Statistical Methodology for the National Virtual Observatory

G. Jogesh Babu
Department of Statistics, 319 Thomas Building, Pennsylvania State University, University Park, PA 16802

E. D. Feigelson
Department of Astronomy & Astrophysics, 518 Davey Laboratory, Pennsylvania State University, University Park, PA 16802

Abstract. Large-scale NVO multi-wavelength surveys present a variety of challenging statistical problems. Multivariate clustering and analysis methods, such as principal components analysis, can be very helpful to astronomers and are briefly described. Newer statistical methods such as projection pursuit, multivariate splines, and visualization tools such as XGobi are 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

Large astronomical surveys from new high-throughput detectors and observatories are powerful motivators for more effective statistical techniques. Astronomical observatories now frequently generate gigabytes of information every day, with terabyte-size raw databases which produce reduced catalogs of \(10^6 - 10^9\) objects. The National Virtual Observatory may become an important institution where these databases and collected and used for astronomical study. These catalogs or raw databases, which may include up to dozens of observational properties of each object, often contain heterogeneous populations which must be isolated prior to detailed analysis.

Quite an amount of work has also been directed to the automated analysis and classification of objects or 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, asymmetry), 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 multi-resolution methods (White 1997; Bijaoui, Rué, & Savalle 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.

But after images are characterized and other preliminary reduction of raw data is performed, the intermediate scientific product of astronomical surveys is frequently a large table with rows representing individual stars, galaxies, sources or locations and columns representing observed or inferred properties. Statistical characterization of such databases is the domain of multivariate analysis. We therefore concentrate on multivariate statistical methodology in the following sections.

2. Multivariate Analysis and Clustering

A sample obtained from one or more multi-wavelength surveys often will not constitute a single type of astronomical object. A multivariate database should be viewed as vectors in $p$-space which can have any form of structure, not just planar correlations parallel to the axes. Apparent relationships between the variables may elucidate astrophysical processes, or may arise from heterogeneity of the sample. 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 statistical justification (i.e., no probabilistic measures of merit) and there is little mathematical guidance which produces ‘better’ clusters.

Hierarchical clustering methods produces small clusters within larger clusters. One such procedure, ‘percolation’ or the ‘friends-of-friends’ algorithm is a favorite among astronomers. In statistical parlance, it is called single linkage clustering and can be easily 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 due to the choice of 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 some mathematical foundation is $k$-means partitioning. It finds the combination of $k$ groups that minimizes intragroup variance. However, it is necessary to specify the number $k$ of groups in advance.

For each homogeneous class, the study of pairwise relationships between variables provides valuable insights into the data structure. The sample covariance matrix $S$ contains information for this 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 (PCA). A quick search of astronomical literature database yielded hundreds of papers in astronomy using PCA. The 1st principal component is $e_1^T X$ where $e_k$ is the eigenvector of $S$ corresponding to the $k^{th}$ 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.

PCA is also used in analyzing the covariance function of a timeseries data. However, caution should be exercised in using PCA. The two timeseries, the so-called Ramp Function and the Brownian Bridge (i.e., tied-down Brownian Motion, tied-down at 1), both have identical covariance function. So PCA cannot distinguish the data from these two timeseries. Ramp function $X$ is obtained by choosing a random point $\omega$ on the unit interval $[0,1]$ and defining

$$X(t) = \begin{cases} 
  t & \text{if } 0 \leq t \leq \omega \\
  t - 1 & \text{if } \omega \leq t \leq 1.
\end{cases}$$

Note that the stochastic process generated by the Ramp function is highly regulated, while the Brownian Bridge is a white noise Gaussian process. Without additional model assumptions such as Gaussian structure PCA is not of much use. PCA also fails for non-linear relations (e.g., $X$ Gaussian and $Y = X^2$).

3. Application to Gamma-ray Bursts Data

As only a few gamma-ray burst (GRB) sources have astronomical counterparts at other wavebands, empirical studies of GRBs have been largely restricted to the analysis of their bulk properties such as fluence and spectral hardness, and evolution of these properties within a burst event. While bursts exhibit a vast range of complex temporal behaviors, their bulk properties appear simpler and amenable to straightforward statistical analyses. Examination of whether GRB bulk properties comprise a homogeneous population or are divided into distinct classes may lead to astrophysical insights of their origins.

