A Framework for Statistical Inference in Astrophysics

Chad M. Schafer

Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; email: cschafer@cmu.edu

Annu. Rev. Stat. Appl. 2015. 2:141–62

First published online as a Review in Advance on December 17, 2014

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-022513-115538

Keywords

astrophysics, complex models, complex data, parameter estimation, measurement error

Abstract

The rapid growth of astronomical data sets, coupled with the complexity of the questions scientists seek to answer with these data, creates an increasing need for the utilization of advanced statistical inference methods in astrophysics. Here, focus is placed on situations in which the underlying objective is the estimation of cosmological parameters, the key physical constants that characterize the Universe. Owing to the complex relationship between these parameters and the observable data, this broad inference goal is best divided into three stages. The primary objective of this article is to describe these stages and thus place into a coherent framework the class of inference problems commonly encountered by those working in this field. Examples of such inference challenges are presented.

1. INTRODUCTION

Statistical inference plays a significant role in modern astrophysics research. Current astronomical surveys take rich, multidimensional measurements of millions of stars, galaxies, and other celestial objects, and future surveys will push measurement counts into the billions. Concurrent with (and because of) this flood of information, the scientific questions that one seeks to answer with these data are growing in complexity. This combination of massive, complex data sets and challenging inference problems creates an opportunity for statisticians to make deep contributions of significant scientific interest.

The growth in the number of observed quasars—the luminous, active centers of some massive, distant galaxies—provides an illustrative example. Early discoveries (Matthews & Sandage 1963) led to careful study of individual objects in order to understand the nature of these phenomena. As methods for quasar detection improved, and as the sample size grew, focus shifted from understanding individual objects to inferring properties of the population of quasars, most importantly estimation of the distribution of quasar magnitude (intrinsic brightness), the quasar luminosity function (QLF). Throughout the 1980s and 1990s, researchers used progressively larger quasar catalogs to constrain the QLF: Marshall et al. (1984) used 35 quasars, Boyle et al. (1988) used 420, and Pei (1995) used 1,200. The Sloan Digital Sky Survey (SDSS) (Eisenstein et al. 2011) marked a key moment in the study of quasars. The first SDSS quasar catalog (Schneider et al. 2002) contained 3,814 quasars, but this number quickly grew to over 46,000, the number Richards et al. (2006) used to estimate the QLF. The most recent SDSS quasar catalog contains over 166,000 objects with high-quality spectroscopic measurements (Pâris et al. 2014), and the Million Quasars (MILLIQUAS) Catalog (**http://heasarc.gsfc.nasa.gov/W3Browse/all/milliquas.html**) combines these with lower-quality SDSS quasars and those from other surveys to create a catalog of over a million objects. Larger samples enable more precise estimation of the QLF, but such estimation requires novel, sophisticated statistical methods that account for sample selection effects and measurement error.

The above example is typical of the evolution seen in many areas of study in astrophysics. A drive to improve our understanding of the Universe and its evolution has motivated a shift from the study of individual objects to that of population-targeted inference. Different theories predict different forms for the QLF, so comparing the best estimate of it with these predictions is crucial. Fortunately, owing to the deep understanding of the Universe developed by theoretical astrophysicists, many key questions have been largely reduced to the estimation of physical constants often referred to as cosmological parameters.

The broad statistical inference goal is thus often the estimation of these cosmological parameters using observations of astronomical objects. This article presents a three-step framework that supports typical inference problems of this form. In the first stage, one must estimate the properties of individual objects from the raw observed data. In the second stage, this catalog of measurements is used to estimate a relevant, population-level summary of these objects. (The estimation of the QLF occurs at this second stage.) Finally, in the third stage, the cosmological parameters are inferred using one or more of these summaries. Thus, as with most statistical inference problems, estimation in astrophysics is built upon parameters, data, and the probability models that relate them. Although a wide range of statistical methods are employed at these three steps, there are recurrent challenges in observation-based estimation of cosmological parameters, including dealing with the multidimensional, structured nature of the observed data; the biases that result from our limited, noisy view of the Universe; and the complexity of the relationships between the cosmological parameters and the probability distributions that model the summary statistics. Instead of attempting an exhaustive exploration

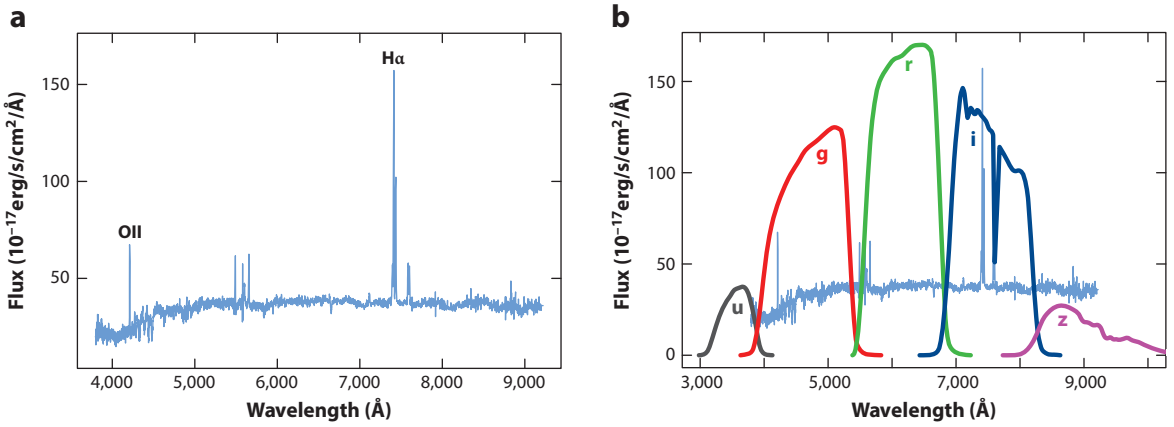

Figure 1

(*a*) Galaxy spectrum as measured by the Sloan Digital Sky Survey (SDSS). The large Hα line results from the abundance of hydrogen in galaxies, and the significant OII line is a signature of active star formation within a galaxy (Gilbank 2010). (*b*) Same spectrum as in panel *a*, with the filter response functions from the SDSS ugriz filter system superimposed. u, g, r, i, and z are standard names attached to the filters.

of the topic, this review focuses on these broad challenges in the context of some important examples.

2. RAW OBSERVABLES

In their raw state, most modern astronomical data take the form of intensity measurements at different wavelengths on the electromagnetic spectrum or possibly through different filters that accumulate these intensities over a range of wavelengths. For example, **Figure 1*a*** shows the emission spectrum of a galaxy as measured by the SDSS. By its tenth data release in 2013, this unprecedented astronomical survey had measured the spectra of roughly 3.3 million astronomical objects, including 1.8 million galaxies (Ahn et al. 2013). An individual spectrum is characterized by a smooth continuum with prominent emission and absorption lines. **Figure 1** highlights two prominent emission lines: The large Hα line results from the abundance of hydrogen in galaxies, and the significant OII line is a signature of active star formation within a galaxy (Gilbank et al. 2010). These are just two examples of how spectrum features encode valuable information regarding an object.

Measuring a spectrum requires targeting with a spectrograph. As this operation is relatively time consuming, spectra are measured for only some of the astronomical objects detected during an initial imaging survey. An imaging survey consists of taking photographs of the sky using charge-coupled devices (CCDs) through a small number of filters, often called the bands. For example, the SDSS uses 5 bands and an imaging camera with a 5×6 arrangement of 30 CCDs in which each row employs the same filter (Gunn et al. 1998). The light intensity (called the flux) of the object as measured through this filter system can be thought of as measuring a coarse version of the full spectrum. (The flux measurements are typically transformed into either an apparent or absolute magnitude; see the sidebar titled Intensity Units.) In particular, instead of measuring the full spectrum, one observes this spectrum convolved with a collection of response functions. **Figure 1*b*** displays the filter response functions for the SDSS; this camera uses a standard ugriz filter system.

INTENSITY UNITS

The intensity of an astronomical object in a particular band is often expressed as its apparent magnitude, usually denoted m. Apparent magnitude is calculated relative to a standard, so the difference in apparent magnitudes between any two objects is

$$m_1 - m_2 = -2.5 \log_{10}(F_1/F_2),$$

where F denotes the flux. The absolute magnitude, denoted M, is the apparent magnitude of an object as viewed at a distance of 10 parsecs, and is calculated as

$$M = m - 5 \log_{10}(d/10),$$

where d is the distance to the object measured in parsecs. Absolute magnitudes have the advantage of placing magnitudes on a standard scale that is not tied to the position of Earth. This comes at the cost, however, of needing to estimate the distance d. A related quantity is the luminosity (L), which represents the amount of energy the object emits per second and is given by $L = (4\pi d^2)F$. The bolometric luminosity is the luminosity measured over all wavelengths. The difference between the apparent and absolute magnitudes, called the distance modulus, is a commonly used distance metric in cosmology (Sparke & Gallagher 2007).

