On the non-Gaussian errors in high-z supernovae type Ia data

Meghendra Singh[1], Ashwini Pandey[2], Amit Sharma[2], Shashikant Gupta[2] and Satendra Sharma[3]

[1] Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh, Lucknow 226021, India;
   meghendrasingh_db@yahoo.co.in
[2] GD Goenka University, Gurgaon, Haryana 122103, India
[3] Yobe State University, Damaturu, Yobe State, Nigeria

Abstract The nature of random errors in any data set is Gaussian, which is a well established fact according to the Central Limit Theorem. Supernovae type Ia data have played a crucial role in major discoveries in cosmology. Unlike in laboratory experiments, astronomical measurements cannot be performed in controlled situations. Thus, errors in astronomical data can be more severe in terms of systematics and non-Gaussianity compared to those of laboratory experiments. In this paper, we use the Kolmogorov-Smirnov statistic to test non-Gaussianity in high-z supernova data. We apply this statistic to four data sets, i.e., Gold data (2004), Gold data (2007), the Union2 catalog and the Union2.1 data set for our analysis. Our results show that in all four data sets the errors are consistent with a Gaussian distribution.

Key words: cosmology — data analysis — statistics — probability

1 INTRODUCTION

The light curves of Type Ia supernovae (SNeIa) have been used as cosmological distance indicators (Riess et al. 1998; Perlmutter et al. 1999) to trace the expansion history and detect cosmic acceleration. The overall picture of the Universe is consistent with a model known as ΛCDM, consisting of around one quarter baryonic and dark matter and three quarters dark energy. The dark energy can be treated as a cosmic-fluid with equation of state $P = w\rho$, where the pressure (P) is allowed to be negative. The SNIa data can be used to constrain the equation of state parameter (w) which is the key to study dark energy (Freedman et al. 2009; Hicken et al. 2009; Rest et al. 2014; Scolnic et al. 2014).

However, many alternative explanations exist for dark energy and its exact nature is also unknown. For instance, a classical fixed cosmological constant, Λ, yields $w = -1$, whereas other models (e.g. quintessence) yield values of $w > -1$ (Huterer & Turner 2001). To overcome this difficulty, precise enough data are required to detect fluctuations in the dark energy. The data should also cover a wide range of redshifts to constrain the detailed behavior of dark energy with time. Presently, data fulfilling the above criteria are obtained by observations of SNeIa. Determination of supernova (SN) distances having high precision and tiny systematic errors is crucial for the above purpose; and we would like to be certain that their statistics are well understood. Furthermore, if the Central Limit Theorem holds (Kendall & Stuart 1977), then statistical uncertainties in

SNIa data should follow a normal distribution. The systematics, if present, have to be identified and removed separately. Treatment of the errors becomes more important in astronomy since it is hard to repeat or perform the related experiments in a controlled way, unlike laboratory experiments in other fields. In the present paper, we use the Kolmogorov-Smirnov test (hereafter KS test) in an elegant way to detect non-Gaussian uncertainties in SNIa data.

This paper aims to address the above mentioned problems. The rest of the paper is structured as follows: In Section 2, we illustrate the different data sets used for our analysis, while Section 3 contains a detailed description of methodology used. In Section 4, we continue and put forward our results for various data sets and lastly Section 5 is reserved for conclusions.

2 DATA

The Gold data GD04 (Riess et al. 2004) containing 157 SNe, GD07 (Riess et al. 2007) consisting of 182 SNe along with the more recent and larger data sets Union2 (Amanullah et al. 2010) and Union2.1 (Suzuki et al. 2012) composed of 557 and 580 SNe respectively are used to carry out our investigation. The redshift z and distance modulus μ are the measured quantities in the data. If m is the apparent magnitude and M is the absolute magnitude, then distance modulus is defined as

$$\mu(z) = m(z) - M . \tag{1}$$

The apparent magnitude $m(z)$ and hence distance modulus $\mu(z)$ depend on the intrinsic luminosity of an SN, its

redshift z and the cosmological parameters. The distance modulus $\mu(z)$ and luminosity distance d_{L} are related by

