scheme to the Lorenz series with no noise to judge the effectiveness in distinguishing between an ordered and a disordered system. For this we investigate the effectiveness of the measure in differentiating between the Lorenz series, the phase randomised surrogate, the shuffled surrogate series, and a series of white Gaussian noise with the same mean and variance.

Data Set	*KernEn* Value	σ Value
Lorenz Series	0.8365	0.1244
Phase-Randomised Surrogate	1.3251	0.1725
Shuffled Surrogate	1.3048	0.1298
Gaussian Noise	2.1564	0.3071

Table 4.3: Kernel entropy and bandwidth values calculated using the Bayesian bandwidth selection scheme on the Lorenz series

The results of the bandwidth selection procedure, given in Table 4.3 were encouraging. The highly-ordered Lorenz series clearly gives a lower kernel entropy value than its surrogates. This is consistent with the results we have seen before, and the bandwidth selection returns a suitable value in each case. One point of note is that the bandwidth for the surrogate data is higher than either of the other two data sets which gives the surrogate a higher kernel



(a) $\gamma = 0.01$ (b) $\gamma = 0.04$ (c) $\gamma = 0.09$

(d) $\gamma = 0.16$ (e) $\gamma = 0.25$

Figure 4.21: The resulting distributions obtained by running the Bayesian bandwidth estimation for different proposal distribution variances γ.

entropy value than might be expected if the optimal bandwidth were constant. This is of great interest because, as we have seen previously, the surrogate can be more difficult to distinguish from the deterministic signal which indicates that this variable bandwidth approach may be effective in a real world application where the nature of a series is unknown.

Secondly, we applied kernel entropy with the bandwidth selection scheme to the series with increasing additive Gaussian noise to judge its consistency under the influence of noise. Figure 4.22 shows the kernel entropy values for the Lorenz series with increasing noise plotted with the values for the shuffled series. The plot shows that as the series becomes more noisy, it approaches the entropy values for the disordered series, which is exactly what you would expect in the statistic.

4.3.4 Quantifying the level of disorder in a system

Some applications require the comparison of the level of disorder. To investigate this we used the four Duffing-Van der Pol oscillator series as they display varying levels of disorder depending on the parameters used. Before applying the bandwidth selection, we should investigate how the statistics behave for different r and σ values when applied to each series. In this case we used the full series with no additive noise. Both statistics were calculated

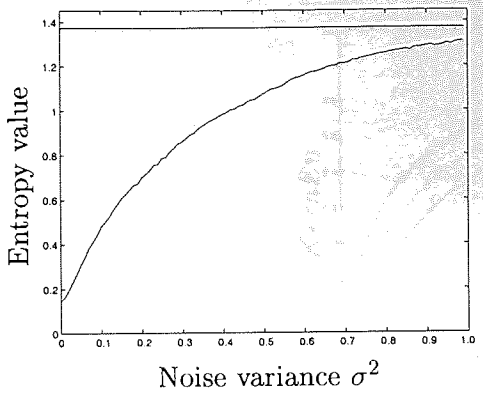


Figure 4.22: Kernel entropy calculated for the Lorenz series (blue) and the shuffled surrogate (black) for increasing noise variance. The bandwidth is calculated separately for each noise value with the Bayesian MCMC approach.

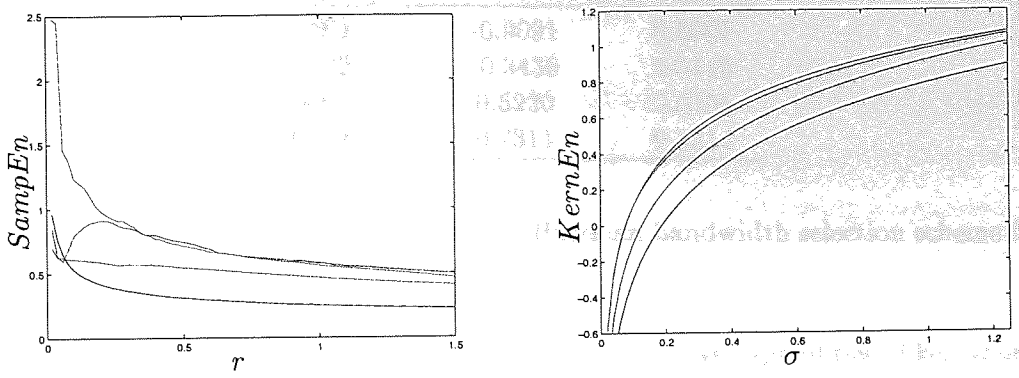
with $m = 2$ and r and σ between 0.02 and 1.5.

Figure 4.23 shows the results of both entropy measures for a range of r and σ values for all four Duffing-Van der Pol series. From the phase diagrams of the Duffing-Van der Pol oscillators (Figure 4.5) we would expect DVP1 to exhibit the least disorder, followed by DVP2, then DVP3 and DVP4. However, this is not always the case in these plots. The sample entropy values in Figure 4.23(a) are inconsistent at low r but, as we have seen, this is to be expected. What is more unusual is that both measures return a lower value for DVP4 than for DVP3 as r and σ get larger despite the previous assumption that DVP4 is more disordered than DVP3. For sample entropy, this occurs at $r > 0.35$ and for kernel entropy, this occurs at $\sigma > 0.088$ which is a very small kernel size. This indicates that DVP4 may not be as disordered (in a mathematical sense) as DVP3.

Data Set	DVP1	DVP2	DVP3	DVP 4
H	2.1737	3.5563	4.0517	3.895

Table 4.4: Information entropy values for the four Duffing-Van der Pol oscillator series.

This claim would appear to be supported by the results in Table 4.4 which shows the information entropy for the four series. DVP4, does indeed exhibit less disorder than DVP3. However, this is in odds with the Lyapunov exponents given in table 4.1 which indicates that DVP4 exhibits chaotic behaviour where DVP3 does not. The information entropy does not take into account time information which is what we are trying to incorporate. Hence, the information entropy only tells us that the values in the series are less disordered in DVP4, not the *order* of the values in the series are less disordered. As a result, it should not be surprising



(a) Sample Entropy

(b) Kernel Entropy

Figure 4.23: Sample entropy results for increasing r, and kernel entropy results for increasing σ, for series DVP1 (blue), DVP2 (red), DVP3 (green) and DVP4 (magenta).

that kernel entropy and, to a lesser extent, sample entropy, require smaller kernel width values as there is no noise in the system, assuming a large kernel size is assuming that there is an amount of noise. This again highlights the importance of suitable bandwidth selection and that the statistics should possibly be thought of as regularity measures as opposed to disorder measures.

The calculation of the bandwidth may also mean that we can approximate the entropy rate (Equation 4.7 with the kernel entropy). Pesin's formula [Pesin, 1977] tells us that the sum of the positive Lyapunov exponents gives us the entropy rate (Chapter 2). If we investigate this using the Duffing-Van der Pol Oscillator system then only DVP4 has a positive Lyapunov exponent which is 0.0235. The corresponding kernel entropy estimate with the Bayesian bandwidth scheme is 0.7311. These results are very different implying that the kernel entropy is not applicable as an estimator of the entropy rate, possibly due to the dependence on fixed m and the fact that we are calculating the Renyi entropy as opposed to the information entropy.

We can now gauge the effectiveness of the bandwidth selection scheme. To do this, we started by calculating kernel entropy using the Bayesian bandwidth selection procedure with $m = 2$. The results can be seen in Table 4.5.

The results show that kernel entropy combined with the Bayesian bandwidth selection can clearly discriminate between different levels of disorder with the most ordered series (DVP1) having the lowest kernel entropy value followed by DVP2, DVP3 and DVP4. This shows that

Data Set	$KernEn$ Value	σ Value
DVP1	-0.9091	0.0067
DVP2	-0.3439	0.0145
DVP3	0.5220	0.0701
DVP4	0.7311	0.1314

Table 4.5: Kernel entropy values from using the Bayesian bandwidth selection scheme for the four Duffing-Van der Pol oscillator series when $m = 2$.

the Bayesian bandwidth selection does affect the discriminative capabilities of kernel entropy to quantify disorder as it assigns a higher bandwidth size and therefore a higher entropy value for DVP4 than DVP3 which is consistent with the maximal Lyapunov exponent values given in Table 4.1. The more irregular the series, the greater the bandwidth value; the bandwidth selection algorithm will assign a bandwidth that can best describe the probability density of a system. This result is of significance as it shows that using a variable bandwidth may be able to give a more suitable result. This means that although the kernel entropy cannot be considered an estimator of the entropy rate, it still may be applicable as an *indicator* of the level of chaos and disorder in a system.