The above example establishes two standard types of data available for astronomical objects: spectroscopic and photometric. Spectroscopic data (those consisting of spectra) are rich in scientific information but less abundant. In contrast, photometric data (magnitude measurements) are of much lower resolution, but they are measured easily and in greater quantity. Future surveys will focus more heavily on photometric measurements, owing to both improved CCD technology and advances in statistical methods that can utilize these lower-resolution data. For example, the Large Synoptic Survey Telescope (LSST) currently under construction in Chile will gather photometric data in six bands, and the quantity of data gathered will dwarf that of SDSS. This instrument will generate a catalog of photometric data on approximately 10 billion galaxies and an additional 10 billion stars (Zhang et al. 2013).

The LSST will capture 15 terabytes (TB) of images of the Universe on a nightly basis, including repeated visits to the same regions, thereby enabling exploration of the time variability of astronomical phenomena to an extent that has not previously been possible. For example, supernovae, stars in the fiery final stage of their lives, can be studied only with repeated viewing of the same area of the sky over a period of at least 30 days. Changes in the magnitude of an object over time are summarized in a light curve; **Figure 2*a*** shows an example of the light curves measured from a single supernova (Jha et al. 2006). Currently, roughly 1,000 such supernovae are available for study, but the LSST will push this count into the millions (LSST Science Collaboration et al. 2009).

The cameras used in these imaging surveys are of sufficiently high resolution that they also create catalogs of images of individual objects, thereby generating an additional source of scientific information. For example, the Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS) (Grogin et al. 2011, Koekemoer et al. 2011) utilized the Hubble Space Telescope to build a catalog of over 250,000 galaxy images, one of which is shown in **Figure 2*b***. This image depicts the intensity of the galaxy as measured through a single filter. As is the case with spectra and light curves, this high-dimensional image contains significant low-dimensional structure that is related to properties of the depicted object. Thus, an initial step in utilizing these raw data is to exploit this structure to estimate object-specific parameters.

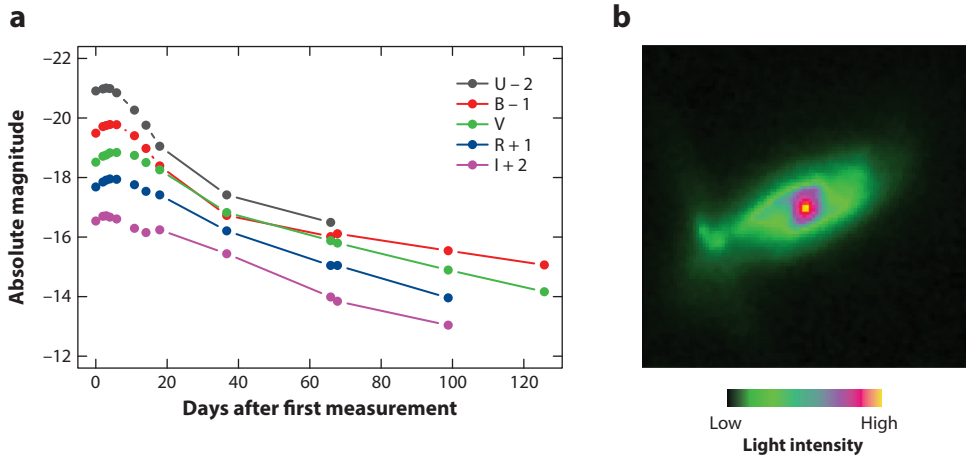

Figure 2

Temporal and spatial variation in flux measurements. (*a*) The light curves showing a supernova measured in five bands (UBVRI). U, B, V, R, and I are standard names attached to the filters. Note that the absolute magnitudes are shifted to improve readability (shifts are indicated by the numbers ±1 or ±2 in the legend). Because intensity increases as absolute magnitude decreases, it is customary to create these plots with the vertical axis reversed. Data are from Jha et al. (2006). (*b*) The spatial variation in intensity in a single band (J) for a galaxy as measured by the Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS).

3. THE STAGES OF STATISTICAL INFERENCE

This section presents the three stages of inference introduced above: from the raw observables to the object-specific parameters, from the object-specific parameters to the canonical parameters, and from the canonical parameters to the cosmological parameters. Although I present this framework within the context of a few examples, a wide range of estimation problems fit into it. The examples included here have been chosen to provide some breadth and background in this area.

3.1. From Raw Data to Object-Specific Parameters

As described in Section 2, the raw data from astronomical surveys consist of flux measurements at different wavelengths or, more commonly, at different bands along the spectrum. These data are observed at a fixed point in time, over a period of time (creating light curves), or at different spatial locations (creating images). The next step in any of these cases is to transform these measurements into estimates of parameters that specify properties of the object under study. A fundamental parameter to estimate is the type of the object. The SDSS, for example, classifies spectra into quasars, galaxies, and stars, and each of these broad classes has a number of subclasses. The classification is performed by comparing the observed spectrum with a family of template spectra, minimizing the weighted sum of the squared deviations between a smoothed version of the raw spectrum and the templates (Bolton et al. 2012). This weighting accounts for differences in the errors in the measurements. Even relatively basic functions of the observables (for example, ratios of or differences in the fluxes of different bands, called hardness ratios and colors, respectively) present challenging statistical issues because of such errors (Park et al. 2006). In the following subsections, I consider two additional inference problems of this type: redshift estimation and classification based on light curves.

THE EXPANDING UNIVERSE

Edwin Hubble's (1929) discovery that galaxies at a greater distance had greater redshift was the crucial initial observation that lent support to the Big Bang cosmological model. This observation implied both the expansion of the Universe and the existence of a time when objects were at a much smaller distance. In astrophysics, the scale factor $a(t)$ is generally defined as the ratio of the distance between two objects (on cosmological scales) at time t relative to the distance between them today. The current rate of change in $a(t)$ is called the Hubble parameter and is denoted H_0. The second derivative of $a(t)$ is also estimated to be positive; this acceleration is attributed to dark energy.

3.1.1. Redshift estimation. The comparison of template and observed spectra must account for the Doppler shift in the wavelengths of light resulting from the fact that on large scales, all astronomical objects are moving away from the observer (see the sidebar titled The Expanding Universe). Hence, the ratio of the wavelength of the observed light ($\lambda_{\text{observed}}$) to its wavelength when it was emitted (λ_{emitted}) is greater than one. This ratio is quantified by the redshift, denoted z:

$$z = \frac{\lambda_{\text{observed}}}{\lambda_{\text{emitted}}} - 1.$$

The search for the best-fitting template spectrum involves minimizing over candidate values for z with the observed spectrum shifted appropriately. (A spectrum that has been transformed to adjust for the Doppler shift is said to be in the rest frame.) By the standards of astrophysical problems, the estimation of redshift using spectra is relatively easy: Most SDSS galaxy redshift estimates have quoted standard errors of less than 0.03%, and the accuracy of the stated errors has been well established (Bolton et al. 2012).

Redshifts are a crucial ingredient in many inference problems in astrophysics, as the redshift is a proxy for the distance of an object from us, which in turn is a proxy for how far into the past our view of the object is being taken. Therefore, redshifts need to be estimated for all objects, not only those with spectra. The estimation of redshifts using only photometric observations is referred to as the photometric redshift estimation problem, and this problem is among the most widely explored inference challenges in astrophysics. Recall that photometry implies the availability of only approximately five magnitude measurements for each object, each corresponding to a particular band of wavelengths. Given the relationship between these measurements and the spectrum, it is not surprising that these magnitudes could be used to estimate redshift. However, it should also not come as a surprise that the errors in photometric estimates will greatly exceed those made when using the full spectrum. The absorption and emission lines that are so useful for aligning observed and template spectra are largely smoothed over by the filter response functions (see **Figure 1**).

The earliest proposed approaches to this problem were based on comparing the observed fluxes with those that would be derived from template spectra; this is the so-called template-fitting approach (Butchins 1981, Koo 1985). The general recipe is as follows: A family of templates is built using rest frame spectra of galaxies with known types and varying amounts of redshift applied. These templates are then subjected to the filter response functions to construct the expected magnitudes[1]: anticipated measurements under each known combination of object type and redshift. The observed colors are then compared with the training sample to obtain an estimate.

[1]Technically, they used colors instead of magnitudes. Colors are the differences between magnitudes in adjacent bands.

Butchins (1981) simply used the nearest neighbor in color space, but recent methods have used least squares measures. For example, Barro et al. (2011) used the following:

$$\chi^2(j, z, A) = \sum_{i=1}^{B} \left[\frac{F_{\text{obs},i} - A \cdot F_{\text{temp}\,j,i}(z)}{\sigma_i} \right]^2,$$

where $F_{\text{obs},i}$ is the observed flux in band i, $F_{\text{temp}\,j,i}(z)$ is the flux in band i using the template j shifted to redshift z, σ_i is the estimated standard error in $F_{\text{obs},i}$, A is a normalizing constant, and B is the number of bands. This quantity is minimized over j, z, and A to find the best redshift estimate. Dahlen et al. (2013) provide a listing and comparison of many template-fitting methods applied to CANDELS data.

