Statistical methods in astronomy

JAMES P. LONG[1] AND RAFAEL S. DE SOUZA[2,3]

[1] Texas A&M University, College Station, TX, USA.
[2] MTA Eötvös University, EIRSA "Lendulet" Astrophysics Research Group, Budapest, Hungary.
[3] Instituto de Astronomia, Geofísica e Ciências Atmosféricas, USP, SP, Brazil.
* E-mail: jlong@stat.tamu.edu
* E-mail: rafaeldesouza@alumni.usp.br

We present a review of data types and statistical methods often encountered in astronomy. The aim is to provide an introduction to statistical applications in astronomy for statisticians and computer scientists. We highlight the complex, often hierarchical, nature of many astronomy inference problems and advocate for cross-disciplinary collaborations to address these challenges.

1. INTRODUCTION

Astronomy has a long history of exploiting observational data to estimate parameters and quantify uncertainty in physical models. Problems in astronomy propelled the development of many statistical techniques, from classical least squares estimation[1;2] to contemporary methods such as nested sampling[3;4].

Late 20th century advances in data collection, such as automation of telescopes and use of CCD cameras, resulted in a dramatic increase in data size and complexity, producing a surge in use and development of statistical methodology. Astronomers use these data sets for a diverse range of science goals, including modeling formation of galaxies, finding earth–like planets[5], estimating the metric expansion of space, and classifying transients.

This article reviews common data types and statistical methodology currently in use in astronomy, with the goal of making astronomical applications more accessible to methodological and applied statisticians. A non-exhaustive selection of topics is covered in this article. We refer readers to the "Further Reading" section at the end of this text in which historical and methodological viewpoints of astrostatistics are presented.

In Section 2 we review three common types of astronomical data: images, spectra, and time series. In Section 3 we discuss some statistical methods currently used in astronomy. Many of these methods are under active development within the statistics and computer science research communities. We conclude in Section 4 by describing one astrostatistics challenge, mapping the Milky Way halo with RR Lyrae stars, and the various statistical tools necessary for addressing this problem.

2. ASTRONOMICAL DATA TYPES

A. Image Data

Telescopes take images of the night sky. Figure S1 shows an image taken by the Dark Energy Camera (DECam) as part of the Dark Energy Survey (DES)[6]. DES takes approximately 400 one-gigabyte images per night.[1] Astronomical images are often taken with a photometric filter which blocks certain light wavelengths.

A *photometric pipeline* identifies objects in images and estimates their brightness. These pipelines contain many statistical tools such as machine learning algorithms (see Section E) and hierarchical models (see Section C). The pipeline outputs a *catalog* containing object positions, brightnesses, and classifications (star, galaxy, asteroid, etc.). Catalog data is typically much easier to study and model than the raw image data, so most subsequent analysis is performed on them.

B. Spectral Data

A spectra represents the intensity of light in different wavelengths, providing considerably more information than can be directly inferred from image data. Figure S2 shows the spectrum of the galaxy Messier 77, a barred spiral in the Cetus constellation. Spectra carry information

[1] See https://www.darkenergysurvey.org/the-des-project/survey-and-operations/data-management/.

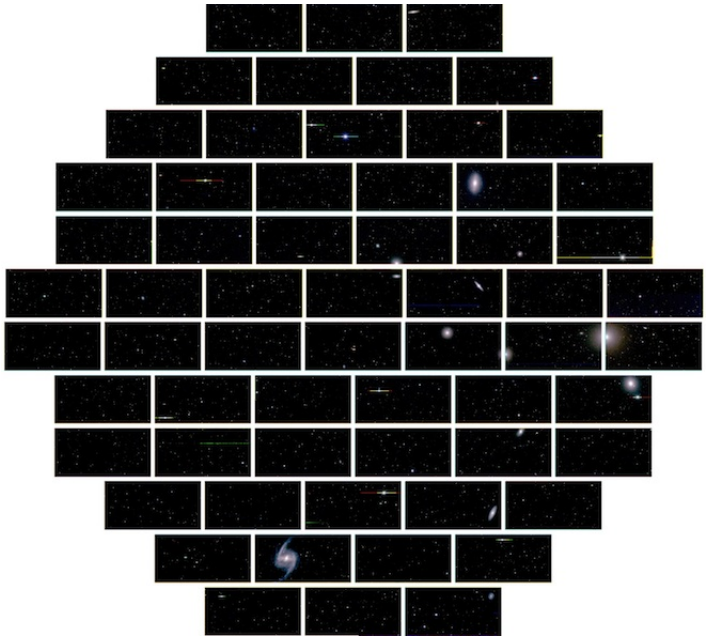


Fig. S1. Image of the night sky taken by the DECam. The white lines are gaps between the CCDs on the detector. Identifying, classifying, and estimating brightness of objects in images is a major statistical challenge in astronomy.

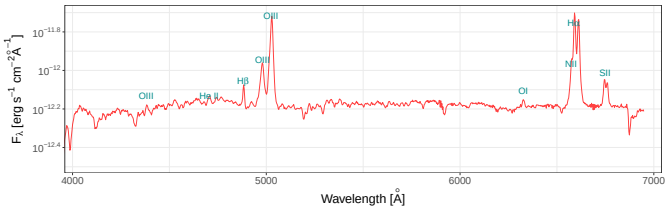


Fig. S2. Example of a galaxy spectra from Messier 77.

about some of the most important physical properties of astronomical objects such as temperature and chemical composition. Additionally, the displacement of spectral features towards longer wavelength (known as redshift) may be used to estimate object distance, thus providing a precious tool for understanding evolution of the universe.

