The reliability of the AIC method in Cosmological Model Selection

Ming Yang Jeremy Tan[1] and Rahul Biswas[2,1]

[1] *Department of Physics,*

University of Illinois at Urbana-Champaign,

1110 W. Green Street, Urbana,

IL 61801, USA and

[2] *Argonne National Laboratory High Energy Division,*

Argonne National Laboratory,

9700 S. Cass Avenue,

Bldg 362, Lemont,

IL 60439, USA

arXiv:1105.5745v1 [astro-ph.CO] 29 May 2011

Abstract

Akaike Information Criterion (AIC) has been used widely as a statistical criterion to compare the appropriateness of different parametric models underlying a particular dataset. Under suitable conditions, AIC is an indirect estimate of the Kullback-Leibler divergence $D(T||A)$ of a candidate model A with respect to the truth T, where we have defined T as an underlying process consisting of a signal with stochastic noise; a particular empirical data is a single realization of this process. Thus, a model with a smaller AIC is ranked as a better model, since it has a smaller Kullback-Leibler discrepancy with T. However, it is an important question whether the difference between the AIC values for two candidate models is statistically significant. This has been addressed in terms of a probability of models using Akaike Weights. But, it is also important to remember that the AIC itself is statistically estimated with the aid of available data. We explore the impact of the possible errors of the estimated AIC in the context of comparing models underlying the late time acceleration of the universe, using SNIa by using a parametric bootstrap technique to study the reliability of the estimated AIC. From the specific example that we study, we find that the estimator uncertainty in the AIC differences can be significant. Therefore, AIC based studies should not only pay attention to Akaike weight based criterion of reliability, but should also consider the impact of estimator uncertainty of AIC.

I. INTRODUCTION

Suppose we wish to choose the 'best' model from a set of theoretical models (or theories) of a natural phenomenon with the aid of the relevant empirical data. How can we objectively accomplish this goal? Many statistical techniques such as Hypothesis Testing [1], Bayesian Evidence [2], Akaike Information Criterion (AIC) [3], Bayesian Information Criterion (BIC) [4], Minimum Description Length (MDL) [5], etc. have been developed to address the question of model selection.

The late time accelerating expansion of the universe has been firmly established [6–18], but there is no consensus on the physics behind this phenomenon. A number of possible explanations such as a small positive cosmological constant or vacuum energy, otherwise unobserved dynamical fields usually called dark energy [19–22], or a modification of General Relativity [23–25] have been proposed as an explanation. With many models still consistent with current data, it is clear that further progress of the field requires the collection of larger and complementary datasets and a definite framework for model selection. Several large new surveys [26–29] have been planned to study this late time acceleration by collecting more data. Of course, even with accumulation of more quality data, the importance of analyzing the model selection process will not diminish, because reliable discriminating methods can always allow us to exploit the available data maximally.

Hence, statistical techniques addressing model selection have been applied to this context [30–40]. Many of these discuss the use of information criteria like AIC and BIC which are easy to calculate. A number of works have applied this to data. In particular, Szydłowski et al. [33] used the method of AIC and BIC, and a compilation of SNIa to compare various late time acceleration cosmological models. Sollerman et al. [38] and Davis et al. [39] compared such models based on the SDSS [16] and ESSENCE Supernova data and high redshift data [11, 13] along with a summary of CMB and BAO data using AIC and BIC. AIC has also been used in comparing principal component based models for dark energy [41]. We wish to explore the use of AIC in the context of selecting the 'best' theoretical model describing the late time acceleration of the universe.

This paper will focus on the use and validity of applying the AIC technique to choose the 'best' late time acceleration model from SNIa data. For our calculation, we use data from the Constitution compilation [14] of the CFA3 sample [42] ESSENCE [43], SNLS [12] and 'High-z' samples [44]. We will study and refine the AIC methodology in this context. In particular, we wish to study its reliability when subjected to estimator uncertainty. With the AIC technique, the usual approach has been to use a single number (called AIC) to rank models; the smaller AIC the better the model. This concept of discrete ranking of models has been extended [45, 46] by paying attention to the differences between actual AIC values. However, it is important to note that the AIC values themselves are empirically estimated from the data and thus have statistical uncertainties. Therefore, the reliability of the estimates of AIC differences is a crucial issue. While there has been work [47] that tries to address the reliability of the AIC technique through analytic calculations, we will instead study the reliability through numerical simulations.

The paper is organized as follows. In Section II, we briefly review the information theoretic origin of AIC, extending this idea further, by reviewing the theoretical implications of the AIC differences in Section III. In Section IV, we study the estimator uncertainty of these AIC differences in the context of cosmological model selection. This involves a comparison of a set of four candidate models by using SNIa, as in the work by Szydłowski and Godłowski [32]. The reliability of these AIC estimates are calculated via a bootstrap method [48] and this will be used to evaluate the validity of the model selection technique. The results of these simulations and the conclusion will then be given in Sections V and VI, respectively.

II. THE AKAIKE INFORMATION CRITERION (AIC)

The Kullback-Leibler (KL) divergence [49] is a commonly used quantity that measures the discrepancy [55] of one probability distributions with respect to another probability distribution. We denote the KL divergence of a candidate model A (which gives the probability distribution p_A) with respect to the truth T (with probability distribution p_T) as $D(T||A)$, where we have defined T as an underlying process consisting of a signal with stochastic

noise; a particular empirical data is a single realization of this process. If we denote X as the set of all possible outcomes that can be generated by both A and T, and x as an element in this set, we can define $D(T||A)$ as

$$D(T||A) = \int_{x \in X} dx \; p_T(x) \log \frac{p_T(x)}{p_A(x)}.$$

In this paper, we define a model class as the totality of model probability distributions with the same parametric form (but with different parameter values). Within each model class, there is a set of parameters (the 'best' model) that gives the lowest KL divergence with respect to the truth. Thus, to choose the 'best' model class we first choose the 'best' model (parameter set) from a particular model class as the representative of the model class. This is done by the maximum likelihood criterion. The model selection strategy is thus obtained by comparing the KL divergence of the representative models of the individual model classes.

However, the truth is unknown a priori, so for a given model A, $D(T||A)$ cannot be evaluated directly. We can solve this problem by computing the AIC value [3] which is an unbiased estimator for $D(T||A)$ (up to a fixed offset that is independent of the candidate models). Since the fixed offset is independent of the model classes, a comparison of the AIC values is a useful surrogate for the strategy of comparing $D(T||A)$ for different candidate models A. We can compute the AIC of the representative ('best') models of various model classes, which in the asymptotic limit is known [3] to be

$$\text{AIC} = 2k - 2\log(\text{ML}), \tag{1}$$

where ML denotes the maximum likelihood, and k is the number of free parameters in the candidate model. When we make a further assumption that the distribution of errors follows a Gaussian distribution, this further reduces to

$$\text{AIC} = 2k + \chi^2_{ML}, \tag{2}$$

where χ^2_{ML} is the usual chi-square evaluated at the maximum likelihood estimate of the model parameters. The AIC values are subsequently ranked by the smallness of their values; the model with the smallest value is determined to be the 'best' model.

It is worth noting that the AIC estimate (Eqn. 2) of the KL divergence assumes that the number of data points is sufficiently large. AIC in the form written above is an unbiased estimator for large data sets. For smaller data sets, the $2k$ term can be corrected by an additional $\frac{2k(2k+1)}{N-k-1}$ term to approximately correct for this effect, where N is the number of data points in a single data set. While further studies to obtain a more accurate expression for this term are possible, for the cases we shall consider, this correction is always less than 0.06 (which will be seen to be negligible for our purposes) and will only decrease in importance when more data is collected. We shall therefore ignore this correction altogether in this paper.