Regression and supervised learning methods are also widely used for estimating redshifts on the basis of photometric data; such methods build on a training sample of objects with known (spectroscopic) redshifts. Connolly et al. (1995) fit a succession of linear models, from a simple linear model up to a fourth-order model, using the magnitudes four bands as the predictors. Since then, a wide range of estimation methods have been attempted, including neural networks (Collister & Lahav 2004, Firth et al. 2003, Laurino et al. 2011), random forests (Carliles et al. 2010, Carrasco Kind & Brunner 2013), k-nearest neighbors (Ball et al. 2008, Zhang et al. 2013), spectral connectivity analysis (Freeman et al. 2009), and boosted decision trees (Gerdes et al. 2010), among others. As spectroscopic sample sizes have grown, the feasibility and success of these approaches have increased, but they still suffer from the assumption that the training sample is representative of the photometric sample for which redshift estimates are required. This assumption is of particular concern given that spectroscopy is typically more readily available for closer, higher-quality objects; active learning may be useful for dealing with this challenge (Richards et al. 2012).

The SDSS utilizes the algorithm of Csabai et al. (2007) to estimate redshifts using photometry. This algorithm simply fits a local polynomial in a neighborhood surrounding the targeted predictors and uses that polynomial to make an estimate and to approximate the error in that estimate. The search for training objects that are in the neighborhood is facilitated by a k-dimensional (k-d) tree (Bentley 1975). **Figure 3*a*** shows the performance of this algorithm on the 6,514 galaxies in a particular region for which spectroscopic redshifts are available. The left plot shows that there is, in general, strong agreement between the estimate and the spectroscopic redshift. The root mean squared error for this sample is 0.0495.

Adequate quantification of the uncertainty in photometric redshift estimates is also critical. **Figure 3*b*** compares the nominal errors in the photometric redshift quoted by SDSS, along with the observed absolute error in the estimate; the observed errors seem to be in line with what is expected by the model. As there has been a shift toward Bayesian methods in astrophysics, calculation and reporting of a posterior distribution for each unknown redshift are becoming a standard practice [see, for instance, Budavári (2009)]. The primary advantage of this practice is that it creates a natural way to propagate this uncertainty forward into the donwstream analyses. Some of the estimators that yield posteriors are built around a classic Bayesian approach to the problem (Benítez 2000, Xia et al. 2009), but specification of the likelihood can be challenging. The technique utilized by SDSS is to simply utilize the distribution of training sample redshifts for the nearest neighbors in magnitude space. [The reader is referred to Sheldon et al. (2012) for the application to SDSS; the method is originally described and tested in Lima et al. (2008) and Cunha et al. (2009).]

3.1.2. Classification of variable sources. The LSST will make repeated observations of the same fields of view and hence will have the additional capacity to identify and measure variable objects,

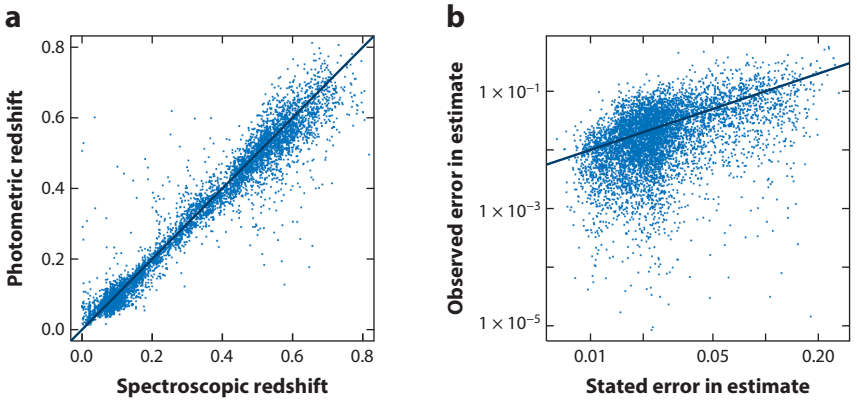

Figure 3

Sloan Digital Sky Survey (SDSS) photometric redshift estimation algorithm performance for 6,514 galaxies. Data are from SDSS Data Release 10. Panel *a* compares the redshift as estimated using only photometry with the more reliable spectroscopic estimates. There is strong agreement between the two estimates, but there are also several large errors. More importantly, the size and distribution of errors are a function of redshift. To illustrate this, panel *b* compares the actual error (the difference between the photometric and spectroscopic redshifts) with the quoted standard errors attached to each photometric estimate. Note that the use of log axes creates a false impression of a skewed distribution around the line of agreement.

i.e., phenomena that are changing in time. Identification will happen by taking the difference of images of the same part of the sky; the LSST is expected to report on over 10,000 such objects per evening, and a stated goal of the project is to be able to identify at least 1,000 of them in real time from a single field of view (LSST Science Collaboration et al. 2009). Simply identifying interesting phenomena in the massive collection of light curves requires sophisticated analyses (Blocker & Protopapas 2013), and classification into different object types (supernovae, blazars, or cataclysmic variables, to name a few) can proceed on the basis of measured light curves and any additional available local information. The capacity to do this classification in real time is important because, as stated by Bloom et al. (2012, p. 1176), "the vast majority of science conducted with time-variable objects . . . comes when more data are accumulated about the objects of interest."

One challenging and important problem in this domain is the classifcation of supernovae into one of several types. Of particular interest is the separation of Type Ia supernovae from others, as these supernovae are valuable sources of information about the nature of dark energy. **Figure 2*a*** shows an example of the light curves of a Type Ia supernova. The light curves of supernovae of other types will show subtle differences, mostly in the rate at which the magnitude drops off following the peak. Construction of a suitable training sample is a challenging problem in itself. Although some theoretical modeling of Type Ia supernova light curves has been done (Chatzopoulos et al. 2012), most of the modeling is based on simulations of the physical process (Blinnikov et al. 2006, Woosley et al. 2007), as is becoming increasingly common in astrophysics. Observational limitations (partially observed light curves, irregular spacing of observations, measurement errors) must also be incorporated into a training sample. The software package SNANA (Supernova Analysis Software) (Kessler et al. 2009) is a valuable tool for simulating realistic light curves and was used to generate a training sample for the *Supernova Photometric Classification Challenge* (Kessler et al. 2010b), which allowed the performances of a wide range of classification techniques to be compared (Kessler et al. 2010a).

A supernova is a transient object: It can be observed over only a limited window of time. Other variable objects show periodic behavior in their light curves, owing to either their internal

physics (e.g., Cepheids and RR Lyrae) or their interactions with their surroundings (e.g., eclipsing binaries). It is natural to estimate the periodogram to study the period structure in light curves, but the irregular time spacing must be taken into account. Long gaps are often present owing to periods when objects are hidden or not within the field of view of the telescope. A popular tool for estimating the periodogram in the case of irregularly spaced observations, the Lomb–Scargle periodogram, was developed by researchers in astrophysics (Lomb 1976, Scargle 1982). This estimator uses least squares to determine the best-fitting family of sine waves to the observed series; this family is then transformed into the power at different periods. Although this estimator continues to be widely used, there has been a significant growth in the use of machine learning approaches for classification based on any available information on the object, including the periodogram. There is growing recognition of the need to incorporate contextual information, i.e., properties of the neighborhood surrounding the variable object (Mahabal et al. 2011).

3.2. From Object-Specific Parameters to Canonical Parameters

Although the physical constants that parameterize the models for the Universe and its evolution remain the ultimate inferential target, the estimation of intermediate parameters in a step between the estimation of object-specific parameters and that of the cosmological parameters is almost always required and/or advantageous. These intermediate parameters are referred to here as canonical parameters, as these quantities naturally retain the important information in the objects under study while providing a significant compression of the often massive catalogs. In the ideal case, the estimator for the canonical parameter will be a minimally sufficient statistic for the cosmological parameters under study, although it is only realistic to claim there is approximate sufficiency. This step places significant demands on adequately accounting for observational limitations. For example, Akritas & Bershady (1996) developed a technique for performing linear regression in the case where there are dependent errors on both the predictor and explanatory variables, with a main application being the estimation of the Tully–Fisher relation (Tully & Fisher 1977). Below is described the estimation of two other canonical parameters, the aforementioned luminosity functions and angular correlation functions.

3.2.1. Luminosity functions. Perhaps the most fundamental example of a canonical parameter is the luminosity function. A standard definition used in astrophysics is that a luminosity function is the "number of objects per unit volume, per unit luminosity" (Sarjeant 2010, p. 135), but from a statistical standpoint, the estimation of the luminosity function is effectively a rescaled density estimation problem. Luminosity functions are calculated for all classes of objects, for example, quasars, galaxies, stars, and white dwarfs. Note that although they are called luminosity functions, the distributions are often expressed in terms of a related quantity, the absolute magnitude (M). The characteristic shape of a luminosity function is shown in **Figure 4**: At brighter magnitudes (smaller M), there is a drop in the number of objects. This is an estimate of the galaxy luminosity function based on SDSS Data Release 6 (Montero-Dorta & Prada 2009). The estimate here is based on a simple parametric form for the luminosity function put forth by Schechter (1976):

$$\phi_S(M) = n^* 10^{-0.4(\alpha+1)(M-M^*)} \exp\left(-10^{-0.4(M-M^*)}\right).$$

However, as stated by Binney & Merrifield (1998, p. 163), "[t]his formula was initially motivated by a simple model of galaxy formation (Press & Schechter 1974), but it has proved to have a wider range of application than originally envisaged . . . With larger, deeper surveys, the limitations of the simple Schechter function start to become apparent."

