MNRAS **000**, 1–6 (2015) Preprint 11 April 2018 Compiled using MNRAS LATEX style file v3.0

Predicting extragalactic distance errors using Bayesian inference in multi-measurement catalogs

Germán Chaparro-Molano,[1⋆] Juan Carlos Cuervo,[2] Oscar Alberto Restrepo Gaitán[1,3] Sergio Torres Arzayús[4]
[1] *Vicerrectoría de Investigación, Universidad ECCI, 111311 Bogotá, Colombia*
[2] *Department, Institution, Street Address, City Postal Code, Country*
[3] *Radio Astronomy Instrumentation Group, Universidad de Chile, Santiago de Chile, Chile*
[4] *Centro Internacional de Física, Bogotá, Colombia*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
This is a simple template for authors to write new MNRAS papers. The abstract should briefly describe the aims, methods, and main results of the paper. It should be a single paragraph not more than 250 words (200 words for Letters). No references should appear in the abstract.

Key words: Galaxies: distances – keyword2 – keyword3

1 INTRODUCTION

Efforts to reduce the uncertainty in the estimate the Hubble constant are single-method such as SNIa Dhawan et al. (2018). Bayesian analysis of systematic uncertainties for cosmology when using SNIa-derived galactic parameters (heterogeneous errors) Rubin et al. (2015). Hubble constant MCMC estimation based on Cepheids distance determination for NGC 4258 Humphreys et al. (2013). Freedman & Madore (2010) is the important hubble paper, although the original hubble estimation from redshift independent distances is Freedman et al. (2001). Barris & Tonry (2004) estimates distances using snia but no redshifts.

The `emcee` affine invariant MCMC ensemble sampler Foreman-Mackey et al. (2013) has been widely used due to its usability and efficiency. Markov Chain Monte Carlo (MCMC) samplers such as Foreman-Mackey et al. (2013) have been widely used for fitting data to models. Recently, Zhang & Shields (2018) used `emcee` for model assessment using Bayesian and Akaike Information Criteria along with Bayes factors, focusing on small datasets, where it is not relevant to reproduce the original variance of the data. `emcee` has also been proved to be useful in recovering probabilistic models for photometric redshifts Speagle & Eisenstein (2017a,b) . Said et al. (2016) has used emcee for Tully-Fisher in the southern Zone of Avoidance.

Tully Fisher relation Obreschkow & Meyer (2013)

GW searches White et al. (2011), we should improve redshift independent distance determination .

Springob et al. (2014) attempt to predict redshift-derived distance errors within a Bayesian framework yields 26%.

Exploration of prior discrepancy modeling in model estimation Ling et al. (2014)

Importance of distance and catalogs Dabringhausen & Fellhauer (2016); Hernan-Caballero et al. (2016) in:

Other discrepancy measures for model selection de la Horra & Teresa Rodriguez-Bernal (2012). Chi2 model selection De la Horra (2008)

? is the 2.4 percent determination of hubble constant

The NASA/IPAC Extragalactic Distance (NED) catalog of Redshift-Independent Distances Mazzarella & Team (2007); Steer et al. (2017) is a catalog with the following properties. HyperLEDA Makarov et al. (2014) is a catalogue for nearby extragalactic distances, which also includes redshift-independent distance measurements, but it is much smaller than NED, and does not give a prescription for treating errors. Changes in Hubble constant estimation using TF relation without Cepheids Mould & Sakai (2008)

Determining whether a galaxy belongs to a group by analyzing their common properties Kourkchi & Tully (2017)

Determining the spatial distribution of galaxies in order to study large-scale structure Roman & Trujillo (2017) or local universe peculiar velocities Sorce et al. (2014). Kinematics of nearby galaxies in void Nasonova & Karachentsev (2011)

Studies of anisotropy be it of morphological types Javanmardi & Kroupa (2017) or density-velocity Ma et al. (2013) which also need to take into account instrument detection limits Fan et al. (2013) and source identification Budavari & Szalay (2008). Anisotropy hubble from HST data McClure & Dyer (2007)

Tully et al. (2016) distance errors are weighted stan-

⋆ E-mail: gchaparrom@ecci.edu.co

dard deviations across different methods. Individual errors are not available for all methods. Courtois et al. (2012) local universe structure reconstruction Cosmicflows-1

Kelly (2007) is a widely used Bayesian linear regressor which uses a Gaussian Mixture Model to approximate the distibution of unobserved "true" data values and from this information estimate the regression coefficients.