$$\mu(z) = 5 \log \left(d_{\mathrm{L}}(z)\right) + 25 \,, \qquad (2)$$

where the luminosity distance is measured in Mpc and follows

$$d_{\mathrm{L}}(z) = \frac{c(1+z)}{H_0} \int_0^z \frac{dx}{h(x)} \,, \qquad (3)$$

where $h(z; \Omega_{\mathrm{M}}, \Omega_{\mathrm{X}}) = H(z; \Omega_{\mathrm{M}}, \Omega_{\mathrm{X}})/H_0$; hence it is independent of H_0 and only depends on densities of dark matter (Ω_{M}) and dark energy (Ω_{X}). The variation of Ω_{X} with redshift is already encoded in the associated cosmological models; for instance Ω_{X} is a constant in the ΛCDM model. Although the nature of the relation between μ and M is linear, that with luminosity distance is logarithmic. This also implies the logarithmic dependence of μ on the Hubble parameter H_0.

3 METHODOLOGY

We now give an introduction to the method used in our analysis. Originally, this method was described in Singh et al. 2015 (hereafter GS15) to find non-Gaussianity in the HST Key Project Data.

If the correct theoretical value of the distance modulus of the ith supernova at redshift z is $\mu_i^{\mathrm{th}}(z)$, then the observed value μ_i^{obs} will be

$$\mu_i^{\mathrm{obs}} = \mu_i^{\mathrm{th}}(z) \pm \sigma_i, \qquad (4)$$

where σ_i is the error in the measurement of distance modulus. We expect these errors to be completely random; however, there could be some undesired contribution from systematic effects. For the time being we assume that the systematic part in the errors is zero. We show in the next paragraph that the presence of systematic errors will not affect our analysis. Furthermore, the Central Limit Theorem suggests that the random part of the errors should be Gaussian in nature with mean value zero. Now we define a quantity χ_i such that

$$\chi_i = \frac{\mu_i^{\mathrm{obs}} - \mu_i^{\mathrm{th}}(z)}{\sigma_i}. \qquad (5)$$

Clearly χ_i should follow the standard normal distribution $N(0,1)$, i.e., a Gaussian distribution with zero mean and unit variance. The effect of random errors is to scatter the data around the true value and that of systematics is to shift the average away from the true value. If systematics are present then they will just shift the average, hence one should subtract the best-fit value rather than the true theoretical value in Equation (5). Thus Equation (5) takes the following form for a given SN

$$\chi_i = \frac{\mu_i^{\mathrm{obs}} - \mu_i^{\mathrm{bestfit}}(z)}{\sigma_i}, \qquad (6)$$

where $\mu_i^{\mathrm{bestfit}}(z)$ is calculated using the best-fit values of cosmological parameters. Statistical independence among SNe in our analysis is an obvious assumption. χ_i defined in

Equation (6) should follow a standard normal distribution, i.e., Gaussian with zero mean and unit standard deviation.

We use the flat ΛCDM cosmology in our analysis, since it fits the SNe data well. However, other cosmological models could also be investigated using a similar approach. In order to get best-fit values of cosmological parameters, we minimize χ^2 which is defined as

$$\chi^2 = \Sigma_{i=1}^N \left[\frac{\mu^i - \mu^{\Lambda \mathrm{CDM}}}{\sigma_i}\right]^2. \qquad (7)$$

Once again, we emphasize that Equation (7) is used to find the best-fit values of cosmological parameters and it is then used in Equation (6) to calculate χ_i.

As argued earlier, χ_i should follow the standard normal distribution. To check this, we use the KS test to determine whether or not a given sample follows a Gaussian distribution (Press 2007). For this we define our null hypothesis as: "The errors in the SNe data are drawn from a Gaussian distribution." Thus χ_i values in Equation (6) would follow the standard normal distribution. We apply the KS test to calculate the test statistic and the p value which is the probability of attaining the observed sample results when the null hypothesis is true.