Although parametric forms more complex than the Schechter model have been proposed [e.g., by Boyle et al. (2000)], nonparametric estimators for luminosity functions appear to be more

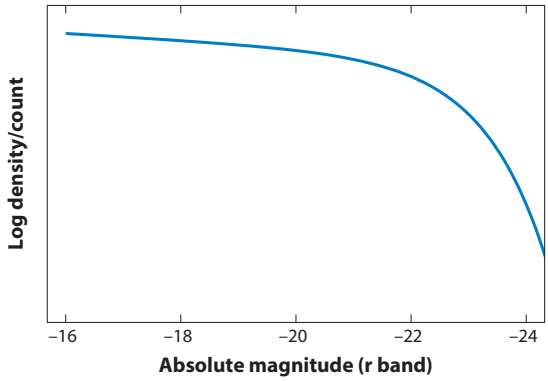

Figure 4

The Schechter form for the luminosity function of galaxies, estimated from r band measurements from Sloan Digital Sky Survey (SDSS) Data Release 6 (Montero-Dorta & Prada 2009).

promising. In fact, the earliest estimators were scaled histograms; the $1/V_{\max}$ method proposed by Schmidt (1968) took its name from the fact that observations were weighted by the inverse of the maximum volume at which that object could have been observed. This is a simple adjustment for Malmquist bias: Inherently dim objects at a great distance are more difficult to detect, and hence an observer would get the incorrect impression that objects at that distance are almost all bright. With this method, a dim object will have a small value for $V_{\max}$, and its contribution to the histogram will therefore be scaled up.

One way to address Malmquist bias is to treat the data as being subjected to truncation in apparent magnitude: Only objects brighter than a cutoff are assumed to be observed, creating a magnitude-limited sample. The cutoff is chosen such that one can believe that almost all objects that satisfy the bound will be observed. It is, however, a standard practice to further adjust estimates by considering the selection function of the survey, which simply reflects the varying probability of observing a given object as a function of brightness and distance. The challenge is that truncation in apparent magnitude creates an irregular truncation boundary in absolute magnitude. **Figure 5** shows a sample of 13,391 quasars from SDSS Data Release 3 (Richards et al. 2006) with apparent magnitudes between 15.0 and 19.1.[2] Distant quasars (those at large redshift) that are intrinsically dim are unobservable, whereas nearby quasars that are bright are difficult to distinguish from other objects.

Lynden-Bell (1971) introduced into the astronomy literature the nonparametric maximum likelihood estimator (NPMLE) for the case of one-sided truncation of absolute magnitude, and Woodroofe (1985) derived some of the asymptotic properties of this estimator. Efron & Petrosian (1999) extended the NPMLE to the case of double truncation of absolute magnitude. Each of these papers assumes that absolute magnitude and redshift are independent (and, hence, that the luminosity function does not evolve with redshift). The density estimate (or distribution function estimate) that results from a NPMLE procedure places all of the probability on observed data values, but with the current volume of data this does not seem to be a limiting factor. Efron & Petrosian (1999) also developed a permutation test for independence of the two variables. Independence of absolute magnitude and redshift is a strong assumption, and evidence suggests that it is not justified (see Boyle et al. 2000).

[2]The slight nonsmoothness in the boundary is caused by K-corrections.

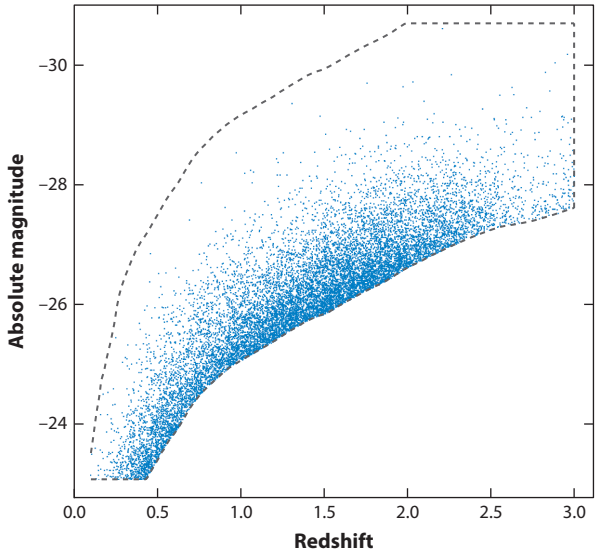

Figure 5

Measurements of 13,391 quasars from Sloan Digital Sky Survey (SDSS) Data Release 3. Quasars outside of the dashed boundary are truncated to create a sample for which there is a greater degree of confidence in the completeness within the retained region.

Recent methods have moved beyond the independence assumption and allowed for modeling of the evolution of the luminosity function with redshift. Schafer (2007) used a semiparametric approach in which the bivariate luminosity function $b(z, M)$ was decomposed into

$$\log b\,(z, M) = \mathbf{f}\,(z) + \mathbf{g}\,(M) + \mathbf{h}\,(z, M, \theta)\,,$$

where $\mathbf{h}(z, M, \theta)$ takes an assumed parametric form intended to model the dependence between the two random variables. For example, a physical parametric model for the evolution of the luminosity function that could be incorporated into $\mathbf{h}(z, M, \theta)$ may exist; alternatively, a simple first-order approximation has proven useful. The functions $\mathbf{f}(\cdot)$ and $\mathbf{g}(\cdot)$ are estimated nonparametrically, and bandwidth parameters are used to control the amount of smoothness in the estimate. Kelly et al. (2008) adopted a Bayesian approach, constructing the posterior for the parameters of the bivariate luminosity function modeled as a mixture of Gaussian densities. A major challenge moving forward is estimation of luminosity functions in the case in which only photometric estimates of the redshifts are available, introducing significant heteroskedastic measurement error.

3.2.2. Angular correlation functions. Homogeneity and isotropy are important assumptions underlying inferences regarding the Universe because they imply that the Earth receives a representative view of current and past structure, that we can take our observations as being a random draw from the larger population of interest. It follows naturally that the processes that generated observable structure are invariant to the direction of view. As a result, in a variety of situations, the properties of observed fields can be summarized via an angular correlation function that relates the covariance in that field to only the angle of separation between two directions of observation. The correlation is often naturally expressed via its Fourier transform, the power spectrum.

A leading example arises from the model for the cosmic microwave background radiation (CMB). The CMB is composed of photons that began to travel approximately 380,000 years after

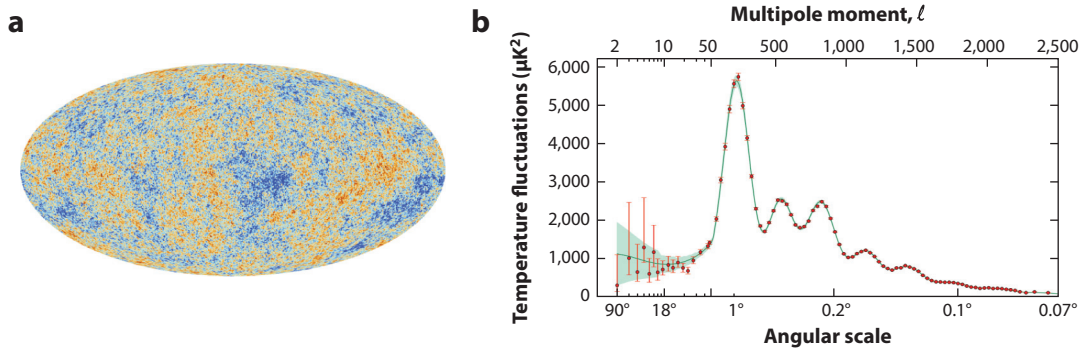

Figure 6

Results from measurements of the cosmic microwave background radiation (CMB) from the Planck satellite (Planck Collab. 2014a). Panel *a* depicts the measured CMB anisotropy field. The red areas represent regions where the photon temperature exceeds the mean, and the blue areas represent areas where the temperature is below the mean. (*b*) The estimate of the power spectrum of the field shown in panel *a*. The shaded region shows the uncertainty inherent in estimating the power spectrum owing to the fact that we can only observe a single realization of the CMB. Note that this is high at low ℓ but quickly diminishes. Images courtesy of ESA and the Planck Collaboration.

the Big Bang; prior to this, temperatures in the Universe were so high that photons could not travel freely. The small (on the order of 200 μK) but measurable fluctuations in the temperature of the CMB were seeded by the same processes that led to the wide range of structure visible in the present Universe. Theory relates the stochastic properties of this field to important cosmological parameters, and the discovery and measurement of the CMB have revolutionized the understanding of the Universe. **Figure 6*a*** depicts the CMB anisotropy field based on measurements taken by the Planck satellite (Planck Collab. 2013). Planck is the most recent in a sequence of increasingly precise instruments to measure the CMB; important predecessors include the Cosmic Background Explorer (COBE) (Bennett et al. 1996), for which George Smoot and John Mather were awarded the Nobel Prize, and the Wilkinson Microwave Anisotropy Probe (WMAP) (Bennett et al. 2013).