As already mentioned explicitly, when we are comparing different model classes, we are actually comparing the best fit models taken from each model class. These best fit models are derived from the maximum likelihood estimate of the model parameters given the data. Thus, for the rest of this paper, when we refer to a model, we actually mean the (best fit) model in its model class.

III. MODEL COMPARISON AND AIC DIFFERENCES

Since AIC is essentially the measure of discrepancy of a model distribution from the truth, it is intuitively obvious that the smaller the AIC difference between two models, the harder it becomes to judge which model is better; even if the AIC estimate of this difference in the KL divergence can be obtained without estimator error, the small difference would make it difficult to tell the two probabilistic models apart for a small number of observations. Hence, there is a need to associate a confidence level for distinguishing between a model A and another model B using the AIC difference between them.

Let $P(M)$ be the probability that model M is true. We wish to relate $P(A)/P(B)$ to the AIC difference

$$\Delta_{A,B} \equiv \mathrm{AIC}(A) - \mathrm{AIC}(B) \approx 2\left[D(T||A) - D(T||B)\right],$$

where $\approx$ implies the asymptotic relation. This asymptotic relation may not necessarily be realized as the number of samples is small. Using a certain set of extra assumptions, Akaike

[45] showed that for a pair of models A and B,

$$P(A)/P(B) \approx \exp[-\Delta_{A,B}/2]. \tag{3}$$

Since the AIC differences are estimates of the differences of KL divergences, one can also use the idea of distinguishing probability distributions to justify Eqn. 3 when the number of number of samples are sufficiently large (Appendix A). Assuming the truth to be in the set of candidate models, Burnham and Anderson [46] extended Akaike's idea by normalization of the ratios and defined the Akaike weight w_i of model i as given by

$$w_i = \frac{\exp[-\Delta_{i,b}/2]}{\sum_j^n \exp[-\Delta_{j,b}/2]}, \tag{4}$$

where the candidate models are numbered as $i = 1, \cdots, n$ and b denotes the best model which has the smallest AIC value among all the candidate models.

Either way, Eqn. 3 forms the basis of the claim [46] that in order to distinguish between two models A and B on the basis of a certain data set, the AIC difference must be greater than a pre-determined cut-off [46] where the odds ratio is given by Eqn. 3. As a rule of thumb, a cut-off value of 5 has been mentioned by Liddle [40] as the minimum AIC difference between two models needed to make a 'strong' assertion that one model is better than the other.

IV. IMPACT OF AIC UNCERTAINTIES IN FINITE DATA SETS

In the preceding section, we have discussed the AIC differences $\Delta_{A,B}$ without any regard for the fact that AIC is a statistical estimate. The associated uncertainty in the AIC estimate may not be negligible and may be dependent on the realization of noise in a particular data set. Thus, there must be a statistical uncertainty in the value of $\Delta_{A,B}$ [56] even when estimated from a data set of similar quality [57]. The ensemble of such observations defines an empirical probability distribution of $\Delta_{A,B}$, and the particular value of $\Delta_{A,B}$ obtained from the current SN data sets is actually a sample value drawn from this probability distribution.

Ideally, we should be able to study the probability density distribution of $\Delta_{A,B}$ under repeated observations of results with sample size N: $P(\Delta_{A,B}|E_N)$, where E_N denotes

a collection of observation data which individually consists of N observation points and are drawn from the underlying truth process. Because producing a large subset of E_N is impossible, in this paper we instead use a bootstrap approach [48] to generate 'mock' empirical data sets and estimate the probability distribution of $\Delta_{A,B}$.

Perhaps the most frequently used method to produce 'mock' empirical data is the bootstrap method proposed by Efron [48]. When we apply this approach to regression models, we need a probability model that specifies the distribution of residuals. In our case in particular, the distance modulus μ is related to the red shift z, so a regression model has the following structure $\mu = f(z) + \epsilon$, where ϵ is the residual (the error term). Since we do not know the true relation f, we cannot obtain the purely empirical distribution of the residuals [58]. Thus, for regression models, the standard bootstrap method always involve using model-dependent probability distribution of residuals. Davis et al. [39] extended this method to compare between two regression models and find the standard deviation of their BIC differences; we will further extend this idea and show that studying the structure of the distribution of AIC differences is important.

When we wish to check the reliability of a particular model, we could use the probability distribution of residuals based on the model itself. Since we cannot have any model-free bootstrap data, we must choose a particular model to produce the residual distribution. Therefore, to estimate $\Delta_{A,B}$, we need some model C as a reference probability model that is used to generate the bootstrap data. Let us denote the estimate of $\Delta_{A,B}$ based on a data set d as $\Delta^d_{A,B}$. We wish to produce $\{\Delta^d_{A,B}|d \in C_N\}$, where C_N denotes a collection of parametric bootstrap data generated by model C, which individually consist of N observation points with the same data quality as our empirical data. Although the needed details are in Appendix B, our approach may be outlined as follows.

Suppose we have a single data set consisting of N observed results $\{(z_1, \mu_1, \sigma_1), (z_2, \mu_2, \sigma_2), \cdots, (z_N, \mu_N, \sigma_N)\}$, where the z_i values denote the i^{th} observed redshift, μ_i the i^{th} generated distance modulus, and σ_i the observed error bars of μ_i. We can create a bootstrap sample C_N consisting of N observation points based on model C. C_N relates the same set of coordinates by the relation: $\mu_i = f(z_i) + \epsilon_i$, $i = 1, \cdots, N$, where

ϵ_i is a stochastic term obeying $N(0, \sigma_i^2)$, and $f(z)$ is the maximum likelihood estimate of model C (ΛCDM, DGP, etc.) (Appendix C).

As mentioned above, we wish to simulate the distribution of $\{\Delta^d_{A,B}|d \in C_N\}$ as a proxy for $P(\Delta_{A,B}|E_N)$. To proceed, we choose a subset of models that have been often studied in the literature [32] and are listed in Table I. We also assumed that the universe is flat and set the curvature term Ω_k in the Hubble function $H(z)$ to zero. Three of these models ΛCDM, wCDM, CPL [50, 51] are dark energy models with different parametrizatons of the equations of state $w(z) = p(z)/\rho(z)$, where $p(z)$ and $\rho(z)$ are the pressure and density of dark energy, respectively. These models are nested: setting $w_a = 0$ in the CPL model, we obtain the wCDM model; setting $w_0 = -1$ in the latter gives the ΛCDM model. We also use the flat DGP model which is a modified gravity model and cannot be nested in the previous classes of models.

To make contact with observational data, we choose candidate models with parameter sets that are 'best' for the Constitution compilation of SNIa data ($f(z)$ being the maximum likelihood estimation derived model), where we use the distance moduli and the error bars in data (Appendix C). Since we need to find estimates of cosmological parameters for different models by maximizing likelihoods, we use the results from the more appropriate MLCS light curve fitter (for $R_V = 1.7$) in [14]. The details of finding the maximum likelihood and the corresponding AIC value is given in Appendix C; 371 SNIa events were used. The Hubble function, as a function of cosmological parameters in each of these models, along with the free parameters and AIC values (calculated from the Constitution compilation) are shown in Table I.

For the nested models, the parameter values that fit the data best turn out to be close to the ΛCDM model. Consequently, one does not gain much in terms of a lower χ^2_{ML}, while the extra free parameters are penalized to give higher values of AIC for wCDM and CPL. The DGP model gives the best AIC value, which is only slightly better than the ΛCDM model. It is known that the simultaneous use of Cosmic Microwave Background(CMB) data from WMAP [38, 39] and Large Scale Structure (LSS) data disfavors the DGP model compared to ΛCDM, since the parameter subspaces that provide

TABLE I: Different model classes considered in this paper: We show the evolution of the Hubble function $H(z)$ with redshift z, the Hubble constant H_0 and other free parameters in the models. k is the number of free parameters. The respective AIC values were obtained from the Constitution compilation.