10% estimated uncertainty for photometrically derived distance scale ladder Tully & Pierce (2000)

Are Jarrett et al. (2000) sources important in our analysis?

Get better feeling of linmix fit. Do stuff to hyperleda catalog? Arm-wave my way out of that?

Remember that Cepheids and SNIa are primary distance indicators. TF FP are secondary

Springob et al. (2007)

Chaparro Molano et al. (2018), Torres & Cuervo (2018), Gelman et al. (1996) Brooks et al. (2000) Tully & Fisher (1977)

This is a simple template for authors to write new MNRAS papers. See `mnras_sample.tex` for a more complex example, and `mnras_guide.tex` for a full user guide.

All papers should start with an Introduction section, which sets the work in context, cites relevant earlier studies in the field by Speagle & Eisenstein (2017b), and describes the problem the authors aim to solve (e.g. Speagle & Eisenstein 2017a).

2 METHODS, OBSERVATIONS, SIMULATIONS ETC.

From here on, when we mention distance measurements in the NED-D catalog, we will be excluding from our analysis measurements that require the target redshift to calculate the distance, as indicated in the `redshift (z)` column.

In the NED-D database, $\sim$ 16000 galaxies ($\sim$ 9%) have more than one distance measurement, $\sim$ 1800 galaxies ($\sim$ 1%) have more than 12 distance measurements, and 180 galaxies ($\sim$ 0.1%) have distances measurements using more than 6 different methods. Even though our analysis for error estimation can be used across different distance determination methods for single galaxies, we think that it is more informational to separate the analysis by method.

In this section we will focus on galaxies whose distances have been measured using the Tully-Fisher method. The reason is that out of all distance determination methods listed in the NED-D database, it has the largest number of galaxies with non-reported distance modulus errors (818).

For many galaxies, the random error for each distance modulus measurement ϵ_i (for $i = 1, ..., N$, where N is the number of distance measurements per galaxy) is not representative of the scatter across measurements, even when considering the same method for determining distances. In addition, distance modulus distributions for each measurement (which are assumed to be Gaussian) are transformed to log-normal distributions in distance space. For this reason, we consider that the best approach to consider the effects of random and scattering errors in catalog-wide, multi-method distance analyses is to take bootstrap samples of the posterior distribution for each extragalactic distance $P(D_G)$. The posterior distribution can be obtained by drawing distance modulus samples from $P(\mu)$, which is the unweighted mixture of normal distributions corresponding to each distance modulus measurement μ_i,

$$\mu \sim \sum_i^N \mathcal{N}(\mu_i, \epsilon_i^2) \ ,$$

and then converting to metric distance,

$$D_G = 10^{\frac{\mu}{5}+1} \ .$$

Therefore,

$$D_G \sim \sum_i^N \text{lognormal}(M_i, \sigma_{M_i}^2) \ .$$

Here $M_i = \ln D_i$ and $\sigma_{M_i} = \epsilon_i \cdot \ln 10$.

However, this method is not very efficient for a standardized treatment of errors. It is more convenient to treat each extragalactic metric distance D_G as a normal random variable with a single-valued σ_D as a measure of the uncertainty in the estimation of an extragalactic distance,

$$D_G \sim \mathcal{N}(D, \sigma_D^2)$$

For this reason we compare four methods for estimating the D, σ_D pair. Two of these methods use robust measures of the posterior distribution of each extragalactic distance (H, M), and the other two use measures based on propagation of errors (P, Q).