There are significant challenges in the processing and analysis of the observations of the CMB [see Cabella & Marinucci (2009) for a discussion]. Here, focus is placed on the model for the constructed map of the CMB anisotropy. Earth is at the center of the sphere depicted in **Figure 6*a***. If $z(s)$ and $z(t)$ are the observed CMB anisotropy in directions s and t, then the standard model for the CMB assumes that $(z(s), z(t))$ is a realization of a bivariate Gaussian pair with mean zero and the following covariance:

$$\mathbf{N}(s,t) + \sum_{\ell=1}^{\infty} \left(\frac{2\ell+1}{4\pi} \right) C_\ell(\theta)\, P_\ell(s \cdot t). \qquad (1)$$

Here, $\mathbf{N}(s,t)$ is a contribution from measurement error, and the remaining portion depends on s and t only through the cosine of the angle separating the pixels. This second term is a Legendre decomposition; $P_\ell(u)$ is the ℓth Legendre polynomial, defined for $-1 \le u \le 1$. Hence, the CMB anisotropy field is modeled as an isotropic Gaussian process on the sphere. In the absence of measurement error, such a process is fully characterized by its spherical harmonic power spectrum $\{C_\ell(\theta)\}_{\ell=1}^{\infty}$, which depends on the values of the cosmological parameters θ. [Extensive testing of the Gaussianity assumption has been conducted; the reader is referred to Marinucci (2004) for an overview of methods.]

In an idealized experiment with perfect ability to measure the CMB anisotropy,

$$\frac{1}{2\ell+1}\left(\sum_{m=-\ell}^{\ell}|a_{\ell m}|^2\right) \tag{2}$$

is the maximum likelihood estimator of C_ℓ, where $a_{\ell m}$ is the (ℓ, m) coefficient in the spherical harmonic transform of the field. There are significant observational limitations, however. In addition to error in each of the measurements, a large band exists within which the Milky Way obscures observation of approximately a third of the CMB. As the observable has a multivariate normal distribution with well-defined mean and covariance, derivation of the likelihood function that takes into account the incompletely observed field is not difficult. Unfortunately, however, direct maximization is computationally prohibitive. Following an idea promoted by Efstathiou (2004), Planck utilized a Markov chain Monte Carlo (MCMC)-based algorithm for numerically maximizing the likelihood only for estimating the power spectrum at large angular scales ($\ell \leq 50$) and the pseudo-C_ℓ method of Hivon et al. (2002) for higher ℓ (Planck Collab. 2013). For this method, one computes the spherical harmonic coefficients on an appropriately weighted version of the field:

$$\tilde{a}_{\ell m} = \int z(\mathbf{u})W(\mathbf{u})Y^*_{\ell m}(\mathbf{u})\mathrm{d}\mathbf{u},$$

where $W(\mathbf{u})$ is the weight in direction $\mathbf{u}$ and $Y_{\ell m}$ is the (ℓ, m) spherical harmonic. The weight is chosen to downweight noisy pixels and mask the unobserved regions. The integral is over the sphere but approximated by a sum over pixels. [CMB analysis motivated the creation of an equal-area pixelization scheme for the sphere, HEALPix (Górski et al. 2005).] The pseudo-C_ℓ, denoted $\tilde{C}_\ell$, are estimated from the $\tilde{a}_{\ell m}$ as in Equation 2. This form is chosen because it is possible to construct a nonsingular matrix $\mathbf{M}$ such that

$$\tilde{C}_\ell = \sum_{\ell'} M_{\ell,\ell'}C_{\ell'}.$$

[The reader is referred to appendix A2 in an article by Hivon et al. (2002) for the derivation.] The estimator follows naturally from inverting this relationship. The inefficiency of this plug-in estimator relative to the maximum likelihood estimator (MLE) is mitigated by the large sample of coefficients that are averaged at high ℓ.

Figure 6*b* shows the Planck estimate (Planck Collab. 2014a). Theory predicts a smooth power spectrum, and the variability in the estimate at low ℓ is due to cosmic variance, the fact that we have only a single CMB to measure. Nonparametric regression can be used to reduce this variance, provided care is taken to avoid smoothing over important features (Genovese et al. 2004). The ultimate step of estimating the cosmological parameters is facilitated by an adequate approximation to the distribution of these estimators; this process is a relatively complicated one that must take into account a wide range of potential sources of correlation and error in the estimator. For example, the Planck team had to account for potential error sources such as cosmic variance, instrumental noise, the masking caused by our galaxy, and contamination from sources other than the Milky Way, among others. As was the case for the estimator, a separate derivation is performed at low and high ℓ values. The complexity requires that the full likelihood be made available as a subroutine (see **http://wiki.cosmos.esa.int/planckpla/index.php/Main_Page** for details).

3.3. From Canonical Parameters to Cosmological Parameters

Current observations are largely consistent with the Lambda Cold Dark Matter (ΛCDM) model, a particularly simple model for the Universe consisting of fewer than 20 free parameters. (The

exact number depends on the particular data set being modeled.) For example, under the ΛCDM model, the power spectrum for the CMB can be characterized by 6 free parameters. More complex models require more parameters and, in some cases, functional quantities such as the equation of state, which are best estimated using flexible nonparametric methods (Genovese et al. 2009). As stated previously, estimation of cosmological parameters typically relies on the theory that relates these parameters to canonical parameters, so estimation for the cosmological parameters proceeds by treating the estimate of the canonical parameter as the data. In the ideal case, this estimator would be minimally sufficient for the cosmological parameters.

For example, the anisotropy observed in the CMB evolved from perturbations in the initial density field, which, under a simple model, is assumed to have power law spectrum $A(k)$, parameterized by A_s and n_s. Then the power spectrum for the CMB is

$$C_\ell(\theta) \propto \int \left(\frac{g(\ell, k, \theta)^2 A(k)}{k} \right) \mathrm{d}k, \tag{3}$$

where g is the transfer function relating the two spectra. The remaining parameters in θ include those that decompose the total matter or energy density of the Universe into important constituents; their total is denoted Ω. The components of Ω express density relative to the critical density, which is the amount of matter/energy needed to eventually force the Universe to stop expanding and collapse. If $\Omega = 1$, the Universe is flat. A leading theory is that the early Universe had a period of very rapid expansion, called inflation (Guth 1981); this theory explains observational evidence that suggests the Universe is flat, or nearly so. Recent analysis showed that Planck observations are consistent with inflation (BICEP2 Collab. 2014). The components of Ω include Ω_m, the density of total matter, and Ω_Λ, the density of dark energy. The ΛCDM model assumes that $\Omega_\Lambda + \Omega_m = \Omega = 1$. Examples of other important parameters include the optical depth (τ), which characterizes the probability that a CMB photon would have reionized, and the Hubble parameter (H_0), which represents the current rate of expansion of the Universe.

3.3.1. Classic approaches. A long-standing approach to parameter estimation in astrophysics is least squares, finding the value of the parameter θ that minimizes the sum of squared deviations between the estimated canonical parameter and its theoretical prediction under θ:

$$\chi^2(\theta) = \sum_{i=1}^{n} \left(\frac{\hat{f}_i - f_i(\theta)}{\sigma_i} \right)^2,$$

where $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n)$ is the estimated canonical parameter, $\mathbf{f}(\theta)$ is its theoretical prediction under cosmological parameters θ, and σ_i is the error in $\hat{f}_i$. (Assumed covariance structure in $\hat{\mathbf{f}}$ is often incorporated in the natural way.) The literature is filled with references to χ^2 values (the sums of squared deviations for a fit) and reduced χ^2 values (these sums of squares divided by the degrees of freedom). The development is well presented in the standard text of Bevington (1969). Uncertainties in the parameters are calculated based on the Hessian of the χ^2 function at its minimum. Avni (1976) assumed that χ^2 has the chi-squared distribution in order to construct a widely used procedure for constructing confidence regions for θ: The region consists of all parameter values θ' for which $\chi^2(\theta') - \chi^2(\hat{\theta})$ is less than some constant.

This least squares construction clearly performs best in cases in which $\hat{\mathbf{f}}$ is approximately Gaussian with mean $\mathbf{f}(\theta)$ and covariance consistent with the assumed form; in this case, the use of the curvature in the χ^2 surface matches the well-grounded usage of MLEs. The limitations of these approximations are evident as more data become available and greater precision is expected

Table 1 The seven parameters in the Lambda Cold Dark Matter (ΛCDM) model[a]

Parameter	Description	MLE	68% credible interval
Ω_b	Baryonic matter density	0.0490	0.0490 ± 0.0073
Ω_m	Total matter density	0.3175	0.314 ± 0.020
Ω_Λ	Dark energy density	0.6825	0.686 ± 0.020
H_0	Rate of expansion (km/s/Mpc)	67.11	67.4 ± 1.4
τ	The optical depth	0.0925	0.097 ± 0.038
A_s	Amplitude of initial spectrum ($\times 10^9$)	2.215	2.23 ± 0.16
n_s	Spectral index of initial spectrum	0.9624	0.9616 ± 0.0094

[a]The parameters that define the ΛCDM model, as estimated by Planck (Planck Collab. 2013). Under this model the Universe is assumed to be flat, so $\Omega_m + \Omega_\Lambda = 1$. Hence there are six free parameters. Abbreviation: MLE, maximum likelihood estimate.

in estimates. As a result, there has been a shift toward the appropriate application of maximum likelihood estimation, which has emphasized constructing Fisher information matrices for quoting uncertainties and confidence regions. This naturally places an increased burden on the adequate modeling of the relevant likelihood function.

