Astronomical Data Mining with Neural Networks

Simone Scaringi

A Thesis presented for the degree of
Master of Philosophy

Astronomy Group
Department of Physics and Astronomy
University of Southampton
September 2006

Dedicated to

My family...and anyone fighting the evil tyranny of religion against science.

Astronomical Data Mining with Neural Networks

Simone Scaringi

Submitted for the degree of Master of Philosophy

September 2006

Abstract

We give a brief overview of artificial neural networks (ANNs), focusing on Kohonen networks (KNs). The two kinds of KNs will be described in detail: the unsupervised self-organizing map (SOM) and the supervised learning vector quantization (LVQ). We then apply these algorithms to two astronomical classification problems: the classification of broad absorption line quasars (BALQSOs) and of gamma-ray bursts (GRBs). In the context of BALQSOs, we find a BALQSO fraction of 10.4%, and compile a catalogue from the Sloan Digital Sky Survey (SDSS) using the supervised LVQ. This is currently the most complete BALQSO catalogue. We then apply the unsupervised SOM to GRB light curves obtained from the Burst and Transient Source Experiment (BATSE). Using only shape-dependent variables, we find that two classes are recovered: single-pulsed bursts (SPBs) and multi-pulsed bursts (MPBs). We show that

these two network classes also have different observational properties that are independent of light curve shape (T90 and fluence), suggesting an intrinsic difference between the two. We conclude with some attempts to correlate our GRB result to previous studies and suggest improvements for future work.

Declaration

The work in this thesis is based on research carried out at the Astronomy Group at the University of Southampton, England. No part of this thesis has been submitted elsewhere for any other degree or qualification. The work presented in this thesis is my own work (with a lot of help from Christian Knigge) unless stated otherwise in the text.

Acknowledgements

I would like to greatly acknowledge Dr. Christian Knigge for his great tutelage and infinite patience. Without him, this work would have not been possible. His immense enthusiasm in the field has been extremely contagious and inspiring, and I am forever in debt for the opportunities he has provided me with.

I also wish to thank the University of Southampton for the funding provided in this last year for the creation of this thesis.

I thank the Sloan Digital Sky Survey for their hard work in producing enormous amounts of data for the astronomical community.

Moreover, I wish to thank Dr. Mike Goad for his invaluable assistance and expert comments.

Finally, I would like to share my gratitude with the Astronomy Group in Southampton for their immense expertise and passion in taking astronomy research a step further. I am proud to be remaining within the group for (at least!) the next three years.

Contents

List of Figures

Chapter 1

Neural Networks In Astronomy

> *Every historical period has its godword. There was an Age of Faith, and Age of Reason, an Age of Discovery. Our time has been nominated to be the Age of Information* - Theodor Roszack (1986), The Cult of Information

A new model for the way information is processed has emerged during the last century[1]: Artificial Neural Networks (ANNs). Inspired by the way biological nervous systems work, the key element of this model is the novel structure of its processing system. ANNs are composed of a large number of interconnected processing elements, called neurons (by analogy to the brain), which work together to solve a specific problem such as pattern recognition or classification. Just like animals, ANNs can be trained to solve such problems, since they are able to learn by example. This new paradigm arose mainly due to the ever

[1]For a good review of the history of Artificial Neural Networks refer to Anderson D. and McNeil G. [2]

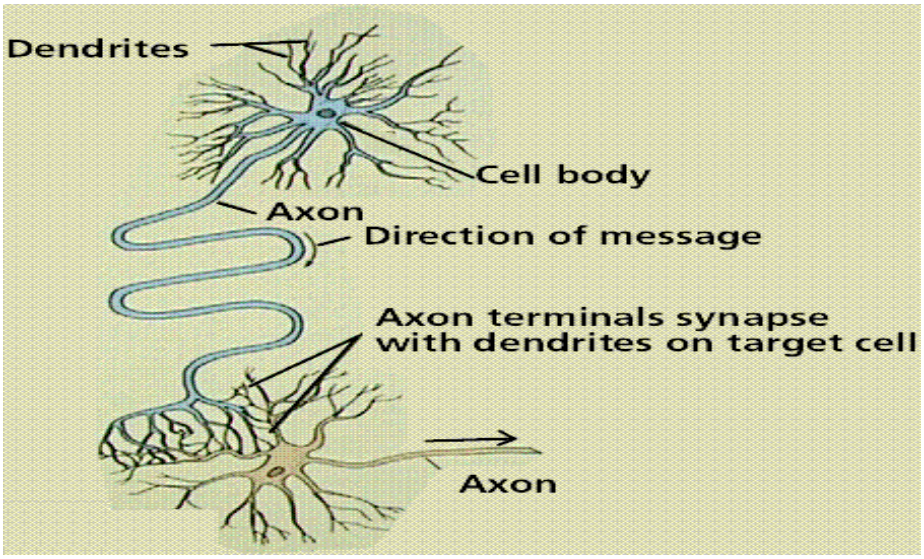

Figure 1.1: Schematic of brain neurons as taken from [27].

increasing amounts and sheer complexity of data that need to be analyzed.

In order to explain the basic principles governing ANNs, let us consider the analogy to the human brain, since this was the basic inspiration for their invention. Much is still unknown about how the brain "trains" itself and processes information, but the basic principles seem to be understood. The cell responsible for our learning is the neuron (Figure 1.1). It collects electrical signals (information) through a host of fine structures named dendrites. It then "communicates" with other neurons by sending electrical activity through a connecting structure called an axon, which eventually splits into many smaller branches, which terminate in so-called synapses. It is these synapses that communicate with the dendrites of other neurons, and the information is thus propagated to many more neurons. Learning takes place by changing the effectiveness of the synapses so that the influence of one neuron on another changes.

In ANNs a similar scheme is adopted. An input is presented to one or more hidden layers of neurons, which process the information and communicate with

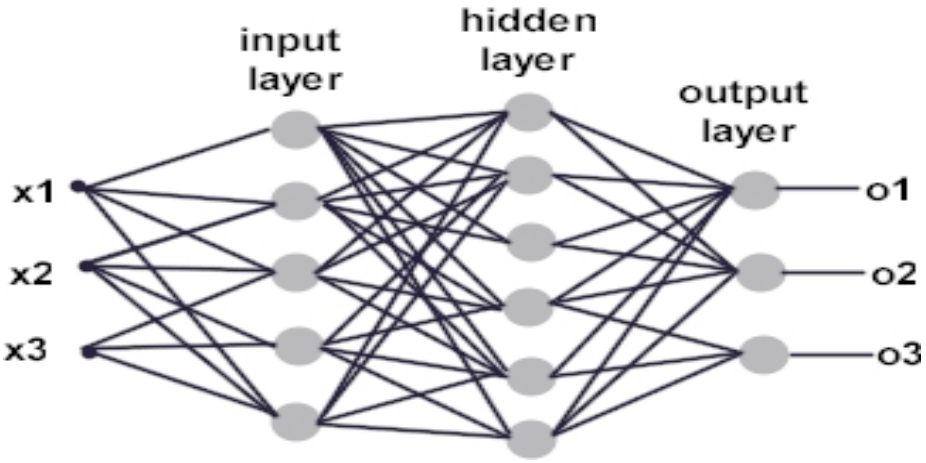

Figure 1.2: Schematic of a ANN as taken from [28].

each other using weights; these weights are the analogues of synapses. As the weights change in response to the inputs, learning takes place. The result from all neurons are then collected in the last layer, the output layer. (Actually, this distinction between output and hidden layer is not always clear. In fact, this thesis will be dealing with an ANN in which these two layers are the same.)

The main difference between the brain and an ANN is the number of neurons and synapses/weights. A typical ANN contains a few hundred to a few thousand weights, compared to the 10^{14} synapses in the human brain. Moreover, there is also a difference in processing power and speed. In particular, the human brain can parallelise (solve multiple tasks at the same time) tasks neatly and therefore speed up processing. This level of parallelism is not possible with contemporary ANNs.

In the context of data mining and classification, ANN schemes can come in two main types: supervised or unsupervised. In the former case, the user knows in advance what the dataset consists of and has a training set available for the network. This training set will be tagged into various classes as determined by the user. The ANN will then process the information in a reward-punishment

scheme. Every time an input gets connected to the correct neuron the network will "learn", while it will "unlearn" every time the connection is wrong. The aim is to determine the set of weights which minimise the error. In doing so, particular neurons will be tuned to recognise particular input patterns, which could then be applied to unseen data for recognition purposes.

On the other hand, unsupervised ANNs work in a way that is more closely analogous to the brain. No external factor, other than the inputs, can affect the performance of the ANN. Thus there is no separate "training set". Learning instead becomes a process by which the neuron weights collect "experiences" from past inputs and "compete" for representation. This form of ANN is more robust against outliers, and particularly useful for locating new, unseen or overlooked, patterns within a dataset.

A particular unsupervised ANN has received particular attention, especially in the context of data classification: self-organizing maps (SOMs). Also known as a type of Kohonen network, this form of unsupervised learning is being used in a variety of fields and will be used in this thesis. In addition to being a very effective ANN, it also enables the user to easily interpret the results (outputs), contrary to many other ANNs. Ever since its first application in speech recognition [9], the algorithm has been used for a wide variety of problems, and has particularly flourished in astronomy over the past few years. For example, SOMs have been used for star/galaxy classification by Miller et al [11], and for galaxy morphology classification by Naim et al [13]. SOMs have also been applied to the problem of gamma-ray burst classification by Rajaniemi et al [16], an area we will revisit later in the thesis. More recently SOMs have also been applied successfully to the automated classification of light curves from interacting binaries by Brett et al [4].

The last 50 years have been demanding increasingly powerful computers. As

Djorgovski [34] says,

> *...the world is drowning in a tidal wave of data, which increases exponentially both in volume and complexity.*

With the amount of data growing in astronomy with a doubling time scale of ~ 1.5 years [34], the potential usefulness of ANNs is also increasing. Surveys like the Sloan Digital Sky Survey (SDSS) and the 2 Degree Field Survey (2DF) are already producing incredible amounts of images and spectra, and conventional data exploration techniques will become less and less feasible. Astronomers will find it harder to visually inspect most of this data and consider it for classification. The Wide Angle Search for Planets (SuperWASP), for example, uses ultra-wide-field surveys techniques to observe the sky, producing over 40 Gb per night. The analysis of this data must be automated, and ANNs can provide a well-tested, powerful and convenient approach to achieving this.

In the following chapter, we will present a detailed description of the two types of ANNs used. Both are Kohonen Networks, but one is the unsupervised Self-Organising Map (SOM), whereas the other is the supervised Learning Vector Quantisation (LVQ) scheme. In Chapter 3 we will apply both SOM and LVQ to spectra obtained from the SDSS in order to identify broad absorption line quasars (BALQSOs). In Chapter 4 we will concentrate on GRBs, and use SOMs with data taken from the Compton Gamma Ray Observatory (CGRO), to search for potential classes within the GRB population. The summary and conclusions will be presented in the last chapter together with some potential ANN applications in astronomy.

Chapter 2

Kohonen Networks

> *We are modelling our experiences all the time. Our thinking is based on mental images and ideas, which are projections of some internal representation from the brain to the exterior world. In that process our nervous system carries out modelling of various occurrences. In the history of mankind, mathematical modelling was first used in counting, then in geometry relating the land use and astronomy, and finally in all exact and even less exact sciences.* - Teuvo Kohonen (1982), Self-Organizing Maps

In this chapter we will describe in more detail the two ANNs used in this thesis. The unsupervised SOM will be introduced first, giving an informal explanation together with a more technical one. This algorithm will then be applied in both Chapters 3 and 4. The supervised version of the SOM, LVQ, will also be introduced, and this is used in Chapter 3.

2.1 Self-Organizing Maps: An Overview

The SOM, created by Teuvo Kohonen [35] in the early 80's, is a particular kind of ANN with the advantage of displaying the result to the user in an easily interpretable way. As described in the introduction, all ANNs are characterised by a layer of hidden nodes and output nodes, but the connections differ between different types of ANNs. In the SOM, the output nodes are connected in such a way that the results are presented on a 2D (or 3D) map of neurons. In this scenario, the hidden and output nodes are actually the same, as shown in Figure 2.1. In this "neuron map", after training, each individual neuron will be specialised at recognising some particular input pattern. Neurons representing similar patterns are located close to each other in the output map, which has the topology of a torus. In this way, there are no borders to the map, and every neuron is allowed to communicate with every other neuron. This last statement is key to the SOM, in that when one neuron is activated, and changed by an input, it will communicate this to all other neurons through weights, and thus affect even potentially distant neighbours.

The whole learning process starts by creating the weights responsible for the communication between the inputs and the map. This is done by randomising the initial neuron weights. The inputs are then presented one by one to the map. Each input stimulates a particular neuron (the most similar to the input) which will then take the responsibility of "learning" from the input and "teaching" neighbouring neurons. These processes are then iterated, i.e. the same inputs are presented to the map over and over. The main parameters affecting training are the so-called learning factor and neighbourhood kernel. The learning factor determines the amount "learned" by the neurons during each iteration, whilst the neighbourhood kernel is responsible for controlling who learns on the map. At first, all neurons will learn regardless of position on the map. As the iteration

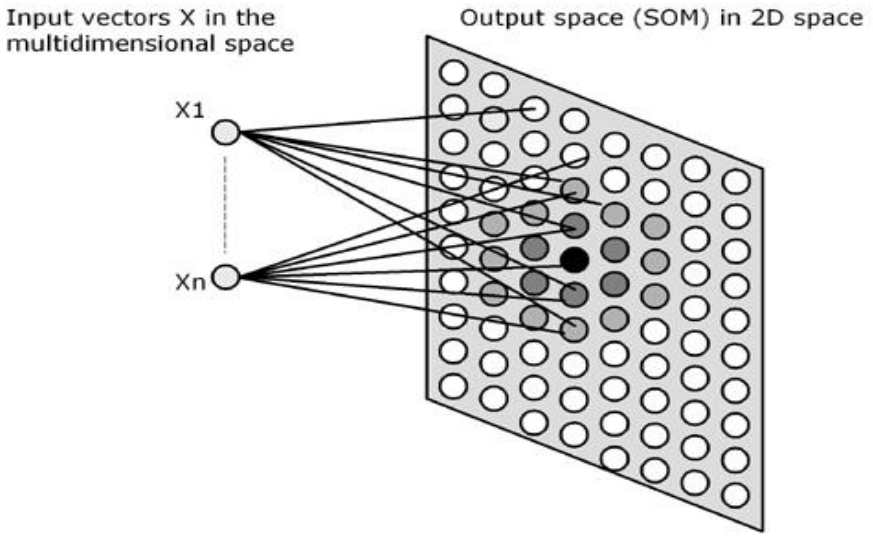

Figure 2.1: Schematic of the SOM. Note the hidden and output layers are the same for this ANN as taken from [29].

process proceeds, the learning factor decreases, and so does the neighbourhood kernel. By the final iteration only close neighbours of a stimulated neuron will learn. The combination of the two parameters ensures that the map will learn the gross structure of the input data early on during training, whilst focusing on the fine structure during the later stages.

The whole process can also be considered as a "model fitting" minimisation process by which the neurons are trying to find the best fit to the data. In doing so, the resulting neurons will act like composites to the data at the end of training. This is nice, in that the user will then have to only check the neuron weights rather than the input data, thus decreasing the amount of visual inspection to be carried out. In addition to performing an elegant regression, the neurons develop into specific "decoders" of their respective inputs and organise themselves on the map in a meaningful order, thus allowing the user to project and display higher dimensional data on a 2D grid.

