View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by RERO DOC Digital Library

Mon. Not. R. Astron. Soc. **421,** 913–925 (2012) doi:10.1111/j.1365-2966.2011.20147.x

Parameter estimation with Bayesian estimation applied to multiple species in the presence of biases and correlations

J. Newling,[1,2★] B. Bassett,[1,2,3] R. Hlozek,[4] M. Kunz,[5] M. Smith[2,6] and M. Varughese[7]

[1]*Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch 7701, South Africa*
[2]*African Institute for Mathematical Sciences, 6-8 Melrose Road, Muizenberg 7945, South Africa*
[3]*South African Astronomical Observatory, PO Box 9, Observatory 7935, South Africa*
[4]*Department of Astrophysics, Oxford University, Oxford OX1 3RH*
[5]*Département de Physique Théorique, Université de Genève, Genève CH1211, Switzerland*
[6]*Astrophysics, Cosmology and Gravity Centre, University of Cape Town, Rondebosch 7701, South Africa*
[7]*Department of Statistical Sciences, University of Cape Town, Rondebosch 7701, South Africa*

Accepted 2011 November 7. Received 2011 November 5; in original form 2011 July 29

ABSTRACT

The original formulation of Bayesian estimation applied to multiple species (BEAMS) showed how to use a data set contaminated by points of multiple underlying types to perform unbiased parameter estimation. An example is cosmological parameter estimation from a photometric supernova sample contaminated by unknown Type Ibc and II supernovae. Where other methods require data cuts to increase purity, BEAMS uses all of the data points in conjunction with their probabilities of being each type. Here we extend the BEAMS formalism to allow for correlations between the data and the type probabilities of the objects as can occur in realistic cases. We show with simple simulations that this extension can be crucial, providing a 50 per cent reduction in parameter estimation variance when such correlations do exist. We then go on to perform 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 future photometric supernova surveys and BEAMS, which lead to specific recommendations for future supernova surveys.

Key words: methods: data analysis – methods: statistical – supernovae: general – cosmological parameters.

1 INTRODUCTION

Type Ia supernovae (SNeIa) provided the first widely accepted evidence for cosmic acceleration in the late 1990s (Riess et al. 1998; Perlmutter et al. 1999). While they were based on relatively small numbers of spectroscopically confirmed SNeIa, those results have since been confirmed by independent analyses of other data sets (Eisenstein et al. 2005; Percival et al. 2007, 2010; Fu et al. 2008; Giannantonio et al. 2008; Mantz et al. 2010; Komatsu et al. 2011).

Next-generation supernova (SN) surveys such as LSST will be fundamentally different, yielding thousands of high-quality candidates every night for which spectroscopic confirmation will probably be impossible. Creating optimal ways of using this excellent photometric data is a key challenge in SN cosmology for the coming decade. There are two ways in which one can imagine using photometric candidates. The first approach is to try to classify the candidates into Type Ia, Ibc or II SNe (Johnson & Crotts 2006; Kuznetsova & Connolly 2007; Poznanski, Maoz & Gal-Yam 2007; Rodney & Tonry 2009) and then use only those objects that are believed to be SNeIa above some threshold of confidence. This has recently been discussed by Sako et al. (2011), 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 an SNIa. This is the approach exemplified by the Bayesian estimation applied to multiple species (BEAMS) formalism, which accounts for the contamination from non-Ia (nIa) SN data using the appropriate Bayesian framework, as presented in Kunz, Bassett & Hlozek (2007, 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

★E-mail: james.newling@gmail.com

© 2012 The Authors
Monthly Notices of the Royal Astronomical Society © 2012 RAS

been done so in other fields. In this paper, 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 functions 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: $\boldsymbol{X} = X_{1,\ldots,N}$. An observation of X we will denote by x, so that the full set of N observations is denoted by $\boldsymbol{x} = x_{1,\ldots,N}$. For SNe, the observations $\boldsymbol{x}$ are the photometric data of the N SNe. As such, for SNe the probability density function (pdf) $f_{X|\Theta}(\boldsymbol{x}|\theta)$ 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 programme. The SDSS-II SN survey (Abazajian et al. 2009) 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 follow-up programme 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,

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

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 $\boldsymbol{p} = p_{1,\ldots,N}$ and we will call them τ_A-probabilities. The τ_A-probability for an SN is thus the probability of being type Ia, conditional on knowing the subset x_P of 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 realization denoted by τ. A summary of all the variables used in the paper is given in Table 1.

Attempts to approximate τ_A-probabilities include those of Poznanski et al. (2002), Newling et al. (2011) and Richards et al. (2011) and as implemented in SALT2 (Guy et al. 2007). 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 also that there is no sense in which one set of τ_A-probabilities is *the* correct set, as this 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 7. 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.

Table 1. A description of all the random variables used in this paper.

Random variables	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: $T \in \{\text{Ia, nIa}\}$.
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.
$\bar{P}$	$\bar{p}$	This is exactly P if the object is unconfirmed and 1 or 0 if confirmed, depending on type.

In 2009, the Supernova Photometric Classification Challenge (SNPCC) was run to encourage work on SN classification by light curves alone (Kessler et al. 2010b). 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 working 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 recognize SNeIa, but would also need to provide precise, unbiased probabilities of the object being an SNeIa.

In brief, this paper consists of three more or less independent parts. In Section 2, we present an extension of BEAMS to the case in which certain correlations, which were ignored in KBH, are present. In Section 3, we discuss the relevance of τ_A-probabilities in a broader context, and specifically their importance in BEAMS. Then, in Sections 4–6, we perform simulations to better understand the importance of sample sizes, nearness of population distributions, biases of τ_A-probabilities and decisiveness of τ_A-probabilities (to be defined). Finally, in Section 7 we present new ideas from the machine learning literature describing when and how τ_A-probability biases emerge and how to correct for them. This is then discussed in the context of the SNPCC in Section 8.