There are several astronomical surveys collecting spectral information such as the RAdial Velocity Experiment [7], one of the largest spectroscopic surveys of Milky Way stars publicly available. It enables study of Milky Way morphology and history through stellar spectroscopic observations and astrometric databases. The SDSS-IV MaNGA Survey [8] collects $\sim$ 10,000 spectral measurements for nearby galaxies, enabling construction of two-dimensional maps of physical properties throughout each galaxy.

C. Time Series and Functional Data

Images and spectra represent two common forms of "raw" astronomical data, which together with photometric information, the integrated flux through a given filter, provide the basis to derive several data types. For example, many light sources vary in brightness as a function of time. Astronomical surveys which image the same area of the sky repeatedly over time produce a time series or *light curve* for each object, permitting analysis of temporal brightness variation.

Figure S3a shows a time series for a star observed by the Optical Gravitational Lensing Experiment (OGLE) [9]. The data was collected in two filters, represented by orange crosses and blue circles over the course of approximately 10 years. The *cadence*, or time spacing between observations, is irregular, a typical feature in astronomy data. OGLE has collected approximately 400,000 of these light curves, all of which are publicly available.[2] The statistical challenges with this data include modeling shape variation and classifying sources based on the astrophysical reason for brightness variation. For example, the star in Figure S3a is varying in brightness periodically over time. From this data, one can estimate a period and plot magnitude versus time modulo period (see Figure S3b). Methods for estimating periods for this type of data are under active development [10;11]. For comparison, in Figure S4 we show a supernova spectra as a function of time with the epoch of maximum brightness highlighted in red. [12]

3. STATISTICAL METHODOLOGY IN ASTRONOMY

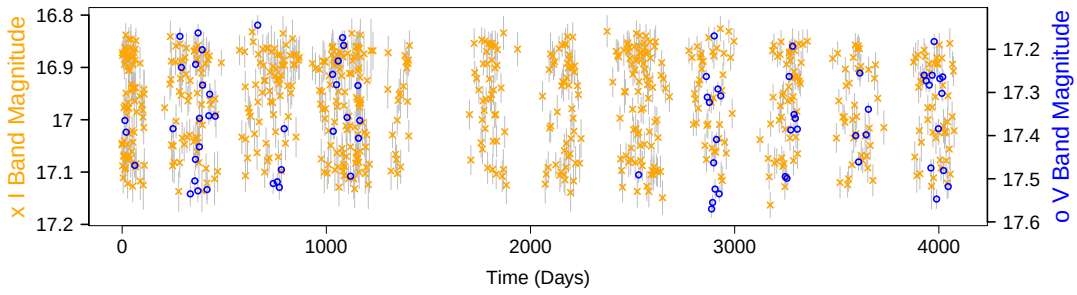
Astrostatisticians use a wide range of statistical methods to analyze these complex data sets. We now discuss several areas of statistical methodology with recent applications within astronomy.

A. Measurement Error Models

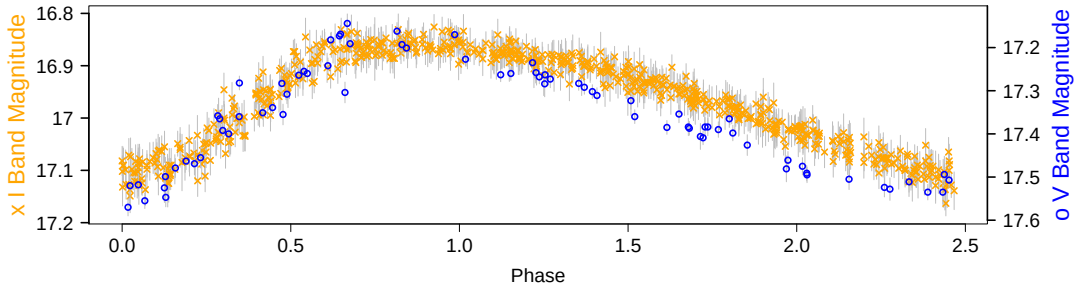
Models typically assume homoskedasticity (i.e. errors with the same variance). Predictor (i.e. independent) variables in regression models are often assumed to be measured without error. However in Astronomy, heteroskedastic errors are the norm. Further, it is common practice to have estimates of the measurement error variances available through modeling of uncertainties inherent to the detection procedure [13]. In cases where the measurement error is large, explicit errors-in-variables models are necessary to avoid biased estimates, particularly in regression models. These models often have a hierarchical structure in which the true predictor values are treated as parameters.

Approaches commonly used in astronomy to solve this problem include the bivariate correlated errors and intrinsic scatter model [14] and hierarchical Bayesian models [15;16;2;17]. Examples of applications include the devel-

[2] http://ogledb.astrouw.edu.pl/~ogle/CVS/



(a) Light Curve



(b) Folded Light Curve

Fig. S3. (a) Light curve of a variable star observed by OGLE. Models from the time series and functional data analysis literature are often used for studying these objects. (b) The light curve in a) is produced by a periodic variable star. From the data in a) one can estimate a period ($\approx$ 2.48 days) and plot the folded light curve, magnitude versus phase (= time modulo period).

