Astrostatistics: The Final Frontier

Peter Freeman, Joseph Richards, Chad Schafer, and Ann Lee

How did the universe form? What is it made of, and how do its constituents evolve? How old is it? And, will it continue to expand? These are questions cosmologists have long sought to answer by comparing data from myriad astronomical objects to theories of the universe's formation and evolution. But, until recently, cosmology was a data-starved science. For instance, before the 1990 launch of the Hubble Space Telescope, the Hubble constant—a number representing the current expansion rate of the universe—could only be inferred to within a factor of two, and cosmologists had to make do performing simple statistical analyses. Since that time, technological advances have led to a flood of new data, ushering in the era of precision cosmology. (The Sloan Digital Sky Survey alone has collected basic data for more than 200 million objects.)

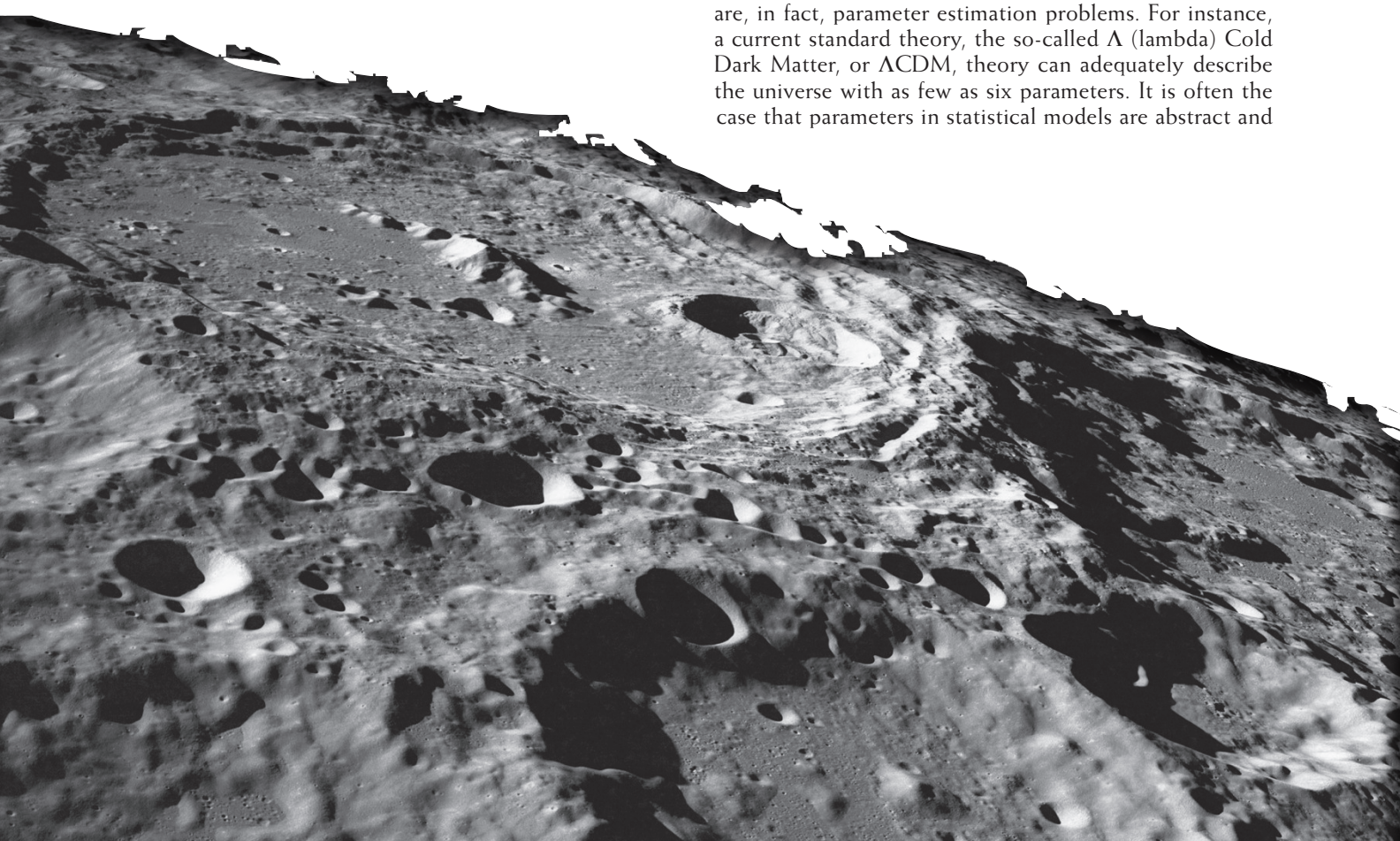
To help make sense of all this data, cosmologists have increasingly turned to statisticians, and a new interdisciplinary field has arisen: astrostatistics. Work in astrostatistics uses a wide range of statistical methods, but there are a few particular statistical issues that are prevalent. Here, we focus on two broad challenges: parameter estimation using complex models and data analysis using noisy, nonstandard data types.

