Statistical Science
2001, Vol. 16, No. 3, 199–231

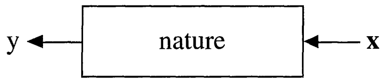
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

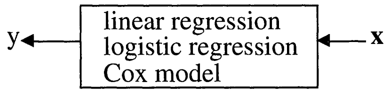
There are two different approaches toward these goals:

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

$$\text{response variables} = f(\text{predictor variables, random noise, parameters})$$

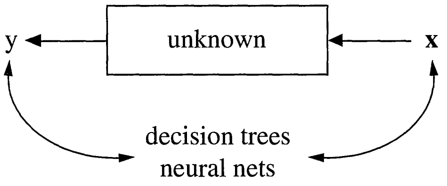
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;

Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720-4735 (e-mail: leo@stat.berkeley.edu).

- Kept statisticians from using more suitable algorithmic models;
- Prevented statisticians from working on exciting new problems;

I will also review some of the interesting new developments in algorithmic modeling in machine learning and look at applications to three data sets.

2. ROAD MAP

It may be revealing to understand how I became a member of the small second culture. After a seven-year stint as an academic probabilist, I resigned and went into full-time free-lance consulting. After thirteen years of consulting I joined the Berkeley Statistics Department in 1980 and have been there since. My experiences as a consultant formed my views about algorithmic modeling. Section 3 describes two of the projects I worked on. These are given to show how my views grew from such problems.

When I returned to the university and began reading statistical journals, the research was distant from what I had done as a consultant. All articles begin and end with data models. My observations about published theoretical research in statistics are in Section 4.

Data modeling has given the statistics field many successes in analyzing data and getting information about the mechanisms producing the data. But there is also misuse leading to questionable conclusions about the underlying mechanism. This is reviewed in Section 5. Following that is a discussion (Section 6) of how the commitment to data modeling has prevented statisticians from entering new scientific and commercial fields where the data being gathered is not suitable for analysis by data models.

In the past fifteen years, the growth in algorithmic modeling applications and methodology has been rapid. It has occurred largely outside statistics in a new community—often called machine learning—that is mostly young computer scientists (Section 7). The advances, particularly over the last five years, have been startling. Three of the most important changes in perception to be learned from these advances are described in Sections 8, 9, and 10, and are associated with the following names:

Rashomon: the multiplicity of good models;
Occam: the conflict between simplicity and accuracy;
Bellman: dimensionality—curse or blessing?

Section 11 is titled "Information from a Black Box" and is important in showing that an algorithmic model can produce more and more reliable information about the structure of the relationship between inputs and outputs than data models. This is illustrated using two medical data sets and a genetic data set. A glossary at the end of the paper explains terms that not all statisticians may be familiar with.

3. PROJECTS IN CONSULTING

As a consultant I designed and helped supervise surveys for the Environmental Protection Agency (EPA) and the state and federal court systems. Controlled experiments were designed for the EPA, and I analyzed traffic data for the U.S. Department of Transportation and the California Transportation Department. Most of all, I worked on a diverse set of prediction projects. Here are some examples:

Predicting next-day ozone levels.
Using mass spectra to identify halogen-containing compounds.
Predicting the class of a ship from high altitude radar returns.
Using sonar returns to predict the class of a submarine.
Identity of hand-sent Morse Code.
Toxicity of chemicals.
On-line prediction of the cause of a freeway traffic breakdown.
Speech recognition
The sources of delay in criminal trials in state court systems.

To understand the nature of these problems and the approaches taken to solve them, I give a fuller description of the first two on the list.

3.1 The Ozone Project

In the mid- to late 1960s ozone levels became a serious health problem in the Los Angeles Basin. Three different alert levels were established. At the highest, all government workers were directed not to drive to work, children were kept off playgrounds and outdoor exercise was discouraged.

The major source of ozone at that time was automobile tailpipe emissions. These rose into the low atmosphere and were trapped there by an inversion layer. A complex chemical reaction, aided by sunlight, cooked away and produced ozone two to three hours after the morning commute hours. The alert warnings were issued in the morning, but would be more effective if they could be issued 12 hours in advance. In the mid-1970s, the EPA funded a large effort to see if ozone levels could be accurately predicted 12 hours in advance.

Commuting patterns in the Los Angeles Basin are regular, with the total variation in any given daylight hour varying only a few percent from one weekday to another. With the total amount of emissions about constant, the resulting ozone levels depend on the meteorology of the preceding days. A large data base was assembled consisting of lower and upper air measurements at U.S. weather stations as far away as Oregon and Arizona, together with hourly readings of surface temperature, humidity, and wind speed at the dozens of air pollution stations in the Basin and nearby areas.

Altogether, there were daily and hourly readings of over 450 meteorological variables for a period of seven years, with corresponding hourly values of ozone and other pollutants in the Basin. Let $\mathbf{x}$ be the predictor vector of meteorological variables on the nth day. There are more than 450 variables in $\mathbf{x}$ since information several days back is included. Let y be the ozone level on the $(n+1)$st day. Then the problem was to construct a function $f(\mathbf{x})$ such that for any future day and future predictor variables $\mathbf{x}$ for that day, $f(\mathbf{x})$ is an accurate predictor of the next day's ozone level y.

To estimate predictive accuracy, the first five years of data were used as the training set. The last two years were set aside as a test set. The algorithmic modeling methods available in the pre-1980s decades seem primitive now. In this project large linear regressions were run, followed by variable selection. Quadratic terms in, and interactions among, the retained variables were added and variable selection used again to prune the equations. In the end, the project was a failure—the false alarm rate of the final predictor was too high. I have regrets that this project can't be revisited with the tools available today.

3.2 The Chlorine Project

The EPA samples thousands of compounds a year and tries to determine their potential toxicity. In the mid-1970s, the standard procedure was to measure the mass spectra of the compound and to try to determine its chemical structure from its mass spectra.

Measuring the mass spectra is fast and cheap. But the determination of chemical structure from the mass spectra requires a painstaking examination by a trained chemist. The cost and availability of enough chemists to analyze all of the mass spectra produced daunted the EPA. Many toxic compounds contain halogens. So the EPA funded a project to determine if the presence of chlorine in a compound could be reliably predicted from its mass spectra.

Mass spectra are produced by bombarding the compound with ions in the presence of a magnetic field. The molecules of the compound split and the lighter fragments are bent more by the magnetic field than the heavier. Then the fragments hit an absorbing strip, with the position of the fragment on the strip determined by the molecular weight of the fragment. The intensity of the exposure at that position measures the frequency of the fragment. The resultant mass spectra has numbers reflecting frequencies of fragments from molecular weight 1 up to the molecular weight of the original compound. The peaks correspond to frequent fragments and there are many zeroes. The available data base consisted of the known chemical structure and mass spectra of 30,000 compounds.

The mass spectrum predictor vector $\mathbf{x}$ is of variable dimensionality. Molecular weight in the data base varied from 30 to over 10,000. The variable to be predicted is

$$y = 1\text{: contains chlorine,}$$

$$y = 2\text{: does not contain chlorine.}$$

The problem is to construct a function $f(\mathbf{x})$ that is an accurate predictor of y where $\mathbf{x}$ is the mass spectrum of the compound.