2.2 Self-Organizing Maps: Technical description

In the following section, we will explore in more detail the equations that govern the SOM algorithm [35]. Moreover we will show some simple example runs to give a flavour of the abilities (and limitations) of the ANN for classification purposes.

Suppose we want to classify a large dataset, all consisting of n data points, $F(x) \in \Re^n$. The SOM should then be initialised accordingly, with neurons having weights with the same dimension as the input data, $m_i(x) \in \Re^n$, where i is the neuron index. These neurons should be organised on a torus-like map, so that no borders exist between them, and all are able to communicate. The number of these neurons is determined by the user, but one must be careful not to have too few of them. If the user knows in advance there are 3 evident patterns in a dataset, there is no point in running a SOM with 2 neurons: the result would be uninterpretable and wrong. It is always best to include more neurons than one would expect to need, as we will show later.

Once the neuron weights and input data are ready, we start by presenting the inputs to the map one by one. Each time we ask the ANN to locate the neuron most similar to the input data. This is done by fitting the input to all neurons and evaluating their norm[1]. Obviously, the most similar neuron will be the one with the lowest value. Let us assume this neuron to be defined by the index c in the following

$$\|F(x) - m_c\| = min_i \|F(x) - m_i\| \tag{2.1}$$

. This neuron is called the Best Matching Unit (BMU) for the specific input and will take the responsibility of "learning" and "teaching" other neurons. This is done by updating itself and others according to the following equation

[1]The norm is the Euclidean distance between two vectors (similar to χ^2).

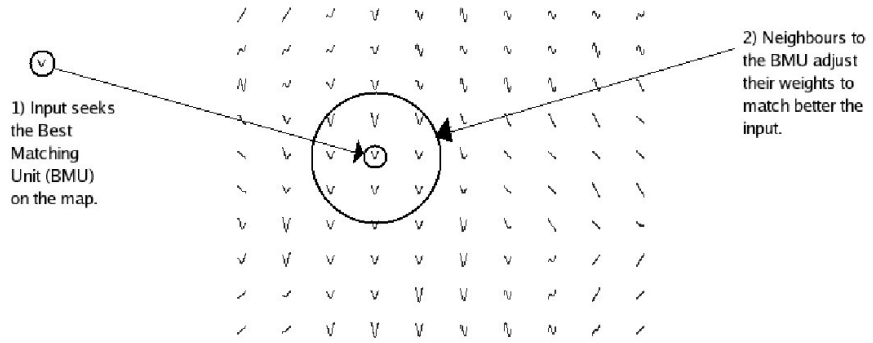

Figure 2.2: Snapshot of the neuron weights on a 64 node (8×8) map. The input seek the most similar neuron weight on the map. Neuron weights are then updated according to their position in map space: the closer to the BMU the more it's weight will be updated.

$$m_i(t+1) = m_i(t) + h_{ci}(t)\left[F(x) - m_i(t)\right] \tag{2.2}$$

, where $m_i(t+1)$ is the updated neuron and $m_i(t)$ the old one. Note the new parameter $h_{ci}(t)$, or neighbourhood kernel, which is responsible for the self-organisation of the whole map. This parameter is responsible for the amount learned at iteration t, together with how much a neuron learns given its distance in map space from the BMU. It in turn is controlled by two parameters, namely the learning factor $\alpha(t)$, and the width of the kernel $\sigma(t)$, both taken to be decreasing with time. The definition we use is

$$h_{ci}(t) = \alpha(t) \cdot exp\left(-\frac{\|r_c - r_i\|}{2\sigma^2(t)}\right) \tag{2.3}$$

, where r_c and r_i are respectively the coordinates in map space (2D) of the BMU and of the neuron under consideration. It is this last equation which enables the SOM to learn in a competitive environment, as neurons further away will learn less. This idea is sketched in Figure 2.2.

Since learning is a stochastic process, the final statistical accuracy depends

on the number of iterations. In general, the more, the better, but it is especially important to have a long final convergence, in which the map concentrates on fine-tuning the neuron weights. Note that compared to other ANNs, the algorithm is extremely efficient, and if only a few input samples are available they must be recycled for the desired number of iterations. As noted above, $\alpha(t)$ is conventionally taken to be a decreasing function of time, but this does not have to be the case. One can even keep the learning factor to a very low constant throughout the whole process, and only vary the neighbourhood width, thus ensuring that the fine structure of the data is considered throughout. On the other hand, $\sigma(t)$ cannot be constant. A good starting point is to set it at half the neuron map size, so that during the first iteration, all neurons will learn. This function must be decreasing with time so that by the end only close neighbours to the BMU will be updated.

Once training is ended, the neuron map is ready for inspection. In order to best interpret the results obtained by the SOM, we use the so-called U-matrix devised by Ultsch [24]. This has the same size as the neuron map and is usually used to visually identify statistically different clusters within the data set. It encapsulates the rate of change between neuron weights. In other words, each element within the matrix will represent the average goodness-of-fit between that neuron and its neighbours. In our case, each neuron has four neighbours, but if one adopts hexagonal neurons, it will have 6. This matrix can then be colour coded for ease of interpretation.

2.3 Self-Organizing Maps: Examples

In this section we will consider the ability of the SOM to organise four trivial cases: sine, cosine, negative and positive gradient curves. These will contain 30 points varying from 0 to 2π, and be in the range between -1 to 1. We will

create a dataset consisting of 500 inputs (125 of each), adding some noise to the curves in order to disguise them, and make it harder for the SOM to classify them. More specifically, we will add some normally distributed numbers to each data point in the input sample, increasing the standard deviation until the SOM has trouble in distinguishing the four classes. In other words, we will decrease the "signal-to-noise" (S/N) ratio of our sample until the data is to "noisy" to be recognised by the algorithm.

In Figure 2.3 we present some examples taken from four datasets created. The first (second, third and fourth) set has normally distributed numbers with a 0.4 (0.5, 0.7 and 0.85) standard deviation added on to them. As we will see the SOM will have no problem in distinguishing the first three, but by the last, the standard deviation added is too large, and the results become harder to interpret.

Having created our datasets, we prepare the neuron map and weights for each of the four runs. In each map, there will be 400 neurons (20×20), and each neuron will be randomly initialised with data points taken from a normal distribution with mean 0 and standard deviation 1. We initialise the neuron weights this way since we know in advance the average range of our inputs to be between -1 and 1 with some noise added on. By doing this we will help the map to quickly learn the gross scale structure of our data. This initialisation step is not necessary for correct convergence, but it allows us to keep the learning constant fixed to a very low value of 0.01 throughout. This means the map is always focused in on the fine structure present in the dataset and gives us a better feel for the ability of the ANN to self-organise. The width of the kernel on the other hand will begin as half the map size and decrease linearly over the course of the training phase, ending as half a neuron size.

In Figure 2.4 we show the four U-matrices generated by the four runs. We

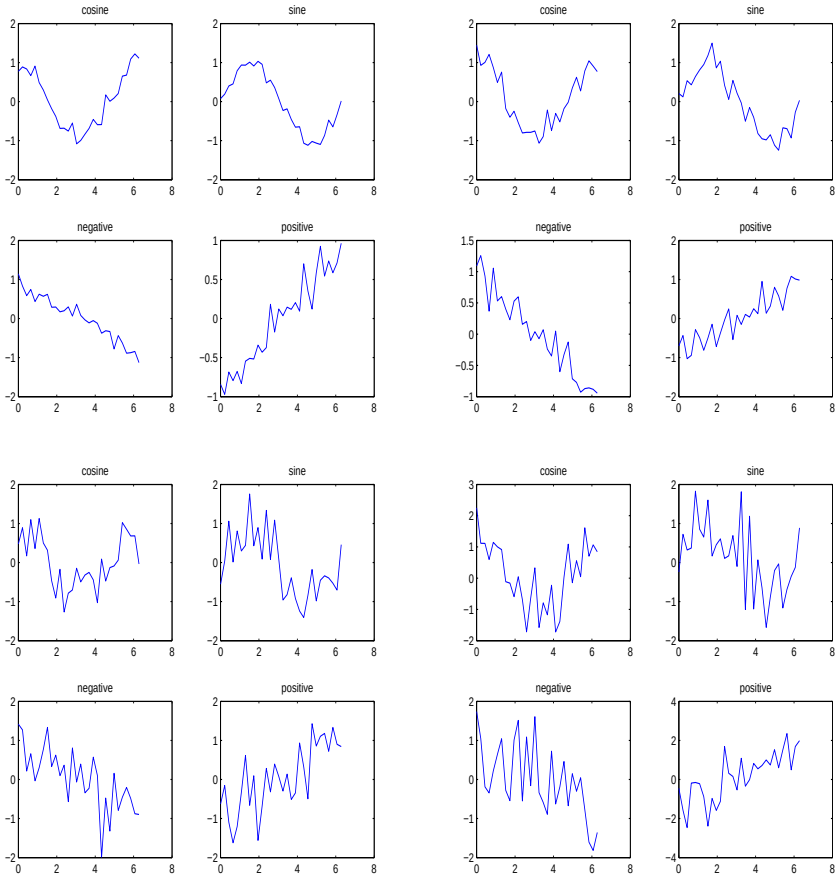

Figure 2.3: Four examples from each of the four datasets created. From top left to bottom right the quadrants are respectively 0.4, 0.5, 0.7 and 0.85 standard deviations curves.

note that all four used 10 learning steps, each of which had 500 files iterated through them. By visual inspection, it is clear from the first map that the SOM was able to distinguish all four classes easily - there are definite borders on the map and no outliers. Also note that since all four classes were present in equal numbers (125 of each), the four classes occupy more or less the same number of neurons in map space. Moving onto the map with 0.5 standard deviation, we find that this is very similar to the first, with definite borders and no outliers. However, at 0.7 standard deviations, the borders between the sine and negative gradient curves tend to disappear, since these two curves can look remarkably similar (as already seen in Figure 2.3). Note, however, that the four classes are still easily distinguished by looking at the density distribution of the mapped inputs (black dots). No outliers seem to appear and there is no confusion either. The last map presented, with 0.85 standard deviations, is a bit more confusing. The sine and negative gradient clusters have merged together, and one cannot tell by visual inspection of the map where the border lies. Moreover, the boundaries for the remaining classes have become blurred, with many inputs being mapped onto borders, without definite classification. It is clear from Figure 2.3, however, that not even a human "inspector" would have classified all these curves correctly, due to the low S/N.

2.4 Supervised Learning: Learning Vector Quantization

As outlined in the introduction, a supervised form of the SOM exists, namely LVQ. Also devised by Kohonen ([35]), this form of ANN is meant to be used for classification purposes when prior knowledge about the classes within a dataset exist. In such cases, one can use a pre-classified or visually tagged dataset to

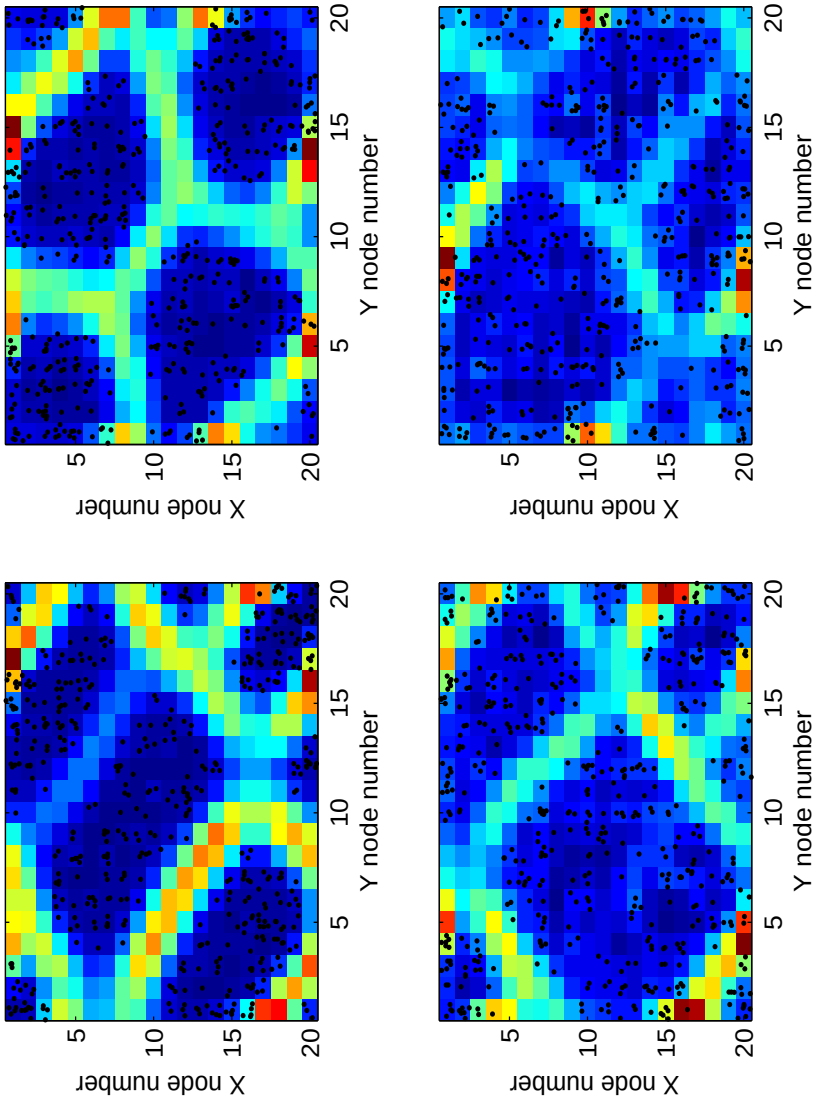

Figure 2.4: The four U-matrices produced by the four runs. The maps are in the same order as for Figure 2.3. The black dots represent the input files assigned, randomized within their host neuron.

train the LVQ, and then present a new dataset as input for classification. LVQ is based on a "reward-punishment" scheme[2] , but the ideas behind this ANN are very similar to that of SOMs. Competitive learning is still present, but the neurons are not allowed to communicate with each other, only with the inputs. We will now explore how the algorithm is structured.

We begin by creating a neuron map, but this time each neuron will have associated with it a tag referring to its class. This can be only two neurons for two classes or ten for each, depending on the feature one is trying to pick out. The number of neurons per class is predetermined by the user, and trial and error should be employed for best performance. Neuron weights can again be initialised at random, taking into account the range of the inputs, or can be initialised with an input from that class; both cases will converge. The only parameter affecting the performance of the algorithm is the learning factor α. This can be set at decrease linearly or kept fixed to a very low value. The effect is the same as for the SOM: learning the gross structure of the dataset with high α, and concentrating on the fine structure with low α.

As stated before, the neurons within the map have no connections between them, and no order exists except for the tags. Thus there is no topology associated with the map. The algorithm begins in the same way as the SOM. The inputs are presented to the neurons one by one, which will compete to find the BMU using the Euclidean distance as a metric. Once the best fit has been located, learning or unlearning takes place using the following equations:

$$m_c(t+1) = m_c(t) + \alpha(t)\,[F(x) - m_c(t)] \tag{2.4}$$

[2]The SOM is never "punished", as there exist only "learning" rules.

if $F(x)$ and m_c belong to the same class,

$$m_c(t+1) = m_c(t) - \alpha(t)\,[F(x) - m_c(t)] \tag{2.5}$$

if $F(x)$ and m_c belong to different classes,

$$m_i(t+1) = m_i(t) \tag{2.6}$$

for all other neurons. Note from equations 2.4, 2.5 and 2.6 that only the BMU will learn or unlearn each time an input is presented, leaving other neurons as they are.

