Novel Methods of Supernova Classification and Type Probability Estimation

James Newling

Thesis presented for the degree of

Master of Science

Department of Mathematics and Applied Mathematics

University of Cape Town

August 2011

Acknowledgements

I have worked with some amazing people over the last year and a half, and I'd like to thank them for their help. The A.I.M.S. cosmology group members past and present whose company I have enjoyed: Bryony, Carolina, Marco, Mat, Melvin, Michelle, Patrice, Petja and Yabebal. Other people who have guided me in my first escapade into cosmology: David Parkinson, Martin Kunz and Renée Hlozek. My U.C.T. officemates who understand in real-time how tough thesis writing can be: Anuj, Dawie, Dino and Terence. Especially I thank my supervisor Bruce Bassett, for showing me how much fun research can be. I am also grateful to the NRF and SKA for providing me with generous funding.

Abstract

Future photometric surveys will provide vastly more supernovae than have presently been observed, the majority of which will not be spectroscopically typed. Key to extracting information from these future datasets will be the efficient use of lightcurves. In the first part of this thesis we introduce two methods for distinguishing type Ia supernovae from their contaminating counterparts, kernel density estimation and boosting. We show that these methods perform well at classifying $20,000$ simulated Dark Energy Survey lightcurves, provided that training is done on a sample which is representative of the unclassified supernovae. However, training on the types of spectroscopic samples currently produced by supernova surveys leads to poor performance, and we recommend that special attention be given to the creation of more representative training samples.

In the second half of this thesis we shift focus from classification to the related problem of type probability estimation, and ask how best to use type probabilities. In their 2007 paper, the inventors of BEAMS (Bayesian Estimation Applied to Multiple Species) showed how to use a contaminated dataset to perform unbiased parameter estimation, using all of the data points in conjuction with their probabilities of being of particular types. We describe an implicit assumption made by the authors of the first BEAMS paper, relating to the independence of data and type probabilities, and present the necessary modifications to deal with correlated data. We also perform a simple 1-D simulation to compare pre- and post- modification BEAMS, and show how the modification provides a 50% reduction in parameter estimation variance. We then perform three tests to quantify the importance of the type probabilities, one of which illustrates the effect of biasing the probabilities in various ways. Finally, a general presentation of the selection bias problem is given, and discussed in the context of supernova cosmology.

Table of Contents

Chapter 1

Introduction

Introducing Supernovae

Supernovae (SNe) have captured the minds of mortal men for millenia. Of all the wonders in the night sky, SNe best illustrate that life and death are not unique to our planet. The brightest apparent magnitude event in recorded history, the SN of 1006 AD so captured the minds of prehistoric native North Americans that they recorded the event by carving its likeness into a rock. That same SN, which was visible during daytime to the naked eye with an apparent magnitude of -7.5, was also recorded by star watchers in Europe, Japan, China and the Middle East [1].

The observations in Europe of Milky Way SNe in 1572 and 1602, the most recent to be observable to the naked eye, had a noticeable impact on European science. Johannes Kepler's observations of the 1602 SN may well have paved the way for the rejection of the idea that the universe is unchanging beyond the moon and planets [2].

In more recent times, SNe featured in the famous Curtis-Shapley debate of 1920 as to the scale of the universe and the nature of nebulae. The faint novae that had been observed in nebulae suggested to Harlow Shapley that the systems containing them must be nearby [3]. Heber Curtis countered that the novae being observed were of previously unheard of magnitudes in distant galaxies. Curtis was proved correct when Edwin Hubble discovered

extremely faint Cepheid variable stars in the nebulae using the 100 inch Hooker Telescope, at the time the largest telescope in the world.

The word supernova was first used by Swiss astrophysicist Fritz Zwicky, who along with Walter Baade carried out the first systematic SN survey in the 1930s. Zwicky hypothesized that a SN was the release of energy from a normal star collapsing in on itself to form a neutron star [4]. This was not the correct explanation for the SNe that Zwicky and Baade studied, but was in fact the correct explanation for type II SNe, which at that time had not been observed. Zwicky and Baade had been observing what we today call type Ia SNe.

SNe are classified into two groups, depending on whether their spectra show hydrogen lines: type I SNe do not show hydrogen lines and type II SNe do. Type II SNe all result from the core collapse of a supergiant star. Type I SNe can be subclassified according to the presence of Silicon absorption lines in their spectrum. Type Ia SNe are those type I SNe which contain Silicon lines, while type Ib and Ic lack Silicon lines. Type Ib and Ic SNe are thought to result from similar processes as those leading to type II SNe, the difference being that the supergiant stars which cause Ib and Ic SNe have previously had their hydrogen lost in intense stellar winds. The spectra of the different types of SNe are illustrated in Figure 1.1. Two good sources of information about SNe are the references [5] and [6].

The currently accepted belief is that type Ia SNe result when a white dwarf gradually increases in mass until it reaches the Chandrasekhar limit of about 1.4 times the mass of the Sun [8]. At this mass the star becomes unstable and carbon and oxygen fuse to form nickel, releasing enormous amounts of energy. The additional matter required to reach the critical mass is provided by a companion star, which loses its mass through a process called Roche overflow [9]. It is the property that type Ia SNe all explode with approximately the same energy that makes them more than just beautiful events but tools of cosmology: Ia SNe are standard candles.

The idea of a standard candle is that, given an object of known absolute magnitude such as a 100 watt tungsten incandescent light bulb, it is possible to calculate its distance from only its observed apparent magnitude. The

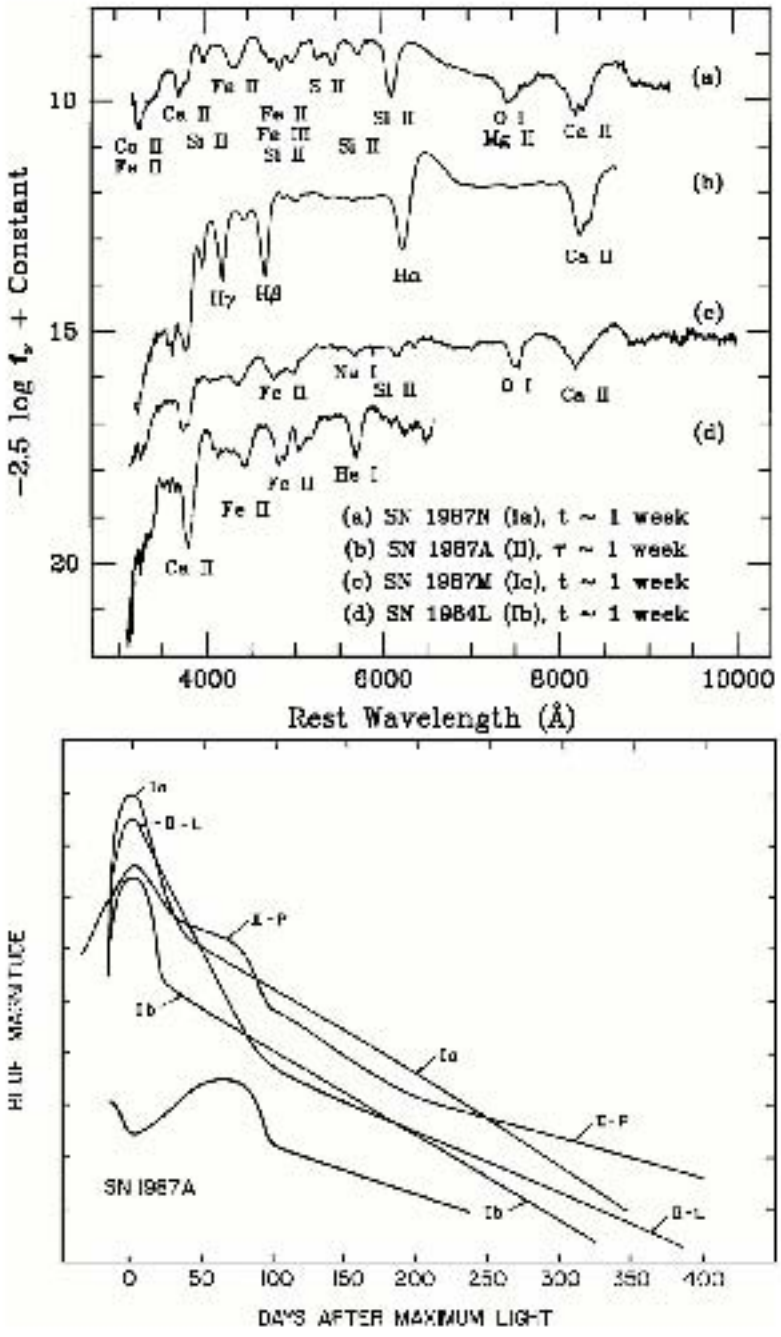

Figure 1.1: (Above) Peak magnitude spectra of type Ia, II, Ib and Ic SNe. Type Ia show pronounced Silicon absorption at 6150 Å, and lack the distinct hydrogen lines of type II SNe. While the top figure above is of frequencies of light at one moment, the bottom figure is of light through a blue filter, recorded over several months. The labels Ia, Ib, IIL, IIB refer to subclasses of SNe. Source: Optical Spectra of Supernovae [7].

apparent magnitude falls as the inverse of the square of the distance: double the traveling time of a beam of light and its apparent magnitude is quartered. Without knowing the absolute magnitude of an object, it is not possible to determine whether it is a faint object nearby or a bright object at a great distance. That type Ia SNe all explode with approximately the same density and mass means that they all release approximately the same amount of energy, and emit the same amount of light. Strong evidence supporting this is that, over the past 40 years, several galaxies have been observed to host two Ia SNe, and in all of these galaxies the two SNe have had almost the same maximum apparent brightness [5]. Other evidence is provided by the use of other confirmed standard candles, in particular Cepheid variable stars [5].

Type Ia SNe are generally not referred to as standard candles, but rather as standardisable candles. While the variance in peak magnitude of type Ia SNe is lower than that of any other type, there is still a significant variance (35%) in absolute magnitude. However, there is a way to reduce this variance which we will discuss later, and thus type Ia are standardisable. Cepheid variable stars are technically also only standardisable candles, as it is only by using a particular period-luminosity relationship that the variance in their magnitudes can be reduced to less than 15%. Type Ia SNe are the brightest standardisable candles available, and have been observed at a time when the universe was less than a quarter of its current age [10]. For this reason, type Ia SNe are greatly important as they allow us to measure the size of our observable universe. But not only do SNe help us measure the size, but also the change in size of the universe through time. To better understand this it is necessary to briefly discuss the relevant ideas from general relativity.

General Relativity and Models of the Universe

Einstein's field equations of general relativity tell us how the distribution of energy and momentum determine the geometry of space and time. The famous field equations of 1916,

$$\mathbf{G} = \frac{-8\pi G}{c^4}\mathbf{T}, \tag{1.1}$$

were spectacularly successful in accurately predicting planetary motion and the deflection of light by the Sun, and were readily adopted as the best description of physics on the large scale. The tensor quantity on the left of (1.1), $\mathbf{G}$, describes the geometry of spacetime, while the tensor quantity on the right, $\mathbf{T}$, describes the distribution of energy in the universe. c is the speed of light and is constant.

When Einstein tried to use these equations to describe the space-time geometry of the entire universe, he realized that in the form (1.1) they would not allow a static universe. Believing that the universe was static, he chose to include an additional term in (1.1), which could be done without violating any of the basic assumptions of his work. This additional term ($\Lambda\mathbf{g}$) allowed for a static universe,

$$\mathbf{G} + \Lambda\mathbf{g} = \frac{-8\pi G}{c^4}\mathbf{T}. \tag{1.2}$$

The tensor $\mathbf{g}$ is the metric tensor, and the term Λ came to be known as the cosmological constant. The effect of a positive cosmological constant is a long-range force which works to oppose gravity and thus prevents a stationary universe from collapsing in on itself. Twelve years after Einstein's 1917 publication which included the cosmological constant, it was discovered that the universe is expanding [11], at which point Einstein famously claimed that the inclusion of the cosmological constant was the 'greatest blunder' of his life.

A central assumption which is often made in modeling the whole universe is what is referred to as the Cosmological Principle. Simply stated, it says that on sufficiently large scales, the universe is homogeneous (the same everywhere) and isotropic (the same in all directions). The universe is also usually modeled as being uniformly filled with a gas or liquid, where one can think of galaxies as being the atoms composing this cosmic fluid. With this assumption, one only needs to specify the average density and pressure at a given time to determine the state of the gas. The pressure and density of the

cosmic fluid of the expanding or contracting universe at time t are usually denoted by $p(t)$ and $\rho(t)$ respectively. The model is usually simplified further by assuming that the pressure is negligible. These are all assumptions which Einstein made in constructing the static model.

Another important feature of Einstein's model of the Universe, not captured in (1.2), is that it has positive curvature, which implies that the universe is spatially finite in extent, and that if one were to travel in a straight line for a sufficiently large distance, one would return to one's starting point. This is analogous to what happens when one travels in a straight line on the surface of a sphere, but in one dimension higher. The curvature (k) is an important feature of a model of the universe, and when assumed to spatially constant can be positive: closed universe with $k = 1$, zero: flat universe with $k = 0$, or negative: open universe with $k = -1$.

The Einstein model is in fact a special case of a solution to the Friedmann equation,

$$\dot{R}^2 = \frac{8\pi G R^2}{3}\left(\rho + \frac{\Lambda\, c^2}{8\pi G}\right) - kc^2, \tag{1.3}$$

which is solved in conjunction with the space-time metric,

$$(\mathrm{ds})^2 = \frac{R\,(t)^2}{\left(1 + \frac{kr^2}{4}\right)}\left[(\mathrm{dx}^2) + (\mathrm{dy}^2) + (\mathrm{dz}^2)\right] - c^2\mathrm{dt}^2 \tag{1.4}$$

to determine the behavior of the universe. An important element of the equation and metric is the scale factor $R(t)$, often referred to as $a(t)$. $R(t_1)/R(t_0)$ is the factor by which the universe has expanded between times t_0 and t_1. Note that $R(t)$ can be interpreted as the radius of a closed universe ($k = 1$), but this interpretation is not valid in the cases of flat and open space, as the universe is then infinite in extent. Appearing also in the Friedmann equation are Λ and $\rho(t)$. In the static Einstein model it is a requirement that Λ takes a value precisely determined by ρ,

$$\Lambda_E \stackrel{\text{def}}{=} \frac{4\pi G\rho}{c^2}. \tag{1.5}$$

This relationship, along with $k = 1$, is the unique combination for which

the solution to the Friedman equation is $R(t) = 1$ for all t, that is the static universe. Shortly after Einstein published in 1917, the Dutch astronomer Willem de Sitter proposed a quite different model for the universe. As is usually done, de Sitter assumed the Cosmological Principle to be true. He differed in two ways in solving the field equations from Einstein, firstly in assuming a flat universe ($k = 0$), and secondly in assuming that in addition to the pressure being negligible, the matter density was negligible. What little matter there was in the de Sitter universe would not affect cosmological expansion or contraction, this would be determined entirely by the cosmological constant. In the de Sitter model, with a positive cosmological constant, the universe would expand forever. This was the key feature of the de Sitter model: it was the the first to describe an expanding universe.

An important quantity in describing an expanding or contracting universe is the Hubble parameter, which is the rate of expansion per unit length:

$$H(t) = \frac{\dot{R}(t)}{R(t)}$$

In an expanding de Sitter universe, the Hubble parameter is constant,

$$H(t) = \sqrt{\Lambda},$$

which describes a situation where the distance between two points which are initially co-moving grows exponentially.

So far we have described two of the possible solutions of the Friedmann equations. We will now give a brief taxonomy of the other spatially isotropic and homogeneous possible solutions, which together are called Friedmann-Robertson-Lemaître-Walker models. The case where $\Lambda < 0$ is not particularly interesting. In such a case, independent of the curvature of the universe and the matter density, the universe starts with a 'big bang' ($R(0) = 0$) and ends in a 'big crunch' after some finite time, with $\dot{H} < 0$ at all times [5].

The models with $\Lambda = 0$ were until recently believed to be the most likely models of our universe. Amongst these models with a cosmological constant of zero, all models necessarily start with a big bang, irrespective of curvature. Only the closed universe ($k = 1$) undergoes a big crunch as well as a big bang,

and this is known as the closed model. In the flat and open cases, the universe continues expanding forever, but differ in the rate at which they do so. In the open universe, the distance between objects grows linearly in the limit ($R(t) \propto t$), while in the flat case the rate of expansion is slower ($R(t) \propto t^{\frac{2}{3}}$). This model, with $k = 0$ and $\Lambda = 0$ is often referred to as the critical model, and is used as a reference for other models as we will describe. Note that with the critical model, the Hubble constant is related to the density by,

$$\rho_{\text{crit}}(t) = \frac{3H^2(t)}{8\pi G}. \tag{1.6}$$

Let us now consider the case where $\Lambda > 0$. We have already met two of these, the Einstein model ($k = 1$) and the de Sitter model ($k = 0$). The Einstein model is in fact a very special case of a closed universe solution with $\Lambda > 0$, where Λ and ρ follow the strict relationship (1.5). The other possibilities for the closed universe with positive Λ are $0 < \Lambda < \Lambda_E$ and $\Lambda_E < \Lambda$. In the case of a lower but non-zero cosmological constant ($0 < \Lambda < \Lambda_E$), for the closed universe there are two possibilities: a big bang - big crunch model, and a model where the universe is infinitely old. In this infinitely old universe model the universe undergoes an infinitely long contraction, before reaching some critical compression and expanding thereafter forever. If our universe really started with a big bang as we believe, this model is ruled out.

The final possibility for $k = 1$ is $\Lambda > \Lambda_E$, known as the Lemaître model. In this model of a finite universe, the universe starts with a big bang and then expands indefinitely. An interesting feature of this model which the de Sitter model does not contain, is a period of reduced expansion, during which matter almost wins out in reversing expansion, but fails to do so.

It should be mentioned that in addition to the Einstein model, there are two other possible solutions when $\Lambda = \Lambda_E$. These correspond to a slightly smaller and slightly larger universe, neither being stationary as is the Einstein model. In the case of a smaller universe, it is necessary that it starts with a big bang, and then asymptotically approaches the Einstein universe in size. If the universe is slightly larger than is the Einstein, it grows exponentially.

We have now discussed all of the closed universe solutions. The two

remaining cases to discuss are with $\Lambda > 0$, and $k = 0 / -1$. These two models are qualitatively the same, in that they both describe a universe which starts with a big bang and continues to expand forever. When there is a non-negligible matter density, they both exhibit an epoch of reduced expansion (like the Lemaître model), during which the gravitational force of matter is dominant. However, after a certain amount of expansion, matter is too sparsely distributed to maintain the repulsive effect of the cosmological constant, and thereafter the universe accelerates forever. The de Sitter model is the limiting case where matter density is zero for $\Lambda > 0, k = 0$, for which the epoch of deceleration does not happen. The case of $k = 0$ and a non-negligible matter component, commonly called the accelerating model, is at present the most likely model of our universe, as we will discuss.

Discovery that the Universe is Expanding and Accelerating

In 1907 there was no evidence that the universe was anything but static. However in 1929, a decade after Einstein laid forth his model of the static universe, strong evidence that we live in a non-static universe was presented. During a public talk, Hubble presented a figure of 24 spiral galaxy distances vs redshifts, showing a significant linear relationship. Who deserves credit for the discovery that the universe is expanding is still currently being debated, but we do know that Hubble calculated the distances to the galaxies by observing Cepheid variables in the galaxies, and the redshift values which Hubble used were known thanks to the meticulous work of Vesto Slipher [12].

The linear relationship Hubble observed agreed well with the de Sitter model of an expanding universe. The light leaving more distant galaxies would take longer to reach Earth, and so in an expanding universe would have more time to be stretched. In the de Sitter model the amount by which light is redshifted is,

$$z = H_0 t = \frac{H_0}{c} d. \tag{1.7}$$

where d is the distance that the light has travelled to reach the observer and H_0 is the Hubble constant, which for the de Sitter model is indeed constant through time. Hubble's presentation of data which obeyed relationship (1.7) was strong evidence against the static model. With the demise of the static model, the two models competing for preference were the de Sitter model and the steady-state model, which we will not discuss here but to say that it lost credibility with the discovery of the Cosmic Microwave Background (CMB) in 1964. Although Hubble was out by a factor of ten in his estimate of H_0, his discovery marks the starting point in the quest to accurately measure the expansion of the universe.

Equation 1.7 is a precise relationship in an exponentially growing de Sitter universe, but it is only a first order approximation in other Friedmann-Lemaître-Robertson-Walker solutions. If one is going to include acceleration in the equation describing expansion or contraction, (1.7) should be replaced by

$$d = \frac{cz}{H_0}\left(1 + \frac{1}{2}(1 - q_0)z\right). \tag{1.8}$$

where H_0 and q_0 are the Hubble and deceleration parameters at the present time. The choice of a deceleration parameter as opposed to an acceleration parameter in (1.8) is explained by the fact that until recently it seemed more plausible that the universe were decelerating than accelerating. The determination of these parameters was for many years the primary objective of observational cosmologists. The deceleration parameter, like the Hubble parameter, is not a constant and is related to the scale factor R by,

$$q(t) = \frac{-R(t)}{\dot{R}(t)^2}\ddot{R}(t).$$

As we have seen, an important parameter in determining the expansion of

the universe is the average cosmic matter density, ρ. It is difficult to measure the current value of ρ, that is ρ_0, as it is believed that approximately five sixths of the matter contributing to the total matter in the universe is dark matter, which can only be detected by its gravitational effects on ordinary matter.

From (1.6), for a given value of the Hubble parameter, we have a corresponding critical density for the critical model. As already mentioned, it is common to use the critical density as a reference for other models, and a much quoted quantity in this regard is the density parameter for matter,

$$\Omega_m(t) \stackrel{\text{def}}{=} \frac{\rho(t)}{\rho_{\text{crit}}(t)} \tag{1.9}$$

If one looks at the Friedmann equation (1.3), one notices that ρ and $\Lambda c^2/8\pi G$ enter in the same way. This suggests that $\Lambda c^2/8\pi G$ is some type of energy density and can be represented in a way analogous to ρ in (1.9). This is indeed what is done, and so another much quoted quantity is the density parameter for the cosmological constant,

$$\Omega_\Lambda(t) \stackrel{\text{def}}{=} \frac{\rho_\Lambda}{\rho_{\text{crit}}(t)} \qquad \text{where } \rho_\Lambda = \Lambda c^2/8\pi G. \tag{1.10}$$

By rearranging the Friedmann equation, it is not difficult to show that a flat universe ($k = 0$) corresponds to the situation where $\Omega_m + \Omega_\Lambda = 1$. The evidence from the power spectrum of the CMB strongly suggests that $k = 0$ at present [13]. While the CMB best constrains $\Omega_m + \Omega_\Lambda$, it is with SNe that the best constraints on $\Omega_m - \Omega_\Lambda$ are obtained. Together, these two constraints can be used to determine what type of universe we live in. It is to the topic of constraints from SNe that we now turn.

Constraining Cosmology with Supernovae

Even before it was discovered in the 1990s that the universe is expanding at an accelerating rate, there was a big problem with unaccelerating models containing $\Lambda = 0$. The most popular model before 1990, the critical model,

predicted that the universe was 9 billion years old, while astrophysicists believed that the oldest observed objects in the universe, globular cluster stars, were at least 11 billion years old [14]. The discovery that the universe was accelerating pushed the predicted age of the universe back, and it currently stands at 13.75 ± 0.11 billion years, which is consistent with the age of globular clusters [15].

The method that Hubble used to estimate the Hubble constant was not powerful enough to detect a signifcant non-zero deceleration parameter. The problem was that the standard candles he was using, Cepheid variables, were not bright enough to be observed beyond a redshift of 0.1 which would be necessary to observe acceleration or deceleration of any significant amount. With the building of the Mount Palomar telescope in the late 1940s, it was hoped that whole galaxies could be used as standard candles at redshifts greater than 0.1. The hope was that, if galaxies had some intrinsic magnitude limit, then the brightest galaxies in clusters would be close to that limit, and thus be standard candles. The result that $q_0 = 3.7 \pm 0.8$ was published in 1956 under the assumption that this reasoning was sound, suggesting that our universe was one with $\Lambda = 0$. By the late 1970s however, it was shown that attempts to use galaxies as standard candles were erroneous [5], as galaxies in the early universe (high redshift) were intrinsically brighter than those of our epoch. High redshift galaxies were therefore more distant than claimed, and therefore their light had taken longer to be redshifted by the amount observed, in agreement with a lower value for q_0. With the realization that galaxies were not standard candles, cosmologists turned to SNe in the hope that they would be able to elucidate the expansion history of the universe.

Cosmological distances are usually given as distance moduli, where the distance modulus (μ) is related to the apparent magnitude (m) and the absolute magniude (M) by,

$$\mu = m - M$$

and it is related to the distance (d) in parsecs by

$$\mu = -5 + 5 \log_{10}(d),$$

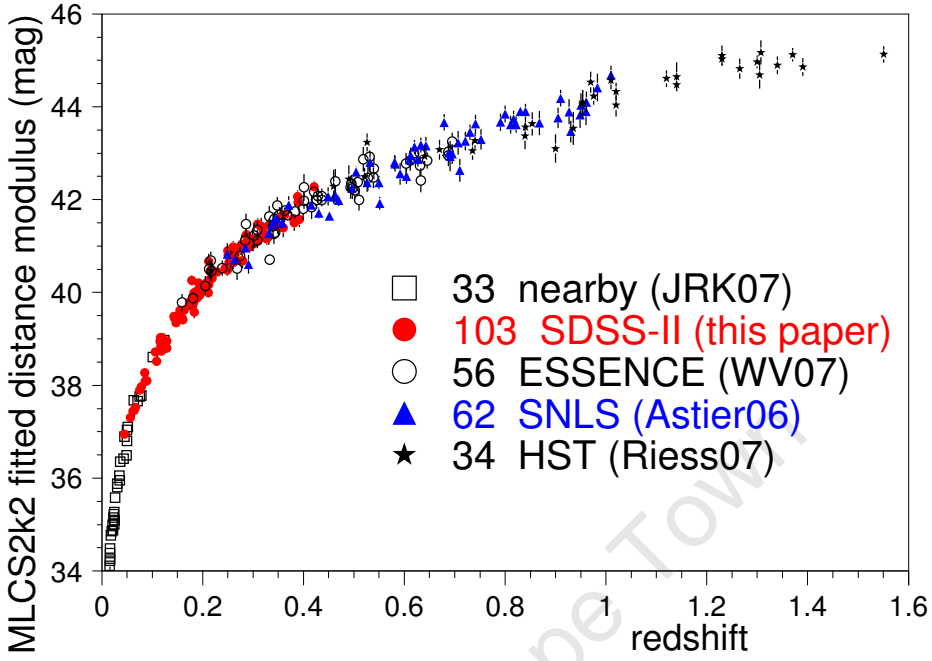

Figure 1.2: Hubble diagram of 288 SNe from the indicated surveys. Taken from SDSS-II [18].

where 1 parsec is $\sim 3 \times 10^{16}$ m. A plot of the distance moduli of several objects against their redshifts is called a Hubble diagram, as presented in Figure 1.2. In 1968, Charles Kowal made the first SN Hubble diagram in an attempt to estimate H_0, using 19 type Ia SNe [16]. There were several sources of error in that first Hubble diagram and the estimate of H_0 was poor, but Kowal was optimistic: "It may even be possible to determine the second-order term in the redshift-magnitude relation when light curves become available for very distant objects." [17]. It would be thirty years before his prediction was realized, as there were several complications with observing SNe at the necessarily high redshifts.

One complication, as already mentioned, is that type Ia SNe are not standard candles, but only standardisable candles. There is a 35% variation in the peak absolute magnitude of uncorrected type Ia SNe, but when the correlation between peak magnitude and time of decay is taken into account, this reduces to about 15%. The correction for this correlation is known as

the Phillips correction, as illustrated in Figure 1.3 [19]. With its discovery it was estimated that by observing and correcting about 30 type Ia SNe with redshifts in the range between 0.5 and 1.0, moderate cosmic acceleration or deceleration should be discovered [20]. Two groups in the mid-1990s did this, the High-z Supernovae Search Team [21] and the Supernova Cosmology Project [22], and they came to the same conclusions. By 1998 the results clearly showed that SNeIa are too faint than would be expected if the universe had been expanding at its current rate since the big bang. Therefore, the light has been traveling for longer than predicted by a non-accelerating universe, and so in the past the universe was expanding at a slower rate. It was an unexpected result at the time, but the data seemed decisive: the universe is accelerating.

It took thirty years for the technology and understanding of SNe to progress to the level where such a discovery could be made, but results since then have only confirmed the finding [24, 25, 26, 27, 28, 29, 15]. That said, there is still a lingering concern amongst some that systematic uncertainties undermine the validity of results, as will be discussed in the next section.

In the view of many, we are now entering an era of precision cosmology [5]. In the coming decades, it is hoped that the true values of the cosmological parameters will be measured more and more accurately. It should be noted that at present there are several competing models as to what Ω_Λ really represents in our universe. The mysterious energy which is causing cosmic acceleration has come to be called *dark energy*. The competing models of dark energy, which we will not discuss here, each manifest themselves slightly differently through the dark energy equation of state. Roughly speaking, this equation relates the density of dark energy (ρ_{DE}) to the pressure of dark energy (p_{DE}). The quantity defining this ratio is w_{DE},

$$w_{DE} = \frac{p_{DE}}{\rho_{DE}.}$$

If the universe is correctly described by Einstein's field equations with a cosmological constant, that is a ΛCDM universe, then $w_{DE} = -1$. Other models such as the quintessence model predict different values w_{DE}. Currently there

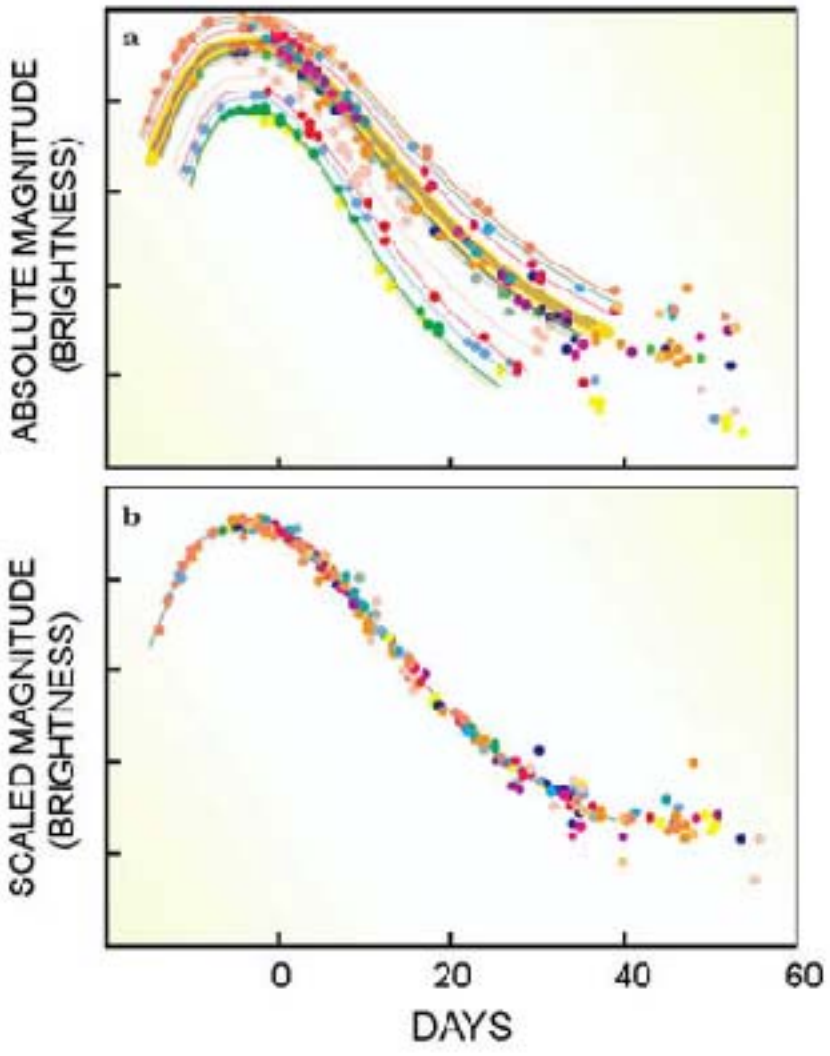

Figure 1.3: Illustration of the Phillip's correction. Intrinsically brighter SNeIa have simultaneously extended lifespans, as illustrated in the top figure. By taking this correlation into account, it is possible to significantly reduce the variance in absolute magnitude estimation. In the bottom figure, fainter SNeIa have been simultaneously stretched along the time and magnitude axis, and the bright SNe have been similarly shrunk. Figure from the Berkeley Labs [23].

is no evidence against $w_{DE} = -1$, and so ΛCDM stands strong. It is through accurately observing type Ia SNe that we hope to further constrain w_{DE} and whittle down the number of competing models of our universe.

Type Ia Supernova Complications

As mentioned already, galaxies cannot be used as standard candles, because their absolute magnitudes are dependent on the age of the universe. Are we certain that this is not the case for type Ia SNe? The physical process giving rise to type Ia SNe suggests that they should be immune to such evolution, but concerns persist. If evolution does exist for type Ia SNe it is significantly less than that of galaxies [30, 31]. That said, one group have found a significant correlation between the absolute magnitude of type Ia SNe and the metallicity of host galaxies [32]. They find that SNe in metal rich galaxies are 0.13 ± 0.06 magnitudes brighter after Phillips corrections, which is in fact in agreement with particular nucleosynthesis models [33].

One of the largest unknowns in estimating distances to SNe is the amount of dust existing between SNe and the Earth, and the effect this dust has on the light as it passes through it. It is known that dust particles found in interstellar gas, of around 10^{-7} m in diameter are very effective at absorbing ultraviolet and visible light, making observations redder. It is not known how much dust there is in other galaxies and whether it has the same composition as dust in our galaxy, although efforts have been made to determine its effect assuming it is of the same composition [34]. One important fact to consider in making dust corrections is that intrinsically faint SNe are also intrinsically redder [20].

A feature of the dust in our galaxy is the reddening effect it has on light, but the possibility of a "grey dust" which causes dimming without reddening has also been considered [35]. It has been suggested that such a dust could be the cause of high redshift SNe appearing dimmer than they would in a dust free non-accelerating universe, but this possibility was ruled out with the observation of what was at the time the most distantly observed SN, at redshift 1.7 [36]. That supernova was too bright relative to other distant SNe, and

it therefore fitted a model of a universe which was undergoing deceleration at redshift 1.7. Recall that when the matter density of the universe is non-neglibible, there is a period during which the universe decelerates. Objects whose light was emitted during a deceleration phase appear brighter than they would be, had deceleration occured. With the detection of the apparent cosmic deceleration, dust became an unviable explanantion, as it would be impossible for dust to increase the flux of light passing through it.

There are observed relationships in our galaxy between reddening and overall loss (extinction) of light. One of the more common of these is referred to as the CCM law. However, even in our galaxy, one important constant in the CCM law has a best fit value which varies by a factor of two depending on the line of sight [37]. Add to this the real possibility of dust evolution with redshift, and one sees that dust is possibly the greatest barrier at present to measuring the cosmological parameters using SNIa.

It is a growing concern that larger samples of distant SNe will not lead to improved knowledge about dark energy [17]. This concern stems from the fact that systematic uncertainties such as dust may now be the dominant source of error. In addition to dust, another important systematic is the uncertainty in the lightcurve fitting methods. Different lightcurve fitters, such as SALT, SALT2, and MCLS2k2 each fit for the peak magnitudes and distances in different ways. SALT2 for example uses a principal component analysis algorithm for missing data [38].

Finally, there is the systematic error which is well known in observational astronomy - the Malmquist bias [39]. Simply put, bright objects in the the sky are more likely to be observed than faint ones. As the SNe with known distances are all relatively nearby, they do not suffer from the Malmquist bias. The more distantly observed SNe which do suffer from the Malmquist bias, are naïvely believed to be nearer than they actually are, biasing the estimated cosmological parameters. This thesis is very much related to the Malmquist bias, as we shall soon discuss.

Other methods of constraining Ω_m and Ω_Λ

Type II SNe can also be used as standard candles, although currently they do not provide as tight bounds as those provided by type Ia SNe [40, 41]. The relationship which is used to standardize type II SNe is the one between their blackbody spectra and temperatures. Given temperature and flux over several nights, both obtained from observations, one can determine the SN's fractional expansion rate. Then, one can calculate the SN's absolute expansion rate from the absorption lines of the expanding sphere of hydrogen. Combining the absolute and fractional expansion, one can obtain the SN's distance.

The constraints from SNe are strongest on $\Omega_m - \Omega_\Lambda$. As already mentioned, the best constraint on $\Omega_m + \Omega_\Lambda$ comes from the precise shape of the power spectrum of the CMB. Evidence for the precise vale of Ω_m came first from galaxy clustering surveys, and then from Baryon Acoustic Oscillations (BAO) [24], whereby the clustering of baryonic matter follows a particular pattern. These methods are all in agreement with each other, in that the currently best fitting cosmological parameters are not in disagreement with any of their results [42]. This agreement leads cosmologists to refer to the ΛCDM model as the concordance model.