To measure predictive accuracy the data set was randomly divided into a 25,000 member training set and a 5,000 member test set. Linear discriminant analysis was tried, then quadratic discriminant analysis. These were difficult to adapt to the variable dimensionality. By this time I was thinking about decision trees. The hallmarks of chlorine in mass spectra were researched. This domain knowledge was incorporated into the decision tree algorithm by the design of the set of 1,500 yes–no questions that could be applied to a mass spectra of any dimensionality. The result was a decision tree that gave 95% accuracy on both chlorines and nonchlorines (see Breiman, Friedman, Olshen and Stone, 1984).

3.3 Perceptions on Statistical Analysis

As I left consulting to go back to the university, these were the perceptions I had about working with data to find answers to problems:

(a) Focus on finding a good solution—that's what consultants get paid for.

(b) Live with the data before you plunge into modeling.

(c) Search for a model that gives a good solution, either algorithmic or data.

(d) Predictive accuracy on test sets is the criterion for how good the model is.

(e) Computers are an indispensable partner.

4. RETURN TO THE UNIVERSITY

I had one tip about what research in the university was like. A friend of mine, a prominent statistician from the Berkeley Statistics Department, visited me in Los Angeles in the late 1970s. After I described the decision tree method to him, his first question was, "What's the model for the data?"

4.1 Statistical Research

Upon my return, I started reading the *Annals of Statistics*, the flagship journal of theoretical statistics, and was bemused. Every article started with

> Assume that the data are generated by the following model: ...

followed by mathematics exploring inference, hypothesis testing and asymptotics. There is a wide spectrum of opinion regarding the usefulness of the theory published in the *Annals of Statistics* to the field of statistics as a science that deals with data. I am at the very low end of the spectrum. Still, there have been some gems that have combined nice theory and significant applications. An example is wavelet theory. Even in applications, data models are universal. For instance, in the *Journal of the American Statistical Association* (*JASA*), virtually every article contains a statement of the form:

> Assume that the data are generated by the following model: ...

I am deeply troubled by the current and past use of data models in applications, where quantitative conclusions are drawn and perhaps policy decisions made.

5. THE USE OF DATA MODELS

Statisticians in applied research consider data modeling as the template for statistical analysis: Faced with an applied problem, think of a data model. This enterprise has at its heart the belief that a statistician, by imagination and by looking at the data, can invent a reasonably good parametric class of models for a complex mechanism devised by nature. Then parameters are estimated and conclusions are drawn. But when a model is fit to data to draw quantitative conclusions:

- The conclusions are about the model's mechanism, and not about nature's mechanism.

It follows that:

- If the model is a poor emulation of nature, the conclusions may be wrong.

These truisms have often been ignored in the enthusiasm for fitting data models. A few decades ago, the commitment to data models was such that even simple precautions such as residual analysis or goodness-of-fit tests were not used. The belief in the infallibility of data models was almost religious. It is a strange phenomenon—once a model is made, then it becomes truth and the conclusions from it are infallible.

5.1 An Example

I illustrate with a famous (also infamous) example: assume the data is generated by independent draws from the model

$$\text{(R)} \qquad y = b_0 + \sum_1^M b_m x_m + \varepsilon,$$

where the coefficients $\{b_m\}$ are to be estimated, ε is $N(0, \sigma^2)$ and σ^2 is to be estimated. Given that the data is generated this way, elegant tests of hypotheses, confidence intervals, distributions of the residual sum-of-squares and asymptotics can be derived. This made the model attractive in terms of the mathematics involved. This theory was used both by academic statisticians and others to derive significance levels for coefficients on the basis of model (R), with little consideration as to whether the data on hand could have been generated by a linear model. Hundreds, perhaps thousands of articles were published claiming proof of something or other because the coefficient was significant at the 5% level.

Goodness-of-fit was demonstrated mostly by giving the value of the multiple correlation coefficient R^2 which was often closer to zero than one and which could be over inflated by the use of too many parameters. Besides computing R^2, nothing else was done to see if the observational data could have been generated by model (R). For instance, a study was done several decades ago by a well-known member of a university statistics department to assess whether there was gender discrimination in the salaries of the faculty. All personnel files were examined and a data base set up which consisted of salary as the response variable and 25 other variables which characterized academic performance; that is, papers published, quality of journals published in, teaching record, evaluations, etc. Gender appears as a binary predictor variable.

A linear regression was carried out on the data and the gender coefficient was significant at the 5% level. That this was strong evidence of sex discrimination was accepted as gospel. The design of the study raises issues that enter before the consideration of a model—Can the data gathered answer the question posed? Is inference justified when your sample is the entire population? Should a data model be used? The deficiencies in analysis occurred because the focus was on the model and not on the problem.

The linear regression model led to many erroneous conclusions that appeared in journal articles waving the 5% significance level without knowing whether the model fit the data. Nowadays, I think most statisticians will agree that this is a suspect way to arrive at conclusions. At the time, there were few objections from the statistical profession about the fairy-tale aspect of the procedure, But, hidden in an elementary textbook, Mosteller and Tukey (1977) discuss many of the fallacies possible in regression and write "The whole area of guided regression is fraught with intellectual, statistical, computational, and subject matter difficulties."

Even currently, there are only rare published critiques of the uncritical use of data models. One of the few is David Freedman, who examines the use of regression models (1994); the use of path models (1987) and data modeling (1991, 1995). The analysis in these papers is incisive.

5.2 Problems in Current Data Modeling

Current applied practice is to check the data model fit using goodness-of-fit tests and residual analysis. At one point, some years ago, I set up a simulated regression problem in seven dimensions with a controlled amount of nonlinearity. Standard tests of goodness-of-fit did not reject linearity until the nonlinearity was extreme. Recent theory supports this conclusion. Work by Bickel, Ritov and Stoker (2001) shows that goodness-of-fit tests have very little power unless the direction of the alternative is precisely specified. The implication is that omnibus goodness-of-fit tests, which test in many directions simultaneously, have little power, and will not reject until the lack of fit is extreme.

Furthermore, if the model is tinkered with on the basis of the data, that is, if variables are deleted or nonlinear combinations of the variables added, then goodness-of-fit tests are not applicable. Residual analysis is similarly unreliable. In a discussion after a presentation of residual analysis in a seminar at Berkeley in 1993, William Cleveland, one of the fathers of residual analysis, admitted that it could not uncover lack of fit in more than four to five dimensions. The papers I have read on using residual analysis to check lack of fit are confined to data sets with two or three variables.

With higher dimensions, the interactions between the variables can produce passable residual plots for a variety of models. A residual plot is a goodness-of-fit test, and lacks power in more than a few dimensions. An acceptable residual plot does not imply that the model is a good fit to the data.

There are a variety of ways of analyzing residuals. For instance, Landwher, Preibon and Shoemaker (1984, with discussion) gives a detailed analysis of fitting a logistic model to a three-variable data set using various residual plots. But each of the four discussants present other methods for the analysis. One is left with an unsettled sense about the arbitrariness of residual analysis.

Misleading conclusions may follow from data models that pass goodness-of-fit tests and residual checks. But published applications to data often show little care in checking model fit using these methods or any other. For instance, many of the current application articles in *JASA* that fit data models have very little discussion of how well their model fits the data. The question of how well the model fits the data is of secondary importance compared to the construction of an ingenious stochastic model.

