of the broader objectives are suspect in the present context. In a sense these are efforts to establish data descriptions that are potentially causal, recognizing that causality, in the sense that a natural scientist would use the term, can rarely be established from one type of study and is at best somewhat tentative.

Professor Breiman takes a rather defeatist attitude toward attempts to formulate underlying processes; is this not to reject the base of much scientific progress? The interesting illustrations given by Beveridge (1952), where hypothesized processes in various biological contexts led to important progress, even though the hypotheses turned out in the end to be quite false, illustrate the subtlety of the matter. Especially in the social sciences, representations of underlying process have to be viewed with particular caution, but this does not make them fruitless.

The absolutely crucial issue in serious mainstream statistics is the choice of a model that will translate key subject-matter questions into a form for analysis and interpretation. If a simple standard model is adequate to answer the subject-matter question, this is fine: there are severe hidden penalties for overelaboration. The statistical literature, however, concentrates on how to do the analysis, an important and indeed fascinating question, but a secondary step. Better a rough answer to the right question than an exact answer to the wrong question, an aphorism, due perhaps to Lord Kelvin, that I heard as an undergraduate in applied mathematics.

I have stayed away from the detail of the paper but will comment on just one point, the interesting theorem of Vapnik about complete separation. This confirms folklore experience with empirical logistic regression that, with a largish number of explanatory variables, complete separation is quite likely to occur. It is interesting that in mainstream thinking this is, I think, regarded as insecure in that complete separation is thought to be a priori unlikely and the estimated separating plane unstable. Presumably bootstrap and cross-validation ideas may give here a quite misleading illusion of stability. Of course if the complete separator is subtle and stable Professor Breiman's methods will emerge triumphant and ultimately it is an empirical question in each application as to what happens.

It will be clear that while I disagree with the main thrust of Professor Breiman's paper I found it stimulating and interesting.

Comment

Brad Efron

At first glance Leo Breiman's stimulating paper looks like an argument against parsimony and scientific insight, and in favor of black boxes with lots of knobs to twiddle. At second glance it still looks that way, but the paper *is* stimulating, and Leo has some important points to hammer home. At the risk of distortion I will try to restate one of those points, the most interesting one in my opinion, using less confrontational and more historical language.

From the point of view of statistical development the twentieth century might be labeled "100 years of unbiasedness." Following Fisher's lead, most of our current statistical theory and practice revolves around unbiased or nearly unbiased estimates (particularly MLEs), and tests based on such estimates. The power of this theory has made statistics the dominant interpretational methodology in dozens of fields, but, as we say in California these days, it is power purchased at a price: the theory requires a modestly high ratio of signal to noise, sample size to number of unknown parameters, to have much hope of success. "Good experimental design" amounts to enforcing favorable conditions for unbiased estimation and testing, so that the statistician won't find himself or herself facing 100 data points and 50 parameters.

Now it is the twenty-first century when, as the paper reminds us, we are being asked to face problems that never heard of good experimental design. Sample sizes have swollen alarmingly while goals grow less distinct ("find interesting data structure"). New algorithms have arisen to deal with new problems, a healthy sign it seems to me even if the innovators aren't all professional statisticians. There are enough physicists to handle the physics case load, but there are fewer statisticians and more statistics problems, and we need all the help we can get. An

Brad Efron is Professor, Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford University, Stanford, California 94305–4065 (e-mail: brad@stat.stanford.edu).

attractive feature of Leo's paper is his openness to new ideas whatever their source.

The new algorithms often appear in the form of black boxes with enormous numbers of adjustable parameters ("knobs to twiddle"), sometimes more knobs than data points. These algorithms can be quite successful as Leo points outs, sometimes more so than their classical counterparts. However, unless the bias-variance trade-off has been suspended to encourage new statistical industries, their success must hinge on some form of biased estimation. The bias may be introduced directly as with the "regularization" of overparameterized linear models, more subtly as in the pruning of overgrown regression trees, or surreptitiously as with support vector machines, but it has to be lurking somewhere inside the theory.

Of course the trouble with biased estimation is that we have so little theory to fall back upon. Fisher's information bound, which tells us how well a (nearly) unbiased estimator can possibly perform, is of no help at all in dealing with heavily biased methodology. Numerical experimentation by itself, unguided by theory, is prone to faddish wandering:

Rule 1. New methods always look better than old ones. Neural nets are better than logistic regression, support vector machines are better than neural nets, etc. In fact it is very difficult to run an honest simulation comparison, and easy to inadvertently cheat by choosing favorable examples, or by not putting as much effort into optimizing the dull old standard as the exciting new challenger.

Rule 2. Complicated methods are harder to criticize than simple ones. By now it is easy to check the efficiency of a logistic regression, but it is no small matter to analyze the limitations of a support vector machine. One of the best things statisticians do, and something that doesn't happen outside our profession, is clarify the inferential basis of a proposed new methodology, a nice recent example being Friedman, Hastie, and Tibshirani's analysis of "boosting," (2000). The past half-century has seen the clarification process successfully at work on nonparametrics, robustness and survival analysis. There has even been some success with biased estimation in the form of Stein shrinkage and empirical Bayes, but I believe the hardest part of this work remains to be done. Papers like Leo's are a call for more analysis and theory, not less.

Prediction is certainly an interesting subject but Leo's paper overstates both its role and our profession's lack of interest in it.

- The "prediction culture," at least around Stanford, is a lot bigger than 2%, though its constituency changes and most of us wouldn't welcome being typecast.
- Estimation and testing are a form of prediction: "In our sample of 20 patients drug A outperformed drug B; would this still be true if we went on to test all possible patients?"
- Prediction by itself is only occasionally sufficient. The post office is happy with any method that predicts correct addresses from hand-written scrawls. Peter Gregory undertook his study for prediction purposes, but also to better understand the medical basis of hepatitis. Most statistical surveys have the identification of causal factors as their ultimate goal.

The hepatitis data was first analyzed by Gail Gong in her 1982 Ph.D. thesis, which concerned prediction problems and bootstrap methods for improving on cross-validation. (Cross-validation itself is an uncertain methodology that deserves further critical scrutiny; see, for example, Efron and Tibshirani, 1996). The *Scientific American* discussion is quite brief, a more thorough description appearing in Efron and Gong (1983). Variables 12 or 17 (13 or 18 in Efron and Gong's numbering) appeared as "important" in 60% of the bootstrap simulations, which might be compared with the 59% for variable 19, the most for any single explanator.

In what sense are variable 12 or 17 or 19 "important" or "not important"? This is the kind of interesting inferential question raised by prediction methodology. Tibshirani and I made a stab at an answer in our 1998 *annals* paper. I believe that the current interest in statistical prediction will eventually invigorate traditional inference, not eliminate it.

A third front seems to have been opened in the long-running frequentist-Bayesian wars by the advocates of algorithmic prediction, who don't really believe in any inferential school. Leo's paper is at its best when presenting the successes of algorithmic modeling, which comes across as a positive development for both statistical practice and theoretical innovation. This isn't an argument against traditional data modeling any more than splines are an argument against polynomials. The whole point of science is to open up black boxes, understand their insides, and build better boxes for the purposes of mankind. Leo himself is a notably successful scientist, so we can hope that the present paper was written more as an advocacy device than as the confessions of a born-again black boxist.