χ² versus median statistics in SNIa data analysis

A. Barreira¹,²† and P.P. Avelino³,²†

¹Centro de Física do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
²Departamento de Física da Faculdade de Ciências da Universidade do Porto,
Rua do Campo Alegre 687, 4169-007 Porto, Portugal
³Centro de Astrofísica da Universidade do Porto, Rua das Estrelas, 4150-762 Porto, Portugal

In this paper we compare the performances of the χ² and median likelihood analysis in the determination of cosmological constraints using type Ia supernovae data. We perform a statistical analysis using the 307 supernovae of the Union 2 compilation of the Supernova Cosmology Project and find that the χ² statistical analysis yields tighter cosmological constraints than the median statistic if only supernovae data is taken into account. We also show that when additional measurements from the Cosmic Microwave Background and Baryonic Acoustic Oscillations are considered, the combined cosmological constraints are not strongly dependent on whether one applies the χ² statistic or the median statistic to the supernovae data. This indicates that, when complementary information from other cosmological probes is taken into account, the performances of the χ² and median statistics are very similar, demonstrating the robustness of the statistical analysis.

I. INTRODUCTION

More than a decade ago, type Ia supernovae (SNIa) provided the first clear evidence in favor of cosmic acceleration. Since then, the availability of ever larger, higher-quality SNIa datasets, as well as measurements using other cosmological probes, such as the Cosmic Microwave Background (CMB) or the Baryonic Acoustic Oscillations (BAO), have been providing overwhelming evidence for the existence of dark energy, a fluid with large negative pressure capable of driving the acceleration of the universe.

In the so-called ΛCDM model, also known as the concordance model, the dark energy role is played by a cosmological constant Λ, responsible for approximately 73% of the energy density of the universe at the present day. The remaining percentage is mainly in the form of cold dark matter, most of which non-baryonic and dark. Radiation is residual at the present time and the universe is spatially flat. Despite the good agreement with observational data, this model has little appeal on theoretical grounds since the value of Λ required to explain the observed cosmic acceleration is off by ~ 120 orders of magnitude from the standard quantum field theory prediction. The coincidence between our observing time and the time of the onset of cosmic acceleration is also puzzling.

Many other dark energy models have been proposed, with the most influential being the ones based on dynamical scalar fields. In these models, the energy density varies with time and suitable choices of the scalar field lagrangian can relax some of the problems associated with the cosmological constant (see, for instance, the review [8]). As a result, an important step towards a better understanding of the dark energy involves further testing of its possible dynamical nature.

The determination of dark energy constraints from observational data requires a robust statistical data analysis. In the particular case of SNIa, the usual procedure is to carry out a standard χ² likelihood analysis. However, in [9, 10] it has been argued that the median statistic could be a more reliable alternative. The use of the median, despite having the drawback of not being as constraining as the χ² analysis, has the strong advantage of requiring weaker assumptions about the data, thus yielding more trustworthy constraints. Moreover, the median is less vulnerable to the presence of “outliers”, which is also a significant advantage given current uncertainties about the physics of SNIa.

In this work we revisit and extend the analysis of [10], updating it using the recent Union 2 SNIa compilation of the Supernova Cosmology Project (SCP) [8]. We compare the performances of χ² and median statistics in the determination of cosmological constraints using SNIa data, considering also CMB and BAO measurements. The layout of this paper is as follows. In Sec. II we describe the application of χ² and median statistics to SNIa data. We also discuss the use of additional information contained in the CMB and BAO. In Sec. III we present and discuss our results. Finally, we conclude in Sec. IV.