Parameter Estimation Using Type Ia Supernovae

A supernova is a violent explosion of a star whose brightness, for a short time, rivals that of the galaxy in which it occurs. There are several classes of supernovae (or SNe). Those dubbed Type Ia result from runaway thermonuclear reactions that unbind white dwarfs, Earth-sized remnants of stars such as our sun. In the 1990s, astronomers used observations of Type Ia SNe to infer the presence of dark energy, a still unknown source of negative pressure that acts to accelerate the expansion of the universe, rather than slowing the expansion, as does normal baryonic matter (e.g., protons and neutrons).

There are many interesting analyses we can carry out with Type Ia SNe data. For instance, we can construct procedures to test different models of the evolution of dark energy properties as a function of time. Or, we may adopt a model for dark energy and see how SNe data, in concert with that model, constrain basic cosmological parameters. We demonstrate the latter analysis here.

One may be surprised to learn that theories regarding the formation and evolution of the universe are sufficiently developed that some of the biggest questions in cosmology are, in fact, parameter estimation problems. For instance, a current standard theory, the so-called Λ (lambda) Cold Dark Matter, or ΛCDM, theory can adequately describe the universe with as few as six parameters. It is often the case that parameters in statistical models are abstract and



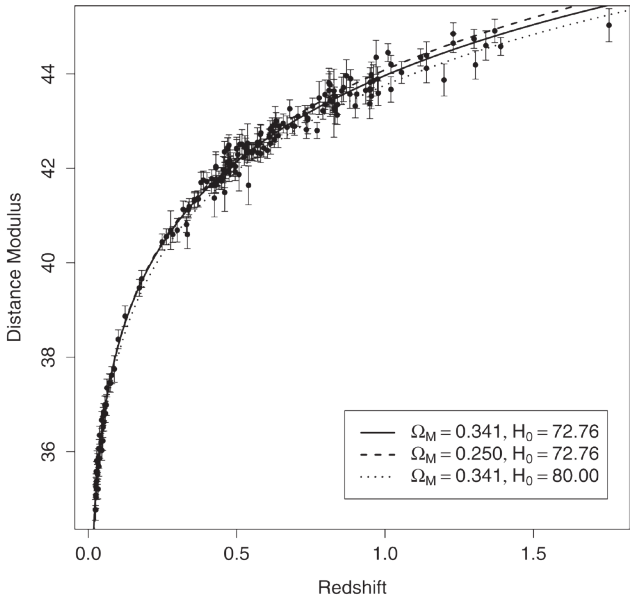


Figure 1. Distance modulus (m) versus redshift (z) for Type Ia supernovae. W_m is the fraction of critical energy density contributed by baryonic matter and dark matter, while H_0 is the Hubble constant. Curves represent theoretical predictions for three parameter values.

