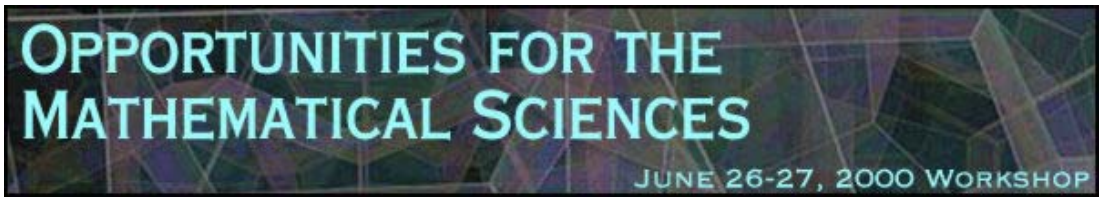

Statistics as the Information Science

P. BICKEL

Statistical models and methods have become prevalent in almost every field in the sciences (physical, life, and social) and engineering. The forces driving this development have been the exponential growth of computing power and the development of innumerable novel devices for generating huge amounts of data rapidly. In the physical sciences we can think of the Hubble space telescope, narrow aperture synthetic radar imaging, satellite photometry; in the life sciences, genetics and the data bases it has spawned and, most recently, gene arrays; in engineering and the social sciences, the Internet. Data is only of interest if we can devise ways of thinking about it, that is, models, implicit or explicit, which enable us to extract useful information and make predictions. In the context of our examples, we want to understand more about the structure of the universe, the topography of the Earth, the role of different proteins in cellular processes, linking between genotype and phenotype to devise better diagnostics for diseases, improving search engines, etc. Common features of much of current data are:

i. Size: Terabytes

ii. Complex structures: Complexity in relations between different parts of the data and the nature of the data itself. Examples might include images, three-dimensional objects as produced for instance by models of the atmosphere or ocean, and four-dimensional temporal series of three-dimensional objects.

iii. Noisiness: Most of the measurement processes we have discussed are inherently subject to random fluctuations.

iv. Imperfect understanding of the basic processes about which we seek information.

v. Indirect observation of what is desired. For instance, gene correlations are observed, not their actual interactions.

Complexity, noisiness and imperfect understanding naturally lead to the construction of statistical models for what is observed.

There are two principal aspects of the use of statistical models, exploratory and confirmatory. In the exploratory phase, models are used to describe various features of the data that may be important for understanding. For instance, hidden Markov models are fit to genetic sequences to identify common features of different sequences and patterns in single sequences. Confirmatory use is the attachment of probabilities to the features observed to assure ourselves that what we are seeing is not due merely to chance. Both aspects are critical. We need to have things drawn to our attention before being able to do anything. But following things up is usually expensive: laboratory experiments have to be carried out, new equipment has to be designed. In the current data environment, models applied to huge and complex data sets lead to all sorts of "interesting" findings, many of which are spurious. A different kind of example pointing to the difficulty of confirmatory analysis is deciding whether the evidence, statistical or otherwise, for global warming is strong enough to take action in some direction.

These questions have always been present, and basic approaches and mathematical techniques have been developed by the developers of probability theory starting with Gauss and Laplace and including Kolmogorov, Levy, Feller, and other important figures in the past century, as well as the founders of modern statistics, R.A. Fisher, J. Neyman, A. Wald, and others. These techniques continue

to have an impact as they are introduced in new arenas, but there has been a phase change which calls for rethinking old ideas and an infusion of new ones.

What has changed is the scale of the problems as well as their internal complexity.

1. Stochastic models have to be studied for very exotic spaces of objects, trees (phylogenetic or other) and other complex combinatorial structures, images, and the three- and four-dimensional types of objects we mentioned earlier.

2. Accounting for extensive searches within classes of procedures is a major issue. We have already mentioned this issue in connection with pattern finding, e.g., gene hunting or hunting for functionally important sites in proteins. Situations where more theory is available are applications of machine learning, or equivalently, nonparametric classification and regression. Complicated classification rules (algorithms) are developed on the basis of a "training set" of data. For instance, computer speech recognition algorithms are trained on the basis of large bodies of speech of one or more speakers. These algorithms are based on many pieces, each complex: hidden Markov models, neural nets, quantization. How do we predict in advance the performance of these algorithms on new speakers? Measuring performance on the training set can be highly misleading since the available number of parameters of the algorithms are such that, in principle, recognition can be made almost perfect. But this typically leads to disastrously poor performance on the test (new) sample, a phenomenon known as "overfitting." Good measures of performance require probabilistic analyses of the algorithms. These are important not only for realistic performance measurement but also for qualitative understanding of the many types of procedures which have been developed. There has been considerable mathematical work in this area (bounds of Vapnik-Chervonenkis type, minmax results) but the bounds obtained are unrealistic and qualitative understanding is frequently spotty.

3. Computer models have arisen and are being developed in many fields, ranging from the atmospheric sciences to transportation to the immune system. The models are sometimes deterministic (the atmosphere), sometimes stochastic (parts of the immune system), and sometimes both (transportation). Often, as in the big atmosphere models, they consist of enormous numbers of coupled partial differential equations solved numerically with inputs or boundary conditions involving noisy data. Validation of these models is largely not performed, or is an art.

4. These problems pose important joint challenges to applied mathematics, numerical analysis, computer science, and statistics. The scale of some of these models, which may run for many days on the most powerful machines, leads to intrinsic compromises. How does one allocate resources between replication of experiments, constructing better algorithms, and building in improvements to basic theory in refining the scale of the models? To make these compromises intelligently involves studying the interaction of statistical uncertainties coming from the data, with numerical error, with model uncertainty, with computational speed, all in the light of what the model is going to be used for. The theory of this type of combination of error analyses still has to be developed and should be done in an interdisciplinary fashion. These questions are becoming more and more central as experimentation in many areas, if not impossible, is ruinously expensive, analytic approximations are largely impossible, and computing is ever cheaper.

5. Computing power has led to the development of many Monte Carlo based techniques for situations where analytic approximation is impossible. For instance, Markov chain Monte Carlo and related techniques, originally introduced in physics, enable the implementation of Bayesian and other analyses in models which are intrinsically designed to take into account the indirect partial nature of most observation. Similarly, "particle filters" are a very promising approach of dealing with a ubiquitous and extremely useful class of models generally referred to as state space or hidden Markov models. The mathematical properties of these algorithms are far from understood.

6. As a recent NRC sponsored workshop on the interface of computer science and the mathematical sciences illustrated, stochastic models enter in many areas: network traffic modeling, computer vision, and "data mining," which is essentially a statistical enterprise. In the first area, scaling laws, which appear phenomenologically, call out for stochastic analysis. In the second area nonparametric statistical models compete with PDE-based models for various aspects of the tasks of object recognition and texture generation.

To sum up, statistical models are at the center of development of a large number of key fields in the sciences, engineering, and public policy. The change in scale and complexity of the types of data and phenomena being studied in all fields poses new mathematical and conceptual challenges.

Last Modified:
Oct 17, 2001

Previous page | Top of this page | Next page