II. COSMOLOGICAL PROBES

A. SNIa

Type Ia supernovae appear to be good “standard candles” and therefore they can serve as useful distance indicators. Constraints arise by comparing the theoretical distance modulus,

$$
\mu_{th} = 5 \log_{10} \left( \frac{d_L(z)}{\text{Mpc}} \right) + 25,
$$

(1)

with the observational distance modulus $\mu_{obs}$ inferred from the data, at the measured supernovae redshifts. In

In this introduction, we emphasize the importance of using median statistics in SNIa data analysis, particularly when compared to χ² analysis. The use of median statistics is argued to be more robust against outliers, providing tighter constraints on dark energy. The paper compares the performances of χ² and median likelihood analyses using type Ia supernovae data from the Union 2 compilation, along with additional cosmological probes such as the Cosmic Microwave Background and Baryonic Acoustic Oscillations. The layout of the paper is outlined, with Sec. II describing the application of statistical analysis to SNIa data and Sec. III presenting the results. The conclusion is drawn in Sec. IV.
Eq. (1), \(d_L(z)\) is the luminosity distance given by,

\[
d_L(z) = \frac{(1 + z)c}{H_0\sqrt{\Omega_{k0}}} S \left( \sqrt{\Omega_{k0}} \int_0^z \frac{dz'}{E(z')} \right),
\]

(2)

where \(z\) is the cosmological redshift, \(H_0\) is the present day value of the Hubble expansion rate and \(E(z) = H(z)/H_0\). The function \(S\) is defined as,

\[
S(x) = \begin{cases} 
\sin x, & \Omega_{k0} < 0 \\
1, & \Omega_{k0} = 0 \\
\sinh x, & \Omega_{k0} > 0
\end{cases},
\]

(3)

with the flat case \((\Omega_{k0} = 0)\) being recovered taking the limit \(\Omega_{k0} \to 0\) in Eq. (2). In this paper we consider,

\[
E(a) = \sqrt{\Omega_{m0}a^{-3} + \Omega_{de0}a^{-3(1+w)} + \Omega_{k0}a^{-2}},
\]

(4)

where \(a = 1/(1 + z)\) is the scale factor, \(\Omega_{m0}\) and \(\Omega_{de0}\) are, respectively, the present day values of the matter and dark energy fractional densities \((\Omega_i = \rho_i/\rho_c\), with \(\rho_c\) being the critical density\) and \(\Omega_{k0} = 1 - \Omega_{m0} - \Omega_{de0}\). The parameter \(w\) is the equation of state parameter of the dark energy (the ratio between the pressure and the energy density) which, for simplicity, we assume to be constant (see however \([11, 12]\)). If the dark energy role is played by a cosmological constant \(\Lambda\) then \(\Omega_{de} = \Omega_\Lambda\) and \(w = -1\).

1. \(\chi^2\) analysis

In the \(\chi^2\) statistical analysis, the likelihood \(P\) of the cosmological parameters \(Q\) is given by \(P \propto \exp(-\chi^2/2)\), with

\[
\chi^2 = \sum_{i=1}^{N} \left( \frac{\mu_{obs}(z_i) - \mu_{th}(z_i, Q)}{\sigma_i} \right)^2.
\]

(5)

In Eq. (4), \(N\) is the number of supernovae in the dataset and \(\sigma_i\) is the observational error associated with \(\mu_{obs}\) at the redshift \(z_i\). The \(\chi^2\) analysis assumes that: (i) the experimental results are statistically independent; (ii) there are no systematic errors; (iii) the statistical errors follow a Gaussian distribution; (iv) the standard deviation of the statistical errors is equal to the observational uncertainty.

Presently, there is no strong evidence supporting that the supernovae magnitude errors are Gaussianly distributed and therefore the hypothesis (iii) and (iv) are quite strong. Moreover, the \(\chi^2\) statistic is highly susceptible to the presence of “outliers” in the datasets. This constitutes an extra concern, in particular due to the uncertainties associated to the calibration of the SNIa light curves.

2. Median analysis

The fewer the assumptions one needs to make about the dataset, the more reliable the results derived from it are. The median statistical analysis discards the assumption of the Gaussianity of the errors, requiring only the use of hypothesis (i) and (ii).