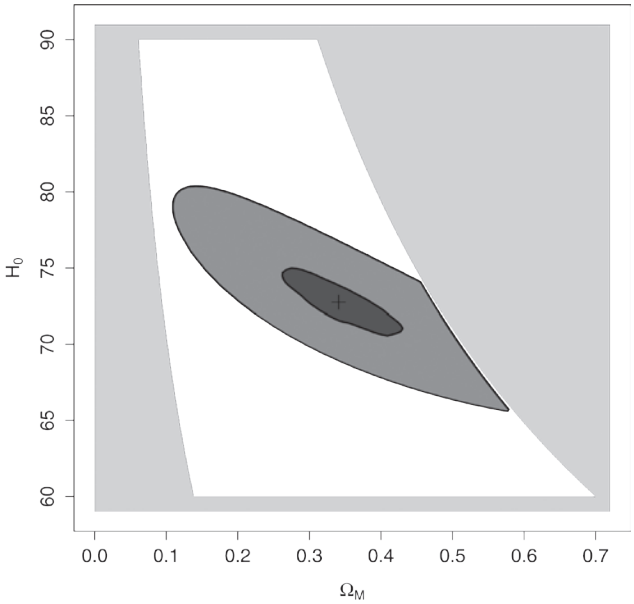


Figure 2. Ninety-five percent joint confidence region for Ω_m and H_0 using SNe data. The larger region is based on performing a χ^2 test at each possible parameter combination. The smaller region is created using a procedure that attempts to optimize the precision of the estimate. Values of (Ω_m H_0) outside the white area are considered implausible based on other observations.

lack intrinsic significance, such as the slope and intercept parameters (β_0 and β_1) in simple linear regression (as it is usually the estimate of the line that is of interest). But this is not the case with cosmological model parameters; the goal is to precisely estimate each of these fundamental constants.

Different types of cosmological data help constrain different sets of cosmological parameters. We can use the data of Type Ia SNe in particular to make inferences about two of them: Ω_m and H_0. The former (pronounced "omega m") is the fraction of the critical energy density (i.e., the amount necessary to make the universe spatially flat) contributed by baryonic matter and dark matter whose presence is inferred by its influence upon baryons, but whose make-up is still unknown. The latter ("h naught") is the aforementioned Hubble constant (and not an indication of a null hypothesis), the current value of the Hubble parameter $H(t)$ that describes the rate at which the universe expands as a function of time.

What truly makes Type Ia SNe special from a data analysis standpoint is that theory holds them to be standard candles: Two SNe at the same distance from us will appear equally bright, so that distance and brightness are monotonically related. To estimate the distance to a supernova, astronomers use its redshift, z, which is a directly observable quantity representing the relative amount by which the universe has expanded since the explosion occurred (which can be up to billions of years ago).

Redshifts are measured by examining the distance between peaks and troughs in light waves, which expand as the universe does. We observe photons emitted from a supernova with wavelength λ_{emit} to have a longer wavelength $\lambda_{obs} = (1+z)\,\lambda_{emit}$. This motivates the term "redshift": visible light emitted from SNe of progressively higher redshift appears progressively redder to us, as red light is at the long-wavelength end of the spectrum of visible light.

Another measure of astronomical distance is the distance modulus, μ, a logarithmic measure of the difference between a supernova's observed brightness and its intrinsic luminosity, which uses the fact that objects farther away will appear fainter. In Figure 1, we show measurements of z and μ for a sample of Type Ia SNe. Measurement errors in z are small ($\lesssim$1%) and are not shown, while estimates of measurement error in μ are given by the vertical bars. Note the deviation from linearity in the trend in z versus μ is due to the accelerated expansion of the universe, which is attributed to dark energy.

Assuming a particular cosmological model, the observables μ and z are linked via a function of the cosmological parameters (Ω_m, H_0):

$$\mu(z|\Omega_m, H_0) = 5\log_{10}\left(\frac{c(1+z)}{H_0}\int_0^z \frac{du}{\sqrt{\Omega_m(1+u)^3+(1-\Omega_m)}}\right)+25 \quad (1)$$

Other theoretical assumptions about the structure of the universe would suggest the use of other functions. The particular function in equation (1) represents a spatially flat universe, where the actual energy density of the universe exactly equals the critical density, separating open universes that expand forever from closed universes that first expand, then contract back to a point. It also contains a contribution from dark energy in the form of a so-called cosmological constant (i.e., the dark energy has a particular form that does not evolve with time); its fractional contribution to the critical energy density is $\Omega_\Lambda \equiv 1-\Omega_m$.