2 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

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

where the p_i are τ_A-probabilities. The summation is over all of the 2^N possible ways so that the N observations can be classified into two classes. We will refer to the expression on the right-hand side of equation (1) as the KBH 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 KBH posterior reduces,

$$\prod_{i=1}^{N}\left[\, f_{D_i|\Theta,T_i}(d_i|\theta,A)\, p_i + f_{D_i|\Theta,T_i}(d_i|\theta,B)\,(1-p_i)\right]. \tag{2}$$

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

$$f_{\boldsymbol{T}}(\boldsymbol{\tau}) = \prod_{\tau_i=A} p_i \prod_{\tau_i=B}(1-p_i). \tag{3}$$

Equation (3) states that the left-hand side prior probability of the SNe having types $\boldsymbol{\tau}$ is given by the product on the right-hand side 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:

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

(i) We will first use the definition of conditional probability to obtain

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

(ii) The term in the numerator can be then written as the sum over all 2^N possible type vectors

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

(iii) The numerator is again modified using the definition of conditional probability

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

(iv) We will now assume that the probability of having τ_A-probabilities and types $\boldsymbol{p}$ and $\boldsymbol{\tau}$, respectively, is independent of Θ. As noted after following equation (4) of KBH, for SNe this assumption rests on the fact that Θ (that is $\Omega_{\rm m}$, 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:

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

(v) Rearranging this, and again using the definition of conditional probability, we obtain

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

(vi) The first term on the above line 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 KBH posterior. Making this assumption, we arrive at

$$\propto f_{\Theta}(\theta) \sum_{\boldsymbol{\tau}} f_{\boldsymbol{D}|\Theta,\boldsymbol{P},\boldsymbol{T}}(\boldsymbol{d}|\theta,\boldsymbol{p},\boldsymbol{\tau}) \prod_{\tau_i=A} p_i \prod_{\tau_j=B}(1-p_j). \tag{4}$$

We will refer to the newly derived expression (4) as the full posterior. Let us now consider the difference between the KBH posterior (equation 1) and the full posterior, and note that in the full posterior, the likelihood of the data $\boldsymbol{D}$ is conditional on Θ, $\boldsymbol{P}$ and $\boldsymbol{T}$, whereas in the KBH posterior, $\boldsymbol{D}$ is only conditional on Θ and $\boldsymbol{T}$. This is the only difference between the two posteriors, and so when $\boldsymbol{D}|\Theta,\boldsymbol{T}$ is independent of $\boldsymbol{P}$, the posterior (equation 4) reduces to the KBH posterior (equation 1), making them equivalent. This is an important result: when $\boldsymbol{D}|\Theta,\boldsymbol{T}$ and $\boldsymbol{P}$ are independent, the KBH and full posteriors are the same.

Our results can be summarized 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 KBH posterior in equation (1) by

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

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 (equation 4). When $\boldsymbol{D}|\Theta,\boldsymbol{T}$ and $\boldsymbol{P}$ are independent, it reduces to the KBH posterior (equation 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 an SN of unknown type adds no information to the estimation of the distance modulus'. Indeed it is difficult to imagine how this could be the case: we know that SNeIa are brighter than 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 an 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 (Guy et al. 2007). This 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 (Sullivan et al. 2006; Sako et al. 2008) 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

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

© 2012 The Authors, MNRAS **421,** 913–925
Monthly Notices of the Royal Astronomical Society © 2012 RAS

Newling et al. (2011) 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 6, 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 (equation 4) reduces to

$$f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p}) \propto \prod_{i=1}^{N}\Big[f_{D_i|\Theta,P_i,T_i}(d_i|\theta,p_i,A)p_i + f_{D_i|\Theta,P_i,T_i}(d_i|\theta,p_i,B)(1-p_i)\Big]. \tag{5}$$

In Section 7, 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. Finally, we mention an extension where whether or not a SN is followed-up is included in the cosmology posterior, with details in Appendix A. As an additional supplement to this section, in Appendix B we derive the posterior type probabilities.

3 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, then the expected total proportion of type A objects is

$$P(T=A) = \langle P\rangle = \int_0^1 p f_P(p)\,\mathrm{d}p. \tag{6}$$

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

$$\mathrm{FPR}(f_P) = P(P>c|T=B) = \frac{\int_c^1 (1-p) f_P(p)\,\mathrm{d}p}{\int_0^1 (1-p) f_P(p)\,\mathrm{d}p}, \tag{7}$$

and the false negative rate (FNR) is

$$\mathrm{FNR}(f_P) = P(P<c|T=A) = \frac{\int_0^c p\, f_P(p)\,\mathrm{d}p}{\int_0^1 p\, f_P(p)\,\mathrm{d}p}. \tag{8}$$

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 Fig. 1.

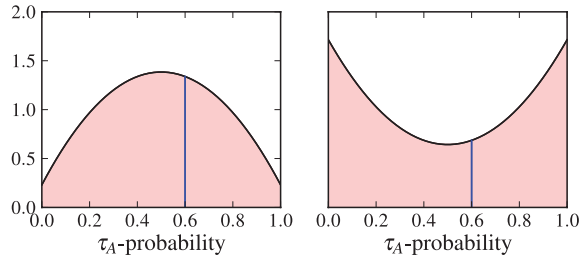

Figure 1. Two τ_A-probability distributions, both with means of 0.5. Using a threshold of 0.6, we have on the left-hand side FPR = 0.17 and FNR = 0.45, and on the right-hand side FPR = 0.15 and FNR = 0.28.

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

$$f^{\mathcal{P}}(p) = \frac{1}{2}\left(\delta_{\mathcal{P}}(p) + \delta_{1-\mathcal{P}}(p)\right), \tag{9}$$

where $\delta_{\mathcal{P}}$ and $\delta_{1-\mathcal{P}}$ are δ-functions centred 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 an observation of P from equation (9) is equivalent to drawing it from $\{1-\mathcal{P}, \mathcal{P}\}$ with equal probability:

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

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

KBH (p. 5) stated that the expected proportion of type A objects (equation 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

$$\sigma_\mu \propto \sqrt{\langle P\rangle N}. \tag{10}$$

It should be noted that the result from KBH (equation 10) is an asymptotic result in N. For small N, the decisiveness of the probabilities plays an important part. If equation (6) were the only factor determining the expected error (σ_μ), then $f^{0.5}$ would be equivalent to f^1 in terms of the expected error. This would mean that perfect type knowledge does not reduce the error, which would be surprising. An example in Section 4.1 illustrates that decisiveness does play a role in determining the error.