Having described the algorithms used in this thesis, we will next apply such methods for BALQSO classification purposes.

Chapter 3

AGN, QSOs and BALQSO

> *Anyone who has never made a mistake has never tried something new* - Albert Einstein

This chapter will deal with the classification of broad absorption line quasars (BALQSOs). The main motivation for this comes from the recent work by Trump et al [23], which found a BALQSO fraction two times times higher than previous studies. This result is striking and needs further confirmation and analysis. However, first we will briefly introduce the quasar (QSO) population, connecting it to active galactic nuclei (AGN). We will then explore the various definitions of "breadth" in BALQSO classification work and closely inspect the catalogue produced by Trump et al [23]. Having established that their work needs revision, we will employ the techniques explained in the previous chapter to produce a more reliable BALQSO catalogue.

3.1 Introduction

In the 1950's, due to the fast technological advances in radio astronomy caused by the war, the first radio surveys were compiled[1]. These revealed a new population of bright, point-like objects that, at the time, due to the poor positional accuracy of the radio dishes, could not be identified with an optical counterpart. About 10 years later, the positional accuracy of radio telescopes was good enough for astronomers to identify the optical counterparts of these radio bright sources. What they found was something like a relatively blue, unresolved 'star' with what looked like a very faint emission around it. Subsequently astronomers found that the spectra of these objects exhibited very broad emission lines, but the point-like appearance of the objects obscured the connection to the already known Seyfert galaxies. At the time, the astronomical community agreed they had discovered a new type of radio-loud star. However, there was still a lot of confusion, since astronomers found these 'stars' not to be moving in position when comparing older photographic plates to each other. Then, in 1962, Maarten Schmidt stared at the optical spectrum of the famous Quasar 3C273 and recognised the Balmer series of hydrogen, but shifted from their normal wavelength by a factor of $(1 + z) = 1.16$. This implied the sources were of extragalactic origin. Very few people had thought of this due to the extremely high luminosities this would imply for these sources. To distinguish them from previous interpretations, these new types of objects were named quasi-stellar radio sources or quasars for short. In fact, it is now believed that quasars are part of the AGN (Active Galactic Nucleus) family. More specifically, these objects represent the high luminosity tail of the AGN distribution and outshine their host galaxies by a large factor. AGN have extreme luminosities $(10^{39} - 10^{47}$ erg/s) which put them amongst the most luminous objects known

[1] For a good review of Active Galactic Nuclei and their history refer to Mushotsky R. [33]

to date. Moreover, observed short timescale variations suggest that the only plausible energy production mechanism is the release of gravitational potential energy from matter deep within the potential well of the super-massive black hole (SMBH; $10^6 - 10^9 M_\odot$) of the galaxy.

Thanks to many multi-wavelength studies, AGN today come in many flavours, with sometimes very different observational characteristics (e.g. broad vs. narrow emission lines or radio loud vs. quiet). However, despite these differences, it is now thought that all of these objects are fundamentally similar, with differences reflecting mainly orientation effects, luminosity variations and differences in the relative luminosities of jets and disks. Quasars (QSOs[2]) also come in many flavours, mainly reflecting observational differences in their spectra. The resonant transitions named broad emission lines (BELs) in QSO spectra are believed to be formed deep within the potential well of the SMBH hosted by the AGN. This region will contain very high density gas moving at high velocities which, through thermal emission, produces the emission lines. The breadth of the lines is thought to be caused by the Doppler effect. In the rest-frame, UV/optical QSO spectra are reasonably well-described by a reddened power law with superimposed permitted emission lines. However, some QSOs show absorption troughs blue-ward of the strong UV resonance lines. These so-called BALQSOs are a sub-class of QSOs that exhibit strong, broad and blue-shifted spectroscopic absorption features. Most BALQSOs (the so-called HiBALs) only display absorption troughs in certain high-ionization lines (e.g. NV, CIV, SiIV), but, in a small percentage of BALQSOs (the so-called LoBALs) some low-ionization lines (most notably MgII) are also affected.

From the Sloan Digital Sky Survey Early Data Release, Reichard et al [19] showed that BALQSOs and BELQSOs appear to be drawn from the same parent

[2]QSO is an abbreviation for quasi-stellar object. Quasar is a short form of quasi-stellar radio source.

population, consistent with the idea that all AGNs are intrinsically similar, differing mainly in viewing angle. However, the dividing line between BALQSOs and non-BALQSOs is still somewhat non trivial. The main problem resides in distinguishing between narrow absorption lines (NALs) and BALs. In practice, Weyman ([25]) arbitrarily set the dividing line to be 2000 km/s in velocity space in order to distinguish the two.

For years, BALs have been regarded as signatures of large scale outflows since only this mechanism can account for the breadth of the absorption. There is some observational evidence (Brotherton et al [5]) that the outflows are predominantly equatorial (along the plane of accretion), but there are already hints that the true geometry may be more complex than a simple outflowing disk (Punsly et al [15]). The driving mechanism for the outflows is still not known, but the ghost of Lyman-α[3] seen in BALQSOs (Arav [3], North et al [14]) implies that radiation pressure mediated through spectral lines contributes in at least some BALQSOs.

In trying to understand the relationship between QSOs and BALQSOs, a quantity known as the BALQSO fraction is of particular significance. More specifically, the BALQSO fraction is defined as the fraction of QSOs that display BALQSO features. Its importance derives from the fact that it allows a simple, geometric interpretation: in the context of simple unified schemes, the BALQSO fraction is the covering fraction of the outflow. Thus an estimate of this fraction can provide strong constraints on the physical model of the accretion processes and associated outflows of AGNs and QSOs.

The biggest obstacle to measuring the BALQSO fraction reliably are NALs. Traditionally these have been thought to be caused by clouds of gas or by a corona orbiting the host galaxy of the AGN and were therefore not associated

[3] A hump manifesting itself at -5900 km/s in the troughs of BALs providing strong evidence for the importance of line driving in powering the outflows of BALQSOs.

with the AGN itself. These clouds might be moving due to the extreme radiation pressure caused by the AGN itself, manifesting themselves as NALs. On the other hand, Elvis [30] proposed a quasar unification model in which the NALs are formed in the same outflow that is responsible for BALs, with the observational difference caused only by viewing angle. Again, if this were to be true, it would elegantly unify all the three different kinds of phenomena, as shown in Figure 3.1. If the AGN is viewed directly through the flow, we see BALs; if, on the other hand, we look across the flow we would observe a NAL; and finally, if we do not view the central engine through any outflow, we simply observe a BEL. This model should therefore explain the incidence of BELs, BALs and NALs amongst QSOs as a result of the outflow geometry. It is therefore extremely important to determine the correct percentage of each subclass.

3.2 The P-Cygni Profile

Having introduced BALQSOs as part of the AGN family, we will now go through the physical processes that are responsible for both the emission and the absorption observed in these objects. The 'P-Cygni profile' specifically refers to broad blue-shifted absorption next to the line emission and is named after the first star in which this phenomenon was observed.

To best understand the physics, we consider the simple scenario of a spherical star emitting pure continuum photons into a non-rotating spherical outflow, and consider only line formation via pure scattering (e.g. by a strong resonance transition). Figure 3.2 from Knigge [1] shows the geometry of the situation. By removing the narrow 'cylinder' of outflowing material from the line of sight, one would only observe a BEL caused by scattering into the line of sight the symmetrical outflow, resulting in a broad emission line. However, with an outflow

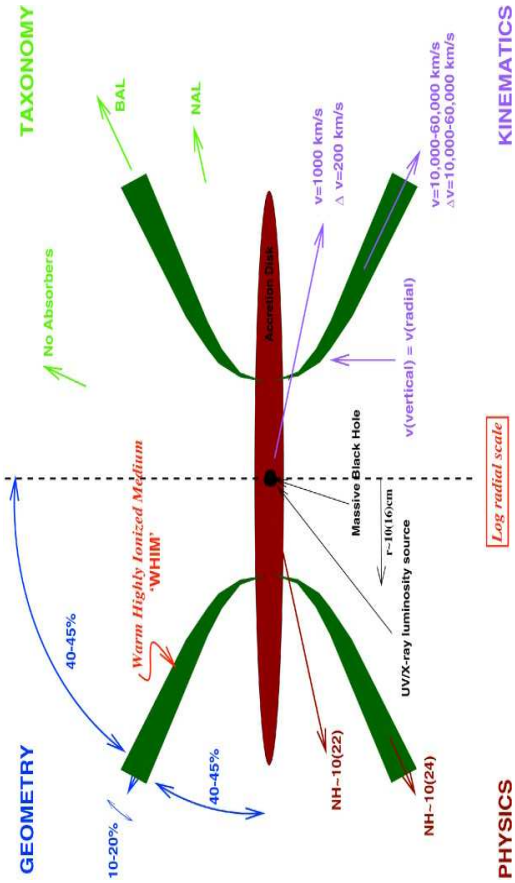

Figure 3.1: Schematic of the simple unifying model for BELQSOs and BALQSOs as taken from Elvis [30]. The red ellipse represents the accretion disk whilst the green regions represent the outflow.

present in the line of sight, the situation changes. The cylinder 'behind' the star is occulted, and photons scattering from it will never reach the observer. However, the outflow "in front of" the star will scatter continuum photons out of the line of sight of the observer. Now, because the outflow is moving towards the observer, we see an absorption blue-ward of the emission, whose width can also be used to infer the maximum outflow velocity.

Figure 3.3 shows a sketch of the line profile expected from such an outflow (Knigge [1]).

One must take care, however, since the observed absorption profiles seen in BALQSOs are far from being described by the spherical star description presented! The central engines of these sources are SMBHs, and it is thought that the emission components might be caused by thermal emission rather than scattering. Nevertheless, the broad, blue-shifted absorption trough would still be caused by viewing the continuum source through an outflow and should contain important information regarding the velocity field in which the absorption takes place.

3.3 The SDSS DR3 Quasar Catalogue

The Sloan Digital Sky Survey (SDSS; York et al [26]) is an imaging and spectroscopic survey which aims to provide the astronomical community with immense amounts of data. In particular, it concentrates on the large scale distribution of galaxies and quasars. The survey is carried out using a CCD camera on a dedicated 2.5 meter telescope at Apache Point observatory, New Mexico. Images in five broad optical bands (ugriz) over approximately $10,000$ deg^2 of the high Galactic latitude sky in the Northern hemisphere are taken. The catalogue contains photometry from 136 different imaging runs between 1998 and 2003 and spectra from 826 spectroscopic plates between 2000 and 2003, covering a

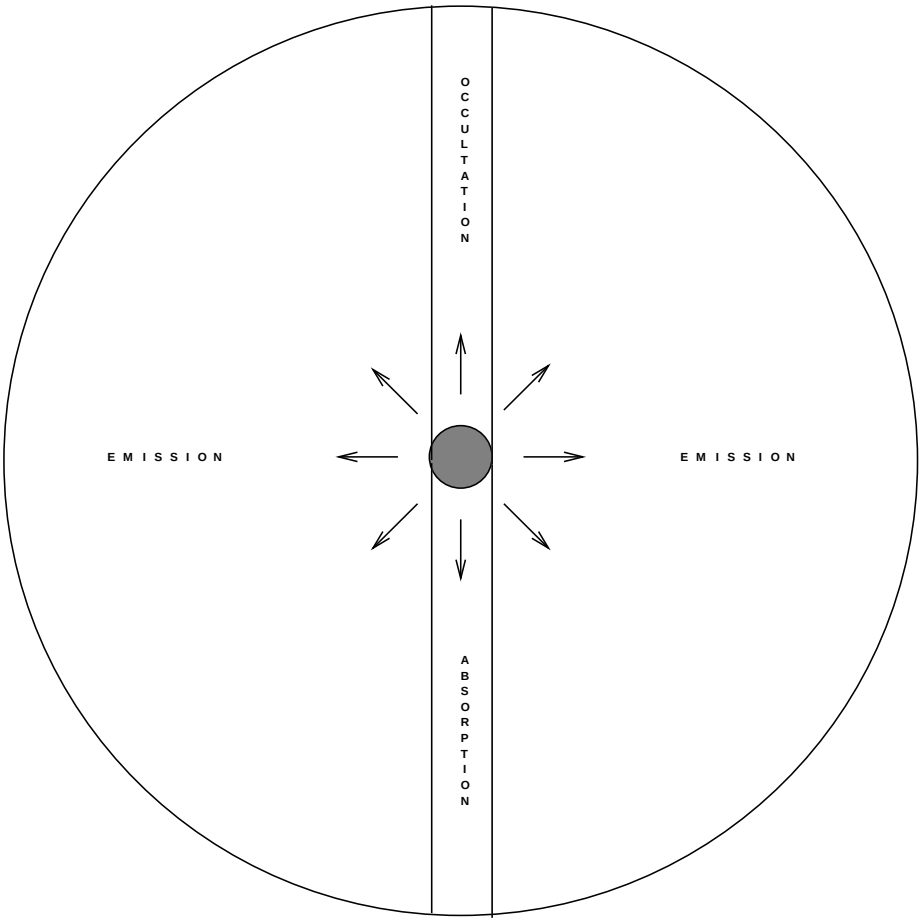

Figure 3.2: Sketch of the spherical outflow causing a P-Cygni profile as taken from Knigge [1]. The direction of the observer is down. Note the column of material flowing towards the observer causing the blue wing absorption.

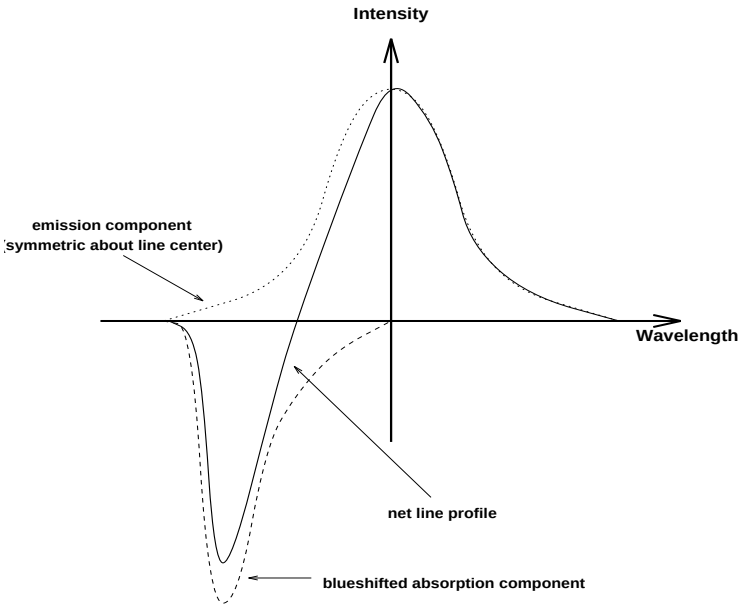

Figure 3.3: Typical P-Cygni profile decomposed into underlying absorption and emission components as taken from Knigge [1].

smaller fraction of the sky (≈ 4188 deg^2). All spectra cover a wavelength range of 3800-9200 Angstroms (in the observed frame). Most of the QSOs found in the SDSS have been identified based on a colour selection algorithm that uses all 5 optical bands. More details of the spectroscopic observations can be found in York et al [26].