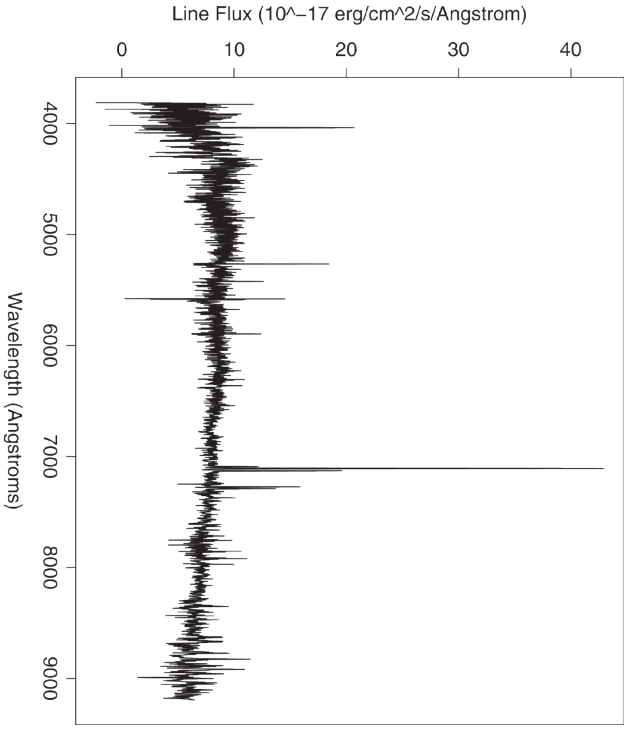


Figure 3. Flux versus wavelength for a typical Sloan Digital Sky Survey (SDSS) galaxy spectrum

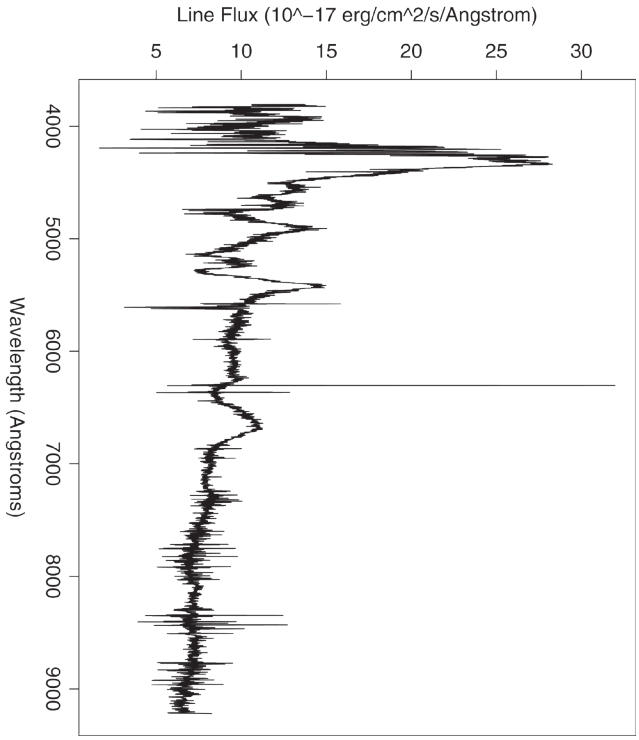


Figure 4. Flux versus wavelength for a typical Sloan Digital Sky Survey (SDSS) quasar spectrum

The cosmological constant was first used by Albert Einstein to make the universe spatially flat and unchanging within the context of his theory of general relativity. Later, after Hubble demonstrated the universe was expanding, Einstein disowned the constant, referring to it as his "biggest blunder."

According to the model, the observed pairs (z_i, Y_i) are realizations of $Y_i = \mu(z_i | \Omega_m, H_0) + \sigma_i \varepsilon_i$, where the ε_i are independent and identically distributed standard normal. The standard deviations σ_i are estimated from properties of the observing instrument, but are generally taken as known. Thus, there is a simple way of constructing a joint confidence region for Ω_m and H_0 by performing a χ^2 test for each possible pair (Ω_m, H_0): using the fact that

$$\sum_{i=1}^{182} \left(\frac{Y_i - \mu(z_i | \Omega_m, H_0)}{\sigma_i} \right)^2$$

has the χ^2 distribution with 182 degrees of freedom when the parameters are correctly specified. Figure 2 shows the 95% confidence region that results from this procedure.

Cosmologists are fond of χ^2-type statistics because of their intuitive appeal, but the resulting confidence regions are needlessly large (imprecise). Ongoing work focuses on tightening the bounds on parameters by improving the statistical procedures.

Figure 2 also depicts a confidence region created using a Monte Carlo–based technique developed by Chad Schafer and Philip Stark that approximates optimally precise confidence regions. These parameters also appear in stochastic models describing other cosmological data sets, and joint analyses of these data sets will lead to yet more precise estimates of the unknown parameters.