Model	$H(z)/H_0$	Free Parameters	k	AIC
ΛCDM	$\sqrt{\Omega_m(1+z)^3+(1-\Omega_m)}$	Ω_m	1	401.35
wCDM	$\sqrt{\Omega_m(1+z)^3+(1-\Omega_m)(1+z)^{3(1+w)}}$	Ω_m,w	2	403.05
CPL	$\sqrt{\Omega_m(1+z)^3+(1-\Omega_m)(1+z)^{3(w_0+w_a+1)}\exp[\frac{-3w_az}{1+z}]}$	Ω_m,w_0,w_a	3	404.66
DGP	$(\sqrt{\Omega_m(1+z)^3+\Omega_{rc}}+\sqrt{\Omega_{rc}})$ $\Omega_{rc}=(1-\Omega_m)^2/4$	Ω_m	1	401.13

the best fits for CMB, LSS and SNIa data do not overlap as much as in the ΛCDM model. We checked that our results are consistent with this, but will ignore the CMB and LSS data to focus on methodology.

In order to calculate the distribution of $\Delta^d_{A,B}$, we adopt the best fit model in a model class C as a reference model to produce 1000 mock data sets of C_{371} which is expected to be similar in quality to the Constitution compilation of Supernova data. All the model classes in Table I are successively chosen as the reference model C. Following our definition of 'similar quality', each simulated data set has the same set of redshift values and error bars as the Constitution compilation, while their apparent magnitudes are those expected from a noisy realization of the reference model. The basic steps are:

1. We produce a mock data set consisting of 1000 realizations of $d \in C_{371}$ for a reference model C as outlined above.
2. Candidate models A and B are fitted to $d \in C_{371}$ by maximizing the likelihood and the AIC values of these models A and B are computed through Eqn. 2.
3. Thus, for each element $d \in C_{371}$ we can make the AIC difference $\Delta^d_{A,B}$. We study $\{\Delta^d_{A,B} \mid d \in C_{371}\}$ for the reference model C by plotting a histogram.

We should note that the probability distribution of $\Delta^d_{A,B}$ is due to errors introduced by the stochastic noise term, but ignores the effect of the uncertainties in the cosmological parameters of the reference model C.

V. BOOTSTRAP STUDY RESULTS

We first consider the issue of statistical self-consistency in the following sense: When the reference model is C, does the bootstrap AIC method outlined above choose model C as a better model than the rest? The distribution of the values of $\Delta^d_{A,C}$ for $d \in C_N$ can tell us about the statistical self-consistency. If $P(A/C) \equiv P(\{d \in C_N | \Delta^d_{A,C} > 0\})$ is larger than some predefined threshold, which must not be less than 1/2, we may say model C is better than A when C is the reference. $P(A/C)$ is summarized in Table II. The P-values (Appendix D) for the table are extremely small.

X\Y	DGP	ΛCDM	wCDM	CPL
DGP	-	58.5 %	82.8 %	88.0%
ΛCDM	61.0%	-	86.5%	92.4%
wCDM	13.9%	18.1%	-	90.2%
CPL	13.6%	16.8%	15.7%	-

TABLE II: Percentage of positive AIC differences $P(Y/X)$. We defined the difference and percentage using the following convention. If we define the bootstrap data as being produced by the reference model X and we are comparing it against model Y, the difference is defined as the AIC value of Y minus the AIC value of X and is denoted by the symbol $\Delta^d_{Y,X}$. The table counts the percentage of positive $\Delta^d_{Y,X}, d \in X_N$. The P-value of this table is given in Appendix D.

We see from Table II that we obtained $\Delta^d_{CPL,DGP} > 0$, 88% of the time, when DGP model is the reference (i.e., $P(\text{CPL/DGP}) = 0.880$), and 86% of the time, when CPL model is the reference (i.e., $1 - P(\text{DGP/CPL}) = 0.864$). This means that the DGP model is significantly favored over the CPL model even if the CPL model is the reference. This means that the AIC method is statistically inconsistent when it is applied to the CPL model, and

would be automatically disqualified under the current number of data points in the sample. In another example, we compared the DGP and ΛCDM models in the table and noticed $\Delta^d_{DGP,\Lambda CDM} > 0$, 61% of the time when ΛCDM model is the reference (i.e., $P(\text{DGP}/\Lambda\text{CDM}) = 0.610$), and 42% of the time when DGP model is the reference (i.e., $1 - P(\Lambda\text{CDM}/\text{DGP}) = 0.415$). This means that if the reference model was either ΛCDM or DGP and we apply AIC to compare between them, AIC will slightly favor the reference. Only the AIC technique that is applied to compare DGP and ΛCDM is statistically self-consistent while we cannot use AIC to self-consistently study the other models under the current level of observation quality. However, a test that gives the right answer 3 out of 5 times is unreliable, since we can only do a single empirical test from our actual data. ΛCDM and DGP cannot be distinguished significantly using either reference models. Thus, looking at both examples, we must conclude that there are insufficient data points to tell reliably the models apart with the AIC. This could be remedied only by collecting more data points.

We could learn more details about the statistical uncertainty of the AIC differences by investigating the shape of the distribution of the AIC differences $\{\Delta^d_{A,B} | d \in C_N\}$ under different reference models C. They roughly fall into two categories. They could either resemble an approximately symmetric Gaussian-like distribution such as the distribution of $\Delta^d_{\Lambda CDM,DGP}$ when $d \in \Lambda\text{CDM}_{371}$ or a strongly asymmetric distribution exponential-like distribution such as the distribution of $\Delta^d_{\Lambda CDM,wCDM}$ when $d \in \Lambda\text{CDM}_{371}$. We see an example of a Gaussian-like distribution in Figure 1 and exponential-like distribution in Figure 2. This is important when we study the spread of the statistical uncertainty of the distribution of the AIC differences since they cannot be specified unless we know the shapes beforehand. Due to the structural differences between the two kinds of distributions, we must define the 'error' bars according to the shape in order to make any useful comments about the uncertainty of the estimate. For the case of the Gaussian-like distribution, we define the error bars to be the standard deviation of the statistical distribution of differences. For the case of the exponential-like distribution, we define an error bar region as the range that begins at the sharp edge of the distribution and stops at the point where the range contains 67% of the differences. We chose 67%, because 0.67 is approximately the probability of finding an outcome within 1 standard deviation of the mean of a Gaussian distribution.