It can be dangerous to look for correlations prior to classifying (or establishing the homogeneity of) the population. While the anticorrelation between hardness ratio and burst duration seen in full samples may be the manifestation of a single astrophysical process, it may alternatively reflect differences between distinct processes. Most multivariate analyses thus begin with a study of homogeneity and classification, and then investigate the variance-covariance structure (i.e., correlations) within each class. We recently performed such an analysis (Mukherjee et al. 1998)

Two multivariate clustering procedures are used on a sample of 797 bursts from the Third BATSE Catalog: a nonparametric average linkage hierarchical agglomerative clustering procedure validated with Wilks’ $\Lambda^*$ and other MANOVA tests; and a parametric maximum likelihood model-based clustering procedure assuming multinormal populations calculated with the EM Algorithm and validated with the Bayesian Information Criterion. The two methods yield very similar results. We find the BATSE GRB population consists of three classes with the following Duration/Fluence/Spectrum bulk properties: Class I with long/bright/soft bursts, Class II with short/faint/hard bursts, and Class III with intermediate/intermediate/soft bursts. Class IV consists of a single point later found to have erroneous processing.

These are frames from the ‘grand tour’ movie of the 5-dimensional dataset provided by the XGobi software where each cluster is ‘brushed’ with a different symbol.
Figure 1. Two snapshots from the XGobi grand tour of the 5-dimensional database with bursts brushed according to the nonparametric Average Linkage clustering results: Class I (●), Class II (×), Class III (□) and Class IV (○).

4. New Methodology Needed for Astronomy

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

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 not well developed for such situations. For example, there does not seem to be a clustering algorithm that weights points by their known measurement errors. Astronomers also need density estimation, k-sample goodness-of-fit tests, spatial point processes, time series and multivariate analyses for such datasets.

Some objects may be undetected at one or many wavebands, leading to upper limits or censoring in one or many variables. A branch of statistics known as survival analysis, used 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. But a full multivariate survival analysis is not yet available.

Astronomical surveys nearly always suffer truncation in one or more variables due to sensitivity limits of the telescopes. Left-truncation is frequently
present in any variable because faint objects are undetected. This can create
spurious structure in the variance-covariance matrix and makes the sample dis-
tribution a biased estimate of the underlying population. As with censoring,
little statistical attention has been directed towards such multivariate datasets,
except for linear regression problems in econometrics (Maddala 1983).

Finally, following the traditions of celestial mechanics of previous centuries,
modern astronomers often seek exploratory structural regression relations be-
tween variables, and wish to constrain parameters of nonlinear astrophysical
models. Multivariate methodology was largely developed to assist social sci-
ences and industry where such modeling does not arise. Classical regression
techniques 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.

While these issues have yet to be adequately addressed by statisticians,
some recent advances can have significant benefits to astronomers. 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 meth-
ods fit the data with multidimensional splines (Friedman 1991). These methods
are based on reasonable, but not unique, procedures for sparsely choosing the
number of parameters that avoid overfitting the data.

Further, astronomers can greatly benefit from visualization tools that per-
mit 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 pro-
jection pursuit options. ExplorN, ViSta, CViz, IVEE and other packages give
d-dimensional grand tour, saturation brushing, parallel coordinate plots and
other functionalities.

Finally, we note that this brief paper omits many topics in statistics with
potential importance for astronomy. These include nonparametric methods,
Bayesian approaches, wavelet analysis, bootstrap resampling, and many aspects
of traditional multivariate analysis. The statistical methodology for understand-
ing databases is vast and constantly growing.

5. Astrostatistics References and Codes

Many statistical methods are briefly reviewed in an astronomical context by
Babu & Feigelson (1996), and multivariate methods are more thoroughly de-
scribed (with Fortran codes) by Murtagh & Heck (1987). The conference pro-
ceedings by Feigelson & Babu (1992), Babu & Feigelson (1997) also have several
articles containing use of multivariate analysis in astronomy. The monographs
by Ripley (1996) and by Johnson & Wichern (1992) would be useful resources
for astronomers.

While commercial statistical packages are the most powerful tools for imple-
menting statistical procedures, a considerable amount of software is in the public
domain on the World Wide Web. An informative essay on statistical software
was written by Wegman (1997). Information on commercial statistical software packages such as SAS, SPSS and S-PLUS is available on-line. Significant archives of on-line public domain statistical software reside at StatLib and the Guide to Available Mathematical Software. StatLib provides many state-of-the-art codes useful to astronomers such as XGobi, loess and MARS. Finally, we maintain a Web metasite called StatCodes that gives links to on-line source codes and packages of multivariate, clustering, visualization and other statistical methods for astronomical research.

References

Murtagh, F., & Heck, A. 1987, Multivariate Data Analysis (Dordrecht: Reidel)
Ripley, B. D. 1996, Pattern Recognition and Neural Networks (Cambridge: Cambridge University Press)

http://www.galaxy.gmu.edu/papers/str1.html
http://www.stat.cornell.edu/compsites.html
http://lib.stat.cmu.edu
http://gams.nist.gov
http://www.astro.psu.edu/statcodes