For this, we use the Matlab function *kstest* $[h, p, k, C_{\mathrm{v}}]$ where p represents the probability of the data errors being drawn from a Gaussian distribution, k is the maximum distance between the two cumulative distribution functions (CDFs), and C_{v} is the critical value which is decided by the significance level α. Different values of α indicate different tolerance levels for a false rejection of the null hypothesis. For instance, $\alpha = 0.01$ means that we allow 1% of repeated trials to reject the null hypothesis when it is true. C_{v} is the critical value of the probability to obtain/generate the data set in question given the null hypothesis; and it can be compared with p. A value of $h = 1$ is returned by the test if $p < C_{\mathrm{v}}$ and the null hypothesis is rejected. However for $p > C_{\mathrm{v}}$, h remains 0 and the null hypothesis is not rejected.

4 RESULTS

We apply the statistic discussed in Section 3 to various SNe data sets and present the results here. Similar analysis was presented in Gupta & Saini 2010 (hereafter GS10) and in Gupta & Singh 2014 (hereafter GS14) using a different method ($\Delta \chi^2$) based on extreme value theory.

As a first check, we calculate the best-fit values of cosmological parameters for all four data sets by minimizing χ^2 which are presented in Table 1. It is clear that both Gold data sets favor higher matter density (Ω_{M}) and consequently smaller expansion rate (H_0) compared to the Union2 and Union2.1 data sets. One important fact is that the χ^2 per degree of freedom generates the smallest value for GD07 while it is the largest for GD04, indicating the overestimation and underestimation of errors in GD07 and GD04 respectively.

We calculate χ_i values as defined in Equation (6) for each data set using the best-fit values presented in Table 1.
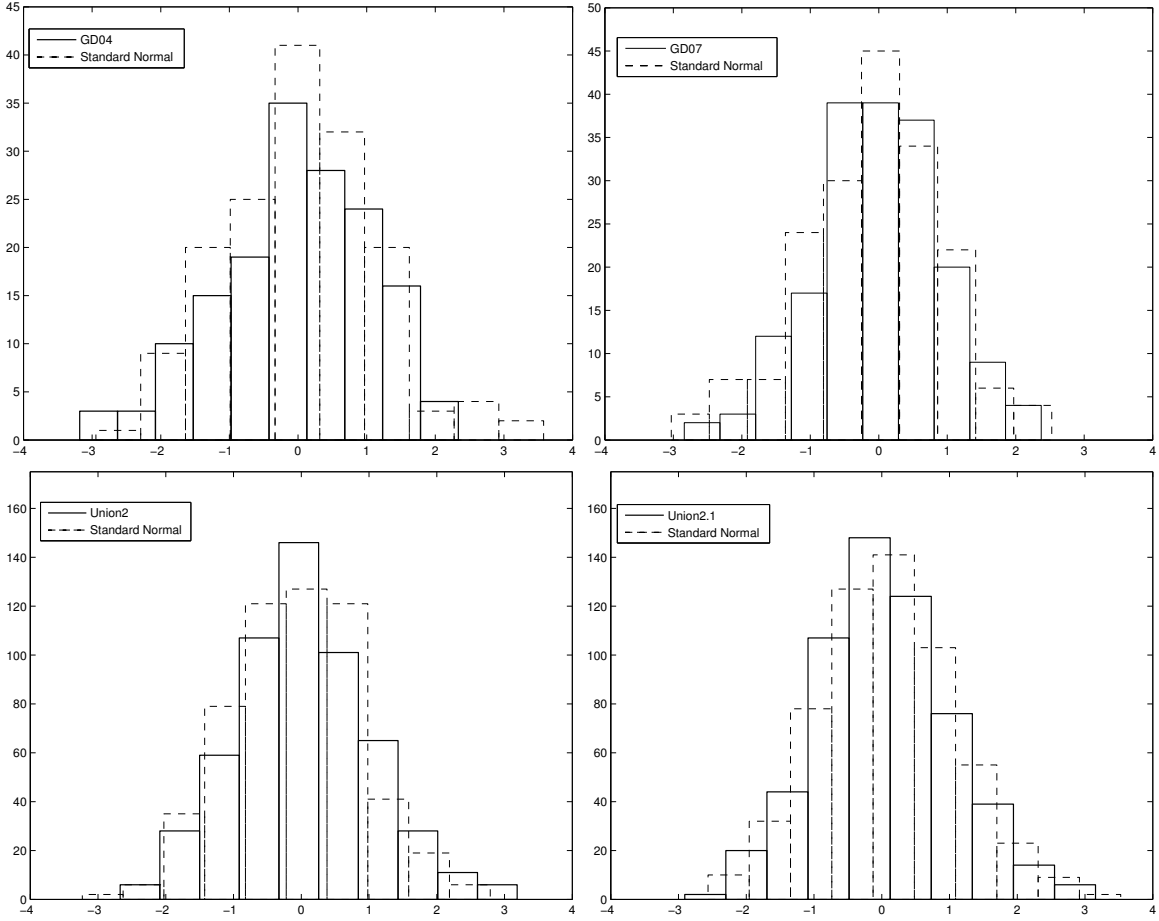
Fig. 1 Histograms of χ_i for various data sets are compared with those of a standard normal distribution.