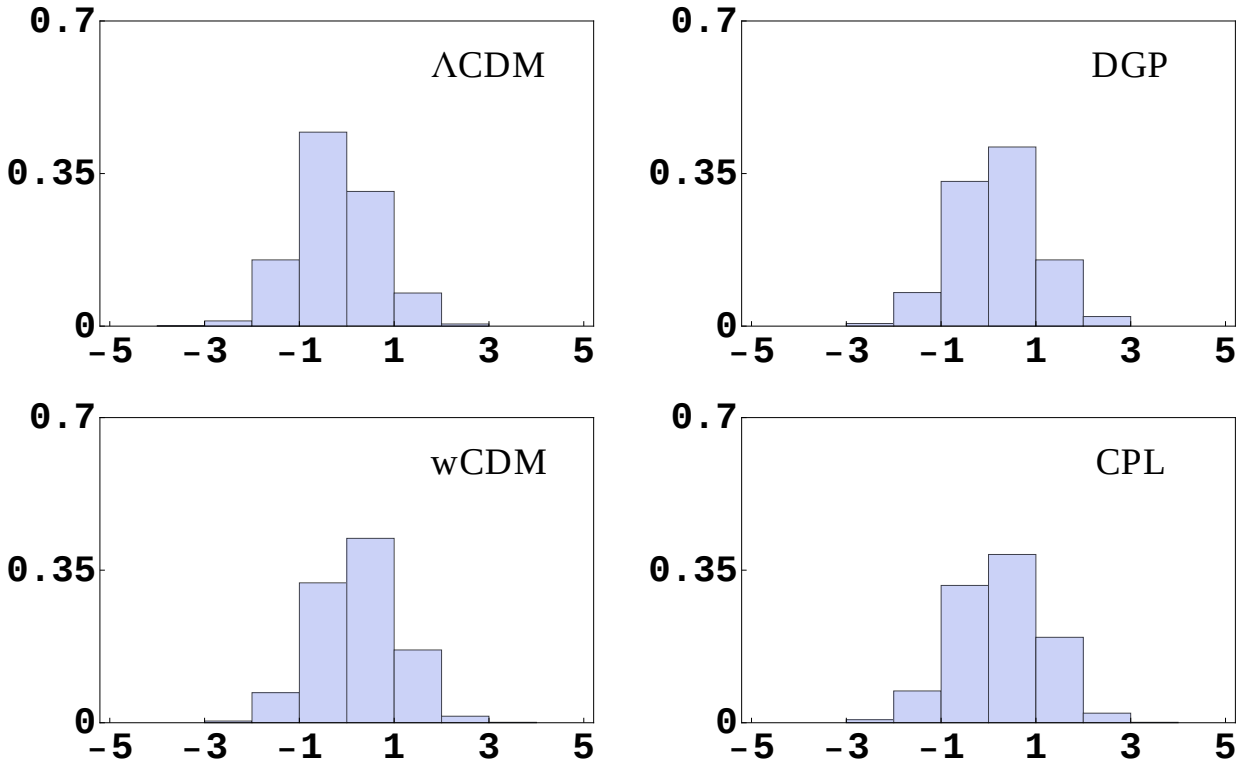


FIG. 1: Probability distributions of the AIC differences between ΛCDM and DGP ($\Delta^d_{\Lambda CDM,DGP}$) for different reference model C. C is used to generate the respective bootstrap samples and is written in the right upper corner of the figures. The horizontal axis indicates the $\Delta^d_{\Lambda CDM,DGP}$ value while the vertical axis indicates their relative frequency. If the process underlying our observations was really the best fit model of class C, then the values of the AIC differences under different realizations of noise would have the histogram distribution $\{\Delta^d_{\Lambda CDM,DGP} \mid d \in C_{371}\}$ shown.

As an illustration, we look at statistical uncertainty of the distribution of $\Delta^d_{\Lambda CDM,DGP}$ for the two separate cases where the DGP and ΛCDM models are the references. The distribution of the differences in both cases are Gaussian-like. We notice that the standard deviation of $\Delta^d_{\Lambda CDM,DGP}$ is 0.89 with DGP model as the reference and 0.83 with ΛCDM model as the reference. The AIC difference between the DGP model and ΛCDM model $\Delta_{\Lambda CDM,DGP}$ in our original AIC analysis is 0.22 and smaller than the error bars. This means that any subsequent analysis based on the value of $\Delta_{\Lambda CDM,DGP}$ observed is unreliable. This could also be significant even if $\Delta_{\Lambda CDM,DGP}$ was larger than the error bars as we would have to modify any subsequent AIC difference analysis to include this uncertainty. When we look at $\Delta^d_{\Lambda CDM,CPL}$ (exponential-like distribution), the error bar region ranges

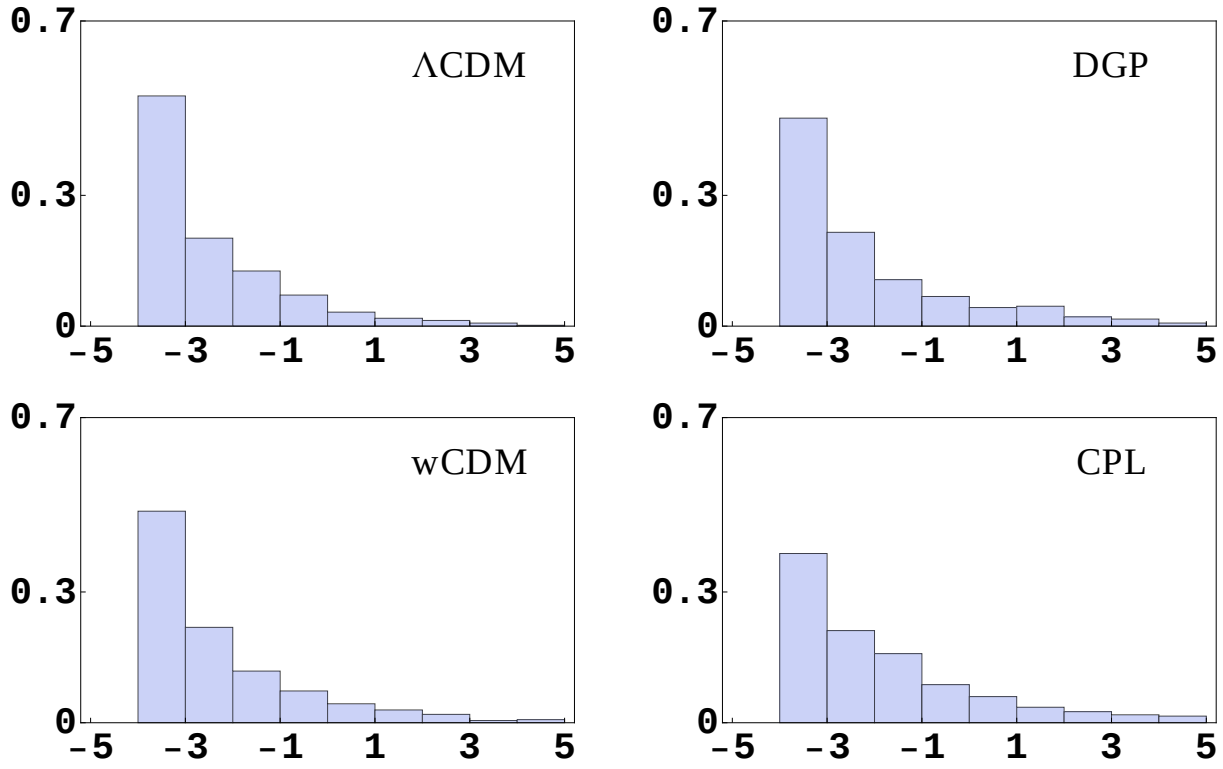


FIG. 2: Probability distributions of the AIC differences between ΛCDM and CPL ($\Delta^d_{\Lambda CDM,CPL}$) for different reference model C. C is used to generate the respective bootstrap samples and is written in the right upper corner of the figures. The horizontal axis indicates the $\Delta^d_{\Lambda CDM,CPL}$ value while the vertical axis indicates their relative frequency. If the process underlying our observations was really the best fit model of class C, then the values of the AIC differences under different realizations of noise would have the histogram distribution $\{\Delta^d_{\Lambda CDM,CPL} \mid d \in C_{371}\}$ shown. The exponential-like distributions observed are due to the fact that ΛCDM and CPL are nested models.

from -4.00 to -2.31 for the case when the reference is the ΛCDM model and ranges from -4.00 to -1.48 when the reference is the CPL model. The $\Delta_{\Lambda CDM,CPL}$ value calculated from the empirical data was found to be -3.31. However, the error bar region range of approximately 2 would make the value of -3.31 less certain. These are just two examples in which one can carry out an analysis to determine the reliability of the model likelihood ratio $P(A)/P(B)$ that is calculated in Eqn. 3.

We note that the statistical uncertainty of the AIC differences obtained above is smaller than the reliability cut-off of 5 already mentioned above [38–40]. However, it is still

significant enough and needs to be considered in our analysis.