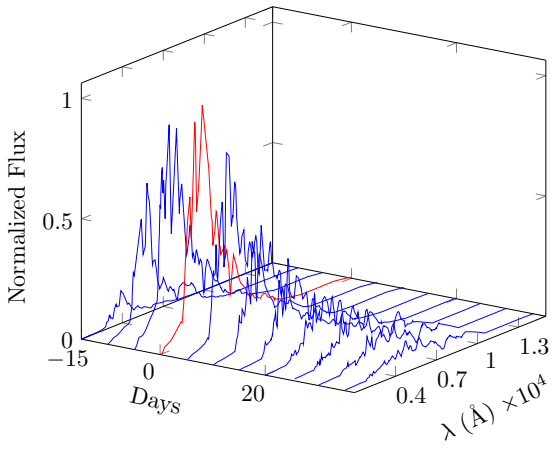


Fig. S4. Example of a supernova spectrum as function of days since maximum brightness.

opment of a Gaussian mixture model for estimating a density from observations subject to measurement error[18], the subsequent application of this model to account for flux uncertainties in probabilistic classification of quasars[19], and use of a hierarchical Bayesian model to handle discrete measurement uncertainties in a negative binomial model to probe the population of globular clusters in galaxies[20].

B. Survival Analysis

Astronomical surveys are, by construction, unable to obtain unbiased samples from the population of objects. Surveys are often magnitude-limited, i.e. brighter objects are more likely to be detected. This results in truncation due to the telescope sensitivity limit, which in astronomy is called *Malmquist bias*[21;22]. In other situations we know an object exists, but some of its features are too faint to be detected, resulting in censored observations. In statistics, solutions for such problems are treated under the general umbrella of survival analysis. Censoring and truncation are often called selection effects in astronomy[13]. Survival analysis challenges in astronomy may involve multivariate data[23], nonparametric density estimation with truncation[24], or selection effects within a regression model[15].

C. Bayesian Models and Computation

Use of Bayesian methodology has grown considerably in astronomy over the past three decades. Active areas of Bayesian research include hierarchical models, posterior samplers, and models for complex data types such as images and functions.

Hierarchical Bayesian models (HBM) are used for problems where individual object parameters and population parameters are unknown. HBM are applied to many types of data in astronomy as for instance to detect and characterize galaxies in astronomical images using a HBM with variational inference to approximate the posterior [25], for modeling of supernovae light curves [26;27], and to fit cosmic ray data [28].

Approximate Bayesian computation (ABC) avoids computationally expensive likelihood evaluations by simulating data sets and comparing the distance between the simulated data and the actual data. ABC is being used in astronomy for inferring cosmological parameters [29;30;31;32] and probing galaxy evolution [33;34]. The growing use of ABC has lead to the development of software packages, such as *cosmoabc*, an ABC sampler via Population Monte Carlo for general astronomical applications [35].

Several flavors and variants of Markov Chain Monte Carlo samplers have been developed by astronomers including an implementation of an affine–invariant ensemble sampler [36] and the Diffusive Nested Sampling, an extension of nested samplers [37].

D. Generalized Linear Models

The ubiquitous linear regression model relies on a number of distributional assumptions which fail to hold when the data come from *exponential family* distributions other than the Gaussian. Generalized linear models (GLMs) [38], assume, through a link function, a linear relationship between the response variable y and set of predictors x. Several problems in astronomy require the use of GLMs and extensions, such as modeling the fraction of Seyfert galaxies in terms of environment (Bernoulli) [39;40], the population of globular clusters as a function of the host galaxy properties (Negative binomial) [20], and the distance of galaxies as a function of their colors (Gamma) [41].

E. Machine Learning

For several astronomy problems, prediction and pattern recognition are more important than parameter estimation. Methodology from the machine learning (ML) community is now routinely used to solve these problems [42;43]. While “off the shelf" machine learning methods are sometimes sufficient, astronomy ML problems may involve additional challenges, such as biased training sets, computationally intensive feature extraction, or real–time classification, which inhibit use of standard methods. We describe some of these challenges below.

ML has seen extensive use in classification of variable source light curves (see Section C for definition). Here astronomers are often more interested in determining the class of the source than estimating source parameters. Since light curves are functions, this is a functional classification problem. A common approach to the problem is to construct a training set of objects of known class (often using some level of human classification), extract features for these objects, and then train a ML classifier on this data. The classifier can then, in principle, be used to classify new objects in other surveys [44;45;46;47]. ML tools are used for several other problems in astronomy including identification of sources in images [48], clustering of spectral data [49], and photometric redshift estimation [50;51].

A critical issue often overlooked is the lack of representativeness between spectroscopic and photometric samples. Cross-validation performance measures have been shown to be misleading in this situation. [52] Mismatches between training and test samples are not exclusive to astronomical problems. Methodology developed by the ML community to address this challenge has been used on several astronomy problems, including domain adaptation [53;52], active learning [54], and a combination of both under the umbrella of adaptive learning techniques [55]. However challenges remain, including incorporating feature measurement error, missing data, censoring and truncation into ML algorithms [56;57].