All of the SDSS quasar spectra have been wavelength calibrated and sky subtracted, but not corrected for Galactic extinction. A redshift estimate is also determined automatically, by cross-correlating the quasar spectra and fitting emission lines to some quasar template.

The number of QSOs found by the SDSS has increased by a factor of ≈ 12 with Data Release 3 (DR3) [20] compared with that in the Early Data Release [17]. This nicely illustrates the need for automated methods when dealing with such a rapid growth in datasets. The DR3 catalogue contains 44221 spectroscop-

ically identified QSOs. Moreover, one can see that robust and reliable methods are also needed since more outliers (and maybe even new sub-populations) will appear. With such a rapid growth, finding ways to reduce computation time and minimise human input to the analysis become important. One would preferably like to spend as little time as possible reducing and classifying the data and avoid having to look at specific outliers and decide which class best fits them.

3.4 Trump et al's BALQSO Catalogue

In March 2006, Trump et al [23] released a data release 3 (DR3) BALQSO catalogue consisting of 4787 BALQSOs out of the 16,883 QSOs (26%) that had the CIV and MgII in the observed spectra. Compared to the Early Data Release (EDR), this is an increase in the BALQSO fraction by over a factor of two. This is mainly due to a new metric adopted in defining BALs. In this section we will first describe the metrics used for BALQSO recognition, then analyse the results obtained by Trump et al and finally assess the reliability of the metrics adopted.

3.4.1 BALQSO Metrics

In the EDR catalogue by Reicard et al [18], the definition of BAL adopted was that of Weyman et al [25], namely the Balnicity Index (BI). This definition came about because of the difficulties in distinguishing BALs from other absorbed QSOs. In particular, narrow absorption line quasars (NALQSOs) exhibit narrow absorption features ($\approx$ 1000 km/s, usually detached from the emission line), making it difficult to distinguish them from BALQSOs. The dividing line between the BALs and NALs was arbitrarily set to 2000 km/s. The BI index is determined as follows:

- Concentrating on CIV emission line, define a continuum as sensibly as

possible between the rest wavelength of CIV and SiIV emission lines.

- Define the systematic rest frame as accurately as possible
- Compute the BI (modified equivalent width of strong absorption in km/s) using the following equation

$$BI = \int_{25000}^{3000} \left[1 - \frac{f(v)}{0.9}\right] C dv \tag{3.1}$$

, where $f(v)$ is defined to be the normalised flux as a function of velocity displacement from line centre. The value C is binary and therefore can only take the values 0 or 1. It is initially set to zero, and is turned to 1 whenever the quantity in brackets has been continuously positive for over 2000 km/s. However C is turned back to zero when the quantity in brackets turns negative again. A BI of zero means no absorption, whilst any QSO with BI greater than zero is considered to be a BAL. Thus the BI gives us a feel of how much broad absorption is present. It turns out that the BI metric is very good at identifying BALs, but one can see that the definition is very conservative and some BALs will be missed (as we will see later in the chapter). Moreover, BALs exhibiting the ghost of Lyman-α [14] may be missed by the BI metric, and wrong redshift determination or bad continuum fitting can also cause missclasifications.

Trump et al [23], on the other hand, adopted a new metric named the Absorption Index (AI), devised by Hall et al [7]. They modified the BI metric slightly, so as to measure all absorption within the limits of every trough, and extended the integration limit to 29,000 km/s. The formal definition is given in the following equation

$$AI = \int_{0}^{29000} [1 - f(v)]\, C dv \tag{3.2}$$

Obviously, as for the BI, the spectrum needs to be normalised. Again in this case, the value C is binary, but it behaves differently. C is set to zero except

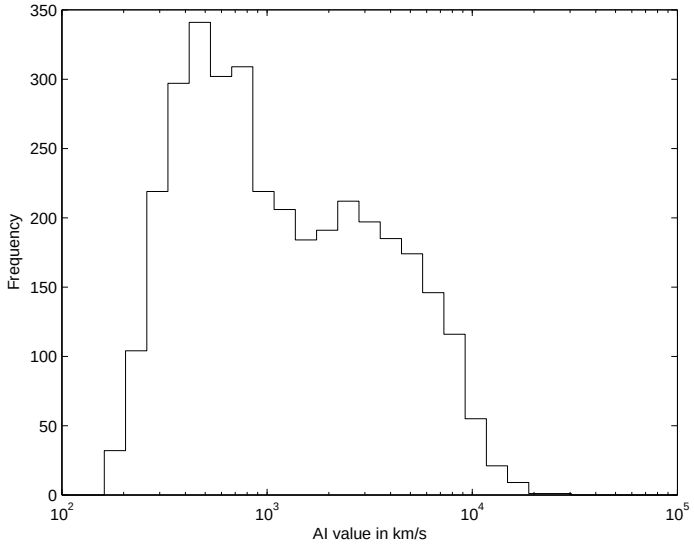

Figure 3.4: Distribution of QSOs with AI>0 from Trump et al [23].

in contiguous troughs which exceed a minimum depth of 10%, and a minimum width of 1000 km/s, in which case it is set to one. By adopting this extra definition, the SDSS team have increased the numbers of BALs in DR3 from 1756 (10.4%) which satisfy BI>0 to 4386 (26.0%) which satisfy BI>0 or AI>0. This is not surprising: the AI is much less conservative than the BI, since the minimum width is only 1000km/s and the index extends to 29000km/s. Having said this, the DR3 BAL catalogue does not address this as a problem at all. Moreover Trump et al [23] discard any QSO with BI>0 and AI=0 (although they still give BI values for each QSO).

When defining the AI, Trump et al [23] chose 1000km/s as their minimum width, after having also tried 450km/s and 750km/s. Their preference for 1000 km/s threshold was based on the fact that the other two cuts were too liberal (but they did not try larger widths).

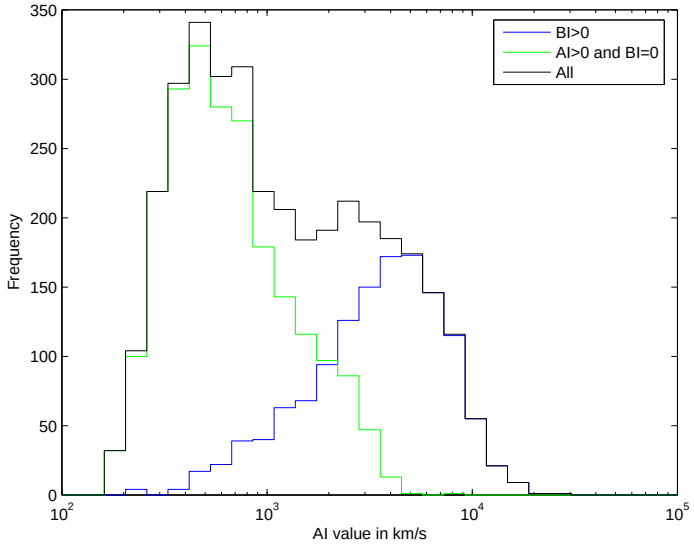

Figure 3.5: Same distribution as in Figure 3.4 broken down for objects with BI>0 in blue and the rest in green.

3.4.2 Trump et al's results: a closer examination

Figure 3.4 shows the AI distribution for QSOs taken from Trump ([23]). It is formed by a population of 11646 QSOs, all of which contained the CIV line from which the AI was measured. The distribution is clearly bimodal. We note that Trump et al [23] produced a similar plot on a linear velocity scale up to 4000km/s. They did not notice this bimodality since the linear scale was inappropriate for such a large range of AI's. The distribution is roughly split between objects with BI>0 and the rest, as shown in Figure 3.5. This result is striking and suggests that the most recent BAL catalogue contains far too many objects.

We will now inspect more closely spectra from the extremes of the distribution together with spectra from the intersection. Figure 3.6 shows four particular objects from the bulk of the green distribution (also the bulk of the whole dis-

tribution) with an AI≈400km/s. It seems clear that none of these objects have any proper CIV absorption in them, never-mind BALs.

On the other hand, objects with an AI≈4000km/s and a BI>0 seem to contain most of the properties of genuine BALs, as shown in 3.7. Clearly, the AI on its own is inappropriate for BAL recognition. Over half of the objects in the whole distribution are contained in the low-AI region of this bimodal distribution, so serious revision of the catalogue seems necessary.

Let us now inspect the distribution containing objects with BI>0. In Figure 3.8 are presented objects from the low velocity tail of the blue distribution. Again there are no evident signs of absorption. We note that these QSOs have always been included in BAL catalogues.

On the other hand the bulk of the blue distribution does contain genuine BALs, as the ones taken from Figure 3.7. The BI is by no means perfect though. As we see from Figure 3.9, some genuine BALQSOs would have been missed if one had relied solely on the BI.

It is clear that further examination is needed. Next we will try and use some of the methods explained in the previous chapter to help us mine the data and produce a new, more appropriate, BALQSO catalogue.

3.5 Data Selection & Normalisation

Preconditioning the data is of key importance when employing SOMs. If one does not account for redshift, for example, the SOM might classify on that, giving no importance whatsoever to spectral shape. On the other hand, if one does not take into account dust reddening, the SOM might classify on the spectral index, ignoring the absorption issues we are interested in. This next section will explain the procedures adopted to ensure the SOM (and later LVQ) have the best chance to identify BALs.

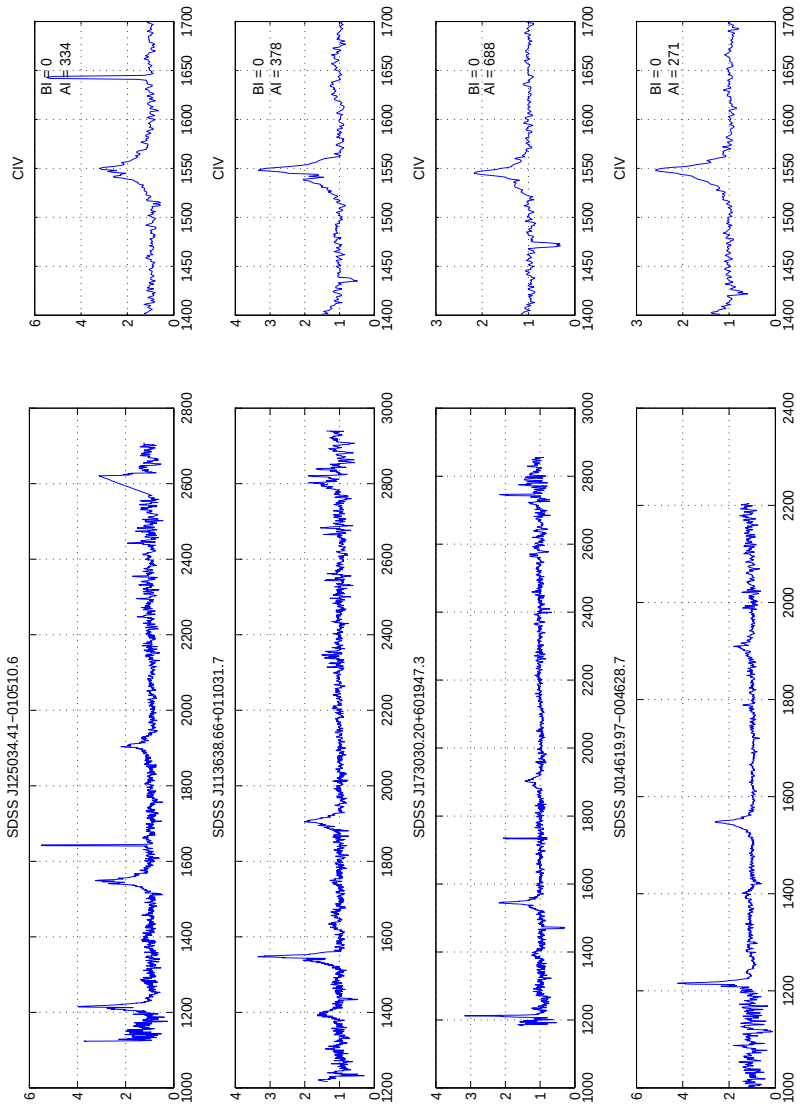

Figure 3.6: Four QSOs from the distribution with AI≈400km/s.

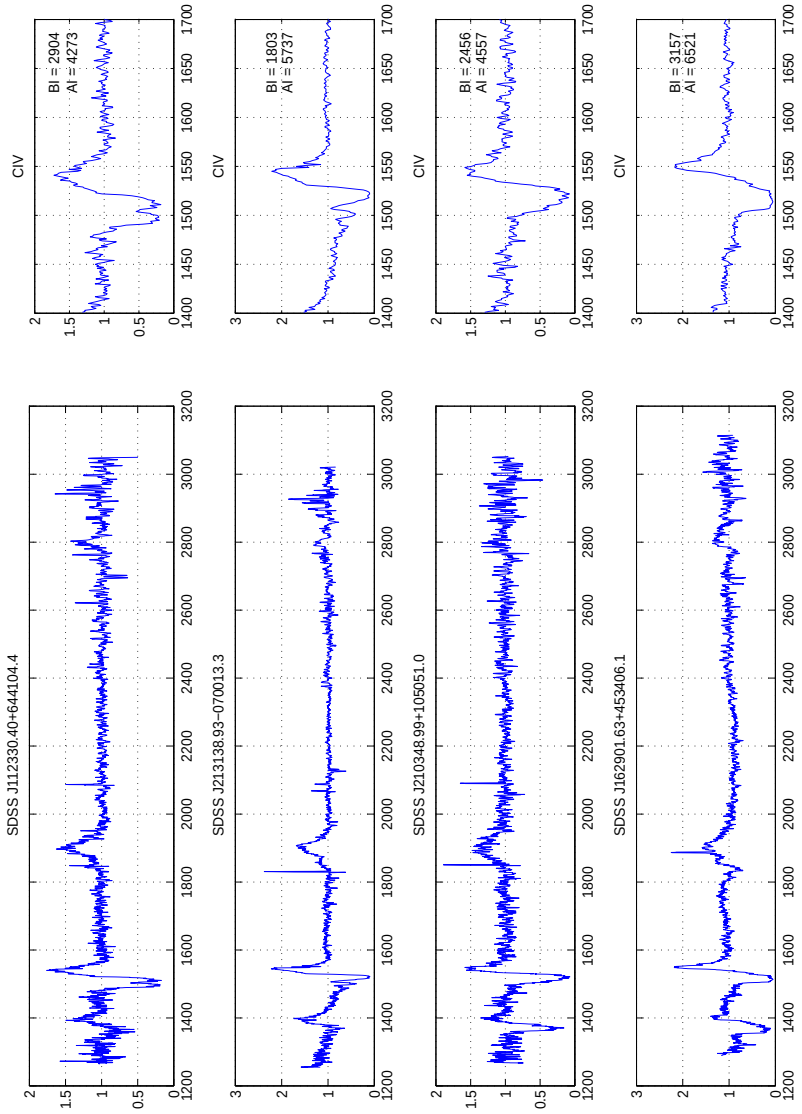

Figure 3.7: QSOs with AI≈4000km/s and BI>0.