We note that a comparison of nested models will always involve exponential-like distribution of AIC differences. This is because the χ^2_{ML} for the general case cannot be smaller than the more specific case. For example, ΛCDM is a special case of the more general wCDM and CPL models. Hence, when we fit these models against data generated from the best fit ΛCDM model and fit the other nested models against this data, we find that the wCDM and CPL χ^2_{ML} values cannot be larger than the ΛCDM χ^2_{ML} value and are usually smaller. This results in a sharp edge that gives the exponential-like distribution. Since AIC $= \chi^2_{ML} + 2k$, the edge of these exponential-like histograms is shifted by twice the difference between their number of parameters. On the other hand, there is no similar constraint relating the χ^2_{ML} values of the ΛCDM and DGP (at the best fit value); consequently the histogram turns out to be Gaussian-like. They have the same number of free parameters so this distribution is roughly centered about zero.

VI. CONCLUSION

AIC has been widely used as a technique for model selection. Most commonly, this has been applied by computing the AIC values for each candidate model through Eqn. 2 and selecting the model with the smallest AIC value as the best model. The issue of considering the magnitudes of AIC differences between the models to indicate the relative plausibility or confidence in the models has also been addressed by Akaike and elaborated by Burnham and Anderson through the use of Akaike weights Eqn. 4. In the field of cosmology, AIC has been used in selecting models underlying the late time acceleration of the universe using data. There have also been suggestions of a rule of thumb that states an AIC difference of 5 or more would give strong evidence for the model with the smallest AIC value. This approximately corresponds to a ratio of model likelihoods of 12 or more.

In this paper, we propose a method for calculating the ratio of the model likelihood between models A and B based on their AIC differences $\Delta_{A,B}$ (Appendix A), using the idea

of probabilistic model distinguishability by Balasubramanian [52]. This is an alternative method of arriving at the odds ratio of Eqn. 3 assuming that the AIC difference between the candidate models is a perfect unbiased estimator of the difference between the models' KL divergences with respect to the truth. A related analysis using the AIC differences in accordance with the Akaike weights also derives the same equation. The analysis of $\Delta_{A,B}$ was extended further by investigating the statistical uncertainty of this estimate. Our focus was not necessarily on the 'best' cosmological theory. Thus, we did not use the most exhaustive data sets, nor explore in detail the systematics associated with the surveys considered.

To this end, we studied the distribution of the differences of AIC estimates given a certain quality of data $P(\Delta_{A,B}|E_N)$. Since we do not know the exact process underlying the empirical data (E_N in Section IV), we approach this problem by studying surrogate processes, where the reference or generating model is assumed to be one of four (best fit) candidate models for the late time acceleration of the universe; following the approach used by Davis et al. [39]. These models were listed in Table I with the best fit free parameters equal to the maximum likelihood values of each of these models. This was obtained by fitting 371 SNIa extracted from the Constitution compilation.

Our simulations have demonstrated that, given the data used, there was insufficient data, in particular, to reliably use AIC to tell all the models apart; in agreement with the general consensus [39]. This was demonstrated, in particular, by calculating $P(A/C) \equiv P(\{d \in C_N|\Delta^d_{A,C} > 0\})$, which is the proportion of positive AIC differences for each given reference model. The success rate of the technique was shown to be unsatisfactory. We also calculated the respective statistical uncertainty ($\sim 1\ \sigma$ error bars) of $\Delta^d_{A,B}$ and showed it was even larger than the observed $\Delta_{A,B}$ between some of the models. As an important example, we considered the ΛCDM and DGP models since they were shown to have the two lowest AIC values in Table I. It was shown that the statistical uncertainty of $\Delta^d_{\Lambda CDM,DGP}$ was larger than the observed $\Delta_{\Lambda CDM,DGP}$, making it difficult to determine the better model between the two. From our simulation, we also showed that the shapes of the distribution of the AIC differences can be quite varied, ranging from a symmetric Gaussian-like distribution to an exponential-like distribution with a sharp edge

and one sided tail. Thus, in order to use AIC reliably, one must pay proper attention to the statistical variation of $\Delta_{A,B}$.

In this paper, we made a number of assumptions to study the AIC technique. All calculations in this paper were only for an assumed reference. Since the empirical data does not give us E_N, there is no way to know the actual distribution of $\Delta_{A,B}$. However, we should note that AIC is a model comparison technique that assumes that one of the model classes contains the reference C. By restricting C to the set of candidate models, we can at least look for statistical self-consistency in that assumption. It should be emphasized again that the reference models used were the best fit models and did not take into account the statistical uncertainty of the individual model parameters. That can be taken into account by sampling the distribution of parameters. As mentioned before, the whole point of this simulation is to highlight the statistical distribution of the AIC differences under different reference models and look for statistical inconsistencies under each of the assumptions. Another issue that should be noted is that the exact variance in the data is unknown and that the error bars in the data may not be reflective of the true error bounds. While we use these in our simulations, we note that the correctness of these error estimates was an assumption of previous AIC computations.

In summary, the reliability of the AIC estimator is an important issue that should be taken into consideration when using the AIC technique to select models. It should also be noted that such considerations are not just restricted to AIC but any technique that relies on the maximum likelihood estimators. This should be borne in mind when applying the techniques to any statistical analysis.

Acknowledgments

We would like to thank Yoshi Oono for checking the manuscript and providing many insights about the nature of AIC and the bootstrap technique. M.Y.J. Tan was funded in part by Yoshi Oono. During the course of this work, R. Biswas was partially supported by NSF AST 07-08849 and NSF AST 09-08693 ARRA and then by Argonne National Laboratory, operated by UChicago Argonne, LLC. Work at Argonne was supported by U.S. Department

of Energy, Office of Science under contract DE-AC02-06CH11357.

Appendix A: Confusion Probability and Model Likelihood

This is an almost identical repeat of Balasubramanian [52] explanation of error probabilities which is framed in the language of hypothesis testing. It is reproduced for the convenience of the reader. Suppose $\{x_1, x_2, ..., x_N\} \in \mathbb{X}^N$ are drawn i.i.d. variables from one of f_1 and f_2 with $D(f_1\|f_2) < \infty$. Let $A_N \subseteq \mathbb{X}^N$ be the acceptance region for the hypothesis that the distribution is f_1 and define the type I and type II error probabilities as $\alpha_N = f_1^N(A_N^C)$ and $\beta_N = f_2^N(A_N)$ respectively. A_N^C is the complement of A_N in $\mathbb{X}^N$, and f^N denotes the product distribution on $\mathbb{X}^N$ describing N i.i.d outcomes drawn from f. In this definition α_N is the probability that f_1 is mistaken for f_2, and β_N is the probability of the opposite error. Stein's lemma tells us how low we can make β_N given a particular value of α_N. Indeed, let us define $\beta_N^\epsilon = \min_{A_N \subseteq \mathbb{X}^N, \alpha_N \leq \epsilon} \beta_N$ for a positive ϵ. Then Stein's lemma tells us

$$\lim_{\epsilon \to 0} \lim_{N \to \infty} \frac{1}{N} \ln \beta_N^\epsilon = -D(f_1\|f_2). \tag{A1}$$

To prove Stein's lemma, we refer to the proof by Cover and Thomas [49], which is provided almost verbatim here for convenience sake. Defining $\delta \in \mathbb{R}^+$, we first state A_N more explicitly as:

$$A_N = \left\{ x \in \mathbb{X}^N : \exp[N\,(D(f_1\|f_2) - \delta)] \leq \frac{f_1(x)}{f_2(x)} \leq \exp[N\,(D(f_1\|f_2) + \delta)] \right\} \tag{A2}$$

Then, we have the following properties:

1. $f_1^N(A_N) \to 1$.