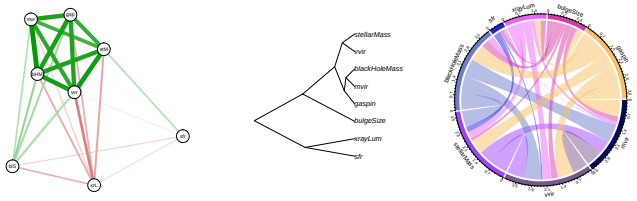


Fig. S5. Three visualizations (left to right: graph, cladogram and a chord diagram) of the same galaxy catalog from a N-body/hydro cosmological simulation.

F. Information Visualization

Visualization methods exploit the human visual system to optimize intuitive insight into data structure. Whilst the role of visualization belongs to the groundwork of astronomical analysis, new paradigms for multidimensional data visualization are yet to be fully utilized. Patterns and non-trivial correlations that might go undetected in tabular-based data, can be unfolded if the proper tools are applied [58]. Among the methods that have been developed to facilitate the exploration of multivariate astronomical data are phylogenetic trees [59], graphs, chords [58], and starfish diagrams [60]. Figure S5 shows a dataset from a N-body/Hydro cosmological simulation [61] visualized with three different techniques.

4. COMPLEX INFERENCE CHALLENGES IN ASTRONOMY: AN EXAMPLE

The process of turning data into scientific knowledge discovery typically requires the use of many statistical tools,

often in innovative ways. Some of the most challenging statistical questions that arise in astronomy relate to how to merge these tools into a data analysis pipeline that permits valid statistical inferences while remaining computationally feasible.

As an illustrative example, consider the challenge of mapping the Milky Way halo, the region of space that surrounds our galaxy. This problem has attracted much recent attention.[62;63;64;65;66] Astronomers would like to produce maps of the locations of stars in the halo and identify structures, such as collections of gravitationally bound stars. This has important consequences for the Λ Cold Dark Matter (ΛCDM) cosmological model, our current framework for understanding how the universe was born and developed. Creating halo maps is difficult because it is impossible to determine the distance to most stars. We can, however, estimate distances to a small subset of stars, known as RR Lyrae (RRL), due to their all having similar luminosities (standard candles in astronomy). The locations of these stars trace the structure in the halo. Inference on the Milky Way halo requires:

1. Identifying the RRL stars among all stars observed in an astronomical survey. Recalling that variable stars are, as data, irregularly sampled functions (see Figure S3a), this is a large functional data classification problem. Once the RRL have been identified, we estimate their distance.

2. Using the estimated locations of the RRL, we estimate the local density of objects in order to identify structure. Often RRL locations are viewed as a realization from a Poisson process in three dimensional space. Errors from the previous step, including misclassified stars and uncertainty on distance estimates impact this map estimate.

3. Finally, one can compare the observed structure in the halo map to predictions made by the ΛCDM cosmological model. Different halo structures provide evidence for different values of the free parameters in ΛCDM. These comparisons could be heuristic or more quantitative (e.g. optimizing parameters in a cosmological simulation to produce halo structure which most closely resembles observations).

Schafer[67] argues that cosmological inference problems are best divided into three stages: inference on object parameters, inferences on class parameters, and finally inferences on the fundamental cosmological parameters. The three steps above roughly correspond to these stages. Each stage requires many statistical decisions. Uncertainty must be propagated through the stages while at the same time approximations must be made to keep the analysis pipeline computationally feasible.

Upcoming astronomical sky surveys, such as the Transiting Exoplanet Survey Satellite (TESS)[3], the James Webb Space Telescope (JWST)[4], and the Large Synoptic Survey Telescope (LSST)[5], promise ever larger data sets with more challenging inference problems. Interdisciplinary collaborations of statisticians and astronomers will be essential for developing the new statistical methodology necessary for fully realizing the science potential of these projects.

[3]https://tess.gsfc.nasa.gov/

[4]https://www.jwst.nasa.gov/

[5]https://www.lsst.org/

RELATED ARTICLES

- E. D. Feigelson, *Astronomy, Statistics In*. John Wiley & Sons, Inc., 2004
- A. A. Goodman, "Principles of high-dimensional data visualization in astronomy," *Astronomische Nachrichten*, vol. 333, p. 505, June 2012
- T. J. Loredo, "Bayesian astrostatistics: a backward look to the future," in *Astrostatistical challenges for the new astronomy*, pp. 15–40, Springer, 2013
- J. M. Hilbe, *Astrostatistics*. John Wiley & Sons, Ltd, 2014
- J. M. Hilbe, J. Riggs, B. D. Wandelt, R. S. de Souza, E. E. O. Ishida, J. Cisewski, V. Surdin, M. Killedar, R. Trotta, B. Bassett, Y. Fantaye, and C. Impey, "Life, the universe, and everything," *Significance*, vol. 11, no. 5, pp. 48–75, 2014
- E. Cameron, "What we talk about when we talk about fields," in *Statistical Challenges in 21st Century Cosmology* (A. Heavens, J.-L. Starck, and A. Krone-Martins, eds.), vol. 306 of *IAU Symposium*, pp. 9–12, May 2014
- C. M. Schafer, "A framework for statistical inference in astrophysics," *Annual Review of Statistics and Its Application*, vol. 2, pp. 141–162, 2015
- E. D. Feigelson, "The changing landscape of astrostatistics and astroinformatics," in *IAU Symposium*, vol. 325 of *IAU Symposium*, pp. 3–9, June 2017
- S. Sharma, "Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy," *arXiv:1706.01629*, June 2017