Method H takes D as the median of the posterior and σ_D as the half-distance (H) between the 84th and 16th percentiles of the posterior. Method M takes D as the median of the posterior and σ_D as the median absolute deviation (MAD) of the posterior. Method P consists on calculating D from the weighted mean distance modulus $\bar{\mu}^*$ with weights $w_i = \epsilon_i^{-2}$. σ_D is calculated by propagation (P) of measurement errors i.e. from the uncertainty of the weighted mean (Tully et al. 2016),

$$\sigma_D^P = 0.461 \, \bar{D}^* \left(\sum_i^N w_i \right)^{-1/2} \ , \tag{1}$$

Method P does not take into account the scatter in distance measurements for single galaxies, which is why method Q calculates D same as method P, but σ_D is calculated as the sum in quadrature (Q) of the propagated uncertainty of the weighted mean and the propagated unbiased weighted sample variance σ_D^*:

$$\sigma_D^Q = \left[\left(\sigma_D^P\right)^2 + \left(\sigma_D^*\right)^2 \right]^{1/2} \ . \tag{2}$$

Here σ_D^* is calculated as (Brugger 1969),

$$\sigma_D^* = 0.461 \, \bar{D}^* \sqrt{\frac{N}{N-1.5} \frac{\sum_i^N w_i (\mu_i - \bar{\mu}^*)^2}{\sum_i^N w_i}} \ . \tag{3}$$

Fig. 1 shows that the width in the posterior distribution of each extragalactic distance is best explained using the H method, whereas the less robust P and Q methods

Table 1. This is an example table. Captions appear above each table. Remember to define the quantities, symbols and units used.

A	B	C	D
1	2	3	4
2	4	6	8
3	5	7	9

under-predict the variance for galaxies in the whole distance range. The M method also under-predicts the variance, but being a robust method, it is not as sensitive to outliers as the mehods P and Q, as seen in the case of NGC 1558 in Fig. 1. It should be noted that the shape of $P(D)$ for the galaxies shown in Fig. 1 is not representative of the distance range.

Figure 2 shows that distance errors grow linearly with distance. Furthermore, the quadrature (Q) and propagation (P) methods underpredict distance errors for most galaxies in the sample. Figure 4 shows that method Q underpredicts distance errors with respect to the median absolute deviation method (M), which also yields a tighter linear correlation to extragalactic distance due to its robustness.

Given that σ_D calculated using the H method is obtained from many realizations from the posterior distribution of extragalactic distances, it is also possible to calculate its variance as the half-distance between the 84th and 16th percentile of σ_D realizations. Figure **??** shows that the variance of the estimated error is proportional to the error for the H and M methods. This will be relevant in Section XXX when we construct a predictive model for unknown errors.

talk about this using only TF

Normally the next section describes the techniques the authors used. It is frequently split into subsections, such as Section 2.1 below.

2.1 Maths

Simple mathematics can be inserted into the flow of the text e.g. $2 \times 3 = 6$ or $v = 220\,\mathrm{km\,s^{-1}}$, but more complicated expressions should be entered as a numbered equation:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{4}$$

Refer back to them as e.g. equation (4).

2.2 Figures and tables

Figures and tables should be placed at logical positions in the text. Don't worry about the exact layout, which will be handled by the publishers.

Figures are referred to as e.g. Fig. **??**, and tables as e.g. Table 1.

3 CONCLUSIONS

The last numbered section should briefly summarise what has been done, and describe the final conclusions which the authors draw from their work.

ACKNOWLEDGEMENTS

This research has made use of the NASA/IPAC Extragalactic Database (NED), which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