3.3.2. The rise of Bayesian methods. A prominent challenge in these inference problems is the complexity of the relationships between the cosmological parameters and the likelihood for the estimator of the canonical parameter. This complexity is largely a result of the indirect relationship between these parameters and the observables. For example, as presented above, the power spectrum of the CMB is derived from input cosmological parameters by modeling the effect of these parameter values on the properties of the Universe, which in turn affect the paths of the CMB photons. Hence, the mapping from cosmological parameters to a power spectrum is available only via sophisticated computer subroutines, the most widely used being CMBFAST (Seljak & Zaldarriaga 1996, 1999).

In large part owing to these limitations, Christensen et al. (2001) and Knox et al. (2001) promoted the use of MCMC techniques for cosmological parameter estimation and for the CMB case in particular. Bayesian methods had been utilized and promoted in astrophysics prior to this time [see Loredo (2013) for background], but the computational advantages of MCMC brought them into much wider use. The random walks that underlie MCMC can proceed in a natural manner in cosmological parameter space, requiring only the calculation of the forward mapping from these parameter values into the needed likelihood. The Bayesian approach also allows for the easy construction of one-dimensional and multidimensional region estimates of the unknowns via marginalizations of the posterior. By the time the parameter estimates from the Wilkinson Microwave Anisotropy Probe (WMAP) (Verde et al. 2003) were released, MCMC had become the widely accepted approach to estimating cosmological parameters, and the Bayesian philosophy had been embraced.

Significant effort has been placed into the construction of tools that ease the application of MCMC in cosmology. Planck utilized the software package CosmoMC (Lewis 2013, Lewis & Bridle 2002) to implement MCMC for cosmological parameter estimation. **Table 1** shows the MLEs and 68% credible intervals for the key parameters of the ΛCDM model (Planck Collab. 2014a).[3] Careful consideration is also placed on appropriate parameterization in which to

[3]Note that it is standard in astrophysics to construct 68% intervals owing to their connection with error bars of size equal to the standard error intervals. It is also common to call these "confidence intervals" even though they are based on the posterior.

implement the chains of the MCMC; there are significant degeneracies between parameters that can lead to poor performance if not addressed (Kosowsky et al. 2002). CosmoPMC (Kilbinger et al. 2011) implements the population Monte Carlo (PMC) approach developed by Cappé et al. (2008) for use in the same contexts as an alternative to MCMC for approximating the posterior for cosmological parameters. The reader is referred to Wraith et al. (2009) for a comparison of the methods.

Improvements in the quantity and resolution of the available data have resulted in a corresponding need to improve the quality of the approximations to the likelihood functions that relate the cosmological parameters to the distribution for the estimators for the canonical parameters. A growing challenge (opportunity) is the utilization of sophisticated simulation models that recreate relevant aspects of the data-generation process. It is increasingly becoming the case that such models reflect the best understanding of the relationships between parameters of interest and the observables. In describing the GADGET-2 simulation model, Springel (2005, p. 1105) stated that "[w]ithout numerical simulations, the ΛCDM model may arguably not have developed into the leading theoretical paradigm . . . because direct simulation is often the only available tool to compute accurate theoretical predictions in the highly non-linear regime of gravitational dynamics and hydrodynamics." GADGET-2 and its descendants were used to create the Millennium Simulations (Angulo et al. 2012, Boylan-Kolchin et al. 2009, Springel et al. 2005). The resulting catalogs are studied in much the same way as observed data: Comparisons are made between the resulting estimates of luminosity functions and other canonical parameters to validate the parameter estimates that result from observational studies [see, for example, Guo et al. (2011)]. There are significant practical limitations to the use of these simulations models to constrain cosmological parameters, as exploration of the cosmological parameter space is not computationally feasible. Schneider et al. (2008, p. 1) describe methods for constructing Gaussian process emulators of complex simulation models, motivating their work by pointing out that "[t]he computational demands for future observations will only increase as more accurate theoretical predictions are required to match the reduced errors in the data." Many smaller-scale simulation codes exist (e.g., SNANA) that can feasibly be used for parameter estimation; such codes have the advantage that observational effects can be incorporated into the process (Weyant et al. 2013).

4. DISCUSSION

The above examples are only a sampling of the ways in which statistical inference plays a role in astrophysics. Nevertheless, these examples have been chosen to illustrate some important challenges in this domain. First, methods must be able to handle massive data sets in order to be practical. Future surveys such as the LSST will gather data on billions of objects, and taking full advantage of these data sets will place unprecedented demands on statistical methods. Second, observational limitations play an important role. The isotropy and homogeneity assumptions allow properties of the Universe to be inferred from the limited view from the Earth, but significant errors that would result from not adjusting for limitations such as Malmquist bias remain. Third, methods need to be able to deal with heteroskedastic measurement error. Careful attention must be paid to the errors in the estimates of the object-specific parameters, especially at the stage of estimating canonical parameters. These estimates will inevitably be subject to an amount of error that depends on other properties of the object. Fourth, the ultimate target, the cosmological parameters, are related to the observables in complex ways. As the CMB example illustrates, even though the observations constrain the cosmological parameters, the nature of the relationship is often indirect. In many cases of interest, the best hope may be to model the data-generating process using realistic simulation models. Finally, a shift from a variance-dominated era to a bias-dominated era has occurred. As sample sizes have increased, there has been a move from concern with the

variance of estimators to concern with bias, often referred to as systematic errors by astrophysicists. These errors arise from a range of sources, including likelihood function misspecification, observational biases, and so forth. Greater attention must be focused on methods for dealing with this source of error; for instance, Lee et al. (2011) present a general method for modeling the bias due to instrumental errors. Clearly, astrophysics will continue to grow as a source of challenging and interesting statistical inference problems.

SUMMARY POINTS

1. Statistical inference is a key component of modern research in astrophysics, as the field is becoming increasingly data driven.
2. Many astrophysical studies have the ultimate goal of estimating key cosmological parameters.
3. Inference challenges in astrophysics are characterized by complexity: The relationship between cosmological parameters and the observable data is complex, and these data are of complex form. For this reason, it is useful to break the full inference process into three distinct stages.
4. The first stage of inference, estimating properties of celestial objects from the raw observables, is often challenging owing to limitations in what can be observed and errors in the measurements that are available.
5. The second stage of inference, estimating canonical parameters from the object-specific properties, is often made difficult by our Earth-centered, noisy, limited view of the Universe.
6. The third stage of inference, estimating the cosmological parameters from the estimates of the canonical parameters, requires careful consideration of the distribution of these estimates. Markov chain Monte Carlo is a popular technique for this stage.
7. At all stages, great care must be taken to adequately model the joint distribution of errors in the estimates and then to carry these errors forward into the subsequent analysis.

FUTURE ISSUES

1. Because of ongoing surveys such as the Sloan Digital Sky Survey (SDSS), as well as future surveys such as the Large Synoptic Survey Telescope (LSST), the already substantial volume of data available to astronomers is only growing. Focus must be placed on the development of procedures that can take advantage of this, avoiding data reduction steps that discard useful physical information.
2. The LSST will be an exclusively photometric survey, meaning that statistical inference procedures must be adapted to handle the low-resolution nature of this information. In particular, techniques for accurately estimating redshift and classifying observed objects are crucial.
3. Cosmological simulation models continue to grow in their accuracy and resolution, but to fully exploit these tools, inference procedures must be able to adapt to situations in which a simple likelihood function is unavailable.

4. As the volume of data grows, the field shifts from variance-dominated challenges to bias-dominated challengs. Statistical procedures must adapt by adequately modeling and incorporating errors in the measurements and estimates derived from those data. Nonparametric procedures will play a central role at this stage.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Thanks go to Jessi Cisewski, Peter Freeman, Michael Vespe, and Larry Wasserman for their help and comments on this work. Funding is from National Science Foundation (NSF) grant DMS-1106956.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the US Department of Energy Office of Science. The SDSS-III website is **http://www.sdss3.org/**. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

This work is based on observations taken by the CANDELS Multi-Cycle Treasury Program with the NASA/ESA HST, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555.