Assuming that the experimental results are statistically independent and that there are no systematic errors present, one expects that after performing a sufficiently large number of measurements, approximately, half of the values obtained will be above (or below) the correct mean value. In particular, if we perform \(N\) measurements, the probability that \(k\) of them will be above (or below) the median is given by the binomial distribution,

\[
P(k, N) = \frac{2^{-N}N!}{k!(N - k)!}.
\]

(6)

This way, given the data set \(\mu_{obs}(z_i) (i = 1..N)\), the likelihood of the cosmological parameters \(Q\) is obtained by counting the number of observational values that fall above (or below) the curve given by Eq. (1).

Despite being associated with less tighter constraints than the \(\chi^2\) statistic, the median statistic is not very sensitive to the presence of a few “outlier” SNIa objects. In \([9]\), it has been shown how the presence of one or very few “ill” data points could severely distort a \(\chi^2\) analysis, while the median results remained approximately the same.

On the other hand, one should notice that the likelihood computed with the median statistic only accounts for the number of experimental points above or below the expected correct curve, not differentiating between the various ways in which these points could be distributed. For instance, a set where the first half of the data points is above the expected curve and the second half is below has the same likelihood (when computed with the median) as a set where the first point is above, the second below, the third above, the fourth below, and so forth. These two cases should not be indistinct since the first could turn out to be a terrible fit to the data. In \([10]\) it has been shown how some modifications of the median statistic could alleviate this problem. In short, the modifications involved taking into account the size of the largest continuous sequence found above (or below) the model’s prediction, or the number of total continuous sequences obtained.

Here, we propose an alternative way to cope with this problem. Instead of counting the number of points that are above (or below) the model’s prediction in the entire redshift range of the data set, one may divide the dataset into redshift intervals with a certain number of SNIa objects and perform the counting in each interval. Suppose we divide the dataset into \(n\) intervals, with \(N_j\) being the number of SNIa objects in the \(j\)-th interval \((j = 1..n)\). This way, the overall likelihood of the parameters \(Q\) is
given by

\[ P = \prod_{j=1}^{n} P(k_j, N_j), \quad (7) \]

where \( k_j \) is the number of points that, in the \( j \)-th interval, are above (or below) the theoretical curve given by the parameters. This way, by properly dividing the dataset into groups of supernovae, we are more likely to avoid pathological situations in which very large sequences of SNIa above (or below) the median value are present.

\[ \Omega_m \approx 0.3, \Omega_k \approx 0.7 \]

\[ B. \text{ The CMB shift parameter} \]

The CMB shift parameter \( R \) is defined by

\[ R_{th} \equiv \frac{\sqrt{\Omega_m} H_0 \, d_L(z_{dec})}{c \, (1 + z_{dec})} = \]

with the redshift of decoupling \( z_{dec} \approx 1090.97 \) [5, 13]. Following the WMAP 7-year results [5] we take \( R_{obs} = 1.725 \) with an error \( \sigma_R = 0.018 \). The likelihood derived from the shift parameter is then \( P_{CMB} \approx \exp \left( -\chi^2_{CMB}/2 \right) \), with

\[ \chi^2_{CMB} = \left( \frac{R_{th} - R_{obs}}{\sigma_R} \right)^2. \quad (9) \]

It turns out that assuming that \( R \) is Gaussianly distributed around \( R_{obs} \) with standard deviation \( \sigma_R \) provides an efficient summary of the information encoded in the full CMB data [14, 15].
C. The BAO scale

The baryonic acoustic oscillations imprinted in the CMB manifest themselves today in the large-scale distribution of galaxies. The BAO signature on large scales was found in [16], when a small “bump” in the two-point correlation function of red-luminous galaxies was measured. Cosmological constraints arise via the position of the “bump”, which is related to the quantity

\[ A_{th}(z) = \sqrt{\Omega_{m0}} E(z)^{-\frac{1}{2}} \times \]
\[ \times \left[ \frac{1}{z\sqrt{|\Omega_{k0}|}} S\left( \frac{|\Omega_{k0}|}{z} \int_0^z \frac{dz'}{E(z')} \right) \right]^{\frac{3}{2}} \]  \( .(10) \)