KBH (p. 8) mentioned that the effect of biases in τ_A-probabilities on BEAMS can be catastrophic. They consider the case in which 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 there is a real probability $p_i - a$ that it is type 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

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

transform the claimed priors ($\tilde{p}$) by

$$p = 0.2 + 0.6\,\tilde{p}. \tag{11}$$

Introducing a bias such as the one defined by equation (11) will have no effect on the optimal FPR and FNR, provided the probability threshold is chosen optimally. This is because equation (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 equation (11) does have an effect on BEAMS parameter estimation, as we show in Section 5. In Section 7 we discuss how to guarantee that the τ_A-probabilities are free of bias.

4 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, a distance modulus at a given redshift z_0 and $\boldsymbol{d}$, the fitted distance moduli of SNe at z_0. Furthermore, $\boldsymbol{D}|\Theta$, $\boldsymbol{T}$ and $\boldsymbol{P}$ will be independent, such that the KBH and full posterior are equivalent.

4.1 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

$$f_{D|T}(d, \tau) = \text{Normal}(\mu_\tau, 1), \tag{12}$$

where $\mu_A = -1$ and $\mu_B = +1$, as illustrated in Fig. 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 are 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 $f^{\mathcal{P}}$. Secondly, the type of the observation is chosen; with probability p it is chosen as A, and with probability $1 - p$ it is chosen as B. Finally, the data (d) are drawn from equation (12). Note that $D|T$ is independent of P and so the KBH 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 (FoM) to compare the performance with different sample sizes (N) and decisiveness ($\mathcal{P}$):

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

where $\hat{\mu}_A$ is the maximum likelihood estimate of μ_A using the KBH 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 Fig. 2 the performance of BEAMS for various (N, $\mathcal{P}$) combinations. A good approximation to the FoM h in Fig. 2 appears to be

$$h(N, \mathcal{P}) \approx N \left(0.32 + 1.44 \left(\mathcal{P} - \frac{1}{2}\right)^3\right), \tag{13}$$

although this 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 Fig. 2. This says that given a completely blind sample ($\mathcal{P} = 0.5$),

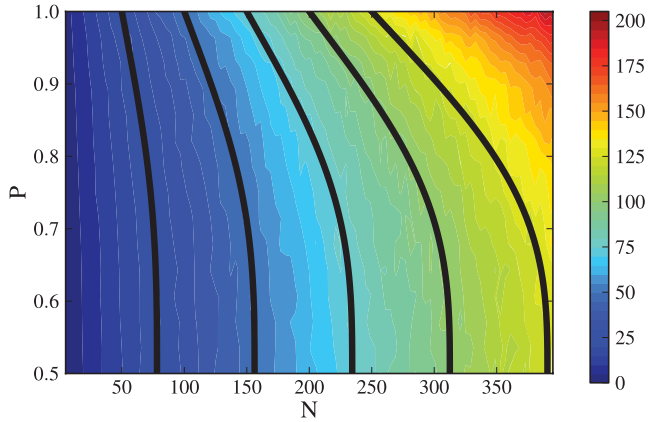

Figure 2. Contour plot of $h(N, \mathcal{P})$. The solid lines are approximations to lines of constant h of the form (13).

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 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.

4.2 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 Fig. 3. In this simulation, we leave both the population means as free parameters to be fitted for. 20 objects are drawn from the types A and B, with the τ_A-probabilities drawn from $f^{\mathcal{P}}$. The simulation is done with five different $\mathcal{P}$ values. The τ_A-probabilities are illustrated in Fig. 4, and the approximate shape of the posterior marginals of μ_A for each $\mathcal{P}$ value is illustrated in Fig. 5 by MCMC chain counts.

There are two interesting results from this simulation. First, there is a negligible difference in the performance of $\mathcal{P} = 1$ and 0.7, so that having a 30 per cent type uncertainty for all objects as opposed to absolute type knowledge does not weaken the results. Secondly, as $\mathcal{P}$ approaches 0.5, BEAMS still correctly locates the population means but is unsure which mean belongs to which population.

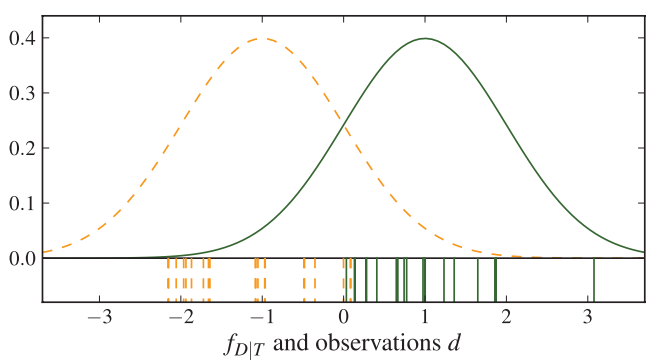

Figure 3. Above are the population A (left) and population B (right) distributions, with (for simulation in Section 4.2) the observed values of D drawn from these distributions shown as vertical lines beneath.

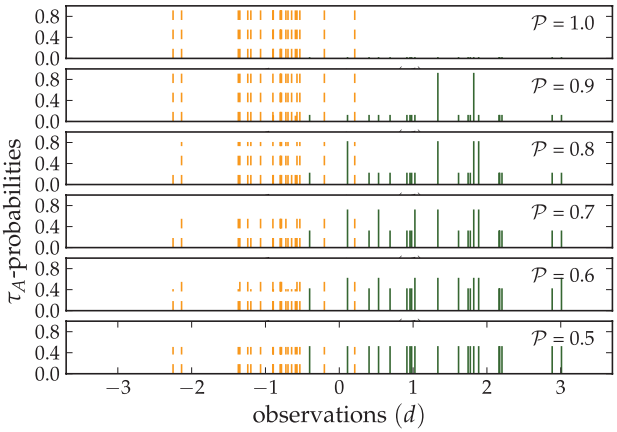

Figure 4. For values of $\mathcal{P}$ from 1 (above) to 0.5 (below), a τ_A-probability of $\mathcal{P}$ or $1 - \mathcal{P}$ is attached to each observation.