Statistics and The Contribution of this Thesis

The work which I will present in this thesis was originally in the form of two papers. The first of these papers [43], is presented in Chapter 2, and the second of these papers which has not yet been published, is presented in Chapter 3. The summaries of these chapters which will follow are based on the introductions of these two papers.

Thus far we have discussed in broad terms the history of the relevent cosmological ideas to which this thesis pertains. While the story told has been an interesting one, it would have been perhaps equally relevent to present a history of the relevent ideas from statistics, in particular the theorem of a Reverend Thomas Bayes:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A) \times \mathrm{P}(A)}{\mathrm{P}(B)}.$$

However, the statistical tools we will use in this thesis consist of a mixed bag, a group of ideas which would not sit as neatly aligned as perhaps the preceding overview of supernova cosmology does. For this reason, we have preferred to introduce the relevent statistical ideas as the necessity arises.

Chapter 2: Photometric Supernova Classification

There are two distinct ways in which a SN can be observed. The first of these is photometrically, whereby the apparent magnitude of the SN is observed over a series of nights in a variety of colourbands. Using this method it is cheap to observe a SN, but the observed photometric lightcurves do not in general provide sufficient information with which to classify a SNe as being of type Ia or otherwise. The second, relatively expensive way to observe a SN is spectroscopically. As already mentioned, type Ia SNe are unique in lacking hydrogen lines and containing the broad absorption lines of nickel and silicon, and so by obtaining the spectrum of a SN, one can say with certainty whether it of type Ia or otherwise.

We have already discussed that SNeIa provided the first widely accepted evidence for cosmic acceleration in the late 1990's [21, 22]. Based on small numbers of spectroscopically-confirmed type Ia SNe, those results have been confirmed by independent analyses and by a series of steadily improving SNIa surveys [24, 25, 26, 27, 28, 29, 15]. These modern SNIa surveys have acquired about an order of magnitude more Ia SNe than the early survets, now covering redshifts out to $z \sim 1.5$ [44, 45, 46, 47, 48, 49]. In addition, these surveys now have excellent photometric lightcurve coverage with rolling search strategies and multi-frequency lightcurve data with significantly better control of photometric errors due to the use of a single telescope to acquire the data in each major survey.

The next generation of SNIa surveys will be integrated into major photometric surveys, such as the Dark Energy Survey (DES) [50], PanSTARRS [51], SkyMapper [52] and LSST [53]. These next generation surveys promise

to catalyse a new revolution in SNIa research due to the sheer number of high-quality SNIa candidates that will be discovered: tens of thousands and perhaps millions of good SNIa candidates over the next decade. Spectroscopic followup will probably be limited to a very narrow subset of these candidates and so finding ways to best choose the followup subset to utilize the photometric data is a key challenge in SN cosmology.

In Chapter 2 we are interested in methods that can be used to accurately identify type Ia SNe from their lightcurves alone. This is a departure from traditional studies of type Ia SNe where all SNe used in cosmological parameter estimation have had their type confirmed via one or more spectra. Previous endeavors to use lightcurves for classification include those of Poznanski et al. [54], Brett et al. [55] and Rodney et al. [56]. In addition template-based photometric typing was used in the SDSS II SN survey [57] to select the most likely Ia candidates for spectroscopic followup with high confidence.

The origin of Chapter 2 was an entry into the Supernova Photometric Classification Challenge (SNPCC) run by Rick Kessler at the University of Chicago [58]. The SNPCC provided a simulated spectroscopic training data sample of approximately 1000 known SNe. The challenge was then to predict the types of approximately 20 000 other objects from their lightcurves alone. The challenge is now over, and the results from the different contributors have been summarised as a paper [59]. In Chapter 2 we present the details of a number of approaches to this problem, and their successes and failures. In it, we will discuss methods we have implemented to go from multi-band lightcurves to a classification, and the performance of the methods in the SNPCC. In particular, we will discuss in detail Kernel Density Estimation and boosting.

One of the main results which we discover and which we highlight in Chapter 2 is not specific to any particular classification technique, but a general problem faced in classification problems. That is, how a non-representative training sample negatively affects the performance of classification algorithms. For SN classification, spectroscopically confirmed SNe are brighter than unconfirmed photometric SNe, and so the training set is brighter than the ob-

jects which we wish to classify. While this problem is illustrated in Chapter 2, it is only in Chapter 3 that a solution is presented.

Chapter 3: BEAMS and debiasing

There are two ways that one can imagine using photometric candidates to measure cosmic expansion. The first approach is to use all of the SNe, irrespective of how likely they are to actually be of type Ia. This is the approach exemplified by the BEAMS formalism, which accounts for the contamination from non-Ia SN data using the appropriate Bayesian framework [60]. We consider BEAMS in greater detail in Chapter 3. The more conservative approach is to try to classify the candidates into Ia, Ib, Ic or II SNe, and then only use those objects that are believed to be SNeIa above some threshold of confidence, which is the essence of Chapter 2.

Where the emphasis in Chapter 2 is on absolute classification for parameter estimation, Chapter 3 focuses on the obtaining of accurate type probabilities for parameter estimation within the BEAMS framework. BEAMS (Bayesian Estimation Applied to Multiple Species) is a method for using unclassified data from distinct population subgroups to perform unbiased parameter estimation. Where other methods would require cuts to increase the purity of the contaminated set, BEAMS uses all of the data points in conjunction with their probabilities of being 'pure'. For application to SNe, the data points are the lightcurves of each SN, and the probabilities are the probabilities of being type Ia or another type of SN.

Chapter 3 consists roughly of three parts. First, we describe an implicit assumption made by the authors of the first BEAMS paper, relating to the independence of data and type probabilities, and present the necessary modifications to extend BEAMS to the case where the assumption does not hold. Then we perform a simple 1-D simulation to compare pre- and post- modification BEAMS, and show how the modification provides a 50% reduction in parameter estimation variance. Second, we perform three tests to quantify the importance of the type probabilities, one of which illustrates the effect of biasing the probabilities in various ways. Third, a general presentation

of the selection bias problem is given, and discussed in the context of SN cosmology.

Chapter 2

Supernova Classification

In this chapter we are interested in methods that can be used to identify SNeIa from their lightcurves alone. The work presented here was performed for the SNPCC, which provided a simulated spectroscopic training data sample of approximately 1000 known SNe and 20 000 other objects with lightcurves alone. We will here present the details of a number of approaches to this problem, and their successes and failures. We will discuss methods we have implemented to go from multi-band lightcurves to a final classification, and the performance of the methods in the SNPCC. In particular, we will discuss in detail Kernel Density Estimation and Boosting.

One of the main results in this chapter is that a non-representative training sample negatively affects the performance of the SN classification algorithms. For SN classification, spectroscopically confirmed SNe are brighter than unconfirmed photometric SNe, and so the training set is brighter than the objects which we wish to classify.

2.1 The Lightcurve Data

2.1.1 The Supernova Challenge Data

The SN data used in this thesis consists of approximately 20 000 simulated SN lightcurves with associated SN types released after the SNPCC[1]. The SNPCC data [2] are only relevant in our discussion of competition scores. Our reason for using the post-competition data is that it has numerous improvements and bug-fixes and is a more accurate simulation. The simulation was based on a DES-like survey [61]. The SNPCC dataset consists of a mixture of SN types (Ia, II, Ib, Ic), sampled randomly with proportions given by their expected rates as a function of redshift.

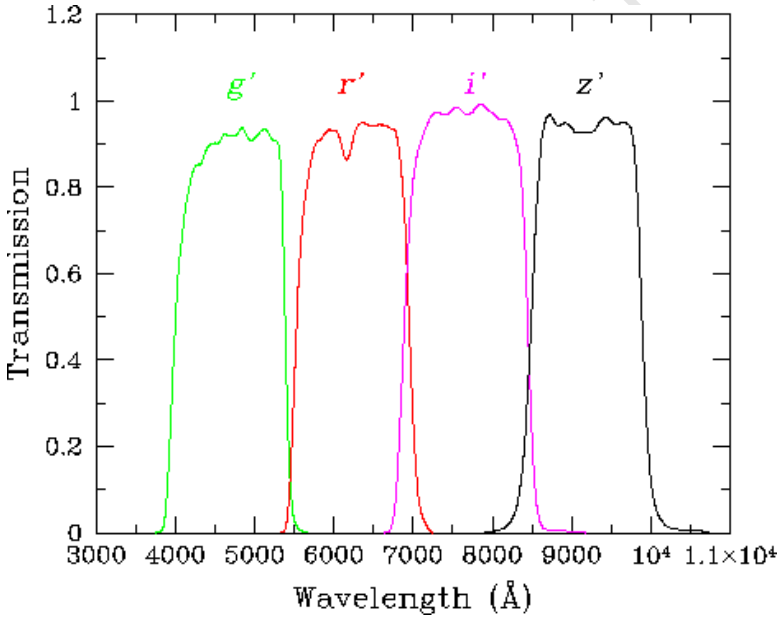

Figure 2.1: Filters used for the DES survey, illustrating the proportion of light detected through each of four filters at different frequencies. Figure taken from Bartelmann, M. & White, S. [62]

Each simulated SN consists of flux measurements in the *griz* filters illus-

[1]These post-SNPCC lightcurves are available at http://sdssdp62.fnal.gov/sdsssn /SIMGEN_PUBLIC/

[2]These competition lightcurves are available from http://www.hep.anl.gov/ SNchallenge/

trated in Figure 2.1 [63]. In effect, the flux of an object in a particular colour band is the integral over all frequencies of the filtered flux. Distances were calculated assuming a standard ΛCDM cosmology ($\Omega_M = 0.3, \Omega_\Lambda = 0.7$ and $w = -1$), with anomalous scatter around the Hubble diagram drawn from a Gaussian distribution with $\sigma_m = 0.09$. The SNPCC data includes two selection criteria, that is criteria on being photometrically observed. Each object is required to have at least one observation with a signal-to-noise ratio (S/N) above 5 in any filter, and must also have at least 5 observations after explosion. A complete summary of the SNPCC is given in the challenge and results papers [58, 59].

We (the authors of [43]) took part in two of the SNPCC challenges. In the first (`+HOSTZ`) challenge, participants were provided with photometric host galaxy redshift estimates, based on simulated galaxies analysed using the methods discussed in Oyaizu et al. [64] and asked to return the type of each SN candidate. In the second (`-HOSTZ`) challenge, no redshift estimates for simulated SNe were provided. Both challenges are considered in this chapter, but with emphasis on the `+HOSTZ` challenge. We did not attempt to distinguish between non-Ia sub-types (such as type II and type Ib/c SNe).

Figure 2.3 shows the multi-band lightcurve data for a randomly selected Ia and non-Ia SN. To these measurements, a parametric curve has been fitted as discussed in Section 2.1.2.

Training Samples

The aim of the SNPCC was for the participants to classify each of the simulated SNe into Ia or non-Ia (and non-Ia sub-classes if they desired) with the aim of minimising false Ia detections and maximising correct Ia detections. A spectroscopic training sample of $\sim$ 1000 SNe of known type was provided which was a simulation of spectroscopic observations on expected telescopes. The imagined telescopes were set to have limiting magnitudes, that is mimimum observable fluxes, through r band and i band filters of 21.5 and 23.5 respectively. Because spectroscopy is harder than photometry, the distribution of SNe in this spectroscopic sample is much brighter on average

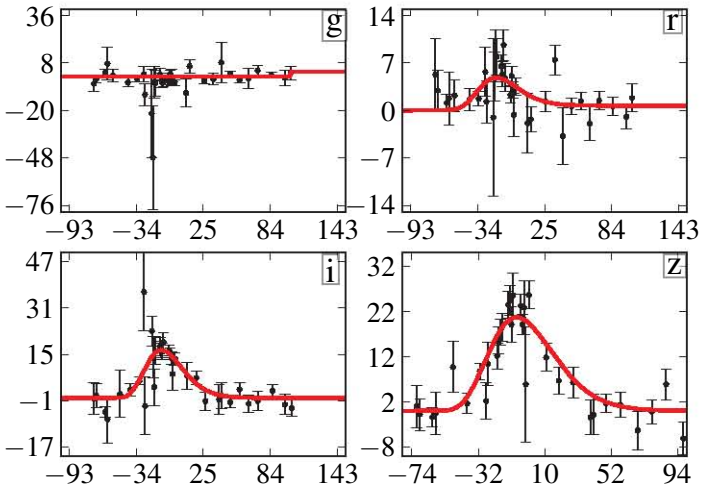

Figure 2.2: A typical well-sampled SNIa lightcurve, in this case at redshift z = 0.694. Overplotted is the best-fitting curves using Eq. 2.1.

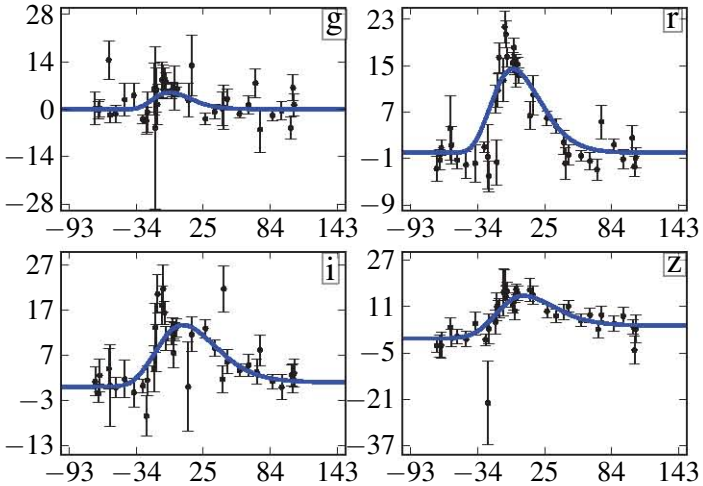

Figure 2.3: The lightcurve of a typical well-sampled non-Ia SN at z = 0.663. Overplotted is the best-fitting curves using Eq. 2.1.

than the full photometric sample, and hence is not representative of the full sample. This is a crucial point to appreciate and as a result in this chapter we refer to this sample as the *non-representative training sample*.

We will often compare with the results from a representative sample, generated by spectroscopically following up a sample of objects that is representative of the full photometric SN population. To produce an unbiased training sample, at the conclusion of the SNPCC when the types of each SNPCC object were revealed, we randomly selected $\sim$ 1000 SNe from the entire SNPCC dataset, and considered the effect of using this as our training sample. This is referred to in the text as the *representative training sample*. We refer to the SNe that require classification as the *unclassified set*.

2.1.2 Post-processed Data

Fitting a parameterised curve

In the provided photometric data the number, sampling times, frequency and accuracy of the sampled magnitudes varies greatly for each SN, as illustrated in Figure 2.3. In order to standardize the raw data, we fit by weighted least squares, a parameterised function to the lightcurves in each of the four colour bands. Our parameters are $(A, \phi, \psi, k, \sigma)$ and the flux in each band is taken to be [3]:

$$F(t) = A\left(\frac{t-\phi}{\sigma}\right)^{k} \exp\left(-\frac{t-\phi}{\sigma}\right) k^{-k} e^{k} + \Psi(t). \tag{2.1}$$

The five parameters to be fit in each band have the following interpretations: $A + \psi$ is the peak flux, ϕ is the starting time of the explosion, k determines relative rise and decay times and σ is a temporal stretch term. An additional parameter τ, the time of peak flux, is determined by these parameters via $\tau = k \cdot \sigma + \phi$. The function Ψ is a "tail" function such that $F(t) \to \psi$ as $t \to \infty$. The exact form (illustrated in Figure 2.4) of Ψ is:

[3] This function has a single maximum and therefore cannot fit examples which have a double peak. However, for the SN data we consider this turns out not to be an important limitation.

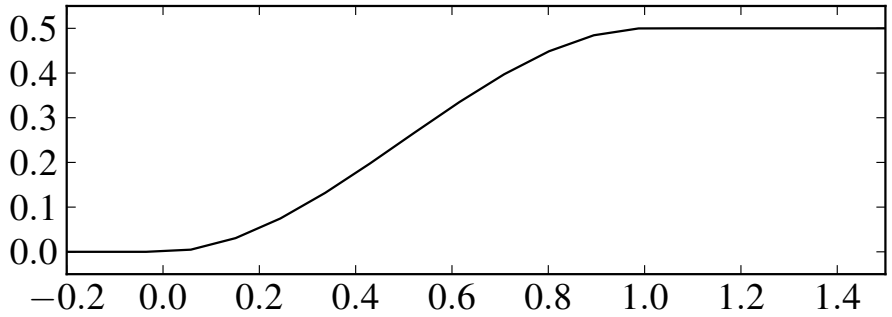

Figure 2.4: The tail function Ψ, which is used in fitting Eq. 2.1. Parameters (ψ, ϕ, τ) are kept fixed at $(0.5, 0, 1)$ here.

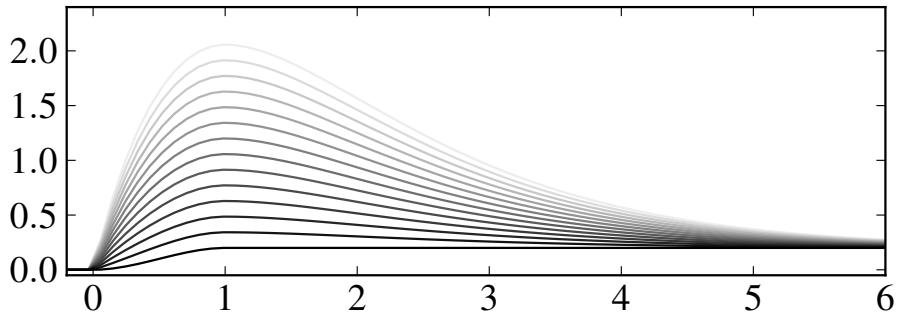

Figure 2.5: The effect of varying A on the function $F(t)$ from low (dark) to high (light). We keep the parameters (k, σ, ϕ, ψ) fixed at $(1, 1, 0, 0)$.

$$\Psi(t) = \begin{cases} 0 & -\infty < t < \phi \\ \text{cubic spline} & \phi < t < \tau \\ \psi & \tau < t < \infty \end{cases}$$

where the cubic spline is uniquely determined to have zero derivative at $t = \phi$ and $t = \tau$. The effect of each parameter is illustrated in Figures 2.5 to 2.8. We have also posted two files at our website [65], each containing 200 randomly selected and fitted SNe to illustrate the range of fits possible. With five free parameters, A, ψ, ϕ, k and σ in each colour band and a host redshift (in `+HOSTZ` challenge), we have 21 parameters specifying each SN. We do not

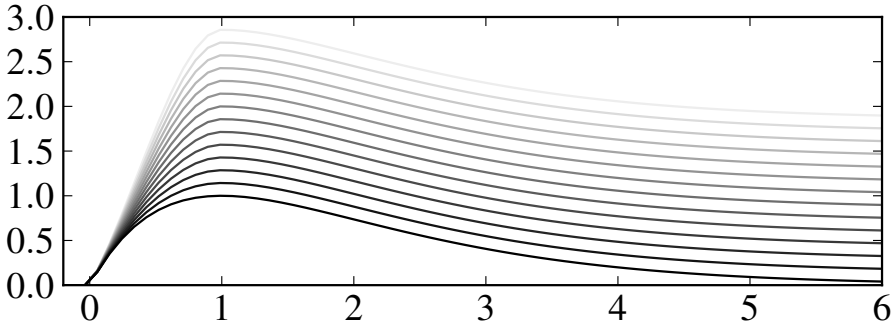

Figure 2.6: The effect of varying ψ on the function $F(t)$ from low (dark) to high (light). We keep the parameters (k, σ, ϕ, A) fixed at $(1, 1, 0, 1)$.

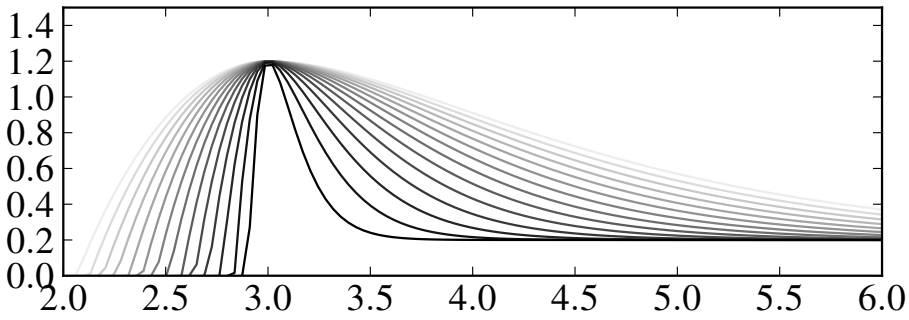

Figure 2.7: The effect of varying σ on the function $F(t)$ from 0.1 (dark) to 1.0 (light). Increasing σ linearly stretches the curve away from the $t = \phi$. We keep the parameters (A, ϕ, k, τ) fixed at $(1, 0, 1, 3)$.

require that there be any correlation between the derived parameters in any band, e.g. between explosion time, time at peak or stretch. It would be a natural extension to allow for such correlations, as a reduction in the number of dimensions of the KDE would reduce it's variance. We leave this to future work.

Sparse datasets

About 5% of all the SNe had fewer than 8 observations in one or more of the four bands. To avoid overfitting, we did not fit these SNe with Eq. 2.1.

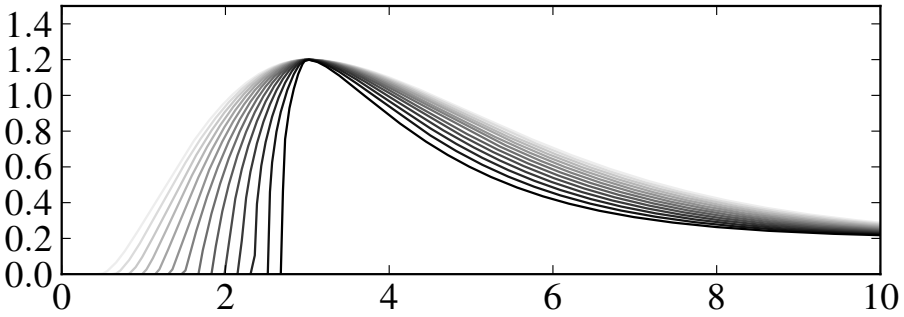

Figure 2.8: The effect of varying k on the function $F(t)$ from 0.2 (dark) to 1.8 (light). Increasing k decreases the ratio of rise to decay time (Rapid rise relative to decay means low k). We keep the parameters (A, σ, ϕ, τ) fixed at $(1, 1.5, 0, 3)$.

Instead, these sparsely sampled SNe were each fit to a 5 dimensional point - the maximum flux in each of the four colour bands plus the host redshift. The KDE and boosting methods (Section 2.2) were applied to these SNe in the same way as was done in the 21 dimensional case (Sections 2.3.1, 2.3.2). Unless otherwise stated, discussions and illustrations will all reference the 95% of SNe which had 8 or more observations in all bands and hence were fit with 21 parameters.

SALT fits

In Section 2.3.4, we consider classification methods that require information on the distances to SNe to constrain their type. Distance moduli for all SNPCC SNe were derived using the publicly available lightcurve fitter SALT2 [66]. Fits were carried out using the g, r and i passbands (i.e. z colour band data was not included). All available SNe were considered, which is significantly more liberal than the usual data-quality cuts applied during past SN cosmology analyses [48]. In this way, we maximized the number of SNe available for this work. We applied SALT2 to 1256 SNe available in the non-representative training sample. Immediately, we found that 165 SNe failed to pass through SALT2 with the reported error of the lightcurve either

having too low a S/N or missing g-band data. We did not investigate these errors further and simply exclude these SNe. Furthermore, when the S/N is low, SALT2 fits some SNe but returns a default upper limit magnitude of 99 and is unable to produce meaningful parameters from the lightcurve fit. This problem affected 62 SNe in the training sample, which were also removed from the sample. For the 1029 SNe that were successfully fitted, SALT2 returned a best fit value for the four parameters x_0, x_1 and c for each event. These parameters relate to the peak magnitude and stretch/colour corrections of the lightcurve, details of which can be found in Guy et al. [66]. The best-fit Ia model lightcurve was also returned in the observer frame, which we used to calculate a χ^2 goodness of fit value for each SN in each passband (g,r,i) which are used in Section 2.3.4 to classify SNe. Distance moduli are calculated using the best fit parameters using

$$\mu = (m_B - M) + \alpha x_1 - \beta c. \tag{2.2}$$

where we used values of $\alpha = 0.1$, $\beta = 2.77$ and $M = 30.1$ to calculate the distance moduli, as discussed in Hicken et al.[67]. These values are consistent with those found in other analyses and were not expected to significantly affect our results. Figure 2.9 shows the Hubble diagrams for the two training samples considered in this analysis. Also shown is the best-fit cosmology to each Ia dataset assuming a flat ΛCDM model. In the non-representative training sample, non-Ia SNe are predominately found at lower redshifts than the representative training sample due to the effective magnitude cuts coming from the spectroscopic requirement of the non-representative sample.

2.2 New Classification Methods

We now describe in very general terms the classification algorithms we have used. In order to classify a given object Y as either Ia or non-Ia, one would like the posterior probabilities $\mathrm{P}\,(Y = \mathrm{Ia}|x)$ and $\mathrm{P}(Y = \text{non-Ia}|x) = 1 - \mathrm{P}(Y = \mathrm{Ia}|x)$. Here x are the parameters or features that characterize the

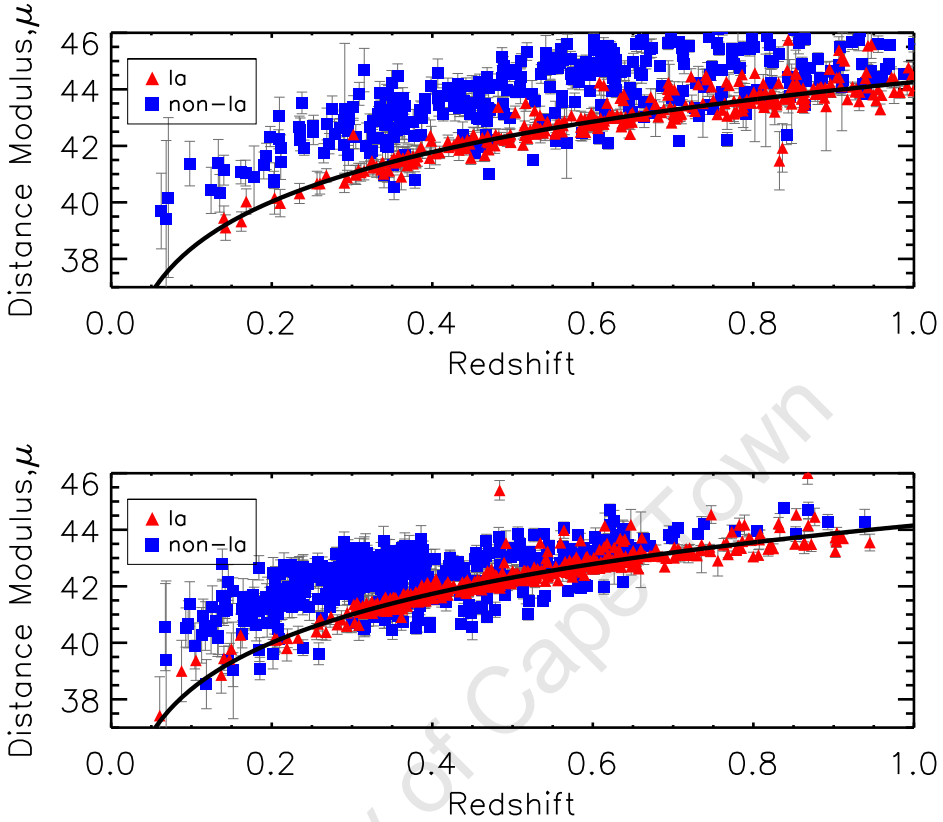

Figure 2.9: Hubble diagrams for the 2 training samples considered in this chapter. SNeIa are shown as red triangles, while non-Ia SNe are plotted as blue squares. Also shown is the best-fit cosmology to each SNIa sample. (*Above*) The representative training sample, with $\Omega_m = 0.23$. (*Below*) The non-representative training sample, as provided for the SNPCC, with $\Omega_m = 0.3$.

SN. Knowing these posterior probabilities is equivalent to knowing the *odds*:

$$odds(x) = \frac{\mathrm{P}(Y = \mathrm{Ia}|x)}{\mathrm{P}(Y = \text{non-Ia}|x)}.$$

Now one classifies Y as a Ia for example if $odds(x) > 1$, i.e if $\mathrm{P}(Y = \mathrm{Ia}|x) > 0.5$. The two methods we discuss in this Section approximate the *odds* in different ways:

1) Kernel Density Estimation estimates $\mathrm{P}(x|Y = \mathrm{Ia})$ and $\mathrm{P}(x|Y = \text{non-Ia})$, the density of the features in classes Ia and non-Ia respectively, and then uses Bayes' formula,

$$\mathrm{P}(Y = Ia|x) = \frac{\mathrm{P}(x|Y = \mathrm{Ia})\mathrm{P}(Y = \mathrm{Ia})}{\mathrm{P}(x)},$$

to express the *odds* in terms of these quantities:

$$odds(x) = \frac{\dfrac{\mathrm{P}(x|Y = \mathrm{Ia})\mathrm{P}(Y = \mathrm{Ia})}{\mathrm{P}(x)}}{\dfrac{\mathrm{P}(x|Y = \text{non-Ia})\mathrm{P}(Y = \text{non-Ia})}{\mathrm{P}(x)}}$$

$$= \frac{\mathrm{P}(x|Y = \mathrm{Ia}) \cdot \mathrm{P}(Y = \mathrm{Ia})}{\mathrm{P}(x|Y = \text{non-Ia}) \cdot \mathrm{P}(Y = \text{non-Ia})}.$$

2) Boosting directly estimates $odds(x)$ through regression methods, as a sum of small trees built by a type of functional gradient descent. These methods are discussed in detail below.

2.2.1 Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a non-parametric method for estimating the probability density function (pdf) of a random variable. Within this chapter, the probability densities of the post-processed data described in Sections 2.1.2 - 2.1.2 are used for classification. Pdfs are useful as we may base a classification rule upon the relative probabilities that a candidate SN is either type Ia or not type Ia. Such a classification rule will require both the Ia and the non-Ia probability densities for the observed SN data. KDE enables us to derive these pdfs in a fairly model-independent manner, as we now discuss.

Suppose we have a set of d observables and that we would like to estimate the value of the pdf at a point $\vec{x}$ in this d-dimensional space. Given a training set with n observations, i.e. n points $\vec{X}_i$ in this d-dimensional space, the

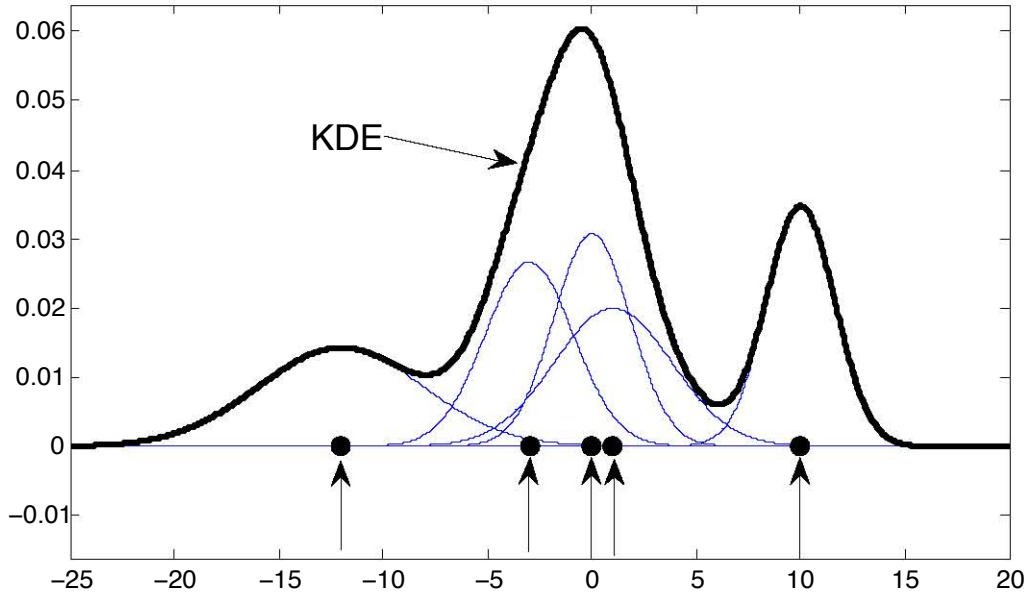

Figure 2.10: Schematic figure illustrating the idea of a KDE in one dimension. The training data points are shown as dark points with arrows. The Gaussian kernels are shown together with the sum of the kernels. Note that the KDE is not normalized in this figure and is thus close to what we actually used.

Kernel Density Estimate (KDE) is given by

$$\hat{f}_h(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K_i\left(\frac{\vec{x}-\vec{X}_i}{h}\right), \tag{2.3}$$

where $\hat{f}_h(\vec{x})$ is the KDE, $\vec{X}_i$ is the i-th training observation, K_i is the kernel function for the i-th training observation and h is the global kernel bandwidth. h is a tuning parameter: the kernels become more "peaked" about the training observations as h becomes smaller. The optimal bandwidth may be obtained by cross-validation (see Appendix A1). The choice of kernel is arbitrary, except that any proposed kernel should satisfy the following two conditions:

- $\int K(\vec{x})d\vec{x} = 1$

- $K(-\vec{x}) = K(\vec{x})$

The first condition ensures that the KDE integrates to unity and that all observations carry equal weight, whilst the second condition ensures that the KDE is unbiased and that each kernel is centered about one of the n d-dimensional training data points. The basic idea of the KDE method is illustrated in Figure 2.10 in a simple 1D example. A commonly used kernel (and the kernel that we will use in this chapter) is a multivariate Gaussian, normalized to unit volume:

$$K\left(\frac{\vec{x}-\vec{X}_i}{h}, \Sigma_i\right) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left\{-\frac{1}{2}\left(\frac{\vec{x}-\vec{X}_i}{h}\right)^T \Sigma_i^{-1} \left(\frac{\vec{x}-\vec{X}_i}{h}\right)\right\}. \tag{2.4}$$

Here $\vec{x}$ and $\vec{X}_i$ are d dimensional vectors and Σ_i is a $d \times d$ covariance matrix that changes the orientation and shape of the kernel around each training observation i; for example the covariance matrix Σ_i can be estimated from the nearest ℓ neighbours of a training data point, which is what we do, as described in Section 2.3.1 and as illustrated in Figure 2.11. This approach provides the possibility of adapting the kernel to local variation. In contrast the bandwidth parameter h affects the global behaviour of the kernels. While it is more common to choose the covariances to be equal, for the SNPCC and the current application this would have been a bad choice (as described in Section 2.3.1).

Integration over data errors

In order to classify a SN with lightcurve measurements $\vec{x}$, we must evaluate the KDEs at $\vec{x}$. However in our case we are not sure where $\vec{x}$ lies in parameter space as the lightcurve measurements have errors and are not perfectly sampled.

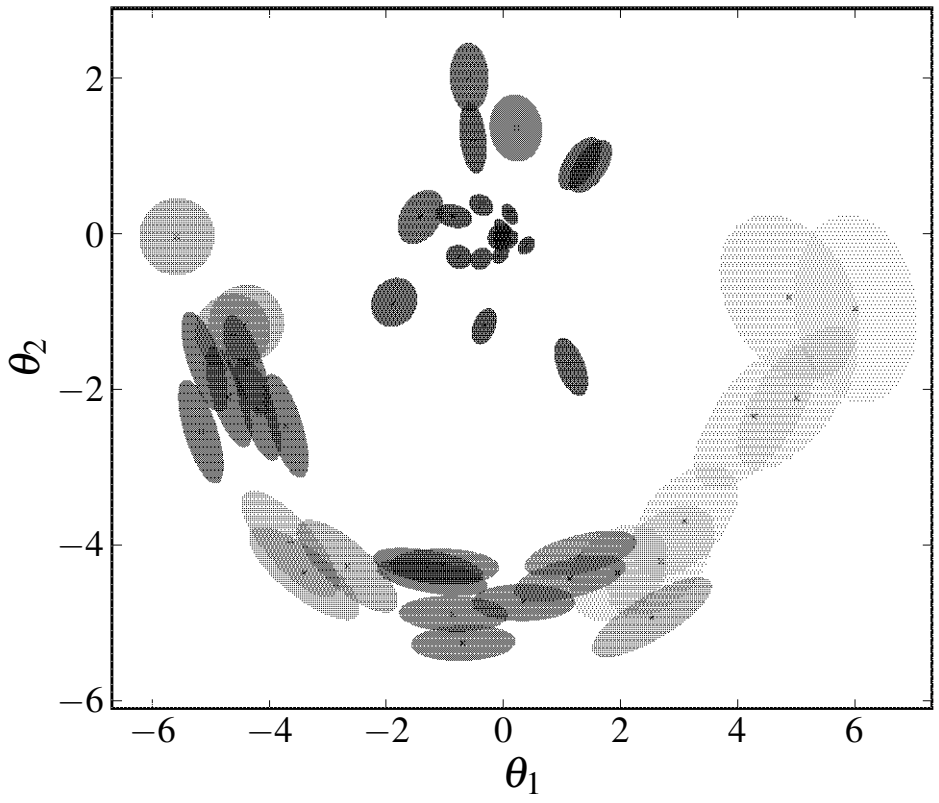

Figure 2.11: A realisation of fifty points from an unusual distribution. Around each observed point a kernel is constructed. The axes of each kernel are the eigenvalues of the point's (2×2) Σ_i matrix (Eq. 2.4). Each Σ_i is the covariance matrix of the nearest ℓ points multiplied by the global bandwidth, h. Here $h = 0.6$ and $\ell = 10$.