5.3 The Multiplicity of Data Models

One goal of statistics is to extract information from the data about the underlying mechanism producing the data. The greatest plus of data modeling is that it produces a simple and understandable picture of the relationship between the input variables and responses. For instance, logistic regression in classification is frequently used because it produces a linear combination of the variables with weights that give an indication of the variable importance. The end result is a simple picture of how the prediction variables affect the response variable plus confidence intervals for the weights. Suppose two statisticians, each one with a different approach to data modeling, fit a model to the same data set. Assume also that each one applies standard goodness-of-fit tests, looks at residuals, etc., and is convinced that their model fits the data. Yet the two models give different pictures of nature's mechanism and lead to different conclusions.

McCullah and Nelder (1989) write "Data will often point with almost equal emphasis on several possible models, and it is important that the statistician recognize and accept this." Well said, but different models, all of them equally good, may give different pictures of the relation between the predictor and response variables. The question of which one most accurately reflects the data is difficult to resolve. One reason for this multiplicity is that goodness-of-fit tests and other methods for checking fit give a yes–no answer. With the lack of power of these tests with data having more than a small number of dimensions, there will be a large number of models whose fit is acceptable. There is no way, among the yes–no methods for gauging fit, of determining which is the better model. A few statisticians know this. Mountain and Hsiao (1989) write, "It is difficult to formulate a comprehensive model capable of encompassing all rival models. Furthermore, with the use of finite samples, there are dubious implications with regard to the validity and power of various encompassing tests that rely on asymptotic theory."

Data models in current use may have more damaging results than the publications in the social sciences based on a linear regression analysis. Just as the 5% level of significance became a de facto standard for publication, the Cox model for the analysis of survival times and logistic regression for survive–nonsurvive data have become the de facto standard for publication in medical journals. That different survival models, equally well fitting, could give different conclusions is not an issue.

5.4 Predictive Accuracy

The most obvious way to see how well the model box emulates nature's box is this: put a case $\mathbf{x}$ down nature's box getting an output y. Similarly, put the same case $\mathbf{x}$ down the model box getting an output y'. The closeness of y and y' is a measure of how good the emulation is. For a data model, this translates as: fit the parameters in your model by using the data, then, using the model, predict the data and see how good the prediction is.

Prediction is rarely perfect. There are usually many unmeasured variables whose effect is referred to as "noise." But the extent to which the model box emulates nature's box is a measure of how well our model can reproduce the natural phenomenon producing the data.

McCullagh and Nelder (1989) in their book on generalized linear models also think the answer is obvious. They write, "At first sight it might seem as though a good model is one that fits the data very well; that is, one that makes $\hat{\mu}$ (the model predicted value) very close to y (the response value)." Then they go on to note that the extent of the agreement is biased by the number of parameters used in the model and so is not a satisfactory measure. They are, of course, right. If the model has too many parameters, then it may overfit the data and give a biased estimate of accuracy. But there are ways to remove the bias. To get a more unbiased estimate of predictive accuracy, cross-validation can be used, as advocated in an important early work by Stone (1974). If the data set is larger, put aside a test set.

Mosteller and Tukey (1977) were early advocates of cross-validation. They write, "Cross-validation is a natural route to the indication of the quality of any data-derived quantity.... We plan to cross-validate carefully wherever we can."

Judging by the infrequency of estimates of predictive accuracy in *JASA*, this measure of model fit that seems natural to me (and to Mosteller and Tukey) is not natural to others. More publication of predictive accuracy estimates would establish standards for comparison of models, a practice that is common in machine learning.

6. THE LIMITATIONS OF DATA MODELS

With the insistence on data models, multivariate analysis tools in statistics are frozen at discriminant analysis and logistic regression in classification and multiple linear regression in regression. Nobody really believes that multivariate data is multivariate normal, but that data model occupies a large number of pages in every graduate textbook on multivariate statistical analysis.

With data gathered from uncontrolled observations on complex systems involving unknown physical, chemical, or biological mechanisms, the a priori assumption that nature would generate the data through a parametric model selected by the statistician can result in questionable conclusions that cannot be substantiated by appeal to goodness-of-fit tests and residual analysis. Usually, simple parametric models imposed on data generated by complex systems, for example, medical data, financial data, result in a loss of accuracy and information as compared to algorithmic models (see Section 11).

There is an old saying "If all a man has is a hammer, then every problem looks like a nail." The trouble for statisticians is that recently some of the problems have stopped looking like nails. I conjecture that the result of hitting this wall is that more complicated data models are appearing in current published applications. Bayesian methods combined with Markov Chain Monte Carlo are cropping up all over. This may signify that as data becomes more complex, the data models become more cumbersome and are losing the advantage of presenting a simple and clear picture of nature's mechanism.

Approaching problems by looking for a data model imposes an a priori straight jacket that restricts the ability of statisticians to deal with a wide range of statistical problems. The best available solution to a data problem might be a data model; then again it might be an algorithmic model. The data and the problem guide the solution. To solve a wider range of data problems, a larger set of tools is needed.

Perhaps the damaging consequence of the insistence on data models is that statisticians have ruled themselves out of some of the most interesting and challenging statistical problems that have arisen out of the rapidly increasing ability of computers to store and manipulate data. These problems are increasingly present in many fields, both scientific and commercial, and solutions are being found by nonstatisticians.

7. ALGORITHMIC MODELING

Under other names, algorithmic modeling has been used by industrial statisticians for decades. See, for instance, the delightful book "*Fitting Equations to Data*" (Daniel and Wood, 1971). It has been used by psychometricians and social scientists. Reading a preprint of Gifi's book (1990) many years ago uncovered a kindred spirit. It has made small inroads into the analysis of medical data starting with Richard Olshen's work in the early 1980s. For further work, see Zhang and Singer (1999). Jerome Friedman and Grace Wahba have done pioneering work on the development of algorithmic methods. But the list of statisticians in the algorithmic modeling business is short, and applications to data are seldom seen in the journals. The development of algorithmic methods was taken up by a community outside statistics.

7.1 A New Research Community

In the mid-1980s two powerful new algorithms for fitting data became available: neural nets and decision trees. A new research community using these tools sprang up. Their goal was predictive accuracy. The community consisted of young computer scientists, physicists and engineers plus a few aging statisticians. They began using the new tools in working on complex prediction problems where it was obvious that data models were not applicable: speech recognition, image recognition, nonlinear time series prediction, handwriting recognition, prediction in financial markets.

Their interests range over many fields that were once considered happy hunting grounds for statisticians and have turned out thousands of interesting research papers related to applications and methodology. A large majority of the papers analyze real data. The criterion for any model is what is the predictive accuracy. An idea of the range of research of this group can be got by looking at the *Proceedings of the Neural Information Processing Systems Conference* (their main yearly meeting) or at the *Machine Learning Journal*.

7.2 Theory in Algorithmic Modeling

Data models are rarely used in this community. The approach is that nature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknowable. What is observed is a set of $\mathbf{x}$'s that go in and a subsequent set of $\mathbf{y}$'s that come out. The problem is to find an algorithm $f(\mathbf{x})$ such that for future $\mathbf{x}$ in a test set, $f(\mathbf{x})$ will be a good predictor of $\mathbf{y}$.

The theory in this field shifts focus from data models to the properties of algorithms. It characterizes their "strength" as predictors, convergence if they are iterative, and what gives them good predictive accuracy. The one assumption made in the theory is that the data is drawn i.i.d. from an unknown multivariate distribution.

There is isolated work in statistics where the focus is on the theory of the algorithms. Grace Wahba's research on smoothing spline algorithms and their applications to data (using cross-validation) is built on theory involving reproducing kernels in Hilbert Space (1990). The final chapter of the CART book (Breiman et al., 1984) contains a proof of the asymptotic convergence of the CART algorithm to the Bayes risk by letting the trees grow as the sample size increases. There are others, but the relative frequency is small.