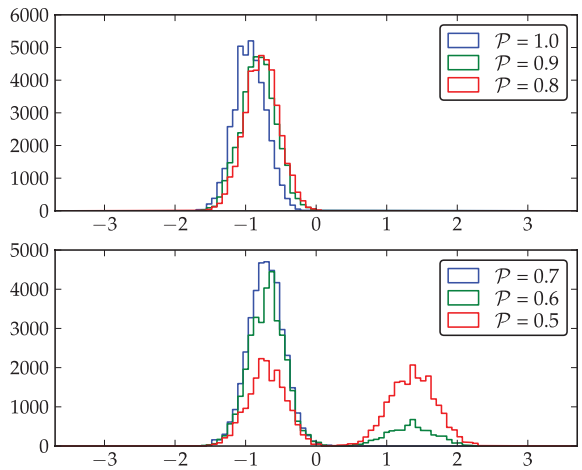

Figure 5. MCMC chain counts, approximating the posterior distributions of μ_A for the different values of decisiveness, $\mathcal{P}$.

5 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

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

Recall 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 such manner as follows.

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

(ii) $\mathcal{P} \overset{+-}{\Rightarrow} \{0.4, 0.6\}$. Here the decisiveness of the τ_A-probabilities is underestimated, so that $p = 0.8 \rightarrow p = 0.6$ and $p = 0.2 \rightarrow p = 0.4$.

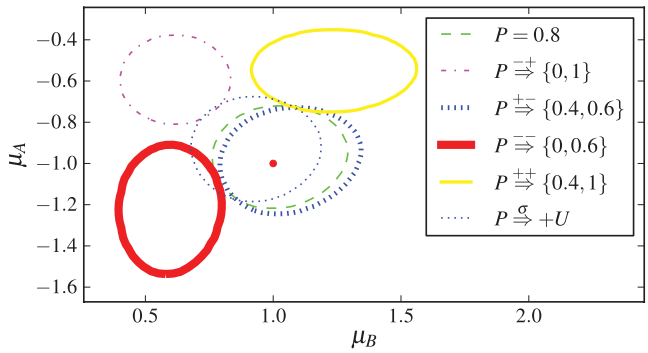

Figure 6. The 99 per cent posterior confidence regions using the five biasing numbers of the τ_A-probabilities, as described in Section 5.

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

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

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

The 99 per cent posterior confidence regions obtained using these biased τ_A-probabilities in a simulation of 400 points are illustrated in Fig. 6. The underestimation of decisiveness (item ii) has little effect on the final confidence region, but overestimating the τ_A-probability decisiveness (item i) 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 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 the 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 Fig. 6. Note 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 noticeable increases in the size of the 99 per cent confidence region.

The effect of the flat τ_A-probability shifts (items iii and iv) introduces biases larger than 4σ. This case was considered in KBH where, as 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 was the case in simulations summarized in table II of KBH, is that a flat τ_A-probability shift in confidence towards being type B (item iii) 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 in which BEAMS is not used.

Finally, we note that in this simulation the addition of unbiased noise to the τ_A-probabilities (item v) has an insignificant effect. This suggests that systematic biases should be the primary concern of future work on the estimation of τ_A-probabilities.

6 WHEN GIVEN TYPE, THE DATA ARE STILL DEPENDENT ON τ-PRIORS

In this section, we consider for the first time a simulation in which the data are 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 Fig. 7. To clarify the difference between this simulation and the previous ones, prior to this data were simulated as follows:

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

where at the last step, the data were 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 $f^{0.7}$,

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

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

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

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

$$f_{D|T}(d|A) = \text{Normal}\,(-1, 1) \tag{14}$$

and

$$f_{D|T}(d|B) = \text{Normal}\,(1, 1), \tag{15}$$

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

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

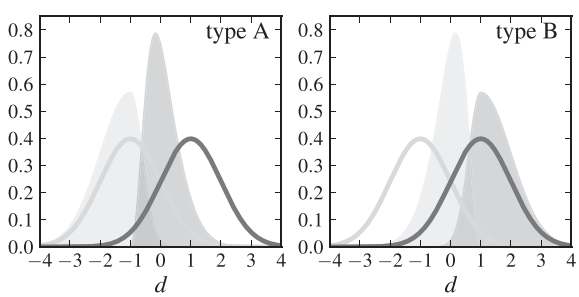

Figure 7. Plots of $f_{D|P,T}(d|p, \tau)$ (filled curves) for $p = 0.7$ (light) and $p = 0.3$ (dark), and for type A (left) and type B (right). Overlying are $f_{D|T}(d|A)$ (light) and $f_{D|T}(d|B)$ (dark).

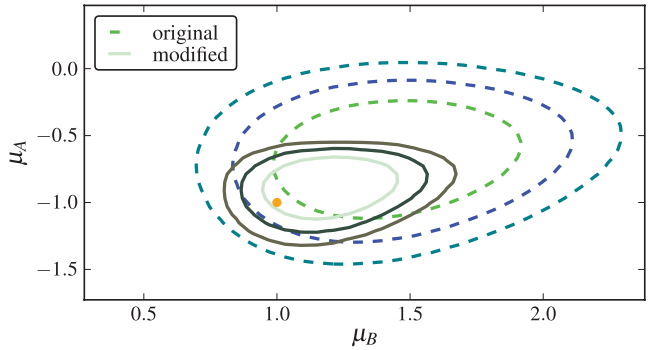

Figure 8. Posterior distributions on the parameters (μ_A, μ_B) using the correct posterior (equation 4) (solid) and the KBH posterior (equation 1) (dashed). The KBH posterior assumes independence between $D|T$ and P. Plotted are the 80, 95 and 99 per cent confidence levels. The true parameters (orange point) lie within the 95 per cent confidence regions of both posteriors.

where K is a normalizing constant. The marginal $f_{D|P,T}(d|0.3, A)$ is then constructed to guarantee (14). The above construction guarantees that the population of A objects with low τ_A-probabilities (0.3) lies on average closer to the B mean than that 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 Fig. 7.