	$m = 3$		$m = 4$	
Data Set	$KernEn$ Value	σ Value	$KernEn$ Value	σ Value
DVP1	-0.8870	0.0068	-0.8808	0.0067
DVP2	-0.3859	0.0145	-0.5441	0.0145
DVP3	0.4142	0.0691	0.3336	0.0700
DVP4	0.6707	0.1331	0.6202	0.1332

Table 4.6: Kernel entropy values from using the Bayesian bandwidth selection scheme for the four Duffing-Van der Pol oscillator series when $m = 3$ and $m = 4$.

Table 4.6 shows the results of kernel entropy with Bayesian bandwidth selection with $m = 3$ and $m = 4$. The first thing to notice is that the value given for the bandwidth values is very consistent across all three m values studied. This is encouraging as it indicates that the MCMC process is consistent and that the dimensionality of the kernel has little effect on its variance which is as you would expect as they are derived from the same series with no noise. Again, when $m = 3$ and $m = 4$, the more irregular the system, the higher the kernel entropy value. However, the values seem to become closer as m increases. The investigation of the kernel entropy value as m gets large may be of some interest but for comparison with sample entropy and in evaluating its usefulness for cardiac applications (which is the motivation for development of kernel entropy), this is outside the scope of this thesis.

4.3.5 Conclusion

We have introduced a novel measure based on the entropy rate for quantifying regularity of a time series known as kernel entropy. This was done by investigating how existing measures may be improved by approaching the problem from a signal processing perspective. This perspective led us to the use of Parzen windows for density estimation in the measure which allowed us to employ certain mathematical properties to ensure theoretical robustness and computational tractability.

The use of Parzen windows also allows us to select a method for automatically determining the size of the window, something which was not possible in the previous approaches. To this end, we used a Bayesian approach to create an appropriate posterior which we sampled from using a Markov chain Monte Carlo algorithm. This method performed well on synthetic data.

Kernel entropy was tested against a previous measure known as sample entropy in a variety of experiments. This showed that kernel entropy

1. is more robust to noise,
2. is more consistent and predictable,
3. has a wider range of appropriate values for σ than the equivalent term in sample entropy but the 'optimal' values are generally lower.
4. is more consistent over varying series lengths, particularly when N is small,
5. can be combined with an automatic bandwidth selection scheme to remove ambiguity in its choice without reduction in effectiveness.

Therefore, we can see that kernel entropy, with or without the bandwidth selection scheme, outperformed sample entropy in these tests. These indicate that kernel entropy will be a beneficial addition to the catalogue of regularity measures.

Chapter 5

Application of the Methods

As we have developed a method for quantifying regularity in a time series, we can apply it to a series created from the P-wave data. In Section 5.1 we show how we created such a series. This gives us two objectives in this chapter,

1. evaluate the effectiveness of kernel entropy,
2. determine the value of the P-wave series.

Of the three experiments outlined in Section 5.2, the first experiment only investigates kernel entropy; judging the effectiveness of the bandwidth selection on the classification results of kernel entropy when compared to the sample entropy. The other two experiments were designed to investigate both of these aims simultaneously. To this end we investigate the effectiveness of kernel entropy by applying it alongside the sample entropy, the information entropy and the Fisher information (see Chapter 4) to see how it performed in comparison with these measures. We applied the measures to both the RRI and the PWL series, which allowed us to judge the effectiveness of the latter.

The results are given in Section 5.3 with a brief discussion on each experiment which is expanded upon in the conclusion 5.4. We then end the chapter with Section 5.5; a discussion of issues arising pertaining to these results.

5.1 Creation of the Time Series

An RR-interval is calculated by subtracting the time, t, of the occurrence of the previous R-point ($R_{previous}$) from the time of occurrence of the current R-point ($R_{current}$) thus

$$RR_t = t_{R_{current}} - t_{R_{previous}}. \tag{5.1}$$

The series of these values forms the the RR-interval (RRI) series.

The P-waves were extracted from the data sets using the method detailed in Chapter 3. The P-wave duration was calculated simply as

$$P_t = t_{P_{end}} - t_{P_{start}}, \tag{5.2}$$

where P_t is the duration of the P-wave, P_{start} is the first upward deflection of the signal from the baseline, and P_{end} is the point when the signal rejoins the baseline. We call the time series formed by the sequence of these the *P-wave Length* (PWL) series, to avoid confusion with the standard P-wave duration and P-wave dispersion statistics.

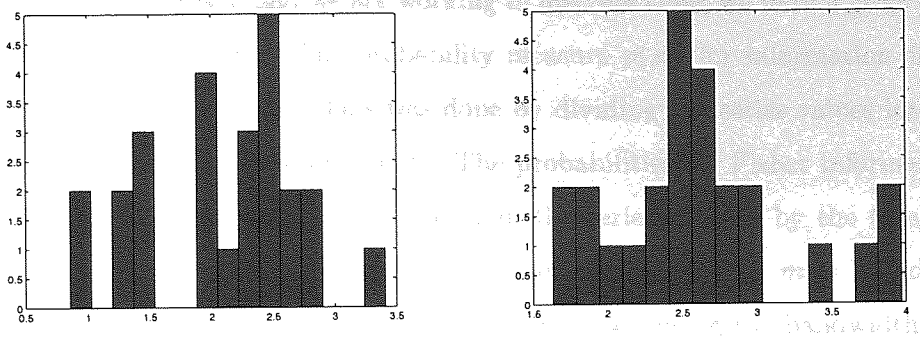
5.2 Experiments

To gauge the effectiveness of kernel entropy and the the P-wave length series, we propose three experiments. In the first, we apply only sample entropy and kernel entropy to the RR-interval series. In the final two experiments, we apply information entropy, Fisher information, sample entropy and kernel entropy to both the RR-interval and P-wave length series.

5.2.1 Experiment 1

The goal of this experiment was to evaluate the usefulness of the bandwidth selection method on real data. The data set in question is the paroxysmal atrial fibrillation prediction challenge data set. The data in this case comes as pairs of recordings from the same patient, with one recording immediately prior to an AF episode and one recording distant from an episode. There were 25 patients, so there were 50 distinct recordings. The task is to classify which of the pair of recordings for each patient is immediately prior to an atrial fibrillation episode and which is distant. Here, only the RR-interval series was used (as the usefulness of the P-wave length series has not been determined and its use is therefore superfluous to the goal of the experiment).

The first stage was to calculate the sample entropy for a range of r values in 0.01 increments ($r = [0.15, 0.16, \ldots, 0.3]$). It was and note down which of the pair gives the highest and lowest sample entropy. This range was chosen as it encompasses the range recommended by Pincus in [Pincus, 1991]. It was noted that the sample entropy values were more often *lower* for the records immediately prior to AF and so for each pair (see Figure 5.1), the recording with the lower value was classified as the one immediately prior to AF.



(a) Group P_c (b) Group P_d

Figure 5.1: Histograms of the values of sample entropy for Groups P_c and P_d for Experiment 1.

The second stage was to calculate kernel entropy using the bandwidth selection procedure and then classify the recordings in the same way. We then compared the classification results to judge the efficacy of the bandwidth selection procedure.

5.2.2 Experiment 2

As before, the measures were applied to series derived from the paroxysmal atrial fibrillation prediction challenge database with the same classification aim. However, in this experiment, the main aim was to judge the effectiveness of the P-wave length series so aside from kernel and sample entropy, we also used information entropy and Fisher information. Also, for comparison with conventional P-wave techniques, these results were compared to the standard P-wave statistics calculated for the P-waves; the P-wave [average] duration and P-wave dispersion.

As well as comparing the results between the RR-interval and P-wave length series, we can investigate the accuracy that can be achieved by combining the two series. We need to determine if the series provide the same information since, if so, there is little point in using the PWL series as the RRI series is easier to compute robustly. We can get an indication of the potential for classification improvement by combining the series by using them as inputs to a neural network classifier. In this case we used an MLP with a logistic sigmoidal activation function and 8 hidden units trained using the Bayesian evidence procedure. All the data was normalised to zero mean and unit variance and a 0-1 encoding was used for the outputs. The experiment was carried out using leave-one-out cross validation.

