Part 7. Data Processing Techniques

https://doi.org/10.1017/S0074180900129043 Published online by Cambridge University Press

STATISTICAL METHODOLOGY FOR LARGE ASTRONOMICAL SURVEYS

E.D. FEIGELSON[1] AND G.J. BABU[2]
[1] *Dept. of Astron & Astrophys, Pennsylvania State University*
[2] *Dept. of Statistics, Pennsylvania State University*

Abstract. Multiwavelength surveys present a variety of challenging statistical problems: raw data processing, source identification, source characterization and classification, and interrelations between multiwavelength properties. For these last two issues, we discuss the applicability of standard and new multivariate statistical techniques. Traditional methods such as ANOVA, principal components analysis, cluster analysis, and tests for multivariate linear hypotheses are underutilized in astronomy and can be very helpful. Newer statistical methods such as projection pursuit, multivariate splines, and visualization tools such as XGobi are briefly introduced. However, multivariate databases from astronomical surveys present significant challenges to the statistical community. These include treatments of heteroscedastic measurement errors, censoring and truncation due to flux limits, and parameter estimation for nonlinear astrophysical models.

1. Introduction

Between the 16th and 19th centuries, astronomy and statistics were closely allied fields. Many of the foundations of mathematical statistics were laid by astronomers such as Tycho Brahe, Galileo, Tobias Mayer and Adrien Legendre (Stigler 1986). But this relationship weakened during the late 19th century, as statistics turned to applications in the social sciences and industry, astronomy reaped benefits from mathematical physics. A byproduct of this shift is that most astronomers are trained by physicists and receive little or no formal education in statistics. Most astronomers are thus only vaguely aware of the tremendous advances in statistical theory and practice of the last few decades. Similarly, with the notable exception

B.J. McLean et al. (eds.), New Horizons from Multi-Wavelength Sky Surveys, 363–370.

of galaxy clustering studies by Jerzy Neyman and Elizabeth Scott in the 1950–60s, statisticians became unaware of the tremendous developments in astronomy.

Mutual interest in astrostatistics has reemerged during the past decade. The comparison of astronomical data to astrophysical questions is becoming increasingly complex, outpacing the capabilities of traditional statistical methods. About 500 astronomical papers annually have 'statistics' or 'statistical' in their abstracts, yet they rarely refer to contemporary statistical texts or monographs for methodological guidance. Statistical procedures implemented in *Numerical Recipes* (Press *et al.* 1992) are used on a daily basis.

Recent cross-disciplinary efforts in astrostatistics have produced valuable resources. A number of conferences have been held in Europe (*e.g.*, Rolfe 1983; Jaschek & Murtagh 1990; Subba Rao 1997) and the U.S. (Feigelson & Babu 1992; Babu & Feigelson 1997), astrostatistical sessions at large meetings are being organized, an introductory monograph on astrostatistics has emerged (Babu & Feigelson 1996), and the Statistical Consulting Center for Astronomy is active (Feigelson *et al.* 1995; `http://www.stat.psu.edu/scca`). A monograph on multivariate data analysis, with FORTRAN codes and bibliography of astronomical applications, is very relevant to the issues discussed here (Murtagh & Heck 1987).

2. Statistics and Astronomical Surveys

Large astronomical surveys from new high-throughput detectors and observatories are powerful motivators for more effective statistical techniques. Observatories now frequently generate gigabytes of information every day, with terabyte-size raw databases which produce reduced catalogues of 10^6–10^9 objects. These catalogues, which may include up to dozens of observational properties of each object, often contain heterogeneous populations which must be isolated prior to detailed analysis. Although there are many types of astronomical surveys with many different goals, the statistical problems arising in their analysis can often be divided into three stages. We treat the first two stages very briefly here to concentrate on the final phase.