REFERENCES

1. S. M. Stigler, *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986.
2. J. M. Hilbe, R. S. de Souza, and E. E. O. Ishida, *Bayesian Models for Astrophysical Data: Using R, JAGS, Python, and Stan*. Cambridge University Press, 2017.
3. J. Skilling, "Nested Sampling," in *American Institute of Physics Conference Series* (R. Fischer, R. Preuss, and U. V. Toussaint, eds.), vol. 735 of *American Institute of Physics Conference Series*, pp. 395–405, Nov. 2004.

4. F. Feroz and M. P. Hobson, "Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses," *Monthly Notices of the Royal Astronomical Society*, vol. 384, pp. 449–463, Feb. 2008.

5. D. Foreman-Mackey, D. W. Hogg, and T. D. Morton, "Exoplanet Population Inference and the Abundance of Earth Analogs from Noisy, Incomplete Catalogs," ApJ, vol. 795, p. 64, Nov. 2014.

6. B. Flaugher, H. Diehl, K. Honscheid, T. Abbott, O. Alvarez, R. Angstadt, J. Annis, M. Antonik, O. Ballester, L. Beaufore, *et al.*, "The dark energy camera," *The Astronomical Journal*, vol. 150, no. 5, p. 150, 2015.

7. A. Kunder, G. Kordopatis, M. Steinmetz, T. Zwitter, P. J. McMillan, L. Casagrande, H. Enke, J. Wojno, M. Valentini, C. Chiappini, G. Matijevič, A. Siviero, P. de Laverny, A. Recio-Blanco, A. Bijaoui, R. F. G. Wyse, J. Binney, E. K. Grebel, A. Helmi, P. Jofre, T. Antoja, G. Gilmore, A. Siebert, B. Famaey, O. Bienaymé, B. K. Gibson, K. C. Freeman, J. F. Navarro, U. Munari, G. Seabroke, B. Anguiano, M. Žerjal, I. Minchev, W. Reid, J. Bland-Hawthorn, J. Kos, S. Sharma, F. Watson, Q. A. Parker, R.-D. Scholz, D. Burton, P. Cass, M. Hartley, K. Fiegert, M. Stupar, A. Ritter, K. Hawkins, O. Gerhard, W. J. Chaplin, G. R. Davies, Y. P. Elsworth, M. N. Lund, A. Miglio, and B. Mosser, "The Radial Velocity Experiment (RAVE): Fifth Data Release," *Astronomical Journal*, vol. 153, p. 75, Feb. 2017.

8. K. Bundy, M. A. Bershady, D. R. Law, R. Yan, N. Drory, N. MacDonald, D. A. Wake, B. Cherinka, J. R. Sánchez-Gallego, A.-M. Weijmans, D. Thomas, C. Tremonti, K. Masters, L. Coccato, A. M. Diamond-Stanic, A. Aragón-Salamanca, V. Avila-Reese, C. Badenes, J. Falcón-Barroso, F. Belfiore, D. Bizyaev, G. A. Blanc, J. Bland-Hawthorn, M. R. Blanton, J. R. Brownstein, N. Byler, M. Cappellari, C. Conroy, A. A. Dutton, E. Emsellem, J. Etherington, P. M. Frinchaboy, H. Fu, J. E. Gunn, P. Harding, E. J. Johnston, G. Kauffmann, K. Kinemuchi, M. A. Klaene, J. H. Knapen, A. Leauthaud, C. Li, L. Lin, R. Maiolino, V. Malanushenko, E. Malanushenko, S. Mao, C. Maraston, R. M. McDermid, M. R. Merrifield, R. C. Nichol, D. Oravetz, K. Pan, J. K. Parejko, S. F. Sanchez, D. Schlegel, A. Simmons, O. Steele, M. Steinmetz, K. Thanjavur, B. A. Thompson, J. L. Tinker, R. C. E. van den Bosch, K. B. Westfall, D. Wilkinson, S. Wright, T. Xiao, and K. Zhang, "Overview of the SDSS-IV MaNGA Survey: Mapping nearby Galaxies at Apache Point Observatory," *Astrophysical Journal*, vol. 798, p. 7, Jan. 2015.

9. A. Udalski, M. Szymanski, I. Soszynski, and R. Poleski, "The optical gravitational lensing experiment. final reductions of the ogle-iii data," *Acta Astronomica*, vol. 58, pp. 69–87, 2008.

10. J. P. Long, E. C. Chi, R. G. Baraniuk, *et al.*, "Estimating a common period for a set of irregularly sampled functions with applications to periodic variable star data," *The Annals of Applied Statistics*, vol. 10, no. 1, pp. 165–197, 2016.

11. J. T. VanderPlas and Ž. Ivezic, "Periodograms for multiband astronomical time series," *The Astrophysical Journal*, vol. 812, no. 1, p. 18, 2015.