Using a Gaussian kernel, we write the KDE as

$$\hat{f}(x) = \frac{1}{n} \sum_i \frac{1}{h^d} K\left(\frac{x - X_i}{h}, \Sigma_i\right) \tag{2.5}$$

For simplicity we suppress vector notation but all quantities (other than n and h) are d dimensional vectors or matrices, and the index i runs over the points in the training set.

Now assume that the location of a point in the d dimensional space is not known exactly and is instead given by a Gaussian pdf. We take the mean to be x and the covariance matrix to be Y. The KDE value is then given by integrating the KDE over the unknown pdf of the point being classified:

$$\int dz K\left(z-x, Y\right) \hat{f}(z) = \frac{1}{n} \sum_i \frac{1}{h^d} \int dz K\left(z-x, Y\right) K\left(\frac{z-X_i}{h}, \Sigma_i\right).$$

We notice that this reverts back to the original value if K is a delta-function located at x. Further, the function being integrated is a product of two Gaussians, which is itself another Gaussian. The KDE value then simplifies to

$$\hat{f}(x) = \frac{1}{n} \sum_i \frac{1}{h^d} K\left(\frac{x-X_i}{h}; \Sigma_i + h^{-2d} Y\right), \tag{2.6}$$

i.e. the KDE kernels simply have an increased variance, given by the sum of their covariance matrix and the covariance matrix of the point being evaluated, scaled by h^{-2d}. The importance of including this increased variance for uncertain observations should not be ignored, especially when the variances of the points being classified are large (as is the case in our situation). Correctly implementing Eq. (2.6) can significantly improve classification performance. In Section 2.3.1 we compare analyses on the SN data including and ignoring the covariance Y.

2.2.2 Boosting

Boosting is a learning algorithm for classification [68]. Until recently the most commonly used boosting algorithm was AdaBoost [69]. AdaBoost works by combining weak-classifiers into a committee, whose combined decision is significantly better than that of individual weak-classifiers. The precise workings behind AdaBoost's success remained hazy until it was shown that boosting produces the powerful committee by sequentially adding together weak-classifiers calculated by steepest descent [70]. The further ideas of slow learning [71] and bagging [72] were later introduced to boosting, culminating eventually in the Gradient Boosting Machine (GBM) algorithm. The algo-

rithm, implemented as a package in the statistical programming language R[4], is described in this section. A brief discussion of trees and loss functions is first presented in preparation for the presentation of the GBM algorithm.

Tree functions

The most widely used weak-classifiers (a.k.a. basis functions) in boosting are trees. Trees are discontinuous functions which take discrete values in different regions of a domain. That is to say, a tree T has the form:

$$T(\vec{x}) = \begin{cases} z_1 & \text{if } \vec{x} \in R_1 \\ \vdots & \\ z_K & \text{if } \vec{x} \in R_n \end{cases}$$

where the K distinct regions $R_1 \cdots R_K$ together partition $\vec{x}$-space. The region boundaries can be described through the branchings of a tree, as illustrated in Figure 2.12. For boosting, it is common to only use trees of a very simple form, that is only trees with branchings of the form $x^{(i)} < v$, where $x^{(i)}$ is one of the dimensions of $\vec{x}$-space and v is a real number. In the case of the SNPCC, $\vec{x}$ are the parameters fitted to the lightcurves in Section 2.1.2.

Loss function for classification

Suppose we have observed n training points, each consisting of data and type: $(\vec{X}_i, \tau_i)$, where the data $\vec{X}_i$ is a d-dimensional vector, and the type τ_i is ± 1, corresponding to the two classes. Suppose that we are required to find a function $F : R^d \to R$ which minimises the following *loss function*:

$$L(F) = \sum_{i=1}^{n} \log\left(1 + \exp\left[-2F(\vec{X}_i)\tau_i\right]\right). \tag{2.7}$$

The specific form chosen for the loss function 2.7 can be explained by considering its partial derivatives with respect to $F(\vec{X}_i)$. Doing so, it can be

[4]R and its associated packages can be downloaded from http://www.r-project.org.

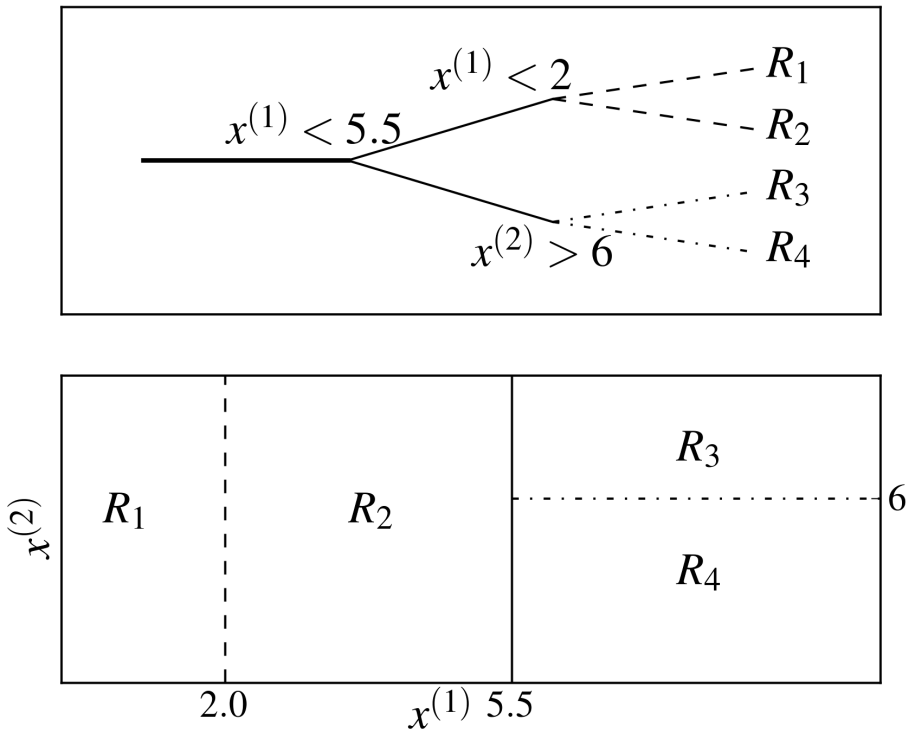

Figure 2.12: (*Above*) A tree of depth 2 for classifying an object into one of 2^2 regions. (*Below*) The tree domain containing 2^2 distinct regions as defined by the tree.

shown that the form of F which minimises (2.7) is given by [73]:

$$F(\vec{X}_i) = \frac{1}{2}\log \frac{\#\text{ observations: } \vec{X}_i, \tau = 1}{\#\text{ observations: } \vec{X}_i, \tau = -1} \tag{2.8}$$

Eq. (2.8) is an approximation to half the *log odds* (the log of the *odds*):

$$\textit{log odds} \equiv \log \frac{\mathrm{P}(\tau_i = 1|\vec{x} = \vec{X}_i)}{\mathrm{P}(\tau_i = -1|\vec{x} = \vec{X}_i)}. \tag{2.9}$$

Eq. (2.9) provides a key result: a function which minimises the loss func-

tion (2.7) is a good approximation to half the *log odds*. A good approximation to the *log odds* is exactly what is needed for classification problems. The boosting algorithm aims to approximately minimise this loss function and in so doing arrive at an approximation of the *log odds* which can then be used for classification.

If you have observations at every possible data point, you can directly approximate the *log odds* through (2.8). In reality, you will not have observations at all possible data points, and so cannot do this. This corresponds to not having observed all possible lightcurves, and so needing to make inferences from similar lightcurves. Boosting does this inference through constrained minimisation of the loss function, as described in the following section.

The Gradient Boosting Machine

The Gradient Boosting Machine [71] works by sequentially adding new trees to a function F, each addition reducing $L(F)$ (2.7) and so improving the approximation of F to half the *log odds*.

The trees, which have depth D, are appended to F at each of the M iterations of the GBM algorithm. Choosing larger M and D values results in a final $L(F)$ nearer to the global minimum value (2.8). However, our end objective is not to reach the global minimum but to construct a good approximation to the *log odds*, and trees of lower depth are generally better suited to this end, being less prone to fitting noise.

Algorithm 1 (below) outlines an implementation of the GBM. A few subtleties have been omitted from it here, and we refer you to Appendix A5 for a fuller description. We recommend watching our demonstrative animation of the algorithm while reading Algorithm 1. The animation can be found at our website [65].

Algorithm 1 - Gradient Boosting Machine

· Input: $\vec{X}_i, \tau_i$ for observations $i = 1$ to n.

· Initialize: $F_0(\vec{x}) \leftarrow \frac{1}{2}\log\frac{1+\bar{\tau}}{1-\bar{\tau}}$ where $\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n}\tau_i$.

· Initialize: $z_i \leftarrow 0$ for observations $i = 1$ to n. The z_i's will measure how much of a "misfit" each observation is.

· Choose tree depth D and number of trees M.

· for $m = 1$ to M:

1) for $i = 1$ to n, update z_i:

$$z_i \leftarrow -\frac{\partial L}{\partial F_{m-1}\left(\vec{X}_i\right)} = \frac{2\tau_i}{1 + \exp\left[2F(\vec{X}_i)\tau_i\right]}$$

2) Fit by least squares T_m, the new tree: $z_i \sim T_m(\vec{X}_i)$.
(where T_m has regions $R_{m,1} \cdots R_{m,2^D}$ fitted to minimise the ingroup variance: see Appendix A3 for details.)

3) Choose constants $\gamma_{m,1} \cdots \gamma_{m,2^D}$ for $R_{m,1} \cdots R_{m,2^D}$. (chosen to minimise $L\left(F_{m-1} + T_m\right)$)

4) $F_m \leftarrow F_{m-1} + T_m$

· Finally, $F \leftarrow F_M$.

F is our final approximation to half the *log odds*, and it can now be used to classify with a simple rule of the form:

IF $F(\vec{x}_i) > v \Rightarrow \tau_i = 1$; ELSE $\tau_i = -1$,

where the optimal v depends on the Figure of Merit.

Notice that the variable z_i, updated in step 1, is positive if $\tau_i = 1$ and negative if $\tau_i = -1$. For this reason, when T_m is fit to the z_i's at step 2, observations of the same type are more likely to fall into the same region of T_m. Moreover, observations with large z_i's carry more weight while fitting T_m, and hence are even more likely to be placed with objects of the same type. This acts to place special attention on unusual objects, or objects whose type is not clear.

While values are fitted for each tree region in step 2 (as described in Appendix A3), these values will not necessarily result in a reduced $L\left(F_{m-1} + T_m\right)$. Hence at step 3 of the algorithm, $\gamma_{m,k}$ values are explicitly chosen to minimise $L\left(F_{m-1} + T_m\right)$. In effect, only the tree *shape* is taken from step 2.

2.3 Results

The entries in the SNPCC were evaluated using the Figure of Merit (FoM):

$$\begin{aligned} f(N_{Ia}^{\checkmark}, N_{non-Ia}^{\times}) &= \text{efficiency} \times \text{pseudo-purity} \\ &= \left(\frac{N_{Ia}^{\checkmark}}{N_{Ia}^{TOT}}\right) \times \left(\frac{N_{Ia}^{\checkmark}}{N_{Ia}^{\checkmark} + 3 \cdot N_{non-Ia}^{\times}}\right), \end{aligned} \tag{2.10}$$

where $N_{Ia}^{\checkmark}$ is the number of correctly classified SNeIa, $N_{non-Ia}^{\times}$ is the number of non-Ia SNe classified as SNeIa, and N_{Ia}^{TOT} is the total number of SNeIa. Had the coefficient of $N_{non-Ia}^{\times}$ in the denominator of the pseudo-purity term been 1 and not 3 the term would have been true purity, i.e. the proportion of SNeIa in the final Ia-classified group. How relevant the FoM (2.10) is to cosmology is not absolutely clear, but it is a robust measure of how well a classification algorithm penalizes both missed detections and false discoveries. For applications such as BEAMS [60] a FoM which takes type probabilities as inputs would be more useful.

In this section we discuss the implementation and performance of each of our methods. Unless stated otherwise, the scores given in this section refer to the SNPCC, while all figures are using the post-SNPCC data described in Section 2.1.1. Of particular interest to us is the comparison of results obtained when the training is done with representative and non-representative samples. We also briefly mention applications that these methods have previously found in cosmology and related fields.

2.3.1 21D KDE

Application

Kernel Density Estimators have been used before in astronomy for estimating the probability density function from a discrete or noisy data set [74, 75, 76], identifying groups [77] and clusters [78] in galaxy surveys, and determining the timings of millisecond pulsars [79] and gamma-ray bursts [80], to name a few examples.

In Section 2.1.2 we described how we fit the SN lightcurves in each of the 4

bands using the parameterised function (2.1), resulting in twenty lightcurve parameters. With the addition of host redshift in the case of the `+HOSTZ` challenge, each SN is described by a 21 dimensional (21D) point. We use KDE to approximate the 21D Ia and non-Ia probability density functions based on the training data.

We allowed the 21D training points to have different covariance matrices, as described in Section 2.2.1. As previously mentioned a single global covariance is most common for KDE, but in cases where a pdf has large regions of high and low probability, this can be problematic. In low probability regions the kernel density will be too "spikey" while in high probability regions it will be too smooth. To understand this, consider what would happen if, in Figure 2.11, the ellipses were constrained to all be of the same size (chosen too small and the low probability region would have "bumps", too large and the high probability region would lose features.). The 21D points for the SNPCC are not uniformly distributed, as illustrated by the cumulative plots of Appendix A6, and so are susceptible to this problem. Using cross-validation as described in Appendix A1 we chose $\ell = 10$ and $h = 0.6$ (using the notation from Section 2.2.1).

Having constructed two KDEs from a training sample, each unclassified SN may be classified as follows:

1. Fit Eq. 2.1 to each of the four lightcurves thus obtaining a 21D point for the candidate.

2. Evaluate the Ia and non-Ia kernel probabilities derived from the training sample at the 21D point, and then evaluate the *odds*.

3. If the *odds* (or *log odds*) is above some threshold, classify as Ia.

In cases where one or both of the KDEs are a poor representation of the underlying pdf, it may be preferable to modify step (3). For example if one of the KDEs is particularly inaccurate, one may prefer to classify by using only the other KDE. For the SNPCC leaving step (3) unchanged was advisable, as can be deduced from Figure 2.13. The lines in Figure 2.13 are lines of constant *odds*. If KDEs are accurate approximations to pdfs, a line

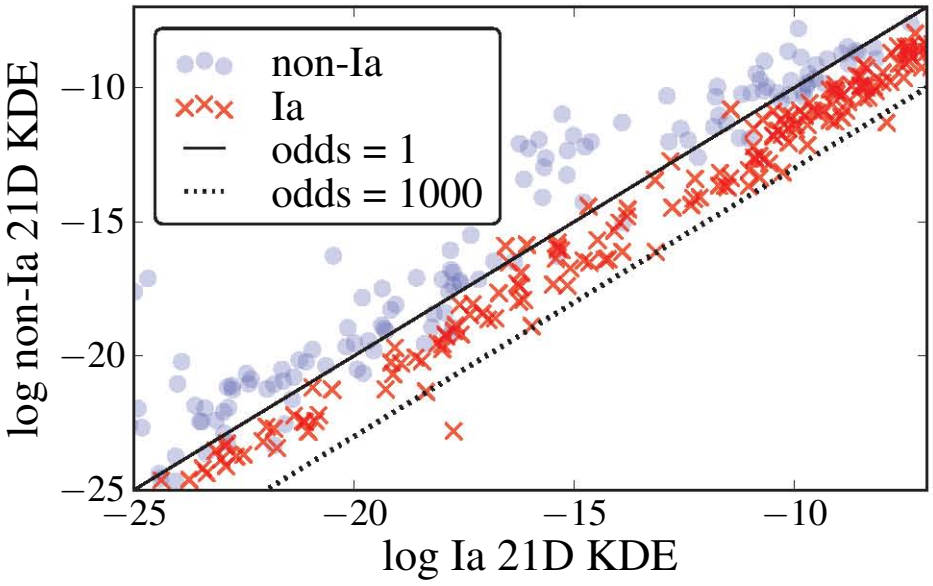

Figure 2.13: Ia (red crosses) and non-Ia (blue circles) in the non-representative training sample. The KDE values at calculated using tenfold cross-validation.

of constant *odds* is optimal for discriminating between Ia's (below the line) and non-Ia's (above the line), irrespective of the FoM used. Furthermore, if the KDEs are accurate approximations to pdfs, there should be an equal number of Ia's and non-Ia's on the line *odds* = 1 and 1000 times more Ia's as non-Ia's on the line *odds* = 1000. This is roughly observed in Figure 2.13 and so we can proceed to choose the *odds* line which maximizes the SNPCC FoM.

For the entry in the SNPCC, we failed to include the parameter covariance matrices when calculating KDE values (in effect, we set Y to be a matrix of zeros in Eq. 2.6). Our final score suffered as a result - the benefit of correctly implementing the calculation (2.6) is illustrated in Figure 2.15, where we see from both the histograms and the cumulative plots an increased separation between Ia's and non-Ia's when Eq. (2.6) is correctly implemented. We find a 15% increase in score when correctly implemented on the post-SNPCC data.

The KDE method still obtained the second and third highest scores in the `-HOSTZ` and `+HOSTZ` competitions respectively, with scores of 0.37 and 0.39. Of interest is that the 20D KDE (`-HOSTZ`) is almost as good at classifying as the 21D KDE (`+HOSTZ`). The winning competition scores were 0.51 (`-HOSTZ`) and 0.53 (`+HOSTZ`) [59].

Non-representative vs representative

As with all of our methods, we constructed classifiers using both the non-representative sample provided and a representative sample of equal size, as described in Section 2.1.1. In each case, the remaining unclassified SNe were used as a test of the performance of the classifier.

Figure 2.14 carries useful information about the performance of the non-representatively trained KDEs and representatively trained KDEs. For example, the efficiency of classifying Ia's with a *log odds* threshold of 2 is simply 1 minus the cumulative value of the unclassified Ia's (solid red) at *log odds* $= 2$. For both representatively and non-representatively trained KDEs $1 - F_{\mathrm{Ia}}(2) \approx 0.75$, meaning that about 25% of SNeIa are correctly classified when a threshold of *log odds* $= 2$ is used.

To obtain high purity, the *log odds* threshold must be chosen such that the non-Ia cumulative frequency is high compared to the Ia cumulative frequency. To obtain high efficiency, the *log odds* threshold must be chosen such that the Ia cumulative frequency is low. Putting these together, to obtain both high purity and high efficiency, a *log odds* threshold must be found at which the non-Ia cumulative frequency is high and the Ia cumulative frequency is low.

The dashed lines in Figure 2.14 are the cumulatives of the training data using tenfold cross-validation. In the case of representative training, we see that these are accurate predictors of the true cumulatives. But in the case of non-representative training, the non-Ia cumulatives of training and unclassified SNe are vastly different. If in the case of non-representative training one assumed that the training samples were in fact representative, one would predict a non-Ia misclassification rate of under of 10% using a *log odds* cutoff

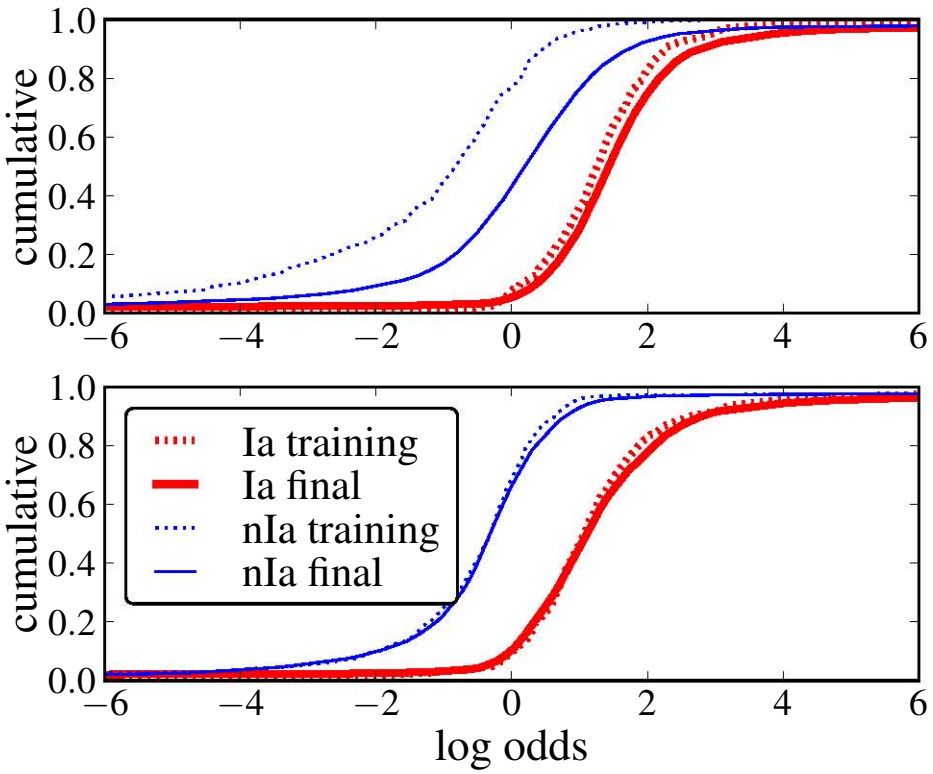

Figure 2.14: The cumulative frequency of *log odds* for non-Ia (blue) and Ia (red) SNe, for the training (dashed) and unclassified (solid) samples. The training *log odds* were calculated using tenfold cross-validation. (*Above*) Using non-representative training and (*Below*) using representative training.

of 1. In reality it is 30%. Such dangerous predictions are impossible to make if a representative sample is used in KDE construction, as illustrated by the hugging of the solid lines to the dotted lines.

2.3.2 Boosting

Application

Boosting has been used in particle physics, for example by the MiniBooNE neutrino oscillation experiment [81] and is implemented in the photometric

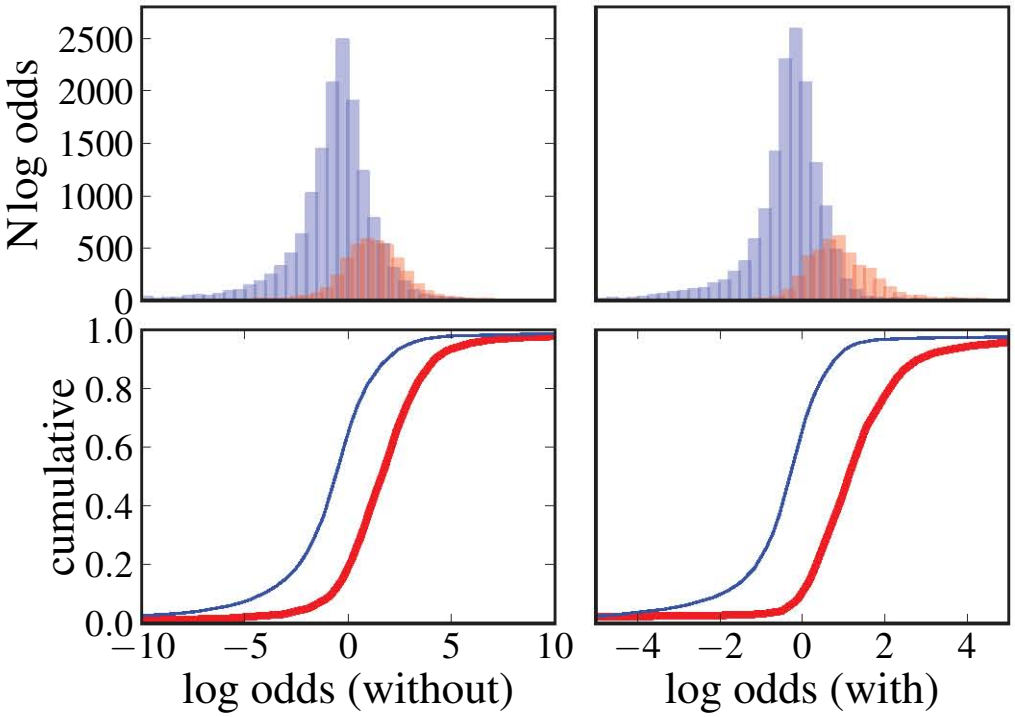

Figure 2.15: (*Above*) Histograms and (*Below*) cumulative plots of the 21D (representatively constructed) KDE *log odds*. (*Left*) The parameter covariance matrix is not included in KDE evaluation as proposed in Section 2.2.1. (*Right*) The parameter covariance matrix is included in KDE evaluation.

redshift package AborZ [82]. In the SNPCC we applied boosting to the twenty fitted lightcurve parameters for the `-HOSTZ` competition, and the twenty-one parameters for the `+HOSTZ` competition. Using tenfold cross-validation we chose to use 4000 trees to maximize the FoM (2.10). We chose the learning rate to be 0.05 and the bagging fraction to be 0.5 (these parameters are described in Appendix A5).

During the training phase of the SNPCC we expected, based on the idea that the training sample was representative, that boosting would significantly outperform the 21D KDE. In reality boosting performed more poorly than the 21D KDE, obtaining scores of 0.20 (`-HOSTZ`) and 0.25 (`+HOSTZ`) [59] strongly suggesting that the 21D KDE method is more robust to biases in

the training set than boosting.

In the case of the post-SNPCC data, the score obtained with non representative training is even lower (0.15) (`+HOSTZ`) due to bugs in the original SNPCC data such as too dim non-Ias which made classification easier, as described in the results paper[59]. As a result, comparison of scores in this chapter with those in the competition cannot be made directly.

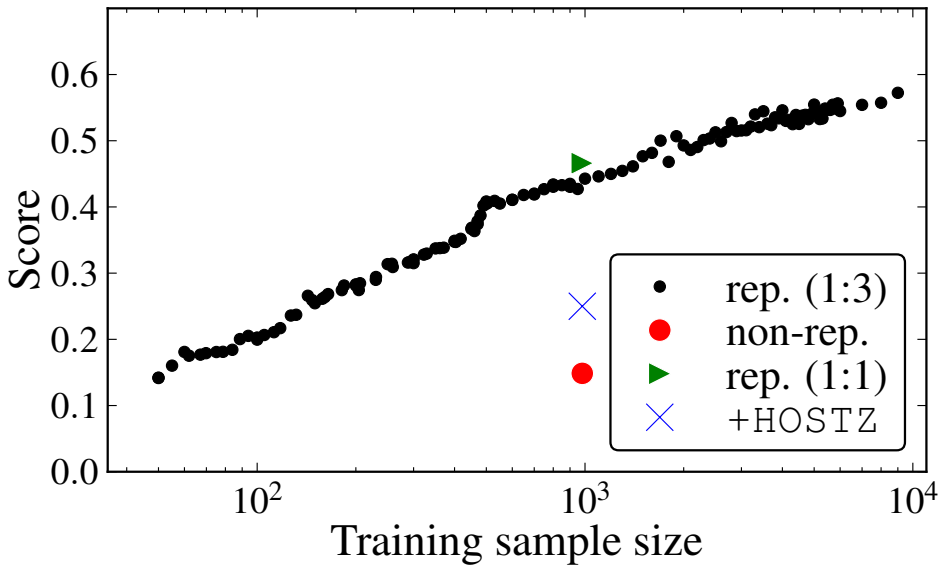

Figure 2.16: (Small black circles) The score obtained by boosting when trained with random representative samples of varying size (100 to 6000 SNe). (Large red circle) Training on the given non-representative sample. (Blue cross) The score obtained in the `+HOSTZ` competition. (Green triangle) The performance when trained with a "random" sample with non-random Ia:non-Ia ratio of 1:1 as opposed to true ratio Ia:non-Ia $\sim$ 1:3.

Non-representative vs representative

Our failure to correctly predict our score in the SNPCC was a result of the biases in the training sample. Boosting appears to be even more sensitive to training sample bias than the 21D KDE method. This is illustrated by

the large deviation in Figure 2.17 of the unclassified non-Ia curve from the training non-Ia curve with non-representative training.

While boosting is more sensitive to bias in the training sample than the 21D KDE, it is a superior classifier when a representative training sample is used. This is illustrated in Figure 2.17 by the large vertical separation between non-Ia and Ia cumulative curves when a representative sample is used. The vertical separation between the Ia and non-Ia curves is larger in the case of the boosting than the 21D KDE, resulting in a lower contamination rate and higher efficiency when boosting is used.

We see from Figure 2.16 that training with 1000 representative SNe results in a score 3 times greater than training with 1000 non-representative SNe. We also see from Figure 2.16 that training with a non-representative sample of size 1000 can be matched by training with only 50 representative SNe. The score obtained when 500 representative Ia and 500 representative non-Ia SNe are used for training, as opposed to the truly representative case where the Ia:non-Ia training ratio is 1:3, is only slightly higher; the advantage of extra Ia's at the cost of non-Ia's is marginal.

We did not include the parameter covariance matrices in any way in boosting. It is not clear how this inclusion would best be done, but the noticeable improvement to the 21D KDE score when the covariance is included suggests that it is worthwhile considering this question for future implementations. Two possibilities are a) 'supersampling' - converting each training point into 100 training points drawn from a distribution with covariance given by the parameter covariance matrix, and b) including the covariance matrix determinant as a 22nd boosting parameter.

We find that with boosting if a non-representative training sample is used the cumulative frequency lines of the unclassified SNe do not follow those of the training sample. On the other hand if a representative sample is used, tenfold cross-validation provides accurate predictions for the unclassified SNe boosting values, as illustrated by the close hugging of training and unclassified cumulative lines in Figure 2.17.

We see that boosting the 21D lightcurve parameters with a representative sample results in a robust photometric classifier. To illustrate this point we

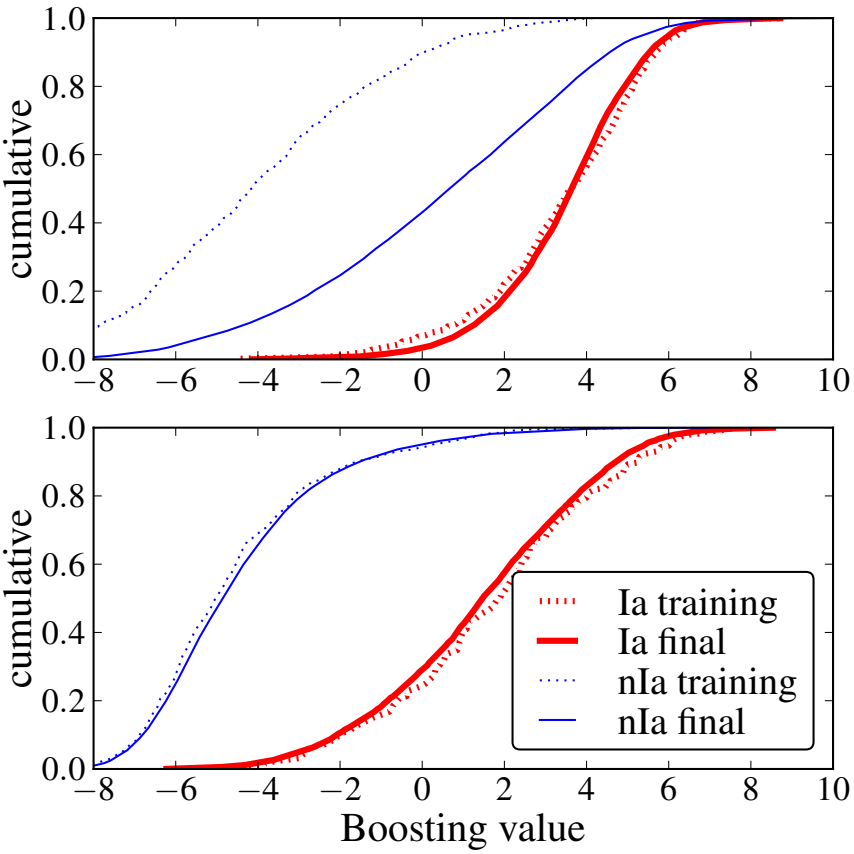

Figure 2.17: Boosting values obtained using (*Above*) the non-representative training sample and (*Below*) the representative training sample. The boosting values are approximations to $\frac{1}{2}$ *log odds*.

have created on online archive of 200 randomly selected unclassified SNe, and labeled them according to boosting's output [65]. In some cases it is difficult to identify obvious Ia or non-Ia features, yet the algorithm classifies correctly.

2.3.3 Parameter importance

One advantage of the boosting algorithm is its ability to quantify the importance of parameters for classification (see Appendix A4 for details). In this section we consider this special property of boosting in an effort to discover which fitted parameters are most useful for classification. We also ask which are the parameters that distinguish the non-representative training sample from the representative training sample, i.e. what makes the non-representative Ia's and non-Ia's a biased sample. We answer this question by performing boosting on a sample of representative and non-representative Ia's, as if the SNPCC had been a competition to determine if a SN attains a spectrum or not.

Figure 2.19 illustrates which parameters are most useful in distinguishing Ia from non-Ia in the representative training sample. One interesting feature illustrated in Figure 2.19 is that every parameter appears to carry information.

The third most important parameter (after redshift and and A in z-band) is the parameter k in the i-band. To interpret this piece of information, we first see in Figure A.7 that non-Ia SNe have on average lower k values than Ia's. From this we then infer from Figure 2.8 that Ia's have a higher rise-time to decay-time ratio than non-Ia SNe.

The equivalent figure for the non-representative training (Figure A.2 in Appendix A4 paints a similar picture with one noticeable difference: The information for distinguishing between Ia and non-Ia SNe in the non-representative training sample is carried almost exclusively in the r-band.

We now turn to the comparison of representative and non-representative SNe. Figure 2.20 suggests that the most biased parameter in the non-representative training sample is redshift. This is not surprising given that we know that the non-representative SNeIa are at lower redshift than the true Ia population (Figure 2.18). Indeed, we see from Figure 2.18 70% of Ia SNe in the non-representative training set are at a redshift of less than 0.6, while only 20% of Ias in the unclassified set are within this redshift.

In the case of non-Ia SNe (Figure A.3 in Appendix A4) boosting allocates

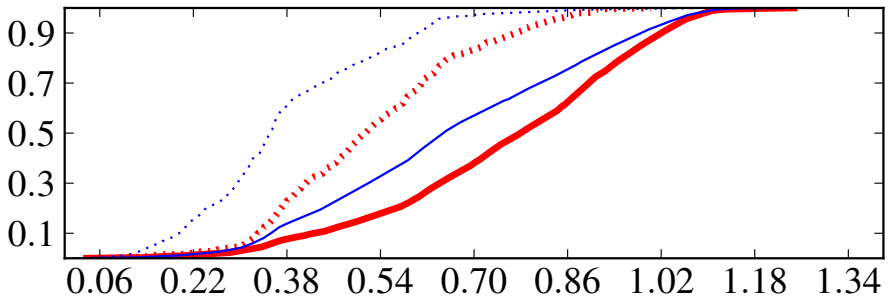

Figure 2.18: Cumulative plot of redshift, non-representative training (dashed) vs unclassified (solid) and Ia (red, thick) vs non-Ia (blue, thin).

the majority of the bias in the non-representative sample to the A's. This is also unsurprising given that we are more likely to obtain a spectrum from bright objects than dim objects. It is not clear to us why boosting designates non-Ia bias to the A's and Ia bias to redshift.

2.3.4 Hubble KDE

Applications

An alternative method for using the idea of KDEs is to use the SALT2 lightcurve fitter (with $\alpha = 0.1$ and $\beta = 2.77$ as in Hicken et al.[67]) to estimate distance moduli, μ_i and errors σ_i for all the objects in both the training and unclassified data, assuming that all the data are SNeIa. We can then construct two 2D KDEs for the training data: one consisting of all the known SNeIa and one from all the non-Ia data. Each kernel is normalised to have a total volume of unity and we use a slight modification of the standard KDE formalism because we do not normalise the KDE. Instead the heights of the summed KDEs are proportional to the number of SNeIa and non-Ia respectively. In this way we include prior information related to the SN rates. A redshift range where there are many more SNeIa than non-Ia will automatically tend to lead to a larger Ia KDE as a result.