Reducing raw data into images The treatment of the raw data from the telescope or satellite observatory can be very complex, and has embedded within it many choices of statistical methods. These methods are typically described in internal technical memoranda which are rarely published or publically examined, and sometimes are invisible except for comments in source code. The IRAS Faint Source Survey Explanatory

Supplement (Moshir *et al.* 1992) offers a glimpse into this complex netherworld: a median filter is applied to reduce noise; outliers are detected to remove particle events; overlapping scans are combined and interpolated; fluxes are estimated with a trimmed mean; signal is extracted with a $S/N \geq 3.5$ criterion; distinct sources are devined by a complicated source merging procedure; sky positions are derived from recursive Kalman filtering and connected polynomial segmant fitting to satellite gyroscope time series data. The IRAS analysis benefits from robust statistcal procedures, such as the median and trimmed mean rather than the usual mean, which have been developed by statisticians over the past 20 years (*e.g.*, Hoaglin *et al.* 1983). The problems addressed here are specific to each instrument and survey, and general advice has limited value.

Reducing images to catalogues The analysis of astronomical images can be very complicated. In sparsely occupied images from photon-counting detectors (as in X-ray and gamma-ray astronomy), efforts concentrate on detecting sources above an uninteresting background. Methods include maximum likelihood analysis based on the Poisson distribution, matched filtering and Voronoi tesselations. In fully occupied grey-scale images, a wide variety of image restoration methods have been applied to deconvolve point spread functions and reduce noise: least squares fitting; Lucy-Richardson method; maximum entropy and other Bayesian methods; neural networks, Fourier and wavelet filtering (*e.g.*, Narayan & Nityananda 1986; Perley *et al.* 1989; Hanisch & White 1993; Starck & Murtagh 1994; Lahav *et al.* 1995). Many of these methods rest upon developments in statistical methodology.

Much work has also been directed to the automated analysis and classification of objects on images, particularly the discrimination of stars from galaxies on optical band photographic plates and CCD images. Each object is characterized by a number of properties (*e.g.*, moments of its spatial distribution, surface brightness, total brightness, concentration, assymetry), which are then passed through a supervised classification procedure. Methods include multivariate clustering, Bayesian decision theory, neural networks, k-means partitioning, CART (Classification and Regression Trees) and oblique decision trees, mathematical morphology and related multi-resolution methods (Bijaoui *et al.* 1997; White 1997). Such procedures are crucial to the creation of the largest astronomical databases with 1–2 billion objects derived from digitization of all-sky photographic surveys.

The scientific product of multi-wavlength surveys is frequently a large table with rows representing individual stars, galaxies, sources or locations and columns representing observed or inferred properties. Often a single survey effort will produce multi-wavelength results, as in the four infrared bands of IRAS, the five photometric colors of the Sloan Digital Sky Survey,

or spectral bands in the ROSAT All-Sky Survey. Analysis of such data is the domain of *multivariate analysis*. We therefore concentrate on multivariate statistical methodology in the following sections.

3. Fundamentals of Multivariate Analysis and Clustering

A multivariate analysis often begins with the computation of simple statistics of the sample: the mean and standard deviation of each variable; linear (Pearson's r) or rank (Spearman's ρ or Kendall's τ) correlation coefficients between pairs of variables. Statisticians often divide each value by the sample standard deviation for that variable (known as 'standardizing' or 'Studentizing' the sample), while astronomers often take a log transform or consider the ratio of two variables with the same units.

Study of pair-wise relationships between variables provides a valuable but fundamentally limited view of the data. A multivariate database should be viewed as a cloud of points (or vectors) in p-space which can have any form of structure, not just planar correlations parallel to the axes. The sample covariance patris S contains information for this more general approach, and lies at the root of many methods of multivariate analysis developed during the 1930–60s. The method most widely used in astronomy is *principal components analysis*. Here the 1st principal component is $e_1^T X$ where e_k is the eigenvector of S corresponding to the kth largest eigenvalue. This is equivalent to finding by the direction in p-space where the data are most elongated using least-squares to minimize the variance. The second component finds the elongation direction after the first component is removed, and so forth. Important applications in astronomy include the stellar spectral classification (Deeming 1964), eludication of Hubble's tuning-fork spiral galaxy classification system (Whitmore 1984), and characterization of relationships between emission lines, broad absorption lines and the continuum in quasar spectra (Francis *et al.* 1992).