REFERENCES

Barris B., Tonry J., 2004, ASTROPHYSICAL JOURNAL, 613, L21
Brooks S. P., Catchpole E. A., Morgan B. J. T., 2000, Statistical Science, 15, 357
Brugger R. M., 1969, The American Statistician, 23, 32
Budavari T., Szalay A. S., 2008, ASTROPHYSICAL JOURNAL, 679, 301
Chaparro Molano G., Restrepo Gaitán O. A., Cuervo Marulanda J. C., Torres Arzayus S. A., 2018, in Revista Mexicana de Astronomia y Astrofisica Conference Series. pp 63–63
Courtois H. M., Hoffman Y., Tully R. B., Gottloeber S., 2012, ASTROPHYSICAL JOURNAL, 744
Dabringhausen J., Fellhauer M., 2016, MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY, 460, 4492
De la Horra J., 2008, COMMUNICATIONS IN STATISTICS-THEORY AND METHODS, 37, 1412
Dhawan S., Jha S. W., Leibundgut B., 2018, ASTRONOMY & ASTROPHYSICS, 609
Fan D., Budavari T., Szalay A. S., Cui C., Zhao Y., 2013, PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC, 125, 218
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC, 125, 306
Freedman W. L., Madore B. F., 2010, in Blandford R., Faber S., van Dishoeck E., Kormendy J., eds, Annual Review of Astronomy and Astrophysics, Vol. 48, ANNUAL REVIEW OF ASTRONOMY AND ASTROPHYSICS, VOL 48. pp 673–710, doi:10.1146/annurev-astro-082708-101829
Freedman W. L., et al., 2001, The Astrophysical Journal, 553, 47
Gelman A., li Meng X., Stern H., 1996, Statistica Sinica, 6, 733
Hernan-Caballero A., Spoon H. W. W., Lebouteiller V., Rupke D. S. N., Barry D. P., 2016, MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY, 455, 1796
Humphreys E. M. L., Reid M. J., Moran J. M., Greenhill L. J., Argon A. L., 2013, ASTROPHYSICAL JOURNAL, 775
Jarrett T. H., Chester T., Cutri R., Schneider S., Skrutskie M., Huchra J. P., 2000, The Astronomical Journal, 119, 2498
Javanmardi B., Kroupa P., 2017, ASTRONOMY & ASTROPHYSICS, 597
Kelly B. C., 2007, ASTROPHYSICAL JOURNAL, 665, 1489
Kourkchi E., Tully R. B., 2017, ASTROPHYSICAL JOURNAL, 843
Ling Y., Mullins J., Mahadevan S., 2014, JOURNAL OF COMPUTATIONAL PHYSICS, 276, 665
Ma Y.-Z., Taylor J. E., Scott D., 2013, MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY, 436, 2029
Makarov D., Prugniel P., Terekhova N., Courtois H., Vauglin I., 2014, ASTRONOMY & ASTROPHYSICS, 570