12. R. Kessler, B. Bassett, P. Belov, V. Bhatnagar, H. Campbell, A. Conley, J. A. Frieman, A. Glazov, S. González-Gaitán, R. Hlozek, S. Jha, S. Kuhlmann, M. Kunz, H. Lampeitl, A. Mahabal, J. Newling, R. C. Nichol, D. Parkinson, N. Sajeeth Philip, D. Poznanski, J. W. Richards, S. A. Rodney, M. Sako, D. P. Schneider, M. Smith, M. Stritzinger, and M. Varughese, "Results from the Supernova Photometric Classification Challenge," PASP, vol. 122, p. 1415, Dec. 2010.

13. E. Feigelson and G. Babu, *Modern Statistical Methods for Astronomy: With R Applications*. Cambridge University Press, 2012.

14. M. G. Akritas and M. A. Bershady, "Linear Regression for Astronomical Data with Measurement Errors and Intrinsic Scatter," ApJ, vol. 470, p. 706, Oct. 1996.

15. B. C. Kelly, "Some aspects of measurement error in linear regression of astronomical data," *The Astrophysical Journal*, vol. 665, no. 2, p. 1489, 2007.

16. T. J. Loredo, "Accounting for Source Uncertainties in Analyses of Astronomical Survey Data," in *American Institute of Physics Conference Series* (R. Fischer, R. Preuss, and U. V. Toussaint, eds.), vol. 735 of *American Institute of Physics Conference Series*, pp. 195–206, Nov. 2004.

17. S. Andreon and B. Weaver, *Bayesian Methods for the Physical Sciences: Learning from Examples in Astronomy and Physics*. Springer Series in Astrostatistics, Springer International Publishing, 2015.

18. J. Bovy, D. W. Hogg, and S. T. Roweis, "Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations," *The Annals of Applied Statistics*, pp. 1657–1677, 2011.

19. J. Bovy, J. F. Hennawi, D. W. Hogg, A. D. Myers, J. A. Kirkpatrick, D. J. Schlegel, N. P. Ross, E. S. Sheldon, I. D. McGreer, D. P. Schneider, *et al.*, "Think outside the color box: Probabilistic target selection and the sdss-xdqso quasar targeting catalog," *The Astrophysical Journal*, vol. 729, no. 2, p. 141, 2011.

20. R. S. de Souza, J. M. Hilbe, B. Buelens, J. D. Riggs, E. Cameron, E. E. O. Ishida, A. L. Chies-Santos, and

M. Killedar, "The overlooked potential of generalized linear models in astronomy - III. Bayesian negative binomial regression and globular cluster populations," *Monthly Notices of the Royal Astronomical Society*, vol. 453, pp. 1928–1940, Oct. 2015.

21. K. G. Malmquist, "On some relations in stellar statistics," *Meddelanden fran Lunds Astronomiska Observatorium Serie I*, vol. 100, pp. 1–52, Mar. 1922.

22. A. Sandage, *Malmquist Bias and Completeness Limits*. Nov. 2000.

23. C. M. Schafer, "A statistical method for estimating luminosity functions using truncated data," *The Astrophysical Journal*, vol. 661, no. 2, p. 703, 2007.

24. B. Efron and V. Petrosian, "Nonparametric methods for doubly truncated data," *Journal of the American Statistical Association*, vol. 94, no. 447, pp. 824–834, 1999.

25. J. Regier, A. Miller, J. McAuliffe, R. Adams, M. Hoffman, D. Lang, D. Schlegel, and M. Prabhat, "Celeste: Variational inference for a generative model of astronomical images," in *International Conference on Machine Learning*, pp. 2095–2103, 2015.

26. K. S. Mandel, G. Narayan, and R. P. Kirshner, "Type ia supernova light curve inference: Hierarchical models in the optical and near-infrared," *The Astrophysical Journal*, vol. 731, no. 2, p. 120, 2011.

27. N. Sanders, M. Betancourt, and A. Soderberg, "Unsupervised transient light curve analysis via hierarchical bayesian inference," *The Astrophysical Journal*, vol. 800, no. 1, p. 36, 2015.

28. K. Soiaporn, D. Chernoff, T. Loredo, D. Ruppert, I. Wasserman, *et al.*, "Multilevel bayesian framework for modeling the production, propagation and detection of ultra-high energy cosmic rays," *The Annals of Applied Statistics*, vol. 7, no. 3, pp. 1249–1285, 2013.

29. C. M. Schafer and P. E. Freeman, "Likelihood-free inference in cosmology: Potential for the estimation of luminosity functions," in *Statistical Challenges in Modern Astronomy V*, pp. 3–19, Springer, 2012.

30. A. Weyant, C. Schafer, and W. M. Wood-Vasey, "Likelihood-free cosmological inference with type ia supernovae: approximate bayesian computation for a complete treatment of uncertainty," *The Astrophysical Journal*, vol. 764, no. 2, p. 116, 2013.

31. J. Akeret, A. Refregier, A. Amara, S. Seehars, and C. Hasner, "Approximate Bayesian computation for forward modeling in cosmology," *Journal of Cosmology and Astroparticle Physics*, vol. 8, p. 043, Aug. 2015.

32. E. Jennings and M. Madigan, "astroABC : An Approximate Bayesian Computation Sequential Monte Carlo sampler for cosmological parameter estimation," *Astronomy and Computing*, vol. 19, pp. 16–22, Apr. 2017.