Furthermore, we generate four sets of random numbers following a Gaussian distribution with zero mean and unit standard deviation. Figure 1 presents a comparison of histograms of Gaussian random numbers with those of χ_i values for each data set.

Secondly, the results of the KS test, which are arrived at by comparing the calculated CDFs for χ_i values with those of Gaussian distributions, are presented in Table 2. The second, third and fourth columns in Table 2 denote values of p, k and C_v respectively. Since $p > C_v$ in all cases gives $h = 0$, this means that we cannot reject the null hypothesis that the errors are drawn from a Gaussian distribution. This is shown explicitly by Figure 2.

5 CONCLUSIONS

We have used the method presented in GS15 to detect non-Gaussianity in the error bars in SNe data. Our main conclusions for this part of our work are as follows: (a) The errors are probably underestimated in GD04 and overestimated in GD07. In this sense, both of the sets represent extreme positions. (b) For a flat ΛCDM cosmology, GD07 favors slightly higher matter density and this can be verified by the fact that in GD07 the distances are smaller compared to those in the GD04 set for common SNe. (c)

Table 1 The Best-fit Values for Various Data Sets

Data Set	# SNe	Ω_M	H_0	χ^2/dof
(1)	(2)	(3)	(4)	(5)
GD04	157	0.30	64.5	1.143
GD07	182	0.33	63.0	0.883
Union2	557	0.27	70.0	0.975
Union2.1	580	0.28	70.0	0.973

Table 2 Results of the KS Test for Various Data Sets

Data Set	p value	k	C_v
(1)	(2)	(3)	(4)
GD04	0.9280	0.0425	0.1073
GD07	0.7872	0.0475	0.0997
Union2	0.7328	0.0288	0.0572
Union2.1	0.6764	0.0296	0.0561