Since our data is digitised, and we are working in discrete time, we have a large but finite number of possible data values. The probability measure $\rho(x_i)$ for information entropy in Equation 2.20 used binned values. This was done by dividing the series values into 10 bins and using this to calculate the probabilities. The probabilities for Fisher information were simply the number of occurrences of a value x_i in the series divided by the total number of time points in the series, N. Sample entropy was calculated with $m = 2$ and $r = 0.2$ and kernel entropy was calculated with $m = 2$ and σ chosen using the bandwidth selection procedure.

5.2.3 Experiment 3

The third experiment aimed to investigate the value of multiple time series when distinguishing several heart conditions and the effectiveness of the measures when applied to them. This is to test how the methods may be of use to aid initial diagnosis.

The first dataset for this experiment was the atrial fibrillation dataset used above, only including the 25 records close to AF (group P_c). The second dataset consisted of thirty-minute samples taken at random from the Apnoea-ECG Database (35 records). The final data set was taken from the BIDMC Congestive Heart Failure (CHF) Database (15 records) and was again thirty-minute samples taken at random from longer (up to 24 hour) recordings.

The four information theoretic measures were calculated for the P-wave length and RR-interval series derived from each of these data sets.

We compared the two series by plotting the results for each measure when applied to the RR-interval against the results of the same measure applied to the P-wave length series. In this way we could identify patterns in the data and see which series leads to better discrimination between the conditions. As the application of such an approach in practice would be to aid diagnosis of a cardiac condition, normal ECG data was not included.

Further investigation was carried out by using the measures from the both series as inputs to a MLP neural network with 6 hidden units and a softmax activation function. This allowed us to estimate classification accuracy using leave-one-out cross validation. The decision boundaries are also plotted for clarity.

5.3 Results

5.3.1 Experiment 1 - Effectiveness of the Bandwidth Selection Procedure

The results in Table 5.1 show that kernel entropy combined with automated bandwidth selection performs as well as the best performing sample entropy. Although the results are not particularly encouraging on their own (68% is not a particularly high classification percentage), this is almost inconsequential as this experiment only aims to gauge the value of Bayesian bandwidth selection for use in conjunction with kernel entropy.

Sample Entropy r	Correctly Classified
0.15	52
0.16	52
0.17	52
0.18	56
0.19	60
0.20	60
0.21	64
0.22	**68**
0.23	60
0.24	64
0.25	64
0.26	64
0.27	64
0.28	64
0.29	60
0.30	64
Kernel Entropy	**68**

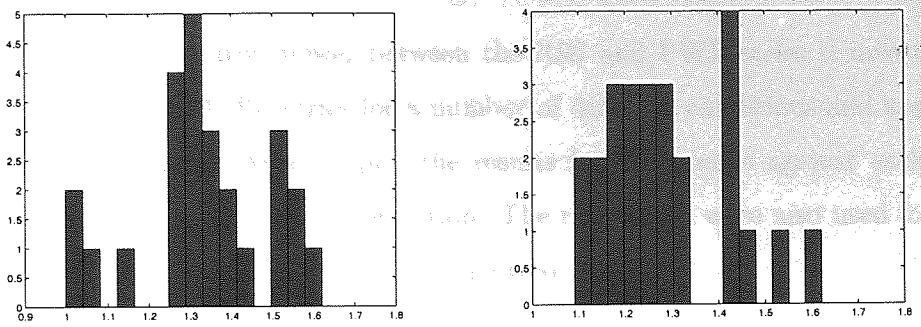
Table 5.1: Correct classification percentage of the 25 pairs of patient-specific atrial fibrillation data when sample entropy is applied for a range of r values, and kernel entropy is applied using the Bayesian bandwidth estimation.

The results are particularly encouraging as they show the variation in the results of sample entropy; indicating how important it is to select r carefully.

Figure 5.2 shows that, similar to sample entropy, the results of the statistic are lower for the recordings prior to an atrial fibrillation episode.

5.3.2 Experiment 2 - Performance on Atrial Fibrillation Prediction

Since the time series were from pairs of recordings, the values of the entropy measures could be directly compared to each other. The results were then examined to see if the recordings immediately prior to AF displayed a higher or lower value than their counterpart. This



(a) Group P_c (b) Group P_d

Figure 5.2: Histograms of the values of kernel entropy for Groups P_c and P_d for Experiment 1.

showed that, in general, the values *decreased* before AF. Therefore, the recording from each pair that had the lower value was classified as that being immediately prior to AF onset.

These results were compared to the conventional P-wave measures and the correct classification percentage is shown in Table 5.2. The poor performance of the P-wave dispersion is due to strong agreement in both recordings for each patient.

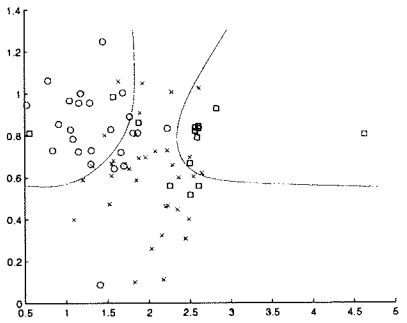
Combining the measures of both series yielded better results than each series on its own. This implies that the information obtained from the series is not highly correlated. It is worth noting that Fisher information, sample entropy and kernel entropy all achieved the same results. This is likely to be because there is not sufficient information for this classification task available in the series themselves and so the measures are all performing as well as is possible, regardless of the measure used.

Measure	RRI	PWL	Classifier
Information Entropy	60	64	92
Fisher Information	56	72	100
Sample Entropy	56	72	100
Kernel Entropy	56	72	100
P-wave Duration	-	64	-
P-wave Dispersion	-	44	-

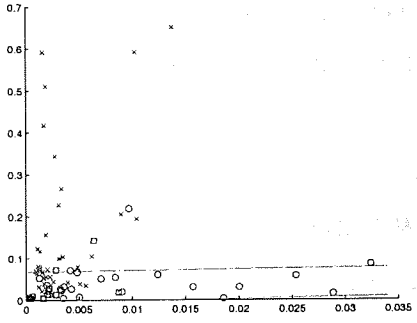
Table 5.2: Correct classification percentage for each statistic applied to each time series derived from the atrial fibrillation dataset, and both series as inputs in a neural network.

5.3.3 Experiment 3 - Discriminative Potential of the Series

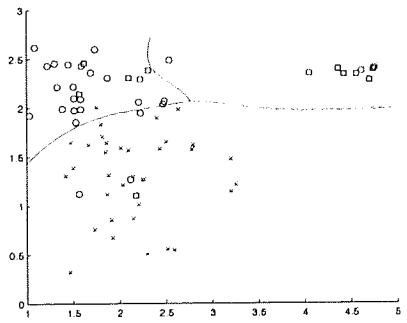
The disparity in discriminative power between the RRI and PWL series is investigated by determining the RRI and PWL series for a number of different conditions and applying the various measures to them. We then plot the results for each series against each other to determine if they provide the same information. The raw values were also used for training an MLP classifier and the decision boundaries plotted.



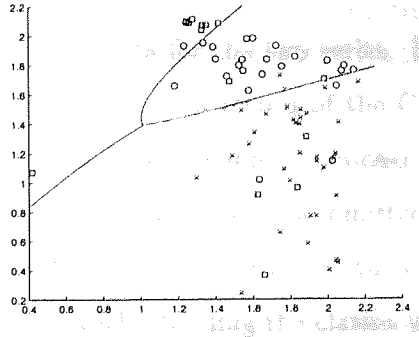
(a) Information Entropy



(b) Fisher Information



(c) Sample Entropy



(d) Kernel Entropy

Figure 5.3: Comparison of the results of the RRI series (x-axis) and the PWL series (y-axis) for each of the information theoretic measures. The three classes are atrial fibrillation (circle), sleep apnoea (cross) and chronic heart failure (square).