The 2D Gaussian kernel chosen for the Hubble KDE algorithm had a

fixed bandwidth (standard deviation) in the redshift direction of 0.05 (chosen simply to avoid being too peaked or smooth) while the bandwidth in the μ direction was determined by the error σ_i in the distance modulus coming from SALT2 along with the 0.12 mag intrinsic dispersion error as mentioned already. This variation means that points with large errors contribute very stretched out, low amplitude humps to the final KDE, while points with small errors are much more peaked, reflecting our confidence in that point. For illustrative purposes we plot the difference $\mathrm{KDE}_{Ia} - \mathrm{KDE}_{non-Ia}$ of the two KDEs in Figure 2.21. Positive values correspond to places where the Ia KDE dominates, negative values to where the non-Ia KDE dominates. In addition we plot the training data used to construct the KDEs.

Classification using these KDEs is then simple. For any candidate object, we run it through SALT2 to give an estimated μ and σ. We can only use this approach on the data with a redshift estimate, z, unlike the 20D KDE and boosting algorithms which do not require a redshift. We then simply find the values of the two KDEs at that (μ, z) to yield probabilities of the object being a Ia or non-Ia. As in the other KDE method, one should fold in the error σ on the candidates which, assuming Gaussianity, is simple, as described in Section 2.2.1. The result of this analysis is that each candidate has a pair of probabilities: (P_{Ia}, P_{non-Ia}) that can be used to classify.

Non-representative training sample

We applied this method to the whole sample of unknown SNe supplied. In total, we lost 4578 SNe as junk because of SALT2 failures previously mentioned (of which 2619 were complete failures and 1959 failed to return meaningful parameters from the Ia lightcurve fit), leaving 12487 SNe for further analysis.

Essentially this Hubble KDE approach simply checks whether or not an object lies close to the true cosmology curve on the Hubble diagram (defined by the Ia KDE) at that redshift. This may seem circular, as it is the cosmology which we are wanting to calculate from SNeIa, but in fact there is no assumed cosmology using this approach and therefore no inbuilt bias to a particular set of cosmological parameters. There are many non-Ia's which

lie close to the true cosmology curve, and as a result one either has to be very strict with cuts (and therefore lose many true Ia's) or one has to accept a large number of false positives: non-Ia's that are classified as Ia's.

Because there are so many non-Ia's this, and similar Hubble-diagram based methods (such as the Portsmouth entry to the SNPCC), are less competitive as classifiers. In addition they also require a redshift estimate for the SNe and are hence doubly inferior compared with the 21D KDE and boosting.

2.3.5 Combining 21D and Hubble KDEs

In Section 2.3.1 we described the 21D KDE approach, and in Section 2.3.4 we described the Hubble KDE approach. In this section, we describe how one combines these approaches. As outlined in Appendix A2, there are several ways of combining *odds* from different algorithms to construct a better combined classifier. For our combination entries in the SNPCC, we constrained our classifier to be of the form:

$$(\text{Hubble odds})^{\alpha} \cdot (\text{21D odds})^{\beta} > \eta. \tag{2.11}$$

This corresponds to a straight line in Figure 2.22. The scores for the combination entry was 0.28. Surprisingly, this was less than the score obtained using the 21D KDE alone, and so we believe that the line chosen for the SNPCC was poor. A straight line does seem to be a good choice for the distribution of values in Figure 2.22, but perhaps a better choice would be of the form:

$$\text{Hubble odds} > \gamma_1 \text{ and } \text{21D odds} > \gamma_2. \tag{2.12}$$

A pure 21D *odds* classifier would rely on a vertical decision line, and a pure Hubble *odds* classifier would rely on a horizontal line, but it is clear from Figure 2.22 that a classifier of the form 2.11 (dashed) or 2.12 (solid) should work better. Figure 2.22 shows the separation of Ia's and non-Ia's that come from using the Hubble KDE *odds* and 21D KDE *odds* with the integration of errors presented in Section 2.2.1. The optimal lines of forms (2.11) and (2.12)

result in scores of 0.24 and 0.22 respectively in the case of non-representative training and 0.45 and 0.42 respectively in the case of representative training. These scores are calculated using a purely 21D *odds* classification for the $\sim$ 8000 SNe without SALT2 fits, and a 21D-Hubble combination for the remaining $\sim$ 12500 SNe with SALT2 fits. As with boosting, the 21D KDE classifier is significantly worse using the post-SNPCC data as previously discussed in section 2.3.2, and so comparison between these post-SNPCC scores and other SNPCC scores should not be made until further analyses have been done.

To be in the top-right corner of Figure 2.22, and therefore be classified as Ia, requires that a candidate must simultaneously lie close to the true cosmology distance modulus and have multiband lightcurves that have the right shape; a very natural approach to SNIa classification. It would be interesting to combine the Hubble *odds* with 21D boosting instead of 21D KDE, as boosting the twenty parameters produces better results, as seen in Section 2.3.2.

An obvious extension, if one wanted to combine the outputs from more than two classifiers, would be to use them as inputs to a new boosting analysis. The *odds* from the 21D KDE, the Hubble KDE, the 21D boosting, and indeed any classifier of sufficient ability can be used as weak classifiers in boosting. A reason to exercise caution in using boosting or a neural network as a final classifier in this way is the possibility of overtraining, but this can be prevented by using tenfold cross-validation.

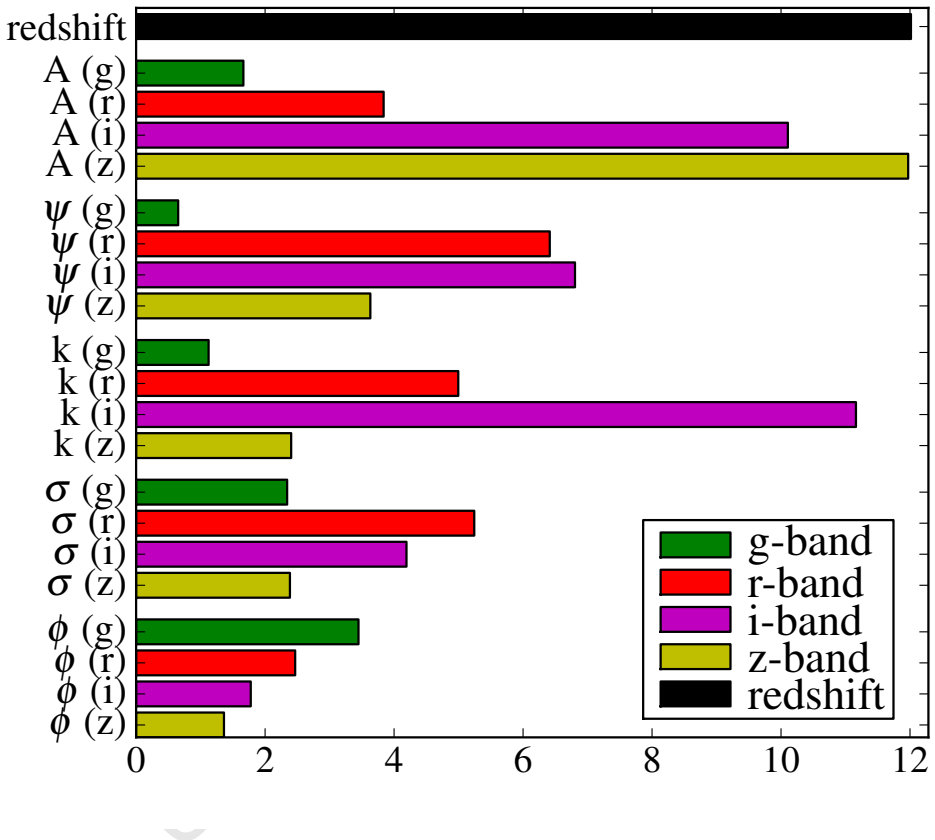

Figure 2.19: The importance of each of the 21 parameters in classifying SNe as Ia (or not) using boosting on the representative training sample. The precise value on the x-axis is calculated according to how variables appear in the trees, as described in Appendix A4.

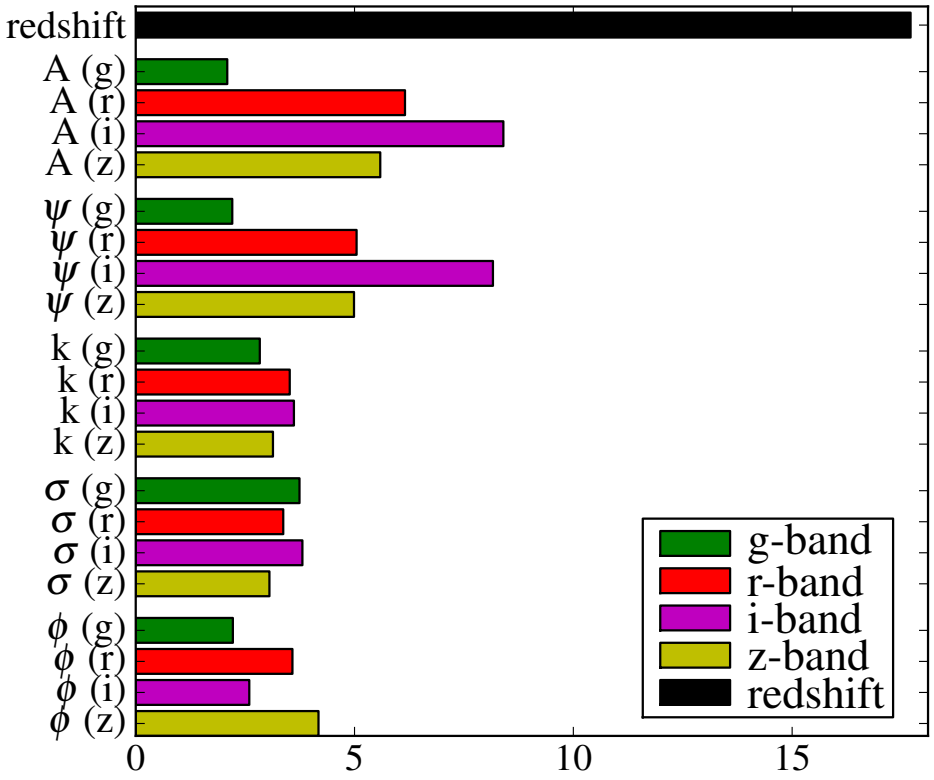

Figure 2.20: The importance of parameters in distinguishing representative from non-representative SNeIa using boosting.

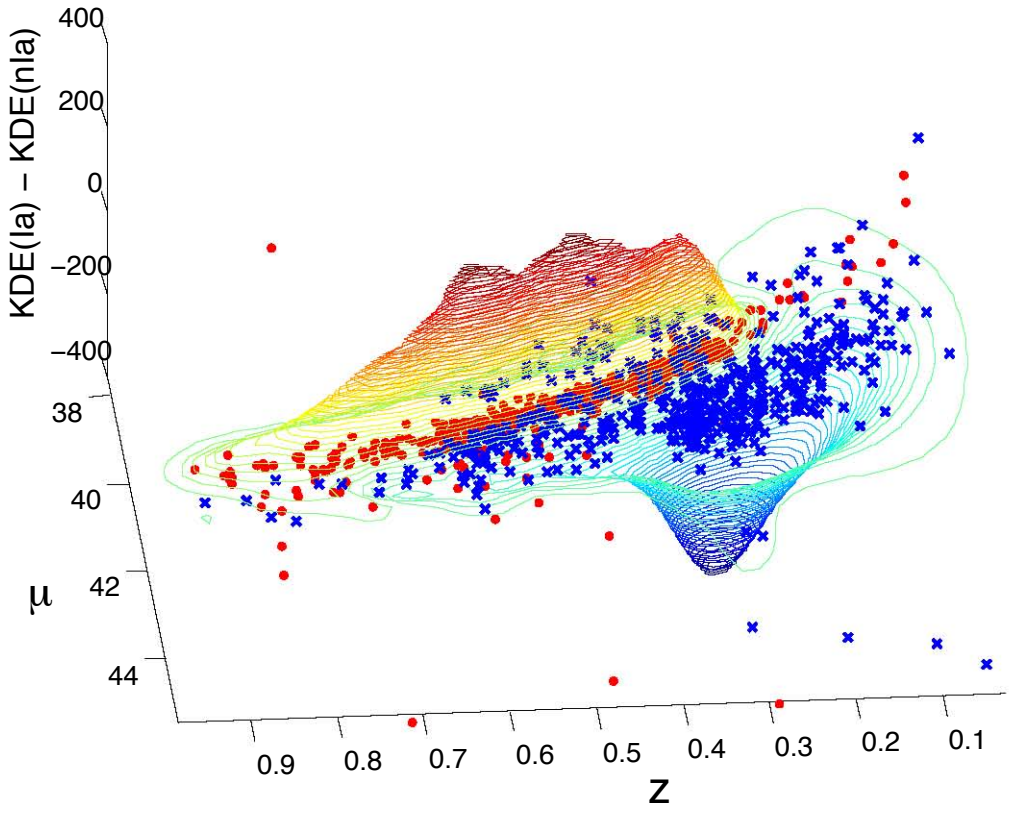

Figure 2.21: 3D contours of the difference between the Ia and non-Ia Hubble diagram KDEs as a function of redshift and distance modulus (μ) together with the actual non-representative training data used to produce the KDEs. The data used to construct the KDEs are also shown: Ia data as red circles and non-Ia data as blue crosses. There is a clear offset in the two KDEs reflecting the fact that in this training data the non-Ia's are fainter, hence predominantly at lower redshift and with a much larger scatter than the Ia's.

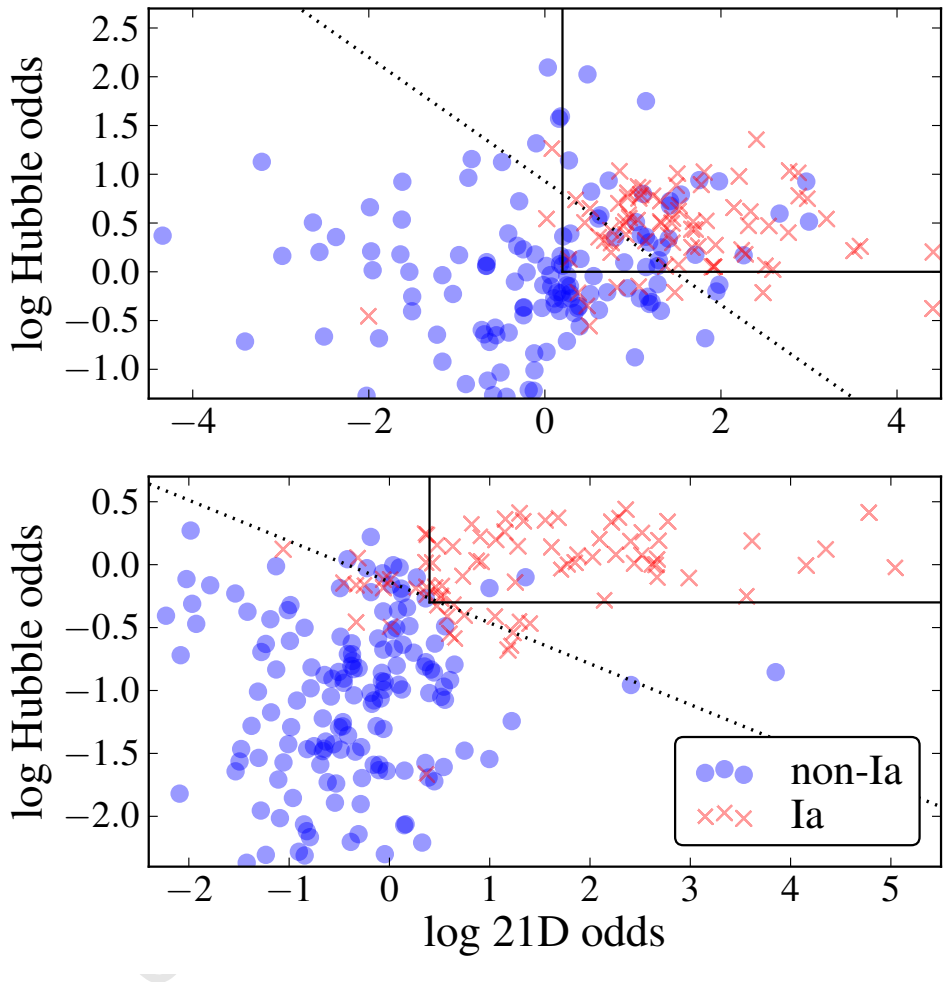

Figure 2.22: SNe of type Ia (cross) and non-Ia (circle), located according to their 21D *odds* (x-axis) and Hubble *odds* (y-axis). (*Above*) *Odds* were calculated from KDEs constructed using the non-representative training sample. (*Below*) A corresponding plot where KDEs were constructed with the representative training sample. We see that the separation obtained is smaller when non-representative training is used, and indeed the score obtained in the non-representative case is significantly lower. Note that the SNe in this figure are a random sample of the $\sim$ 12500 with a meaningful SALT2 fit.

2.4 Dealing with bias

We include this section here for completeness, as it was published in the same paper as the rest of Chapter 2. However, in the process of writing Section 3.7 we realised that much of what was previously written in the section was wrong, and has therefore been removed!

We see in Figure 2.16 that representative training samples with more than 50 objects outperforms the 1000 strong non-representative training sample. In light of this astonishing fact we ask: how large a representative sample can be extracted from the non-representative sample? The correct answer to this question is a sample of size zero, as there are no spectroscopically confirmed SNe as dim as the dimmest unconfirmed SNe, and therefore it is impossible to extract a representative sample for the unconfirmed SNe. A correct treatment of this topic will be presented in the following chapter in Section 3.7.

2.5 Discussion and Conclusions

In this chapter we have discussed the problem of classifying SNe into sub-classes (Type Ia or non-Ia) based on photometric lightcurve data alone, that is, multi-band fluxes as a function of time. This classification will be necessary for future surveys which will detect vastly more candidates than will be possible to follow up spectroscopically.

We have investigated two novel classes of classification algorithms, Kernel Density Estimation (KDE) and boosting, and applied them to simulated SNe lightcurve data, finding that the methods performed impressively as long as they were trained on a representative sample. Using the KDE approach, we considered both a 21 dimensional case based on lightcurve parameters from all bands and a 2 dimensional version based on fits to the Hubble diagram, using redshift information and an estimate of the distance modulus obtained using standard lightcurve fitting software.

A key issue for the classification methods we used was the issue of the training data sets. We compared the results based on training on two very

different data sets: the first, a non-representative set, mimicking the kind of spectroscopic sample available as part of the follow-up program of a typical current-generation SN survey. The second was a representative sample of the same size where training objects were selected at random from the full sample.

In general we found that the training on the representative sample produced exceptionally good results and that cross-validation on the training sample was able to accurately predict the purity and efficiency of the method on the full sample. On the other hand, training on the non-representative sample lead to relatively poor performance on the full data set. The importance of having an unbiased, representative sample is illustrated by the fact that for boosting, representative samples larger than about 50 objects outperformed the full non-representative sample of 1000 objects, as shown in Figure 2.16.

Our primary result and recommendation therefore is that boosting and KDE are powerful methods for SN classification, with remarkably little astrophysical input. However, they require training samples that are as unbiased and representative as possible. Further, we found that a small unbiased training sample outperforms a much larger, but biased, training sample.

Our other main result is that neither boosting nor the 21D KDE method suffered particularly when the SN redshift information was unavailable. This is particularly gratifying given that accurate SN/host galaxy redshifts will not be available for most candidates in the future and that methods based on the Hubble diagram critically require redshift information to perform successfully.

While the algorithms we have presented were successful, there are modifications to our boosting implementation that should be experimented with, for example different choices of lightcurve parameterisation. Finally it is perhaps useful to comment on how our methods compare to the winner of the SNPCC (the methods we described in this chapter finished second and third in the competition) which used a template-based method and performed very well. Our first comment is that comparison is hard because there was an overlap between the templates used by Sako and those used to generate

the SNPCC, as described in the results paper [59], so it is not clear how the method would perform on completely independent data. Secondly, it is not known how the various classification methods would perform with different Figures of Merit. For cosmological applications one might prefer to use a Figure of Merit which looks at the bias in recovered cosmological parameters such as the w_0, w_a dark energy equation of state parameters. Investigating this important issue is left to future work. It is clear that finding the best approach to SN classification, and the best way to combine results from different classifiers, will be an active area of research in the coming decade.

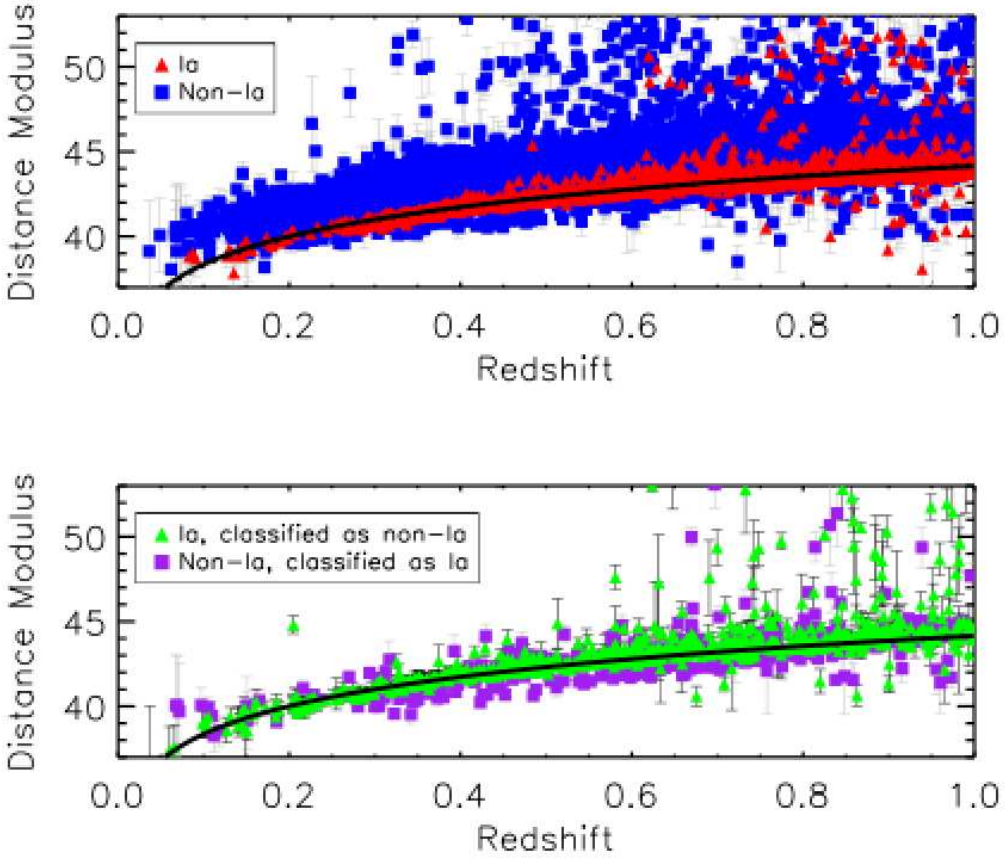

Figure 2.23: Hubble diagrams for the boosting results using the representative training sample. (*Above*) Objects that were correctly identified by the boosting method. SNeIa are plotted as red triangles, with non-Ia SNe shown as blue squares. (*Below*) SNe that were incorrectly typed by boosting. SNeIa that were considered to be non-Ia SNe by boosting are shown as green triangles, with incorrectly typed non-Ia SNe shown as purple squares. Overplotted on each graph is the best fitting cosmological model inferred from the representative training sample.

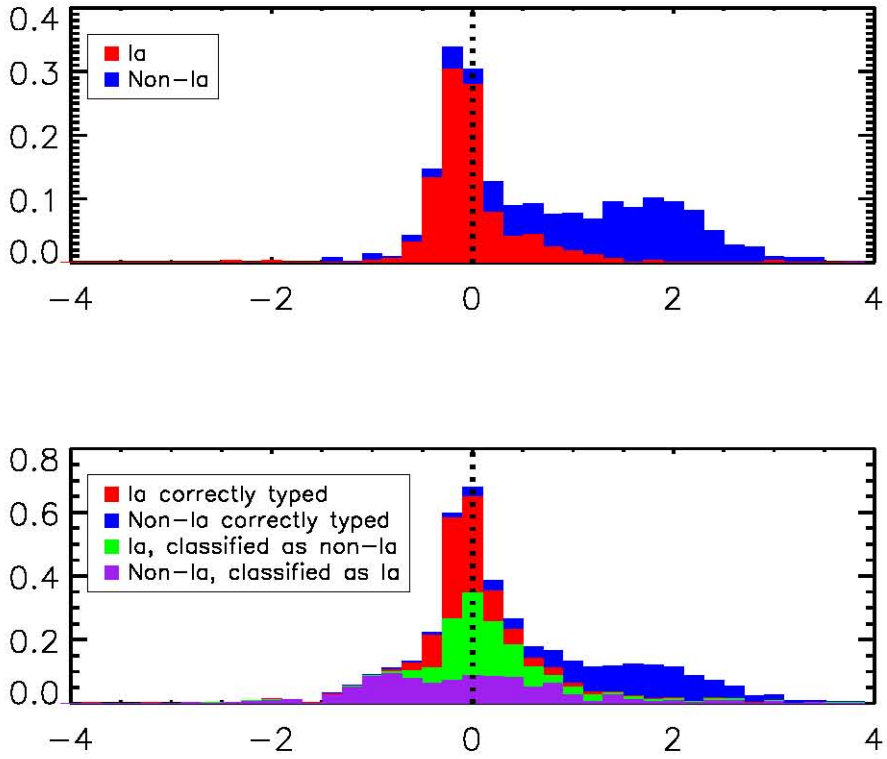

Figure 2.24: Cumulative histograms of the residuals from the best-fit Hubble diagram, determined using the SNeIa in the representative training sample. (*Above*) Residuals for the representative training sample. SNeIa are plotted in blue, with non-Ia SNe shown in red. (*Below*) Residuals for the boosting results. SNeIa that were correctly typed are shown in red, with correctly typed non-Ia SNe shown in blue. SNeIa that were considered to be non-Ia SNe by boosting are shown in green, with incorrectly typed non-Ia SNe shown in purple.

Chapter 3

BEAMS and Debiasing

As we have mentioned, there are two ways that one can imagine using photometric candidates for estimating cosmological parameters. The first approach is to try to classify the candidates into Ia, Ibc or II SNe [83, 84, 85, 86] and then use only those objects that are believed to be SNeIa above some threshold of confidence. This has recently been discussed by [87] who showed that photometric cuts could achieve high purity. Nevertheless it is clear that this approach can still lead to biases and systematic errors from the small contaminating group when used in conjunction with the simplest parameter estimation approaches such as the maximum likelihood method. A second approach is to use all the SNe, irrespective of how likely they are to actually be a SNIa. This approach is exemplified by the BEAMS formalism, which accounts for the contamination from non-Ia SN data using the appropriate Bayesian framework, as presented in [60], hereafter referred to as KBH.

In KBH, two threads are woven: a general statistical framework, and a discussion of how it may be applied to SNeIa. As noted in KBH, the general framework can be applied to any parameter estimation problem involving several populations, and indeed may have already been done so in other fields. In this chapter we take the same approach as in KBH of keeping the notation general enough for application to other problems, while discussing its relevance to SNe.

We will attempt to use the same notation as in KBH, but differ where

we consider it necessary. For example, we write conditional probability densities as $f_{\Theta|D}(\theta|d)$. The quantity $f_{\Theta|D}(\theta|d)\Delta\theta$, should be interpreted as the probability that Θ lies in the interval $(\theta, \theta + \Delta\theta)$, conditional on $D = d$ (for small $\Delta\theta$).

We preserve capital letters for random variables and lowercase letters for their observed values. In the BEAMS framework, one wishes to estimate parameter(s) Θ from N observations of the random variable X. We will use the boldface $\boldsymbol{X}$ to denote a vector of N such random variables: $\boldsymbol{X} = X_{1\cdots N}$. An observation of X we will denote by x, so that the full set of N observations is denoted by $\boldsymbol{x} = x_{1\cdots N}$. For SNe, the observations $\boldsymbol{x}$ are the photometric data of the N SNe. As such, for SNe the probability density function (pdf) $f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)$ is the likelihood of observing the photometric data $\boldsymbol{x}$ assuming some cosmological parameters θ, which we will discuss. The relationship between raw photometric data (X) and the true cosmological parameters (Θ) is highly intricate, resulting in a pdf which cannot realistically be worked with, and so one first reduces each observation x to a single feature d for which there is a direct Θ-dependent model. For SNe, if the parameters Θ are for example Ω_Λ and Ω_m, then d will consist of an estimated luminosity distance and redshift. If the parameter of interest Θ is a luminosity distance at a given redshift, then d will be simply a fitted distance modulus. Unless stated otherwise, this is the case.

The correct treatment of redshifts will be important to BEAMS as applied to future SN surveys. Future surveys will likely have only photometric information for the SNe but will have a spectroscopic redshift for the host galaxy obtained by chance (because of overlap with existing surveys) or through a targeted follow-up program. The SDSS-II supernova survey [88] is an example of both of these. There were host redshifts available from the main SDSS galaxy sample and there was also a targeted host followup program as part of the BOSS survey. Future large galaxy surveys like SKA, EUCLID or BigBOSS will likely provide a very large number of host galaxy redshifts for free.

BEAMS is unique in that the underlying types of the observations are not assumed known. In the case where there are two underlying types ($T \in$

$\{A, B\}$), each observation has an associated type probability (P) of being type A,

$$P \stackrel{\text{def}}{=} \mathrm{P}(T = A|X_P),$$

where X_P is a subset of features of X. In other words, X_P is the component of the raw data X on which type probabilities are conditional. Note that we treat P as a random variable: while the value of P is completely determined by X_P, which in turn is completely determined by X, X is a random variable and therefore so too is P. The realizations of the type probabilities $\boldsymbol{P}$ of the N observations are denoted by $\boldsymbol{p} = p_{1\cdots N}$, and we will call them τ_A-probabilities. The τ_A-probability for a SN is thus the probability of being type Ia, conditional on knowing the subset x_P of the the photometric data. x_P may be the full photometric time-series, the earliest segment of the SN's light curve, a fitted shape parameter, or any other extracted photometric information. Finally, we mention that the type of the SN (T) is a random variable with realisation denoted τ. A summary of all the variables used in the chapter is given in Table 3.

Attempts to approximate τ_A-probabilities include those of [54, 43, 89] and as implemented in SALT2 [66]. Note that values obtained using these methods are only approximations of τ_A-probabilities, as the algorithms are trained on only a handful of spectroscopically confirmed SNe. Note too that there is no sense in which one set of τ_A-probabilities is *the* correct set, as 'correct' depends on what X_P is. Obtaining unbiased estimates of τ_A-probabilities is not easy, and we will consider the problems faced in doing so in Section 3.6. For SNe, the problem is made especially difficult by the fact that spectroscopically confirmed SNe, which are used to train τ_A-probability estimating algorithms, are brighter than unconfirmed photometric SNe.

In 2009 the Supernova Photometric Classification Challenge (SNPCC) was run to encourage work on SN classification by lightcurves alone [58]. Performance of the classification algorithms was judged according to the final purity and efficiency of extracted Ia samples. While the processing of photometric data is essential to the workings of BEAMS for SNe, the classification of objects is not required. It would be interesting to hold another

competition where entrants are required to calculate τ_A-probabilities for SNe. Algorithms would then not only need to recognise SNeIa, but would also need to provide precise, unbiased probabilities of the object being an SNeIa.

In brief, this chapter consists of three more or less independent parts. In Section 3.1 we present an extension of BEAMS to the case where particular correlations, which were ignored in KBH, are present. In Section 3.2, we discuss the relevance of τ_A-probabilities in a broader context, and specifically the importance to of them in BEAMS. Then is Sections 3.3, 3.4 and 3.5, we perform simulations to better understand the importance of sample sizes, nearness of population distributions, biases of τ_A-probabilities and decisivenesses of τ_A-probabilities (to be defined). Finally, in Section 3.6 we present new ideas from the machine learning literature describing when and how τ_A-probability biases emerge and how to correct for them. These ideas are then discussed in the context of the SNPCC in Section 3.7.

Random Variables		
R.V.	Data	Definition
P	p	The probability of being type A conditional on X_P. We call P the τ_A-probability.
D	d	A particular feature of an object whose distribution depends directly on the parameter(s) we wish to approximate using BEAMS. SNe: D is luminosity distance.
T	τ	The type of an object, $T \in \{A, B\}$ SNe: $T \in$ {Ia, nIa}
X	x	All the features observed of an object. SNe: X is the photometric data.
X_F	x_F	That part of the features which affects confirmation probability. SNe: X_F are peak apparent magnitudes.
X_P	x_P	That part of the features used to determine the τ_A-probability. SNe: X_P could be any reduction of X.
F	f	Whether the object is confirmed or not. For SNe: $F = 1$ if a spectroscopic confirmation is performed.
$\bar{P}$	$\bar{p}$	Is exactly P if the object is unconfirmed and 1 or 0 if confirmed, depending on type.

3.1 Introducing and Modifying the Beams equations

The posterior probability on the parameter(s) Θ, given the data $\boldsymbol{D}$, is derived in Section II of KBH as

$$f_{\Theta|\boldsymbol{D}}(\theta|\boldsymbol{d}) \propto f_{\Theta}(\theta) \quad \times \sum_{\boldsymbol{\tau}\in[A,B]^N} f_{\boldsymbol{D}|\Theta,\boldsymbol{T}}(\boldsymbol{d}|\theta,\boldsymbol{\tau}) \prod_{\tau_i=A} p_i \prod_{\tau_j=B} (1-p_j), \qquad (3.1)$$

where the p_is are τ_A-probabilities. The summation is over all of the 2^N possible ways that the N observations can be classified into two classes. We will refer to the expression on the right of (3.1) as the original posterior. When the N observations are assumed to be independent, that is when

$$f_{\boldsymbol{D}|\Theta,\boldsymbol{T}}(\boldsymbol{d}|\theta,\boldsymbol{\tau}) = \prod_{i=1}^{N} f_{D_i|\Theta,T_i}(d_i|\theta,\tau_i),$$

the original posterior reduces to,

$$\prod_{i=1}^{N} \left[f_{D_i|\Theta,T_i}(d_i|\theta,A)\, p_i + f_{D_i|\Theta,T_i}(d_i|\theta,B)\,(1-p_i) \right]. \tag{3.2}$$

There is one substitution in the derivation of the original posterior on which we would like to focus, given in KBH as eqn. (5) on page 3:

$$f_{\boldsymbol{T}}(\boldsymbol{\tau}) = \prod_{\tau_i=A} p_i \prod_{\tau_i=B} (1-p_i)\,. \tag{3.3}$$

Equation (3.3) states that the l.h.s. prior probability of the SNe having types $\boldsymbol{\tau}$ is given by the product on the r.h.s. involving τ_A-probabilities. We argue that this product should not be treated as the prior $f_{\boldsymbol{T}}$, but rather as the conditional $f_{\boldsymbol{T}|\boldsymbol{P}}$. In effect, we argue that KBH should not use the τ_A-probabilities $\boldsymbol{p}$ unless $\boldsymbol{P}$ is explicitly included as a conditional parameter. It is to this end that we now rederive the posterior on Θ, taking $f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p})$ as a starting point, discussing at each line what has been used.