The likelihood is \( P_A \propto \exp \left( -\chi^2_A / 2 \right) \), with

\[ \chi^2_A = \left( \frac{A_{th}(z_{BAO}) - A_{obs}}{\sigma_A} \right)^2 , \]  \( .(11) \)

where \( z_{BAO} = 0.35 \), \( A_{obs} = 0.469 \) and \( \sigma_A = 0.017 \) [17, 18]. Just like in the case of the shift parameter, it is usually safe to assume that \( A \) is Gaussianly distributed [16, 17].

III. RESULTS

A. SNIa constraints

Following the discussion in Sec. IIA, we divide the SNIa Union 2 dataset [4], which contains a total of \( N = 307 \) SNIa objects, in groups of approximately 75 SNIa each. The supernovae are ordered by increasing redshift with the first three groups having 76 supernovae and the forth group 79. Several other ways to divide the set are possible. However, this choice is not critical for our conclusions.
TABLE I: Constraints on $\Omega_m$, $\Omega_{\Lambda}$ and $w$, obtained from SNIa ($\chi^2$ and median analysis), CMB and BAO data, as well as their combinations. The constraints are at the 1σ level.

| Fit                                          | $\Omega_m$          | $\Omega_{\Lambda}$ | $w$   |
|----------------------------------------------|----------------------|----------------------|-------|
| SNIa ($\chi^2$ + CMB)                       | 0.254; 0.310         | -0.043; 0.056        | -1(fixed) |
| SNIa (median) + CMB                          | 0.260; 0.342         | -0.061; 0.083        | -1(fixed) |
| SNIa ($\chi^2$ + CMB + BAO)                 | 0.269; 0.333         | -0.031; 0.049        | -1(fixed) |
| SNIa (median) + CMB + BAO                    | 0.275; 0.333         | -0.042; 0.064        | -1(fixed) |
| SNIa ($\chi^2$ + CMB + BAO)                 | 0.264; 0.308         | 0(fixed)             | -1.084; -0.955 |
| SNIa (median) + CMB + BAO                    | 0.266; 0.312         | 0(fixed)             | -1.058; -0.902 |

Figures 1 and 2 show the 68.3%, 95.4% and 99.7% confidence level contours on $\Omega_m$ and $\Omega_{\Lambda}$, obtained from the Union 2 SNIa dataset using the $\chi^2$ and median statistical analysis, respectively. Figures 3 and 4 are similar to Figs. 1 and 2 except that now the SNIa constraints are on the $(\Omega_m, w)$ plane and the condition $\Omega_{\Lambda} = 0$ was assumed. In both cases the Hubble parameter today was held fixed at $H_0 = 70.2$ km/s/Mpc. As expected, the constraints using the $\chi^2$ statistic are tighter, consequence of the stronger assumptions it makes about the data. Nevertheless, the constraints obtained using the median statistical analysis are not as bad as one could originally fear. The median contours are more “stretched” than the $\chi^2$ ones, but their “width” is similar. The median statistical constraints are in principle more reliable, since they do not assume Gaussianity of the SNIa magnitude error distribution.

B. SNIa + CMB + BAO constraints

Figures 5 and 6 show the 68.3%, 95.4% and 99.7% confidence level contours on $\Omega_m$ and $\Omega_{\Lambda}$, obtained from the Union 2 SNIa dataset (using the $\chi^2$ and median statistical analysis, respectively), combined with constraints on the CMB shift parameter $R$. The differences between the results obtained with the $\chi^2$ and median statistics are now significantly reduced. This is related to the fact that, as shown in Figs. 1 and 2, the “width” of the SNIa contours, obtained using the $\chi^2$ and median analysis, is very similar. The CMB contours “cross” the SNIa contours, rendering similar combined constraints in both cases. Considering the BAO data (not shown) tightens the constraints slightly (see Table 1).