33. E. Cameron and A. N. Pettitt, "Approximate Bayesian Computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift," *Monthly Notices of the Royal Astronomical Society*, vol. 425, pp. 44–65, Sept. 2012.

34. C. Hahn, M. Vakili, K. Walsh, A. P. Hearin, D. W. Hogg, and D. Campbell, "Approximate Bayesian computation in large-scale structure: constraining the galaxy-halo connection," *Monthly Notices of the Royal Astronomical Society*, vol. 469, pp. 2791–2805, Aug. 2017.

35. E. Ishida, S. Vitenti, M. Penna-Lima, J. Cisewski, R. de Souza, A. Trindade, E. Cameron, V. Busti, C. collaboration, *et al.*, "cosmoabc: likelihood-free inference via population monte carlo approximate bayesian computation," *Astronomy and Computing*, vol. 13, pp. 1–11, 2015.

36. D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, "emcee: The mcmc hammer," *Publications of the Astronomical Society of the Pacific*, vol. 125, no. 925, p. 306, 2013.

37. B. J. Brewer, L. B. Pártay, and G. Csányi, "Diffusive nested sampling," *Statistics and Computing*, vol. 21, no. 4, pp. 649–656, 2011.

38. J. Neuhaus and C. McCulloch, "Generalized linear models," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 5, pp. 407–413, 2011.

39. R. S. de Souza, E. Cameron, M. Killedar, J. Hilbe, R. Vilalta, U. Maio, V. Biffi, B. Ciardi, and J. D. Riggs, "The overlooked potential of Generalized Linear Models in astronomy, I: Binomial regression," *Astronomy and Computing*, vol. 12, pp. 21–32, Sept. 2015.

40. R. S. de Souza, M. L. L. Dantas, A. Krone-Martins, E. Cameron, P. Coelho, M. W. Hattab, M. de Val-Borro, J. M. Hilbe, J. Elliott, A. Hagen, and COIN Collaboration, "Is the cluster environment quenching the Seyfert activity in elliptical and spiral galaxies?," *Monthly Notices of the Royal Astronomical Society*, vol. 461, pp. 2115–2125, Sept. 2016.

41. 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*, vol. 10, pp. 61–72, Apr. 2015.

42. N. M. Ball and R. J. Brunner, "Data Mining and Machine Learning in Astronomy," *International Journal of Modern Physics D*, vol. 19, pp. 1049–1106, 2010.

43. M. Brescia, S. Cavuoti, G. S. Djorgovski, C. Donalek, G. Longo, and M. Paolillo, *Extracting Knowledge from Massive Astronomical Data Sets*, p. 31. 2012.

44. J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard, "On machine-learned classification of variable stars with sparse and noisy time-series data," *The Astrophysical Journal*, vol. 733, no. 1, p. 10, 2011.

45. E. E. O. Ishida and R. S. de Souza, "Kernel PCA for Type Ia supernovae photometric classification," *Monthly Notices of the Royal Astronomical Society*, vol. 430, pp. 509–532, Mar. 2013.

46. J. Faraway, A. Mahabal, J. Sun, X.-F. Wang, Y. G. Wang, and L. Zhang, "Modeling lightcurves for improved classification of astronomical objects," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 9, no. 1, pp. 1–11, 2016.

47. M. Lochner, J. D. McEwen, H. V. Peiris, O. Lahav, and M. K. Winter, "Photometric Supernova Classification with Machine Learning," ApJS, vol. 225, p. 31, Aug. 2016.

48. H. Brink, J. W. Richards, D. Poznanski, J. S. Bloom, J. Rice, S. Negahban, and M. Wainwright, "Using machine learning for discovery in synoptic survey imaging data," *Monthly Notices of the Royal Astronomical Society*, p. stt1306, 2013.

49. M. Sasdelli, E. E. O. Ishida, R. Vilalta, M. Aguena, V. C. Busti, H. Camacho, A. M. M. Trindade, F. Gieseke, R. S. de Souza, Y. T. Fantaye, and P. A. Mazzali, "Exploring the spectroscopic diversity of Type Ia supernovae with DRACULA: a machine learning approach," *Monthly Notices of the Royal Astronomical Society*, vol. 461, pp. 2044–2059, Sept. 2016.

50. T. Budavári, "Photometric redshifts," *Advances in Machine Learning and Data Mining for Astronomy*, p. 323, 2012.

51. P. Freeman, J. Newman, A. Lee, J. Richards, and C. Schafer, "Photometric redshift estimation using spectral connectivity analysis," *Monthly Notices of the Royal Astronomical Society*, vol. 398, no. 4, pp. 2012–2021, 2009.

52. R. Beck, C.-A. Lin, E. E. O. Ishida, F. Gieseke, R. S. de Souza, M. V. Costa-Duarte, M. W. Hattab, and A. Krone-Martins, "On the realistic validation of photometric redshifts," *Monthly Notices of the Royal Astronomical Society*, vol. 468, pp. 4323–4339, July 2017.

53. A. A. Mahabal, D. Crichton, S. G. Djorgovski, E. Law, and J. S. Hughes, "From Sky to Earth: Data Science Methodology Transfer," in *IAU Symposium*, vol. 325 of *IAU Symposium*, pp. 17–26, June 2017.