In *canonical analysis*, the variables are divided into two preselected groups and the eigenvectors of the cross-sample covariance matrix $S_{11}^{-1/2} S_{12} S_{22}^{-1} S_{21} S_{11}^{-1/2}$ gives the principal linear relationships between the two sets of variables. This might be used to relate stellar metallicity variables with kinematic variables to study Galactic chemical evolution, or stellar magnetic activity indicators with bulk star properties to study dynamo theory.

A sample, collected from one or more multiwavelength surveys, often will not constitute a single type of astronomical object. Variance-covariance structure residing within the matrix S may thus reflect heterogeneity of the sample, rather than astrophysical processes within a homogeneous class. It is thus important to search for groupings in p-space using multivariate clustering or classification algorithms. Dozens of such methods have been

proposed. Unfortunately, most are procedural algorithms without formal statistics (*i.e.*, no probabilistic measures of merit) and there is little mathematical guidance which produces 'better' clusters.

Hierarchical clustering procedures produces small clusters within larger clusters. One such procedure, 'percolation' or the 'friends-of-friends' algorithm is a ffavorite among astronomers. It is called *single linkage clustering* obtained by successively removing the longest branches of the unique *minimal spanning tree* connecting the n points in p-space. Single linkage produces long stringy clusters. This may be appropriate for galaxy clustering studies, but researchers in other fields usually prefer *average or complete linkage* algorithms which produce more compact clusters. The many varieties of hierarchical clustering arise because the scientist must chose the metric (*e.g.*, should the 'distance' between objects be Euclidean or squared?), weighting (*e.g.*, how is the average location of a cluster defined?), and criteria for merging clusters (*e.g.*, should the total variance or internal group variance be minimized?).

An alternative method with a more rigorous mathematical foundation is *k-means partitioning.* It finds the combination of k groups that minimizes intragroup variance. However, it is necessary to specify k in advance.

4. Methodological Challenges from Astronomical Surveys

Many astronomical surveys are not amenable to traditional multivariate analysis and classification, and present serious needs for methodological advances by statisticians. Four major difficulties are outlined here.

First, fluxes or other measured quantities are subject to **heteroscedastic measurement errors with known variances**. That is, each variable of each object has an associated measurement of the variable uncertainty, and these uncertainties can differ for each object. Surprisingly, statistical methodology is very poorly developed for such situations. For instance, there is no clustering algorithm that weights points by their known measurement errors. Only the LISREL model of the multivariate linear regression problem can begin to treat known heteroscedastic measurement errors (Jöreskog & Sörbom 1989).

Second, objects may be undetected at one or many wavebands, leading to upper limits or **censored data** in one or many variables. A mature field of statistics known as survival analysis, developed principally for biomedical and industrial reliability applications, has been developed for censored datasets. A suite of survival methods is now widely used in astronomy (Feigelson 1992). However, most survival statistics apply only to univariate problems; Cox regression, the principal multivariate technique, permits censoring only in the single dependent variable. A more general partial cor-

relation coefficient based on Kendall's τ, which permits censoring in any or all variables, has recently been developed for astronomers (Akritas & Siebert 1996). But a full multivariate survival analysis is not yet available.

Third, astronomical surveys are virtually always suffer **truncation** in one or more variables due to sensitivity limits of the telescopes. This can create spurious structure in the variance-covariance matrix and makes the sample distribution a biased estimate of the underlying population. As with censoring, little statistical attention has been directed towards such datasets, except for linear regression problems in econometrics (Maddala 1983).