LITERATURE CITED

Ade PAR, Aghanim N, Alves MIR, Armitage-Caplan C, Arnaud M, et al. (Planck Collab.). 2014a. *Planck* 2013 results. I. Overview of products and scientific results. *Astron. Astrophys.* 571:A1

Ade PAR, Aghanim N, Armitage-Caplan C, Arnaud M, Ashdown M, et al. (Planck Collab.) 2014b. *Planck* 2013 results. XV. CMB power spectra and likelihood. *Astron. Astrophys.* 571:A15

Ade PAR, Aghanim N, Armitage-Caplan C, Arnaud M, Ashdown M, et al. (Planck Collab.) 2014c. *Planck* 2013 results. XVI. Cosmological parameters. *Astron. Astrophys.* 571:A16

Ade PAR, Aikin RW, Barkats D, Benton SJ, Bischoff CA, et al. (BICEP2 Collab.) 2014. BICEP2 I: detection of B-mode polarization at degree angular scales. *Phys. Rev. Lett.* 112:241101

Ahn CP, Alexandroff R, Allende Prieto C, Anders F, et al. 2013. The tenth data release of the Sloan Digital Sky Survey: first spectroscopic data from the SDSS-III Apache Point Observatory Galactic Evolution Experiment. arXiv:1307.7735 [astro-ph.IM]

Akritas MG, Bershady MA. 1996. Linear regression for astronomical data with measurement errors and intrinsic scatter. *Astrophys. J.* 470:706

Angulo RE, Springel V, White SDM, Jenkins A, Baugh CM, Frenk CS. 2012. Scaling relations for galaxy clusters in the Millennium-XXL simulation. *Mon. Not. R. Astron. Soc.* 426:2046–62
Avni Y. 1976. Energy spectra of X-ray clusters of galaxies. *Astrophys. J.* 210:642–46
Ball NM, Brunner RJ, Myers AD, Strand NE, Alberts SL, Tcheng D. 2008. Robust machine learning applied to astronomical data sets. III. Probabilistic photometric redshifts for galaxies and quasars in the SDSS and GALEX. *Astrophys. J.* 683:12–21
Barro G, Pérez-González PG, Gallego J, Ashby MLN, Kajisawa M, et al. 2011. UV-to-FIR analysis of *Spitzer*/IRAC sources in the extended Groth strip. II. Photometric redshifts, stellar masses, and star formation rates. *Astrophys. J. Suppl.* 193:30
Benítez N. 2000. Bayesian photometric redshift estimation. *Astrophys. J.* 536:571–83
Bennett CL, Banday AJ, Gorski KM, Hinshaw G, Jackson P, et al. 1996. Four-Year COBE DMR cosmic microwave background observations: maps and basic results. *Astrophys. J. Lett.* 464:L1
Bennett CL, Larson D, Weiland JL, Jarosik N, Hinshaw G, et al. 2013. Nine-year *Wilkinson Microwave Anisotropy Probe* (*WMAP*) observations: final maps and results. *Astrophys. J. Suppl.* 208:20
Bentley JL. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* 18:509–17
Bevington P. 1969. *Data Reduction and Error Analysis for the Physical Sciences*. New York: McGraw-Hill
Binney J, Merrifield M. 1998. *Galactic Astronomy*. Princeton, NJ: Princeton University Press
Blinnikov SI, Röpke FK, Sorokina EI, Gieseler M, Reinecke M, et al. 2006. Theoretical light curves for deflagration models of type Ia supernova. *Astron. Astrophys.* 453:229–40
Blocker A, Protopapas P. 2013. Semi-parametric robust event detection for massive time-domain databases. In *Statistical Challenges in Modern Astronomy V*, ed. ED Feigelson, GJ Babu, pp. 177–87. New York: Springer
Bloom JS, Richards JW, Nugent PE, Quimby RM, Kasliwal MM, et al. 2012. Automating discovery and classification of transients and variable stars in the synoptic survey era. *Publ. Astron. Soc. Pac.* 124:1175–96
Bolton AS, Schlegel DJ, Aubourg É, Bailey S, Bhardwaj V, et al. 2012. Spectral classification and redshift measurement for the SDSS-III Baryon Oscillation Spectroscopic Survey. *Astron. J.* 144:144
Boylan-Kolchin M, Springel V, White SDM, Jenkins A, Lemson G. 2009. Resolving cosmic structure formation with the Millennium-II Simulation. *Mon. Not. R. Astron. Soc.* 398:1150–64
Boyle BJ, Shanks T, Croom SM, Smith RJ, Miller L, et al. 2000. The 2df QSO Redshift Survey—I. The optical luminosity function of quasi-stellar objects. *Mon. Not. R. Astron. Soc.* 317:1014–22
Boyle BJ, Shanks T, Peterson BA. 1988. The evolution of optically selected QSOs. II. *Mon. Not. R. Astron. Soc.* 235:935–48
Budavári T. 2009. A unified framework for photometric redshifts. *Astrophys. J.* 695:747–54
Butchins SA. 1981. Predicted redshifts of galaxies by broadband photometry. *Astron. Astrophys.* 97:407–9
Cabella P, Marinucci D. 2009. Statistical challenges in the analysis of cosmic microwave background radiation. *Ann. Appl. Stat.* 3:61–95
Cappé O, Douc R, Guillin A, Marin JM, Robert CP. 2008. Adaptive importance sampling in general mixture classes. *Stat. Comput.* 18:447–59
Carliles S, Budavári T, Heinis S, Priebe C, Szalay AS. 2010. Random forests for photometric redshifts. *Astrophys. J.* 712:511–15
Carrasco Kind M, Brunner RJ. 2013. TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Mon. Not. R. Astron. Soc.* 432:1483–501
Chatzopoulos E, Wheeler JC, Vinko J. 2012. Generalized semi-analytical models of supernova light curves. *Astrophys. J.* 746:121
Christensen N, Meyer R, Knox L, Luey B. 2001. Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements. *Class. Quantum Gravity* 18:2677–88
Collister AA, Lahav O. 2004. ANNz: estimating photometric redshifts using artificial neural networks. *Publ. Astron. Soc. Pac.* 116:345–51
Connolly AJ, Csabai I, Szalay AS, Koo DC, Kron RG, Munn JA. 1995. Slicing through multicolor space: galaxy redshifts from broadband photometry. *Astron. J.* 110:2655
Csabai I, Dobos L, Trencséni M, Herczegh G, Józsa P, et al. 2007. Multidimensional indexing tools for the virtual observatory. *Astron. Nachr.* 328:852

Cunha CE, Lima M, Oyaizu H, Frieman J, Lin H. 2009. Estimating the redshift distribution of photometric galaxy samples II. Applications and tests of a new method. *Mon. Not. R. Astron. Soc.* 396:2379–98

Dahlen T, Mobasher B, Faber SM, Ferguson HC, Barro G, et al. 2013. A critical assessment of photometric redshift methods: a CANDELS investigation. *Astrophys. J.* 775:93

Efron B, Petrosian V. 1999. Nonparametric methods for doubly truncated data. *J. Am. Stat. Assoc.* 94:824–34

Efstathiou G. 2004. Myths and truths concerning estimation of power spectra: the case for a hybrid estimator. *Mon. Not. R. Astron. Soc.* 349:603–26

Eisenstein DJ, Weinberg DH, Agol E, Aihara H, Allende Prieto C, et al. 2011. SDSS-III: massive spectroscopic surveys of the distant Universe, the Milky Way, and extra-solar planetary systems. *Astron. J.* 142:72

Firth AE, Lahav O, Somerville RS. 2003. Estimating photometric redshifts with artificial neural networks. *Mon. Not. R. Astron. Soc.* 339:1195–202

Freeman PE, Newman JA, Lee AB, Richards JW, Schafer CM. 2009. Photometric redshift estimation using spectral connectivity analysis. *Mon. Not. R. Astron. Soc.* 398:2012–21

Genovese CR, Freeman P, Wasserman L, Nichol RC, Miller C. 2009. Inference for the dark energy equation of state using type Ia supernova data. *Ann. Appl. Stat.* 3:144–78

Genovese CR, Miller CJ, Nichol RC, Arjunwadkar M, Wasserman L. 2004. Nonparametric inference for the cosmic microwave background. *Stat. Sci.* 19:308–21

Gerdes DW, Sypniewski AJ, McKay TA, Hao J, Weis MR, et al. 2010. ArborZ: photometric redshifts using boosted decision trees. *Astrophys. J.* 715:823–32

Gilbank DG, Baldry IK, Balogh ML, Glazebrook K, Bower RG. 2010. The local star formation rate density: assessing calibrations using [OII], Hα and UV luminosities. *Mon. Not. R. Astron. Soc.* 405:2594–614

Górski KM, Hivon E, Banday AJ, Wandelt BD, Hansen FK, et al. 2005. HEALPix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere. *Astrophys. J.* 622:759–71

Grogin NA, Kocevski DD, Faber SM, Ferguson HC, Koekemoer AM, et al. 2011. CANDELS: The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey. *Astrophys. J. Suppl.* 197:35

Gunn JE, Carr M, Rockosi C, Sekiguchi M, Berry K, et al. 1998. The Sloan Digital Sky Survey photometric camera. *Astron. J.* 116:3040–81

Guo Q, White S, Boylan-Kolchin M, De Lucia G, Kauffmann G, et al. 2011. From dwarf spheroidals to cD galaxies: simulating the galaxy population in a ΛCDM cosmology. *Mon. Not. R. Astron. Soc.* 413:101–31

Guth AH. 1981. Inflationary universe: a possible solution to the horizon and flatness problems. *Phys. Rev. D* 23:347–56