To compare the use of the KBH BEAMS posterior (equation 2) with the full conditional posterior (equation 5), we randomly draw 40 data points from the above distribution and construct the respective posterior distributions, as illustrated in Fig. 8. Observe that the KBH posterior is significantly wider than the full posterior. Indeed, approximately half of the interior of the 80 per cent region of the KBH posterior is ruled out to 1 per cent by the full posterior. Note that while the KBH posterior is wider than the full posterior, it is not biased. This result goes against our intuition; we believed that the KBH 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 KBH posterior is used, or if there can exist dependencies between P and D for which the use of equation (1) leads to a bias, remains an open question.

Fig. 8 illustrates one realization from the distribution we have described, but repeated realizations show that on average, the variance in the maximum likelihood estimator using the KBH 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,

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

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.

7 OBTAINING UNBIASED τ_A-PROBABILITIES

In this section, we investigate likely sources of τ_A-probability biases such as those presented in Section 5, 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 2.

7.1 Selection bias

With respect to classification methods, selection bias refers to the situation in which the confirmed data are 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 Bickel, Brückner & Scheffer (2007). 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 an 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, certain 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 certain circumstances, as we will discuss in the following section.

7.1.1 Formalism

Following where possible the notation of Fan et al. (2005), in what follows we assume that variables (X, T, F) are drawn from $\mathcal{X} \times \mathcal{T} \times \mathcal{F}$, where

(i) $\mathcal{X}$ is the feature space,
(ii) $\mathcal{T} = \{A, B\}$ is the binary type space, and
(iii) $\mathcal{F} = \{0, 1\}$ is the binary confirmation space, where $F = 1$ if confirmed (F for $f$ollowed-up).

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

test set $\stackrel{\text{def}}{=} \{(x, \tau, f) \text{ s.t. } f = 0\}$,
training set $\stackrel{\text{def}}{=} \{(x, \tau, f) \text{ s.t. } f = 1\}$.
For SN cosmology, it could be that $\mathcal{X}$, $\mathcal{T}$ and $\mathcal{F}$ are, respectively,

(i) $\mathcal{X}$ is the space of all possible photometric data, where an SN's photometric data consist of apparent magnitudes and observational standard deviations in four colour bands over several nights;
(ii) $\mathcal{T} = \{\text{Ia, nIa}\}$, type Ia and nIa SNe; and
(iii) $\mathcal{F} = \{0, 1\}$, 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 features and type:

$$P(F = 1|X = x, T = \tau) = P(F = 1). \tag{16}$$

When the training set is unbiased, the 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 (16) can break down, resulting in a *biased* training set, as described in Zadrozny (2004) and listed below. By *removing bias* from a training set, we mean re-weighting the training points such that the training set becomes unbiased.

(i) Confirmation is independent of features only when conditioned on type: $F|T$ and X are independent. This is the simplest kind of biasing, and there are methods for correcting it (Bishop 1996; Elkan 2001). This is not the bias which exists in SN data.

(ii) Confirmation is independent of type only when given features $F|X$ and T 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 is probably the bias which exists in SN data, and there are methods for correcting it, as we will discuss.

(iii) 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 bias is known.

The decision to confirm an SN can be dictated by different features, examples include Sullivan et al. (2006) and Sako et al. (2008), 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 bands, as we will discuss in Section 8. In reality, there are other factors which affect the confirmation decision such as the weather and telescope availability, but these are independent of SN type. Therefore, the type (ii) bias above is the bias which exists in the SN data. Thus, for the remainder of this section, we will assume the type (ii) bias, that is

$$P(F = 1|X = x, T = \tau) = P(F = 1|X = x). \tag{17}$$

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

$$P(F = 1|X = x, T = \tau) = P(F = 1|X_F = x_F), \tag{18}$$

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

In the following subsection, we will describe how to correctly obtain τ_A-probabilities under the assumption of a bias described by equation (18).

7.2 Correctly obtaining τ_A-probabilities

Let us remind the reader as to how we defined τ_A-probabilities in Section 1:

$$\tau_{\text{A}}\text{-probability} \stackrel{\text{def}}{=} P(T_i = A|X_{P,i} = x_{P,i}) = p_i, \tag{19}$$

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 Poznanski et al. (2002), Guy

et al. (2007), Newling et al. (2011) and Richards et al. (2011). Recall 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 (19), to take into account that a biased follow-up may result in an additional conditional dependence on F:

$$\tau_{\mathrm{A}}\text{-probability} \stackrel{\text{def}}{=} P(T_i = A | F_i = f_i, X_{P,i} = x_{P,i}) = P_i. \tag{20}$$

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

$$X_P = X: \quad p_i = P(T = A | F = 0, X = x). \tag{21}$$

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 equation (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 equation (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, Newling et al. (2011) chose $\mathcal{X}_P$ to be a 20-dimensional space of parameters obtained by fitting light curves.

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,

$$f_{T|F,X_P}. \tag{22}$$

Again, for equation (22) we prefer not to use the standard mass function notation, in order to neaten certain integrals which follow. The τ_A-probability of a test set object can now be expressed in the following way:

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

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,

$$\hat{f}(x_P) \approx f_{T|F,X_P}(A|1, x_P). \tag{23}$$

Using the estimate $\hat{f}$ in equation (23), one can estimate the τ_A-probabilities for the training set objects:

$$P(T = A | F = 1, X_P = x_P) \approx \hat{f}(x_P). \tag{24}$$

The estimate (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 $f_{T|F=0,X_P}$ (equation 22) depends on the relationship between X_F (the data which determine 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 $\mathcal{X}_F \subset \mathcal{X}_P$, is when the data which determine confirmation probabilities are completely contained in the data used to calculate τ_A-probabilities. That is,

