Using gamma regression for photometric redshifts of survey galaxies

J. Elliott, R. S. de Souza, A. Krone-Martins, E. Cameron, E. E. O. Ishida, J. Hilbe

Abstract Machine learning techniques offer a plethora of opportunities in tackling *big data* within the astronomical community. We present the set of Generalized Linear Models as a fast alternative for determining photometric redshifts of galaxies, a set of tools not commonly applied within astronomy, despite being widely used in other professions. With this technique, we achieve catastrophic outlier rates of the order of $\sim 1\%$, that can be achieved in a matter of seconds on large datasets of size $\sim 1,000,000$. To make these techniques easily accessible to the astronomical community, we developed a set of libraries and tools that are publicly available.

1 Introduction

Generalized Linear Models [GLMs; 10] are widely used throughout other scientific disciplines such as: biology [1], medicine [8], and economics [12], and is available within the overwhelming majority of contemporary statistical software packages. However, they have been very little used within the astronomical community

There are plenty of opportunities to apply GLMs within astronomy, and one particularly important problem is the estimation of photometric redshifts (photo-z).

Galaxy spectra are made up of many of its physical properties, including morphology, age, metallicity, star formation history, merging history, and a host of other

J. Elliott
Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA, Max-Planck-Institut für extraterrestrische Physik, Giessenbachstraße 1, 85748, Garching, Germany, e-mail: jonathan.elliott@cfa.harvard.edu

R. S. de Souza
MTA Eötvös University, EIRSA "Lendulet" Astrophysics Research Group, Budapest 1117, Hungary e-mail: rafael.2706@gmail.com

A. Krone-Martins
SIM, Faculdade de Ciências, Universidade de Lisboa, Ed. C8, Campo Grande, 1749-016, Lisboa, Portugal e-mail: algol@sim.ul.pt

E. Cameron
Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford, OX1 3PS, United Kingdom e-mail: dr.ewan.cameron@gmail.com

E. E. O. Ishida
Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85748 Garching, Germany e-mail: emille@mpa-garching.mpg.de

J. Hilbe
Arizona State University, 873701,Tempe, AZ 85287-3701, USA, Jet Propulsion Laboratory, 4800 Oak Grove Dr., Pasadena, CA 91109, USA e-mail: j.m.hilbe@gmail.com

for the COIN collaboration

confounding factors in addition to its redshift. This makes robust estimation of photo-zs a difficult task. Estimation is usually done in two ways, by template fitting, or by using machine learning techniques.

There exist several studies that have investigated the advantages of the publicly available codes that estimate the photo-z of galaxies [for a glimpse on the diversity of existent methods, see 7, and references therein]. The overall performance of most codes is good, demonstrating catastrophic errors from 5 to 9%, which is considered reliable within the field. There are also a number of growing techniques that implement a hybrid approach of template and machine learning techniques [3].

Despite the current advancement within this field, there still exist large practical difficulties. In the next years there are a large number of surveys that will start having *big data* catalogues, e.g., the *Large Synoptic Survey Telescope*[1] [9], *EUCLID*[2] [13] or the *Wide-Field Survey Infrared Telescope*[3] [6]. Current techniques will become difficult to employ if they require large training sets, and as such, this warrants the need for fast and reliable photo-z methods that are capable of robustly estimating redshifts quickly, and on large training datasets.

We introduce GLMs as a new technique to quickly and robustly estimate galaxy photo-zs. We show that it can run in a matter of seconds on a single core computer, even for millions of objects. As part of the COsmostatistics INitiative (COIN[4]) collaboration, we created and distributed easy to use software, and web-applications for use of the wider community in estimating photo-zs[5].

2 Methodology

We will not go into the details of deriving GLMs or the formula that is used in our technique, we instead encourage the reader to see the details in Elliott et al. [5] and references therein. However, we note the importance of GLMs, such that they allow you to choose the type of distribution you want to model. GLMs are applicable to the entire set of *exponential families* of distributions: Gaussian/normal, gamma, inverse Gaussian, Bernoulli, binomial, Poisson, and negative binomial. For example, in this study, we want to predict the photo-zs of galaxies from multi-wavelength photometry. For such a study, the gamma distribution is favourable, as a redshift is positive and continuous, as is the gamma distribution. To then use the gamma distribution to predict redshifts, we utilised the following machine learning methodology:

1. The data was randomly split into training and test sets.
2. Robust principal component analysis (e.g., Candès et al. 2, de Souza et al. 4) was carried out on the complete data set.
3. We utilised a gamma family distribution to reflect the fact that measured redshifts are positive and continuous.

