MNRAS 000, 1–6 (2015) Preprint 12 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

Understanding the uncertainties in redshift-independent extragalactic distance measurements is absolutely necessary before reporting statistically sound conclusions regarding the structure of the local universe (Nasonova & Karachentsev 2011; Courtois et al. 2012; Ma et al. 2013; Springob et al. 2014; Sorce et al. 2014; Said et al. 2016; Kourkchi & Tully 2017), large scale structure (McClure & Dyer 2007; Roman & Trujillo 2017; Javanmardi & Kroupa 2017; Torres & Cuervo 2018; Jesus et al. 2018), and events like transient gravitational wave detections (White et al. 2011). Hubble constant estimations have been using increasingly sophisticated statistical tools for primary distance determination methods, such as SNIa (Barris & Tonry 2004; Rubin et al. 2015; Dhawan et al. 2018), Cepheids Humphreys et al. (2013) or both (Riess et al. 2016). Although most estimates of the Hubble constant use Cepheid calibration for calibrating secondary methods (Tully & Pierce 2000; Freedman et al. 2001; Freedman & Madore 2010), Mould & Sakai (2008) have explored changes in Hubble constant estimation using the Tully-Fisher relation (TF) relation without Cepheid calibration. Secondary methods for extragalactic distance determination like the TF relation, or the Fundamental Plane (FP) have recently become more precise thanks to increasing volumes of data from surveys like 6dF (Springob et al. 2014) and 2MASS (Jarrett et al. 2000; Springob et al. 2007) together with Spitzer data (Sorce et al. 2013), along with improved statistical methods (Obreschkow & Meyer 2013).

As of 2018, three multi-measurement catalogs including a substantial amount of redshift-independent extragalactic distance measurements have been released: HyperLEDA (Makarov et al. 2014), NED-D (Mazzarella & Team 2007; Steer et al. 2017), and Cosmicflows-3 (Tully et al. 2016). HyperLEDA includes a homogenized catalog for extragalactic distances in the nearby universe, with 12866 distance measurements for 518 galaxies to date. NED-D is the NASA/IPAC Extragalactic Distance catalog of Redshift-Independent Distances, which compiles XXX distance measurements for XXX galaxies, for which ~ 1800 galaxies ($\sim 1\%$) have more than 12 distance measurements, and 180 galaxies ($\sim 0.1\%$) have distance measurements using more than 6 different methods. Cosmicflows-3 is the most up-to-date catalog, which reports distance measurements for 10616 galaxies for up to four distance determination methods, and calibrated with supernova luminosities. However, unlike HyperLEDA or NED-D, Cosmicflows-3 only reports the latest distance measurement for each method. In HyperLEDA, NED-D and Cosmicflows-3 errors are reported as one standard deviation from the reported distance modulus. Treatment of errors for combining distance moduli across methods or across measurements is suggested by Mazzarella & Team (2007) and Tully et al. (2016) to be based on weighted estimates such as the uncertainty of the weighted mean, albeit with caution due to the hererogeneous origin of the compiled data. In the case of NED-D, this is complicated by the fact that many errors are not reported or are reported as zero. In fact, the TF relation method has the largest number of galaxies with non-reported distance modulus errors (818 to date). Even though extragalactic distances measured using the TF relation were originally reported to have a relative error

⋆ E-mail: gchaparrom@ecci.edu.co

in distance modulus of $10-20\%$ (Tully & Fisher 1977), we consider that this conservative estimate can be improved upon by using a predictive model based on the distance error of galaxies that use the same distance determination method. This requires a robust estimation of the variance of extragalactic distances based on the available data.

For many galaxies in all three catalogs, the random error for each distance modulus measurement ϵ_i (for $i=1,...,N$, for N 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 metric distance space. We improve upon previous methods by robustly estimating the underlying variance across measurements and distance determination methods for the three catalogs by bootstrap sampling the posterior distribution of each extragalactic distance (Chaparro Molano et al. 2018), and comparing our results to more commonly used frequentist methods, such as the weighted estimates mentioned above. Furthermore, we build a predictive Bayesian model for the 818 galaxies in the NED-D catalog whose distances were measured using the TF relation but have non-reported errors.

Here go the sections.

2 POSTERIOR DISTRIBUTION FOR EXTRAGALACTIC DISTANCES

As mentioned in the Introduction, the best approach to consider the effects of random and scattering errors in catalog-wide, multi-method distance analyses is to perform a robust estimation of the variance of the posterior distribution of each extragalactic distance. The posterior distribution of the distance to a given galaxy 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,

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

Therefore,

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

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 (H, M) use robust measures of the posterior distribution of each extragalactic distance, and the other two (P, Q) use measures based on propagation of errors.