Fourth, following the traditions of celestial mechanicians of previous centuries, modern astronomers often seek to constrain **parameters of nonlinear astrophysical models**. Multivariate methodology was largely developed to assist social sciences and industry where such modeling does not arise. While least-squares regression techniques can be extended from linear to non-linear functions (*e.g.*, the orthogonal distance regression package ODRPACK), such methods fail in the presence of heteroscedastic measurement errors, censoring and truncation. Often the model is so complex, particularly if survey selection effects are included within it, that the results are available only through Monte Carlo simulation. A possible approach to such parameter estimation problems is through half-space projections (Babu & Feigelson 1996).

While these issues have yet to be adequately addressed by statisticians, some recent methodological advances can have significant benefits to astronomers. First, a number of approaches have emerged to facilitate both linear and nonlinear modeling of multivariate datasets. **Projection pursuit** regression uses local linear fits and sigmoidal smoothers to model nonlinear behavior (Huber 1985; Friedman 1987). **Multivariate Adaptive Regression Splines** (MARS) and a variety of similar methods fit the data with multidimensional splines (Friedman 1991). These methods are based on reasonable, but not unique, proceedures for parsimoniously choosing the number of parameters that avoid overfitting the data.

Second, astronomers can greatly benefit from visualization tools that permit powerful exploration of complex multivariate datasets. **XGobi** provides a 2-dimensional 'grand tour' of the database by displaying various projections of the data, with flexible interactive choice of variables, color brushing and projection pursuit options. **ExplorN**, operating on Silicon Graphics computers, gives a d-dimensional grand tour, saturation brushing and parallel coordinate plots. **XNavigator** travels through the database along local principal components.

Finally, we note that this brief paper omits many topics in multivariate statistics with potential importance for astronomy. These include non-

parametric methods, Bayesian approaches, outlier detection and robust methods, multicollinearity and ridge regression, goodness-of-fit measures, nonparametric density estimation, wavelet analysis, bootstrap resampling and cross-validation, mathamatical morphology, and many aspects of traditional multivariate analysis. The methodology for understanding multivariate databases is vast and constantly growing.

5. Astrostatistics References and Codes

Multivariate statistics are briefly reviewed in an astronomical context by Babu & Feigelson (1996), and are more thoroughly described (with FORTRAN codes) by Murtagh & Heck (1987). Many monographs presenting multivariate statistics are available, such as Johnson & Wichern (1992).

While commercial statistical packages are the most powerful tools for implementing statistical procedures, a cosiderable amount of software is in the public domain on the World Wide Web. An informative essay on statistical software by Wegman (1997) can be found at

`http://www.galaxy.gmu.edu/papers/astr1.html`.

Information on commercial statistical software packages such as SAS, SPSS and S-PLUS is available at

`http://www.stat.cornell.edu/compsites.html`.

Significant archives of on-line public domain statistical software reside at StatLib (`http://lib.stat.cmu.edu`) and the *Guide to Available Mathematical Software* (`http://gams.nist.gov`). StatLib provides many state-of-the-art codes useful to astronomers such as XGobi, ODRPACK, loess and MARS. Penn State operates the *Statistical Consulting Center for Astronomy* (`http://www.stat.psu.edu/scca`) for astronomers with statistical questions, and is initiating a site with links to statistical software on the Web (`http://www.astro.psu.edu/statcodes`).

Acknowledgements

This work was supported by NSF DMS 9626189, NASA NAGW-2120 and NAS 5-32669.

References

Akritas, M.G. and Siebert, J. (1996) Testing for partial association using Kendall's τ with censored astronomical data. Mon.Not.R.astron.Soc, in press.

Babu, G.J. and Feigelson, E.D. (1996) Astrostatistics. Chapman & Hall, London.

Babu, G.J. and Feigelson, E.D. (1997) Statistical Challenges in Modern Astronomy II. Springer-Verlag, New York.

Bijaoui. A., Rué, F and Savalle, R. (1997), in Statistical Challenges in Modern Astronomy II. Springer-Verlag, New York.