Theory resulted in a major advance in machine learning. Vladimir Vapnik constructed informative bounds on the generalization error (infinite test set error) of classification algorithms which depend on the "capacity" of the algorithm. These theoretical bounds led to support vector machines (see Vapnik, 1995, 1998) which have proved to be more accurate predictors in classification and regression then neural nets, and are the subject of heated current research (see Section 10).

My last paper "Some infinity theory for tree ensembles" (Breiman, 2000) uses a function space analysis to try and understand the workings of tree ensemble methods. One section has the heading, "My kingdom for some good theory." There is an effective method for forming ensembles known as "boosting," but there isn't any finite sample size theory that tells us why it works so well.

7.3 Recent Lessons

The advances in methodology and increases in predictive accuracy since the mid-1980s that have occurred in the research of machine learning has been phenomenal. There have been particularly exciting developments in the last five years. What has been learned? The three lessons that seem most important to one:

Rashomon: the multiplicity of good models;
Occam: the conflict between simplicity and accuracy;
Bellman: dimensionality—curse or blessing.

8. RASHOMON AND THE MULTIPLICITY OF GOOD MODELS

Rashomon is a wonderful Japanese movie in which four people, from different vantage points, witness an incident in which one person dies and another is supposedly raped. When they come to testify in court, they all report the same facts, but their stories of what happened are very different.

What I call the Rashomon Effect is that there is often a multitude of different descriptions [equations $f(\mathbf{x})$] in a class of functions giving about the same minimum error rate. The most easily understood example is subset selection in linear regression. Suppose there are 30 variables and we want to find the best five variable linear regressions. There are about 140,000 five-variable subsets in competition. Usually we pick the one with the lowest residual sum-of-squares (RSS), or, if there is a test set, the lowest test error. But there may be (and generally are) many five-variable equations that have RSS within 1.0% of the lowest RSS (see Breiman, 1996a). The same is true if test set error is being measured.

So here are three possible pictures with RSS or test set error within 1.0% of each other:

Picture 1

$$y = 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12} - 2.1x_{17} + 3.2x_{27},$$

Picture 2

$$y = -8.9 + 4.6x_5 + 0.01x_6 + 12.0x_{15} + 17.5x_{21} + 0.2x_{22},$$

Picture 3

$$y = -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 + 3.4x_{11} + 7.2x_{28}.$$

Which one is better? The problem is that each one tells a different story about which variables are important.

The Rashomon Effect also occurs with decision trees and neural nets. In my experiments with trees, if the training set is perturbed only slightly, say by removing a random 2–3% of the data, I can get a tree quite different from the original but with almost the same test set error. I once ran a small neural net 100 times on simple three-dimensional data reselecting the initial weights to be small and random on each run. I found 32 distinct minima, each of which gave a different picture, and having about equal test set error.

This effect is closely connected to what I call instability (Breiman, 1996a) that occurs when there are many different models crowded together that have about the same training or test set error. Then a slight perturbation of the data or in the model construction will cause a skip from one model to another. The two models are close to each other in terms of error, but can be distant in terms of the form of the model.

If, in logistic regression or the Cox model, the common practice of deleting the less important covariates is carried out, then the model becomes unstable—there are too many competing models. Say you are deleting from 15 variables to 4 variables. Perturb the data slightly and you will very possibly get a different four-variable model and a different conclusion about which variables are important. To improve accuracy by weeding out less important covariates you run into the multiplicity problem. The picture of which covariates are important can vary significantly between two models having about the same deviance.

Aggregating over a large set of competing models can reduce the nonuniqueness while improving accuracy. Arena et al. (2000) bagged (see Glossary) logistic regression models on a data base of toxic and nontoxic chemicals where the number of covariates in each model was reduced from 15 to 4 by standard best subset selection. On a test set, the bagged model was significantly more accurate than the single model with four covariates. It is also more stable. This is one possible fix. The multiplicity problem and its effect on conclusions drawn from models needs serious attention.

9. OCCAM AND SIMPLICITY VS. ACCURACY

Occam's Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately, in prediction, accuracy and simplicity (interpretability) are in conflict. For instance, linear regression gives a fairly interpretable picture of the $\mathbf{y}, \mathbf{x}$ relation. But its accuracy is usually less than that of the less interpretable neural nets. An example closer to my work involves trees.

On interpretability, trees rate an A+. A project I worked on in the late 1970s was the analysis of delay in criminal cases in state court systems. The Constitution gives the accused the right to a speedy trial. The Center for the State Courts was concerned

TABLE 1
Data set descriptions

Data set	Training Sample size	Test Sample size	Variables	Classes
Cancer	699	—	9	2
Ionosphere	351	—	34	2
Diabetes	768	—	8	2
Glass	214	—	9	6
Soybean	683	—	35	19
Letters	15,000	5000	16	26
Satellite	4,435	2000	36	6
Shuttle	43,500	14,500	9	7
DNA	2,000	1,186	60	3
Digit	7,291	2,007	256	10

that in many states, the trials were anything but speedy. It funded a study of the causes of the delay. I visited many states and decided to do the analysis in Colorado, which had an excellent computerized court data system. A wealth of information was extracted and processed.

The dependent variable for each criminal case was the time from arraignment to the time of sentencing. All of the other information in the trial history were the predictor variables. A large decision tree was grown, and I showed it on an overhead and explained it to the assembled Colorado judges. One of the splits was on District N which had a larger delay time than the other districts. I refrained from commenting on this. But as I walked out I heard one judge say to another, "I knew those guys in District N were dragging their feet."

While trees rate an A+ on interpretability, they are good, but not great, predictors. Give them, say, a B on prediction.

9.1 Growing Forests for Prediction

Instead of a single tree predictor, grow a forest of trees on the same data—say 50 or 100. If we are classifying, put the new **x** down each tree in the forest and get a vote for the predicted class. Let the forest prediction be the class that gets the most votes. There has been a lot of work in the last five years on ways to grow the forest. All of the well-known methods grow the forest by perturbing the training set, growing a tree on the perturbed training set, perturbing the training set again, growing another tree, etc. Some familiar methods are bagging (Breiman, 1996b), boosting (Freund and Schapire, 1996), arcing (Breiman, 1998), and additive logistic regression (Friedman, Hastie and Tibshirani, 1998).

My preferred method to date is random forests. In this approach successive decision trees are grown by introducing a random element into their construction. For example, suppose there are 20 predictor variables. At each node choose several of the 20 at random to use to split the node. Or use a random combination of a random selection of a few variables. This idea appears in Ho (1998), in Amit and Geman (1997) and is developed in Breiman (1999).

9.2 Forests Compared to Trees

We compare the performance of single trees (CART) to random forests on a number of small and large data sets, mostly from the UCI repository (ftp.ics.uci.edu/pub/MachineLearningDatabases). A summary of the data sets is given in Table 1.

Table 2 compares the test set error of a single tree to that of the forest. For the five smaller data sets above the line, the test set error was estimated by leaving out a random 10% of the data, then running CART and the forest on the other 90%. The left-out 10% was run down the tree and the forest and the error on this 10% computed for both. This was repeated 100 times and the errors averaged. The larger data sets below the line came with a separate test set. People who have been in the classification field for a while find these increases in accuracy startling. Some errors are halved. Others are reduced by one-third. In regression, where the

TABLE 2
Test set misclassification error (%)

Data set	Forest	Single tree
Breast cancer	2.9	5.9
Ionosphere	5.5	11.2
Diabetes	24.2	25.3
Glass	22.0	30.4
Soybean	5.7	8.6
Letters	3.4	12.4
Satellite	8.6	14.8
Shuttle $\times 10^3$	7.0	62.0
DNA	3.9	6.2
Digit	6.2	17.1

forest prediction is the average over the individual tree predictions, the decreases in mean-squared test set error are similar.

