

Reflections on SCMA III

John Rice
Department of Statistics
University of California at Berkeley

September 11, 2001

1 Introduction

It has been a great privilege to participate in this fascinating meeting and a great challenge to be asked to comment on the wide variety of issues that have arisen. I will try to place the papers we have heard here in some perspective and to outline some current and future challenges and opportunities lying in the intersection of statistics and astronomy. I ask the reader to bear in mind that the "seeing conditions" are poor.

2 A spectrum of statistical methodology

In considering the wide variety of statistical methodology relevant to astronomy, it may be helpful to view it on a spectrum ranging from procedurally based methods to methods based on highly specified stochastic models. Different regions on this spectrum are relevant to different types of problems and individual statisticians often have "personal equations" that influence where their contributions fall.

The following example is illustrative: Smoothing splines were first proposed as a procedure for passing a smooth curve through a noisy scatter plot of observations (y_i, x_i) , $i = 1, \dots, n$. It was desired that the curve not be of any simple parametric form, such as a low degree polynomial, but merely be "smooth." No explicit stochastic structure was assumed for the data. In [9] Reinsch proposed choosing the curve, $g()$ as the minimizer of

$$\sum_{i=1}^n (y_i - g(x_i))^2 \tag{1}$$

subject to the constraint $\int [g''(x)]^2 dx \leq \Omega$. (This basic idea had been around for some time—see [16] for more complete references). Using a Lagrange multiplier, the problem can be written as that of choosing $g()$ to minimize

$$\sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int [g''(x)]^2 dx. \tag{2}$$

Although this optimization problem has some heuristic appeal, the proposal would not have had much impact were it not the case that the minimizing $g(\cdot)$ is a cubic spline and that there are fast and stable numerical algorithms for its computation. The solution depends upon the choice of the smoothing parameter, λ : small values of λ give rise to highly oscillatory functions and large values to very smooth ones. It was left to the user to interactively determine a satisfactory choice.

The next stage in the study of this method was an examination of its properties by statisticians with frequentist personal equations. Thus, an explicit stochastic element was added to the structure to produce a statistical model: it was assumed that the data were of the form $y_i = g_0(x_i) + \epsilon_i$ where the ϵ_i are independent random variables with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. Note that this itself is quite an idealization. Frequentist properties of the spline estimate were the subject of intensive theoretical and numerical research and are now quite well understood.

If the random errors are modeled as Gaussian, the procedure can be viewed as an example of penalized likelihood in which the log likelihood is the first term in (2) and the second term penalizes rough solutions. It is a canonical example of nonparametric regression, in which there is a tradeoff between bias and variance that does not typically occur in parametric models. (It could of course be argued that any parametric model is an approximation and hence may well give rise to bias, so that the distinction between parametric and nonparametric models is illusory). One of the widely used ways of selecting λ in a data-driven way to achieve this balance is cross-validation [17].

Finally, a statistician with a Bayesian personal equation examining (2) will see the sum of a log likelihood and the log of a prior. Wahba [15] identified the prior as a doubly integrated Brownian motion where λ is the variance parameter of this stochastic process. Now another layer of idealization has been introduced and can be formally used to construct posterior credible regions, for example.

The bottom line: what really matters is how well a method works. Are there efficient and stable computational algorithms? How well does it work on a suite of simulated data? On a variety of real data sets? How is it affected by outliers? How is it affected by spacings in the x_i ? How does it compare to alternative methods for doing nonparametric regression? Such assessments are made in a variety of ways, and not only with respect to a single figure of merit, such as integrated squared error. Whether procedurally or model generated, a method must be assessed by its effectiveness.

I find it helpful to think about statistical contributions to astronomy as being arranged along this spectrum as well. At the risk of oversimplification, this can be exemplified by many of the presentations at this meeting. At the procedural end there was the presentation of Cook, showing us some wonderful tools for exploring multivariate data. The presentations of Breiman, Freeman, Murtagh, and Starck showed us some widely applicable procedures that are based on rather minimal modeling structures. Similarly Djogorvsky and Nichols were primarily interested in procedures rather than models. The papers of Shafer, Szalay, Wasserman, and Martinez were primarily at frequentist wavelengths,

while the Bayesian frequency band was occupied by the presentations of Berger, Bretthorst, Connors, van Dyk, Jaffe, Kolaczyk, Loredó, and Scargle.