2.1 Estimating the variance of $P(D_G)$

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),

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

If the P and Q methods, which are not robust, are representative of the variance of the posterior, they should yield similar results as the H method. Next section shows that this is not the case.

2.2 Comparison of variance estimation methods

Without loss of generality, we will focus on galaxies whose distances have been measured using the Tully-Fisher method in the NED-D catalog because it is the method with the most non-reported errors in the database. 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)` field.

Even though our analysis for error estimation can be used to combine distance measurements using different methods for single galaxies, we think that it is more meaningful to separate the analysis by method. A full discussion of our error estimation method applied to multi-method measurements in the HyperLEDA, NED-D and Cosmicflows-3 is given in the appendix.

Fig. 1 shows that the center and variance of the posterior distribution of each extragalactic distance is best explained using the H method, whereas the less robust

P and Q methods 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. For the more symmetrical posterior distribution of UGC 12792, the M and Q methods predict the same center and variance.

Distance errors grow linearly with distance, as seen in Fig. 2. This means that there is a strong systematic component in the variance of $P(D_G)$. Furthermore, the quadrature (Q) and propagation (P) methods underpredict distance errors for most galaxies in the sample. Fig. 3 shows that method Q underpredicts distance errors with respect to the median absolute deviation method (M), which also shows a tighter linear correlation 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. Fig. 4 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 3 when we construct a predictive model for non-reported errors.

3 PREDICTIVE BAYESIAN MODEL FOR MISSING ERRORS

As seen in Figs. 2 and 3, TF distance errors estimated using the robust methods H and M grow linearly with distance. This means that we should be able to predict missing distance errors for galaxies in NED-D. For this reason we try out several Bayesian models using the `emcee` affine invariant Markov Chain Monte Carlo (MCMC) ensemble sampler (Foreman-Mackey et al. 2013), which has been widely used due to its usability and efficiency. Recently,`emcee` has also been proved to be useful in recovering probabilistic models for photometric redshifts Speagle & Eisenstein (2017a,b). Since we want to be able to predict non-reported errors, our model selection is based on posterior predictive checks, i.e. we rely on models that can create synthetic datasets similar to the original dataset (Gelman et al. 1996). This allows us to reproduce the original variance of the error (Fig. 4. Other Bayesian analysis tools like LINMIX (Kelly 2007), which is widely used in astronomy for approximating unobserved data does not use posterior predictive checks. This is also the case in the work of Zhang & Shields (2018) and Jesus et al. (2018), where they used `emcee` for model assessment using Bayesian and Akaike Information Criteria along with Bayes factors while focusing on small datasets, wthout attempting to reproduce the original variance of the data.

We assume that the probability distribution of σ_D for each galaxy is normal, with variance σ_σ and mean $\hat{\sigma}_D$,

$$P(\sigma_D) = \mathcal{N}(\hat{\sigma}_D, \sigma_\sigma^2) \ .$$

Our likelihood function is the joint probability that the full distance-error dataset with m galaxies is generated by the above probability, for $\hat{\sigma}_D, \sigma_\sigma$ calculated from a single model

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

depending on the distance D_G and a set of parameters $\boldsymbol{\theta}$,

$$P(\sigma_D|D_G, \boldsymbol{\theta}) = \prod_i^m P(\sigma_D|D_{Gi}, \boldsymbol{\theta}) \ .$$

Following Bayes' theorem we can compute the posterior probability up to a constant,

$$P(\boldsymbol{\theta}|D_G, \sigma_D) \propto P(\boldsymbol{\theta})P(\sigma_D|D_G, \boldsymbol{\theta}) \ .$$

Due to the simplicity of the models used here, we will only use conservative (flat) priors.

Our first model is based on the (somewhat naive) hypothesis that there are are distinct systematic and random contributions to the distance measurement error, both of which are normally distributed. For this reason, they are added in quadrature,

$$\sigma_D^2 = \sigma_s^2 + \sigma_r^2 \ .$$

If the systematic error is a scale factor error, as Fig. 2 suggests, $\sigma_r = fD$, where the scale factor f and the random error σ_r are constant. We then use `emcee` to sample the posterior over the parameter set $\boldsymbol{\theta} = (f, \sigma_r)$.

Several models in order to reach a predictive model. Central limit theorem?

Brooks et al. (2000). Exploration of prior discrepancy modeling in model estimation Ling et al. (2014), maybe not needed? Other discrepancy measures for model selection de la Horra & Teresa Rodriguez-Bernal (2012). Chi2 model selection De la Horra (2008). Gelman (2003) and Chambert, Rotella & Higgs (Chambert et al.) and for using posterior predictive checks for inference and prediction

4 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

The authors would like to thank O. L. Ramírez-Suárez and J. E. Forero-Romero for their valuable input during the early stages of this work. 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.