9.3 Random Forests are A + Predictors

The Statlog Project (Mitchie, Spiegelhalter and Taylor, 1994) compared 18 different classifiers. Included were neural nets, CART, linear and quadratic discriminant analysis, nearest neighbor, etc. The first four data sets below the line in Table 1 were the only ones used in the Statlog Project that came with separate test sets. In terms of rank of accuracy on these four data sets, the forest comes in 1, 1, 1, 1 for an average rank of 1.0. The next best classifier had an average rank of 7.3.

The fifth data set below the line consists of 16×16 pixel gray scale depictions of handwritten ZIP Code numerals. It has been extensively used by AT&T Bell Labs to test a variety of prediction methods. A neural net handcrafted to the data got a test set error of 5.1% vs. 6.2% for a standard run of random forest.

9.4 The Occam Dilemma

So forests are A+ predictors. But their mechanism for producing a prediction is difficult to understand. Trying to delve into the tangled web that generated a plurality vote from 100 trees is a Herculean task. So on interpretability, they rate an F. Which brings us to the Occam dilemma:

- Accuracy generally requires more complex prediction methods. Simple and interpretable functions do not make the most accurate predictors.

Using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why. In fact, Section 10 points out that from a goal-oriented statistical viewpoint, there is no Occam's dilemma. (For more on Occam's Razor see Domingos, 1998, 1999.)

10. BELLMAN AND THE CURSE OF DIMENSIONALITY

The title of this section refers to Richard Bellman's famous phrase, "the curse of dimensionality." For decades, the first step in prediction methodology was to avoid the curse. If there were too many prediction variables, the recipe was to find a few features (functions of the predictor variables) that "contain most of the information" and then use these features to replace the original variables. In procedures common in statistics such as regression, logistic regression and survival models the advised practice is to use variable deletion to reduce the dimensionality. The published advice was that high dimensionality is dangerous. For instance, a well-regarded book on pattern recognition (Meisel, 1972) states "the features... must be relatively few in number." But recent work has shown that dimensionality can be a blessing.

10.1 Digging It Out in Small Pieces

Reducing dimensionality reduces the amount of information available for prediction. The more predictor variables, the more information. There is also information in various combinations of the predictor variables. Let's try going in the opposite direction:

- Instead of reducing dimensionality, increase it by adding many functions of the predictor variables.

There may now be thousands of features. Each potentially contains a small amount of information. The problem is how to extract and put together these little pieces of information. There are two outstanding examples of work in this direction, *The Shape Recognition Forest* (Y. Amit and D. Geman, 1997) and *Support Vector Machines* (V. Vapnik, 1995, 1998).

10.2 The Shape Recognition Forest

In 1992, the National Institute of Standards and Technology (NIST) set up a competition for machine algorithms to read handwritten numerals. They put together a large set of pixel pictures of handwritten numbers (223,000) written by over 2,000 individuals. The competition attracted wide interest, and diverse approaches were tried.

The Amit–Geman approach defined many thousands of small geometric features in a hierarchical assembly. Shallow trees are grown, such that at each node, 100 features are chosen at random from the appropriate level of the hierarchy; and the optimal split of the node based on the selected features is found.

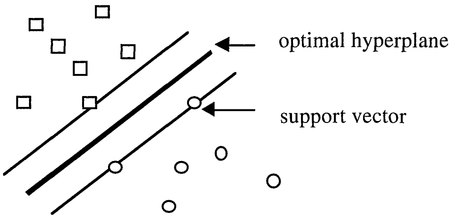
When a pixel picture of a number is dropped down a single tree, the terminal node it lands in gives probability estimates $p_0, \dots, p_9$ that it represents numbers $0, 1, \dots, 9$. Over 1,000 trees are grown, the probabilities averaged over this forest, and the predicted number is assigned to the largest averaged probability.

Using a 100,000 example training set and a 50,000 test set, the Amit–Geman method gives a test set error of 0.7%–close to the limits of human error.

10.3 Support Vector Machines

Suppose there is two-class data having prediction vectors in M-dimensional Euclidean space. The prediction vectors for class #1 are $\{\mathbf{x}(1)\}$ and those for class #2 are $\{\mathbf{x}(2)\}$. If these two sets of vectors can be separated by a hyperplane then there is an optimal separating hyperplane. "Optimal" is defined as meaning that the distance of the hyperplane to any prediction vector is maximal (see below).

The set of vectors in $\{\mathbf{x}(1)\}$ and in $\{\mathbf{x}(2)\}$ that achieve the minimum distance to the optimal separating hyperplane are called the support vectors. Their coordinates determine the equation of the hyperplane. Vapnik (1995) showed that if a separating hyperplane exists, then the optimal separating hyperplane has low generalization error (see Glossary).



In two-class data, separability by a hyperplane does not often occur. However, let us increase the dimensionality by adding as additional predictor variables all quadratic monomials in the original predictor variables; that is, all terms of the form $x_{m1}x_{m2}$. A hyperplane in the original variables plus quadratic monomials in the original variables is a more complex creature. The possibility of separation is greater. If no separation occurs, add cubic monomials as input features. If there are originally 30 predictor variables, then there are about 40,000 features if monomials up to the fourth degree are added.

The higher the dimensionality of the set of features, the more likely it is that separation occurs. In the ZIP Code data set, separation occurs with fourth degree monomials added. The test set error is 4.1%. Using a large subset of the NIST data base as a training set, separation also occurred after adding up to fourth degree monomials and gave a test set error rate of 1.1%.

Separation can always be had by raising the dimensionality high enough. But if the separating hyperplane becomes too complex, the generalization error becomes large. An elegant theorem (Vapnik, 1995) gives this bound for the expected generalization error:

$$\text{Ex(GE)} \leq \text{Ex(number of support vectors)}/(N-1),$$

where N is the sample size and the expectation is over all training sets of size N drawn from the same underlying distribution as the original training set.

The number of support vectors increases with the dimensionality of the feature space. If this number becomes too large, the separating hyperplane will not give low generalization error. If separation cannot be realized with a relatively small number of support vectors, there is another version of support vector machines that defines optimality by adding a penalty term for the vectors on the wrong side of the hyperplane.

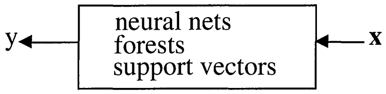
Some ingenious algorithms make finding the optimal separating hyperplane computationally feasible. These devices reduce the search to a solution of a quadratic programming problem with linear inequality constraints that are of the order of the number N of cases, independent of the dimension of the feature space. Methods tailored to this particular problem produce speed-ups of an order of magnitude over standard methods for solving quadratic programming problems.

Support vector machines can also be used to provide accurate predictions in other areas (e.g., regression). It is an exciting idea that gives excellent performance and is beginning to supplant the use of neural nets. A readable introduction is in Cristianini and Shawe-Taylor (2000).