$f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p})$

$\rightarrow$ We will first use the definiton of conditional probability to obtain,

$$= \frac{f_{\Theta,\boldsymbol{D},\boldsymbol{P}}(\theta,\boldsymbol{d},\boldsymbol{p})}{f_{\boldsymbol{D},\boldsymbol{P}}(\boldsymbol{d},\boldsymbol{p})}.$$

→ The term in the numerator can then be written as the sum over all 2^N possible type vectors,

$$= \sum_{\boldsymbol{\tau}} \frac{f_{\Theta,\boldsymbol{D},\boldsymbol{P},\boldsymbol{T}}(\theta, \boldsymbol{d}, \boldsymbol{p}, \boldsymbol{\tau})}{f_{\boldsymbol{D},\boldsymbol{P}}(\boldsymbol{d}, \boldsymbol{p})}.$$

→ The numerator is again modified using the definition of conditional probability,

$$= \sum_{\boldsymbol{\tau}} \frac{f_{\boldsymbol{D}|\Theta,\boldsymbol{P},\boldsymbol{T}}(\boldsymbol{d}|\theta, \boldsymbol{p}, \boldsymbol{\tau}) f_{\boldsymbol{\Theta},\boldsymbol{P},\boldsymbol{T}}(\theta, \boldsymbol{p}, \boldsymbol{\tau})}{f_{\boldsymbol{D},\boldsymbol{P}}(\boldsymbol{d}, \boldsymbol{p})}.$$

→ We will now assume that the probability of having τ_A-probabilities and types $\boldsymbol{p}$ and $\boldsymbol{\tau}$ respectively are independent of Θ. As noted following eqn.(4) in KBH, for SNe this assumption rests on the fact that Θ (that is Ω_m, Ω_Λ) describes large scale evolution, while the SN types $\boldsymbol{\tau}$ depend on local gastrophysics, with little or no dependence on perturbations in dark matter.

$$= \sum_{\boldsymbol{\tau}} \frac{f_{\boldsymbol{D}|\Theta,\boldsymbol{P},\boldsymbol{T}}(\boldsymbol{d}|\theta, \boldsymbol{p}, \boldsymbol{\tau}) f_{\Theta}(\theta) f_{\boldsymbol{P},\boldsymbol{T}}(\boldsymbol{p}, \boldsymbol{\tau})}{f_{\boldsymbol{D},\boldsymbol{P}}(\boldsymbol{d}, \boldsymbol{p})}.$$

→ Rearranging this, and again using the definition of conditional probability, we obtain,

$$= \frac{f_{\boldsymbol{P}}(\boldsymbol{p})}{f_{\boldsymbol{D},\boldsymbol{P}}(\boldsymbol{d}, \boldsymbol{p})} f_{\Theta}(\theta) \sum_{\boldsymbol{\tau}} f_{\boldsymbol{D}|\Theta,\boldsymbol{P},\boldsymbol{T}}(\boldsymbol{d}|\theta, \boldsymbol{p}, \boldsymbol{\tau}) f_{\boldsymbol{T}|\boldsymbol{P}}(\boldsymbol{\tau}|\boldsymbol{p}).$$

→ The first term on the line above is constant with respect to Θ, and so is absorbed into a proportionality constant. We now make one final weak assumption: $f_{\boldsymbol{T}|\boldsymbol{P}}(\boldsymbol{\tau}|\boldsymbol{p}) = \prod_{i=1}^{N} f_{T_i|P_i}(\tau_i|p_i)$. This assumption will be necessary to make a comparison with the original posterior. Making this assumption we arrive at,

$$\propto f_{\Theta}(\theta) \sum_{\boldsymbol{\tau}} f_{\boldsymbol{D}|\Theta,\boldsymbol{P},\boldsymbol{T}}(\boldsymbol{d}|\theta, \boldsymbol{p}, \boldsymbol{\tau}) \prod_{\tau_i=A} p_i \prod_{\tau_j=B} (1 - p_j).$$

(3.4)

We will refer to the newly derived expression (3.4) as the full posterior. Let us now consider the difference between the original posterior (3.1) and the full posterior, and notice that in the full posterior, the likelihood of the data $\boldsymbol{D}$ is conditional on $\Theta, \boldsymbol{P}$ and $\boldsymbol{T}$, whereas in the original posterior $\boldsymbol{D}$ is only conditional on Θ and $\boldsymbol{T}$. This difference in conditional variables is the only difference between the two posteriors, and so when $\boldsymbol{D}|\Theta, \boldsymbol{T}$ is independent of $\boldsymbol{P}$, the posterior (3.4) reduces to the original posterior (3.1), making them equivalent. This is an important result: when $\boldsymbol{D}|\Theta, \boldsymbol{T}$ and $\boldsymbol{P}$ are independent, the original and full posteriors are the same. Our results can be summarised as follows,

(1) As the posterior $f_{\Theta|\boldsymbol{D}}(\theta|\boldsymbol{d})$ is not conditional on τ_A-probabilities it should be independent of τ_A-probabilities, and we thus prefer to replace the original posterior in (3.1) by

$$f_{\Theta|\boldsymbol{D}}(\theta|\boldsymbol{d}) \propto f_{\Theta}(\theta) \quad \times \sum_{\boldsymbol{\tau}\in[A,B]^N} f_{\boldsymbol{D}|\Theta,\boldsymbol{T}}(\boldsymbol{d}|\theta,\boldsymbol{\tau}) \prod_{\tau_i=A} \pi \prod_{\tau_j=B} (1-\pi),$$

where π is an estimate of the global proportion of type A objects.

(2) $f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p})$ is always given by the full posterior (3.4). When $\boldsymbol{D}|\Theta, \boldsymbol{T}$ and $\boldsymbol{P}$ are independent, it reduces to the original posterior (3.1).

It is worth discussing for SNe the statement, "$\boldsymbol{D}|\Theta, \boldsymbol{T}$ and $\boldsymbol{P}$ are not independent". One incorrect interpretation of this statement is, "given that we know the cosmology is Θ, observing[1] $\boldsymbol{P}$ for a SN of unknown type adds no information to the estimation of the distance modulus." Indeed it is difficult

[1]Of course we mean "observing" in the statistical sense, that is obtaining the realistation of the τ_A-probability (p) with some software

to imagine how this could be the case: we know that SNeIa are brighter that other SNe, and therefore obtaining a τ_A-probability close to 1 shifts the estimated distance modulus downwards (towards being brighter).

A correct interpretation of the statement is, "given the cosmology Θ, observing $\boldsymbol{P}$ of a SN of known type adds no information to the estimation of the distance modulus." It may seem necessarily true that a τ_A-probability contributes no new information if the type of the SN is already known, but this is not in general the case; it depends on the method by which τ_A-probabilities are obtained.

Currently for SNe, fitted distance moduli and approximations of τ_A-probabilities are frequently obtained simultaneously, using for example SALT2 [66]. This simultaneity in itself suggests that $\boldsymbol{D}|\Theta, \boldsymbol{T}$ and $\boldsymbol{P}$ will not be independent. In some cases however, τ_A-probabilities are calculated from the early stages of the light curves [90, 91] while the distance modulus is estimated from the peak of the light curve, and so the dependence may be weak. As another example, in Section 4.4 of [43] τ_A-probabilities are obtained directly from a Hubble diagram. Objects lying in regions of high relative SNIa density are given higher τ_A-probabilities than objects lying in low relative SNIa density. As a result, at a given redshift, brighter nIa SNe have higher τ_A-probabilities than faint nIa SNe. Similarly, at a given fitted distance modulus (fitted assuming type Ia), nIa will lie on average at lower redshifts than Ia. Both of these cases, (distance modulus | Θ, type) being correlated with P, and (redshift | Θ, type) being correlated with P, are precisely when $\boldsymbol{D}|\Theta, \boldsymbol{T}$ and $\boldsymbol{P}$ are dependent. In Section 3.5 a simulation illustrating this dependence is presented. For completeness, we mention that in the case of independent observations, that is when,

$$f_{\boldsymbol{D}|\Theta,\boldsymbol{P},\boldsymbol{T}}(\boldsymbol{d}|\theta, \boldsymbol{p}, \boldsymbol{\tau}) = \prod_{i=1}^{N} f_{D_i|\Theta,P_i,T_i}(d_i|\theta, p_i, \tau_i),$$

the full posterior (3.4) reduces to,

$$f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p}) \propto \prod_{i=1}^{N} \big[\, f_{D_i|\Theta,P_i,T_i}\left(d_i|\,\theta,p_i,A\right)p_i \;+\; f_{D_i|\Theta,P_i,T_i}(d_i|\theta,p_i,B)\left(1-p_i\right)\big]. \tag{3.5}$$

In Section 3.6 we will make suggestions as to what functional form may be chosen for $f_{D_i|\Theta,P_i,T_i}$ when using BEAMS for independent SNe.

3.2 Rating τ_A-probabilities

An object's τ_A-probability is the expected proportion of other objects with its features which are type A. In other words, if an object has features x, its τ_A-probability is the expected proportion of objects with features x which are type A. Suppose that the global distribution of P is f_P. The expected total proportion of type A objects is then

$$\mathrm{P}(T=A) = \langle P \rangle = \int_0^1 p f_P(p)\, dp. \tag{3.6}$$

In some circumstances, it is necessary to go beyond calculating τ_A-probabilities and commit to an absolute classification, as was the case in the SNPCC. In such cases the optimal strategy moving from a τ_A-probability to an absolute type (A or B) is to classify objects positively (A) when the τ_A-probability is above some threshold probability (c). The False Positive Rate (FPR) using such a strategy is

$$\begin{aligned} \mathrm{FPR}(f_P) &= \mathrm{P}(P > c|T=B) \\ &= \frac{\int_c^1 (1-p) f_P(p)\, dp}{\int_0^1 (1-p) f_P(p)\, dp}, \end{aligned} \tag{3.7}$$

and the False Negative Rate is

$$\mathrm{FNR}(f_P) = \mathrm{P}(P < c|T=A)$$

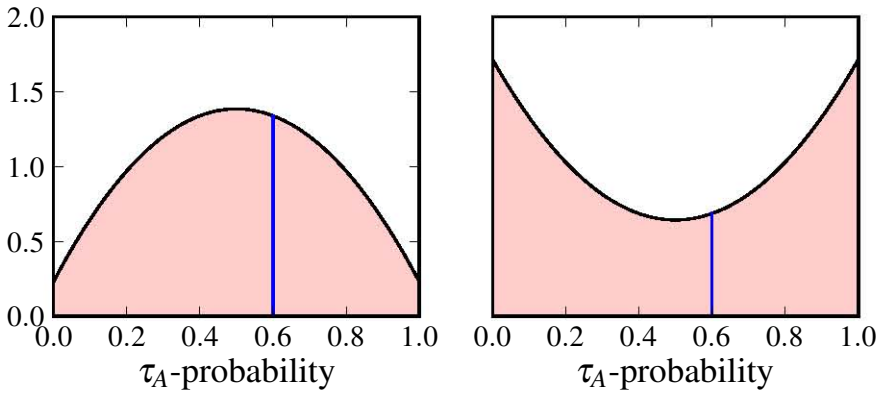

Figure 3.1: Two τ_A-probability distributions, both with means of 0.5. Using a threshold of 0.6, we have on left: FPR = 0.17, FNR = 0.45 and on right: FPR: 0.15, FNR = 0.28

$$= \frac{\int_0^c p\, f_P(p)\, dp}{\int_0^1 p\, f_P(p)\, dp}. \tag{3.8}$$

For SNe the FPR is the proportion of nIa SNe which are misclassified, while the FNR is the proportion of SNeIa which are misclassified (missed).

Intuition dictates that for classification problems, a useful f_P will be one whose mass predominates around 0 and 1. That is, an f_P which with high probability attaches decisive[2] τ_A-probabilities to observations. To minimize the FPR and FNR this is optimal, as illustrated in Figure 3.1.

We will be presenting a simulation illustrating how the decisiveness of τ_A-probabilities affects the parameter estimation of BEAMS. To simplify our study of the effect of the decisiveness of τ_A-probabilities on BEAMS, we introduce a family of distributions: For each $\mathcal{P} \in [0.5, 1]$ we have the distribution

$$f^{\mathcal{P}}(p) = \frac{1}{2}\left(\delta_{\mathcal{P}}(p) + \delta_{1-\mathcal{P}}(p)\right) \tag{3.9}$$

where $\delta_{\mathcal{P}}$ and $\delta_{1-\mathcal{P}}$ are δ-functions centered at $\mathcal{P}$ and $1-\mathcal{P}$ respectively. It is worth mentioning that we will be drawing probabilities from this distribution, which is potentially confusing. Drawing a observation of P from (3.9) is

[2] we say p_1 is more decisive than p_2 if $|p_1 - 0.5| > |p_2 - 0.5|$.

equivalent to drawing it from $\{1-\mathcal{P}, \mathcal{P}\}$ with equal probability:

$$\mathrm{P}(P = p) = \begin{cases} 0.5 & \text{if } p = \mathcal{P} \\ 0.5 & \text{if } p = 1 - \mathcal{P}. \end{cases}$$

If $\mathcal{P}_1$ is more decisive than $\mathcal{P}_2$, we say that the distribution $f^{\mathcal{P}_1}$ is more decisive than $f^{\mathcal{P}_2}$.

On page 5 of KBH it is stated that the expected proportion of type A objects (3.6) determines the expected error in estimating a parameter which is independent of population B. Specifically, they present the result that the expected error when estimating a parameter μ with N objects using BEAMS is given by,

$$\sigma_\mu \propto \sqrt{\langle P \rangle \, N}. \tag{3.10}$$

It should be noted that the the result from KBH (3.10) is an asymptotic result in N. For small N, the decisiveness of the probabilities plays an important part. If (3.6) were the only factor determining the expected error (σ_μ), then $f^{0.5}$ would be equivalent to f^1 in terms of expected error. This equivalence would mean that perfect type knowledge does not reduce error, which would be surprising. An example in Section 3.3.1 illustrates that decisiveness does play a role in determining the error.

As mentioned on page 8 of KBH, the effect of biases in τ_A-probabilities on BEAMS can be catastrophic. They consider the case where there is a uniform bias (a) of the τ_A-probabilities. That is, if observation i has a claimed τ_A-probability p_i of being type A, then the correct probability of it being type A is $p_i - a$. KBH show how, by including a free global shift parameter, such a bias is completely removed. However it is not clear what to do if the form of the bias is unknown. For example, it could be that there is an 'overconfidence' bias, where to obtain the true τ_A-probabilities one needs to transform the claimed priors ($\tilde{p}$) by

$$p = 0.2 + 0.6\,\tilde{p}. \tag{3.11}$$

Introducing a bias such as the one defined by (3.11) will have no effect

on the optimal FPR and FNR, provided the probability threshold is chosen optimally. This is because (3.11) is a one-to-one biasing, and so a threshold ($\tilde{c}$) on biased probabilities results in exactly the same partitioning as a threshold in the unbiased space of $0.2 + 0.6\tilde{c}$. However, introducing a bias such as (3.11) does have an effect on BEAMS parameter estimation, as we show in Section 3.4. In Section 3.6 we discuss how to guarantee that the τ_A-probabilities are free of bias.

3.3 Effects of Decisiveness and sample size on beams

In this section we will perform simulations to better understand the key factors in BEAMS. The data generated will have the following cosmological analogy: Θ - distance modulus at a given redshift z_0; $\boldsymbol{d}$ - the fitted distance moduli of SNe at z_0. Furthermore, $\boldsymbol{D}|\Theta, \boldsymbol{T}$ and $\boldsymbol{P}$ will be independent, such that the original and full posterior are equivalent.

3.3.1 Simulation 1: Estimating a population mean

This simulation was performed to see how the performance of BEAMS is affected by the decisiveness of τ_A-probabilities, and by the size of the data set. The two populations (A and B) were chosen to have distributions,

$$f_{D|T}(d, \tau) = \text{Normal}(\mu_\tau, 1), \tag{3.12}$$

where $\mu_A = -1$ and $\mu_B = +1$, as illustrated in Figure 3.3. The τ_A-probability distribution is chosen to be $f^{\mathcal{P}}$, so that about half of the observations have a τ_A-probability of $\mathcal{P}$, with the remaining observations having τ_A-probabilities of $1 - \mathcal{P}$. By varying $\mathcal{P}$ we vary the decisiveness.

Let us make it clear how the data for this simulation is generated. First, a τ_A-probability (p) is selected to be either $\mathcal{P}$ with probability 0.5 or $1-\mathcal{P}$ with probability 0.5, that is according to $f^{\mathcal{P}}$. Second, the type of the observation is chosen, with probability p it is chosen as A, and with probability $1 - p$ it

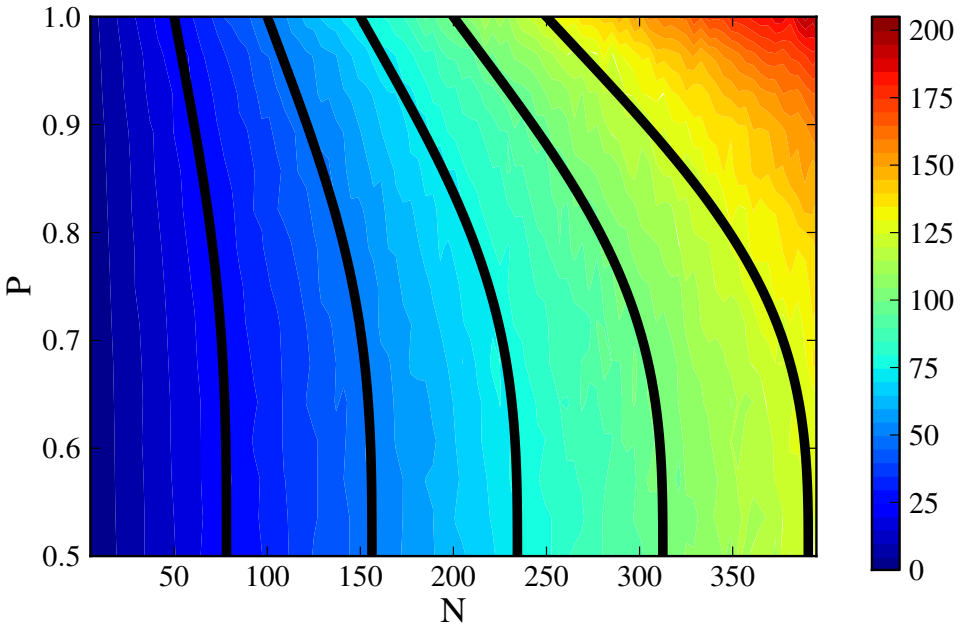

Figure 3.2: Contour plot of $h(N, \mathcal{P})$. The solid lines are approximations to lines of constant h, of the form (3.13).

is chosen as B. Finally, the data (d) is drawn from (3.12). Notice that $D|T$ is independent of P, and so the original posterior is equivalent to the full posterior.

In this simulation we only estimate μ_A, with all other parameters known. We use the following Figure of Merit (h) to compare the performance with different sample sizes (N) and decisivenesses $(\mathcal{P})$:

$$h(N, \mathcal{P}) = \frac{1}{\langle \hat{\mu}_A - \mu_A \rangle^2},$$

where $\hat{\mu}_A$ is the maximum likelihood estimate of μ_A using the original posterior on a sample of size N with τ_A-probabilities from $f^{\mathcal{P}}$, and $\langle \cdot \rangle$ denotes an expectation. Values of h were obtained by simulation, illustrating in Figure 3.2 the performance of BEAMS for various $(N, \mathcal{P})$ combinations. A

good approximation to the FoM h in Figure 3.2 appears to be

$$h(N, \mathcal{P}) \approx N \left(0.32 + 1.44(\mathcal{P} - \frac{1}{2})^3 \right), \tag{3.13}$$

although this approximation is an ad hoc observation. One interesting observation is that $h(N, \mathcal{P} = 1) \approx h(1.5N, \mathcal{P} = 0.5)$ in the region illustrated in Figure 3.2. This relationship tells us that given a completely blind sample ($\mathcal{P} = 0.5$), and the option to either double its size ($N \to 2N$) or to discover the hidden types ($\mathcal{P} : 0.5 \to 1$), doubling its size will provide more information about μ_A. It is important to reiterate that, according to the previously mentioned result of KBH, in the limit of $N \to \infty$ we do not expect $\mathcal{P}$ to play any part in determining $h(N, \mathcal{P})$. That is, for N sufficiently large, the FoM will be independent of $\mathcal{P}$.

While this simulation is too simple to make extrapolations about cosmological parameter estimations from, it may suggest that the information contained in unconfirmed photometric data may be currently underestimated.

3.3.2 Simulation 2: Estimating two population means

The two population distributions for this simulation are the same as those presented in Simulation 1 and as illustrated in Figure 3.3. In this simulation, we leave both the population means as free parameters to be fitted. Twenty objects are drawn from the types A and B, with the τ_A-probabilities are drawn from $f^{\mathcal{P}}$. The simulation is done with five different $\mathcal{P}$ values. The τ_A-probabilities are illustrated in Figure 3.4, and the approximate shape of the posterior marginals of μ_A for each $\mathcal{P}$ value are illustrated in Figure 3.5 by MCMC chain counts.

There are two interesting results from this simulation. The first is that there is negligible difference in performance between $\mathcal{P} = 1$ and $\mathcal{P} = 0.7$, so that having a 30% type uncertainty for all objects as opposed to absolute type knowledge does not weaken the results. The second is that as $\mathcal{P}$ approaches 0.5, BEAMS still correctly locates the population means but is unsure which mean belong to which population. We finally mention that the slight offset

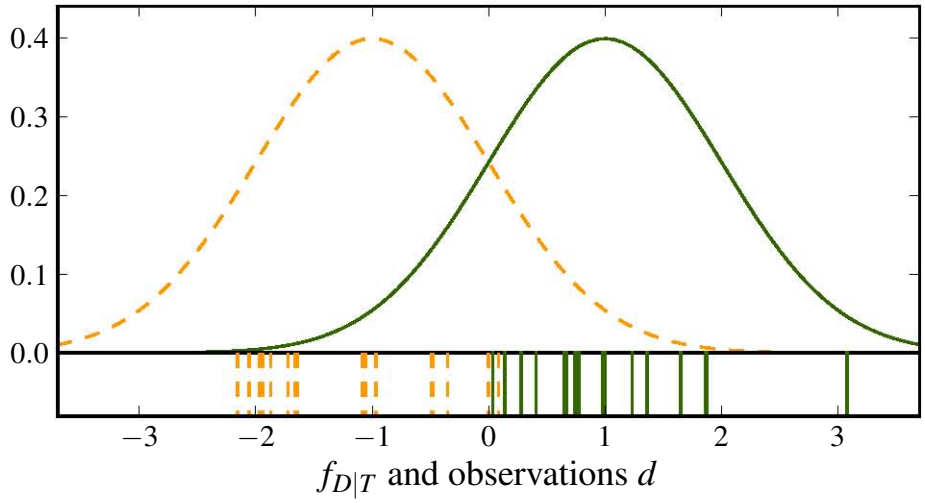

Figure 3.3: Population A (left) and population B (right) distributions, with (for Simulation 3.3.2) the observed values of D drawn from these distributions shown as vertical lines beneath.

of the distribution to below -1 is an artifact; the mean of the posterior is expected to vary about the true mean.

3.4 Effects of τ_A-probabilities bias on BEAMS

In the previous section we considered the effect of the decisiveness of τ_A-probabilities on the performance of BEAMS. In this section we will consider the effect of using incorrect τ_A-probabilities. We will again be estimating μ_A and μ_B where they are -1 and 1 respectively, and the population variances are again both known to be 1. The true τ_A-probability distribution will be $f^{0.8}$, that is

$$\mathrm{P}(P = p) = \begin{cases} 0.5 & \text{if } p = 0.8 \\ 0.5 & \text{if } p = 0.2 \end{cases}$$

It is worth reminding the reader that we are drawing probabilities from a probability distribution, an unusual thing to do. To generate a τ_A-probability from this distribution, one could flip a coin, and return $p = 0.2$ if H and $p = 0.8$ if T. We consider the effect of biasing τ_A-probabilities generated in

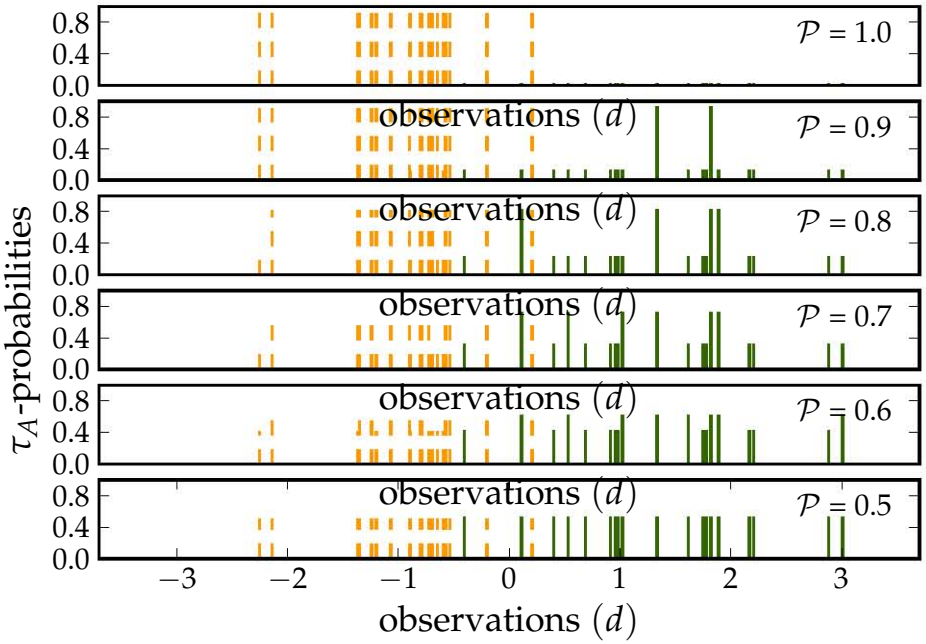

Figure 3.4: For values of $\mathcal{P}$ from 1 (above) to 0.5 (below), a τ_A-probability of $\mathcal{P}$ or $1 - \mathcal{P}$ is attached to each observation.

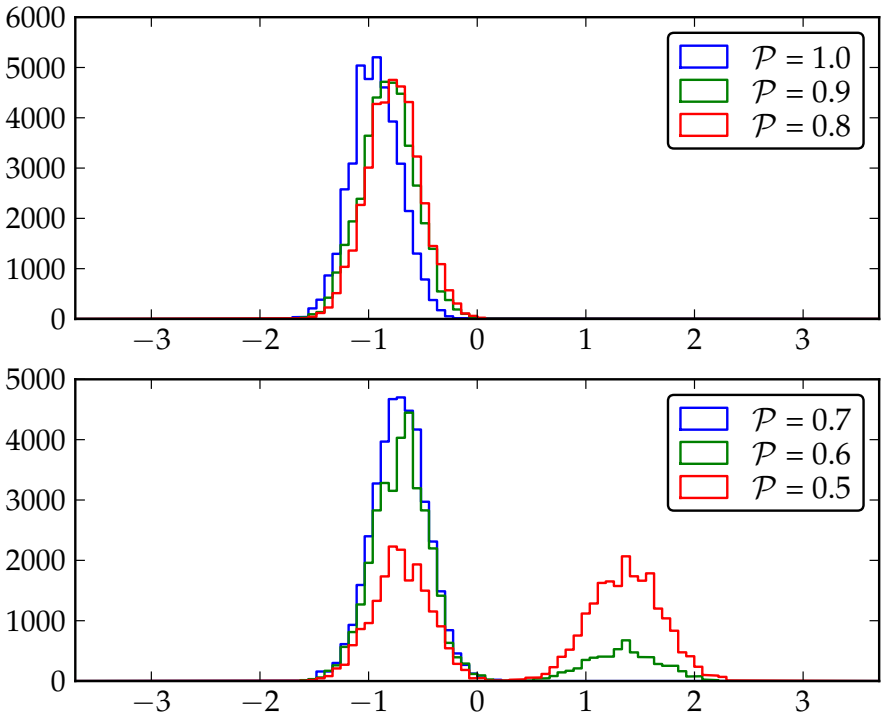

Figure 3.5: MCMC chain counts, approximating the posterior distributions of μ_A for the different values of decisiveness, $\mathcal{P}$.

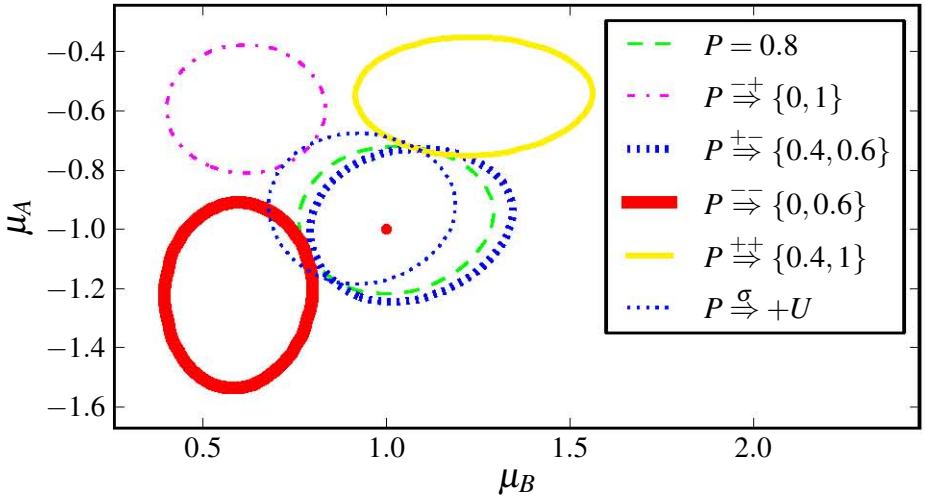

Figure 3.6: The 99 % posterior confidence regions using the five biasings of the τ_A-probabilities, as described in Section 3.4.

such a manner in the following ways:

1. $\mathcal{P} \overset{-+}{\Rightarrow} \{0,1\}$. Here the decisiveness of the τ_A-probabilities is overestimated, so that $p = 0.8 \to p = 1$ and $p = 0.2 \to p = 0$.

2. $\mathcal{P} \overset{+-}{\Rightarrow} \{0.4,0.6\}$. Here the decisiveness of τ_A-probabilities are underestimated, so that $p = 0.8 \to p = 0.6$ and $p = 0.2 \to p = 0.4$.

3. $\mathcal{P} \overset{--}{\Rightarrow} \{0,0.6\}$. Here the τ_A-probabilities are underestimated by 0.2, so that $p = 0.8 \to p = 0.6$ and $p = 0.2 \to p = 0$.

4. $\mathcal{P} \overset{++}{\Rightarrow} \{0.4,1\}$. Here the τ_A-probabilities are overestimated by 0.2, so that $p = 0.8 \to p = 1$ and $p = 0.2 \to p = 0.4$.

5. $\mathcal{P} \overset{\sigma}{\Rightarrow} U$. Here, to each τ_A-probability a uniform random number from $[-0.2, 0.2]$ is independently added.

The 99% posterior confidence regions obtained using these biased τ_A-probabilities in a simulation of 400 points are illustrated in Figure (3.6). The underestimation of decisiveness (2) has little effect on the final confidence

region, but overestimating the τ_A-probability decisiveness (1) results in a 6σ bias. Note that overestimating decisiveness results in the estimate $(\hat{\mu}_A, \hat{\mu}_B)$ being biased towards (μ_B, μ_A). This effect is caused by type B objects which are too confidently believed to be type A, which pull $\hat{\mu}_A$ towards μ_B, and type A objects which are too confidently believed to be type B, which pull $\hat{\mu}_B$ towards μ_A.

The contrast in effect between underestimating and overestimating the decisiveness of τ_A-probabilities is interesting, and not easy to explain. One suggestion we have received is to consider the cause of the observed effect as being analogous to the increased contamination rate induced by overestimating the decisiveness in the case BEAMS is not used. With an increased contamination rate comes an increased bias, precisely as observed in Figure 3.6. It is worth mentioning that underestimating the decisiveness is not entirely without effect, as simulations with more pronounced drops in $\mathcal{P}$ $(0.95 \rightarrow 0.55)$ result in noticable increases in the size of the 99% confidence region.

The effects of the flat τ_A-probability shifts (3) and (4) introduce biases larger than 4σ. This case was considered in KBH where, as we have already mentioned, it was shown that simultaneously fitting for this bias completely compensates for it. While this is a pleasing result, one would prefer to know that the τ_A-probabilities are correct, as one cannot be sure what form the biasing will take.

One phenomenon which is observed in this simulation, as it was in simulations as summarised in Table II on page 8 of KBH, is that a flat τ_A-probability shift in confidence towards being type B (3) does not bias the estimate of μ_A as much as it does the estimate of μ_B, and vice versa. In other words, underestimating the probabilities that objects are type A will result in less biased population A parameters than overestimating the probabilities. This result may also be understood in light of an analogy to increased contamination versus reduced population size in the case where BEAMS in not used. Finally, we notice that in this simulation the addition of unbiased noise to the τ_A-probabilities (5) has no significant effect. This suggests that systematic biases should be the primary concern of future work on the estimation

of τ_A-probabilities.

3.5 When given type, the data is still dependent on τ-priors

In this section we consider for the first time a simulation in which the data is not drawn from $f_{D|T}$, but from $f_{D|T,P}$, so that there is a dependence of the data on the τ_A-probability even when the type is known. The conditional pdfs are shown in Figure 3.7. To clarify the difference between this simulation and the previous ones, prior to this data was simulated as follows:

$$P \to T|P \to D|T,$$

where at the last step, the data was generated with a dependence only on type. Now it will be simulated as:

$$P \to T|P \to D|P,T.$$

More specifically, to generate data we start by drawing a τ_A-probability from $f^{0.7}$,

$$\mathrm{P}(P = p) = \begin{cases} 0.5 & \text{if } p = 0.7 \\ 0.5 & \text{if } p = 0.3. \end{cases}$$

Note that the above distribution guarantees that $\mathrm{P}(T = A) = \frac{1}{2}$. When the τ_A-probability (p) has been generated, we draw a type (τ) from $\{A, B\}$ according to

$$\mathrm{P}(T = \tau) = \begin{cases} p & \text{if } \tau = A \\ 1 - p & \text{if } \tau = B. \end{cases}$$

Once we have p and τ, we generate d. The marginals $f_{D|P,T}(d|p, \tau)$ have

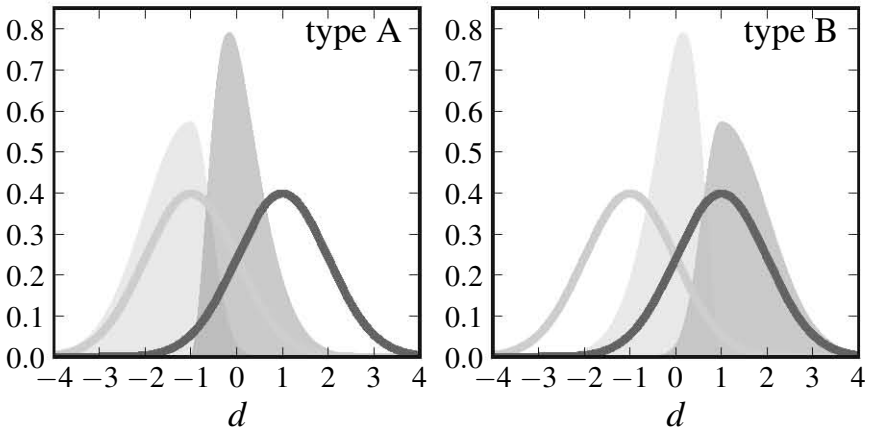

Figure 3.7: Plots of $f_{D|P,T}(d|p,\tau)$ (filled curves) for $p = 0.7$ (light) and $p = 0.3$ (dark), and for type A (left) and type B (right). Overlying are $f_{D|T}(d|A)$ (light) and $f_{D|T}(d|B)$ (dark).

been chosen such that we have

$$f_{D|T}(d|A) = \text{Normal}\,(-1, 1) \tag{3.14}$$

$$f_{D|T}(d|B) = \text{Normal}\,(1, 1), \tag{3.15}$$

as before. The marginal $f_{D|P,T}(d|0.7, A)$ is composed of the halves of two Gaussian curves with different σs, chosen such that the tail away from the B population is longer than the one towards the B population. Specifically,

$$f_{D|P,T}(d|0.7, A) = \\ = \begin{cases} K \exp -\frac{1}{2}(d+1)^2 & \text{if } d < -1 \\ K \exp -\frac{100}{32}(d+1)^2 & \text{if } d > -1 \end{cases}$$

where K is a normalizing constant. The marginal $f_{D|P,T}(d|0.3, A)$ is then constructed to guarantee (3.14). The above construction guarantees that the population of A objects with low τ_A-probabilities (0.3) lie on average closer to the B mean than do objects with high (0.7) τ_A-probabilities. The marginals of the B population are constructed to mirror exactly the A population marginals, as illustrated in Figure 3.7.

To compare the use of the original BEAMS posterior (3.2) with the full conditional posterior (3.5), we randomly draw 40 data points from the above distribution and construct the respective posterior distributions, as illustrated in Figure 3.8. Observe that the original posterior is significantly wider than the full posterior. Indeed, approximately half of the interior of the 80% region of the original posterior is ruled out to 1% by the full posterior. It is interesting to note that, while the original posterior is wider than the full posterior, it is not biased. This result goes against our intuition; we believed that the original posterior would result in estimates for μ_A and μ_B which exaggerated $|\mu_A - \mu_B|$. Whether it is a general result that no bias exists when the original posterior is used, or if there can exist dependencies between P and D for which the use of (3.1) leads to a bias, remains an open question.

Figure 3.8 illustrates one realistation from the distribution we have described, but repeated realisations show that on average, the variance in the maximum likelihood estimator using the original posterior is ~ 3 times larger than the variance using the modified posterior. While these simulations are too simple to draw conclusions about cosmological parameter estimation from, they do suggest that where correlations between τ_A-probabilities and distance moduli exist within a class of SNe, it may be worthwhile accounting for it by using the modified posterior. Currently it is most common when modelling SNe for cosmology, to assume that the likelihood $f_{D|\Theta,T}(d|\theta,\tau)$ is a Gaussian with unknown mean and variance,

$$D|\theta, T = \text{Normal}(\mu(\theta, T), \sigma(T)^2).$$

If one wishes to include the τ_A-probabilities in the likelihood, one could include a linear shift in P for the mean or variance. That is,

$$D|\theta, P, T = \text{Normal}(\mu(\theta, T) + c_1 P, \sigma(T)^2 + c_2 P).$$

Of course this is just one possibility, and one would need to analyse SN data to get a better idea of how P should enter into the above equation.