Figure 5.3 shows the results of applying the four information theoretic measures to both series for three heart conditions. The discrete points are the values obtained directly from application of the information theoretic measures; no classifiers have been applied at this stage. The green lines indicate the decision boundary calculated using an MLP with 6 hidden units and a softmax activation function trained on all of the data. Figure 5.4 shows the class

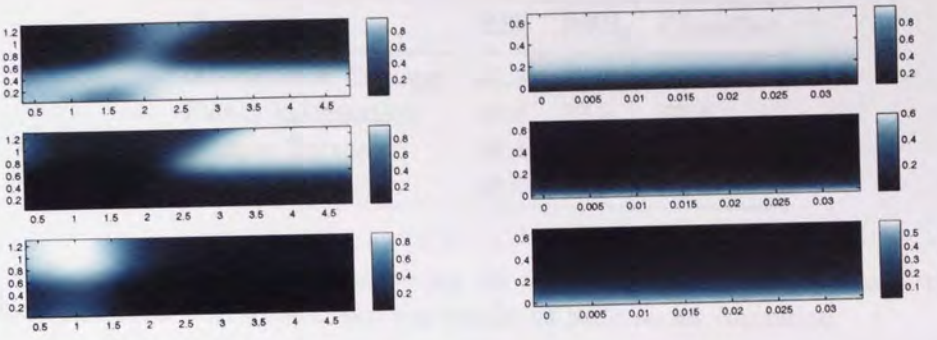
conditional probability densities as estimated by the MLP. The results of MLP classification using leave-one-out cross validation are shown in Table 5.3.

Figure 5.3(a) shows the plot of information entropy when applied to the two series. The classes do not show particularly good separation with a fair amount of overlap. The decision boundaries are also indistinct so any value falling in the area where they overlap is likely to me misclassified. If we were to compare just the RR-interval series results (x-axis) then we can see that there is very little discriminative potential with the classes overlapping considerably. The same is true for the P-wave length series results with considerable overlap between the results. This would account for its relatively poor performance in the classifier (Table 5.3)

The Fisher information plot in Figure 5.3(b) shows variability in the apnoea class for PWL data while having a relatively small variation in the RRI data. Also, it returns a wide range of values from the RRI series of the AF class but only a small range for the PWL series. The CHF data clusters very well with hardly any variation in the PWL values. However, despite the encouraging visual separation, the classifier performs poorly. This is possibly due to the values being too close for the classifier to meaningfully distinguish between and it appears to show a complete failure to distinguish any of the class structure. This can be seen more clearly in Figure 5.4(b).

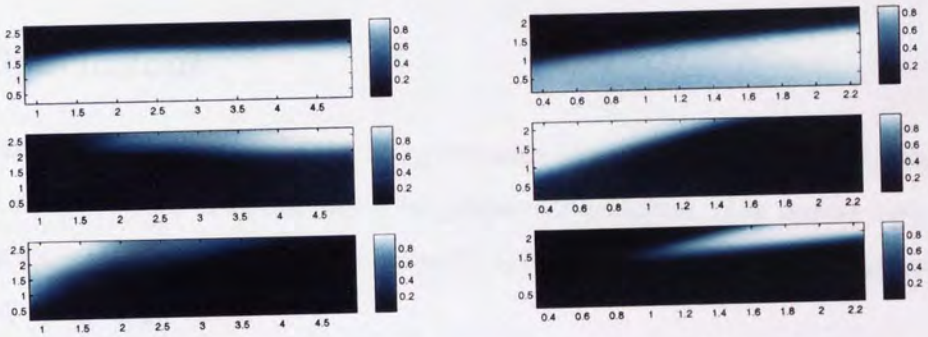
Figure 5.3(c) shows the plot for the sample entropy results for the two series. It should also be noted that the sample entropy was undefined for the 14th record of the CHF data for the RR-interval series. This led to the neural network being unable to process the data and so the results given for sample entropy are with the undefined value omitted. Aside from this, the classes are more separated than with information entropy, and the clustering is improved. This is shown by the decision boundaries clearly dividing the classes with very little ambiguity excepting the triangular area in the centre. This is again demonstrated by the densities shown in Figure 5.4(c). The distinct separation also leads to an improvement in the classifier over the previous measures as can be seen in Table 5.3. If we look at each axis in turn then we can see there is some slight variability in the sample entropy value for the RRI series for apnoea and AF but there is a more distinct difference in the values for the PWL series for these conditions. The main cluster of the CHF results has a slightly larger sample entropy for the RRI series.

Figure 5.3(d) shows the plot for kernel entropy. This also shows good class separation with the individual classes quite tightly condensed. Less class variability is shown in the values for the RRI series, although the PWL series is still distinct. The main cluster of the CHF values



(a) Information Entropy

(b) Fisher Information



(c) Sample Entropy

(d) Kernel Entropy

Figure 5.4: The probability densities for each of the information theoretic measures. The three classes are sleep apnoea (top), chronic heart failure (middle) and atrial fibrillation (bottom). A light colour indicates a high probability that the results will fall in this area.

are closer to the apnoea and AF clusters in the RRI series values than in the other statistics which is not so beneficial for classification based on a single series. This is highlighted by the relatively poor performance of kernel entropy for the RRI series in Table 5.3. However, kernel entropy displays very good class separation in the P-wave length series which is quantitatively shown in its performance for PWL. The results of both series combined give good separation (as evidenced by Figure 5.4(d)) with only one area of ambiguity which only contains two data values.

If we look at the results of Table 5.3 on their own, we see that the P-wave length series outperforms the RRI series for all of the statistics except information entropy. This indicates that the PWL series could have individual diagnostic value. However, the results when the series are combined are also higher in all the statistics, except the Fisher information, than

Measure	RRI	PWL	Classifier
Information Entropy	66.6	60.0	70.6
Fisher Information	46.6	57.3	57.3
Sample Entropy	69.3	69.3	80.0
Kernel Entropy	60.0	81.3	82.6

Table 5.3: Correct classification percentage for each statistic applied to each time series derived from the three conditions, and the results of both series combined.

the P-wave length on its own. Therefore, it would appear that using both series together offers the best procedure for reducing the chance of misclassification.

5.4 Conclusions

The first experiment indicated that by using Bayesian bandwidth selection with kernel entropy is a valid alternative to cross-validation for parameter selection. The results also show how important correct parameter selection actually is for sample entropy, a phenomenon rarely mentioned in the literature.

In the second experiment, the raw results suggest that using the P-wave length series is a viable alternative to the RR-interval series in discriminating between the two types of atrial fibrillation recordings. Also, all these measures applied to the PWL series outperformed the two conventional P-wave statistics which further indicates the potential in this approach. Combining the two series highlights how effective a classifier based on multiple features is and shows that the two series have a degree of independence. This is useful as it shows that the two series, while certainly not completely uncorrelated, may complement each other in diagnostic applications.

This is confirmed by the third experiment which indicates that the P-wave length series, when used in conjunction with the RRI series provides more discriminative power than the RRI series alone. The good class separation displayed by sample entropy and kernel entropy are encouraging, particularly as they incorporate temporal information which may mean that the series of P-waves has a dynamical structure and could therefore be examined using techniques designed to determine chaos such as in [Gottwald and Melbourne, 2005]. Kernel entropy does not perform so well when used on the RRI series, but gave a defined result for each of the recordings, unlike the sample entropy. It performs particularly well when combined with the PWL series which led to kernel entropy outperforming the other measures overall.

Therefore, from these experiments we can say that both the P-wave length series and kernel entropy merit further study as they both display potential for clinical use.

5.5 Discussion

The results indicate that the two novel techniques introduced in this thesis show promise in the field of cardiology. Unfortunately, this cannot be seen as a fully rigorous study as these experiments were performed on limited data sets due to the difficulty of obtaining reliable, consistent data.

In the first experiment, the automated bandwidth estimation led kernel entropy to equal the correct classification percentage of sample entropy. Again, with more data, the test could be repeated to see if this is a reliable result or merely a lucky one. However, it shows the usefulness of the bandwidth selection as often cardiac measurements are taken in a constantly changing, dynamic environment. In this case, it may not be feasible to select the parameters manually and so the automated method may be employed to consistently update the parameters as appropriate. Also, large data sets that exhibit non-stationarity may require windowing and so the σ values will need to be updated accordingly.

In the first two experiments, we can see that the information measures for the series decrease as the patient approaches AF. However, as the records are only quantified as 'distant' or 'close', it is difficult to predict how any clinical application might be implemented. From inspection of the values themselves, there does not appear to be a global threshold that would indicate that the record comes from someone immediately prior to AF; any device employing this method would have to be calibrated for the patient under normal beat conditions. Further study could be undertaken using a number of timed ECG recordings and investigating the temporal evolution of the statistics as the patients approach an AF episode. This would give an insight into the mechanisms of AF as well as furthering the understanding of how the measures and time series change.

In the third experiment, the three classes are naturally quite distinct so it is encouraging that an acceptable degree of class separation was achieved. The experiment does show that the use of the PWL series improves discrimination between classes compared to that achieved with the RRI series. This was done to highlight the potential of the PWL series and advocate its use as a complementary analysis method to the RRI series rather than a substitute.