11. INFORMATION FROM A BLACK BOX

The dilemma posed in the last section is that the models that best emulate nature in terms of predictive accuracy are also the most complex and inscrutable. But this dilemma can be resolved by realizing the wrong question is being asked. Nature forms the outputs $\mathbf{y}$ from the inputs $\mathbf{x}$ by means of a black box with complex and unknown interior.

Current accurate prediction methods are also complex black boxes.



So we are facing two black boxes, where ours seems only slightly less inscrutable than nature's. In data generated by medical experiments, ensembles of predictors can give cross-validated error rates significantly lower than logistic regression. My biostatistician friends tell me, "Doctors can interpret logistic regression." There is no way they can interpret a black box containing fifty trees hooked together. In a choice between accuracy and interpretability, they'll go for interpretability.

Framing the question as the choice between accuracy and interpretability is an incorrect interpretation of what the goal of a statistical analysis is.

The point of a model is to get useful information about the relation between the response and predictor variables. Interpretability is a way of getting information. But a model does not have to be simple to provide reliable information about the relation between predictor and response variables; neither does it have to be a data model.

- The goal is not interpretability, but accurate information.

The following three examples illustrate this point. The first shows that random forests applied to a medical data set can give more reliable information about covariate strengths than logistic regression. The second shows that it can give interesting information that could not be revealed by a logistic regression. The third is an application to a microarray data where it is difficult to conceive of a data model that would uncover similar information.

11.1 Example I: Variable Importance in a Survival Data Set

The data set contains survival or nonsurvival of 155 hepatitis patients with 19 covariates. It is available at ftp.ics.uci.edu/pub/MachineLearning-Databases and was contributed by Gail Gong. The description is in a file called hepatitis.names. The data set has been previously analyzed by Diaconis and Efron (1983), and Cestnik, Konenenko and Bratko (1987). The lowest reported error rate to date, 17%, is in the latter paper.

Diaconis and Efron refer to work by Peter Gregory of the Stanford Medical School who analyzed this data and concluded that the important variables were numbers 6, 12, 14, 19 and reports an estimated 20% predictive accuracy. The variables were reduced in two stages—the first was by informal data analysis. The second refers to a more formal (unspecified) statistical procedure which I assume was logistic regression.

Efron and Diaconis drew 500 bootstrap samples from the original data set and used a similar procedure to isolate the important variables in each bootstrapped data set. The authors comment, "Of the four variables originally selected not one was selected in more than 60 percent of the samples. Hence the variables identified in the original analysis cannot be taken too seriously." We will come back to this conclusion later.

Logistic Regression

The predictive error rate for logistic regression on the hepatitis data set is 17.4%. This was evaluated by doing 100 runs, each time leaving out a randomly selected 10% of the data as a test set, and then averaging over the test set errors.

Usually, the initial evaluation of which variables are important is based on examining the absolute values of the coefficients of the variables in the logistic regression divided by their standard deviations. Figure 1 is a plot of these values.

The conclusion from looking at the standardized coefficients is that variables 7 and 11 are the most important covariates. When logistic regression is run using only these two variables, the cross-validated error rate rises to 22.9%. Another way to find important variables is to run a best subsets search which, for any value k, finds the subset of k variables having lowest deviance.

This procedure raises the problems of instability and multiplicity of models (see Section 7.1). There are about 4,000 subsets containing four variables. Of these, there are almost certainly a substantial number that have deviance close to the minimum and give different pictures of what the underlying mechanism is.

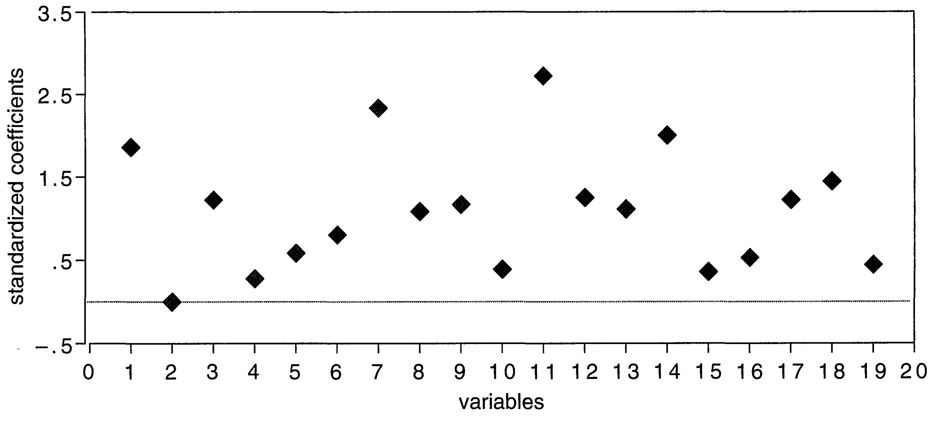


FIG. 1. *Standardized coefficients logistic regression.*

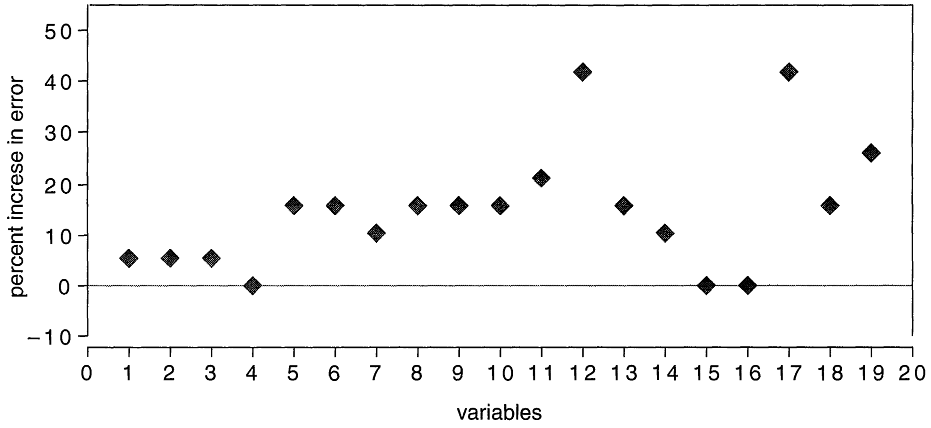


FIG. 2. *Variable importance-random forest.*

Random Forests

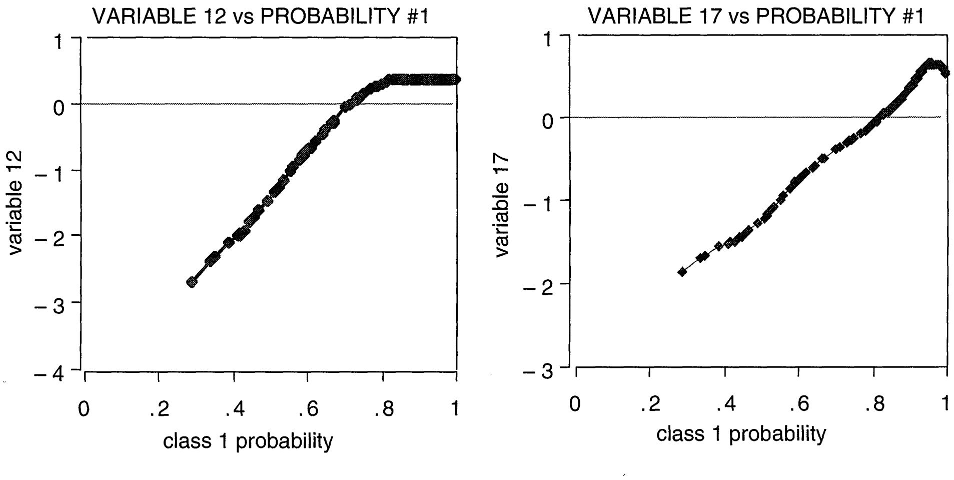
The random forests predictive error rate, evaluated by averaging errors over 100 runs, each time leaving out 10% of the data as a test set, is 12.3%—almost a 30% reduction from the logistic regression error.