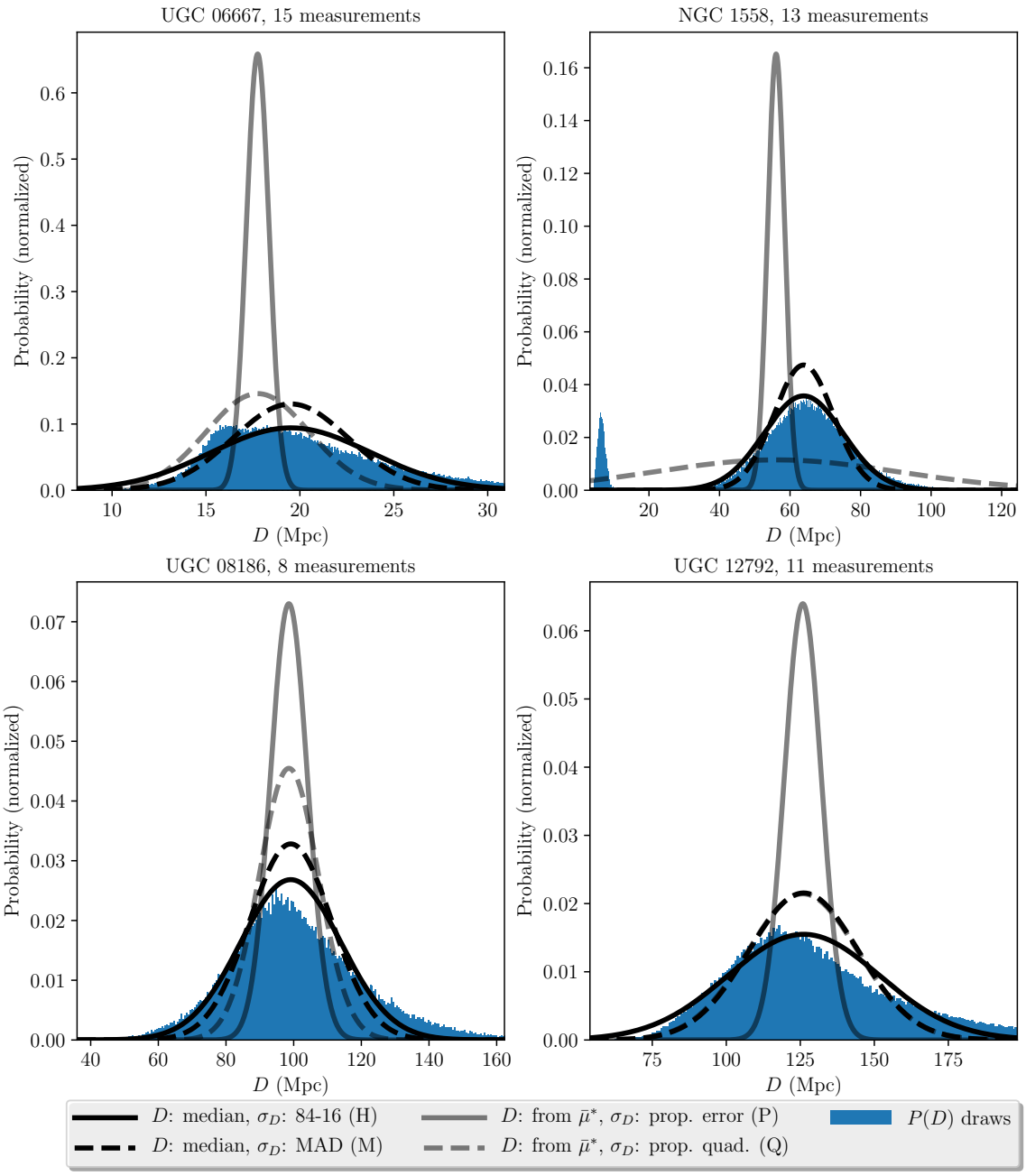

Figure 1. Comparison of extragalactic distance posterior distribution draws $P(D_G)$ and modeled distributions $\mathcal{N}(D, \sigma_D^2)$ for UGC 06667, NGC 1558, UGC 08186, and UGC 12792 using the Tully-Fisher Method for distance determination in NED-D. The galaxies shown here were chosen from the most populated distance range in the NED-D TF sample ($0 - 140$ Mpc) due to them being near four equidistant distance points: 20, 60, 100, and 140 Mpc. The four methods used for approximating the posterior distribution (H, M, P, and Q) are described in the text.

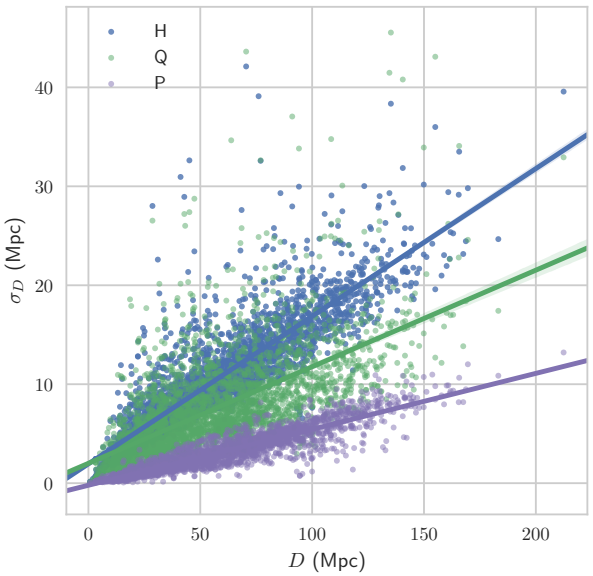

Figure 2. Median extragalactic distance vs. predicted extragalactic distance errors for galaxies with more than 5 TF distance measurements in NED-D according to the H, Q, P error models.

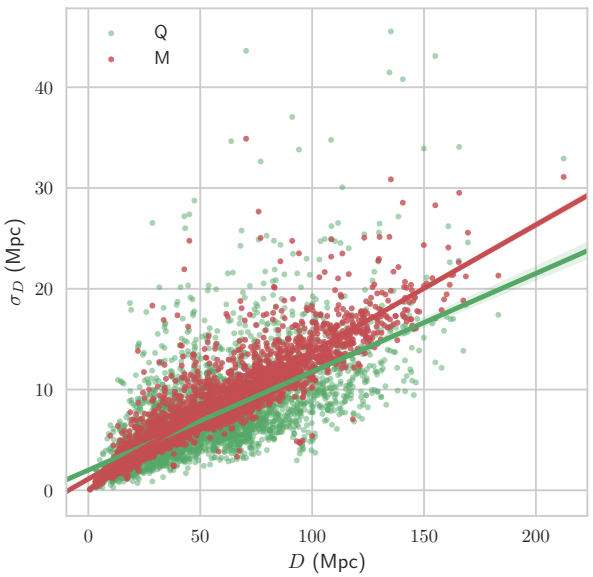

Figure 3. Median extragalactic distance vs. predicted extragalactic distance errors for galaxies with more than 5 TF distance measurements in NED-D according to the Q, M error models.