The presentations of Raftery and Johnstone contained a mix of Bayesian and frequentist perspectives. Johnstone’s exemplified how these two perspectives can enrich each other. In the smoothing spline example above, the Bayesian formulation was rather an afterthought, but in Johnstone’s use of priors on different resolution levels we see the Bayesian formalism being adopted for the purpose of generating smoothing procedures which can be explored from a frequentist perspective or merely viewed as empirical procedures. Clearly, no one would take these priors seriously as quantification of personal belief in a game with a bookie; rather, they are devices that hopefully generate useful methods of averaging (what statistics is all about). The procedures presented by Kolaczyk stemmed from a Bayesian formalism and can be viewed in a similar way—his parameters α and β are quite analogous to the λ for a smoothing spline and I can imagine turning a (α, β) knob to explore differing degrees of smoothing without introspection as to the state of my “belief” about (α, β) .

Statistical models may be viewed as filters through which data are analyzed, and, as William James wrote, “We must be careful not to confuse data with the abstractions we use to analyze them.” But we need these filters/abstractions: as George Box wrote, “All models are wrong, but some are useful.” Models are useful and effective to the degree that they provide a mechanism for accurately extracting information of scientific interest from the data. More elaborate models tend to be more fragile. As one moves towards the Bayesian end of the spectrum, models become more detailed and highly specified, as can be seen in the contributions to this meeting.

We really need to be careful in using models, especially in situations in which there is such a large quantity of data that model accuracy cannot be readily checked visually. Despite extensive effort, “de-glitching” may be incomplete. Even beyond glitches, there may be sources of noise not properly accounted for in the model—do the Gaussian or Poisson variables in the model really reflect all the sources of noise, such as cosmic rays, image motion, and crowding, for example?

For these reasons, robustness has a long and honorable tradition in statistics and is increasingly relevant in this age of data floods. Figure 1 shows light curves in the form of fourth order trigonometric polynomials fit to phased observations at the fundamental frequency (0.46 days) of an RR Lyrae of type “d.” The individual observations of magnitude were accompanied by error bars and, for one particular point, the error bars were far too small. The weighted least squares fits of the light curves are formed in accord with the model, but are very sensitive to deviations from it. The outlying point (just one observation out of 845!) pulls the curve locally and causes global rippling. The robust fitting procedure is designed to do reasonably well if the model holds and to be resistant to outliers.

Even without such phenomena, one needs to be concerned about the biases that are incurred by analyzing the data only through the filters of the model. One needs to ask how crucial the assumptions are. Are the important conclu-

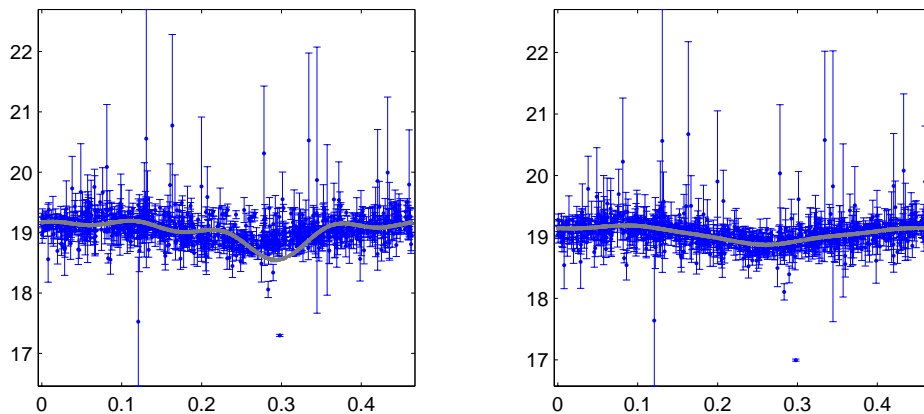


Figure 1: Weighted least squares fit (left panel) and robust fit (right panel) to phased data from an RR Lyrae with period 0.46 days.

sions sensitive to distributional assumptions and assumptions of independence in a frequentist model? In the case of a Bayesian analysis, one should seriously examine the consequences of the choice of a prior. This is not easy for complex hierarchical models and often receives only cursory attention. In his contribution to this meeting Jaffe makes some references to the difficulty of choosing a prior and the influence the prior has on inferences about key cosmological constants. The smoothed lightcurves that Berger displays resulting from priors on wavelet coefficients produce rather suspicious structure precisely in the regions where there is no data (near zero).