Random forests consists of a large number of randomly constructed trees, each voting for a class. Similar to bagging (Breiman, 1996), a bootstrap sample of the training set is used to construct each tree. A random selection of the input variables is searched to find the best split for each node.

To measure the importance of the mth variable, the values of the mth variable are randomly permuted in all of the cases left out in the current bootstrap sample. Then these cases are run down the current tree and their classification noted. At the end of a run consisting of growing many trees, the percent increase in misclassification rate due to noising up each variable is computed. This is the measure of variable importance that is shown in Figure 1.

Random forests singles out two variables, the 12th and the 17th, as being important. As a verification both variables were run in random forests, individually and together. The test set error rates over 100 replications were 14.3% each. Running both together did no better. We conclude that virtually all of the predictive capability is provided by a single variable, either 12 or 17.

To explore the interaction between 12 and 17 a bit further, at the end of a random forest run using all variables, the output includes the estimated value of the probability of each class vs. the case number. This information is used to get plots of the variable values (normalized to mean zero and standard deviation one) vs. the probability of death. The variable values are smoothed using a weighted linear regression smoother. The results are in Figure 3 for variables 12 and 17.


VARIABLE 12 vs PROBABILITY #1

VARIABLE 17 vs PROBABILITY #1

FIG. 3. *Variable* 17 *vs. probability* #1.

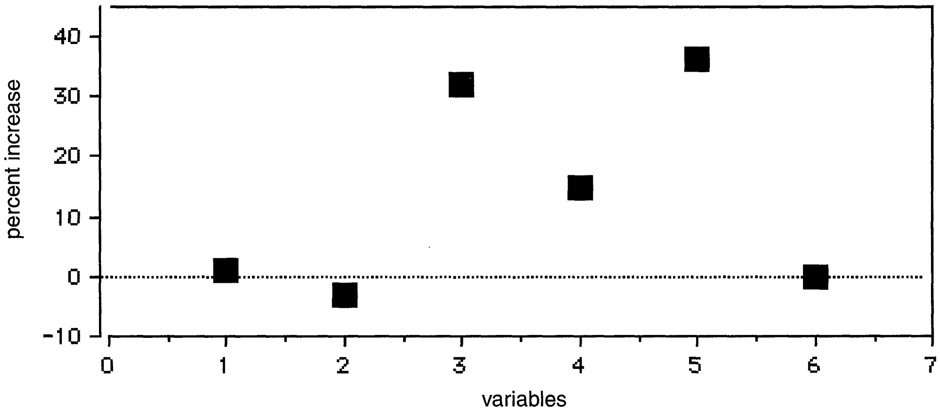


FIG. 4. *Variable importance—Bupa data.*

The graphs of the variable values vs. class death probability are almost linear and similar. The two variables turn out to be highly correlated. Thinking that this might have affected the logistic regression results, it was run again with one or the other of these two variables deleted. There was little change.

Out of curiosity, I evaluated variable importance in logistic regression in the same way that I did in random forests, by permuting variable values in the 10% test set and computing how much that increased the test set error. Not much help—variables 12 and 17 were not among the 3 variables ranked as most important. In partial verification of the importance of 12 and 17, I tried them separately as single variables in logistic regression. Variable 12 gave a 15.7% error rate, variable 17 came in at 19.3%.

To go back to the original Diaconis–Efron analysis, the problem is clear. Variables 12 and 17 are surrogates for each other. If one of them appears important in a model built on a bootstrap sample, the other does not. So each one's frequency of occurrence is automatically less than 50%. The paper lists the variables selected in ten of the samples. Either 12 or 17 appear in seven of the ten.

11.2 Example II Clustering in Medical Data

The Bupa liver data set is a two-class biomedical data set also available at ftp.ics.uci.edu/pub/MachineLearningDatabases. The covariates are:

1.	mcv	mean corpuscular volume
2.	alkphos	alkaline phosphotase
3.	sgpt	alamine aminotransferase
4.	sgot	aspartate aminotransferase
5.	gammagt	gamma-glutamyl transpeptidase
6.	drinks	half-pint equivalents of alcoholic beverage drunk per day

The first five attributes are the results of blood tests thought to be related to liver functioning. The 345 patients are classified into two classes by the severity of their liver malfunctioning. Class two is severe malfunctioning. In a random forests run,

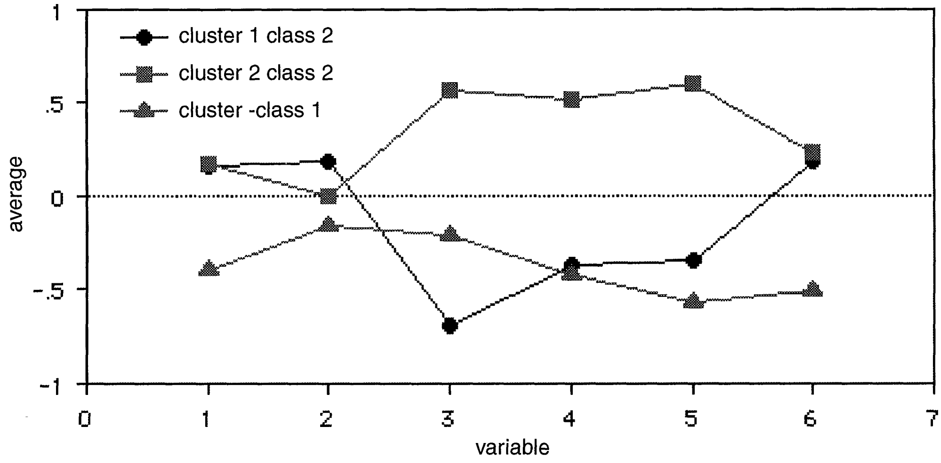


FIG. 5. *Cluster averages—Bupa data.*

the misclassification error rate is 28%. The variable importance given by random forests is in Figure 4.

Blood tests 3 and 5 are the most important, followed by test 4. Random forests also outputs an intrinsic similarity measure which can be used to cluster. When this was applied, two clusters were discovered in class two. The average of each variable is computed and plotted in each of these clusters in Figure 5.

An interesting facet emerges. The class two subjects consist of two distinct groups: those that have high scores on blood tests 3, 4, and 5 and those that have low scores on those tests.

11.3 Example III: Microarray Data

Random forests was run on a microarray lymphoma data set with three classes, sample size of 81 and 4,682 variables (genes) without any variable selection [for more information about this data set, see Dudoit, Fridlyand and Speed, (2000)]. The error rate was low. What was also interesting from a scientific viewpoint was an estimate of the importance of each of the 4,682 gene expressions.

The graph in Figure 6 was produced by a run of random forests. This result is consistent with assessments of variable importance made using other algorithmic methods, but appears to have sharper detail.

11.4 Remarks about the Examples

The examples show that much information is available from an algorithmic model. Friedman (1999) derives similar variable information from a different way of constructing a forest. The similarity is that they are both built as ways to give low predictive error.