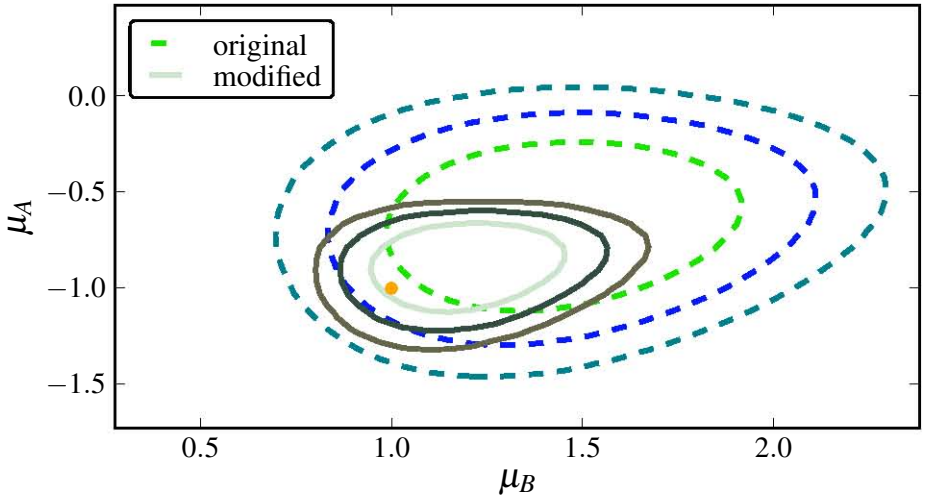

Figure 3.8: Posterior distributions on the parameters (μ_A, μ_B) using the correct posterior (3.4) (solid) and the original posterior (3.1) (dashed). The original posterior assumes independence between $D|T$ and P. Plotted are the 80%, 95% and 99% confidence levels. The true parameters (orange point) lie within the 95 % confidence regions of both posteriors.

3.6 Obtaining Unbiased τ_A-probabilities

In this section we investigate likely sources of τ_A-probability biases such as those presented in Section 3.4, and discuss how to detect and remove them. For SNe, one source of τ_A-probability bias could be the failure to take into account the preferential confirmation of bright objects. This type of bias has been considered in the machine learning literature under the name of selection bias, and we here present the relevant ideas from there. We end the section with a brief discussion on how one could model the pdfs $f_{D|\Theta,\boldsymbol{P},\boldsymbol{T}}$ and $f_{D|\Theta,\boldsymbol{F},\boldsymbol{P},\boldsymbol{T}}$, which are the likelihoods appearing in the extended posteriors introduced in Section 3.1.

3.6.1 Selection Bias

With respect to classification methods, selection bias refers to the situation where the confirmed data is a non-representative sample of the unconfirmed

data. A selection bias is sometimes also referred to as a covariate shift although the two are defined slightly differently, as described in [92]. With selection bias, the confirmed data set is first randomly selected from the full set, and then at a second stage it is non-randomly reduced. Such is the situation with a population census, where at a first stage, a random sample of people is selected from the full population, and then at a second stage, people of a certain disposition cooperate more readily than others, resulting in a biased sample of respondees.

A form of selection bias which is well known in observational astronomy is the Malmquist bias, whereby magnitude limited surveys lead to the preferential detection of intrinsically bright (low apparent magnitude) objects. In the case of SN cosmology, the bias is also towards the confirming of bright SNe. A reason for this bias is that the telescope time required to accurately classify a SN is inversely proportional to the SN's brightness. It is therefore relatively cheap to confirm bright objects and expensive to confirm faint ones.

If the SN confirmation bias is ignored, particular inferences made about the global population of SNe are likely to be inaccurate. In particular, estimates of a classifier's False Positive and False Negative Rates will be biased, and the estimated τ_A-probabilities will be biased in particular circumstances, as we will discuss in the following section.

Formalism

Following where possible the notation of [93], in what follows we assume that variables (X, T, F) are drawn from $\mathcal{X} \times \mathcal{T} \times \mathcal{F}$, where

1. $\mathcal{X}$ is the feature space,
2. $\mathcal{T} = \{A, B\}$ is the binary type space,
3. $\mathcal{F} = \{0, 1\}$ is the binary confirmation space, where $F = 1$ if confirmed (F for followed-up).

A realisation (x, τ, f) lies in either the test set or the training sets, defined respectively as:

$\underline{\text{test set}} \stackrel{\text{def}}{=} \{(x, \tau, f) \quad \text{s.t.} \quad f = 0\}$
$\underline{\text{training set}} \stackrel{\text{def}}{=} \{(x, \tau, f) \quad \text{s.t.} \quad f = 1\}.$

For SN cosmology it could be that $\mathcal{X}, \mathcal{T}$ and $\mathcal{F}$ are respectively,

1. $\mathcal{X}$ is the space of all possible photometric data, where a SN's photometric data consists of apparent magnitudes and observational standard deviations in four colour bands over several nights.

2. $\mathcal{T} = \{\text{Ia}, \text{nIa}\}$, type Ia and non-Ia SNe.

3. $\mathcal{F} = \{0, 1\}$, where $F = 1$ if the SN has been spectroscopically confirmed and thus has its type known.

By having a training set be unbiased we mean that it is a representative sample of the test set, specifically that F is independent of both X and T. That is, the probability of confirmation is independent of both features and type:

$$\mathrm{P}(F = 1 | X = x, T = \tau) = \mathrm{P}(F = 1) \text{ (for all } x \text{ and } \tau) \tag{3.16}$$

When the training set is unbiased, training set and test set objects are drawn from the same distribution over $\mathcal{X} \times \mathcal{T}$. This distribution over $\mathcal{X} \times \mathcal{T}$ can be estimated from the training set, so directly providing an estimate of the more useful test set distribution.

There are three important ways in which the independence relation (3.16) can break down, resulting in a biased training set, as described in [94] and listed below. By removing bias from a training set, we mean reweighting the training points such that the training set becomes unbiased.

1. Confirmation is independent of features when conditioned on type: X and $F | T$ are independent only. This violation of joint independence is the simplest kind of biasing, and there are methods for correcting for it [95], [96]; we do not have this type of bias in SN data.

2. Confirmation is independent of type when condtional on features: T and $F|X$ are independent. If the decision to confirm is based on X

and perhaps some other factors which are independent of T, this is the bias which exists. This the bias which probably exists in SN data, and there are methods for correcting for it, as we will discuss.

3. Confirmation depends on both features and type simultaneously. In this case, it is not possible to remove the bias from the data unless the exact form of the selection bias is known.

The decision to confirm a SN can be dictated by different features, examples include [91, 90], all of which are contained in the photometric data X. Such was the also case in the SNPCC where the probability of confirmation was based entirely on the peak magnitude in the r and i f, as we will discuss in Section 3.7. In reality, there are other factors which affect the confirmation decision such as the weather and telescope availability, but these are factors independent of SN type. Therefore the type 2 bias above is the bias which exists in the SN data. Thus, for the remainder of this section we will assume the type 2 bias, that is

$$\mathrm{P}(F = 1|X = x, T = \tau) = \mathrm{P}(F = 1|X = x) \text{ (for all } \tau) \tag{3.17}$$

The assumption of the type 2 bias can be made stronger. The decision to confirm an object does not in general depend on all of X but only a low-dimensional component (X_F) within X, and so we have

$$\mathrm{P}(F = 1|X = x, T = \tau) = \mathrm{P}(F = 1|X_F = x_F), \tag{3.18}$$

where X_F is contained in X. For SNe, X_F could be the peak apparent magnitude in particular colour bands.

In the following subsection we will describe how to correctly obtain τ_A-probabilities under the assumption of a bias described by (3.18).

3.6.2 Correctly obtaining τ_A-probabilities

Let us remind the reader as to how we defined τ_A-probabilities in the introduction:

$$\tau_A\text{-probability} \stackrel{\text{def}}{=} \mathrm{P}(T_i = A | X_{P,i} = x_{P,i}) = p_i, \tag{3.19}$$

where $X_{P,i}$ is an observable feature of the ith object, extracted from X_i. Estimates of p_i values can be obtained using several methods, of which those mentioned previously are [54, 43, 89, 66] It is worth rementioning that these different methods attempt to estimate different probability functions, as they each condition on different SN features. Thus there is no sense in which one set of τ_A-probabilities estimates is *the* correct set.

We now make an adjustment to definition (3.19), to take into account that biased follow-up may result in an additional conditional dependence on F:

$$\tau_A\text{-probability} \stackrel{\text{def}}{=} \mathrm{P}(T_i = A | F_i = f_i, X_{P,i} = x_{P,i}) = P_i. \tag{3.20}$$

The most informative τ_A-probabilities one could use would be those conditional on all of the features at one's disposal,

$$X_P = X: \quad p_i = \mathrm{P}(T = A | F = 0, X = x). \tag{3.21}$$

However, when $\mathcal{X}$ is a high-dimensional non-homogeneous space, as is the case with photometric SN data, it can be difficult to approximate (3.21) accurately. It is for this reason that it is necessary to reduce the features to a lower dimensional quantity $X_P \in \mathcal{X}_P$, so that the τ_A-probabilities are calculated from a subspace ($\mathcal{X}_P$) of the full feature space, as described by (3.20). The subspace $\mathcal{X}_P$ should be chosen to retain as much type specific information as possible while being of a sufficiently low dimension. In the SNPCC [43] chose $\mathcal{X}_P$ to be a 20-dimensional space of parameters obtained by fitting lightcurves.

The job of obtaining estimated τ_A-probabilities for test set objects ($F = 0$) is one of obtaining an estimate of the type probability mass function,

$$f_{T|F,X_P}. \tag{3.22}$$

Again, for (3.22) we prefer not to use the standard mass function notation, in order to to neaten particular integrals which follow. The τ_A-probability of a test set object can now be expressed in the following way,

$$\mathrm{P}(T = A|F = 0, X_P = x_P) = f_{T|F,X_P}(A|0, x_P).$$

Using kernel density estimation, boosting, or any other method of approximating a probability function, one can construct an approximation ($\hat{f}$) of the type probability function for training set objects,

$$\hat{f}(x_P) \approx f_{T|F,X_P}(A|1, x_P). \tag{3.23}$$

Using the estimate $\hat{f}$ in (3.23) one can estimate the τ_A-probabilities for the training set objects:

$$\mathrm{P}(T = A|F = 1, X_P = x_P) \approx \hat{f}(x_P). \tag{3.24}$$

The estimate (3.24) is not directly important as the training set object types are known exactly. But it is only through the training set objects that we can learn anything about the types of the test set objects.

How $\hat{f}$ from the training set is related to $f_{T|F=0,X_P}$ (3.22) depends on the relationship between X_F (the data which determines confirmation probability) and X_P (the data used to calculate τ_A-probabilities). There are two cases to consider. The first, which we write as $\mathcal{X}_F \subset \mathcal{X}_P$, arises when the data which determines confirmation probabilities is completely contained in the data used to calculate τ_A-probabilities. That is,

$$X_F \subset X_P \quad \overset{\text{def}}{\leftrightarrow} \quad \mathrm{P}(F = 1|X_P = x_P) = \mathrm{P}(F = 1|X_F = x_F).$$

The second case, when $\mathcal{X}_F \not\subset \mathcal{X}_P$ is when not all confirmation information is contained in X_P,

$$\mathcal{X}_F \not\subset \mathcal{X}_P \quad \overset{\text{def}}{\leftrightarrow} \quad \mathrm{P}(F = 1|X_P = x_P) \neq \mathrm{P}(F = 1|X_F = x_F).$$

In the case of $\mathcal{X}_F \subset \mathcal{X}_P$, it can be shown that,

$$\mathrm{P}(F = 1|T = \tau, X_P = x_P) = \mathrm{P}(F = 1|X_F = x_F). \tag{3.25}$$

$\mathcal{X}_F \subset \mathcal{X}_P$

We will show that in the case of $\mathcal{X}_F \subset \mathcal{X}_P$, a type probability function approximating the training population ($\hat{f}$) is an unbiased approximation for the type probability function of the test population ($F = 0$). To show this we start with the type probability of a test object:

$$\mathrm{P}(T = \tau|F = 0, X_P = x_P).$$

$\rightarrow$ Using Bayes' Theorem, we have

$$= \frac{\mathrm{P}(F = 0|T = \tau, X_P = x_P) \cdot \mathrm{P}(T = \tau|X_P = x_P)}{\mathrm{P}(F = 0|X_P = x_P)}$$

$\rightarrow$Then using (3.25), we have

$$\begin{aligned} &= \frac{\mathrm{P}(F = 0|X_F = x_F) \cdot \mathrm{P}(T = \tau|X_P = x_P)}{\mathrm{P}(F = 0|X_F = x_F)} \\ &= \mathrm{P}(T = \tau|X_P = x_P). \end{aligned} \tag{3.26}$$

$\rightarrow$ Using the same steps as above but in reverse and with $F = 1$, we arrive at

$$P(T = \tau|X_P = x_P) = \mathrm{P}(T = \tau|F = 1, X_P = x_P).$$

$\rightarrow$ On the right side is the type probability function for training set objects, and it can be approximated:

$$\approx \hat{f}(x_P). \tag{3.27}$$

This equation provides a useful result, as it says that $\hat{f}$ is not only an approximation of the type probability function of the training data, but also of the test set. Thus, $\hat{f}$ should provide unbiased τ_A-probabilities within the test set when $\mathcal{X}_F \subset \mathcal{X}_P$.

It should be noted that for $\hat{f}$ to be a good approximation for the test set, it is necessary that the training set covers all regions of $\mathcal{X}_P$ where there are test points. That is, if there are values of x_P for which $\mathrm{P}(X_P = x_P|F = 1) = 0$ and $\mathrm{P}(X_P = x_P|F = 0) \neq 0$, then the approximation $\hat{f}$ will not converge to $f_{T|F=0,X_P}$ as the training set size grows. One can refer to [93] for a full treatment of this topic.

With respect to SNe, the requirement of the preceding paragraph is that, if a SN is too faint to be confirmed and to enter the training set, it should not enter the test set either. We will return to this point again in Section 3.7.

One important question which we do not attempt to answer here is, how many SNe of different apparent magnitudes should be confirmed to obtain as rapid as possible convergence of $\hat{f}$ to $f_{T|F=0,X_P}$. An interesting method for deciding which SNe to confirm may be one based on the real-time approach proposed in [97], where the decision to add an object to the training set is based on the uncertainty of its type using the currently fitted $\hat{f}$. In Section 3.7 we discuss this further.

$\mathcal{X}_F \not\subset \mathcal{X}_P$

If $\mathcal{X}_F \not\subset \mathcal{X}_P$ we will not be able to use $\hat{f}$ to estimate the τ_A-probabilities in the test set, as (3.27) required $\mathcal{X}_F \subset \mathcal{X}_P$. In addition to this problem of not being able to use $\hat{f}$ to obtain unbiased τ_A-probabilities for the test set objects, if $\mathcal{X}_F \not\subset \mathcal{X}_P$ then

$$\mathrm{P}(T = \tau|X_P = x_P) \neq \mathrm{P}(T|X_P = x_P, X_F = x_F).$$

This inequality tells us that there is additional type information to be obtained from X_F, and so by not including X_F one is wasting type information. For this reason we recommend reconstructing the τ_A-probabilities based on redefined features, $X_P \leftarrow (X_F, X_P)$.

However, it is possible that one prefers not to use X_F in calculating τ_A-probabilities. This preference may be the case if one wishes to reduce the dependence between D and P, as presented in Section 3.1. For SNe, this reduction may involve obtaining τ_A-probabilities from shape alone, independent of magnitude, so that $\mathcal{X}_P$ is a space whose dimensions describe only shape and not magnitude. In this case, as we cannot use $\hat{f}$, we need to use the relationship derived in [98],

$$\begin{aligned} &\mathrm{P}(T = \tau | F = 0, X_P) \\ &= \int_{\mathcal{X}_F} f_{T,X_F|F,X_P}(\tau, x_F|1, x_P) \cdot w(x_F, x_P)\, dx_F, \end{aligned} \tag{3.28}$$

where the weight function is defined as

$$w(x_F, x_P) = \frac{f_{F|X_F}(0|x_F) f_{F|X_P}(1|x_P)}{f_{F|X_F}(1|x_F) f_{F|X_P}(0|x_P)}. \tag{3.29}$$

Notice that if $\mathcal{X}_F \subset \mathcal{X}_P$, then $w(x_F, x_P) = 1$ and so (3.28) reduces to the type probability function for training set objects, approximated by $\hat{f}$ as expected from (3.27). When $w(x_F, x_P) \neq 1$, the training set type probability function $\hat{f}$ cannot be used directly as an approximation to the test set type probability function. However, if each training set object is weighted using (3.28), then an unbiased test set type probability function approximation can be obtained.

The weight function (3.29) does not require any type information and so can be estimated as a first step. This additional step of estimation introduces additional error into the final estimate of (3.22), a theoretical analysis of which is presented in [99]. An alternative to the two-stage approach would be to fit the two terms in (3.28) simultaneously, as suggested and described by [92]. The use of (3.29) was first suggested in [98], where a detailed analysis of the asymptotic behaviour of its approximation is given. Therein, it is suggested that (3.29) be approximated by kernel density estimation.

In the case where F and X_P are independent, the weight function reduces

to one of only X_F,

$$w(X_F = x_F) = \frac{\mathrm{P}(F = 0|X_F = x_F)\mathrm{P}(F = 1)}{\mathrm{P}(F = 1|X_F = x_F)\mathrm{P}(F = 0)}. \tag{3.30}$$

This reduction in dimension may be valuable in approximating the weight function.

3.6.3 Detecting and removing biases in τ_A-probabilities

In the previous section we presented the correct way in which to estimate τ_A-probabilities in the case $\mathcal{X}_F \not\subset \mathcal{X}_P$. In this section we will present an example illustrating this process, but in the context of bias removal.

Suppose that we have a program which outputs scalar values ($\tilde{p}$), which are purported τ_A-probabilities. We believe that the output values have some unspecified bias, which we wish to remove. An assumption we make is that the $\tilde{p}$ values are calculated in the same way for training and test sets. That is that the program does not process cases $F = 0$ and $F = 1$ differently. It may seem strange to be interested in what the program does when $F = 1$, but as already mentioned it is only from the training set that we can learn anything about the test set. The idea now is to treat the received $\tilde{p}$ values as the x_P's from the previous section, and not directly as τ_A-probabilities.

For this example, we choose $\mathcal{X}_F = [0, 1]$. To now transform a test set value $\tilde{p} \in [0, 1]$ into an unbiased τ_A-probability using (3.28), one needs to estimate previous probability functions using kernel density estimation. The necessary functions we see from (3.28) and (3.29) are $f_{T,X_F|F,X_P}(\tau, x_F|1, \tilde{p})$, $f_{F|X_F}(1, x_F)$, $f_{F|X_P}(0, \tilde{p})$, $f_{F|X_F}(0, x_F)$ and $f_{F|X_P}(0, x_P)$.

It is an interesting and important question as to how accurately these probability functions can be approximated with few data points, but for this example we assume them known,

$$f_{T,X_F|F,X_P}(A, x_F|1, \tilde{p}) = \begin{cases} x_F & \text{if } \frac{1}{2}x_F^2 < \tilde{p} < 1 - \frac{1}{2}x_F^2, \\ 2 \cdot x_F & \text{if } \tilde{p} > 1 - \frac{1}{2}x_F^2 \\ 0 & \text{if } \tilde{p} < \frac{1}{2}x_F^2. \end{cases}$$

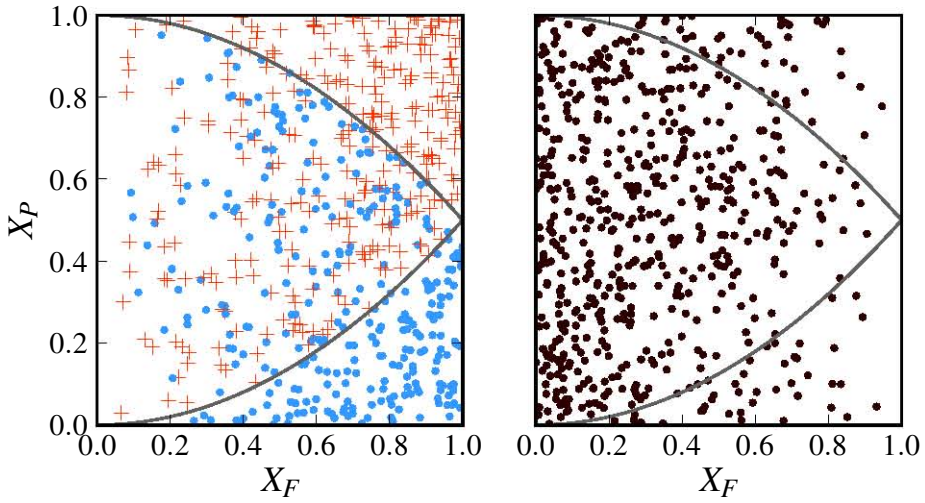

Figure 3.9: Realisations of a training set (left) containing type A (red pluses) and type B (blue points) objects, and a test set (right), drawn according to (3.31). Overlaid are faint lines delineating the discrete regions described by (3.31)

$$\begin{aligned} f_{F|X_F}(0, x_F) &= (1 - x_F), \\ f_{F|X_F}(1, x_F) &= x_F, \\ f_{F|X_P}(1|\tilde{p}) &= f_{F|X_P}(0|\tilde{p}) = \frac{1}{2}. \end{aligned} \tag{3.31}$$

Realisations from the above distribution are illustrated in Figure 3.9. By integrating x_F out of $f_{T,X_F|F,X_P}(A, x_F|1, \tilde{p})$ in (3.31), we have that

$$\mathrm{P}(T = A|F = 1, X_P = \tilde{p}) = \tilde{p}. \tag{3.32}$$

That is, in the training set $\tilde{p}$ is an unbiased estimate of a τ_A-probability. The τ_A-probabilities for objects in the test set we estimate using (3.28),

$$\begin{aligned} &\mathrm{P}(T = A|F = 0, X_P = \tilde{p}) \\ &\qquad = \int_{\mathcal{X}_F} f_{T,X_F|F,X_P}(\tau, x_F|0, x_P)\, dx_F \\ &\qquad = \int_{\mathcal{X}_F} f_{T,X_F|F,X_P}(\tau, x_F|1, x_P)\, w(x_F, \tilde{p})\, dx_F \end{aligned}$$

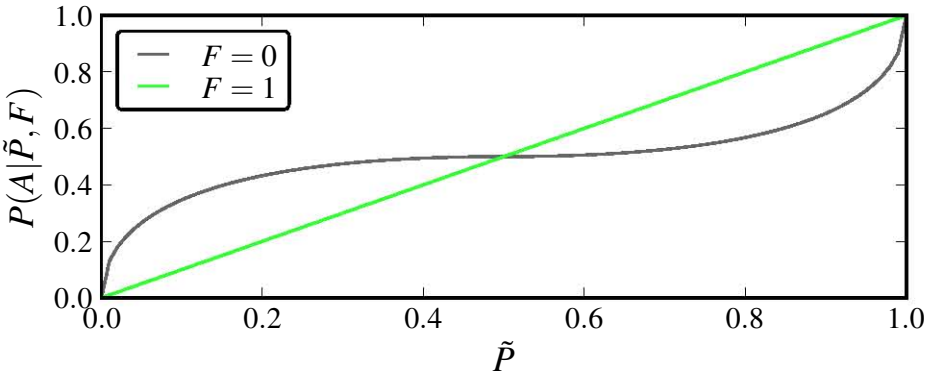

Figure 3.10: Corrected τ_A-probabilities. The disproportionately large number of training SNe with decisive τ_A-probabilities (as depicted in Figure 3.9), causes $\tilde{p}$ values to be too confident as test set τ_A-probability estimates.

$$
\begin{aligned}
&= \int_{\mathcal{X}_F} f_{T,X_F|F,X_P}(\tau, x_F|1, x_P)\, \frac{1 - x_F}{x_F}\, dx_F \\
&= \begin{cases} \sqrt{2\tilde{p}} - \tilde{p} & \text{if } \tilde{p} < 0.5, \\ 2 - \tilde{p} - \sqrt{2 - 2\tilde{p}} & \text{if } 0.5 < \tilde{p}. \end{cases}
\end{aligned} \tag{3.33}
$$

The τ_A-probabilities (3.32) and (3.33) are plotted in Figure 3.10, where we see that $\tilde{p}$ provided accurate τ_A-probabilities for the training set, but not for the test set. This contrast is not unexpected in reality, where the program providing the τ_A-probabilities may have been trained only on the biased training data. It is important to remember that this bias should only arise when $\mathcal{X}_F \not\subset \mathcal{X}_P$.

3.7 Supernova surveys and the SNPCC

The SNPCC provided a simulated spectroscopic training data set of approximately 1000 known SNe. The challenge was then to predict the types of approximately 20 000 other objects[3] from their lightcurves alone. Since the end of the competition, the types of all the simulated SNe have been released,

[3]These lightcurves are available at http://sdssdp62.fnal.gov/sdsssn/SIMGEN_PUBLIC/

making a post competition autopsy relatively easy to perform. In the results paper [59] we see that the probability that a SN was confirmed was based on the r-band and i-band quantities,

$$\epsilon_{\text{spec}}^{\text{band}} = \epsilon_0 \left(1 - x^l\right) \qquad x \stackrel{\text{def}}{=} \frac{m_{\text{peak}}^{\text{band}} - M_{\text{min}}^{\text{band}}}{m_{\text{lim}}^{\text{band}} - M_{\text{min}}^{\text{band}}}.$$

where $m_{\text{peak}}^{\text{band}}$ is the band-specific apparent magnitude of a SN, and $M_{\text{min}}^{\text{band}}$ and $m_{\text{min}}^{\text{band}}$ are constants. In [59] it is given that for the r and i bands,

$$\begin{aligned} \epsilon_{\text{spec}}^{r} &= \epsilon_0 \left(1 - x^5\right) \qquad x \stackrel{\text{def}}{=} \frac{m_{\text{peak}}^{r} - 16.0}{5.5} \\ \epsilon_{\text{spec}}^{i} &= \epsilon_0 \left(1 - x^6\right) \qquad x \stackrel{\text{def}}{=} \frac{m_{\text{peak}}^{i} - 21.5}{2.0} \end{aligned} \tag{3.34}$$

where ϵ_0 is some constant. Once $\epsilon_{\text{spec}}^{i}$ and $\epsilon_{\text{spec}}^{r}$ have been calculated, if a $[0 \to 1]$ uniform random number is less than either of them, confirmation is performed. As confirmation depends only on $\epsilon_{\text{spec}}^{i}$ and $\epsilon_{\text{spec}}^{r}$, we have from (3.26) that

$$\text{P}(T = \tau | F = 0, m_{\text{peak}}^{i}, m_{\text{peak}}^{r}) = \text{P}(T = \tau | F = 1, m_{\text{peak}}^{i}, m_{\text{peak}}^{r}). \tag{3.35}$$

Equation (3.35) can be interpreted as saying that the ratio Ia:nIa is the same in a given $m_{\text{peak}}^{i}, m_{\text{peak}}^{r}$ bin. The manner in which the follow-up was simulated should of course guarantee that (3.35) holds. In theory one should be able to deduce the verity of (3.35) from Figure 3.11, but the redshift bins with large numbers of confirmed SNe are too sparsely populated by unconfirmed SNe to check that the Ia:nIa is invariant. To be in a position where (3.35) can be checked is in general an unrealistic luxury, as without the types of the test objects this is impossible.

In terms of obtaining accurate τ_A-probabilities, a disturbing feature of Figure 3.11 is the absence of training SNe with high apparent magnitudes. With no training SNe with i-band apparent magnitudes greater than 23.5, we cannot infer the types of test SNe with apparent magnitudes greater than 23.5. Indeed there would be no non-astrophysical reason not to believe

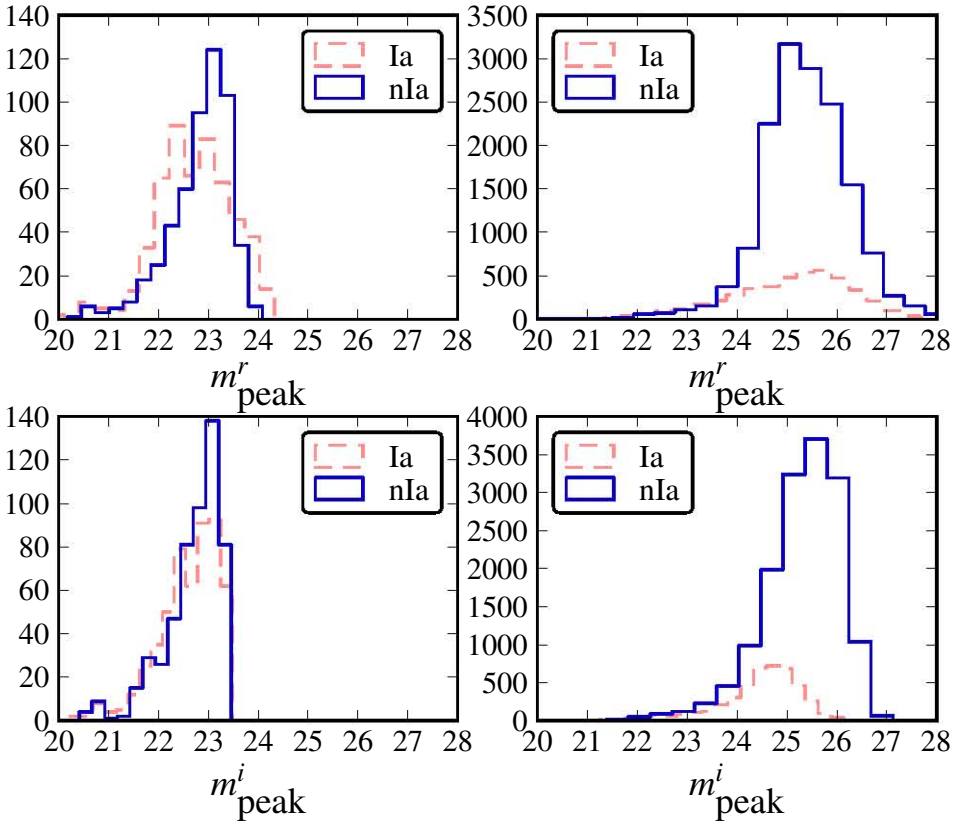

Figure 3.11: Counts of confirmed (left) and not confirmed (right) SNe, Ia (dashed) and non-Ia (solid) as a function of m^r_{peak} (above) and m^i_{peak} (below).

that all SNe with apparent magnitudes greater than 23.5 are non-Ia. As already mentioned in Section 3.6.1, in situations where the training set does not span the test set, one should ignore unrepresented test objects from all analyses. All test SNe other than those for which there are training SNe of comparable peak apparent magnitudes in r and i bands should be removed from a BEAMS analysis, unless there is a valid astrophysical reason not to do so. Cross validation entails ignoring about 95% of unconfirmed SNe; an enormous cut. We therefore consider it important to confirm more faint SNe.

In [43], a comparison is made between training a boosting algorithm on the non-representative spectroscopically confirmed SNe and a representative sample, randomly selected from the unconfirmed SN set. Therein, the authors use twenty fitted lightcurve parameters, including fitted apparent magnitudes in r and i bands. This situation corresponds to the one discussed in Section 3.6.2, where $\mathcal{X}_P \subset \mathcal{X}_F$. For this reason, the probability density function $\hat{f}$ in 3.24 as estimated by their boosting algorithm should be an unbiased estimate for $f_{T|F=0,X_P}$. But being unbiased does not guarantee low error, and when trained on the confirmed SNe, regions of parameter space corresponding to high apparent magnitude had no training SNe with which to learn, and so the approximation of 3.22 was poor. However when trained on the representative set, every region of populated parameter space was represented by the training set, and the approximation of 3.22 was greatly improved.

In their paper, [89] describe their entry in the SNPCC, and they report how a semi-supervised learning algorithm performs better with a few faint training SNe than with many bright ones. The comparison was performed while keeping the total confirmation time constant. Thus their conclusion was the same as ours; that it is important to obtain a more representative SN training sample.

3.8 Conclusions and Recommendations

In this chapter we discussed BEAMS, and extended the original posterior probability function to the case when $D|T$ (distance modulus | type) and P

(type probability) are dependent. In Section 3.5 we considered an example where the dependence between $D|T$ and P is strong, and observed a large reduction in the posterior width using the extended posterior as opposed to the original posterior. No bias is observed when using either the extended or the original posterior.

In Section 3.3 we considered examples where the original posterior is valid, that is when $D|T$ and P are independent. We performed tests to ascertain the importance to BEAMS of i) the decisiveness of the τ_A-probabilities (observations of P), and ii) sample size. In one test (3.3.1), we observed how doubling a sample size reduces error in parameter estimation more than obtaining the true type identity of the objects does. In another test (3.3.2), we observed how BEAMS accurately locates two population means, but fails to match each mean to its population.

We examined the effects of using biased τ_A-probabilities in Section (3.4). The result of KBH, that τ_A-probability biases towards population A affect the population's parameter estimates less than biases in favour of population B, was observed. A similar discovery is that biases towards high decisiveness are more damaging than biases towards low decisiveness. In other words, it is better to be conservative in your prior type beliefs than too confident.

Our recommendations for BEAMS may thus be summarised as follows. Firstly, the inclusion in the likelihood function of τ_A-probabilitiescan dramatically reduce the width of the final posterior, providing tighter constraints on cosmological parameters. Secondly, conservative estimation of τ_A-probabilities is less harmful than too decisive an estimation. Thirdly, it is possible to remove biases in τ_A-probabilities using the techniques described in Section (3.6).

In Section 3.6 we considered the problem of debiasing τ_A-probabilities. Interpreting recent results from the machine learning literature in terms of SN cosmology, we discussed the different ways in which training sets can be biased and how to remove such biases. The key to understanding and correcting biases is the relationship between $\mathcal{X}_F$ and $\mathcal{X}_P$, where $\mathcal{X}_F$ are object features which determine confirmation probability, and $\mathcal{X}_P$ are those features which determine τ_A-probabilities. In brief, when $\mathcal{X}_P$ contains $\mathcal{X}_F$,

τ_A-probabilities should be unbiased, but if this is not the case, there are sometimes ways for correcting the bias.

With respect to future SN surveys, we emphasize the importance of an accurate record as to what information is used when deciding whether or not a SN is confirmed. Using this information, one should in theory be able to remove all the affects of selection bias when $\mathcal{X}_F \not\subset \mathcal{X}_P$. In other words, using all the variables which are considered in deciding whether to follow-up a SN, it will always be possible to obtain unbiased τ_A-probabilities, irrespective of what the τ_A-probabilities are based on. Such follow-up variables may include early segments of light curves, χ^2 goodness of fits, fit probabilities, host galaxy position and type, expected peak apparent magnitude in particular filters, etc.

Our second recommendation for SN surveys is that more faint objects are confirmed. While it not necessary for most machine learning algorithms to have a spectroscopic training set which is exactly representative of the photometric test set, it is necessary that the spectroscopic set at least covers the photometric set. Thus having large numbers of faint unconfirmed objects without any confirmed faint objects is suboptimal.

Conclusions

We will now repeat the most important points from the Conclusions to Chapters 2 and 3. In Chapter 2 we discussed the problem of classifying SNe into sub-classes (Type Ia or non-Ia) based on photometric lightcurve data alone. This approach will be useful for future surveys which will detect vastly more candidates than will be possible to follow up spectroscopically.

We investigated two classes of classification algorithms, Kernel Density Estimation (KDE) and boosting, and applied them to simulated SNe lightcurve data, finding that the methods performed impressively as long as they were trained on a representative sample. Using the KDE approach, we considered both a 21 dimensional case based on lightcurve parameters from all bands and a 2 dimensional version based on fits to the Hubble diagram, using redshift information and an estimate of the distance modulus obtained using the SALT2 lightcurve fitting software.

A key issue for the classification methods we used was the training data sets. We compared the results based on training on two very different data sets: the first, a non-representative set, mimicking the kind of spectroscopic sample available as part of the follow-up program of a typical current-generation SN survey. The second was a representative sample of the same size where training objects were selected at random from the full sample.

In general we found that training on the representative sample produced exceptionally good results and that cross-validation on the training sample was able to accurately predict the purity and efficiency of the method on the full sample. On the other hand, training on the non-representative sample lead to relatively poor performance on the full data set and an inability to pre-

dict purity and efficiency. The importance of having an unbiased, representative sample is illustrated by the fact that for boosting, representative samples larger than about 50 objects outperformed the full non-representative sample of 1000 objects.

Our primary result and recommendation therefore is that boosting and KDE are powerful methods for SN classification, with remarkably little astrophysical input. However, they require training samples that are as unbiased and representative as possible. Our other main result is that neither boosting nor the 21D KDE method suffered particularly when the SN redshift information was unavailable. This outcome is important given that accurate redshifts will not be available for most candidates in the future.

In Chapter 3 we introduced BEAMS, and extended the original posterior probability function presented in [60] to the case when the data and type probabilities are dependent within classes. We considered an example where this dependence is strong, and observed a large difference between the resulting original and extended posterior distributions. No bias is observed when the original posterior is used.

We then considered examples where the original posterior is valid. In particular we performed tests to ascertain the importance to BEAMS of i) the decisiveness of the τ_A-probabilities, and ii) sample size. In one test, we observed that doubling sample size reduces error in parameter estimation more than obtaining the type identity of objects does. In another test, we observed how BEAMS accurately located two population means, but failed to match each mean to its population.