Proof:

$f_1^N(A_N) = f_1^N\left(\frac{1}{N}\sum_{i=1}^N \log\frac{f_1(x_i)}{f_2(x_i)} \in (D(f_1\|f_2) - \delta, D(f_1\|f_2) + \delta)\right) \to 1$ by the law of large numbers, since $D(f_1\|f_2) = E_{f_1}\left(\log\frac{f_1(x)}{f_2(x)}\right)$. Therefore, for any positive ϵ, $\alpha_N < \epsilon$ for large N.

2. $f_2^N(A_N) \leq \exp\left[-N\,(D(f_1\|f_2) - \delta)\right]$

Proof:

$$\begin{aligned} f_2^N(A_N) &= \sum_{A_N} f_2(x), \\ &\leq \sum_{A_N} f_1(x) \exp[-N\,(D(f_1\|f_2) - \delta)], \end{aligned}$$

$$
\begin{aligned}
&= \exp[-N\left(D(f_1\|f_2)-\delta\right)]\sum_{A_N} f_1(x),\\
&= \exp[-N\left(D(f_1\|f_2)-\delta\right)](1-\alpha_N).
\end{aligned}
$$

3. $f_2^N(A_N) \geq \exp\left[-N\left(D(f_1\|f_2)+\delta\right)\right]$

Proof:

$$
\begin{aligned}
f_2^N(A_N) &= \sum_{A_N} f_2(x),\\
&\geq \sum_{A_N} f_1(x)\exp[-N\left(D(f_1\|f_2)+\delta\right)],\\
&= \exp[-N\left(D(f_1\|f_2)+\delta\right)]\sum_{A_N} f_1(x),\\
&= \exp[-N\left(D(f_1\|f_2)+\delta\right)](1-\alpha_N).
\end{aligned}
$$

4. $\lim_{N\to\infty}\frac{1}{N}\log\beta_N = -D(f_1\|f_2)$.

Proof:

From 2. and 3. we know:

$$
\begin{aligned}
\frac{1}{N}\log\beta_N &\leq -D(f_1\|f_2)+\delta+\frac{\log(1-\alpha_N)}{N}.\\
\frac{1}{N}\log\beta_N &\geq -D(f_1\|f_2)-\delta+\frac{\log(1-\alpha_N)}{N}.
\end{aligned}
$$

5. No other sequence of acceptance regions does better.

Proof: Let $B_N \subseteq \mathbb{X}^N$ be any other sequence region with $\alpha_{N,B_N} = f_1^N(B_N^c) < \epsilon$. Let $\beta_{N,B_N} = f_2^N(B_N)$

$$
\begin{aligned}
\beta_{N,B_N} &= f_2^N(B_N),\\
&\geq f_2^N(A_N\cap B_N),\\
&= \sum_{A_N\cap B_N} f_2(x),\\
&\geq \sum_{A_N\cap B_N} f_1(x)\exp[-N\left(D(f_1\|f_2)+\delta\right)],\\
&= \exp[-N\left(D(f_1\|f_2)+\delta\right)]\sum_{A_N\cap B_N} f_1(x),\\
&\geq (1-\alpha_N-\alpha_{N,B_N})\exp[-N\left(D(f_1\|f_2)+\delta\right)],
\end{aligned}
$$

where the last inequality is due to the following:

$$\begin{aligned}\sum_{A_N \cup B_N} f_1(x) &= f_1(A_N \cup B_N), \\ &= 1 - f_1(A_N^c \cap B_N^c), \\ &\geq 1 - f_1(A_N^c) - f_1(B_N^c), \\ &= 1 - \alpha_N - \alpha_{N,B_N}.\end{aligned}$$

Hence, $\frac{1}{N}\log\beta_{N,B_N} \geq -D(f_1\|f_2) - \delta - \frac{\log(1-\alpha_N-\alpha_{N,B_N})}{N}$, and since $\delta > 0$, $\lim_{n\to\infty}\frac{1}{N}\log\beta_{N,B_N} \geq -D(f_1\|f_2)$. Thus no sequence of sets B_N has an exponent better than $D(f_1\|f_2)$.

In summary, property 1. shows that A_N is the sequence that is generated by f_1 in the asymptotic limit. Properties 2., 3. and 4. derive the error probability of Stein's lemma and property 5. shows that A_N is asymptotically optimal and the best error exponent is $D(f_1\|f_2)$.

Thus, we can interpret $\exp\left[-D\left(\text{truth}\|\text{model}\right)\right]$ as the probability of confusing the model with the truth or model probability, using the work of Balasubramanian [52].

This relation allows us to propose a slightly different but related way of interpreting the AIC difference between models as defining the ratio between model probabilities $\frac{P(A)}{P(B)}$ without resorting to Akaike weights. Let us start with 2 models A and B with f_A and f_B as their respective probability distribution functions. Their AIC values are a and b respectively and there is a difference of $\Delta_{A,B} = a - b$ between their AIC values.

$$\begin{aligned}\exp[-\Delta_{A,B}/2] &= \frac{\exp[-a/2]}{\exp[-b/2]}, \\ &\approx \frac{\exp\left[E_{X|\theta_0}[\log f_A(X|\hat{\theta}_A)]\right]}{\exp\left[E_{X|\theta_0}[\log f_B(X|\hat{\theta}_B)]\right]}, \\ &= \frac{\exp\left[-D\left(\text{truth}\|f_A(X|\hat{\theta}_A)\right)\right]}{\exp\left[-D\left(\text{truth}\|f_B(X|\hat{\theta}_B)\right)\right]}, \\ &= \frac{P(A)}{P(B)},\end{aligned}$$

where X represents data sampled from the truth θ_0. $\hat{\theta}_A$ and $\hat{\theta}_B$ are the maximum likelihood parameters of model A and B.

Appendix B: Bootstrap Method

The bootstrap method used in this paper was a parametric bootstrap, where we made certain assumptions about the parametric relationship between the data input (explanatory variable) and the data output (response variable). We start with a set of N actual data points $D = \{(x_1, y_1, \sigma_1), (x_2, y_2, \sigma_2), ..., (x_N, y_N, \sigma_N)\}$, where y_i is the response variable that is observed with the error bar σ_i and x_i the explanatory (input) variable. To produce a bootstrap sample set C_N consisting of N data points, as noted in the text, we use a particular model C: $y = f(x)$ with the needed parameters chosen by maximum likelihood estimation. In the usual bootstrap approach, the obtained set $\{y_i - f(x_i)\}$ is regarded as the estimate of the noise distribution, but in our case, unfortunately, the noise magnitude seems to depend on x. Therefore, we make an example of size N bootstrap data as $\{f(x_i) + \epsilon_i\}$, where x_i are the same as in D and the noise ϵ_i is a Gaussian random variable obeying $N(0, \sigma_i^2)$. With newly generated $\{\epsilon_i\}$ we make a set of N sample bootstrap data set C_N. For each bootstrap sample $d \in C_N$, we estimate the maximum likelihood parameters for model A and B, respectively, and we can compute the set of AIC differences $\{\Delta^d_{A,B} \mid d \in C_N\}$.

Appendix C: Getting χ^2_{ML} from SNIa, marginalizing over H_0

We present the use of AIC in the context of cosmological model selection using SNIa from the Constitution compilation [14]. We fit the theoretical quantity of the distance moduli $\mu(z_i)$ (z is the observed red shift) against its observed value $\mu_i^{obs} = m_i - M_i$, where m_i is the observed apparent magnitude and M_i is the absolute magnitude of the Supernova data. Note that the index i indicates the ith data point.

$\mu(z)$ is calculated by the equation $\mu(z) = 5 \log_{10} \frac{d_L(z)}{10pc}$, where $d_L(z)$ is the luminosity distance. We will assume that the universe is flat, by setting the curvature term Ω_k in the Hubble function $H(z)$ to zero, and under this assumption

$$d_L(z) = (1 + z)c \int_0^z \frac{dz'}{H(z')}, \tag{C1}$$

where c is the speed of light.