In comparison with GS10 and GS14, GD04 was shown to have a non-Gaussian component for errors while the KS test shows the highest probability of being consistent with a Gaussian distribution. (d) The hypothesis that the errors are drawn from a Gaussian distribution cannot be rejected for all of the data sets discussed in the present paper.
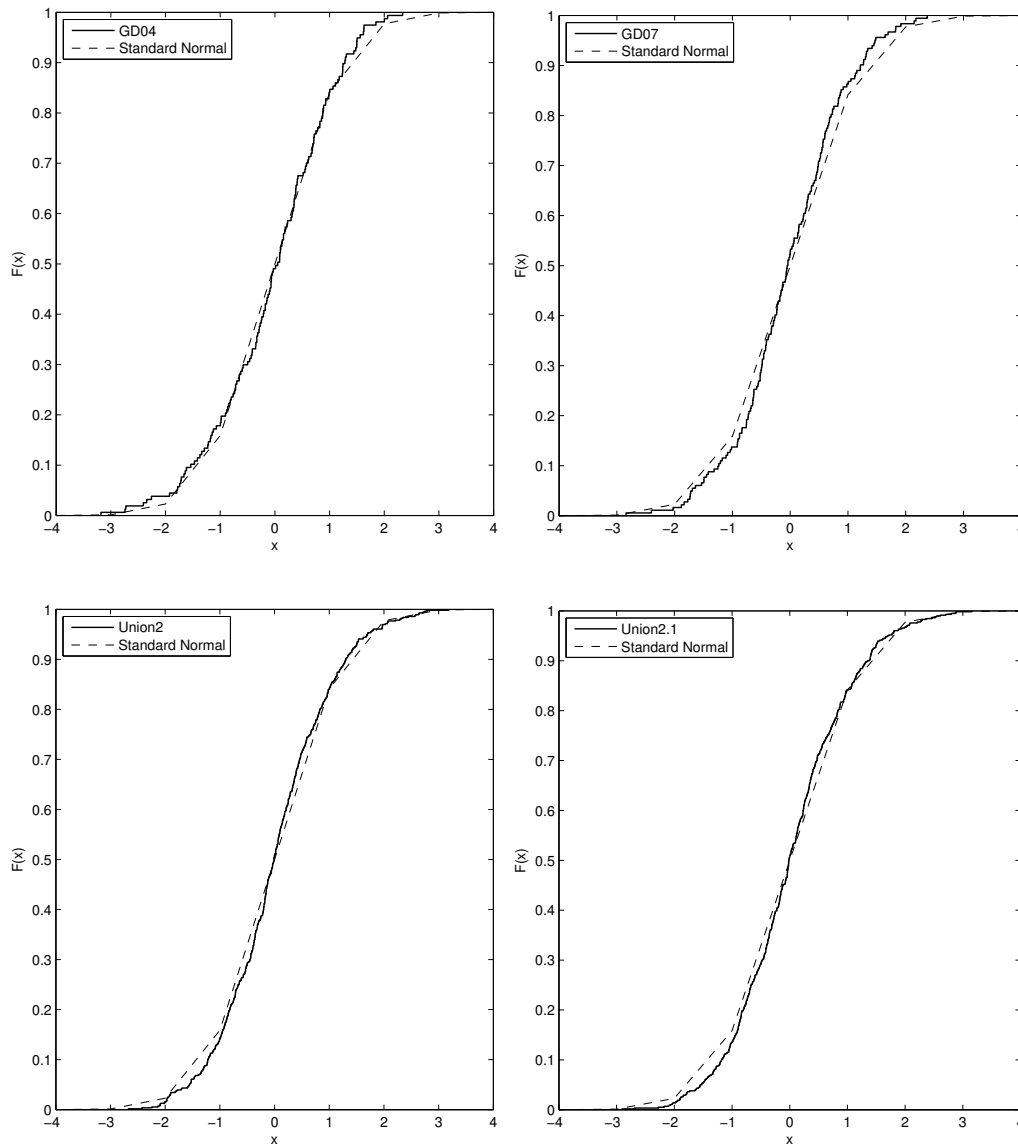
Fig. 2 Comparison of CDF of χ_i for different data sets with their corresponding Gaussian CDF. The smooth curve represents the Gaussian CDF.

References

Amanullah, R., Lidman, C., Rubin, D., et al. 2010, ApJ, 716, 712

Freedman, W. L., Burns, C. R., Phillips, M. M., et al. 2009, ApJ, 704, 1036

Gupta, S., & Saini, T. D. 2010, MNRAS, 407, 651

Gupta, S., & Singh, M. 2014, MNRAS, 440, 3257

Hicken, M., Wood-Vasey, W. M., Blondin, S., et al. 2009, ApJ, 700, 1097

Huterer, D., & Turner, M. S. 2001, Phys. Rev. D, 64, 123527

Kendall, M., & Stuart, A. 1977, The Advanced Theory of Statistics. 1, Distribution Theory (4th edn.; London: Griffin)

Perlmutter, S., Aldering, G., Goldhaber, G., et al. 1999, ApJ, 517, 565

Press, W. H. 2007, Numerical Recipes 3rd Edition: The Art of Scientific Computing (Cambridge: Cambridge Univ. Press)

Rest, A., Scolnic, D., Foley, R. J., et al. 2014, ApJ, 795, 44

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, AJ, 116, 1009

Riess, A. G., Strolger, L.-G., Tonry, J., et al. 2004, ApJ, 607, 665

Riess, A. G., Strolger, L.-G., Casertano, S., et al. 2007, ApJ, 659, 98

Scolnic, D., Rest, A., Riess, A., et al. 2014, ApJ, 795, 45

Singh, M., Gupta, S., Pandey, A., & Sharma, S. 2015, JCAP08(2016)026

Suzuki, N., Rubin, D., Lidman, C., et al. 2012, ApJ, 746, 85