Analysis of the P-wave dynamics could also facilitate investigations into the role of the sino-atrial node and the atrial influence in anomalous cardiac behaviour. This could have

implications in determining pathological influence as well as diagnostic and prognostic applications [Tso et al., 2005].

Ultimately, the robustness of the P-wave length series itself is predominantly dependent on the accuracy of the method used to extract the P-wave. In any ECG recording, noise is always an issue and how the level of noise affects the accuracy of the extraction is undetermined. Our extraction technique was not perfect. However, this can be offset by reliable filtering procedures and by using measures which model noise or are robust to the effects of noise. In the measures used here, only sample and kernel entropy explicitly incorporate noise in the model.

Also, some physiological conclusions may be drawn: as entropy is also a measure of disorder, in the first experiment, it would appear that the disorder decreases as the patient nears AF. This implies that the heart rate gets less erratic before an AF episode. This is consistent with findings of a recent independent study [Tuzcu et al., 2006]. In the second experiment, by the same reasoning, it would appear that people prior to atrial fibrillation and chronic heart failure have a more variable P-wave length than those suffering from apnoea (who are also asleep). This is in line with expectations and highlights the physiological significance of the P-wave length series.

Chapter 6

Summary

In this thesis we have investigated methods to determine cardiac disorders, from the raw ECG data to the classification.

We started by modifying an existing R-point detection algorithm to increase its robustness for a variety of data sets. This was then complimented with a new P-wave extraction technique which performed favourably compared to similar methods in the literature. This was due to careful filtering and baseline wandering removal.

The P wave data was then investigated using standard statistical techniques and then visualising using principal component analysis and NeuroScale. As these proved ineffective, this motivated the need for a different approach.

Of those approaches, some success had been achieved with measures that quantify the regularity in a time series, particularly approximate and sample entropies. However, it was soon noticed that aside from these measures leading to inconsistencies, if one wanted to apply them to varying types of data then the parameter used on one data set may not be applicable to another.

We decided to address these issues by creating kernel entropy which, by using properties of Gaussian kernel Parzen windows and the Renyi entropy, remains computationally efficient as well as being more consistent and robust than the previous measures. The use of Parzen windows also allows us to use techniques to estimate the optimum value for the bandwidth parameter. We chose to use a Bayesian bandwidth selection method utilising the Markov chain Monte Carlo algorithm to estimate the parameter as it is extremely flexible. This not only means that kernel entropy can be confidently applied to different data types, but also that the whole calculation of the measure can be fully automated without needing cross-validation to fine tune the parameters.

The effectiveness of kernel entropy was investigated as compared to sample entropy on synthetic data sets. This showed that kernel entropy was more robust and consistent for noisy data, different series lengths, different window sizes and better at determining the level of disorder. It also showed that the automated bandwidth selection performed well in determining a suitable bandwidth value.

A secondary goal of this thesis was to compare different features derived from the ECG. This led us to construct a time series from the series of sequential P-wave durations which we called the P-wave length series. The usefulness of the series was evaluated by applying kernel and sample entropies to it as well as two other information theoretic techniques for comparison. The results indicate that the use of the P-wave length series alongside the RR-interval series could be of use when trying to differentiate between several conditions. It would be remiss to imply that the PWL series is of equal or more use than the RR-interval series as the latter has been studied and researched to a huge extent and has proved itself as a reliable technique over time. It would take many more years of research and on a much wider range of data sets to develop an appropriate level of confidence for use of the P-wave length series in a clinical setting.

However, the main goal of the thesis was to determine the effectiveness of the kernel entropy when applied to cardiac data. The positive outcomes of the preliminary results on the synthetic data were confirmed in almost all of the experiments on cardiac data. The bandwidth selection method led to kernel entropy achieving an equal classification result to the best sample entropy could achieve despite numerous runs with different parameter values. Kernel entropy again matched the effectiveness of sample entropy in classification involving the P-wave lengths for atrial fibrillation prediction. Kernel entropy also led to better classification accuracy than the other measures when used to determine between a number of conditions.

Therefore, in this thesis, we have shown that kernel entropy is a viable alternative to sample entropy in quantifying regularity. As it is more consistent, and can be combined with a bandwidth selection scheme, it shows much promise for future use.

6.1 Further Research

Aside from validating the methods on a wider range of real-world data sets, other ways the research could be extended is to utilise other data sets. Kernel entropy could be of great use in quantifying the regularity of electroencephalogram (EEG) signals from the brain; for

instance it could be used to investigate how these signals change prior to an epileptic attack.

There could also be a way to automatically determine the value of m in kernel entropy. This was tried briefly, by using the differential entropy in section 2.5. This led to very inconsistent results and so the idea was dropped as erratic m values would lead to very inconsistent kernel entropy values. This is not to say that automatic selection of m is impossible, just that a large degree of robustness would be needed for the method to be feasible.

6.1.1 Multiscale Entropy

The most apparent way that the research could be furthered is by utilisation of kernel entropy in the *Multiscale Entropy* (MSE) formulation which currently uses sample entropy. To understand why kernel entropy may be better suited for use in this approach we need to give an overview of multiscale entropy.

Multiscale entropy is designed to investigate *complexity* as opposed to *regularity*. The difference is that simple periodic signals and completely random ones can both be described compactly, and are therefore not described as "complex" [Costa et al., 2005]. The idea is that investigating the correlations over a number of different structural scales can provide information on the complexity of a signal.

The procedure can be split into three parts

1. A "coarse-graining" process is applied to the time series, giving several coarse-grained time series of different scales,

2. Sample entropy is applied to each coarse-grained time series,

3. The sample entropy values are plotted as a function of the scale factor.

The coarse-grained series are constructed corresponding to a scale factor τ (unrelated to the τ in Section 4.1.1). The time series is divided into non-overlapping windows of length τ and the data points averaged inside each window. For a series such as in Equation 4.1, the mathematical formulation for creation of each coarse-grained data point, $y_j^{(\tau)}$, is

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \tag{6.1}$$

where $1 \leq j \leq N/\tau$.

The sample entropy is then calculated for the series of these $y_j^{(\tau)}$. The process is repeated for different values of τ, up to a limit of the users choice. This has been shown to be of

use for analysis of EEGs in Alzheimer's disease patients [Escudero et al., 2006] and ECGs in RR-interval time series [Costa et al., 2005].

There are obvious benefits of employing kernel entropy in this method as it is more reliable and consistent so confidence in the result is increased. Also, due to the reduction in sample points as the scale is increased, it is known that MSE requires a reasonably large number of data points to produce a consistent results [Costa et al., 2005]. As we know that kernel entropy is more robust to short series lengths, direct replacement of sample entropy with kernel entropy in this formulation would go some way to alleviating this constraint. Moreover, there could be more subtle advantages that could be gained.

In Nikulin and Brismar [2004], the observation is made that the same r value is used for sample entropy in different time scales. This leads to ambiguity in the MSE method as it is unclear if the results are from variation of the coarse-grained series or the underlying complexity. This is answered in Costa et al. [2004] and addressed further by Thuraisingham and Gottwald [2006] where it is noted that the multiscale entropy can have different signatures based on the time scales involved and the nature of the data. The use of kernel entropy with the bandwidth selection could help address these issues and possibly form a more rigorous foundation for the multiscale entropy approach.

Appendix A

Morphological Filters

Mathematical morphology is a formal methodology based in set theory that is usually applied to image processing problems [Maragos and Schafer, 1990]. It involves the use of a *structure* element that is then applied to the data by some pre-defined operation. The structure element is a compact set of small size and simple shape. The choice of the structure element in filtering depends on the noise level and the features in the data one would wish to enhance.

The two fundamental morphological operations are known as *erosion* (denoted by $\ominus$) and *dilation* (denoted as $\oplus$). Mathematically, if X is the data set of interest and B is the structure element then

$$X \ominus B = \bigcap_{b \in B} X - b = \{z : (B + z) \subseteq X\}, \tag{A.1}$$

$$X \oplus B = \bigcup_{b \in B} X + b = \{x + b : x \in X, b \in B\}. \tag{A.2}$$

We can see from this that erosion gives the set of values in X that can contain the complete structure element and dilation gives a new set inclusive of the values outside X that can be contained in X and B.