Mining Spectra for Cosmological Information

Another fundamental challenge in astrostatistics is the extraction of useful information from a large amount of complex data. One example of such data are spectra, measures of photon emission as a function of time, energy, wavelength, etc. They may consist of thousands of measurements, such as the examples of galaxy and quasar data collected by the Sloan Digital Sky Survey (SDSS) that we show in Figures 3 and 4, respectively.

Typical galaxies (e.g., the Milky Way) are agglomerations of billions of stars; whereas, quasars are galaxies going through an evolutionary phase in which a supermassive black hole at the center is actively gobbling up gas and emitting so much light that it effectively drowns out the rest of the galaxy. To probe the physical conditions of galaxies and quasars, we might analyze the global, (nearly) smooth continuum emission and/or the narrow and broad spikes that rise above the continuum (emission lines) or dip below the continuum (absorption lines). Working with entire spectra can be computationally tedious, and working with large groups of spectra even more so. SDSS has so far measured spectra for approximately 800,000 galaxies and 100,000 quasars. It may be that we can construct relatively simple statistics that convey nearly as much information as each spectrum, itself. For instance, to perform galaxy classification, astronomers have conventionally used measurements of ratios of photon emission over particular (small) ranges of wavelengths.

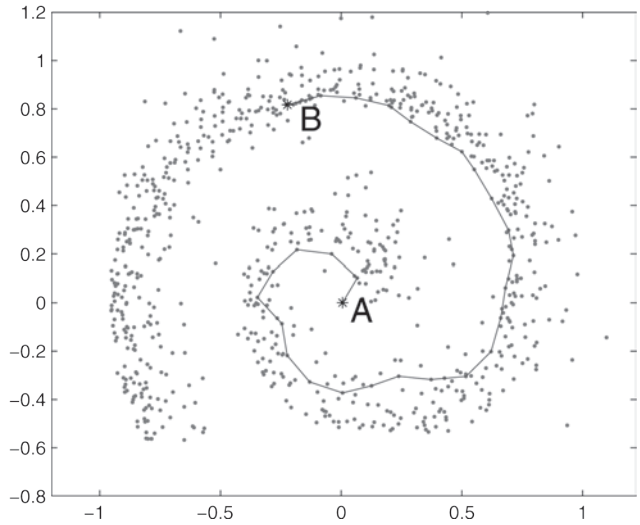


Figure 5. An example of a one-dimensional manifold embedded in two dimensions. The path from A to B is representative of the diffusion distance between A and B, and is a better representation of dissimilarity between them than the Euclidean distance.

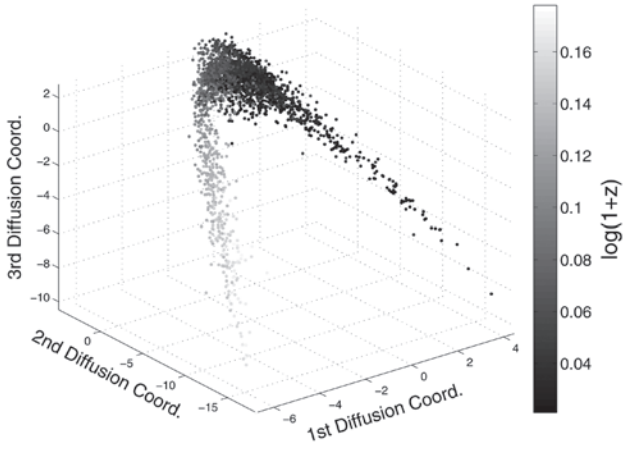


Figure 6. Embedding of a sample of 2,796 SDSS galaxy spectra using the first three diffusion map coordinates, with color representing galaxy redshifts.

We have begun work on the challenge of finding a low-dimensional representation of spectra that retains most of the useful information they encode. The first idea most would think of is to perform a principal components analysis (PCA) of the spectra. The basic idea is that, ideally, each spectrum could be written as the weighted combination of m basis functions, where m is much smaller than p, the length of each spectrum. A spectrum is then represented by the projections onto these m basis vectors; the union of the projections of several spectra form a lower-dimensional embedding of the data often referred to as a principal component (PC) map. PCA was applied to a family of 1,200 SDSS spectra, with 600 from galaxies and 600 from quasars.