3 Challenges and opportunities

The meeting has been very exciting in illustrating many opportunities for application of existing statistical methodology and challenges for the development of new approaches. Let me highlight a few:

Large scale structure: I suspect that there are real opportunities for going beyond two-point and higher order correlation functions for both characterizing structure and discriminating amongst theories. Might not other functionals offer sharper characterization and discrimination? The size of the data, the complexity of coverage patterns, and the presence of selection biases makes this endeavor even more challenging.

Separation of source and background: This problem is omnipresent in astronomical data and we heard about some very interesting developments in the presentations of Freeman and Starck. The problem of removing foreground in studies of the CMB was alluded to by Jaffe. Perhaps because of its high

dimensionality and spatial aspects, this problem does not seem to me to fit very well into our standard paradigms of statistical inference and decision theory. I suspect that some gains can be made by taking more advantage of the fact that the same kind of problem is often faced repeatedly (see the section on empirical Bayes below).

Parameter estimation from massive data sets: Jaffe’s paper gave us a hint of the kinds of problems of this type that will be faced in the near future. How will we meet the corresponding computational challenges? As mentioned by Djorgovski, one possibility is to forgo computing estimates with high precision and/or to forgo notions of statistical optimality (the best may be the enemy of the good). Algorithms derived from the literature on stochastic approximation and on-line gradient methods may turn out to be important. There are close relationships between parameter estimation, coding, information theory and data compression [11]. A sufficient statistic provides marvelous data compression, but these rarely exist. We may need a notion of an “almost sufficient” statistic. Nichol’s use of KD trees is in this spirit. For some current developments on using compression of astronomical data also see [2] and [13]. There has always been a strong interface between statistical inference and computation and the prevalence of massive data sets coupled with the computational power of “the grid” will have profound effects on the nature of the discipline of statistics.

Massive data sets and multivariate analysis: Here we have heard of data sets which would seem to correspond to a multivariate statistician’s dream: enormous n and bounded p , but are we really ready to live out our dreams? The challenges were exhilaratingly described in Djorgovski’s and Strauss’ presentations. The staggering size of the data sets begs for multiscale procedures, for adaptive stratification, for adaptive sequential procedures, and for new methodology.

Although we have heard of some very promising developments from Murtagh and Raftery, I think that there remains a great deal to be done in finding clusters of widely varying morphology and other structures in massive data sets. The complexity and heterogeneity of astronomical data offers further challenges. The structures are likely to be quite different from those encountered in generic market-basket data mining: the strong physical constraints operative in astronomical data and good precision measurements should result in concentrations along low dimensional (nonlinear) manifolds. Local linear embedding [12] and ISOMAP [14] are two interesting recent developments along these lines that may be relevant. Both of these exploit the fact that although nearest neighbors are generically quite distant in high dimensions, they are not if the points lie on relatively low dimensional manifolds. Thus other methods based on nearest neighbors may turn out to be important, too.

How can rare objects be spotted? Can serendipity, so important in the history of astronomy, be automated? This is not just a matter of identifying outliers, although that’s important, too.

Time series analysis: The fascinating irregularly spaced time series found in astronomy have been a stimulus for time series analysis for a long time and challenges remain. The large statistical literature on non-linear time series is rather thin in compelling examples and scientifically plausible analyses and could be enriched and stimulated by confronting such series as those of Miras archived by the AAVSO—see the poster sessions of Foster, Hawkins, and Mattei. Although not discussed in this meeting, I think that the large collections of irregularly spaced time series, such as those of variable stars gathered by microlensing surveys [5] pose methodological challenges for time series.