[1] `http://www.lsst.org/lsst`

[2] `http://sci.esa.int/euclid`

[3] `http://wfirst.gsfc.nasa.gov`

[4] `https://asaip.psu.edu/organizations/iaa/iaa-working-group-of-cosmostatistics`

[5] `https://github.com/COINtoolbox`

4. The predicted photo-z for the test data was calculated using the principal component projections of the test data set and the best-fit GLM using the training sample.
5. To measure how well the photo-zs were estimated, we employed a metric commonly used in the literature, specifically, the catastrophic error.

3 Data Samples

We used two publicly available galaxy datasets to test the technique outlined in the previous section. The first was the *PHoto-z Accuracy Testing* (PHAT), an international intitiative to identify the most promising photo-z methods. We used their publicly available simulated datasets that contains $169,520$ simulated galaxies with redshifts ranging from $z = 0.02 - 2.24$, and magnitudes in 11 filters (u, g, r, i, z, Y, J, H, K, $IRAC1$, and $IRAC2$).

Given that this dataset was purely synthetic, we also used a real dataset acquired from the *Sloan Digital Sky Survey* [SDSS; 14]. For details on the query see Elliott et al. [5]. The sample used contained $1,347,640$ galaxies with a redshift range of $z = 0 - 1.0$, with magnitudes in 5 filters (u', g', r', i', and z'). Comparisons with dereddened values showed no inconsistencies

4 Results

Both data sets were fit using an AMD Athlon X2 Dual-Core QL-64 processor with 1.7 GB RAM on the Ubuntu 10.04 operating system, which represents an old laptop at today's standards. We achieved catastrophic errors of 1.4% for the PHAT0 data set and 8% for the SDSS data set, within ~ 1200, and 10 seconds. Lower catastrophic errors of $\sim 1\%$ could be achieved when using more principal components, but would take longer computational time, ~ 5000 s. We plot the best-fit GLM models for SDSS datasets in Fig. 1.

5 Conclusions

The astronomical community has left Generalized Linear Models relatively untouched, despite its use throughout the academic world. We have demonstrated their ease of use and quick applicability to estimate the photo-zs of galaxies. This technique has been shown to be competitive with current techniques implemented, that can require larger training sets and longer time for their algorithms to learn. Such properties of this technique will become important in the close future when upcoming wide field sky surveys, such as the LSST, will start collecting data at enormous rates per day.

To make GLMs more accessible to the community, we developed a set of software libraries written in Python and R, that can be easily implemented into people's own work. A web application is also available to be used instantly without the need for installation.

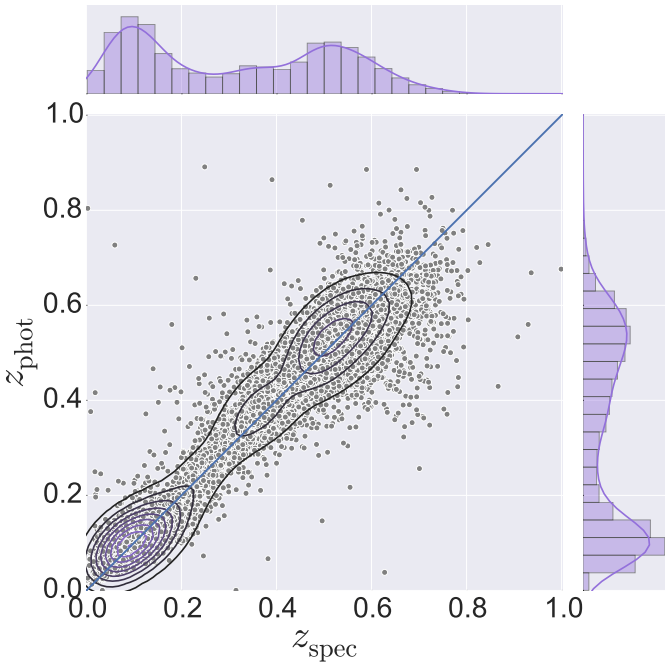

Fig. 1 The 2D probability density of the predicted redshift from the GLM fit vs. the spectroscopic redshift (central plots). The upper and right subplots in each panel depict the redshift distribution along photo-z and $z_{\rm spec}$, respectively.