We examined the effects of using biased τ_A-probabilities. The result of KBH, that τ_A-probability biases away from a population affect the population's parameter estimates less than biases towards the population, was observed. A similar result which is discovered is that biases towards high decisiveness are more damaging than biases towards low decisiveness. In other words, it is better to be conservative in your prior type belief than overconfident.

We considered the problem of debiasing τ_A-probabilities. Interpreting recent results from the machine learning literature in terms of SN cosmology,

we discussed the different ways in which training sets can be biased and how to remove such biases. The key to understanding and correcting biases is the relationship between the data used in deciding to add observations to the training set, and that used to calculate type probability.

The relevence of the results from Chapter 3 to the findings in Chapter 2 was illustrated. It was shown how, under some circumstances, the problem induced by training on a non-representative spectroscopic training set for classification can be removed by reweighting the training SNe correctly. However, in practice, the dearth of faint, high redshift SNe may make reweighting futile. While it may be possible to apply astrophysical insight such as K-corrections [100] to determine how low redshift training set SNe would appear at high redshifts, if one wishes to be astrophysically blind then 90% of the photometric data must be discarded from cosmological parameter estimation. That is unless an exerted effort is made to spectroscopically confirm more high redshift objects.

Thus our overall conclusion is that having large numbers of faint unconfirmed objects without any confirmed faint objects is suboptimal if one wishes to estimate cosmological parameters in an astrophysically blind manner. In addition, if one wishes to use the reweighting scheme presented in Chapter 3, then we recommend that an accurate record is kept of what information is used to decide whether a SN is confirmed. An important point is that the reweighting we presented is not always necessary, and that we do not believe that reweighting would have improved our entries in the SNPCC.

Chapter 4

Declaration

I know the meaning of Plagiarism and declare that all of the work in the document, save for that which is properly acknowledged, is mine. I would like to acknowledge that,

1. Much of the Introduction of [43] was written by Bruce Bassett. As a result page 19 of the Introduction of this thesis is written mostly by Bruce Bassett.

2. The first three paragraphs of Section 2.1.1 were originally written for [43] by Mathew Smith, and at a later stage modified by me.

3. The section entitled Training Sample on page 25 was written by Bruce Bassett.

4. The section entitled SALT fits on page 30 was written by Matthew Smith.

5. Section 2.2.1 was written by Melvin Varughese and myself.

6. The implementation of the Hubble KDE, and the writing of Section 2.3.4 were done by Bruce Bassett.

7. The Discussion and Conclusion section of Chapter 2 was originally written by me, but went through several drafts and now contains sentences from several of the authors.

8. The first paragraph of 2.3.1 and the first paragraph of 2.3.2 were written by David Parkinson.

9. Section A2 was originally written by Martin Kunz.

In addition the following figures were not created by me:

- Mathew Smith: Figures 2.8, 2.22, 2.23
- Bruce Bassett: Figures 2.9, 2.20
- Bryony Martin: Figure 2.11
- other source: Figures 1.1, 1.2, 1.3, 2.1

Finally, it should be noted that much of the unreferenced information presented in the Introduction is from [5], [6], Wikipedia and discussions with Bruce Bassett, Mathew Smith and Patrice Okouma.

Bibliography

[1] P. F. Winkler, G. Gupta and K. S. Long, *The SN 1006 Remnant: Optical Proper Motions, Deep Imaging, Distance, and Brightness at Maximum*, ApJ **585** (Mar., 2003) 324–335 [arXiv:astro-ph/0208415].

[2] J. S. Bloom and J. W. Richards, *Data Mining and Machine-Learning in Time-Domain Discovery & Classification*, arXiv:1104.3142.

[3] S. Simon, Big Bang. Fourth Estate, New York, 2004.

[4] W. Baade and F. Zwicky, *On Super-novae*, Proceedings of the National Academy of Science **20** (May, 1934) 254–259.

[5] M. H. Jones, R. Lambourne, D. J. Adams and O. University, An Introduction to Galaxies and Cosmology. Cambridge University Press, Cambridge, 2004. Published in association with the Open University.

[6] P. Ruiz-Lapuente, Dark Energy: Observational and Theoretical Approaches. Cambridge University Press, 2010.

[7] A. V. Filippenko, *Optical spectra of supernovae*, ARAA **35** (1997), no. 1 309–355 [http://www.annualreviews.org/doi/pdf/10.1146/annurev.astro.35.1.309].

[8] W. Hillebrandt and J. C. Niemeyer, *Type Ia Supernova Explosion Models*, ARAA **38** (2000) 191–230 [arXiv:astro-ph/0006305].

[9] B. Paczyński, *Evolutionary Processes in Close Binary Systems*, ARAA **9** (1971) 183.

[10] N. Benítez, A. Riess, P. Nugent, M. Dickinson, R. Chornock and A. V. Filippenko, *The magnification of SN 1997ff, the farthest known Supernova*, ApJ **577** (2002) [astro-ph/0207097].

[11] E. Hubble, *A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae*, Proceedings of the National Academy of Science **15** (Mar., 1929) 168–173.

[12] D. L. Block, *A Hubble Eclipse: Lemaître and Censorship*, arXiv:1106.3928.

[13] P. de Bernardis, P. A. R. Ade, J. J. Bock, J. R. Bond, J. Borrill, A. Boscaleri, K. Coble, C. R. Contaldi, B. P. Crill, G. D. Troia, P. Farese, K. Ganga, M. Giacometti, E. Hivon, V. V. Hristov, A. Iacoangeli, A. H. Jaffe, W. C. Jones, A. E. Lange, L. Martinis, S. Masi, P. Mason, P. D. Mauskopf, A. Melchiorri, T. Montroy, C. B. Netterfield, E. Pascale, F. Piacentini, D. Pogosyan, G. Polenta, F. Pongetti, S. Prunet, G. Romeo, J. E. Ruhl and F. Scaramuzzi, *Multiple Peaks in the Angular Power Spectrum of the Cosmic Microwave Background: Significance and Consequences for Cosmology*, ApJ **564** (2002), no. 2 559.

[14] L. M. Krauss and B. Chaboyer, *Age Estimates of Globular Clusters in the Milky Way: Constraints on Cosmology*, Science **299** (2003) 65–70.

[15] E. Komatsu et. al., *Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation*, arXiv:1001.4538.

[16] C. T. Kowal, *Absolute magnitudes of supernovae*, ApJ **73** (Dec., 1968) 1021–1024.

[17] R. P. Kirshner, Foundations of supernova cosmology, p. 151. "Cambridge University Press", 2010.

[18] R. Kessler, A. C. Becker, D. Cinabro, J. Vanderplas, J. A. Frieman, J. Marriner, T. M. Davis, B. Dilday, J. Holtzman, S. W. Jha, H. Lampeitl, M. Sako, M. Smith, C. Zheng, R. C. Nichol, B. Bassett, R. Bender, D. L. Depoy, M. Doi, E. Elson, A. V. Filippenko, R. J. Foley, P. M. Garnavich, U. Hopp, Y. Ihara, W. Ketzeback, W. Kollatschny, K. Konishi, J. L. Marshall, R. J. McMillan, G. Miknaitis, T. Morokuma, E. Mörtsell, K. Pan, J. L. Prieto, M. W. Richmond, A. G. Riess, R. Romani, D. P. Schneider, J. Sollerman, N. Takanashi, K. Tokita, K. van der Heyden, J. C. Wheeler, N. Yasuda and D. York, *First-Year Sloan Digital Sky Survey-II Supernova Results: Hubble Diagram and Cosmological Parameters*, ApJS **185** (Nov., 2009) 32–84 [arXiv:0908.4274].

[19] M. M. Phillips, *The absolute magnitudes of Type Ia supernovae*, ApJL **413** (Aug., 1993) L105–L108.

[20] A. G. Riess, W. H. Press and R. P. Kirshner, *A Precise Distance Indicator: Type Ia Supernova Multicolor Light-Curve Shapes*, ApJ **473** (1996), no. 1 88.

[21] A. G. Riess et. al., *Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant*, AJ **116** (Sept., 1998) 1009–1038 [arXiv:astro-ph/9805201].

[22] S. Perlmutter et. al., *Measurements of Omega and Lambda from 42 High-Redshift Supernovae*, ApJ **517** (June, 1999) 565–586 [arXiv:astro-ph/9812133].

[23] P. Preuss, "Supernova magnitude correction." (July, 2011) http://www.lbl.gov/Science-Articles/Archive/sabl/2005/October/04-supernovae.html.

[24] D. J. Eisenstein, I. Zehavi, D. W. Hogg, R. Scoccimarro, M. R. Blanton, R. C. Nichol, R. Scranton, H. Seo, M. Tegmark, Z. Zheng, S. F. Anderson, J. Annis, N. Bahcall, J. Brinkmann, S. Burles, F. J. Castander, A. Connolly, I. Csabai, M. Doi, M. Fukugita, J. A.

Frieman, K. Glazebrook, J. E. Gunn, J. S. Hendry, G. Hennessy, Z. Ivezić, S. Kent, G. R. Knapp, H. Lin, Y. Loh, R. H. Lupton, B. Margon, T. A. McKay, A. Meiksin, J. A. Munn, A. Pope, M. W. Richmond, D. Schlegel, D. P. Schneider, K. Shimasaku, C. Stoughton, M. A. Strauss, M. SubbaRao, A. S. Szalay, I. Szapudi, D. L. Tucker, B. Yanny and D. G. York, *Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies*, ApJ **633** (Nov., 2005) 560–574 [arXiv:astro-ph/0501171].

[25] W. J. Percival, S. Cole, D. J. Eisenstein, R. C. Nichol, J. A. Peacock, A. C. Pope and A. S. Szalay, *Measuring the Baryon Acoustic Oscillation scale using the Sloan Digital Sky Survey and 2dF Galaxy Redshift Survey*, MNRAS **381** (Nov., 2007) 1053–1066 [arXiv:0705.3323].

[26] A. Mantz, S. W. Allen, D. Rapetti and H. Ebeling, *The observed growth of massive galaxy clusters - I. Statistical methods and cosmological constraints*, MNRAS **406** (Aug., 2010) 1759–1772 [arXiv:0909.3098].

[27] L. Fu et. al., *Very weak lensing in the CFHTLS wide: cosmology from cosmic shear in the linear regime*, A&A **479** (Feb., 2008) 9–25 [arXiv:0712.0884].

[28] T. Giannantonio, R. Scranton, R. G. Crittenden, R. C. Nichol, S. P. Boughn, A. D. Myers and G. T. Richards, *Combined analysis of the integrated Sachs-Wolfe effect and cosmological implications*, Phys. Rev. D **77** (June, 2008) 123520–+ [arXiv:0801.4380].

[29] W. J. Percival et. al., *Baryon acoustic oscillations in the Sloan Digital Sky Survey Data Release 7 galaxy sample*, MNRAS **401** (Feb., 2010) 2148–2168 [arXiv:0907.1660].

[30] R. J. Foley, A. V. Filippenko, C. Aguilera, A. C. Becker, S. Blondin, P. Challis, A. Clocchiatti, R. Covarrubias, T. M. Davis, P. M. Garnavich, S. W. Jha, R. P. Kirshner, K. Krisciunas, B. Leibundgut,

B. Li, T. Matheson, A. Miceli, G. Miknaitis, G. Pignata, A. Rest, A. G. Riess, B. P. Schmidt, R. C. Smith, J. Sollerman, J. Spyromilio, C. W. Stubbs, N. B. Suntzeff, J. L. Tonry, W. M. Wood-Vasey and A. Zenteno, *Constraining Cosmic Evolution of Type Ia Supernovae*, ApJ **684** (2008), no. 1 68.

[31] C. Kobayashi and K. Nomoto, *The Role of Type Ia Supernovae in Chemical Evolution. I. Lifetime of Type Ia Supernovae and Metallicity Effect*, ApJ **707** (Dec., 2009) 1466–1484 [arXiv:0801.0215].

[32] K. Konishi, D. Cinabro, P. M. Garnavich, Y. Ihara, R. Kessler, J. Marriner, D. P. Schneider, M. Smith, H. Spinka, J. C. Wheeler and N. Yasuda, *Dependences of Type Ia Supernovae Lightcurve Parameters on the Host Galaxy Star Formation Rate and Metallicity*, arXiv:1101.4269.

[33] F. X. Timmes, E. F. Brown and J. W. Truran, *On Variations in the Peak Luminosity of Type Ia Supernovae*, ApJL **590** (2003), no. 2 L83.

[34] Nearby Supernova Factory, N. Chotard, E. Gangler, G. Aldering, P. Antilogus, C. Aragon, S. Bailey, C. Baltay, S. Bongard, C. Buton, A. Canto, M. Childress, Y. Copin, H. K. Fakhouri, E. Y. Hsiao, M. Kerschhaggl, M. Kowalski, S. Loken, P. Nugent, K. Paech, R. Pain, E. Pecontal, R. Pereira, S. Perlmutter, D. Rabinowitz, K. Runge, R. Scalzo, G. Smadja, C. Tao, R. C. Thomas, B. A. Weaver and C. Wu, *The reddening law of type Ia supernovae: separating intrinsic variability from dust using equivalent widths*, A&A **529** (May, 2011) L4+ [arXiv:1103.5300].

[35] A. Aguirre, *Intergalactic Dust and Observations of Type Ia Supernovae*, ApJ **525** (1999), no. 2 583.

[36] A. G. Riess, P. E. Nugent, R. L. Gilliland, B. P. Schmidt, J. Tonry, M. Dickinson, R. I. Thompson, T. Budavari, S. Casertano, A. S. Evans, A. V. Filippenko, M. Livio, D. B. Sanders, A. E. Shapley, H. Spinrad, C. C. Steidel, D. Stern, J. Surace and S. Veilleux, *The*

Farthest Known Supernova: Support for an Accelerating Universe and a Glimpse of the Epoch of Deceleration, ApJ **560** (2001), no. 1 49.

[37] D. J. Schlegel, D. P. Finkbeiner and M. Davis, *Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds*, ApJ **500** (1998), no. 2 525.

[38] J. Guy, M. Sullivan, A. Conley, N. Regnault, P. Astier, C. Balland, S. Basa, R. G. Carlberg, D. Fouchez, D. Hardin, I. M. Hook, D. A. Howell, R. Pain, N. Palanque-Delabrouille, K. M. Perrett, C. J. Pritchet, J. Rich, V. Ruhlmann-Kleider, D. Balam, S. Baumont, R. S. Ellis, S. Fabbro, H. K. Fakhouri, N. Fourmanoit, S. González-Gaitán, M. L. Graham, E. Hsiao, T. Kronborg, C. Lidman, A. M. Mourao, S. Perlmutter, P. Ripoche, N. Suzuki and E. S. Walker, *The Supernova Legacy Survey 3-year sample: Type Ia supernovae photometric distances and cosmological constraints*, A&A **523** (Nov., 2010) A7+ [arXiv:1010.4743].

[39] M. A. Hendry, J. F. L. Simmons and A. M. Newsam, *What do we mean by "Malmquist Bias"?*, in Cosmic Velocity Fields, pp. 23–+, 1993. arXiv:astro-ph/9310028.

[40] M. A. Hendry, J. F. L. Simmons and A. M. Newsam, *What do we mean by Malmquist Bias?*, in Cosmic Velocity Fields, pp. 23–+, 1993. arXiv:astro-ph/9310028.

[41] D. Poznanski, P. E. Nugent and A. V. Filippenko, *Type II-P Supernovae as Standard Candles: The SDSS-II Sample Revisited*, ApJ **721** (Oct., 2010) 956–959 [arXiv:1008.0877].

[42] D. J. Eisenstein, I. Zehavi, D. W. Hogg, R. Scoccimarro, M. R. Blanton, R. C. Nichol, R. Scranton, H.-J. Seo, M. Tegmark, Z. Zheng, S. F. Anderson, J. Annis, N. Bahcall, J. Brinkmann, S. Burles, F. J. Castander, A. Connolly, I. Csabai, M. Doi, M. Fukugita, J. A. Frieman, K. Glazebrook, J. E. Gunn, J. S. Hendry, G. Hennessy, Z. Ivezić, S. Kent, G. R. Knapp, H. Lin, Y.-S. Loh, R. H. Lupton,

B. Margon, T. A. McKay, A. Meiksin, J. A. Munn, A. Pope, M. W. Richmond, D. Schlegel, D. P. Schneider, K. Shimasaku, C. Stoughton, M. A. Strauss, M. SubbaRao, A. S. Szalay, I. Szapudi, D. L. Tucker, B. Yanny and D. G. York, *Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies*, ApJ **633** (Nov., 2005) 560–574 [arXiv:astro-ph/0501171].

[43] J. Newling, M. Varughese, B. Bassett, H. Campbell, R. Hlozek, M. Kunz, H. Lampeitl, B. Martin, R. Nichol, D. Parkinson and M. Smith, *Statistical Classification Techniques for Photometric Supernova Typing*, MNRAS (2011) [arXiv:1010.1005].

[44] A. V. Filippenko, W. D. Li, R. R. Treffers and M. Modjaz, *The Lick Observatory Supernova Search with the Katzman Automatic Imaging Telescope*, in IAU Colloq. 183: Small Telescope Astronomy on Global Scales, vol. 246 of Astronomical Society of the Pacific Conference Series, p. 121, 2001.

[45] G. Aldering et. al., *Overview of the Nearby Supernova Factory*, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 4836 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, pp. 61–72, Dec., 2002.

[46] P. Astier et. al., *The Supernova Legacy Survey: measurement of Ω_M, Ω_Λ and w from the first year data set*, A&A **447** (Feb., 2006) 31–48 [arXiv:astro-ph/0510447].

[47] A. Clocchiatti et. al., *Hubble Space Telescope and Ground-based Observations of Type Ia Supernovae at Redshift 0.5: Cosmological Implications*, ApJ **642** (May, 2006) 1–21 [arXiv:astro-ph/0510155].

[48] R. Kessler et. al., *First-Year Sloan Digital Sky Survey-II Supernova Results: Hubble Diagram and Cosmological Parameters*, ApJS **185** (Nov., 2009) 32–84 [arXiv:0908.4274].

[49] G. Folatelli et. al., *The Carnegie Supernova Project: Analysis of the First Sample of Low-Redshift Type-Ia Supernovae*, AJ **139** (Jan., 2010) 120–144 [arXiv:0910.3317].

[50] The Dark Energy Survey Collaboration, *The Dark Energy Survey*, arXiv:astro-ph/0510346.

[51] N. Kaiser and Pan-STARRS Team, *The Pan-STARRS Survey Telescope Project*, in Bulletin of the American Astronomical Society, vol. 37 of Bulletin of the American Astronomical Society, p. 1409, Dec., 2005.

[52] B. P. Schmidt, S. C. Keller, P. J. Francis and M. S. Bessell, *The SkyMapper Telescope and Southern Sky Survey*, in Bulletin of the American Astronomical Society, vol. 37 of Bulletin of the American Astronomical Society, p. 457, May, 2005.

[53] J. A. Tyson, *Large Synoptic Survey Telescope: Overview*, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 4836 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, pp. 10–20, Dec., 2002. arXiv:astro-ph/0302102.

[54] D. Poznanski, A. Gal-Yam, D. Maoz, A. V. Filippenko, D. C. Leonard and T. Matheson, *Not Color-Blind: Using Multiband Photometry to Classify Supernovae*, PASP **114** (Aug., 2002) 833–845 [arXiv:astro-ph/0202198].

[55] D. R. Brett, R. G. West and P. J. Wheatley, *The automated classification of astronomical light curves using Kohonen self-organizing maps*, MNRAS **353** (Sept., 2004) 369–376 [arXiv:astro-ph/0408118].

[56] S. A. Rodney and J. L. Tonry, *Fuzzy Supernova Templates. I. Classification*, ApJ **707** (Dec., 2009) 1064–1079 [0910.3702].

[57] J. A. Frieman, B. Bassett, A. Becker, C. Choi, D. Cinabro, F. DeJongh, D. L. Depoy, B. Dilday, M. Doi, P. M. Garnavich, C. J. Hogan, J. Holtzman, M. Im, S. Jha, R. Kessler, K. Konishi, H. Lampeitl, J. Marriner, J. L. Marshall, D. McGinnis, G. Miknaitis, R. C. Nichol, J. L. Prieto, A. G. Riess, M. W. Richmond, R. Romani, M. Sako, D. P. Schneider, M. Smith, N. Takanashi, K. Tokita, K. van der Heyden, N. Yasuda, C. Zheng, J. Adelman-McCarthy, J. Annis, R. J. Assef, J. Barentine, R. Bender, R. D. Blandford, W. N. Boroski, M. Bremer, H. Brewington, C. A. Collins, A. Crotts, J. Dembicky, J. Eastman, A. Edge, E. Edmondson, E. Elson, M. E. Eyler, A. V. Filippenko, R. J. Foley, S. Frank, A. Goobar, T. Gueth, J. E. Gunn, M. Harvanek, U. Hopp, Y. Ihara, Ž. Ivezić, S. Kahn, J. Kaplan, S. Kent, W. Ketzeback, S. J. Kleinman, W. Kollatschny, R. G. Kron, J. Krzesiński, D. Lamenti, G. Leloudas, H. Lin, D. C. Long, J. Lucey, R. H. Lupton, E. Malanushenko, V. Malanushenko, R. J. McMillan, J. Mendez, C. W. Morgan, T. Morokuma, A. Nitta, L. Ostman, K. Pan, C. M. Rockosi, A. K. Romer, P. Ruiz-Lapuente, G. Saurage, K. Schlesinger, S. A. Snedden, J. Sollerman, C. Stoughton, M. Stritzinger, M. Subba Rao, D. Tucker, P. Vaisanen, L. C. Watson, S. Watters, J. C. Wheeler, B. Yanny and D. York, *The Sloan Digital Sky Survey-II Supernova Survey: Technical Summary*, AJ **135** (Jan., 2008) 338–347 [arXiv:0708.2749].

[58] R. Kessler, A. Conley, S. Jha and S. Kuhlmann, *Supernova Photometric Classification Challenge*, arXiv:1001.5210.

[59] R. Kessler et. al., *Results from the Supernova Photometric Classification Challenge*, arXiv:1008.1024.

[60] M. Kunz, B. A. Bassett and R. A. Hlozek, *Bayesian estimation applied to multiple species*, Phys. Rev. D **75** (May, 2007) 103508 [arXiv:astro-ph/0611004].

[61] DES, "Dark energy survey." (July, 2011) http://www.darkenergysurvey.org/ index.shtml.

[62] M. Bartelmann and S. D. M. White, *Cluster detection from surface-brightness fluctuations in SDSS data*, A&A **388** (June, 2002) 732–740 [arXiv:astro-ph/0110647].

[63] M. Fukugita, T. Ichikawa, J. E. Gunn, M. Doi, K. Shimasaku and D. P. Schneider, *The Sloan Digital Sky Survey Photometric System*, AJ **111** (Apr., 1996) 1748–+.

[64] H. Oyaizu, M. Lima, C. E. Cunha, H. Lin, J. Frieman and E. S. Sheldon, *A Galaxy Photometric Redshift Catalog for the Sloan Digital Sky Survey Data Release 6*, ApJ **674** (Feb., 2008) 768–783 [arXiv:0708.0030].

[65] AIMS collaborators, "Cosmology at AIMS, 2010, Boosting for Supernova Classification." (March, 2011) http://cosmoaims.wordpress.com /2010/09/30/boosting-for-supernova-classification/.

[66] J. Guy, P. Astier, S. Baumont and D. Hardin, *SALT2: using distant supernovae to improve the use of type Ia supernovae as distance indicators*, A&A **466** (Apr., 2007) 11–21 [arXiv:astro-ph/0701828].

[67] M. Hicken, W. M. Wood-Vasey, S. Blondin and et al., *Improved Dark Energy Constraints from ˜100 New CfA Supernova Type Ia Light Curves*, arXiv:0901.4804.

[68] Y. Freund and R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, in European Conference on Computational Learning Theory, pp. 23–37, 1995.

[69] Y. Freund and R. E. Schapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,*, Journal of Computer and System Sciences **55** (1997), no. 1 119 – 139.

[70] J. Friedman, T. Hastie and R. Tibshirani, *Special Invited Paper. Additive Logistic Regression: A Statistical View of Boosting*, The Annals of Statistics **28** (2000), no. 2 pp. 337–374.

[71] J. H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, The Annals of Statistics **29** (2001), no. 5 pp. 1189–1232.

[72] J. H. Friedman, *Stochastic gradient boosting*, Computational Statistics & Data Analysis **38** (2002), no. 4 367 – 378.

[73] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: data mining, inference and prediction. Springer, 2 ed., 2009.

[74] D. Fadda, E. Slezak and A. Bijaoui, *Density estimation with non-parametric methods*, A&A **127** (Jan., 1998) 335–352 [arXiv:astro-ph/9704096].

[75] N. Bissantz, L. Dömbgen, H. Holzmann and A. Munk, *Non-parametric confidence bands in deconvolution density estimation*, Journal of the Royal Statistical Society **69** (May, 2007) 483–506.

[76] Y. Ascasibar, *Estimating multidimensional probability fields using the Field Estimator for Arbitrary Spaces (FiEstAS) with applications to astrophysics*, Computer Physics Communications **181** (Aug., 2010) 1438–1443 [arXiv:1006.1296].

[77] M. Balogh et. al., *Galaxy ecology: groups and low-density environments in the SDSS and 2dFGRS*, MNRAS **348** (Mar., 2004) 1355–1372 [arXiv:astro-ph/0311379].

[78] I. Valtchanov, M. Pierre, J. Willis, S. Dos Santos, L. Jones, S. Andreon, C. Adami, B. Altieri, M. Bolzonella, M. Bremer, P. Duc, E. Gosset, C. Jean and J. Surdej, *The XMM-LSS survey. First high redshift galaxy clusters: Relaxed and collapsing systems*, A&A **423** (Aug., 2004) 75–85 [arXiv:astro-ph/0305192].

[79] I. R. Carstairs, A. J. Court, A. J. Dean, N. A. Dipper, M. J. Gorrod, R. A. Lewis, A. Bazzano, P. P. Maggioli, F. Perotti and E. Quadrini, *Kernel density estimators applied to fast timing hard X-ray*

observations of the Crab pulsar, Advances in Space Research **11** (1991) 95–99.

[80] O. C. de Jager, B. C. Raubenheimer and J. W. H. Swanepoel, *Kernel density estimations applied to gamma ray light curves*, A&A **170** (Dec., 1986) 187–196.

[81] B. P. Roe, H. Yang, J. Zhu, Y. Liu, I. Stancu and G. McGregor, *Boosted decision trees as an alternative to artificial neural networks for particle identification*, Nuclear Instruments and Methods in Physics Research A **543** (May, 2005) 577–584 [arXiv:physics/0408124].

[82] D. W. Gerdes, A. J. Sypniewski, T. A. McKay, J. Hao, M. R. Weis, R. H. Wechsler and M. T. Busha, *ArborZ: Photometric Redshifts Using Boosted Decision Trees*, AJ **715** (June, 2010) 823–832 [arXiv:0908.4085].

[83] B. D. Johnson and A. P. S. Crotts, *Photometric Identification of Type Ia Supernovae at Moderate Redshift*, AJ **132** (Aug., 2006) 756–768 [arXiv:astro-ph/0511377].

[84] N. V. Kuznetsova and B. M. Connolly, *A Probabilistic Approach to Classifying Supernovae Using Photometric Information*, ApJ **659** (Apr., 2007) 530–540 [arXiv:astro-ph/0609637].

[85] D. Poznanski, D. Maoz and A. Gal-Yam, *Bayesian Single-Epoch Photometric Classification of Supernovae*, AJ **134** (Sept., 2007) 1285–1297 [arXiv:astro-ph/0610129].

[86] S. A. Rodney and J. L. Tonry, *Fuzzy Supernova Templates. I. Classification*, ApJ **707** (Dec., 2009) 1064–1079 [0910.3702].

[87] M. Sako, B. Bassett, B. Connolly, B. Dilday, H. Campbell, J. Frieman, L. Gladney, R. Kessler, H. Lampeitl, J. Marriner, R. Miquel, R. Nichol, D. Schneider, M. Smith and J. Sollerman,

Photometric SN Ia Candidates from the Three-Year SDSS-II SN Survey Data, ArXiv e-prints (July, 2011) [1107.5106].

[88] K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. Allende Prieto, D. An, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall and et al., *The Seventh Data Release of the Sloan Digital Sky Survey*, ApJS **182** (June, 2009) 543–558 [0812.0649].

[89] J. W. Richards, D. Homrighausen, P. E. Freeman, C. M. Schafer and D. Poznanski, *Semi-supervised Learning for Photometric Supernova Classification*, arXiv:1103.6034.

[90] M. Sullivan et. al., *Photometric Selection of High-Redshift Type Ia Supernova Candidates*, AJ **131** (Feb., 2006) 960–972 [arXiv:astro-ph/0510857].

[91] M. Sako et. al., *The Sloan Digital Sky Survey-II Supernova Survey: Search Algorithm and Follow-Up Observations*, AJ **135** (Jan., 2008) 348–373 [0708.2750].

[92] S. Bickel, M. Brückner and T. Scheffer, *Discriminative learning for differing training and test distributions*, in ICML '07: Proceedings of the 24th international conference on Machine learning, (New York, NY, USA), pp. 81–88, ACM, 2007.

[93] W. Fan, I. Davidson, B. Zadrozny and P. S. Yu, *An improved categorization of classifiers sensitivity on sample selection bias*, in Proceedings of the Fifth IEEE International Conference on Data Mining,, 2005.

[94] B. Zadrozny, *Learning and evaluating classifiers under sample selection bias*, in Proceedings of the twenty-first international conference on Machine learning, ICML '04, (New York, NY, USA), pp. 114–, ACM, 2004.

[95] C. M. Bishop, Neural Networks for Pattern Recognition. Oxford University Press, USA, 1st ed., Jan., 1996.

[96] C. Elkan, *The Foundations of Cost-Sensitive Learning*, in Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, pp. 973–978, 2001.

[97] Y. Freund, H. S. Seung, E. Shamir and N. Tishby, *Selective sampling using the Query by Committee algorithm*, in Machine Learning, pp. 133–168, 1997.

[98] H. Shimodaira, *Improving predictive inference under covariate shift by weighting the log-likelihood function*, Journal of Statistical Planning and Inference **90** (Oct., 2000) 227–244 [http://dx.doi.org/10.1016/S0378-3758(00)00115-4].

[99] C. Cortes, M. Mohri, M. Riley and A. Rostamizadeh, *Sample selection bias correction theory*, CoRR (2008) [arXiv:0805.2775].

[100] D. W. Hogg, I. K. Baldry, M. R. Blanton and D. J. Eisenstein, *The K correction*, arXiv:astro-ph/0210394.

Appendix A

A1. Cross-validation

Cross-validation is a statistical technique that enables one to tune model parameters so as to optimize model prediction. Within the context of the 21D KDE, both the kernel bandwidth h, the number of nearest neighbours k and the odds threshold may be optimized for some figure of Merit (FoM) by tenfold cross-validation. This entails partitioning the training set into 10 roughly equal parts. One may then use nine-tenths of the data to estimate the Ia and non-Ia probability densities and then use these probability densities to classify the remaining one-tenth of the training set. This step may be repeated ten times, predicting the class for each of the ten partitions of the data using the KDEs estimated from the remaining nine partitions. Since we know the SN types of the training set, we can then find a combination of the aforementioned three parameters that maximizes the FoM. Cross-validation can be used in a similar way for boosting. Figure A.4 uses cross-validation to determine that 4000 trees will be near optimal.

A2. Probabilistic interpretation and combination of probabilities

By evaluating each KDE, we may obtain the probability of observing a lightcurve (with the lightcurve data denoted as x) conditioned on the SN being a Ia or not, i.e. we get $p_1 = \mathrm{P}(x|\mathrm{Ia})$ and $p_2 = \mathrm{P}(x|\text{non-Ia})$. The ratio

of p_1 to p_2 is known as the *Bayes factor*, B_{12}. What interests us, however, is the relative probability of the observation x being from a SNIa versus another type. That relative probability, called the *Odds ratio*, $odds(x)$ is

$$\begin{aligned} \mathrm{P}(\mathrm{Ia}|x) &= p_1 \frac{\mathrm{P}(\mathrm{Ia})}{\mathrm{P}(x)} && \text{(A.1)} \\ \mathrm{P}(\text{non-Ia}|x) &= p_2 \frac{\mathrm{P}(\text{non-Ia})}{\mathrm{P}(x)} && \text{(A.2)} \\ odds(x) &= \frac{\mathrm{P}(\mathrm{Ia}|x)}{\mathrm{P}(\text{non-Ia}|x)} = \frac{p_1 \mathrm{P}(\mathrm{Ia})}{p_2 \mathrm{P}(\text{non-Ia})} && \text{(A.3)} \\ &= B_{12} \frac{\mathrm{P}(\mathrm{Ia})}{\mathrm{P}(\text{non-Ia})}. \end{aligned}$$

The probabilities P(Ia) and P(non-Ia) are the prior probabilities to observe a Ia supernova or one of another type respectively.

To convert the relative probability back into absolute probabilities we need use the fact that there are only two possibilities (Ia or not), so that $P_2 = (1 - P_1)$. In this case we have that

$$P_1 = odds(x)/(1 + odds(x)). \tag{A.4}$$

If we have two independent observations x and y then we can update the relative probability $odds(x)$ from observation x:

$$odds(x, y) = odds(x) \frac{\mathrm{P}(y|\mathrm{Ia})}{\mathrm{P}(y|\text{non-Ia})}. \tag{A.5}$$

We can use this to combine for example the probability from the 21 dimensional KDEs with information from the Hubble diagram, but we have to be careful if the 21D KDEs already contain some of the Hubble information implicitly, e.g. through the evolution of the overall amplitudes of the lightcurves as a function of redshift.

It is possible that the KDEs should not be interpreted as probabilities. This may be due to oversmoothing of too wide kernels, or shot noise from too narrow kernels. With a sufficiently large training set one can test how

accurately the KDEs represent probabilities - the proportion of SNeIa in a (calculated) *odds* bin should equal that predicted by Eq. A.4. If it is not, one can consider making a mapping from the calculated *odds* to the true *odds*.

If in combining probabilities one does not want to assume independence, or does not trust the probabilities and doesn't want to make a mapping to true probabilities, there are several alternatives to Eq. A.5. Some of these include capping unreliable *odds* at 1, using linear combinations of *odds* instead of products, using p-values instead of probabilities and down-weighting particularly small/large Bayes' factors. Often an optimal method can be decided on by considering a scatter plot (like Figure 2.22) of the training set. In Section 2.3.5 we considered two new ways of combining *odds*, equations 2.11 and 2.12.

A3. Best trees

Suppose we have some data $\vec{X}_i \in R^2$, $z_i \in R$, and we would like to fit $z_i \sim \vec{X}$ using a tree. To be precise, we would like to find a tree which minimises $\sum_{i=1}^{n}(T(\vec{X}_i) - z_i)^2$, where $T(\vec{X}_i) = v_k$ when $\vec{X}_i$ falls into node k of the tree. We therefore need to find two things, the tree shape and the "leaf" values (the v_ks). Figure A.1 illustrates the idea of "greedy" tree construction. Note that tree show shown here may not be *the* best depth 3 tree. The greedy approach ignores several potential trees. However it is quick and easy, and for boosting where thousands of trees are made it is not necessary to have exactly minimising trees at each step. See also our animation of tree construction on the arxiv at [65].

A4. Calculating parameter importance

To measure how much information each parameter carries in the boosting classifier, we can do the following. For each branching within each tree constructed from the training data, calculate how much total ingroup variance was reduced by this branching. Then for each parameter, for all the branch-

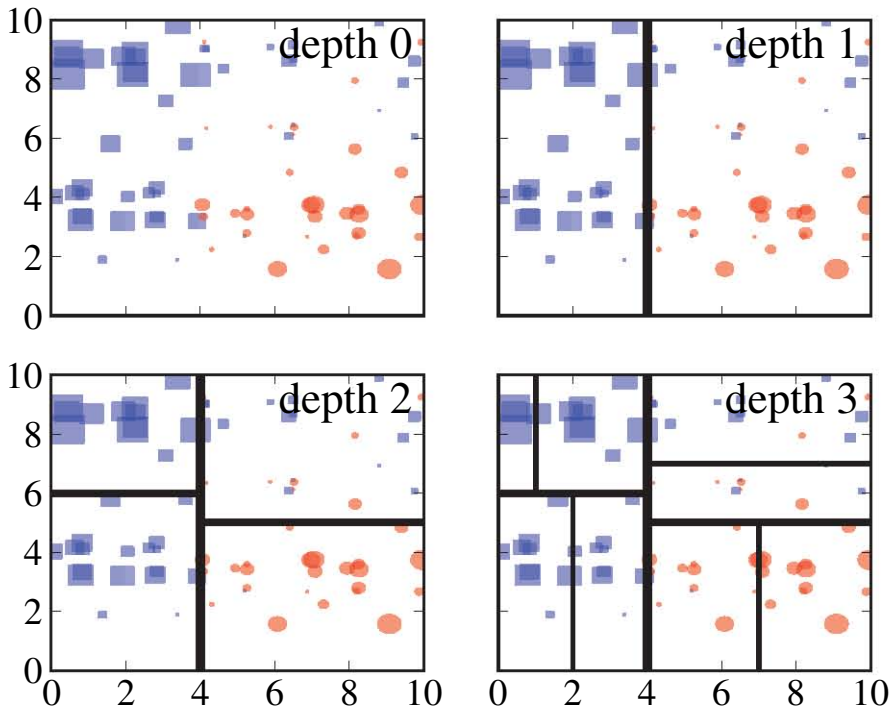

Figure A.1: (*Above left*) At several $\vec{X}_i \in R^2$ we have a value $z_i \in R$, represented by a rectangle if negative and a circle if positive, with the size of the shape being proportional to the magnitude of z_i. We want to split the set of observations by $X^{(1)}$ or $X^{(2)}$ to minimise the average ingroup variance. (*Above right*) After considering all vertical and horizontal lines, we settle on this vertical line as our first "branching" as it minimizes ingroup variance. (*Below left*) Sub-branches are chosen to minimize ingroup variance. (*Below right*) A tree of depth 3.

ings which it defines, add up the ingroup variance reductions. This value is a good indicator of a parameter's importance in classification. This has been done to create the two boosting parameter importance figures: A.2 and A.3.