We start with the assumption that μ_i^{obs} has a Gaussian noise structure. We model it as

$$\frac{1}{\prod_i \sqrt{2\pi\sigma_i^2}} \exp\left[-\sum_i \frac{(\mu_i^{obs} - \mu(z_i))^2}{2\sigma_i^2}\right], \tag{C2}$$

where σ_i are consistent with the error bars associated with μ_i^{obs} in the Constitution compilation. Since μ^{obs} was calculated from the apparent (observed) magnitude m by assuming a fixed absolute magnitude M value which we are actually unsure about. We get around this problem by introducing a nuisance parameter g and integrate it over a flat prior. To do this, we first integrate this over a Gaussian distribution of the nuisance parameter with standard deviation σ_g^2 to get

$$\int_{-\infty}^{\infty} dg \frac{1}{\prod_i \sqrt{2\pi\sigma_i^2}} \exp\left[-\sum_i \frac{(\mu_i^{obs} - \mu(z_i) - g)^2}{2\sigma_i^2}\right] \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left[-\frac{g^2}{2\sigma_g^2}\right]. \tag{C3}$$

We can rewrite this in matrix form:

$$\int_{-\infty}^{\infty} dg \frac{1}{\prod_i \sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(X - gY)^T M (X - gY)}{2}\right] \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left[-\frac{g^2}{2\sigma_g^2}\right], \tag{C4}$$

where X is a n-vector whose ith component is $\mu_i^{obs} - \mu(z_i)$, Y is a n-vector where all the elements are '1s' and M is the inverse of the covariance matrix (which in this case is diagonal). The ^{T} symbol denotes the transpose of a vector. Performing the g integral in C4 and setting σ_g to a large value, the following marginalized function is obtained:

$$\frac{1}{\sqrt{\sigma_g^2 Y^T M Y}} \frac{1}{\prod_i \sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2} X^T \left(M - \frac{M Y Y^T M}{Y^T M Y}\right) X\right]. \tag{C5}$$

This reduces the log likelihood to $\frac{X^T C X}{2} - \frac{1}{2}\log(\sigma_g^2 Y^T M Y) - \sum_i \frac{1}{2}\log(2\pi\sigma_i^2)$. The second term suffers from a log divergence as $\sigma_g \to \infty$. However, since AIC works by comparing relative log likelihood values, we can regularize this term away by setting it to zero. We can also ignore the third term since it is a fixed constant independent of the parameter choice.

This reduces finding the maximum likelihood to minimizing $X^T C X$, where $C = M - \frac{M Y Y^T M}{Y^T M Y}$.

Because of the marginalization against a flat prior, the rank of C is smaller than the rank of M by one and thus C cannot be inverted. The marginalization procedure also implies that the choice of the Hubble parameter H_0 and even the speed of light c is irrelevant to finding the maximum likelihood values of the other parameters.

This leads to the following relative AIC term:

$$X^T(\hat{\theta})CX(\hat{\theta}) + 2k, \tag{C6}$$

where $\hat{\theta}$ is the set of parameters that minimizes $X^T CX$.

The first term corresponds to the maximum likelihood while the second term is the bias correction. The maximum likelihood parameters can then be found using some common minimization procedure. Specifically, these were found using the Gauss-Newton algorithm [53]. This allows us to calculate the AIC values for 4 candidate models. The data used consists of 371 Supernova events taken from the Constitution compilation (MLCS table) [14]. We computed the AIC values for the different models and found that the DGP model has the smallest AIC value among the four models we considered in this paper.

As a very small technical side issue, it should be noted that unlike the other work [54] that used AIC as a model selection tool, as described in the above, we marginalized away the H_0 term against a flat prior which reduces the number of free parameters by one. This technical difference alone should not affect our use of AIC.

Appendix D: P-table

As a sanity check, we tried to ensure that the statistical proportions are unlikely to be generated by random chance. We computed the P-values of the pairs, using a non parametric sign test, in order to test the validity of these percentages. In general, the smaller the P-value, the less likely we are able to get the result through a random fluke. It should be noted that the value 0. means that the P-value is approximately zero up to the machine precision of the computer. We can see from the extremely small P-values that the results on the proportion of differences are reliable. That is due to the fact that we generated 1000

TABLE III: P-Values

X\Y	DGP	ΛCDM	wCDM	CPL
DGP	-	3.81065×10^{-8}	0.	0.
ΛCDM	1.73783×10^{-12}	-	0.	0.
WCDM	0.	0.	-	0.
CPL	0.	0.	0.	-

bootstrap samples.

[1] R. A. Fisher, Nature **123**, 866 (1925).

[2] H. Jeffreys, Theory of Probability (Oxford University Press, USA, 1961), 3rd ed., ISBN 0198503687, URL http://www.worldcat.org/isbn/0198503687.

[3] H. Akaike, IEEE Transactions on Automatic Control **19**, 716 (1974).

[4] G. Schwarz, The Annals of Statistics **6**, 461 (1978), ISSN 00905364, URL http://dx.doi.org/10.2307/2958889.

[5] A. Barron, J. Rissanen, and B. Yu, IEEE Transactions on Information Theory **44**, 2743 (1998), URL http://dx.doi.org/10.1109/18.720554.

[6] S. Perlmutter, G. Aldering, G. Goldhaber, R. A. Knop, P. Nugent, P. G. Castro, S. Deustua, S. Fabbro, A. Goobar, D. E. Groom, et al., Astrophys. J. **517**, 565 (1999), arXiv:astro-ph/9812133.

[7] A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner, et al., Astron. J. **116**, 1009 (1998), arXiv:astro-ph/9805201.

[8] P. M. Garnavich, S. Jha, P. Challis, A. Clocchiatti, A. Diercks, A. V. Filippenko, R. L. Gilliland, C. J. Hogan, R. P. Kirshner, B. Leibundgut, et al., Astrophys. J. **509**, 74 (1998),

arXiv:astro-ph/9806396.

[9] R. A. Knop, G. Aldering, R. Amanullah, P. Astier, G. Blanc, M. S. Burns, A. Conley, S. E. Deustua, M. Doi, R. Ellis, et al., Astrophys. J. **598**, 102 (2003), arXiv:astro-ph/0309368.

[10] J. L. Tonry, B. P. Schmidt, B. Barris, P. Candia, P. Challis, A. Clocchiatti, A. L. Coil, A. V. Filippenko, P. Garnavich, C. Hogan, et al., Astrophys. J. **594**, 1 (2003), arXiv:astro-ph/0305008.

[11] A. G. Riess, L.-G. Strolger, J. Tonry, S. Casertano, H. C. Ferguson, B. Mobasher, P. Challis, A. V. Filippenko, S. Jha, W. Li, et al., Astrophys. J. **607**, 665 (2004), arXiv:astro-ph/0402512.

[12] P. Astier, J. Guy, N. Regnault, R. Pain, E. Aubourg, D. Balam, S. Basa, R. G. Carlberg, S. Fabbro, D. Fouchez, et al., Astron. Astrophys. **447**, 31 (2006), arXiv:astro-ph/0510447.

[13] W. M. Wood-Vasey, G. Miknaitis, C. W. Stubbs, S. Jha, A. G. Riess, P. M. Garnavich, R. P. Kirshner, C. Aguilera, A. C. Becker, J. W. Blackman, et al., Astrophys. J. **666**, 694 (2007), arXiv:astro-ph/0701041.

[14] M. Hicken, W. M. Wood-Vasey, S. Blondin, P. Challis, S. Jha, P. L. Kelly, A. Rest, and R. P. Kirshner, Astrophys. J. **700**, 1097 (2009), 0901.4804.