Empirical Bayes: Astronomers often do the same type of analysis repeatedly: sources are separated from backgrounds, periods and light curves of variable stars are fit, spectra of similar objects are measured. A basic intent of empirical Bayes procedures [4] is to “borrow strength” across objects rather than treating each object *de novo*. Large ensembles of similar objects are being measured and astronomers are often more interested in properties of the ensemble than in the individuals. When interested in better estimating individuals, strength can be borrowed from the ensemble.

The “empirical” in “empirical Bayes” refers to the fact that these procedures attempt to estimate the prior distribution of the ensemble. Estimates of individuals are then constructed using this prior. Suppose one has a noisy measurement of an object of interest: $Y = O + N$, that a collection of templates, T_i , have been empirically constructed for such objects, and that the templates have *a priori* probabilities $P(T_i)$. Then an estimate of the object of interest would be

$$E(O|Y) = \frac{\sum_i T_i P(Y|T_i) P(T_i)}{\sum_i P(Y|T_i) P(T_i)}. \quad (3)$$

To make these notions more concrete, consider an idealized version of the problem of estimating a periodic function from noisy data, where the period is effectively known. The function might be the light curve of a Cepheid, as in the presentation of Berger *et al.* Suppose that there is a whole collection of Cepheid light curves of interest. For simplicity of notation, suppose that there are n time points equally spaced over phase and corresponding observations Y_i . (For a more general setup see [10]). Consider fitting the function as a Fourier series

$$f(x) = \sum_k [A_k \cos(2\pi kx) + B_k \sin(2\pi kx)] \quad (4)$$

where the series is truncated at some point (for simplicity, the mean is taken to be 0). If the measurement errors are modeled as independent with means zero and variances σ_e^2 , the ordinary least squares estimate of a Fourier coefficient is

$$\hat{A}_k = \frac{1}{n} \sum_{j=1}^n Y_j \cos(2\pi kj/n). \quad (5)$$

Taking the point of view that the light curve at hand is drawn from the ensemble, one could estimate A_k by $E(A_k|Y)$, the computation of which would

involve the distribution of A_k over the ensemble. Alternatively, one might consider the best linear approximation to this quantity. In a linear empirical Bayes analysis, the variance parameters σ_e^2 and σ_k^2 are *estimated* from the entire collection of light curves. The linear empirical Bayes estimate of A_k is then

$$E(A_k|Y) = \hat{A}_k \frac{n\sigma_k^2}{n\sigma_k^2 + \sigma_e^2}. \quad (6)$$

The ordinary least squares estimate is thus damped by the ratio of variances, so that high frequency terms with small variances (i.e. those that are typically small) will not contribute much to the estimate of $f()$, especially if n is small. The amount of damping, or tapering, of the Fourier series is determined empirically by the collection of curves at hand.

Contemporary nonparametrics: There is a large literature on nonparametric function estimation: the point of view of this area is that the parameter to be estimated is infinite dimensional, typically a function. Wasserman *et al* gave some examples in their presentation. There has been extensive work on how to choose smoothing parameters automatically.

There has been an explosion of research in high dimensional nonparametric function estimation, discrimination, and clustering, often referred to as “machine learning” in the computer science literature. See the contribution of Breiman to this meeting. Astronomers are generally aware of neural nets and decision trees, but there have been other interesting recent developments, such as support vector machines, bagging, boosting, and graphical models.

There have been developments for semi-parametric estimation [1] in which one is interested in estimating infinite dimensional parameters and scalars. For example, the problem of estimating the period light curve of a variable star can be viewed in this way—the light curve is infinite dimensional and the period is a scalar parameter. See [7] for a detailed analysis.

Model selection: There has been an increasingly explicit recognition in recent statistical literature that models are approximate and not given *a priori*. Rather, there is typically a subtle interplay between data analysis and a set of potential models. Model selection and model averaging [6] are active areas of research in statistics that are likely to have an impact in astronomy. It is interesting that in this meeting we heard about the use of model averaging from two quite different perspectives, those of Berger and Breiman. Of course, the fundamental activity in statistics is figuring out how to average effectively.