PCA, however, can do a poor job for astronomical data. There are two main reasons for this. First, the method is only able to pick out global features in the spectra. It ignores that emission line features spanning a small range of wavelengths can be crucial in, for example, determining whether a spectrum belongs to a galaxy or a quasar. Second, the method is linear. If the data points (the spectra) lie on a linear hyperplane, the method works well, but if the variations in the spectra are more complex, one would be better off with nonlinear dimension reduction methods.

The first issue—local features—has been addressed by astrostatisticians to some extent. Techniques such as wavelets are useful to describe inhomogeneous features in spectra, and these methods are gaining popularity. The second issue—nonlinearity—has largely been ignored in the field of astronomy, however.

In "Exploiting Low-Dimensional Structure in Astronomical Spectra," a paper submitted to *The Astrophysical Journal*, the authors applied a nonlinear approach for dimensionality reduction, embedding the observed spectra within a diffusion map. The basic idea is captured in Figure 5. Ideally, we would like to find a distance metric that measures the distance between points A and B along the spiral direction, and then construct a coordinate system that captures the underlying geometry of the data. Diffusion distances and diffusion maps do exactly this by a clever definition of connectivity and the use of Markov chains.

Imagine a random walk starting at point A that is only allowed to take steps to immediately adjacent points. Start a similar walk from point B. For a fixed scale t, the points A and B are said to be close if the conditional distributions after t steps in the random walk are similar. The diffusion distance between these two points is defined as the difference between these two conditional distributions. The distance will be small if A and B are connected by many short paths through the data.

This construction of a distance measure is also robust to noise and outliers because it simultaneously accounts for all paths between the data points. The path from A to B depicted in Figure 5 is representative of the diffusion distance between A and B and is a better description of the dissimilarity between A and B than, for example, the Euclidean distance from A to B.

In applying this technique for dimensionality reduction, the data set attribute we wish to preserve is the diffusion distance between all points. For the example in Figure 5, a diffusion map onto one dimension ($m=1$) approximately recovers the arc length parameter of the spiral. A one-dimensional PC map, on the other hand, simply projects all the data onto a straight line through the origin. Therefore, the diffusion map technique for dimensionality reduction will be better suited to the analysis of astronomical data, which is often complex, nonlinear, and noisy.

Figure 6 shows the three-dimensional diffusion map constructed from 2,796 SDSS galaxy spectra. Each of these spectra is of length 3500 (i.e., each lies in a 3500-dimensional space) and possesses the sort of noisy, irregular structure seen in Figure 3. Despite the high dimension and noisy data, the presence of a low-dimensional, nonlinear structure is clear. More significantly, this structure can be related to important

properties of the galaxies. The colors of the points in the map give the redshifts of the galaxies. It is evident that using these coordinates, one could predict redshift. Results of this type hold great promise for future exploration of these complicated, high-dimensional data sets.

The Future

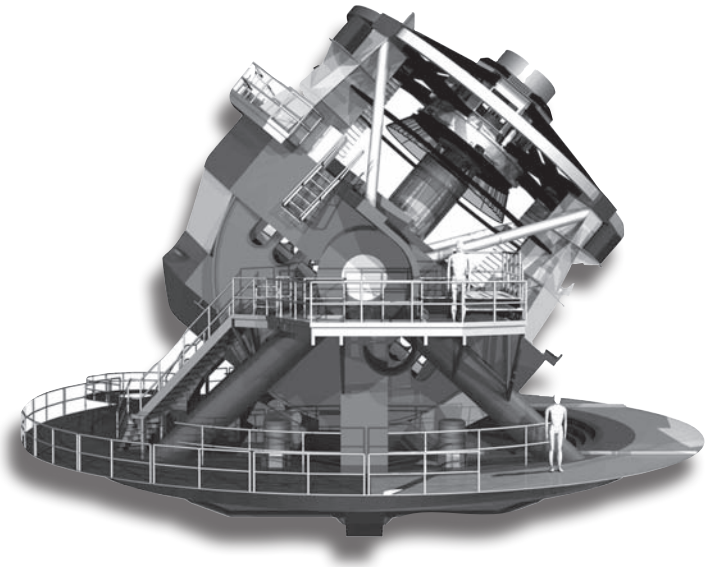
The SDSS has provided the astronomical community with a flood of data, but this flood is minuscule compared to that which will be created by the Large Synoptic Survey Telescope, or LSST. Scheduled to begin observations in 2015, it will repeatedly observe half the sky from its perch on Cerro Pachon in Chile, producing wide-field (3o diameter) snapshots. (SDSS, in contrast, has scanned approximately 20% of the sky without repeated observations.) Repetition will allow LSST to not only probe the nature of dark energy and create unprecedented maps of both the solar system and Milky Way, but also to track how astronomical objects change with time.