We can define a number of combinations of these two operators that also prove useful. We shall limit the discussion to only those used in this thesis, namely *opening* (denoted as $\circ$) and *closing* (denoted as $\bullet$) which are defined as

$$X \circ B = (X \ominus B) \oplus B, \tag{A.3}$$

$$X \bullet B = (X \oplus B) \ominus B. \tag{A.4}$$

In a signal processing sense, both can be seen to 'smooth' the contours of the input signal

by removing sharp features. Opening does this by removing any sharp or small features and closing does this by filling in small gaps between protrusive features.

Appendix B

Surrogate Data

The concept of surrogate data testing in time series analysis was introduced by Theiler [Theiler, 1995] and has been widely used in a number of studies [Small et al., 2000; Small and Judd, 1998; Schreiber and Schmitz, 2000].

The basic principle of surrogate data analysis is to generate a hypothesis that the series that you are investigating has certain properties. Normally several series (known as *surrogates*) have exactly the same statistical properties as the original series, except the properties stated in the hypothesis. The surrogates and the data series are evaluated using a scheme or measure of the users choice. The results then undergo hypothesis testing to see if the data set is suitably different from the surrogate(s) to be classified as exhibiting the hypothesised properties itself. Our use of the surrogate is to determine the degree of 'regularity' that can be determined with the appropriate measures.

We use two surrogates in this thesis: A *shuffled* surrogate which has the same mean and variance but with a random order and a *phase-randomised* surrogate which has the same mean, variance and power spectrum as the original data series but is otherwise random.

The associated null hypothesis of the shuffled surrogates is that the ordering of the data points are random. The alternative hypothesis is that they are not random.

The associated null hypothesis of the phase-randomised surrogates is that the data does not follow an ordered path through phase space. The alternative hypothesis is that they do.

A shuffled surrogate is created by randomly reordering the values of the series. Creation of a phase-randomised surrogate is more difficult and is detailed below.

B.1 Method of Constructing a Phase-Randomised Surrogate

Creation of a surrogate data set has been accomplished in a variety of ways; our method is taken from [Henry et al., 2001] where the Fourier transform is taken of the data, the phases are randomised and then the inverse Fourier transform is taken.

Consider the time series vector $\mathbf{z}$ in the form

$$\mathbf{z} = \mathbf{x} + i\mathbf{y},$$

where $\mathbf{y} = \mathbf{0}$.

We then apply the discrete Fourier transform,

$$Z_m = X_m + iY_m = \frac{1}{N}\sum_{n=1}^{N} z_n \exp\Big(-2\pi i(m-1)(n-1)/N\Big),$$

to get the frequency domain representation of the data. This allows us to construct a set of random phases

$$\phi_m \in [0, \pi], m = 2, 3, \ldots, \frac{N}{2},$$

and apply them to the transformed data thus

$$Z'_m = \begin{cases} Z_m & \text{for } m = 1 \text{ and } m = \frac{N}{2} + 1 \\ |Z_m| \exp\{i\phi_m\} & \text{for } m = 2, 3, \ldots, \frac{N}{2} \\ |Z(N-m+2)| \exp\{-i\phi_{N-m+2}\} & \text{for } m = \frac{N}{2} + 2, \frac{N}{2} + 3, \ldots, N \end{cases}$$

Finally, we finish by applying the inverse Fourier transform,

$$z'_n = x'_n + iy'_n = \frac{1}{N}\sum_{m=1}^{N} Z'_m \exp\Big(2\pi i(m-1)(n-1)/N)\Big),$$

to map back to the time domain.

Bibliography

U. R. Acharaya, N. Kannathal, and S. M. Krishnan. Comprehensive analysis of cardiac health using heart rate signals. *Physiol. Meas.*, 25:1139–1151, 2004.

American College of Cardiology/American Heart Association Task Force on Performance Measures. ACC/AHA clinical performance measures for adults with chronic heart failure. *J. Am. Coll. Cardiol.*, 46(6):1144–1178, 2005.

American College of Cardiology/American Heart Association Task Force on Practice Guidelines. ACC/AHA guidelines for the evaluation and management of chronic heart failure in the adult: Executive summary. *J. Am. Coll. Cardiol.*, 38(7):2102–2113, 2001.

American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Management of patients with st elevated myocardial infarction : executive summary. *J. Am. Coll. Cardiol.*, 44:671–719, 2004.

K. S. Anant, F. U. Dowla, and G. H. Rodrigue. Detection of the electrocardiogram P-wave using wavelet analysis. Technical report, Lawrence Livermore National Laboratory, 2000.

A. C. Bales and M. J. Sorrentino. Causes of congestive heart failure: prompt diagnosis may affect prognosis. *Postgrad. Med.*, 101(1), 1997.

L. Bauwens and M. Lubrano. Bayesian inference on GARCH models using the Gibbs sampler. *Econometrics J.*, 1:C23–C46, 1998.

F. Beckers, B. Verheyden, and A. E. Aubert. Quantifying non-linear behaviour of cardiovascular variability. In *World Congress on Medical Physics and Biomedical Engineering*, August 2003.

C. M. Bishop. *Neural Networks for Pattern Recgonition.* Oxford University Press, 1995.

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

M. Bračič and A. Stefanovska. Nonlinear dynamics of the blood flow studied by Lyapunov exponents. *Bull. Math. Bio.*, 60:417–433, 1998.

B. Brembilla-Perrot. Electrophysiological evaluation of Wolff-Parkinson-White Syndrome. *Indian Pacing Electrophysiol. J.*, 2(4):143–156, 2002.

D. S. Broomhead and G. P. King. Extracting qualitative dynamics from experimental data. *Phys. D*, 20(2-3):217–236, 1986.

R. Brown, P. Bryant, and H. D. I. Abarbanel. Computing the lyapunov spectrum of a dynamical system from an observed time series. *Phys. Rev. A*, 43(6):2787–2806, 1991.

J Carlson, R. Havmöller, A. Herreros, P. Platonov, R. Johansson, and B Olsson. Can orthogonal lead indicators of propensity to atrial fibrillation be accurately assessed from the 12-lead ECG? *Europace*, 7:S39–S48, 2005.

G. Çelebi, I. S. Uzun, M. Pehlivan, M. H. Asyali, A. Türkoğlu, and İ. Soysan. Nonlinear analysis of heart rate variability. In *Nonlinearity and Disorder: Theory and Applications*, pages 387–396. Kluwer Publishers, 2001.

Committee to Develop Guidelines for the Management of Patients With Atrial Fibrillation. ACC/AHA/ESC guidelines for the management of patients with atrial fibrillation. *J. Am. Coll. Cardiol.*, 38(4):1231–1265, 2001.

M. Costa, A. L. Goldberger, and King C.-K. Reply to Nikulin and Brismar [2004]. *Phys. Rev. Lett.*, 92:089804, 2004.

M. Costa, A. L. Goldberger, and King C.-K. Multiscale entropy analysis of biological signals. *Phys. Rev. E*, 71:021906, 2005.

M. Costa and J. A. Healey. Multiscale entropy analysis of complex heart rate dynamics: discrimination of age and heart failure effects. In *Computers in Cardiology*, volume 30, pages 705–708, 2003.

D. Cuesta-Frau, D. Novác, V. Eck, J. C. Pérez-Cortés, and G. Andreu-García. Electrocardiogram baseline removal using wavelet approximations. Technical report, Polytechnic University of Valencia, 2001.

P. Cugini, F. Bernardi, C. Cammarota, D. Cipriani, M. Curione, T. De Laurentis, E. De Marco, R. De Rosa, F. Fallucca, P. Francia, and A. Napoli. Is a reduced entropy in heart rate variability an early finding of silent cardiac neurovegetative dysautonomia in type 2 diabetes mellitus? *J. Clin. Basic. Cardiol.*, 4(4):289–294, 2001.

P. Cugini, M. Curione, C. Cammarotta, F. Bernardi, E. Proietti, L. Cedrone, and C. Danese. Evidence that the information entropy estimating the non-linear variability of human sinusal R-R intervals shows a circadian rhythm. *J. Clin. Basic. Cardiol.*, 2(2):275–278, 1999.

N. Davey, S.P. Hunt, and R.J. Frank. Time series prediction and neural networks. *J. Intell. Robot. Syst.*, 31:91–103, 2001.

P. de Chazal and C. Henegan. Automated assessment of atrial fibrillation. In *Computers in Cardiology*, pages 117–120, 2001.

P. de Chazal, T. Penzel, and C. Heneghan. Automated detection of obstructive sleep apnoea at different time scales using the electrocardiogram. *Physiol. Meas.*, 25:967–983, 2004.

D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith. *Bayesian Methods for Nonlinear Classification and Regression.* Wiley, 2002.