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

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

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

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

$$P(F = 1 | T = \tau, X_P = x_P) = P(F = 1 | X_F = x_F). \tag{25}$$

7.2.1 $\mathcal{X}_F \subset \mathcal{X}_P$

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:

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

(i) Using Bayes' theorem, we have

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

(ii) Then, using equation (25), we have

$$= \frac{P(F = 0 | X_F = x_F) \cdot P(T = \tau | X_P = x_P)}{P(F = 0 | X_F = x_F)}$$
$$= P(T = \tau | X_P = x_P). \tag{26}$$

(iii) Using the same steps as above but in reverse and with $F = 1$, we arrive at

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

(iv) This is the type probability function for training set objects, and it can be approximated as

$$\approx \hat{f}(x_P). \tag{27}$$

This is a useful result, as it says that $\hat{f}$ is an approximation of the type probability function not only of the training data but also of the test set. Thus, $\hat{f}$ should provide unbiased τ_A-probabilities for 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 $P(X_P = x_P | F = 1) = 0$ and $P(X_P = x_P | F = 0) \neq 0$, 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 Fan et al. (2005) for a full treatment of this topic.

With respect to SNe, the requirement of the preceding paragraph is that, if an 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 8.

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 the one based on the real-time approach proposed in Freund et al. (1997), 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 8, we discuss this further.

7.2.2 $\mathcal{X}_F \not\subset \mathcal{X}_P$

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 equation (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

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

This 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 explicitly does not want to use X_F in calculating τ_A-probabilities. This may be the case if one wishes to reduce the dependence between D and P, as presented in Section 2. For SNe, this 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 Shimodaira (2000),

$$P(T = \tau | F = 0, X_P) = \int_{\mathcal{X}_F} f_{T,X_F|F,X_P}(\tau, x_F|1, x_P) \cdot w(x_F, x_P)\, \mathrm{d}x_F, \tag{28}$$

where the weight function is defined as

$$w(x_F, x_P) = \frac{f_{F|X_F}(0|x_F) f_{F|X_P}(1|x_P)}{f_{F|X_F}(1|x_F) f_{F|X_P}(0|x_P)}. \tag{29}$$

Note that if $\mathcal{X}_F \subset \mathcal{X}_P$, then $w(x_F, x_P) = 1$ and so equation (28) reduces to the type probability function for training set objects, approximated by $\hat{f}$ as expected from equation (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 equation (28), then an unbiased test set type probability function approximation can be obtained.

The weight function (equation 29) does not require any type information and so can be estimated as a first step. This additional step of estimation introduces an additional error into the final estimate of equation (22), a theoretical analysis of which is presented in Cortes et al. (2008). An alternative to the two-stage approach would be to fit the two terms in equation (28) simultaneously, as suggested and described by Bickel et al. (2007). The use of equation (29) was first suggested in Shimodaira (2000), where a detailed analysis of the asymptotic behaviour of its approximation is given. Therein, it is suggested that equation (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,

$$w(X_F = x_F) = \frac{P(F = 0|X_F = x_F)P(F = 1)}{P(F = 1|X_F = x_F)P(F = 0)}. \tag{30}$$

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

7.3 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, 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 values from the previous section, and not directly as τ_A-probabilities.

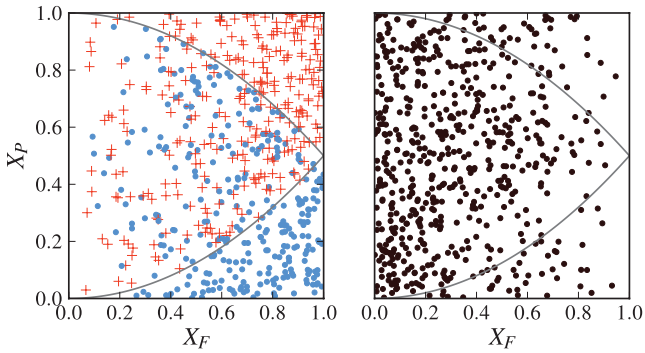

Figure 9. Realizations of a training set (left-hand panel) containing type A (red pluses) and type B (blue points) objects and a test set (right-hand panel), drawn according to equation (31). Overlaid are faint lines delineating the discrete regions described by equation (31).

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 equation (28), one needs to estimate certain probability functions using kernel density estimation. The necessary functions we see from equations (28) and (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,

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

$$f_{F|X_F}(0, x_F) = (1 - x_F),$$

$$f_{F|X_F}(1, x_F) = x_F,$$

$$f_{F|X_P}(1|\tilde{p}) = f_{F|X_P}(0|\tilde{p}) = \frac{1}{2}. \tag{31}$$

Realizations from the above distribution are illustrated in Fig. 9. By integrating x_F out of $f_{T,X_F|F,X_P}(A, x_F|1, \tilde{p})$ in equation (31), we have

$$P(T = A|F = 1, X_P = \tilde{p}) = \tilde{p}. \tag{32}$$

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 equation (28),

$$\begin{aligned} P(T = A|F = 0, X_P = \tilde{p}) &= \int_{\mathcal{X}_F} f_{T,X_F|F,X_P}(\tau, x_F|0, x_P)\, \mathrm{d}x_F \\ &= \int_{\mathcal{X}_F} f_{T,X_F|F,X_P}(\tau, x_F|1, x_P)\, w(x_F, \tilde{p})\, \mathrm{d}x_F \\ &= \int_{\mathcal{X}_F} f_{T,X_F|F,X_P}(\tau, x_F|1, x_P)\, \frac{1 - x_F}{x_F}\, \mathrm{d}x_F \\ &= \begin{cases} \sqrt{2\tilde{p}} - \tilde{p}, & \text{if } \tilde{p} < 0.5, \\ 2 - \tilde{p} - \sqrt{2 - 2\tilde{p}}, & \text{if } 0.5 < \tilde{p}. \end{cases} \end{aligned} \tag{33}$$

The τ_A-probabilities (32) and (33) are plotted in Fig. 10, where we see that $\tilde{p}$ provided accurate τ_A-probabilities for the training set, but not for the test set. This is not unexpected in reality, where the program providing the τ_A-probabilities may have been trained only on the biased training data. Note that this bias should only arise when $\mathcal{X}_F \not\subset \mathcal{X}_P$.