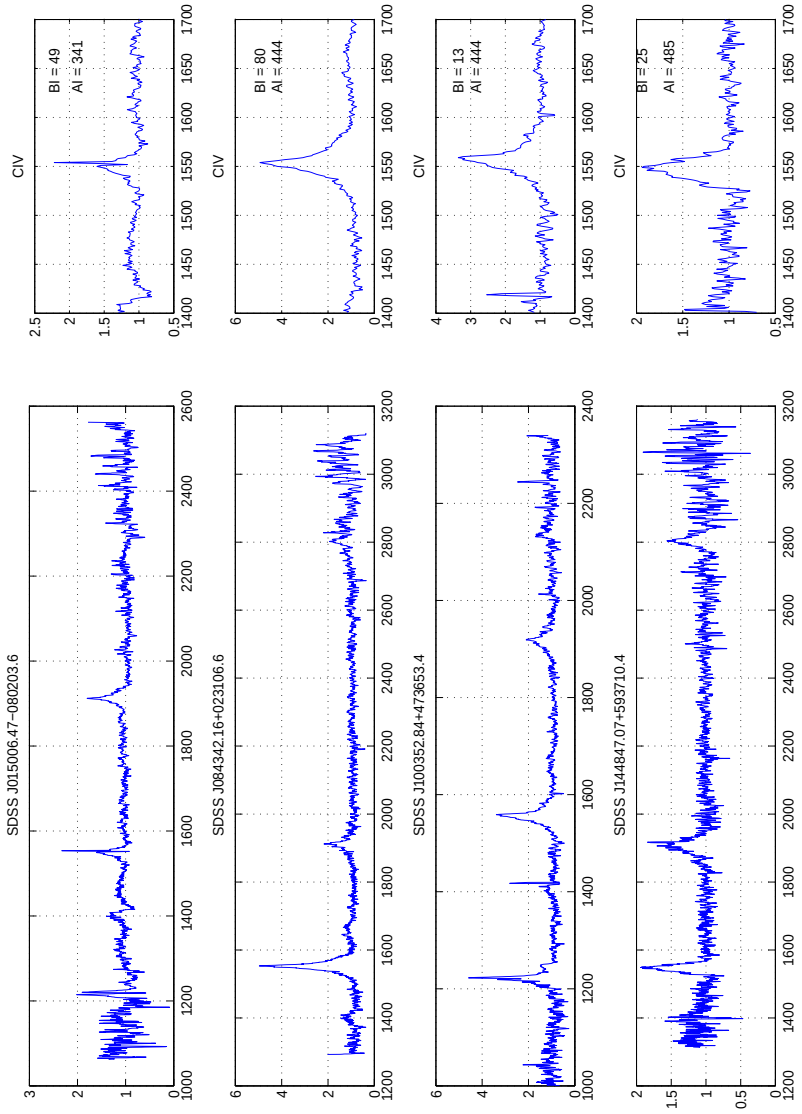

Figure 3.8: QSOs with AI≈400km/s and BI>0.

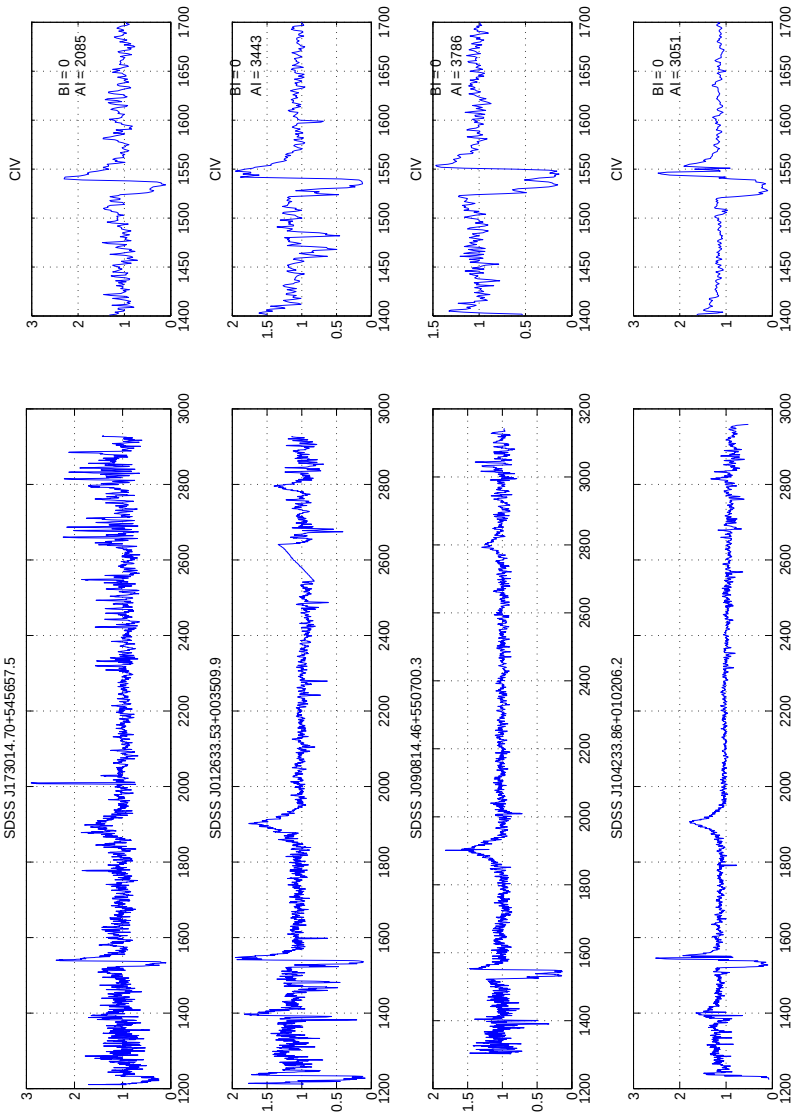

Figure 3.9: BALs with BI=0.

From the 44221 QSOs in [20], we impose a redshift cut of $1.90 \leq z \leq 4.94$ to ensure the presence of the CIV region where we will check for broad absorption. This yields 11646 QSOs which will be carried forward for normalisation.

The first thing that needs to be done is correcting for host galaxy extinction. This affects almost all the spectra of extra-galactic objects. Dust grains within galaxies preferentially scatter wavelengths with increasing intensity at shorter wavelengths. The result is that the observer will see more flux from the red end of a spectrum than from the blue. Although this is unlikely to alter the small scale structure of the spectrum, this phenomenon can definitely alter the slope of the continuum quite drastically, and the ANNs might classify on this rather than on absorption. We will therefore account for this by using the extinction curve tabulated by Pei [37]. Pei examined three particular cases, namely the Milky Way, the SMC (Small Magellanic Cloud) and the LMC (Large Magellanic Cloud). The values obtained for the SMC are commonly used to account for dust reddening in distant QSOs (e.g. [23]), and we will follow this practice here.

Pei's empirical law for dust reddening is expressed as

$$\xi(\lambda) = \left[(E_{\lambda-V}/E_{B-V}) + R_V\right] / \left(1 + R_V\right) \tag{3.3}$$

where $E_{\lambda-V} = A_\lambda - A_V$ is the colour excess in magnitudes, $R_V = A_V/E_{B-V}$ is the ratio of total-to-selective extinction, with the subscript V indicating the visual photometric band. For the SMC, $R_V \approx 2.93$, whilst $E_{\lambda-V}/E_{B-V}$is tabulated in Pei [37] as a function of λ^{-1}.

We now turn our attention to the composite fitting method employed here to normalise the spectra. We assume that all QSO spectra, $f_{int}(\lambda)$, are intrinsically the same, but observed to be at different distances, having different continuum slope characteristics and being differently absorbed. This can be expressed as

$$F_{obs}(\lambda) = K \cdot f_{int}(\lambda) \cdot \lambda^{\alpha} \cdot 10^{-aE \cdot \xi(\lambda)} \tag{3.4}$$

where $a = 0.4(1 + R_V)$, $\xi(\lambda)$ is the Pei SMC extinction curve, λ is the rest wavelength, E is the host galaxy extinction E_{B-V}, K is the constant of proportionality (responsible for distance) and α the spectral index (responsible for reddening). With this assumption, we can create a geometric mean composite of all QSO spectra, whilst still retaining $f_{int}(\lambda)$ as before. This is defined as

$$F_{comp}(\lambda) = K_{comp} \cdot f_{int}(\lambda) \cdot \lambda^{\alpha_{comp}} \cdot 10^{-aE_{comp} \cdot \xi(\lambda)} \tag{3.5}$$

where now α_{comp} and E_{comp} are the arithmetic means of all spectral indeces and host galaxy extinctions in the sample from which the composite was constructed. Subtracting equations 3.5 from 3.4 in log space for all QSO spectra allows us to solve for the constants $\hat{\alpha} = \alpha_{obs} - \alpha_{conp}$, $\hat{E} = E_{obs} - E_{comp}$ and the constant of proportionality $Log(K_{obs}) - Log(K_{comp})$ without actually requiring a formal definition for $f_{int}(\lambda)$. This will then allow us to fit a continuum to the spectra.

The continuum windows adopted here for fitting purposes are those used by North et al [14]. MATLAB was employed for the fitting procedures.

3.6 SOMs application to SDSS

Here we will present our first attempt to apply the methods explained in Chapter 2 for the purposes of BAL recognition. Having normalised our spectra, we focus on the CIV region on which to perform our training. We do this by choosing the rest wavelength region between 1400-1700 Angstroms and smooth it to 1 Angstrom per pixel to ensure the spectrum within this region has the same number of pixels for all QSOs.

We then trained different SOMs with different sets of training data and

different initial conditions. In doing so, we had to account for possible errors in the redshift determination of the QSOs. Without accounting for this, the map would have probably classified on redshift (or more appropriately redshift error) rather than spectral shape. We account for possible redshift errors as follows. During training, each QSO spectrum is compared to each neuron in the map one at a time. To find the best possible fit to the neuron, we allow some shifting of the QSO spectra relative to the neuron. This is done by comparing the neurons to the inputs, allowing for a ± 15 Angstrom shift on the red side of CIV (1535-1575 Angstroms). Only the red side of CIV has been chosen so as to not fit any absorption present on the blue side[4].

A lot of computing time was spent trying to produce a clean map. However, this goal was not achieved. The least confusing map produced was trained on 2000 QSO spectra with a very low learning constant of 0.01. The computation took a day or so, and the map is presented in Figure 3.10.

We inspected the map neurons and established that the top blue cluster is that of BALs. However the U-matrix produced no definite boundaries to the cluster, and no classification was possible with this method. The main reason for this is caused by the metric adopted here to define QSO, the Euclidean distance. This is best illustrated visually. From Figure 3.11 one can see that the blue and red spectra are both BALs, yet in terms of Euclidean distance the blue and green spectrum are more similar. This issue consistently confuses the map. In fact, this problem arises often with SOMs: preconditioning of the data is crucial in identifying clusters. On the other hand it could also be that the two QSO populations are continuous from being absorbed to being non absorbed.

[4]We tried to shift all the spectrum but soon realised that some QSOs were being fit in the absorption window rather than emission.

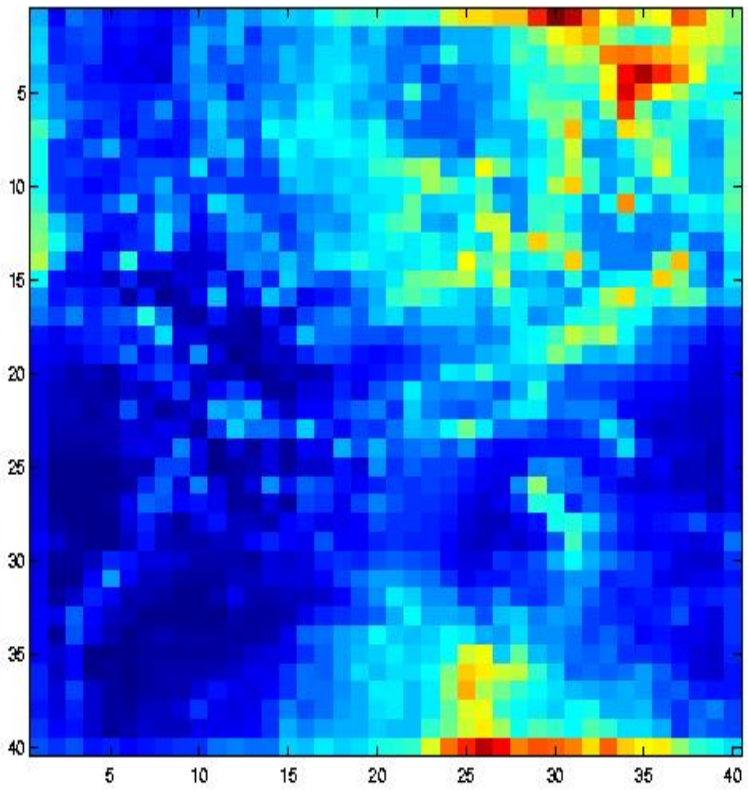

Figure 3.10: SOM trained on 2000 QSOs with $\alpha = 0.01$ throughout. The top blue cluster is that of BALs, however no definite boundaries were produced, and no definite classification was possible.

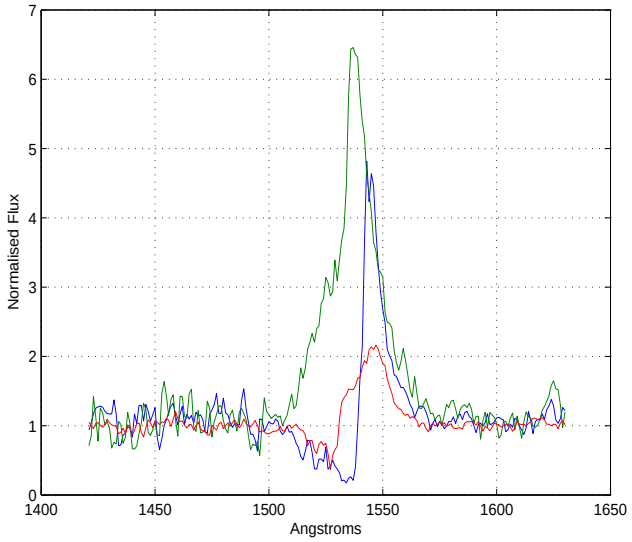

Figure 3.11: Three QSO spectra to demonstrate the Euclidean distance problem. In this case the blue and green spectrum are most similar, however being that only the blue and red are BALQSOs.

3.7 LVQ Classification

Having tried the unsupervised SOM, we now attempt classification using the supervised LVQ. As explained in the previous chapter, this method is noticeably more time consuming as the user needs to tag a training data set. Moreover, if errors may exist within the tags, the user will have to asses the appropriateness of the results. At the same time, however, this method allows us to overcome the problem we had when dealing with SOMs, as we will show later.

3.7.1 Training

The first thing to be done is create a training set for the ANN to work on. As a first step it seems plausible to tag QSOs with BI>0 as likely BALQSOs, since most objects with this index are indeed genuine BALQSOs. We therefore cre-

ate a training set composed of 400 QSOs with BI>0 and 400 QSOs with BI=0. In this way, the map will learn to recognise objects with BI>0, which contain mostly BALs. The key advantage this method retains over a pure BI classification is that the user can ultimately check the map neurons for misclassification and correct them manually. The LVQ can therefore be considered as a mixture of human expert classification and BI classification.