P. A. Doevendans. Hypertrophic cardiomyopathy : do we have the algorithm for life and death? *Circulation*, 101:1224–1226, 2000.

I. A. Dotsinsky and T. V. Stoyanov. Ventricular beat detection in single channel electrocardiograms. *Biomed. Eng. Online*, 3(3), 2004.

J. P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57(3):617–656, 1985.

G. Engel, J. G. Beckerman, V. F. Froelicher, T. Yamazaki, H. A. Chen, K Richardson, R. J. McAuley, E. A. Ashley, S. Chun, and P. J. Wang. Electrocardiographic arrythmia risk taking. *Curr. Probl. Cardiol.*, 29(7):365–432, 2004.

J. Escudero, D. Abásolo, R. Hornero, P Espino, and M. López. Analysis of electroencephalograms in Alzheimer's disease patients with multiscale entropy. *Physiol. Meas.*, 27:1091–1106, 2006.

A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33(2):1134–1140, February 1986.

B. R. Frieden. *Science from Fisher Information: A Unification.* Cambridge University Press, 2004.

R. Furlan, S. Guzetti, W. Crivellaro, S. Dassi, M. Tinelli, G. Baselli, S. Cerutti, F. Lombardi, M. Pagani, and A. Malliani. Continuous 24-hour assessment of the neural regulation of systemic arterial pressure and RR variabilities in ambulant subjects. *Circulation*, 81:537–547, 1990.

Y. Fusheng, H. Bo, and T. Qingyu. Approximate entropy and its application in biosignal analysis. In *Nonlinear Biomedical Signal Processing, Volume II: Dynamic Analysis and Modelling*, chapter 3, pages 72–91. IEEE Press, 2001.

A. Galka, T. Maaß, and G. Pfister. Estimating the dimension of high-dimensional attractors: a comparison between two algorithms. *Physica D*, 121:237–251, 1998.

T. Gautama, D. P. Mandic, and M. M. Van Hulle. A differential entropy based method for determining the optimal embedding parameters of a signal. In *International Conference on Acoustics, Speech and Signal Processing*, volume 6, pages 29–32, April 2003.

T. Gautama, D. P. Mandic, and M. M. Van Hulle. A novel method for determining the nature of time series. *IEEE Trans. Biomed. Eng.*, 51(5):728–736, 2004.

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.

G.A. Gottwald and I. Melbourne. Testing for chaos in deterministic systems with noise. *Physica D*, 212:100–110, 2005.

P. Grassberger and I. Procaccia. Characterisation of strange attractors. *Phys Rev Lett*, 50 (5):346–349, 1983a.

P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9:189–208, 1983b.

R. M. Gray. *Entropy and Information Theory.* Springer-Verlag, 1990.

A. Guerrero and L. A Smith. Towards coherent estimation of correlation dimension. *Phys. Lett. A*, 318:373–379, 2003.

C. Guilleminault, S. J. Connoly, R. Winkle, K. Melvin, and A. Tilkian. Cyclical variation of the heart rate in sleep apnoea syndrome. Mechanisms and usefulness of 24hr electrocardiography as a screening technique. *Lancet*, 321:126–131, 1984.

J. R. Hampton. *The ECG Made Easy (3rd Edition).* Churchill Livingstone, 1986.

T. Harel, I. Gath., and A. Ben-Haim. System response of the sinoatrial node during vagal stimulation. *Physiol. Meas.*, 19:149–157, 1998.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

R. Hecht-Nielsen. Kolmogorov's mapping neural network existence theorem. In *First Annual International Conference on Neural Networks*, 1987.

M. Henon. A two-dimensional mapping with a strange attractor. *Comm.Math.Phys*, 50: 69–77, 1976.

B. Henry, N. Lovell, and F. Camacho. Nonlinear dynamics time series analysis. In *Nonlinear Biomedical Signal Processing, Volume II: Dynamic Analysis and Modelling*, chapter 1, pages 1–28. IEEE Press, 2001.

H. B. Hubert, M. Feinleib, P. M. McNamara, and W. P. Castelli. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham heart study. *Circulation*, 67(5):968–977, 1983.

N. P. Hughes, L. Tarassenko, and S. J. Roberts. Markov models for automated ECG interval analysis. Technical report, University of Oxford, 2003.

A. C. Hunt. T wave alternans in high arrhythmic risk patients: Analysis in time and frequency domains: A pilot study. *BMC Cardiovasc. Disord.*, 2:6, 2002.

I.T Jolliffe. *Principal Component Analysis.* Springer, 1986.

K. Judd. An improved estimator of dimension and some comments on providing confidence intervals. *Physica D*, 56:216–228, 1992.

D. G. Julian. *Cardiology (3rd Edition).* Baillière Tindall, 1978.

L. Kaplan and J. A. Yorke. Chaotic behavior of multidimensional difference equations. In *Lecture Notes in Mathematics*, volume 730, pages 204–227. Springer, 1979.

M. B. Kennel, R. Brown, and H. D. I. Arbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, 45(6), 1992.

J. Y. Ko, J. S. Lih, M. C. Ho, C. C. Tsai, and J. L. Chern. Determinism test, noise estimate and hidden frequency recognition: the singular value decomposition approach. *Chinese J. Phys.*, 37(5):449–465, 1999.

L. F. Kozachenko and N. N. Leonenko. Sample estimate of entropy of a random vector. *Probl. Inform. Transmission*, 23:95–101, 1987.

D. E. Lake. Renyi entropy measures of heart rate Gaussianity. *IEEE Trans. Biomed. Eng.*, 53(1):21–27, 2006.

P. Langley, D. di Bernardo, J. Allen, E. Bowers, F. Smith, S. Vecchietti, and A. Murray. Can paroxysmal atrial fibrillation be predicted? In *Computers in Cardiology*, pages 121–124, 2001.

R. Lepage, J. M. Boucher, J. J. Blanc, and J. C Cornilly. ECG segmentation and P-wave feature extraction: application to patients prone to atrial fibrillation. In *Engineering in Medicine and Biology Society*, volume 1, pages 298–301. IEEE/EMBS, 2001.

G. Y. H. Lip and F. L. Li Saw Hee. Paroxysmal atrial fibrillation. *Q. J. Med.*, 94:665–678, 2001.

J. H. Liu and T. Kao. Removing artifacts from atrial epicardial signals during atrial fibrillation. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 179–183, 2003.

C. R. Loader. Bandwidth selection: Classical or plug-in? *Ann. Stat.*, 27(2):415–438, 1999.

E. N. Lorenz. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20:130–141, 1963.

D. Lowe and M. E. Tipping. Feed-forward neural networks and topographic mappings for exploratory data analysis. *Neural Comput. Appl.*, 4:83–95, 1996.

D. Lowe and M. E. Tipping. Neuroscale: novel topographic feature extraction with radial basis function networks. In *Advances in Neural Information Processing Systems 9*, pages 543–549, 1997.

D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

C. Maier, M. Bauch, and H. Dickhaus. Screening and prediction of atrial fibrillation by analysis of heart rate variability parameters. In *Computers in Cardiology*, pages 129–132, 2001a.

C. Maier, H. Dickhaus, L. M. Khadra, and T. Maayah. Nonlinear behavior of heart rate variability as registered after heart transplantation. In *Nonlinear Biomedical Signal Processing, Volume II: Dynamic Analysis and Modelling*, chapter 5, pages 133–158. IEEE Press, 2001b.

P. Maragos and R.W. Schafer. Morphological systems for multidimensional signal processing. *Proc. IEEE*, 78(4):690–710, 1990.

E. Martin, editor. *Oxford Concise Medical Dictionary.* Oxford University Press, 2003.

P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall, second edition, 1989.

C. R. Meyer and H. N. Keiser. Electrocardiogram baseline noise estimation and removal using cubic splines and state-space computational techniques. *Comput. Biomed. Res.*, 10: 459–470, 1977.

T. P. Minka. Automatic choice of dimensionality for PCA. In *Advances in Neural Information Processing Systems 13*, pages 598–604, 2000.

I. T. Nabney. *Netlab: Algorithms for Pattern Recognition.* Springer, 1999.

F. J. Nieto, T. B. Young, B. K. Lind, E. Shahar, J. M. Samet, S. Redline, R. B. D'Agostino, A. B. Newman, M. D. Lebowitz, and T. G. Pickering. Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study. *J. Am. Med. Assoc.*, 283(14):1829–1836, 2000.