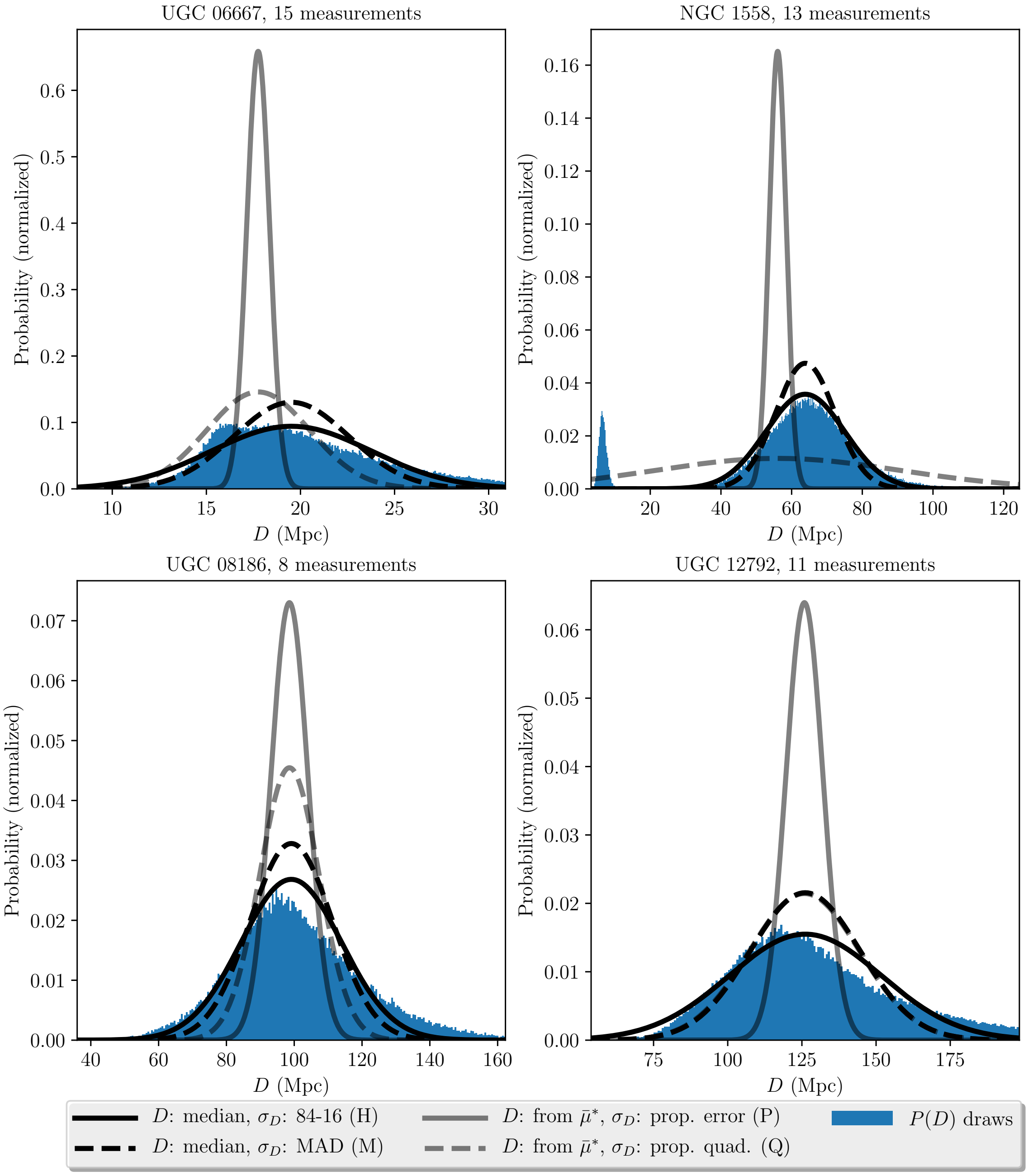

Figure 1. Comparison of extragalactic distance posterior distribution draws and modeled distributions for UGC 06667, NGC 1558, UGC 08186, and UGC 12792 using the Tully-Fisher Method for distance determination in NED-D. The four methods used for approximating the posterior distribution (H, M, P, and Q) are described in the text.