[15] W. L. Freedman, C. R. Burns, M. M. Phillips, P. Wyatt, S. E. Persson, B. F. Madore, C. Contreras, G. Folatelli, E. S. Gonzalez, M. Hamuy, et al., Astrophys. J. **704**, 1036 (2009), 0907.4524.

[16] R. Kessler, A. C. Becker, D. Cinabro, J. Vanderplas, J. A. Frieman, J. Marriner, T. M. Davis, B. Dilday, J. Holtzman, S. W. Jha, et al., Astrophys. J. Suppl. Ser. **185**, 32 (2009), 0908.4274.

[17] T. Schrabback, J. Hartlap, B. Joachimi, M. Kilbinger, P. Simon, K. Benabed, M. Bradač, T. Eifler, T. Erben, C. D. Fassnacht, et al., Astron. Astrophys. **516**, A63+ (2010), 0911.0053.

[18] C. Blake, T. Davis, G. Poole, D. Parkinson, S. Brough, M. Colless, C. Contreras, W. Couch, S. Croom, M. J. Drinkwater, et al., ArXiv e-prints (2011), 1105.2862.

[19] J. A. Frieman, C. T. Hill, A. Stebbins, and I. Waga, Physical Review Letters **75**, 2077 (1995), arXiv:astro-ph/9505060.

[20] B. Ratra and P. J. E. Peebles, Phys. Rev. D **37**, 3406 (1988).

[21] C. Wetterich, Nuclear Physics B **302**, 668 (1988).

[22] I. Zlatev, L. Wang, and P. J. Steinhardt, Physical Review Letters **82**, 896 (1999), arXiv:astro-ph/9807002.

[23] C. Deffayet, Physics Letters B **502**, 199 (2001), arXiv:hep-th/0010186.

[24] G. Dvali, G. Gabadadze, and M. Porrati, Physics Letters B **485**, 208 (2000), arXiv:hep-th/0005016.

[25] S. M. Carroll, A. de Felice, V. Duvvuri, D. A. Easson, M. Trodden, and M. S. Turner, Phys. Rev. D **71**, 063513 (2005), arXiv:astro-ph/0410031.

[26] T. Abbott et al. (Dark Energy Survey) (2005), astro-ph/0510346.

[27] LSST Science Collaborations and LSST Project., ArXiv e-prints (2009), 0912.0201.

[28] D. J. Schlegel et al. (2009), 0904.0468.

[29] URL `http://sci.esa.int/science-e/www/area/index.cfm?fareaid=102`.

[30] A. R. Liddle, Mon. Not. Roy. Astron. Soc. **351**, L49 (2004), astro-ph/0401198.

[31] A. R. Liddle, P. Mukherjee, D. Parkinson, and Y. Wang, Phys. Rev. **D74**, 123506 (2006), astro-ph/0610126.

[32] M. Szydłowski and W. Godłowski, Physics Letters B **633**, 427 (2006), arXiv:astro-ph/0509415.

[33] M. Szydłowski, A. Kurek, and A. Krawiec, Physics Letters B **642**, 171 (2006), arXiv:astro-ph/0604327.

[34] A. R. Liddle, P. Mukherjee, and D. Parkinson, Astron. Geophys. **47**, 4.30 (2006), astro-ph/0608184.

[35] A. R. Liddle, Annual Review of Nuclear and Particle Science **59**, 95 (2009), 0903.4210.

[36] R. Trotta, M. Kunz, and A. R. Liddle (2010), 1012.3195.

[37] R. Trotta, Contemp. Phys. **49**, 71 (2008), 0803.4089.

[38] J. Sollerman, E. Mörtsell, T. M. Davis, M. Blomqvist, B. Bassett, A. C. Becker, D. Cinabro, A. V. Filippenko, R. J. Foley, J. Frieman, et al., Astrophys. J. **703**, 1374 (2009), 0908.4276.

[39] T. M. Davis, E. Mörtsell, J. Sollerman, A. C. Becker, S. Blondin, P. Challis, A. Clocchiatti, A. V. Filippenko, R. J. Foley, P. M. Garnavich, et al., Astrophys. J. **666**, 716 (2007), arXiv:astro-ph/0701510.

[40] A. R. Liddle, Mon. Not. Roy. Astron. Soc. **377**, L74 (2007), arXiv:astro-ph/0701113.

[41] G.-B. Zhao, J.-Q. Xia, H. Li, C. Tao, J.-M. Virey, Z.-H. Zhu, and X. Zhang, Physics Letters B **648**, 8 (2007), arXiv:astro-ph/0612728.

[42] M. Hicken, P. Challis, S. Jha, R. P. Kirshner, T. Matheson, M. Modjaz, A. Rest, W. M. Wood-Vasey, G. Bakos, E. J. Barton, et al., Astrophys. J. **700**, 331 (2009), 0901.4787.

[43] G. Miknaitis, G. Pignata, A. Rest, W. M. Wood-Vasey, S. Blondin, P. Challis, R. C. Smith, C. W. Stubbs, N. B. Suntzeff, R. J. Foley, et al., Astrophys. J. **666**, 674 (2007), arXiv:astro-

ph/0701043.

[44] A. G. Riess, L.-G. Strolger, S. Casertano, H. C. Ferguson, B. Mobasher, B. Gold, P. J. Challis, A. V. Filippenko, S. Jha, W. Li, et al., Astrophys. J. **659**, 98 (2007), arXiv:astro-ph/0611572.

[45] H. Akaike, Bulletin of the International Statistical Institute **44**, :277 (1983).

[46] K. P. Burnham and D. R. Anderson, Sociological Methods & Research **33**, 261 (2004), ISSN 0049-1241, URL `http://dx.doi.org/10.1177/0049124104268644`.

[47] H. Shimodaira, Annals of the Institute of Statistical Mathematics **49**, 395 (1997), URL `http://ideas.repec.org/a/spr/aistmt/v49y1997i3p395-410.html`.

[48] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap (Chapman & Hall, New York, 1993).

[49] T. M. Cover and J. A. Thomas, Elements of information theory (Wiley-Interscience, New York, NY, USA, 1991), ISBN 0-471-06259-6, URL `http://portal.acm.org/citation.cfm?id=129837`.

[50] E. V. Linder, Physical Review Letters **90**, 091301 (2003), arXiv:astro-ph/0208512.

[51] M. Chevallier and D. Polarski, International Journal of Modern Physics D **10**, 213 (2001), arXiv:gr-qc/0009008.

[52] V. Balasubramanian, Neural Comput. **9**, 349 (1997), ISSN 0899-7667, URL `http://dx.doi.org/10.1162/neco.1997.9.2.349`.

[53] Å. Björck, Numerical Methods for Least Squares Problems (SIAM, Philadelphia, 1996).

[54] W. Godłowski and M. Szydłowski, Physics Letters B **623**, 10 (2005), arXiv:astro-ph/0507322.

[55] Note that a KL divergence is a non-symmetric measure that does not obey the triangle inequality.

[56] This is due to a statistical uncertainty in the AIC values coming from the variation of χ^2_{ML}. However, the uncertainty in the AIC values of the models can be correlated, and turns out to be larger than the uncertainty in the AIC differences.

[57] In this paper, two Supernovae data sets are said to have the same quality when they have the same number of data points, the same set of redshift z values and the same set of error bars(standard deviation) that corresponds to the set of z values.

[58] Note that these errors which include light curve fitting errors, intrinsic dispersion and peculiar velocity corrections are also required for calculating quantities like χ^2 for most model selection schemes. Thus, an underlying assumption of the application of the AIC technique as in previous

works is that these error estimates are correct. Since our focus is on the statistical uncertainties in AIC after following other underlying assumptions used in the literature, we also assume that these error estimates are correct.