Figures 7 and 8 show the combined constraints on $\Omega_m$ and $w$ (assuming that $\Omega_{\Lambda} = 0$), obtained from SNIa (using $\chi^2$ and median analysis, respectively), CMB and BAO data. Again, we see that the combined constraints are weakly dependent on whether one derives the constraints from SNIa with the $\chi^2$ or with the median statistical analysis.

Table II presents a summary of the constraints, at the 1σ level, obtained from different data combinations and model assumptions. It shows that the median is almost as constraining as the $\chi^2$ statistical analysis, if combined with additional CMB and BAO constraints. Hence, the strong assumption that the SNIa measurements are Gaussianly distributed is not necessary in order to obtain tight constraints. On the other hand, since the $\chi^2$ and median statistical analysis yield similar cosmological constraints when the SNIa data is combined with other cosmological probes, our results may also be interpreted as providing additional validation of the results obtained using the standard $\chi^2$ statistical analysis.

IV. CONCLUSION

In this paper we compared the cosmological constraints derived from the Union 2 SNIa dataset using $\chi^2$ and median statistics. In the absence of CMB and BAO constraints, we have shown that the $\chi^2$ statistic yields tighter cosmological constraints than the median statistic, as a result of the stronger assumptions it makes about the SNIa magnitude error distribution. On the other hand, when CMB and BAO information is taken into account, the performances of both statistics are very similar. Hence, we conclude that the assumption of the gaussian distribution of the errors in the SNIa analysis does not appear to be critical in the determination of cosmological constraints, provided that complementary information from other cosmological probes is also taken into account.

[1] A. G. Riess et al. (Supernova Search Team), Astron. J. 116, 1009 (1998), astro-ph/9805201.
[2] S. Perlmutter et al. (Supernova Cosmology Project), Astrophys. J. 517, 565 (1999), astro-ph/9812133.
[3] J. Frieman, M. Turner, and D. Huterer, Ann. Rev. Astron. Astrophys. 46, 385 (2008), 0803.0982.
[4] R. Amanullah et al., Astrophys. J. 716, 712 (2010), 1004.1711.
[5] E. Komatsu et al. (WMAP), Astrophys. J. Supp. 192, 18 (2011), 1001.4538.
[6] P. P. Avelino, Phys.Lett. B611, 15 (2005), astro-ph/0411033.
[7] A. Barreira and P. P. Avelino, Phys. Rev. D83, 103001 (2011), 1103.2401.
[8] E. J. Copeland, M. Sami, and S. Tsujikawa, Int. J. Mod. Phys. D15, 1753 (2006), hep-th/0603057.
[9] J. R. Gott, III, M. S. Vogeley, S. Podariu, and B. Ratra, Astrophys. J. 549, 1 (2001), astro-ph/0006103.
[10] P. P. Avelino, C. J. A. P. Martins, and P. Pinto, Astrophys. J. 575, 989 (2002), astro-ph/0111195.
[11] P. P. Avelino, A. M. M. Trindade, and P. T. P. Viana, Phys.Rev. D80, 067302 (2009), 0906.5366.
[12] P. P. Avelino, L. Losano, and J. J. Rodrigues, Phys.Lett. B699, 10 (2011), 1103.1384.
[13] W. Hu and N. Sugiyama, Astrophys. J. 471, 542 (1996), astro-ph/9510117.
[14] Y. Wang and P. Mukherjee, Phys. Rev. D76, 103533 (2007), astro-ph/0703780.
[15] O. Elgaroy and T. Multamaki (2007), astro-ph/0702343.
[16] D. J. Eisenstein et al. (SDSS), Astrophys. J. 633, 560 (2005), astro-ph/0501171.
[17] W. J. Percival et al. (SDSS), Mon. Not. Roy. Astron. Soc. 401, 2148 (2010), 0907.1660.
[18] B. A. Reid et al., Mon. Not. Roy. Astron. Soc. 404, 60 (2010), 0907.1659.