Hivon E, Górski KM, Netterfield CB, Crill BP, Prunet S, Hansen F. 2002. MASTER of the cosmic microwave background anisotropy power spectrum: a fast method for statistical analysis of large and complex cosmic microwave background data sets. *Astrophys. J.* 567:2–17

Hubble E. 1929. A relation between distance and radial velocity among extra-galactic nebulae. *PNAS* 15:168–73

Ivezic Z, Tyson JA, Acosta E, Allsman R, Anderson SF, et al. 2014. LSST: from science drivers to reference design and anticipated data products. arXiv:0805.2366 [astro-ph]

Jha S, Kirshner RP, Challis P, Garnavich PM, Matheson T, et al. 2006. *UBVRI* light curves of 44 type Ia supernovae. *Astron. J.* 131:527–54

Kelly BC, Fan X, Vestergaard M. 2008. A flexible method of estimating luminosity functions. *Astrophys. J.* 682:874–95

Kessler R, Bassett B, Belov P, Bhatnagar V, Campbell H, et al. 2010a. Results from the Supernova Photometric Classification Challenge. *Publ. Astron. Soc. Pac.* 122:1415–31

Kessler R, Bernstein JP, Cinabro D, Dilday B, Frieman JA, et al. 2009. SNANA: a public software package for supernova analysis. *Publ. Astron. Soc. Pac.* 121:1028–35

Kessler R, Conley A, Jha S, Kuhlmann S. 2010b. Supernova Photometric Classification Challenge. arXiv:1001.5210 [astro-ph.IM]

Kilbinger M, Benabed K, Cappé O, Cardoso J-F, Coupon J, et al. 2011. CosmoPMC: cosmology population Monte Carlo. arXiv:1101.0950 [astro-ph.CO]

Knox L, Christensen N, Skordis C. 2001. The age of the Universe and the cosmological constant determined from cosmic microwave background anisotropy measurements. *Astrophys. J. Lett.* 563:L95–98

Koekemoer AM, Faber SM, Ferguson HC, Grogin NA, Kocevski DD, et al. 2011. CANDELS: the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey—the Hubble Space Telescope observations, imaging data products, and mosaics. *Astrophys. J. Suppl.* 197:36

Koo DC. 1985. Optical multicolors: a poor person's Z machine for galaxies. *Astron. J.* 90:418–40

Kosowsky A, Milosavljevic M, Jimenez R. 2002. Efficient cosmological parameter estimation from microwave background anisotropies. *Phys. Rev. D* 66:063007

Large Synoptic Survey Telescope (LSST) Sci. Collab., LSST Proj. 2009. *LSST Science Book*, *Version 2.0*. arXiv:0912.0201 [astro-ph.IM]. **http://www.lsst.org/lsst/scibook**

Laurino O, D'Abrusco R, Longo G, Riccio G. 2011. Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation. *Mon. Not. R. Astron. Soc.* 418:2165–95

Lee H, Kashyap VL, van Dyk DA, Connors A, Drake JJ, et al. 2011. Accounting for calibration uncertainties in X-ray analysis: effective areas in spectral fitting. *Astrophys. J.* 731:126

Lewis A. 2013. Efficient sampling of fast and slow cosmological parameters. *Phys. Rev. D* 87:103529

Lewis A, Bridle S. 2002. Cosmological parameters from CMB and other data: a Monte Carlo approach. *Phys. Rev. D* 66:103511

Lima M, Cunha CE, Oyaizu H, Frieman J, Lin H, Sheldon ES. 2008. Estimating the redshift distribution of photometric galaxy samples. *Mon. Not. R. Astron. Soc.* 390:118–30

Lomb NR. 1976. Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci.* 39:447–62

Loredo T. 2013. Bayesian astrostatistics: a backward look to the future. In *Astrostatistical Challenges for the New Astronomy*, ed. JM Hilbe, pp. 15–40. New York: Springer

Lynden-Bell D. 1971. A method of allowing for known observational selection in small samples applied to 3CR quasars. *Mon. Not. R. Astron. Soc.* 155:95–118

Mahabal AA, Djorgovski SG, Drake AJ, Donalek C, Graham MJ, et al. 2011. Discovery, classification, and scientific exploration of transient events from the Catalina Real-time Transient Survey. *Bull. Astron. Soc. India* 39:387–408

Marinucci D. 2004. Testing for non-Gaussianity on cosmic microwave background radiation: a review. *Stat. Sci.* 19:294–307

Marshall HL, Huchra JP, Tananbaum H, Avni Y, Braccesi A, et al. 1984. A complete sample of quasars at B = 19.80. *Astrophys. J.* 283:50–58

Matthews TA, Sandage AR. 1963. Optical identification of 3C 48, 3C 196, and 3C 286 with stellar objects. *Astrophys. J.* 138:30–56

Montero-Dorta AD, Prada F. 2009. The SDSS DR6 luminosity functions of galaxies. *Mon. Not. R. Astron. Soc*. 399:1106–18

Pâris I, Petitjean P, Aubourg É, Ross NP, Myers AD, et al. 2014. The Sloan Digital Sky Survey quasar catalog: tenth data release. *Astron. Astrophys.* 563:A54

Park T, Kashyap VL, Siemiginowska A, van Dyk DA, Zezas A, et al. 2006. Bayesian estimation of hardness ratios: modeling and computations. *Astrophys. J.* 652:610–28

Pei YC. 1995. The luminosity function of quasars. *Astrophys. J.* 438:623–31

Press WH, Schechter P. 1974. Formation of galaxies and clusters of galaxies by self-similar gravitational condensation. *Astrophys. J.* 187:425–38

Richards GT, Strauss MA, Fan X, Hall PB, Jester S, et al. 2006. The Sloan Digital Sky Survey quasar survey: quasar luminosity function from Data Release 3. *Astron. J.* 131:2766–87

Richards JW, Starr DL, Brink H, Miller AA, Bloom JS, et al. 2012. Active learning to overcome sample selection bias: application to photometric variable star classification. *Astrophys. J.* 744:192

Sarjeant S. 2010. *Observational Cosmology*. Cambridge, UK: Cambridge Univ. Press

Scargle JD. 1982. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J.* 263:835–53

Schafer CM. 2007. A statistical method for estimating luminosity functions using truncated data. *Astrophys. J.* 661:703–13

Schechter P. 1976. An analytic expression for the luminosity function for galaxies. *Astrophys. J.* 203:297–306

Schmidt M. 1968. Space distribution and luminosity functions of quasi-stellar radio sources. *Astrophys. J.* 151:393–409

Schneider DP, Richards GT, Fan X, Hall PB, Strauss MA, et al. 2002. The Sloan Digital Sky Survey quasar catalog. I. Early data release. *Astron. J.* 123:567–77

Schneider MD, Knox L, Habib S, Heitmann K, Higdon D, Nakhleh C. 2008. Simulations and cosmological inference: a statistical model for power spectra means and covariances. *Phys. Rev. D* 78:063529

Seljak U, Zaldarriaga M. 1996. A line-of-sight integration approach to cosmic microwave background anisotropies. *Astrophys. J.* 469:437–44

Seljak U, Zaldarriaga M. 1999. CMBFAST: a microwave anisotropy code. *Astrophys. Source Code Libr.* **http://ascl.net/9909.004**

Sheldon ES, Cunha CE, Mandelbaum R, Brinkmann J, Weaver BA. 2012. Photometric redshift probability distributions for galaxies in the SDSS DR8. *Astrophys. J. Suppl.* 201:32

Sparke LS, Gallagher JS. 2007. *Galaxies in the Universe: An Introduction*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.

Springel V. 2005. The cosmological simulation code GADGET-2. *Mon. Not. R. Astron. Soc.* 364:1105–34

Springel V, White SDM, Jenkins A, Frenk CS, Yoshida N, et al. 2005. Simulating the joint evolution of quasars, galaxies and their large-scale distribution. *Nature* 435:629–36

Tully RB, Fisher JR. 1977. A new method of determining distances to galaxies. *Astron. Astrophys.* 54:661–73

Verde L, Peiris HV, Spergel DN, Nolta MR, Bennett CL, et al. 2003. First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: parameter estimation methodology. *Astrophys. J. Suppl.* 148:195–211

Weyant A, Schafer C, Wood-Vasey WM. 2013. Likelihood-free cosmological inference with type Ia supernovae: approximate Bayesian computation for a complete treatment of uncertainty. *Astrophys. J.* 764:116

Woodroofe M. 1985. Estimating a distribution function with truncated data. *Ann. Stat.* 13:163–77

Woosley SE, Kasen D, Blinnikov S, Sorokina E. 2007. Type Ia supernova light curves. *Astrophys. J.* 662:487–503

Wraith D, Kilbinger M, Benabed K, Cappé O, Cardoso JF, et al. 2009. Estimation of cosmological parameters using adaptive importance sampling. *Phys. Rev. D* 80:023507

Xia L, Cohen S, Malhotra S, Rhoads J, Grogin N, et al. 2009. Improved photometric redshifts with surface luminosity priors. *Astron. J.* 138:95–101

Zhang Y, Ma H, Peng N, Zhao Y, Wu X-B. 2013. Estimating photometric redshifts of quasars via the *k*–nearest neighbor approach based on large survey databases. *Astron. J.* 146:22