Our LVQ map contains 150 neurons, 75 dedicated to objects with BI>0 and the rest to BI=0. At the start, each neuron contains a random QSO from the training set, with wavelength range between 1420-1630 Angstroms, like in the SOM case. The problem we had with SOMs, displayed in figure 3.11, is not an issue anymore, since within the 75 neurons dedicated to BALs we can have various combinations of spectral shape (e.g. strong emission and weak absorption and vice versa). We then employ the LVQ algorithm on the 800 QSOs for 3000 iterations, keeping the learning constant fixed at $\alpha = 0.01$, allowing for the 30 Angstrom shift, as described earlier. In Figure 3.12, we show the average Euclidean distance for each iteration. Convergence was achieved after about 500-1000 iterations. Training took over a day, and the final map weights are presented in Figure 3.13.

Since we know the BI is not a perfect metric, we now need to inspect the neurons within the map. We first search for BAL neurons that contain QSOs with BI=0 in the final map. This yielded 7 neurons. Moreover, visual inspection of the BAL side reveals 6 extra borderline cases. We therefore examined 13 neurons in total on the BAL neurons, together with most QSOs within them. However, we find that only three BAL neurons[5] have been "misclassified" on this side, in the sense that they have learned to recognise non-BALs with BI>0.

We now turn our attention to the non-BAL neurons, here asking the map to return neurons containing objects with BI>0. This yielded 14 neurons plus,

[5]([row,column], [4,5], [5,3] and [3,3])

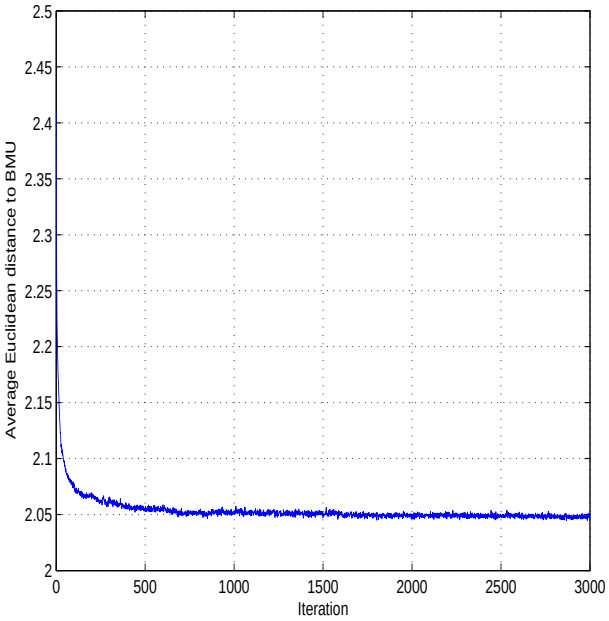

Figure 3.12: Average Euclidean distance plot for the LVQ run

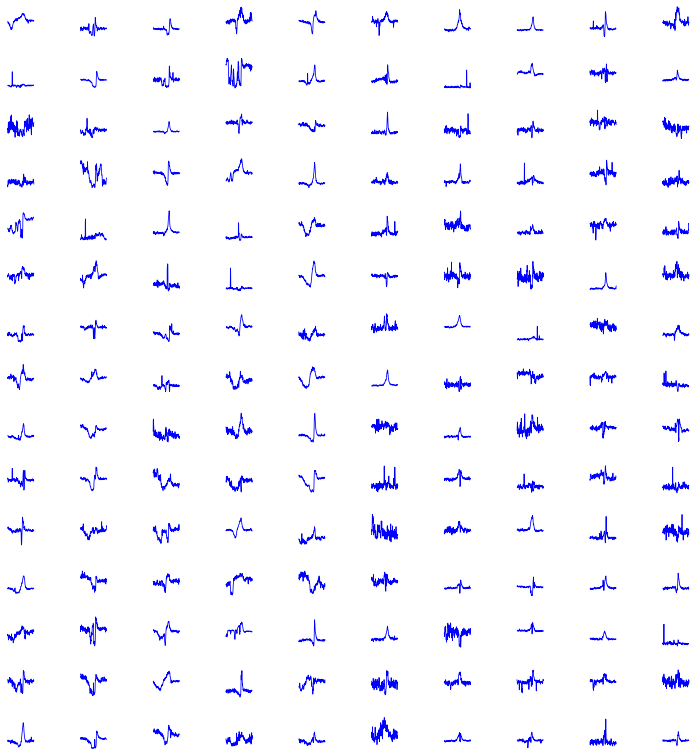

Figure 3.13: Map weights. The upper-left neuron is [1,1] whilst the bottom-right is [15,10]. Columns 1 to 5 were tagged as BI>0 whilst columns 6 to 10 as BI=0.

12 more we decided to check as borderline cases. This is a bigger number than before, and is probably due to the fact that the BI metric misses quite a few BALs (as we have seen from Figure 3.9). Inspection reveals that only 7 of these neurons[6] have been genuinely mislabeled (i.e. contain predominantly BALs). It is worth stressing that this is actually a good thing: we now have known neurons in the map that have actually learned to recognise BALs with BI=0.

3.7.2 Results

Having trained the map and corrected it, we can now easily classify the remaining QSOs (10846). In order to assess the performance of our algorithm, we will now locate some of the spectra presented previously (Figures 3.6, 3.7, 3.8 and 3.9). We note that these spectra have not been included in the training set. We find that all these spectra are classified correctly.

We are now ready to compile our own final BAL catalogue using our LVQ map. This contains 1208 (10.4%) BALs out of the 11643 QSOs. It is interesting to consider the AI distribution of the objects we classify as BALQSOs and "normal" QSOs. This is presented in Figure 3.14. Clearly the skewness of the distributions has diminished, although a low velocity tail of BALs still remains. This might in fact be a characteristic of BALs and will be examined further in future work.

3.8 Conclusion

We have used LVQ to search the DR3 QSO catalogue for BALQSOs. We find a BALQSO fraction of 10.4% for objects exhibiting CIV emission. Our result is in close agreement with that of Reichard et al [18] (based purely on BI), but disagrees with Trump et al [23] (based purely on AI). We think that ours is

[6]([15,10], [12,8], [4,9], [6,7], [12,10], [14,6] and [13,7])

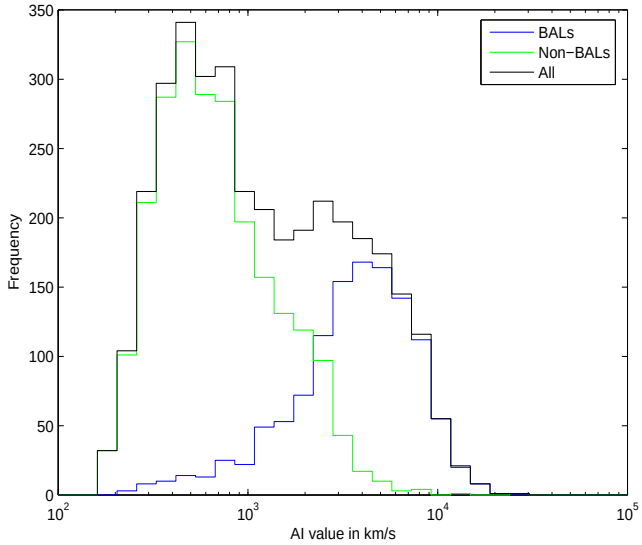

Figure 3.14: Final AI distribution for BALs according to the LVQ classification. To be compared with Figure 3.5. Particulary the blue distribution shows the true distribution of BALQSOs within DR3.

to date the most complete and reliable catalogue of BALQSO systems in the redshift range $1.90 \leq z \leq 4.94$. The neuron map will be posted on the web[7] in form of a MATLAB .mat file together with the catalogue in ASCII format. This can then be used for future reference on the release of future, more complete, QSO catalogues.

[7] http://www.astro.soton.ac.uk/~simo

Chapter 4

Mining Gamma-Ray Bursts

All truths are easy to understand once they are discovered; the point is to discover them - Galileo Galilei

In this chapter we will attempt to use SOMs for mining the GRB data obtained from the Burst And Transient Source Experiment (BATSE) on board the Compton Gamma-Ray Observatory (CGRO). We will first give a brief introduction to GRBs, together with a brief literature review regarding their classification. Next, we will describe the catalogue used together with the background reduction techniques employed to try and extract light curve shape dependent variables. The results from the SOMs will show us a distinction between single-pulsed bursts (SPBs) and multi-pulsed bursts (MPBs), which will be investigated further. We conclude with some suggestions for future work.

4.1 GRBs: An Introduction

Gamma-Ray Bursts are intense and short (0.1-100 seconds) bursts of gamma ray radiation occurring on average once per day at cosmological distances from

Earth[1]. Their distribution in the sky is isotropic. They were first detected by the Vela satellites in the 1960's. Since then, thousands have been observed by missions such as CGRO, BeppoSax, High Energy Transient Explorer (HETE), Konus and many more. Because of the difficulty in detecting such events, GRBs remain poorly understood, with little progress being made until the advent of the Compton Gamma Ray Observatory (CGRO) which proved the isotropy of GRBs, Beppo SAX finding a lot of the optical counterparts and the GRB chaser SWIFT. GRBs are the most luminous sources known, for a few seconds outshining the entire Universe in gamma-rays. Hence GRBs can be seen out to large redshifts, and are thereby strong cosmological probes. They illuminate the intergalactic medium (IGM), and can give us information about star formation, galaxy evolution and the chemical enrichment in the early Universe.

The first detector to produce a statistically meaningful sample of GRBs was BATSE, with over 2700 triggered bursts. The analysis from the duration distribution of such bursts turned out to be bimodal ([10]), with an empirical split at ~ 2 seconds. Since then, the GRB population has been classified into two main groups: short GRBs and long GRBs, with a lot of speculation regarding their origin. However, it is clear that short GRBs are energetically weaker (by a factor of ≈ 100) and have only been seen at relatively low redshift mainly in elliptical galaxies. Indeed some short GRBs may in fact be soft gamma-ray repeaters (SGRs).

For many years, multivariate studies have been employed to study these classes and potential others. Variables such as duration, fluence, maximum peak energy and hardness ratio[2] have been used. In particular, Mukherjee ([12]) suggested the existence of an "intermediate" class between the short and long bursts using such variables. A similar analysis with a different algorithm was

[1]For an excellent review of Gamma-Ray Bursts physics and history refer to Mezaros P. [32]
[2]The hardness ratio is the ratio of the fluences obtained in two different energy channels.

then carried out by Hakkila et al ([6]), with the conclusion that the "intermediate" class is not necessarily a distinct source population. Hakkila also pointed out the fact that this third class could also be an artifact caused by analysis errors. Rajaniemi et al ([16]) employed SOMs to classify the data, using similar variables as Mukherjee and Hakkila, but a different algorithm. Their results, like Hakkila's, suggested the existence of two distinct GRB populations, but did not support the finding of Mukherjee. We note that this last analysis was the first to rely on an unsupervised ANN.

Due to the many multivariate analysis carried out already, we will concentrate solely on light curve shape to try and distinguish possible GRB populations. This kind of analysis is less susceptible to systematic errors either caused by the detector(s) or the analysis. In multivariate studies one usually needs to obtain data from different detectors, increasing the inhomogeneity of the dataset. Concentrating only on light curve shape will yield us a more complete and homogeneous GRB population. Moreover, concentrating solely on light curve shape will allow us to examine variables which have not been included in the analysis. We will employ the unsupervised SOM algorithm for this, but, as usual, we will first have to precondition the data. The information regarding GRB strength and duration will have to be removed after having subtracted the appropriate background.

4.2 The Data & Preconditioning

The data set used within this work is that compiled by Stern et al. ([22]) containing in total 3666 GRBs recorded by BATSE between 23 April 1991 to 19 November 1999. It is the largest sample of such events and was compiled using the observatory's archival data. It consists of 1.024s resolution data covering the whole period of BATSE's lifetime. Almost half of these GRBs did not trigger the

detector, as they were too weak or occurred during dead-time. The catalogue consists of the best-fit time profiles in the four energy channels for each burst, and is currently the best available for the 1.024s resolution. We will be using the sum of the counts in the two brightest energy channels (#2 and #3) in the energy range 50-300 keV. Background still needs to be subtracted, however.

In order to aid the reader with the background reduction steps that follow, we suggest the inspection of Figure 4.2. In it, we show four examples of how the burst data is reduced, and how the variables used within our networks are defined. We will now describe in detail the procedures adopted.

We fit a simple quadratic model to the background using a least squares fit combined with sigma-clipping. The procedure fits a background model to the light curve, while excluding every point lying outside n standard deviations ($n\sigma$). The process is then iterated until no more points are rejected. This enables the model to be fitted only to the background, excluding the burst. A threshold of 2σ has been chosen by trial and error as this visually seems appropriate for most bursts.

Next, in order to eliminate flux and time information, we evaluate the Cumulative Distribution Function (CDF)[3] for each burst and normalise it to the burst fluence and T90[4] duration. However, this is not appropriate from the background-subtracted time profiles, since the noise present before and after the bursts will be a major disturbance, especially for GRBs with a low fluence. This disturbance will result in a "noisy" CDF which will not be always positive and not be monotonically increasing. In order to overcome this problem, we only consider the 3200 GRBs with the greatest fluence, and moreover employ a similar method to sigma-clipping, Chauvenet's criterion [36], to eliminate the background noise contamination. This method was first introduced as a way

[3]The CDF is the cumulative area of the burst light curve.

[4]The T90 duration of a burst is the time it takes for the middle 90% of photons to reach the detector.

of assessing if one piece of experimental data was spurious out of a set of observations. Here, we modify its application slightly. The method assumes that the background noise follows a Gaussian distribution[5] with a corresponding σ. All points lying clearly outside of this distribution must therefore be associated with the burst. In practice, the method proceeds iteratively. In each iteration, the standard deviation of all points is calculated and used to define an exclusion threshold. This threshold is chosen so that less than one point is expected beyond it. Mathematically,

$$\Delta y = \sigma\sqrt{2} erf^{-1}\left(1 - \frac{1}{2M}\right) \tag{4.1}$$

where M is the number of data points. As more and more points get excluded, σ decreases together with M, thus lowering the exclusion threshold every time. This threshold converges asymptotically to the point where only background-noise points are included. The excluded points will then be our background noise free burst.

Having obtained the pure burst light curve, we can produce a clean CDF. This in turn allows us to produce the T90 distribution of such events, as displayed in Figure 4.1, together with fluence and maximum peak intensity. Note that the famous T90 bimodality is not present in our sample, since the data used 1.024 seconds resolution, and the "short GRBs" peak at ~ 0.1 seconds.

Having evaluated the CDF for each burst and normalised it to burst fluence and T90, we define 9 variables in an attempt to encapsulate information regarding burst structure. This was done after taking inspiration from Horvath et al. [8], who used similar variables for GRB classification purposes with a supervised ANN.

We first define T_0, T_{10},T_{20} up to T_{90}, as the interval between the arrival

[5]This is not exactly correct since we know that the background noise is best described by Poisson statistics. However, for sufficiently high number of counts this is not important.