V. V. Nikulin and T. Brismar. Comment on "Multiscale entropy analysis of complex physiologic time series". *Phys. Rev. Lett.*, 92:089803, 2004.

E. Ott. *Chaos in Dynamical Systems.* Cambridge University Press, 1993.

M. I. Owis, A. H. Abou-Zeid, A. B. M. Youssef, and Y. M. Kadah. Study of features based on nonlinear dynamical modelling in ECG arrhythmia detection and classification. *IEEE Trans. Biomed. Eng.*, 49(7):733–736, 2002.

N. Packard, J. Crutchfield, D. Farmer, and R. Shaw (1980). Geometry from time series. *Phys. Rev. Lett*, 45:712–716, 1980.

J. Pan and W. J. Tompkins. A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng.*, 32(3):230–236, 1985.

T. Penzel, J. McNames, P. de Chazal, B. raymond, A. Murray, and G. Moody. Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. *Med. Biol. Eng. Comput.*, 40:402–407, 2002.

Y. B. Pesin. Characteristic lyapunov exponents and smooth ergodic theory. *Russian Math. Surveys*, 32(4):55–114, 1977.

S. M. Pincus. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci.*, 88(6):2297–2301, 1991.

S. Poli, V. Barbaro, P. Bartolini, G. Calcagnini, and F. Censi. Prediction of atrial fibrillation from surface ECG: review of methods and algorithms. *Ann. First Super. Sanità*, 32(2): 195–203, 2003.

J. Príncipe, D. Xu, and J. Fisher. Information theoretic learning. In Simon Haykin, editor, *Unsupervised Adaptive Filtering*, pages 265–319. Wiley, 1999.

R. Quian Quiroga, J. Arnold, K. Lahnertz, and P. Grassberger. Kullback-Leibler and renormalized entropies: applications to electroencephalograms of epilepsy patients. *Phys. Rev. E*, 62(6):8380–8386, December 2000.

A. Rényi. On measures of information and entropy. In *4th Berkeley Symposium on Mathematics, Statistics and Probability 1960*, pages 547–561, 1961.

J. S. Richman and R Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart. Circ. Physiol.*, 278(6):H2039–H2049, 2000.

J. J. Rieta, F. Castells, C. Sánchez, and J. Igual. ICA applied to atrial fibrillation analysis. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 59–64, 2003.

S. J. Roberts, W. Penny, and I. Rezek. Temporal and spatial complexity measures for EEG-based brain-computer interfacing. *Med. Biol. Eng. Comput.*, 37(1):93–98, 1998.

M. T. Rosenstein, J. J. Collins, and C. J. De Luca. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D*, 65:117–134, 1993.

T. Schreiber and A. Schmitz. Surrogate time series. *Physica D*, 142:346–382, 2000.

M. H. Seedhagi. ECG wave detection using morphological filters. *App. Signal Process.*, 5: 182–194, 1998.

M Senni, C. M. Tribouilloy, R. J. Rodeheffer, S. J. Jacobsen, J. M. Evans, K. R. Bailey, and M. M. Redfield. Congestive heart failure in the community: a study of all incident cases in Olmstead County, Minnesota, in 1991. *Circulation*, 98(21):2282–2289, 1998.

F. Shamsham and J. Mitchell. Essentials of the diagnosis of heart failure. *Am. Fam. Physician*, 61(5), 2000.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

M. G. Signorini, R. Sassi, and S. Cerutti. Assessment of nonlinear dynamics in heart rate variability signal. In *Nonlinear Biomedical Signal Processing, Volume II: Dynamic Analysis and Modelling*, chapter 10, pages 263–281. IEEE Press, 2001.

M. Small and K. Judd. Detecting nonlinearity in experimental data. *Int. J. Bifurcat. Chaos*, 8(6):1231–1244, 1998.

M. Small, D. J. Yu, J. Simonotto, and R. G. Harrison. Nonlinear analysis of human ECG during sinus rhythm and arrhythmia. Technical report, Herriot-Watt University, Edinburgh, 2000.

J. S. Steinberg, S. Zelenkofske, S. C. Wong, M. Gelernt, R. Sciacca, and E. Menchavez. Value of the P-wave signal-averaged ECG for predicting atrial fibrillation after cardiac surgery. *Circulation*, 88:2618–2622, 1993.

F. Takens. Detecting strange attractors in turbulence. In D. A. Rand and L.-S. Young, editors, *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, volume 898, pages 366–381. Springer-Verlag, 1981.

Task Force for the Diagnosis and Treatment of Chronic Heart Failure, European Society of Cardiology. Guidelines for the diagnosis and treatment of chronic heart failure. *Eur. Heart J.*, 22:1527–1560, 2001.

Task force of the European Society of Cardiology and the Northern American Society of Pacing and Electrophysiology. Heart rate variability: Standards of measurement, physiological interpretation and clinical use. *Eur. Heart J.*, 17:354–381, 1996.

M. C. Teich, S. B. Lowen, B. M. Jost, K. Vibe-Rheymer, and C. Heneghan. Heart rate variability: Measures and models. In *Nonlinear Biomedical Signal Processing, Volume II: Dynamic Analysis and Modelling*, chapter 6, pages 159–213. IEEE Press, 2001.

J. Theiler. On the evidence for low-dimensional chaos in an epileptic electroencephalogram. *Phys. Lett. A*, 196:335–341, 1995.

R. A. Thuraisingham and G. A. Gottwald. On multiscale entropy analysis for physiological data. *Physica A*, 366:323–333, 2006.

M. E. Tipping and C. M. Bishop. Probabilistic principal componenet analysis. *J. R. Stat. Soc. Ser. B*, 61(3):611–622, 1999.

M. E. Tipping and D. Lowe. Shadow targets: a novel algorithm for topographic projections by radial basis functions. In *IEE Fifth International Conference on Artificial Neural Networks*, pages 101–105, 1997.

H. W. Tso, T. Kao, Y. J. Lin, C. T. Tai, and S. A. Chen. Role of sinus node during atrial fibrillation: A novel insight from regional frequency analysis. In *Computers in Cardiology*, volume 35, pages 65–68, 2005.

V. Tuzcu, S. Nas, T Börklü, and A. Ugur. Decrease in the heart rate complexity prior to the onset of atrial fibrillation. *Europace*, 8(398–402), 2006.

B. van der Pol and J. van der Mark. Frequency demultiplication. *Nature*, 120:363–364, 1927.

S. Vikman, T. H. Mäkikallio, S. Yli-Mäyry, S. Pikkujämasä, A. M. Koivisto, P. Reinikainen, K.E. Juhani Airaksinen, and H. V. Huikuri. Altered complexity and correlation properties of R-R interval dynamics before the spontaneous onset of paroxysmal atrial fibrillation. *Circulation*, 100:2079–2084, 1999.

C. D. Wagner and P. B. Persson. Chaos in the cardiovascular system: an update. *Cardiovasc. Res.*, 40:257–264, 1998.

G. S. Wagner. *Marriott's Practical Electrocardiography.* Lippincott Williams & Wilkins, tenth edition, 2001.

M. P. Wand and M. C. Jones. *Kernel Smoothing.* Chapman and Hall, 1995.

S. M. Weiss and C. A. Kulikowski. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems.* M. Kaufmann Publishers, 1991.

N. Wessel, A. Schumann, A. Schirdewan, A. Voss, and J. Kurths. Entropy measures in heart rate variability data. In *International Symposium on Medical Data Analysis*, pages 78–87. Springer-Verlag, September 2000.

A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano. Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16:285–317, 1985.

T. Wong, P. A. Davlouros, W. Li, C. Millington-Sanders, D. P. Francis, and M. A. Gatzoulis. Mechano-electrical interaction late after fontan operation: relation between P-wave duration and dispersion, right atrial size, and atrial arrhythmias. *Circulation*, 109:2319–2325, 2004.

D. Woodcock and I. T. Nabney. A new entropy measure based on the Renyi entropy rate using Gaussian kernels. Technical report, Aston University, 2006.

D. Xu and J. Príncipe. Learning from examples with quadratic mutual information. In *IEEE Signal Processing Society Workshop*, pages 155–164, 1998.

T. Young, P. E. Peppard, and D. J. Gottlieb. Epidemiology of obstructive sleep apnea. *Am. J. Respir. Crit. Care Med.*, 165:1217–1239, 2002.

X. Zhang, M. L. King, and R. J. Hyndman. A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Comp. Stat. Data Anal.*, 50:3009–3031, 2006.