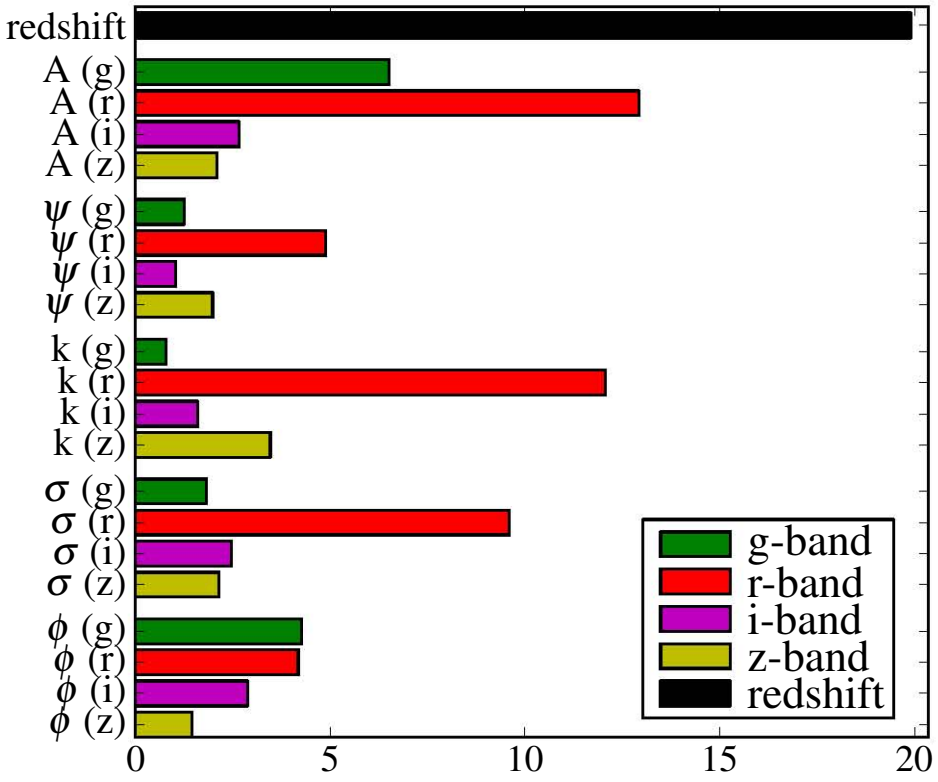

Figure A.2: The importance of parameters in distinguishing Ia from non-Ia in the non-representative training sample.

A5. GBM in full

In this appendix we complete the GBM algorithm presented in Section 2.2.2. There was no mention in Section 2.2.2 of the learning rate ν, or the bagging fraction ϕ. The learning rate $\nu \in [0, 1]$ should appear in step 4 of the main loop. Originally given as,

$$F_k \leftarrow F_{k-1} + T_k$$

step 4 should appear as,

$$F_k \leftarrow F_{k-1} + \nu T_k$$

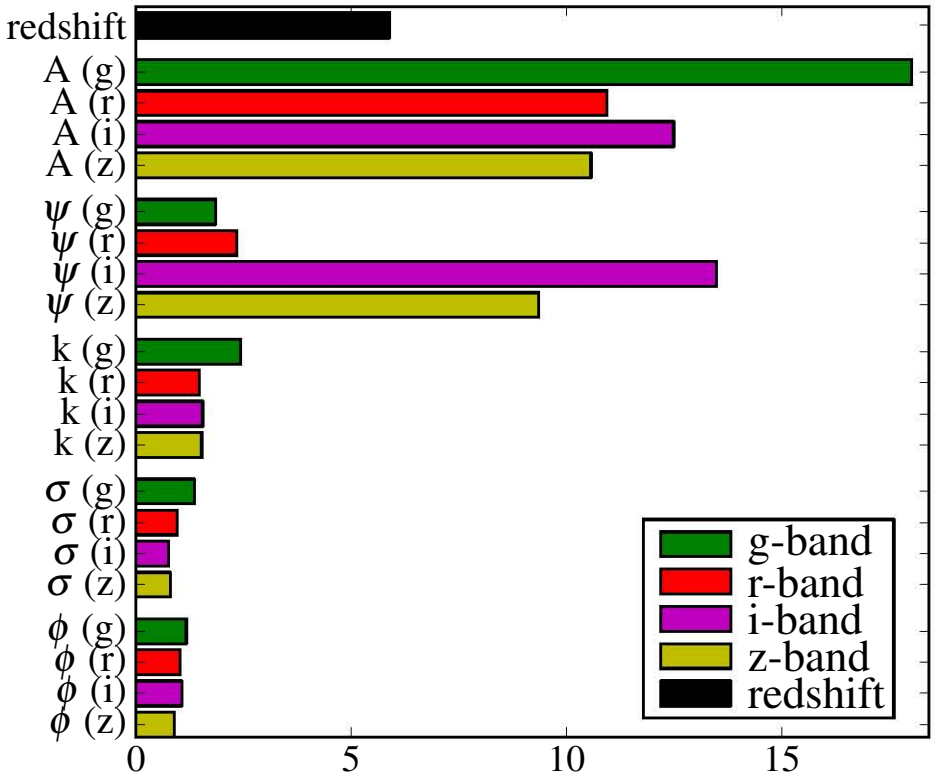

Figure A.3: The importance of parameters in distinguishing non-representative training (with spectrum) from unclassified (without spectrum) non-Ia SNe using boosting.

The learning rate should be set quite low, we used 0.05. It acts to reduce the sensitivity of F to the initial tree choice.

The use of bagging has been shown to improve the efficiency of the GBM algorithm and the accuracy of the final classifier [72]. The idea of bagging is that instead of all the training data being used for every tree construction, a fraction (ϕ) is randomly chosen to fit the tree at each step. For the SNPCC we used $\phi = 0.5$. To include bagging, the inner `for` loop should be modified to read,

for i in {sample of size $\phi \cdot N$ from integers 1 to N}

The last modification that needs to be made to complete the GBM algorithm is at step 3 of the main loop. Full line searches for optimal $\gamma_{k,j}$'s are not used, instead to speed up the algorithm only the initial step of Newton's method is used:

$$\gamma_{k,j} = \frac{\sum_{\vec{X}_i \in R_{k,j}} z_i}{\sum_{\vec{X}_i \in R_{k,j}} |z_i| \left(2 - |z_i|\right)} \tag{A.6}$$

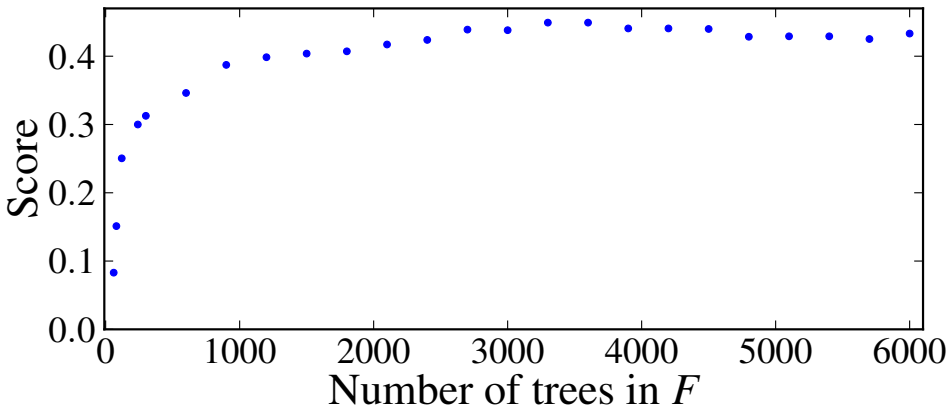

Figure A.4: The score, predicted using tenfold validation, on the representative training sample (1.0). There seems to be no overfitting as more trees are added. We go with number of trees = 4000.

A6. Parameter distributions

In this appendix we examine how the five lightcurve fitting parameters and redshift differ between Ia's and non-Ia's, and between training and unclassified SNe. Ia SNe cumulative frequency lines are red and thick, while non-Ia SNe are blue and thin. The cumulative frequency lines for training SNe are dotted, while the cumulative frequency lines for the unspecified SNe are solid. This appendix comprises Figures A.5 to A.9.

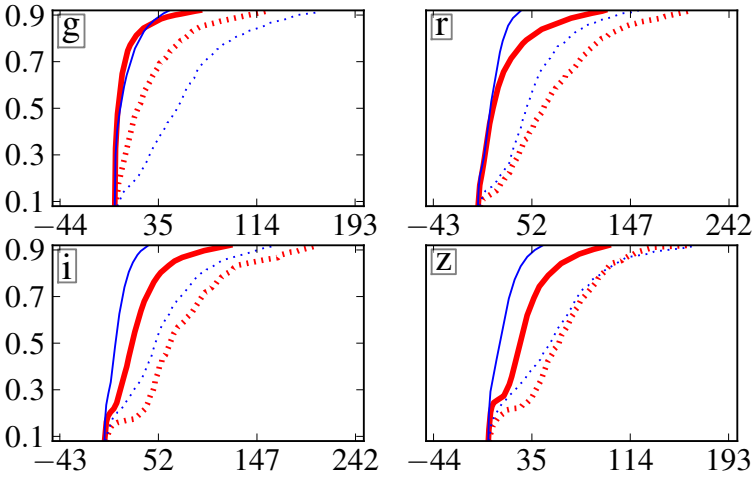

Figure A.5: Cumulative plots of parameter A in bands g,r,i,z. Non-representative training (dashed) vs unclassified (solid) and Ia (red, thick) vs non-Ia (blue, thin). In all bands, the magnitude A of SNe is far larger in the non-representative training set than in the unclassified set.

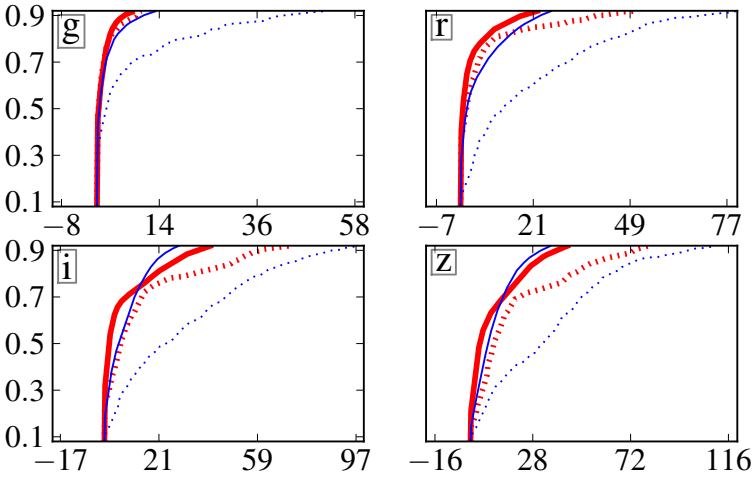

Figure A.6: Cumulative plot of parameter tail in bands g,r,i,z. Non-representative training (dashed) vs unclassified (solid) and Ia (red, thick) vs non-Ia (blue, thin).

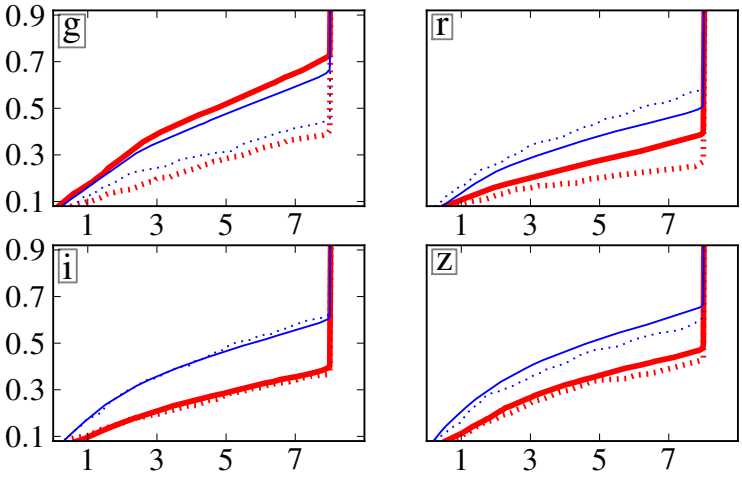

Figure A.7: Cumulative plot of parameter k in bands g,r,i,z. Non-representative training (dashed) vs unclassified (solid) and Ia (red, thick) vs non-Ia (blue, thin).

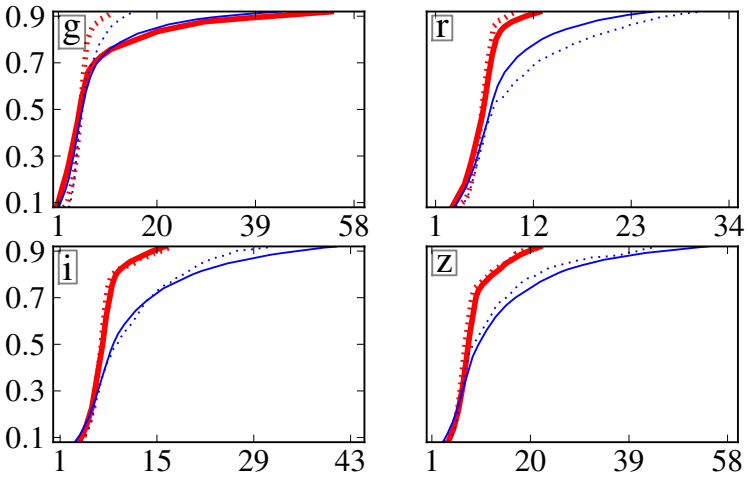

Figure A.8: Cumulative plot of parameter σ in bands g,r,i,z. Non-representative training (dashed) vs unclassified (solid) and Ia (red, thick) vs non-Ia (blue, thin).

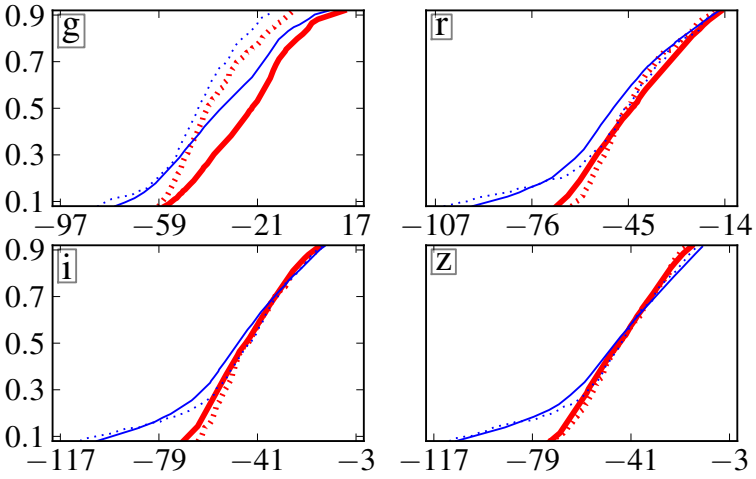

Figure A.9: Cumulative plot of parameter ϕ in bands g,r,i,z. Non-representative training (dashed) vs unclassified (solid) and Ia (red, thick) vs non-Ia (blue, thin).

A7. Random SNe

This appendix consists of a random selection of unclassified Ia and non-Ia SNe and their boosting values from representative training. Also, had a threshold of zero been used on the boosting value, would the classification have been correct (✓) or incorrect (✗). An extension of this appendix (200 SNe) can be found online at [65]. This appendix contains Figures A.10 to A.19

A8. Posterior Type Probabilities

We here derive the posterior type probabilities based on the modifications of Section 3.1. The posterior type probability will be derived, conditional on $\boldsymbol{D}$ and $\boldsymbol{P}$. This derivation can be easily extended to posterior type probabilities conditional on $\boldsymbol{D}, \boldsymbol{F}$ and $\boldsymbol{P}$.

$$f_{T_i|\boldsymbol{D},\boldsymbol{P}}(A|\boldsymbol{d},\boldsymbol{p})$$

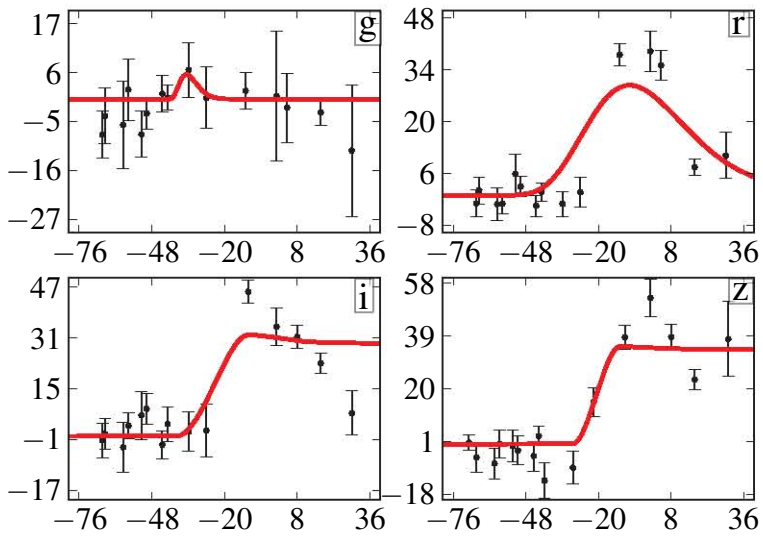

Figure A.10: Ia SN at z = 0.64. Boosting value of 1.78. ✓

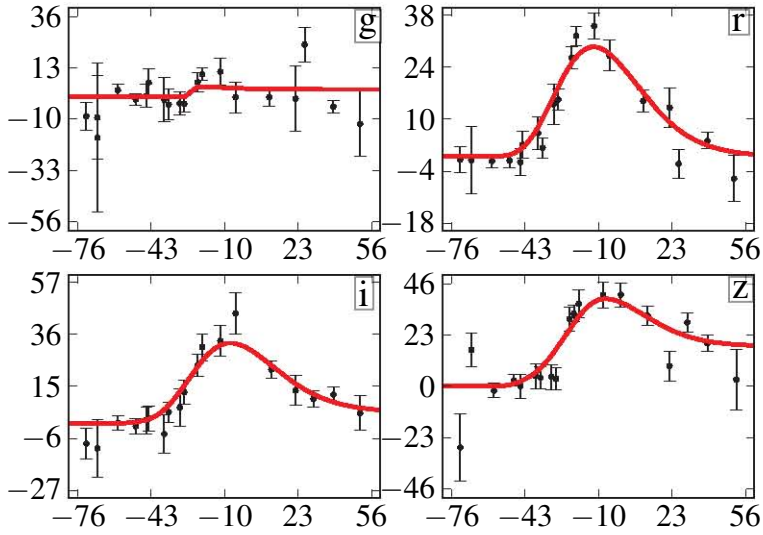

Figure A.11: Ia SN at z = 1.08. Boosting value of 4.29. ✓

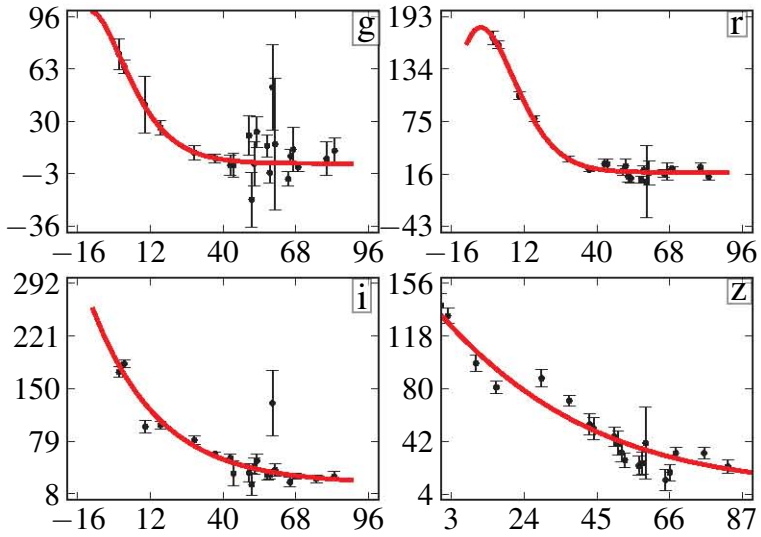

Figure A.12: Ia SN at z = 0.439. Boosting value of 1.15. ✓

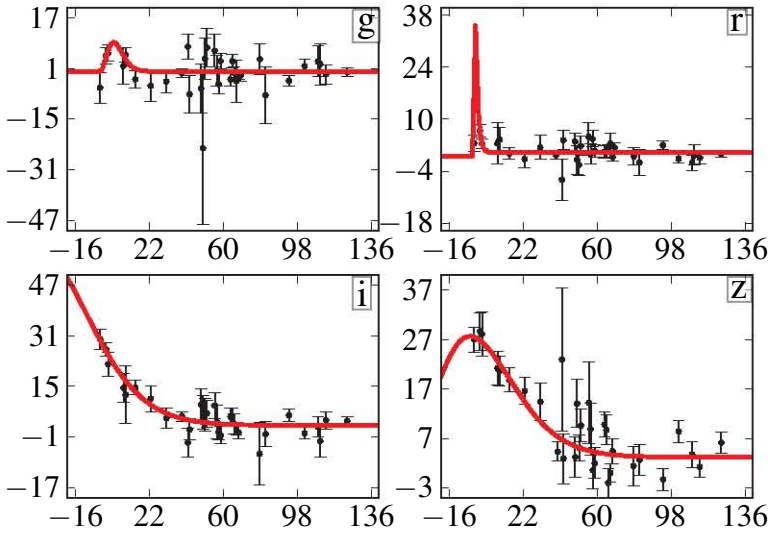

Figure A.13: Ia SN at z = 1.01. Boosting value of 5.94. ✓

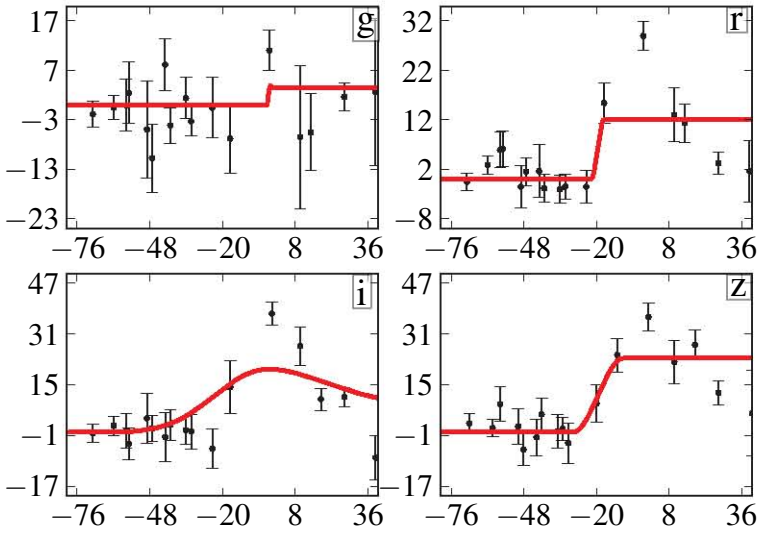

Figure A.14: Ia SN at z = 0.692. Boosting value of -0.869. ✗

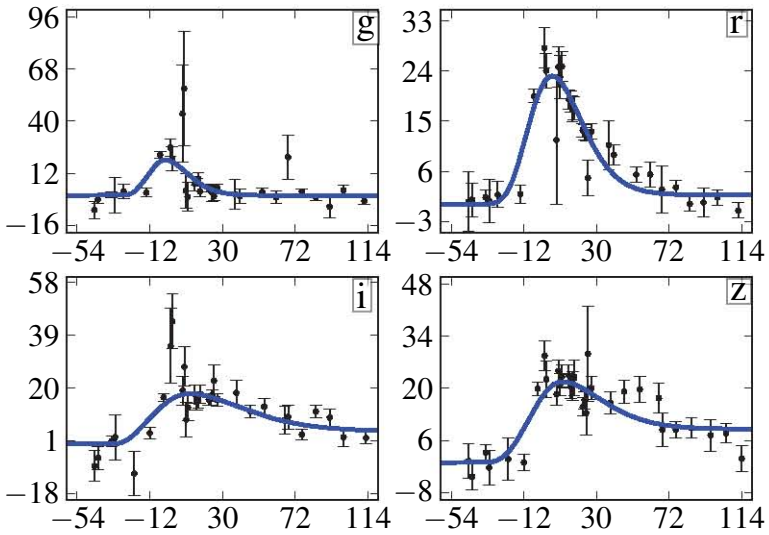

Figure A.15: non-Ia SN at z = 0.578. Boosting value of -5.79. ✓

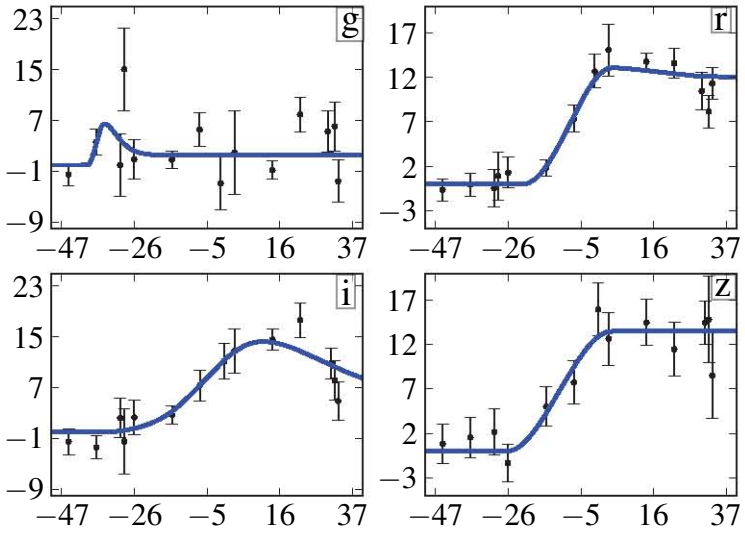

Figure A.16: non-Ia SN at z = 0.674. Boosting value of -3.17. ✓

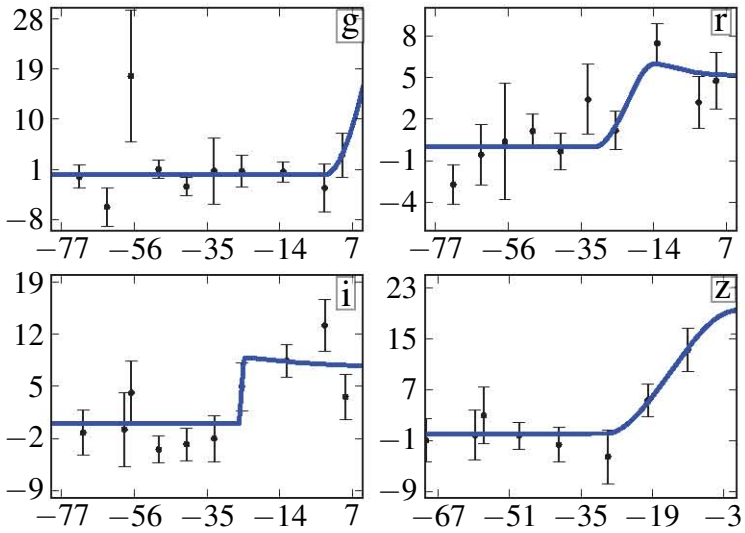

Figure A.17: non-Ia SN at z = 1.04. Boosting value of -1.13. ✓

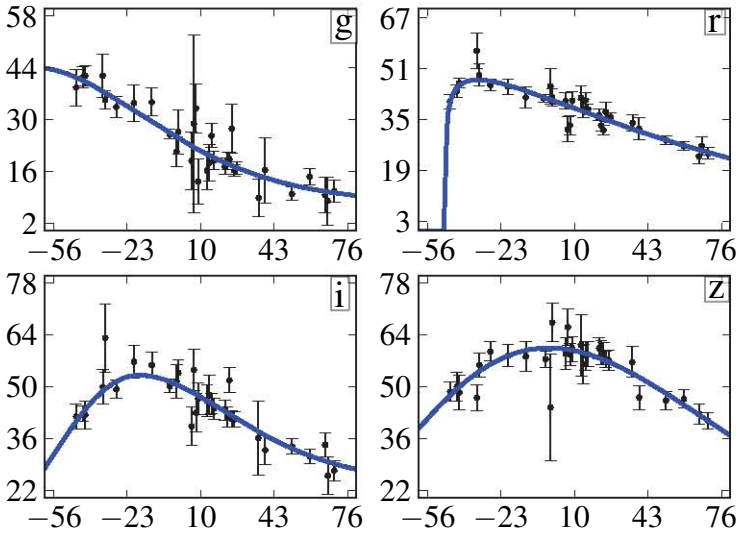

Figure A.18: non-Ia SN at z = 0.722. Boosting value of -3.33. ✓

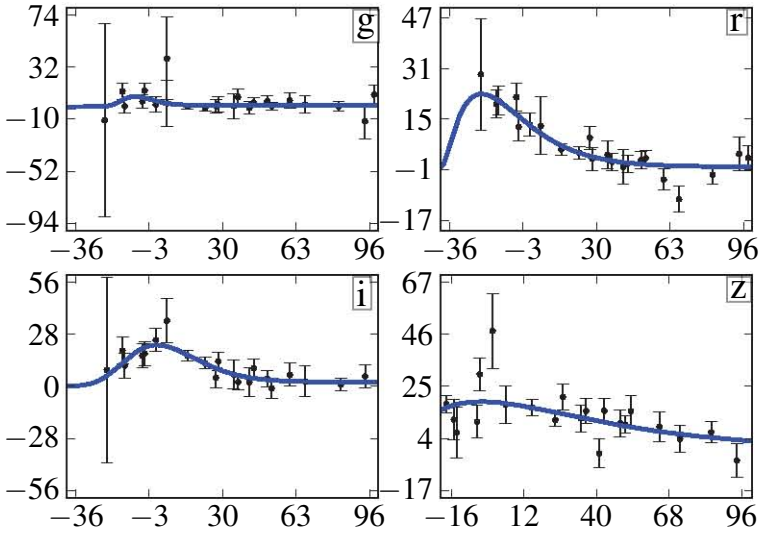

Figure A.19: non-Ia SN at z = 0.688. Boosting value of 0.52. ✗

$$= \int_\theta f_{T_i|\Theta,\boldsymbol{D},\boldsymbol{P}}(A|\theta,\boldsymbol{d},\boldsymbol{p}) f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p})\, d\theta$$
$$= \int_\theta f_{T_i|\Theta,D_i,P_i}(A|\theta,d_i,p_i) f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p})\, d\theta$$

$\rightarrow$ we have assumed that the objects are independent,

$$= \int_\theta \frac{f_{D_i|\Theta,P_i,T_i}(d_i|\theta,p_i,A) f_{T_i|\Theta,P_i}(A|\theta,p_i)}{f_{D_i|\Theta,P_i}(d_i|\theta,p_i)} \times$$
$$\times \quad f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p})\, d\theta$$

$\rightarrow$ we have used Bayes' Theorem,

$$= \int_\theta \left(\frac{A_i}{A_i + B_i}\right) f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p})\, d\theta. \tag{A.7}$$

$\rightarrow$ where $A_i = \mathrm{P}(d_i|\theta,p_i,T_i=A)p_i$, $B_i = \mathrm{P}(d_i|\theta,p_i,\ T_i=B)(1-p_i)$, and we have assumed used that $f_{T_i|\Theta,P_i}(A|\theta,p_i) = p_i$.

If the posterior $f_{\Theta|\boldsymbol{D},\boldsymbol{P}}$ confines θ to a region sufficiently small such that A_i and B_i are approximately constant, then the posterior type probability (A.7) is well approximated by $A_i(\hat{\theta})/\left(A_i(\hat{\theta}) + B_i(\hat{\theta})\right)$ where $\hat{\theta}$ is the maximum likelihood estimator of $f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p})$. Furthermore, the posterior odds ratio,

$$\text{posterior odds ratio} \stackrel{\text{def}}{=} \frac{f_{T_i|\boldsymbol{D},\boldsymbol{P}}(A|\boldsymbol{d},\boldsymbol{p})}{f_{T_i|\boldsymbol{D},\boldsymbol{P}}(B|\boldsymbol{d},\boldsymbol{p})}$$

can be shown to be given by the prior odds ratio multiplied by the Bayes Factor,

$$\text{posterior odds ratio} = \left(\frac{p_i}{1-p_i}\right) \times \left(\frac{f_{D_i|\Theta,P_i,T_i}(d_i|\hat{\theta},p_i,A)}{f_{D_i|\Theta,P_i,T_i}(d_i|\hat{\theta},p_i,B)}\right).$$

A9. Additional conditioning on the confirmation of supernova type

In chapter 3 we did not distinguished between the contributions of unconfirmed and confirmed objects to the posterior. While we can calculate approximate τ_A-probabilities for confirmed objects, these values should not enter the posterior, but be replaced by 0 (if type B) or 1 (if type A). Let us introduce the random variable F to denote whether an object is confirmed, so that $F = 1$ if confirmed and $F = 0$ if unconfirmed. With F introduced, we wish to replace the τ_A-probabilities $\boldsymbol{p}$ by $\bar{\boldsymbol{p}}$, where,

$$\bar{p}_i = \begin{cases} p_i & \text{if } f_i = 0, \\ 1 & \text{if } f_i = 1 \text{ and } \tau_i = A, \\ 0 & \text{if } f_i = 1 \text{ and } \tau_i = B. \end{cases}$$

We must be careful to let the new information which we introduce in $\bar{\boldsymbol{p}}$ be absorbed elsewhere in the posterior. To this end, as we did in the case without F, we start afresh the posterior derivation, explicitly including the vector $(\boldsymbol{f})$ which describes which objects have been followed-up. By this strategy, we arrive at the following posterior distribution

$$f_{\Theta|\boldsymbol{D},\boldsymbol{F},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{f},\boldsymbol{p}) \propto f_{\Theta}(\theta) \times \sum_{\boldsymbol{\tau}} f_{\boldsymbol{D}|\Theta,\boldsymbol{F},\boldsymbol{P},\boldsymbol{T}}(\boldsymbol{d}|\theta,\boldsymbol{f},\boldsymbol{p},\boldsymbol{\tau}) \prod_{\tau_i=A} \bar{p}_i \prod_{\tau_j=B} (1-\bar{p}_j). \tag{A.8}$$

The new information $(\boldsymbol{f})$ has been absorbed into the likelihood, $f_{\boldsymbol{D}|\ldots}$. For a particular application, one may now ask if the addition of $\boldsymbol{F}$ in $f_{\boldsymbol{D}|\ldots}$ is necessary. We have already mentioned that for SNe $\boldsymbol{D}|\theta,\boldsymbol{T}$ is unlikely to be independent of $\boldsymbol{P}$. It is also unlikely that $\boldsymbol{D}|\theta,\boldsymbol{T}$ is independent of $\boldsymbol{F}$, as bright SNe, which have lower fitted distance moduli at a given redshift, are confirmed more regularly than faint ones. However, it is possible that by additionally conditioning $\boldsymbol{D}$ on $\boldsymbol{P}$ this confirmation dependence is broken, so that $\boldsymbol{D}|\theta,\boldsymbol{P},\boldsymbol{T}$ and $\boldsymbol{F}$ are independent. We leave this possibility as an

open question.

In the case of independent SNe, the posterior (A.8) reduces to

$$f_{\Theta|\boldsymbol{D},\boldsymbol{F},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{f},\boldsymbol{p}) \propto \prod_{i=1}^{N} \big[\, f_{D_i|\Theta,F_i,P_i,T_i}\,(d_i|\theta,f_i,p_i,A)\bar{p}_i \;+ \\ f_{D_i|\Theta,F_i,P_i,T_i}(d_i|\theta,f_i,p_i,B)\;(1-\bar{p}_i)\,\big]. \tag{A.9}$$