Statistical computing: There have been recent interesting developments in statistical computing which may well be useful for astronomers. We heard about ggobi from Diane Cook. I recommend that astronomers also check out two open-source projects: R (<http://www.r-project.org/>) and Omega (<http://www.omegahat.org/>). The computational demands posed by modern astronomy will also hopefully act

as an impetus for further developments in statistical computing. Are we ready to compute on “the grid?”

This list is hardly exhaustive. For wider coverage see the recent collection of very readable vignettes [8] which covers a number of areas of current research in statistics.

In summary, the data revolution in modern astronomy offers a rich feast for statisticians, of whom there are relatively few working in the area. It is essential for statisticians to be open-minded, flexible, and creative since, in the words of Leo Breiman, “To a man who only has a hammer, every problem looks like a nail.” There are lots of problems out there, spread over the entire statistical spectrum, and not many of them are best viewed as nails.

4 Conclusion

Contributions of statistics to the analysis of astronomical data are not likely to be limited to straight technology transfer. The discipline of statistics thrives on being confronted with new problems and there are fundamental perspectives underlying statistical methodology that can hopefully be brought to bear to address the exciting challenges of modern astronomy.

The most important and enduring contributions of statistics to astronomy, and of astronomy to statistics, are likely to flow from long, close collaborations. It takes a long time for a statistician to appreciate the contexts of particular problems, what is really important to the science of astronomy, and thus what statistical approaches will be most fruitful. Similarly it takes a long time for an astronomer to understand the language of statistics and understand the intuition and heuristics underlying contemporary statistical methods. Establishing such collaboration is not easy, however: it takes a great deal of time and patience, and time is scarce in our too-busy professional lives.

There is a strong basis for collaboration. The two disciplines have been linked for centuries during which probabilistic ideas have been central to astronomy. They continue to evolve in parallel: now both are confronting the world of massive data sets. The institution of the National Virtual Observatory [3] will hopefully bring us closer together. I hope that we also have more meetings like this one!

References

- [1] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, 1993.
- [2] J. R. Bond, A.H. Jaffe, and L. Knox. Radical compression of cosmic microwave background data. *Astrophysical Journal*, 533:19–37, 2000. astro-ph/9808264.

- [3] R. J. Brunner, S. G. Djorgovski, and A. S. Szalay, editors. *Virtual Observatories of the Future*, volume 225 of *Astronomical Society of the Pacific Conference Series*. Astronomical Society of the Pacific, 2001.
- [4] B. P. Carlin and T.A. Louis. *Empirical Bayes: Past, Present, and Future*, chapter 4, pages 312–318. Volume 93 of Raftery et al. [8], 2002.
- [5] Roger Ferlet, Jean Pierre Maillard, and Brigitte Raban, editors. *Variable Stars and the Astrophysical Returns of Microlensing Surveys*. Editions Frontieres, 1997.
- [6] E. I. George. *The variable selection problem*, chapter 4, pages 350–358. Volume 93 of Raftery et al. [8], 2002.
- [7] P. Hall, J. Reimann, and J. Rice. Nonparametric estimation of a periodic function. *Biometrika*, 87(3):545–557, 2000.
- [8] A. E. Raftery, M. A. Tanner, and M. T. Wells, editors. *Statistics in the 21st Century*, volume 93 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, 2002.
- [9] C. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10:177–183, 1967.
- [10] J. A. Rice and C. O. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57:253–259, 2001.
- [11] J. Rissanen and B. Yu. *Coding and compression: a happy union of theory and practice*, chapter 43, pages 229–236. Volume 93 of Raftery et al. [8], 2002.
- [12] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000.
- [13] M. Tegmark, A. Taylor, and A. Heavens. Karhunen-loeve eigenvalue problems in cosmology: how should we tackle large data sets? *Astrophysical Journal*, 480:22–35, 1997.
- [14] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.
- [15] G. Wahba. Bayesian “confidence intervals” for the cross-validated smoothing spline. *Journal of the Royal Statistical Society B*, 45:133–150, 1983.
- [16] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF regional conference series in applied mathematics*. SIAM, 1990.
- [17] G. Wahba and S. Wold. A completely automatic french curve: fitting spline functions by cross-validation. *Communications in Statistics*, 4:1–17, 1975.