54. J. W. Richards, D. L. Starr, H. Brink, A. A. Miller, J. S. Bloom, N. R. Butler, J. B. James, J. P. Long, and J. Rice, "Active Learning to Overcome Sample Selection Bias: Application to Photometric Variable Star Classification," ApJ, vol. 744, p. 192, Jan. 2012.

55. K. D. Gupta, R. Pampana, R. Vilalta, E. E. O. Ishida, and R. S. de Souza, "Automated supernova ia classification using adaptive learning techniques," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8, Dec 2016.

56. G. J. Babu and A. Mahabal, "Skysurveys, light curves and statistical challenges," *International Statistical Review*, 2015.

57. J. Bloom, J. Richards, P. Nugent, R. Quimby, M. Kasliwal, D. Starr, D. Poznanski, E. Ofek, S. Cenko, N. Butler, *et al.*, "Automating discovery and classification of transients and variable stars in the synoptic survey era," *Publications of the Astronomical Society of the Pacific*, vol. 124, no. 921, p. 1175, 2012.

58. R. S. de Souza and B. Ciardi, "AMADA-Analysis of multidimensional astronomical datasets," *Astronomy and Computing*, vol. 12, pp. 100–108, Sept. 2015.

59. D. Fraix-Burnet, "Concepts of Classification and Taxonomy Phylogenetic Classification," in *EAS Publications Series*, vol. 77 of *EAS Publications Series*, pp. 221–257, May 2016.

60. I. S. Konstantopoulos, "The starfish diagram: Visualising data within the context of survey samples," *Astronomy and Computing*, vol. 10, pp. 116–120, Apr. 2015.

61. Q. Guo, S. White, M. Boylan-Kolchin, G. De Lucia, G. Kauffmann, G. Lemson, C. Li, V. Springel, and S. Weinmann, "From dwarf spheroidals to cD galaxies: simulating the galaxy population in a ΛCDM cosmology," *Monthly Notices of the Royal Astronomical Society*, vol. 413, pp. 101–131, May 2011.

62. B. Sesar, S. R. Banholzer, J. G. Cohen, N. F. Martin, C. J. Grillmair, D. Levitan, R. R. Laher, E. O. Ofek, J. A. Surace, S. R. Kulkarni, T. A. Prince, and H.-W. Rix, "Stacking the invisibles: A guided search for low-luminosity milky way satellites," *The Astrophysical Journal*, vol. 793, no. 2, p. 135, 2014.

63. B. Sesar, N. Hernitschek, S. Mitrović, Ž. Ivezić, H.-W. Rix, J. G. Cohen, E. J. Bernard, E. K. Grebel, N. F. Martin, E. F. Schlafly, *et al.*, "Machine-learned identification of rr lyrae stars from sparse, multi-band data: the ps1 sample," *arXiv preprint arXiv:1611.08596*, 2016.

64. B. Sesar, Ž. Ivezić, S. H. Grammer, D. P. Morgan, A. C. Becker, M. Jurić, N. De Lee, J. Annis, T. C. Beers, X. Fan, *et al.*, "Light curve templates and galactic distribution of rr lyrae stars from sloan digital sky survey stripe 82," *The Astrophysical Journal*, vol. 708, no. 1, p. 717, 2009.

65. R. Zinn, B. Horowitz, A. Vivas, C. Baltay, N. Ellman, E. Hadjiyska, D. Rabinowitz, and L. Miller, "La silla quest rr lyrae star survey: Region i," *The Astrophysical Journal*, vol. 781, no. 1, p. 22, 2013.

66. A. K. Vivas and R. Zinn, "The quest rr lyrae survey. ii. the halo overdensities in the first catalog," *The Astronomical Journal*, vol. 132, no. 2, p. 714, 2006.

67. C. M. Schafer, "A framework for statistical inference in astrophysics," *Annual Review of Statistics and Its Application*, vol. 2, pp. 141–162, 2015.

68. E. D. Feigelson, *Astronomy, Statistics In*. John Wiley & Sons, Inc., 2004.

69. A. A. Goodman, "Principles of high-dimensional data visualization in astronomy," *Astronomische Nachrichten*, vol. 333, p. 505, June 2012.

70. T. J. Loredo, "Bayesian astrostatistics: a backward look to the future," in *Astrostatistical challenges for the new astronomy*, pp. 15–40, Springer, 2013.

71. J. M. Hilbe, *Astrostatistics*. John Wiley & Sons, Ltd, 2014.

72. J. M. Hilbe, J. Riggs, B. D. Wandelt, R. S. de Souza, E. E. O. Ishida, J. Cisewski, V. Surdin, M. Killedar, R. Trotta, B. Bassett, Y. Fantaye, and C. Impey, "Life, the universe, and everything," *Significance*, vol. 11, no. 5, pp. 48–75, 2014.

73. E. Cameron, "What we talk about when we talk about fields," in *Statistical Challenges in 21st Century Cosmology* (A. Heavens, J.-L. Starck, and A. Krone-Martins, eds.), vol. 306 of *IAU Symposium*, pp. 9–12, May 2014.

74. E. D. Feigelson, "The changing landscape of astrostatistics and astroinformatics," in *IAU Symposium*, vol. 325 of *IAU Symposium*, pp. 3–9, June 2017.

75. S. Sharma, "Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy," *arXiv:1706.01629*, June 2017.