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

Chambert T., Rotella J. J., Higgs M. D., , Ecology and Evolution, 4, 1389

Chaparro Molano G., Restrepo Gaitán O. A., Cuervo Marulanda J. C., Torres Arzayus S. A., 2018, in Revista Mexicana de

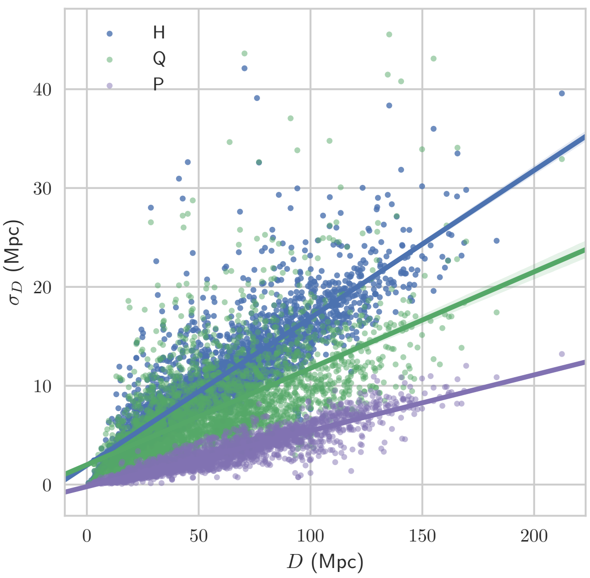

Figure 2. Median extragalactic distance vs. predicted extragalactic distance errors for galaxies with more than 5 TF distance measurements in NED-D according to the H, Q, P error models, showing a linear regression and confidence intervals computed using the `seaborn.regplot` Python function.

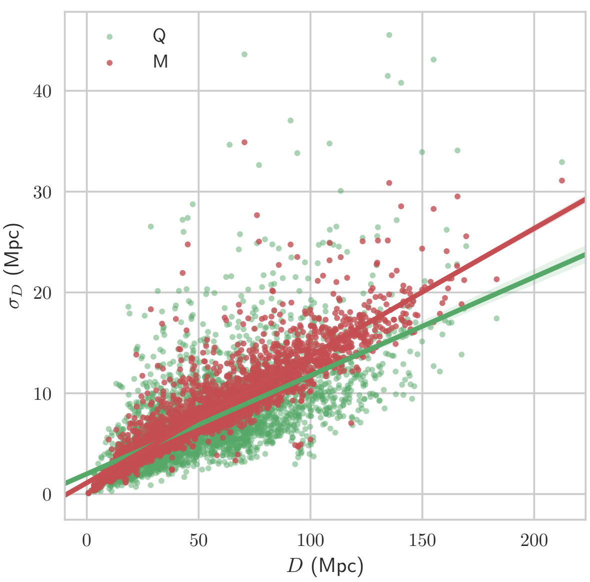

Figure 3. Median extragalactic distance vs. predicted extragalactic distance errors for galaxies with more than 5 TF distance measurements in NED-D according to the Q, M error models, showing a linear regression and confidence intervals computed using the `seaborn.regplot` Python function.

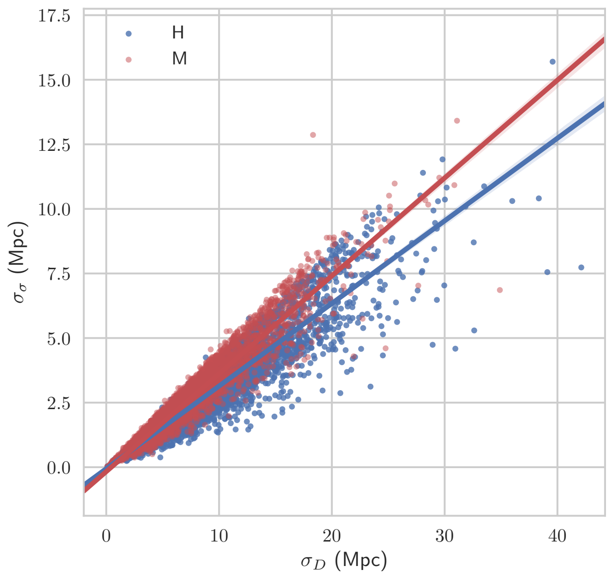

Figure 4. Predicted extragalactic distance errors vs. variance of the error as determined by the H and M methods, showing a linear regression and confidence intervals computed using the `seaborn.regplot` Python function.

Astronomia y Astrofisica Conference Series. pp 63–63
Courtois H. M., Hoffman Y., Tully R. B., Gottloeber S., 2012, ASTROPHYSICAL JOURNAL, 744
De la Horra J., 2008, COMMUNICATIONS IN STATISTICS-THEORY AND METHODS, 37, 1412
Dhawan S., Jha S. W., Leibundgut B., 2018, ASTRONOMY & ASTROPHYSICS, 609
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., 2003, Internat. Statist. Rev., 71, 369
Gelman A., li Meng X., Stern H., 1996, Statistica Sinica, 6, 733
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
Jesus J. F., Gregório T. M., Andrade-Oliveira F., Valentim R., Matos C. A. O., 2018, MNRAS,
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

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
Riess A. G., et al., 2016, The Astrophysical Journal, 826, 56
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., et al., 2013, The Astrophysical Journal, 765, 94
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.