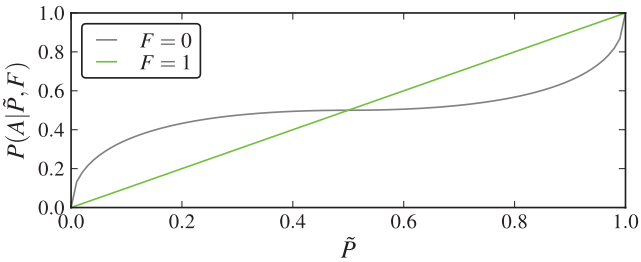

Figure 10. Corrected τ_A-probabilities. The disproportionately large number of training SNe with decisive τ_A-probabilities (as depicted in Fig. 9) causes $\tilde{p}$ values to be too confident as test set τ_A-probability estimates.

8 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 light curves alone. At the end of the competition, the types of all the simulated SNe have been released, making a post-competition autopsy relatively easy to perform. From the results of Kessler et al. (2010a) we see that the probability that an SN was confirmed was based on the r-band and i-band quantities,

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

where $m_{\rm peak}^{\rm band}$ is the band-specific apparent magnitude of an SN, and $M_{\rm min}^{\rm band}$ and $m_{\rm min}^{\rm band}$ are constants. In Kessler et al. (2010a), it is given that for the r and i bands,

$$\begin{aligned} \epsilon_{\rm spec}^{r} &= \epsilon_0 \left(1 - x^5\right), \qquad x \stackrel{\rm def}{=} \frac{m_{\rm peak}^{r} - 16.0}{5.5}, \\ \epsilon_{\rm spec}^{i} &= \epsilon_0 \left(1 - x^6\right), \qquad x \stackrel{\rm def}{=} \frac{m_{\rm peak}^{i} - 21.5}{2.0}, \end{aligned} \tag{34}$$

where ϵ_0 is some constant. Once $\epsilon_{\rm spec}^{i}$ and $\epsilon_{\rm spec}^{r}$ have been calculated, if a $[0 \rightarrow 1]$ uniform random number is less than either of them, confirmation is performed. As confirmation depends only on $\epsilon_{\rm spec}^{i}$ and $\epsilon_{\rm spec}^{r}$, we have from equation (26) that

$$P\left(T = \tau | F = 0, m_{\rm peak}^{i}, m_{\rm peak}^{r}\right) = P\left(T = \tau | F = 1, m_{\rm peak}^{i}, m_{\rm peak}^{r}\right). \tag{35}$$

Equation (35) can be interpreted as saying that the ratio Ia:nIa is the same in a given $m_{\rm peak}^{i}$, $m_{\rm peak}^{r}$ bin. The manner in which the follow-up was simulated should of course guarantee that equation (35) holds. In theory, one should be able to deduce the verity of equation (35) from Fig. 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 equation (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 Fig. 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 that all SNe with apparent magnitudes greater than 23.5 are nIa. As already mentioned in Section 7.1, in situations where the training set does not span the test

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

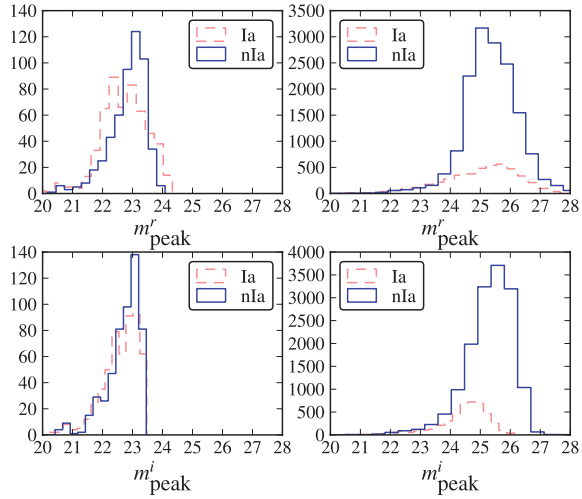

Figure 11. Counts of confirmed (left-hand panels) and not confirmed (right-hand panels) SNe, Ia (dashed) and nIa (solid) as a function of $m_{\rm peak}^{r}$ (above) and $m_{\rm peak}^{i}$ (below).

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 the r and i bands should be removed from a BEAMS analysis, unless there is a valid astrophysical reason not to do so. This entails ignoring about 95 per cent of unconfirmed SNe, which is an enormous cut. We therefore consider it important to confirm more faint SNe.

In Newling et al. (2011), 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 20 fitted light-curve parameters, including fitted apparent magnitudes in the r and i bands. This corresponds to the situation discussed in Section 7.2.1, where $\mathcal{X}_P \subset \mathcal{X}_F$. For this reason, the pdf $\hat{f}$ in equation (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 equation (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 equation (22) was greatly improved.

In their paper, Richards et al. (2011) describe their entry in the SNPCC and report how a semisupervised 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.

9 CONCLUSIONS AND RECOMMENDATIONS

In this paper, we discussed BEAMS and extended the KBH posterior probability function to the case when $D|T$ (distance modulus | type) and P (type probability) are dependent. In Section 6 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 KBH posterior. No bias is observed when using either the extended or the KBH posterior.

In Section 4 we considered examples where the KBH 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 (Section 4.1), 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 (Section 4.2), we observed how BEAMS accurately locates two population means, but fails to match each mean to its population.

We looked at the effects of using biased τ_A-probabilities in Section 5. 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 result which is uncovered 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 summarized as follows. First, the inclusion in the likelihood function of τ_A-probabilities can dramatically reduce the width of the final posterior, providing tighter constraints on cosmological parameters. Secondly, a 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 7.

In Section 7 we considered the problem of de-biasing 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 an 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 an 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 certain filters, etc.

Our second recommendation for SN surveys is that more faint objects are confirmed. While it is 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.