Deeming, T.J. (1964) Stellar spectral classification. I. Application of component analysis, Mon.Not.R.astron.Soc 127, 493.
Feigelson, E.D. (1992) Censoring in astronomical data due to nondetections (with discussion), in Statistical Challenges in Modern Astronomy, (E. D. Feigelson & G. J. Babu, eds.), Springer-Verlag, New York. p. 221.
Feigelson, E.D. and Babu, G.J. (1992) Statistical Challenges in Modern Astronomy. Springer-Verlag, New York.
Feigelson, E.D., Akritas, M., and Rosenberger, J. (1995) Statistical Consulting Center for Astronomy, in Astronomical Data Analysis Software and Systems IV (R.A. Shaw et al., eds.). Astron. Soc. Pacific, San Francisco.
Francis, P.J., Hewett, P.C., Foltz, C.B., and Chaffee, F.H., (1992) An objective classification scheme for QSO spectra, Astrophys.J 398, 476.
Friedman, J.H. (1987) Exploratory projection pursuit, J. Amer. Stat. Assn., 82, 239.
Friedman, J.H. (1991) Multivariate adaptive regression splines (with discussion), Annals of Statistics 19, 1.
Hanisch, R.J. and White, R.L. (eds.) (1993) The Restoration of HST Images and Spectra II, STScI: Baltimore.
Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (eds.) (1983) Understanding Robust and Exploratory Data Analysis. Wiley, New York.
Huber, P.J. (1985) Projection Pursuit, Annals of Statistics 13, 435.
Jaschek, C., and Murtagh, F. (eds.) (1990) Errors, Bias and Uncertainties in Astronomy. Cambridge University Press, Cambridge.
Johnson, R.A. and Wichern, D.W. (1992) Applied Multivariate Statistical Analysis, 3rd edition, NJ: Prentice Hall.
Jöreskog, K. G., and Sörbom, D., 1989, LISREL 7 A Guide to the Program and Applications, SPSS Inc., 444 N. Michigan Ave., Chicago IL 60611.
Lahav, O. *et al.* (1995) Galaxies, human eyes and artificial neural networks, Science 267, 859.
Maddala, G.S. (1983) Limited-dependent and quantitative variables in econometrics. Cambridge University Press, Cambridge.
Moshir, M. *et al.* (1992) IRAS Faint Source Survye Explanatory Supplement Version 2. IPAC, Pasadena CA.
Murtagh, F., and Heck, A. (1987) Multivariate Data Analysis. Reidel, Dordrecht Neth.
Narayan, R. and Nityananda, R. (1986) Maximum entropy image restoration in astronomy, Ann. Rev. Astro. Astrophys. 24, 127.
Perley, R.A., Schwab, F. and Bridle, A.H. (1989) Synthesis Imaging in Radio Astronomy. Astron. Soc. Pacific, San Francisco.
Press, W.H. *et al.* (1992) Numerical Recipes: The Art of Scientific Computing, Cambridge University Press, Cambridge.
Rolfe, E.J. (ed.) (1983) Statistical Methods in Astronomy. ESA SP 201, European Space Agency Scientific & Technical Publications. Noordwijk Neth.
Starck, J.-L. and Murtagh, F. (1994) Image restoration with noise suppression using the wavelet transform, Astron.Astrophys. 288, 342.
Stigler, S.M. (1986) The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, Cambridge.
Subba Rao, T. (ed.) (1997) Applications of Time Series Analysis in Astronomy and Meteorology. Chapman & Hall, London.
White, R.L. (1997) Object classification in astronomical images, in Statistical Challenges in Modern Astronomy II (G.J. Babu & E.D. Feigelson, eds.), Springer, New York.
Whitmore, B.C. (1984) An objective classification systems for spiral galaxies. I. The two dominant dimensions, Astrophys.J 278, 6.
Wegman, E.J., Carr, D.B., King, R.D., Miller, J.J., Poston, W.L., Solka, J.L. and Wallin, J. (1997) Statistical software, siftware and astronomy, in Statistical Challenges in Modern Astronomy II (G.J. Babu and E.D. Feigelson, eds.). Springer-Verlag, New York.