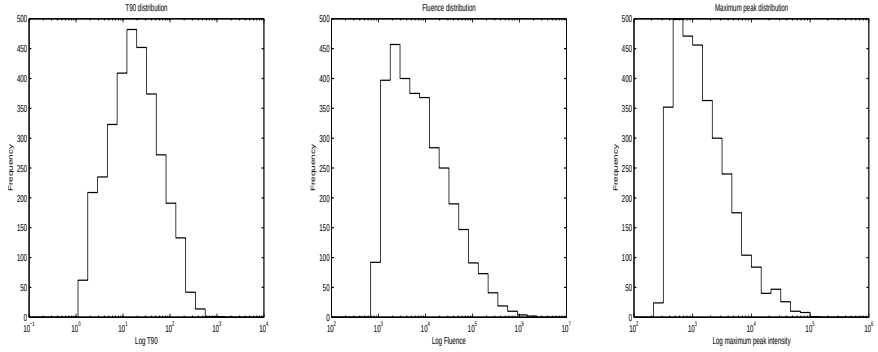

Figure 4.1: Distributions for the 3200 GRB population. From left to right: Log(T90), Log(Fluence) and Log(Maximum peak intensity).

of the middle 0%, 10%, 20% up to 90% of photons. We then define our nine variables as the difference between neighbouring intervals and normalise them to T_{90}. Expressed mathematically,

$$\Delta T_i = (T_{10i} - T_{10(i-1)})/T_{90} \tag{4.2}$$

This will give us a set of variables independent of burst time, since the sum of these will always be 1. The variables are also independent of burst fluence or hardness. This set of variables will then be then fed into the SOM for classification purposes, since now all bursts are expressed in 9 variables.

4.3 SOM Results

We have applied the SOM algorithm to the GRB dataset using as input the 9 variables obtained from Equation 4.2. The map consisted of 225 neurons (15 $\times$ 15), whilst the number of iterations was arbitrarily set to 2000. The neuron weights were initialised with random numbers with 0 mean and 1 standard deviation. This was done so that, on average, the range of the weight values

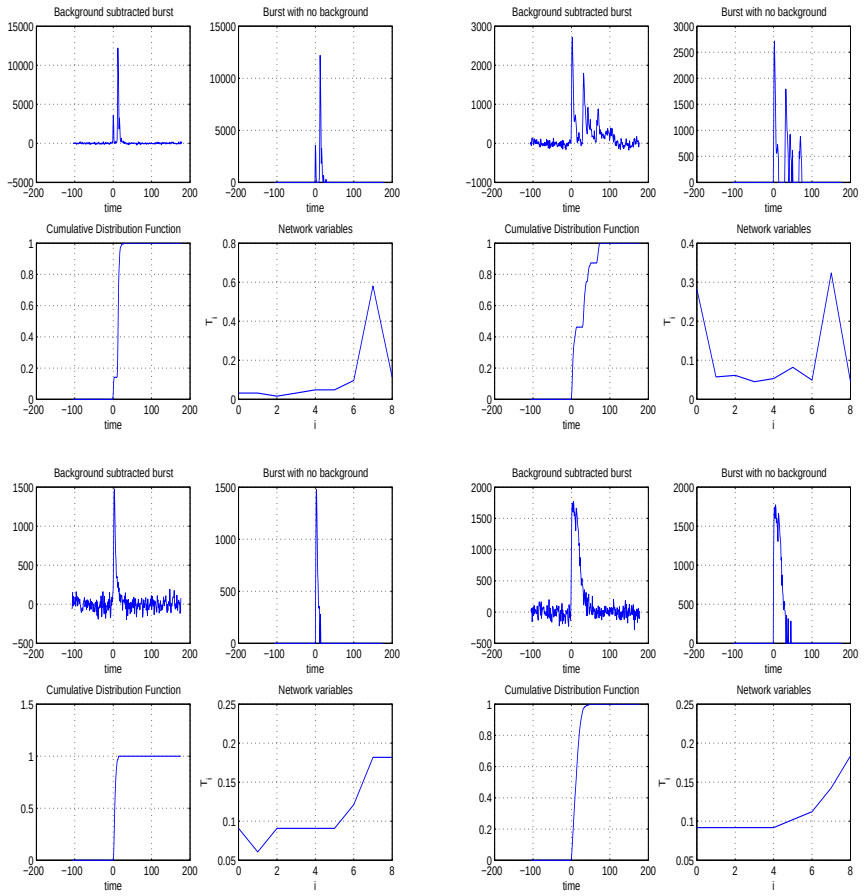

Figure 4.2: Figures to visually assist the explanation of the reduction methods. For each burst we display on the top left the burst after background fitting, subtraction and sigma clipping. The x-axis is time in seconds whilst the y-axis is photon counts in channels 2+3. The top right is the burst without background, as obtained by applying Cheuvenet's criterion. The corresponding burst CDF is presented in the bottom left, but not normalised to T90 for ease of comparison to the real burst. Finally the 9 variables used by the SOM and defined in Equation 4.2 are presented in the bottom right.

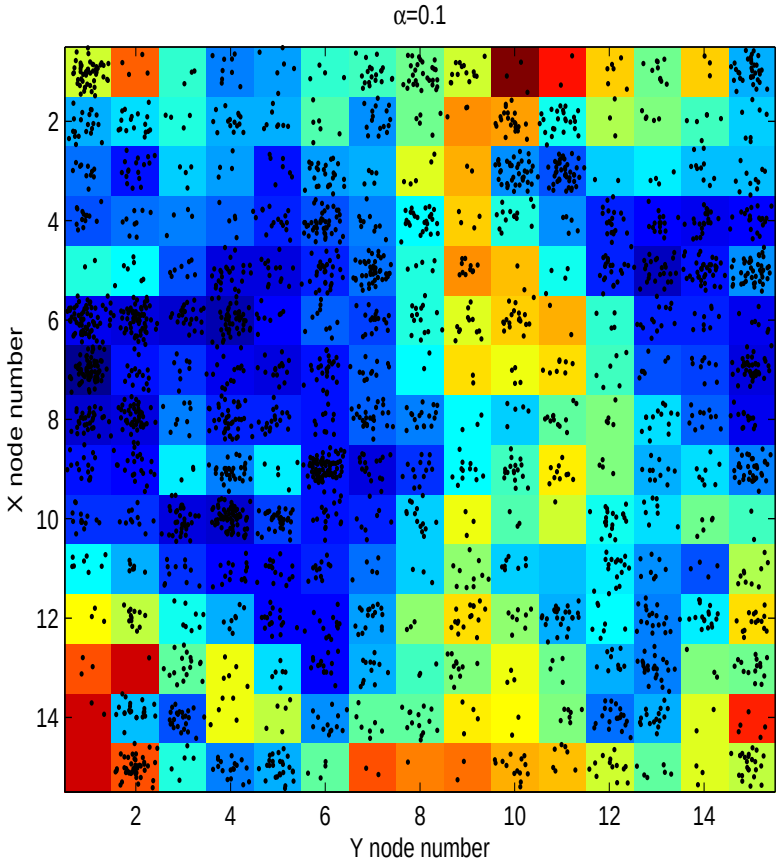

Figure 4.3: U-matrix of the SOM with $\alpha = 0.1$.

were relatively close to the input data[6], and the learning factor could be kept to a low constant. The neighbourhood kernel decreased monotonically from half the map size to a negligible value of $\frac{7.5}{2000} = 3.7 \times 10^{-3}$.

Two runs were carried out: one with $\alpha = 0.1$ throughout and one with $\alpha = 0.01$. The U-matrices for these are presented in Figures 4.3 and 4.5, respectively, together with their final weight vectors in Figures 4.4 and 4.6.

[6]We note that the weights had negative values contrary to any of the inputs. In terms of Euclidean space, however, this distance was on average the same for all inputs across the 9 variables.

Figure 4.4: Weight vectors for the SOM with $\alpha = 0.1$.

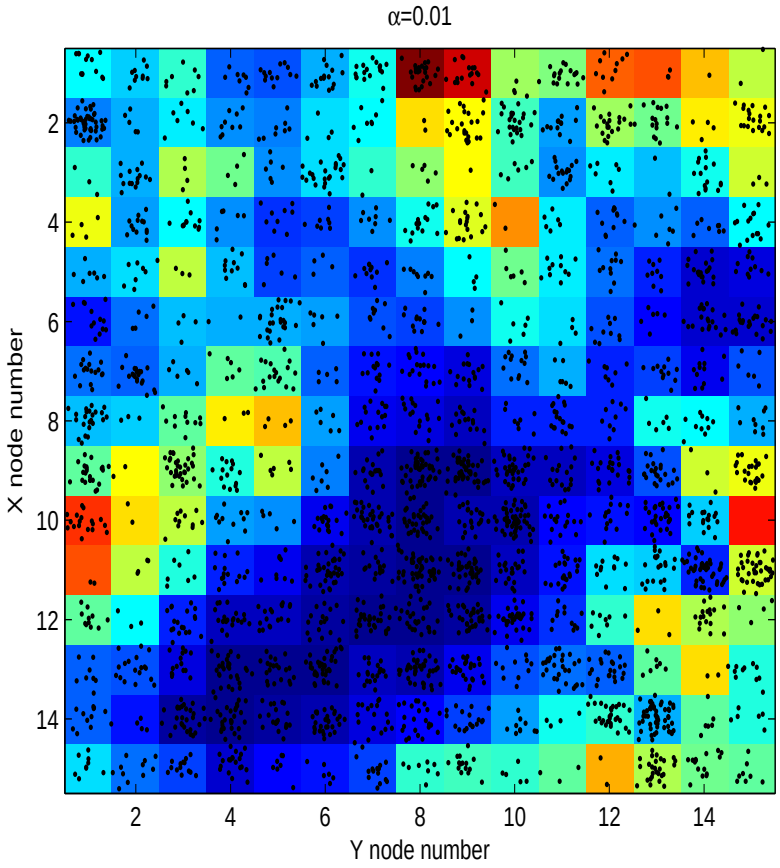

Figure 4.5: U-matrix for the SOM with $\alpha = 0.01$.

Figure 4.6: Weight vectors for the SOM with $\alpha = 0.01$.

The figures show that the U-matrix has identified only two classes, although the boundaries are not completely clear. Moreover, from both the U-matrix and weight vector maps, one can see that a large portion of the neuron weights are dedicated to a particular pattern within the dataset: the one for Single-Pulsed Bursts (SPBs) having a single burst structure in the prompt emission. This class consists of over half of the total population. Having established that SPBs are the only recurring pattern the SOM recognises, we can ask if this is a "real" distinction. To answer this question, we first need to split the SPB class and the remaining Multiple-Pulsed Bursts (MPBs). We can then compare their T90, fluence and maximum peak distributions, since these variables have not been used by the network. Thus any difference between the distributions of these quantities, would suggest that SPBs and MPBs are intrinsically different.

Because the boundaries are not so evident from the U-matrix, strictly speaking we would need to inspect nearly all neuron weights to classify SPB and MPBs. For simplicity, we therefore rerun a smaller SOM with 25 neurons (5×5) and assess those neurons, thus decreasing the amount of visual inspection. This last map was obtained after 500 iterations with a constant α value of 0.1. The U-matrix and weight vectors are displayed in Figure 4.7. Note, however, that for such a small map, the U-matrix has no real meaning, and the figure is included mainly to give a feeling of the hits within the neurons.

We have inspected all 25 neuron weights and conclude that 5 have been dedicated to SPB. These are[7] [4,5], [5,1], [2 3], [3 5] and [5,5], and can be visually checked in Figure 4.7. The total number of SPBs is 1696 whilst the MPBs add up to 1504.

[7] [row,column]

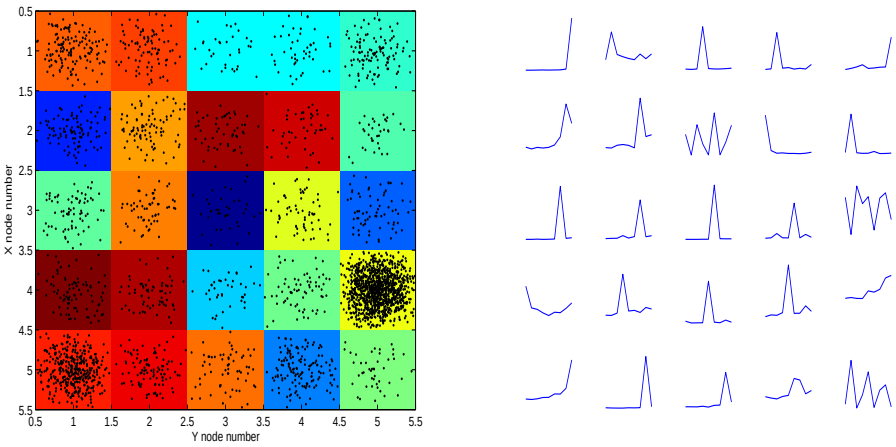

Figure 4.7: U-matrix with hits and weight vector for the 5 by 5 SOM.

4.3.1 Are SPBs and MPBs intrinsically different?

4.3.1.1 Fluence, T90 and Peak Intensity Distributions

Having located the neurons responsible for SPBs and MPBs, we can classify each burst and produce the T90, fluence and peak intensity distributions for the two classes. These can be then compared to each other and to the global distribution in Figure 4.1. These comparisons are presented in Figure 4.8. The T90 distributions seem different for SPBs and MPBs, with the former being characterised by shorter durations than the latter on average. Moreover, the fluence distributions seem to peak in different places, with the MPBs being less energetic on average. These differences suggest that our two network classes do indeed have different intrinsic properties. Only the maximum peak distributions of both classes seem to be quite similar.

However, before we can draw any solid conclusion, we have to consider the possibility of any biasing effects. The two main effects which might have caused this classification scheme to arise are: the possible biasing effect caused by the

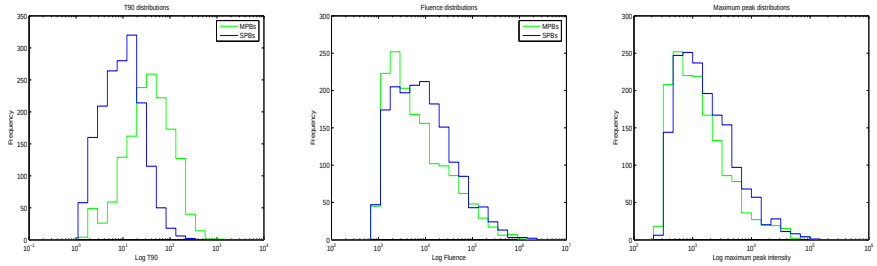

Figure 4.8: Same distributions as in Figure 4.1 but split between SPBs and MPBs.

signal-to-noise (S/N) ratio of the observations and the preconditioning of our data (i.e. the definition of our variables). We will now examine the question of possible biasing effects caused by the observations, and later assess the relevance of preconditioning.