Acknowledgements We thank V. Busti, E. D. Feigelson, M. Killedar, J. Buchner, and A. Trindade for interesting suggestions and comments. JE, RSS and EEOI thank the SIM Laboratory of the *Universidade de Lisboa* for hospitality during the development of this work. Cosmostatistics Initiative (COIN)[6] is a non-profit organisation whose aim is to nourish the synergy between astrophysics, cosmology, statistics and machine learning communities. This work was partially supported by the ESA VA4D project (AO 1-6740/11/F/MOS). AKM thanks the Portuguese agency *Fundação para Ciência e Tecnologia – FCT*, for financial support (SFRH/BPD/74697/2010). EEOI is partially supported by the Brazilian agency CAPES (grant number 9229-13-2). Work on this paper has substantially benefited from using the collaborative website AWOB developed and maintained by the Max-Planck Institute for Astrophysics and the Max-Planck Digital Library. This work was written on the collaborative `WriteLatex` platform, and made use of the GitHub a web-based hosting service and `git` version control software. This work made use of the cloud based hosting platform `ShinyApps.io`. This work used the following public scientific Python packages `scikit-learn v0.15` [11], `seaborn v0.3.1`, and `statsmodels v0.6.0`. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science.

[6] `https://asaip.psu.edu/organizations/iaa/iaa-working-group-of-cosmostatistics`

References

[1] David Brown, Peter Rothery, et al. *Models in biology: mathematics, statistics and computing*. John Wiley & Sons Ltd., 1993.

[2] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL http://doi.acm.org/10.1145/1970392.1970395.

[3] M. Carrasco Kind and R. J. Brunner. Exhausting the information: novel Bayesian combination of photometric redshift PDFs. MNRAS, 442:3380–3399, August 2014. doi: 10.1093/mnras/stu1098.

[4] R. S. de Souza, U. Maio, V. Biffi, and B. Ciardi. Robust PCA and MIC statistics of baryons in early minihaloes. MNRAS, 440:240–248, May 2014. doi: 10.1093/mnras/stu274.

[5] J. Elliott, R. S. de Souza, A. Krone-Martins, E. Cameron, E. E. O. Ishida, and J. Hilbe. The overlooked potential of Generalized Linear Models in astronomy-II: Gamma regression and photometric redshifts. *Astronomy and Computing*, 10:61–72, April 2015. doi: 10.1016/j.ascom.2015.01.002.

[6] J. Green, P. Schechter, C. Baltay, R. Bean, D. Bennett, R. Brown, C. Conselice, M. Donahue, and et al. Wide-Field InfraRed Survey Telescope (WFIRST) Final Report. *arxiv:1208.4012*, August 2012.

[7] H. Hildebrandt, S. Arnouts, P. Capak, L. A. Moustakas, C. Wolf, F. B. Abdalla, R. J. Assef, M. Banerji, and et. al. PHAT: PHoto-z Accuracy Testing. A&A, 523:A31, November 2010. doi: 10.1051/0004-6361/201014885.

[8] J. K. Lindsey. A review of some extensions to generalized linear models. *Statistics in medicine*, 18(17-18):2223–2236, 1999. ISSN 0277-6715. URL http://view.ncbi.nlm.nih.gov/pubmed/10474135.

[9] LSST Science Collaboration, P. A. Abell, J. Allison, S. F. Anderson, J. R. Andrew, J. R. P. Angel, L. Armus, D. Arnett, S. J. Asztalos, T. S. Axelrod, and et al. LSST Science Book, Version 2.0. *arxiv:0912.0201*, December 2009.

[10] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384, 1972.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Robert S Pindyck and Daniel L Rubinfeld. *Econometric models and economic forecasts*, volume 4. Irwin/McGraw-Hill Boston, 1998.

[13] A. Refregier, A. Amara, T. D. Kitching, A. Rassat, R. Scaramella, J. Weller, and f. t. Euclid Imaging Consortium. Euclid Imaging Consortium Science Book. *arxiv:1001.0061*, January 2010.

[14] D. G. York, J. Adelman, J. E. Anderson, Jr., S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, R. Barkhouser, et al., and SDSS Collaboration. The

Sloan Digital Sky Survey: Technical Summary. AJ, 120:1579–1587, September 2000. doi: 10.1086/301513.