Just how much data will LSST collect? Recent projections indicate it will gather 30 terabytes (30×1000^4 bytes) of data per evening of viewing, which is roughly equal to the amount of data collected by SDSS over its entire lifetime. The total amount of LSST data is expected to exceed 10 petabytes (10×1000^5 bytes), and will include observations of 5 billion galaxies and 10 billion stars.

Needless to say, analyses of these data pose interesting challenges that will require the help of statisticians. Two of these challenges are obvious. The first is how to efficiently process the data in a way so as to not discard their important features. The second is how to test ever more sophisticated theories whose development will be motivated by LSST's high-resolution data.

But, there is a third challenge—not necessarily obvious to astronomers—that is likely to arise. As stated, the LSST data will increase the already growing pressure on theoreticians to refine their models. But, could astronomers be getting too much of a good thing? Introductory statistics courses often stress the notion of "practical versus statistical significance" when testing a null hypothesis. We are well aware that with enough data, any hypothesis can be rejected at any reasonable significance level. This will likely occur with the massive influx of cosmological data. The theories, which inevitably contained some level of approximation, will appear to fit the data poorly when subjected to formal statistical testing.

Thus, we envision a growing need for formal model-testing and model-selection tools. We also see an increasing interest in nonparametric and semiparametric approaches, which are useful for finding relationships in the data when we lack a physically motivated, fully parametric model or when such a model is complex and a computationally simpler approach may have similar inferential power. For example, there is no physically motivated model encapsulating how dark energy properties evolve with time. In a paper submitted to *Annals of Applied Statistics*, "Inference for the Dark Energy Equation of State Using Type Ia Supernova Data," the authors assess different nonparametric models of dark energy, fitting them to the Type Ia SNe data described in Figure 1. Importantly, they show how functional properties such as concavity and monotonicity can help sharpen statistical inference. Currently, the number of SNe is insufficient to rule out all but the most extreme dark energy models. These authors show how a factor of 10 increase in the number of SNe, something readily achievable with LSST, will allow us to determine the veracity of many hypotheses, including whether a dark energy model based on Einstein's cosmological constant model is consistent with the data.



Large Synoptic Survey Telescope (LSST), from its perch on Cerro Pachon, Chie can create unprecedented maps of both the solar system and Milky Way.

Design of LSST Telescope
Courtesy of LSST Corporation

This is typical of the outstanding challenges facing statisticians working on inference problems in cosmology and astronomy. Novel methods of data analysis that fully use available computing resources are needed if key questions are to be answered using the soon-to-arrive massive amount of data.

Further Reading

The Center for Astrostatistics, *http://astrostatistics.psu.edu*

Freedman, W. (1992) "The Expansion Rate and Size of the Universe." *Scientific American*, 267:54.

Genovese, C. R.; Freeman, P.; Wasserman, L.; Nichol, R.C.; and Miller, C. (2008) "Inference for the Dark Energy Equation of State Using Type Ia Supernova Data." Submitted to *Annals of Applied Statistics*, *http://arxiv.org/abs/0805.4136*.

The International Computational Astrostatistics (InCA) Group, *www.incagroup.org*

The Large Synoptic Survey Telescope (LSST), *www.lsst.org*

Richards, J.; Freeman, P.; Lee, A.; and Schafer, C. (2008) "Exploiting Low-Dimensional Structure in Astronomical Spectra." Carnegie Mellon University Department of Statistics Technical Report #863. Submitted to *The Astrophysical Journal*, *http://arxiv.org/abs/0807.2900*.

Schafer, C. and Stark, P. (2007) "Constructing Confidence Regions of Optimal Expected Size." Carnegie Mellon University Department of Statistics Technical Report #836. Submitted to the *Journal of the American Statistical Association*.

The Sloan Digital Sky Survey (SDSS), *www.sdss.org*