4.3.1.2 S/N bias?

It is trivial to imagine that many SPBs could actually be MPBs for which we only can only see the strongest spike. In order to understand if this effect might cause our result, we have performed some simulations, in order to determine if the SPB/MPB split is simply determined by burst distance or detection threshold. The way this was carried out was by taking all MPBs and raising their threshold level (decrease their S/N ratio) gradually, until the burst turned into a SPB. In other words, we increased the background level of MPBs until they turned into SPBs. A similar analysis was performed by Schmidt [31] in order to determine the reliability of the T90 measure. Our analysis was performed by "cutting out" the bottom part of the burst, in steps of 5% of the maximum peak of the burst. As a reference for determining if an altered burst is classified as an MPB or a SPB, we used the 5×5 map presented in the previous section. Once we have turned the MPBs into SPBs, we can compare their distributions and inspect them for any changes to assess the biasing effect question. If the distributions

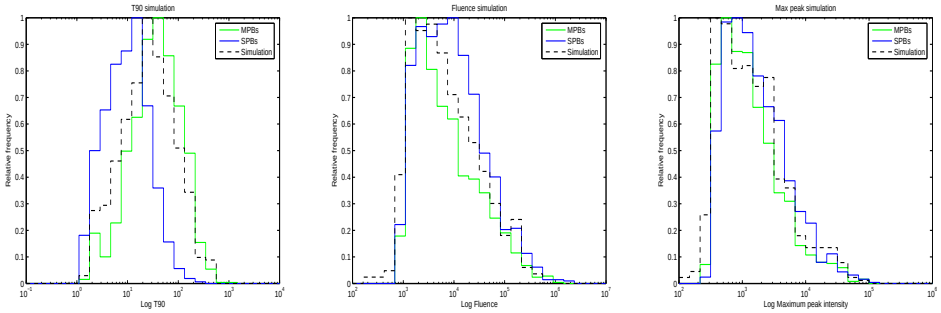

Figure 4.9: Threshold/Distance simulation for MPBs normalised to a common peak. The dashed black line shows the resulting distributions from the simulation described in Section 4.3.1.2. The true SPB and MPB distributions are there for reference.

of the simulation closely follows that of MPBs, then that would suggest biasing effects could not be the cause of the network distinction. On the other if the result of the simulation would closely follow that of SPBs, then it is possible that the SPB/MPB split could be caused by this biasing effect.

Out of 1504 MPBs, only 548 could be turned into SPBs with the method described (i.e. ˜2/3 of all bursts became undetectable before they could be turned into SPBs) . The remaining bursts could not be transformed into SPB by simply increasing the threshold level (equivalent to increasing distance). The distributions of the SPBs created from the MPBs in the simulation are presented in Figure 4.9. We note that most of the 548 MPBs which turned into SPBs did so with a tiny increase in the threshold level (5%-10% of the maximum peak). This would suggest that these are bursts with pre/post cursors or high "spikes" present in the light curve. All three simulation distributions seem to more closely follow the one of the MPBs, although there is a slight shift within the distributions (T90 and fluence) of the simulated MPBs towards the SPB distribution. Overall we think the results show that the SPB population cannot be caused by biasing effects due to distance from the observatory or threshold

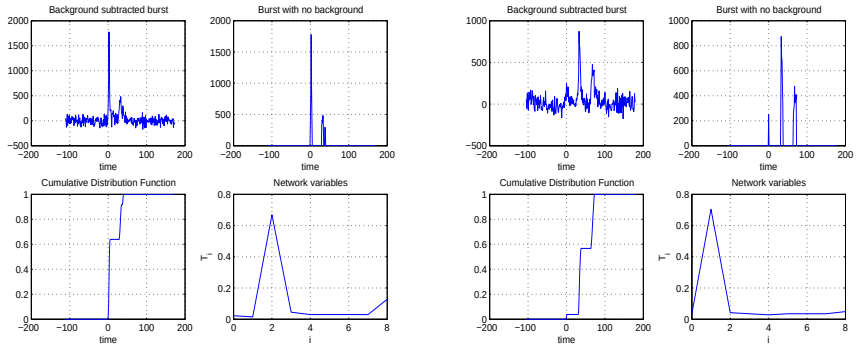

Figure 4.10: Two Multiple bursts with similar ratios between pulses. The smooth changes, as seen from the CDF, happen just before and after 0.6 for each. This change however is not smooth in variable space. One burst peaks at T20-T10 whilst the other at T10-T0.

level. The fact that only about 40% of the MPB population could be turned to SPBs is a particular strong hint towards the fact that these two groups do have indeed different intrinsical properties.

4.3.1.3 Preconditioning bias?

Another possible bias effect could be caused by the particular definition adopted for our network variables. We will now examine this in more detail.

Closer inspection of the neuron weights, together with the hits on the U-matrix, reveals a subtle problem with our network variables, which might be the reason why the MPBs did not cluster properly. Because of the nature of these variables, the change between similar patterns is not a smooth one. This can have a great impact on the ANN, as the Euclidean distance is its measure of "difference". The phenomenon is displayed in Figures 4.10 and 4.11. On the other hand two examples of the SPB group are presented in Figure 4.12.

The figures suggest that the definition of our network variables does indeed explain why the SOM would have found it easier to separate SPBs from MPBs, than to discover sub-classes within the MPB population, for example. How-

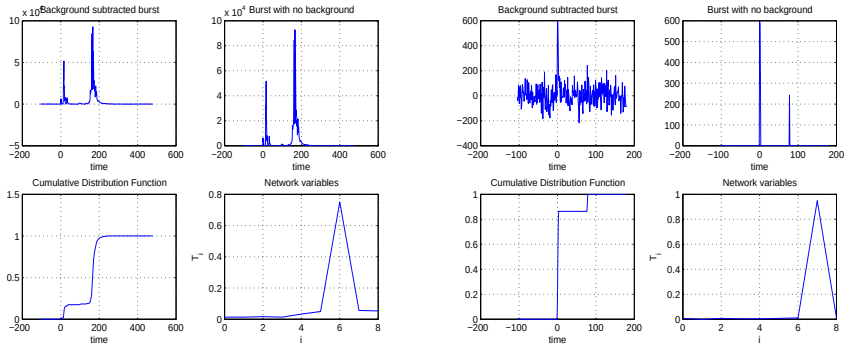

Figure 4.11: Two multiple bursts: One with precursor and one with a post-cursor. Our variables have been defined as to make such events as similar as possible, however the change between the two is still not smooth.

ever, it is important to understand that this does invalidate the evidence for intrinsic differences between SPBs and MPBs, as presented in Sections 4.3.1.1 and 4.3.1.2. Rather, the existence of what may be called "preconditioning bias" means that we cannot rule out that there are, in fact, sub-classes within the MPBs. There could even be a continuous distribution of burst complexity (with SPBs occupying one extreme end), but then our results imply that there must be a correlation between complexity and other burst parameters (such as T90 and fluence).

4.4 Conclusions and future work

This chapter has mainly been focused on mining the GRB dataset using solely light-curve shape dependant variables. This was done with the SOM, finding one main distinction within the set: SPBs and MPBs. The two sets also had the independent characteristic of having different duration and fluence properties.

We note that other authors have examined differences between SPBs and MPBs [6, 21]. This could definitely be investigated further by, for example,

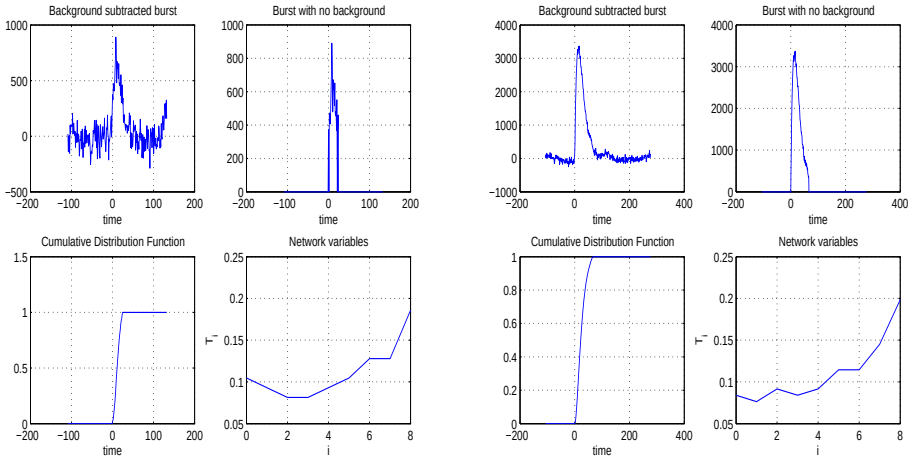

Figure 4.12: Two GRBs mapped onto random blue neurons in maps 4.3 and 4.5. Note the smooth variable rise characteristic of this class of bursts.

including additional variables in the SOM such as spectral lag. Moreover, based on this study, it is clear that concentrating on light curve shape solely requires the definition of some new, smoothly varying, variables. This would give the SOM a better chance in establishing boundaries between classes. One could even run two SOMs, one with the light curve shape variables and one with independent ones, such as T90 and fluence, and inspect the two to identify classes and correlations within the dataset.

Chapter 5

Summary and Conclusions

> *I have never met a man so ignorant that I couldn't learn something from him* - Galileo Galilei
>
> *I have said consistently that global warming something is a serious problem. There is a debate over whether it's manmade or naturally caused* - G.W. Bush
>
> *Two things are infinite: the universe and human stupidity; and I'm not sure about the universe* - Albert Einstein

This thesis has dealt with the application of Kohonen networks to astronomical data mining. In particular, we have shown how important the effect of preconditioning is when applying such algorithms. In chapter 3 we have seen how the metric used in SOMs is not appropriate for spectral shape classifications, but, with the help of the supervised LVQ, we still managed to compile the most complete BALQSO catalogue to date. In fact, when using SOMs for BAL recognition, the metric to adopt is very non-trivial to define, and until one has created an appropriate one, unsupervised methods will always be hard to interpret.

We have also presented an application of SOMs to the GRB dataset obtained by BATSE. Using a new definition of light curve shape driven variables, we were able to identify two major recurring patterns: SPBs and MPBs. However, we have also realised the non-continuous behaviour of these variables, suggesting a possible improvement for future work. We showed that our two "network classes" also have different independent properties (T90 and fluence), which would suggest intrinsic differences between them. The reason for this difference has not been determined, and will definitely be investigated further in future work.

We conclude this thesis by restating the importance of algorithms such as the ones used here for future astronomical data mining. The advent of enormous multi-wavelength surveys is already having a great impact on the astronomical community. The amount of data in astronomy triples every two years, and the dependence on robust algorithms follows accordingly. As an example, in ten years time, the square kilometer array (SKA) will be able to survey the entire radio sky on timescales of weeks, with much higher resolution and sensitivity than is possible today. Moreover, we can already see applications such as the Virtual Observatory taking shape. These will present astronomical data to the community in an extremely effective way and increase the potential for data-mining based research. This will also require classification tools (such as ANNs) that will have to deal with new and unseen patterns.

Bibliography

[1] Knigge C. *Modelling and observations of outflows from accretion disks*. PhD thesis, Oxford University, 1995.

[2] Anderson D. and McNeil G. Artificial neural network tecnology. 1992.

[3] Arav N. et al. Modeling the double-trough structure observed in broad absorption line qsos using radiative acceleration. *1994ApJ...434..479A*, 1994.

[4] Brett D.R. et al. The automated classification of astronomical lightcurves using kohonen self-organizing maps. *2004MNRAS.353...369B*, 2004.

[5] Brotherton M.S. et al. Spectropolarimetry of pks 0040-005 and the orientation of broad absorption line quasars. *Accepted for pubblication in MNRAS*, 2006.

[6] Hakkila J. et al. Gamma-ray burst class properties. *2000ApJ...538..165H*, 2000.

[7] Hall P.B. et al. Unusual broad absorption line quasars from the sloan digital sky survey. *2002ApJS..141..267H*, 2002.

[8] Horvath I. et al. A new definition of the intermediate group of gamma-ray bursts. *2006AA...447...23H*, 2006.

[9] Kohonen T. et al. Phonotopic maps - insightful representation of phonological features for speech recognition. *In Proceedings of 7ICPR, International Conference on Pattern Recognition, pages 182-185.*, 1984.

[10] Kouveliotou C. et al. Identification of two classes of gamma-ray bursts. *1993ApJ...413L.101K*, 1993.

[11] Miller A.S. et al. Star/galaxy classification using kohonen self-organizing maps. *1996MNRAS.279..293M*, 1996.

[12] Mukherjee S. et al. Three types of gamma-ray bursts. *1998ApJ...508..314M*, 1998.

[13] Naim A. et al. Galaxy morphology without classification: Self-organizing maps. *1997ApJS..111..357N*, 1997.

[14] North M. et al. Searching for the ghost of lyman alpha. *2006MNRAS.365.1057N*, 2006.

[15] Punsly B. et al. X-ray absorption in type ii quasars: Implications for the equatorial paradigm of broad absorption line quasars. *2006ApJ...647..886P*, 2006.

[16] Rajaniemi H.J. et al. Classifying gamma-ray bursts using self-organizing maps. *2002ApJ...566..202R*, 2002.

[17] Reichard T. et al. A Catalog of Broad Absorption Line Quasars from the Sloan Digital Sky Survey Early Data Release. *2003AJ....125.1711R*, 2003.

[18] Reichard T. et al. Broad absorption line quasars in the sdss. *2004ASPC..311..219R*, 2004.

[19] Reichard T.A. et al. Continuum and emission-line properties of broad absorption line quasars. *2003AJ....126.2594R*, 2003.

[20] Schneider D.P. et al. The sloan digital sky survey quasar catalogue iii. third data release. *2005AJ....130..367S*, 2005.

[21] Stern B. et al. A complexity-brightness correlation in gamma-ray bursts. *1999ApJ...510..312S*, 1999.

[22] Stern B.E. et al. An off-line scan of the batse daily records and a large uniform sample of gamma-ray bursts. *2001ApJ...563...80S*, 2001.

[23] Trump J.R. et al. A catalog of broad absorption line quasars from the sloan digital sky survey third data release. *2006ApJS..165....1T*, 2006.

[24] Ultsch A. et al. Kohonen' self organising feature maps for exploratory data analysis. *In Proceedings of the International Neural Network Conference (INNC'90), pages 305–308, Dordrecht, Netherlands, 1990*, 1990.

[25] Weymann R.J. et al. Comparisons of the emission-line and continuum properties of broad absorption line and normal quasi-stellar objects. *1991ApJ...373...23W*, 1991.

[26] York D.G. et al. The sloan digital sky survey: Technical summary. *2000AJ....120.1579Y*, 2000.

[27] http://archives.cnn.com/2000/fyi/news/12/01/brain.function/story.neurons.jpg.

[28] http://dms.irb.hr/tutorial/images/ann.jpg.

[29] http://www.ij-healthgeographics.com/content/figures/1476 072X-3-12-7.jpg.

[30] Elvis M. A structure for quasars. *2001AJ....122..549V*, 2001.

[31] Schmidt M. Are durations of weak gamma-ray bursts reliable? *2005NCimC..28..347S*, 2005.

[32] P. Meszaros. Gamma-ray bursts. *2006RPPh...69.2259M*, 2006.

[33] R. Mushotzky. *How are AGN Found?*, pages 53–+. ASSL Vol. 308: Supermassive Black Holes in the Distant Universe, August 2004.

[34] Djorgovski S.G. Virtual astronomy, information technology and the new scientific methodology. *to appear in IEEE Proc. of CAMP05, Coputer Architectures for Machine Perception*, 2005.

[35] Kohonen T. *Self Organizing Maps*. Springer, 1963.

[36] J. R. Taylor. *An introduction to error analysis. The study of uncertainties in physical measurements*. A Series of Books in Physics, Oxford: University Press, and Mill Valley: University Science Books, 1982, 1982.

[37] Pei Y.C. Interstrellar dust from the milky way to the magellanic clouds. *1992ApJ...395..130P*, 1992.