ACKNOWLEDGMENTS

JN has an SKA bursary and MS is funded by an SKA fellowship. BB acknowledges funding from the NRF and Royal Society. MK acknowledges financial support by the Swiss NSF. RH acknowledges funding from the Rhodes Trust.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Bickel S., Brückner M., Scheffer T., 2007, in Ghahramani Z., ed., Proc. 24th Int. Conf. on Machine Learning, Discriminative Learning for Differing Training and Test Distributions. ACM, New York, p. 81
Bishop C. M., 1996, Neural Networks for Pattern Recognition, 1st edn. Oxford Univ. Press, New York
Cortes C., Mohri M., Riley M., Rostamizadeh A., 2008, CoRR, abs/0805.2775
Eisenstein D. J. et al., 2005, ApJ, 633, 560
Elkan C., 2001, in Hayes P. J., ed., Proc. 7th Int Joint Conf. on Artificial Intelligence, The Foundations of Cost-Sensitive Learning. p. 973
Fan W., Davidson I., Zadrozny B., Yu P. S., 2005, in Proc. 5th IEEE Int. Conf. on Data Mining, An Improved Categorization of Classifiers Sensitivity on Sample Selection Bias
Freund Y., Seung H. S., Shamir E., Tishby N., 1997, Machine Learning, 28, 133
Fu L. et al., 2008, A&A, 479, 9
Giannantonio T., Scranton R., Crittenden R. G., Nichol R. C., Boughn S. P., Myers A. D., Richards G. T., 2008, Phys. Rev. D, 77, 123520
Guy J., Astier P., Baumont S., Hardin D., 2007, A&A, 466, 11
Johnson B. D., Crotts A. P. S., 2006, AJ, 132, 756
Kessler R. et al., 2010a, PASP, 122, 1415
Kessler R., Conley A., Jha S., Kuhlmann S., 2010b, preprint (arXiv:1001.5210)
Komatsu E. et al., 2011, ApJS, 192, 18
Kunz M., Bassett B. A., Hlozek R. A., 2007, Phys. Rev. D, 75, 103508 (KBH)
Kuznetsova N. V., Connolly B. M., 2007, ApJ, 659, 530
Mantz A., Allen S. W., Rapetti D., Ebeling H., 2010, MNRAS, 406, 1759
Newling J. et al., 2011, MNRAS, 414, 1987
Percival W. J., Cole S., Eisenstein D. J., Nichol R. C., Peacock J. A., Pope A. C., Szalay A. S., 2007, MNRAS, 381, 1053
Percival W. J. et al., 2010, MNRAS, 401, 2148
Perlmutter S. et al., 1999, ApJ, 517, 565
Poznanski D., Gal-Yam A., Maoz D., Filippenko A. V., Leonard D. C., Matheson T., 2002, PASP, 114, 833
Poznanski D., Maoz D., Gal-Yam A., 2007, AJ, 134, 1285
Richards J. W., Homrighausen D., Freeman P. E., Schafer C. M., Poznanski D., 2012, MNRAS, 419, 1121
Riess A. G. et al., 1998, AJ, 116, 1009
Rodney S. A., Tonry J. L., 2009, ApJ, 707, 1064
Sako M. et al., 2008, AJ, 135, 348
Sako M. et al., 2011, preprint (arXiv e-prints)
Shimodaira H., 2000, J. Stat. Plan. Inference, 90, 227
Sullivan M. et al., 2006, AJ, 131, 960
Zadrozny B., 2004, in Brodley C. M., ed., Proc. 21st Int. Conf. on Machine Learning ICML, Learning and Evaluating Classifiers under Sample Selection Bias. ACM, New York, p. 114

APPENDIX A: ADDITIONAL CONDITIONING ON THE CONFIRMATION OF SUPERNOVA TYPE

In this paper, we did not distinguish 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 this 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 Section 6, we start afresh the posterior derivation, explicitly including the vector $(\boldsymbol{f})$ which describes which objects have been followed-up. Doing this, we arrive at the following posterior distribution:

$$f_{\Theta|\boldsymbol{D},\boldsymbol{F},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{f},\boldsymbol{p}) \propto f_{\Theta}(\theta) \times \sum_{\boldsymbol{\tau}} f_{\boldsymbol{D}|\Theta,\boldsymbol{F},\boldsymbol{P},\boldsymbol{T}}(\boldsymbol{d}|\theta,\boldsymbol{f},\boldsymbol{p},\boldsymbol{\tau}) \prod_{\tau_i=A} \bar{p}_i \prod_{\tau_j=B} (1-\bar{p}_j). \quad \text{(A1)}$$

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 as an open question.

In the case of independent SNe, the posterior (equation B1) reduces to

$$f_{\Theta|\boldsymbol{D},\boldsymbol{F},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{f},\boldsymbol{p}) \propto \prod_{i=1}^{N} [\, f_{D_i|\Theta,F_i,P_i,T_i}(d_i|\theta, f_i, p_i, A)\, \bar{p}_i + f_{D_i|\Theta,F_i,P_i,T_i}(d_i|\theta, f_i, p_i, B)\,(1-\bar{p}_i)]. \quad \text{(A2)}$$

APPENDIX B: POSTERIOR TYPE PROBABILITIES

We here derive the posterior type probabilities based on the modifications of Section 2. 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 $\boldsymbol{P}$:

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

(i) assuming that the objects are independent, we have

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

(ii) using Bayes' theorem, we have

$$= \int_{\theta} \left(\frac{A_i}{A_i+B_i}\right) f_{\Theta|\boldsymbol{D},\boldsymbol{P}}(\theta|\boldsymbol{d},\boldsymbol{p})\,\mathrm{d}\theta, \quad \text{(B1)}$$

where $A_i = P(d_i|\theta, p_i, T_i = A)p_i$, $B_i = P(d_i|\theta, p_i, T_i = B)(1-p_i)$, and we have assumed using 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 (equation A1) is well approximated by $A_i(\hat{\theta})/[A_i(\hat{\theta}) + B_i(\hat{\theta})]$, 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,

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

This paper has been typeset from a TEX/LATEX file prepared by the author.

© 2012 The Authors, MNRAS **421,** 913–925
Monthly Notices of the Royal Astronomical Society © 2012 RAS