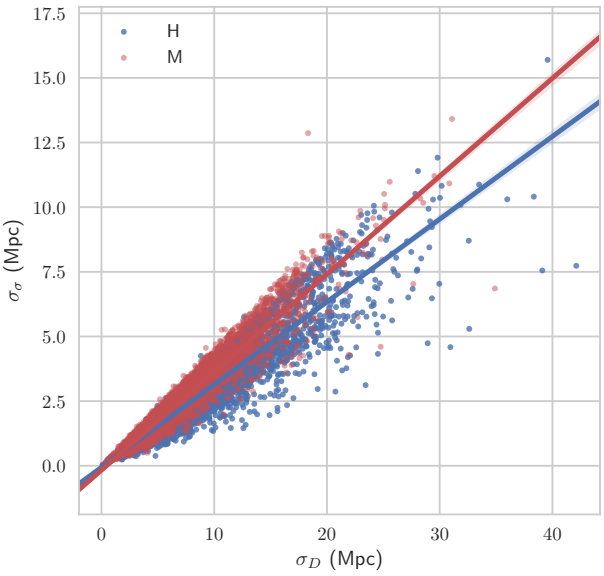

Figure 4. Predicted extragalactic distance errors vs. variance of the error as determined by the H and M methods.

Mazzarella J. M., Team N., 2007, in Shaw R., Hill F., Bell D., eds, ASTRONOMICAL SOCIETY OF THE PACIFIC CONFERENCE SERIES Vol. 376, ASTRONOMICAL DATA ANALYSIS SOFTWARE AND SYSTEMS XVI. pp 153–162
McClure M. L., Dyer C. C., 2007, NEW ASTRONOMY, 12, 533
Mould J., Sakai S., 2008, ASTROPHYSICAL JOURNAL LETTERS, 686, L75
Nasonova O. G., Karachentsev I. D., 2011, ASTROPHYSICS, 54, 1
Obreschkow D., Meyer M., 2013, ASTROPHYSICAL JOURNAL, 777
Roman J., Trujillo I., 2017, MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY, 468, 703
Rubin D., et al., 2015, ASTROPHYSICAL JOURNAL, 813
Said K., Kraan-Korteweg R. C., Staveley-Smith L., Williams W. L., Jarrett T. H., Springob C. M., 2016, MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY, 457, 2366
Sorce J. G., Courtois H. M., Gottlober S., Hoffman Y., Tully R. B., 2014, MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY, 437, 3586
Speagle J. S., Eisenstein D. J., 2017a, MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY, 469, 1186
Speagle J. S., Eisenstein D. J., 2017b, MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY, 469, 1205
Springob C. M., Masters K. L., Haynes M. P., Giovanelli R., Marinoni C., 2007, The Astrophysical Journal Supplement Series, 172, 599
Springob C. M., et al., 2014, MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY, 445, 2677
Steer I., et al., 2017, ASTRONOMICAL JOURNAL, 153
Torres S., Cuervo J. C., 2018, Tecciencia, 24, 53
Tully R. B., Fisher J. R., 1977, A&A, 54, 661
Tully R. B., Pierce M. J., 2000, The Astrophysical Journal, 533, 744
Tully R. B., Courtois H. M., Sorce J. G., 2016, ASTRONOMICAL JOURNAL, 152

White D. J., Daw E. J., Dhillon V. S., 2011, CLASSICAL AND QUANTUM GRAVITY, 28
Zhang J., Shields M. D., 2018, MECHANICAL SYSTEMS AND SIGNAL PROCESSING, 98, 465
de la Horra J., Teresa Rodriguez-Bernal M., 2012, SORT-STATISTICS AND OPERATIONS RESEARCH TRANSACTIONS, 36, 69

APPENDIX A: SOME EXTRA MATERIAL

If you want to present additional material which would interrupt the flow of the main paper, it can be placed in an Appendix which appears after the list of references.

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