There are 32 deaths and 123 survivors in the hepatitis data set. Calling everyone a survivor gives a baseline error rate of 20.6%. Logistic regression lowers this to 17.4%. It is not extracting much useful information from the data, which may explain its inability to find the important variables. Its weakness might have been unknown and the variable importances accepted at face value if its predictive accuracy was not evaluated.

Random forests is also capable of discovering important aspects of the data that standard data models cannot uncover. The potentially interesting clustering of class two patients in Example II is an illustration. The standard procedure when fitting data models such as logistic regression is to delete variables; to quote from Diaconis and Efron (1983) again, "...statistical experience suggests that it is unwise to fit a model that depends on 19 variables with only 155 data points available." Newer methods in machine learning thrive on variables—the more the better. For instance, random forests does not overfit. It gives excellent accuracy on the lymphoma data set of Example III which has over 4,600 variables, with no variable deletion and is capable of extracting variable importance information from the data.

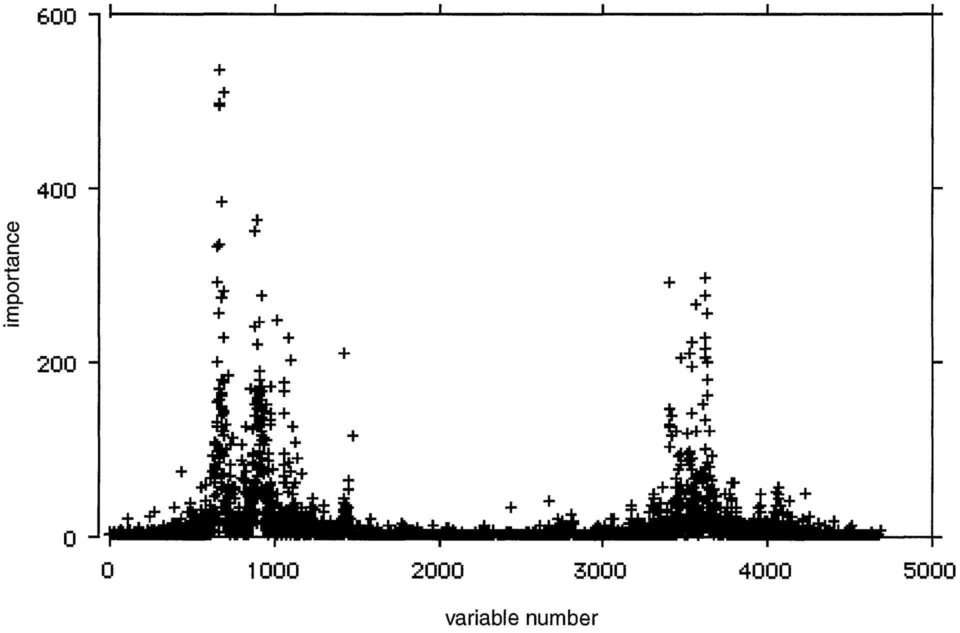


FIG. 6. *Microarray variable importance.*

These examples illustrate the following points:

- Higher predictive accuracy is associated with more reliable information about the underlying data mechanism. Weak predictive accuracy can lead to questionable conclusions.
- Algorithmic models can give better predictive accuracy than data models, and provide better information about the underlying mechanism.

12. FINAL REMARKS

The goals in statistics are to use data to predict and to get information about the underlying data mechanism. Nowhere is it written on a stone tablet what kind of model should be used to solve problems involving data. To make my position clear, I am not against data models per se. In some situations they are the most appropriate way to solve the problem. But the emphasis needs to be on the problem and on the data.

Unfortunately, our field has a vested interest in data models, come hell or high water. For instance, see Dempster's (1998) paper on modeling. His position on the 1990 Census adjustment controversy is particularly interesting. He admits that he doesn't know much about the data or the details, but argues that the problem can be solved by a strong dose of modeling. That more modeling can make error-ridden data accurate seems highly unlikely to me.

Terrabytes of data are pouring into computers from many sources, both scientific, and commercial, and there is a need to analyze and understand the data. For instance, data is being generated at an awesome rate by telescopes and radio telescopes scanning the skies. Images containing millions of stellar objects are stored on tape or disk. Astronomers need automated ways to scan their data to find certain types of stellar objects or novel objects. This is a fascinating enterprise, and I doubt if data models are applicable. Yet I would enter this in my ledger as a statistical problem.

The analysis of genetic data is one of the most challenging and interesting statistical problems around. Microarray data, like that analyzed in Section 11.3 can lead to significant advances in understanding genetic effects. But the analysis of variable importance in Section 11.3 would be difficult to do accurately using a stochastic data model.

Problems such as stellar recognition or analysis of gene expression data could be high adventure for statisticians. But it requires that they focus on solving the problem instead of asking what data model they can create. The best solution could be an algorithmic model, or maybe a data model, or maybe a combination. But the trick to being a scientist is to be open to using a wide variety of tools.

The roots of statistics, as in science, lie in working with data and checking theory against data. I hope in this century our field will return to its roots. There are signs that this hope is not illusory. Over the last ten years, there has been a noticeable move toward statistical work on real world problems and reaching out by statisticians toward collaborative work with other disciplines. I believe this trend will continue and, in fact, *has* to continue if we are to survive as an energetic and creative field.

GLOSSARY

Since some of the terms used in this paper may not be familiar to all statisticians, I append some definitions.

Infinite test set error. Assume a loss function $L(y, \hat{y})$ that is a measure of the error when y is the true response and $\hat{y}$ the predicted response. In classification, the usual loss is 1 if $y \neq \hat{y}$ and zero if $y = \hat{y}$. In regression, the usual loss is $(y - \hat{y})^2$. Given a set of data (training set) consisting of $\{(y_n, \mathbf{x}_n)n = 1, 2, \ldots, N\}$, use it to construct a predictor function $\phi(\mathbf{x})$ of y. Assume that the training set is i.i.d drawn from the distribution of the random vector $Y, \mathbf{X}$. The infinite test set error is $E(L(Y, \phi(\mathbf{X})))$. This is called the generalization error in machine learning.

The *generalization error* is estimated either by setting aside a part of the data as a test set or by cross-validation.

Predictive accuracy. This refers to the size of the estimated generalization error. Good predictive accuracy means low estimated error.

Trees and nodes. This terminology refers to decision trees as described in the Breiman et al book (1984).

Dropping an $\mathbf{x}$ *down a tree*. When a vector of predictor variables is "dropped" down a tree, at each intermediate node it has instructions whether to go left or right depending on the coordinates of $\mathbf{x}$. It stops at a terminal node and is assigned the prediction given by that node.

Bagging. An acronym for "bootstrap aggregating." Start with an algorithm such that given any training set, the algorithm produces a prediction function $\phi(\mathbf{x})$. The algorithm can be a decision tree construction, logistic regression with variable deletion, etc. Take a bootstrap sample from the training set and use this bootstrap training set to construct the predictor $\phi_1(\mathbf{x})$. Take another bootstrap sample and using this second training set construct the predictor $\phi_2(\mathbf{x})$. Continue this way for K steps. In regression, average all of the $\{\phi_k(\mathbf{x})\}$ to get the bagged predictor at **x**. In classification, that class which has the plurality vote of the $\{\phi_k(\mathbf{x})\}$ is the bagged predictor. Bagging has been shown effective in variance reduction (Breiman, 1996b).

Boosting. This is a more complex way of forming an ensemble of predictors in classification than bagging (Freund and Schapire, 1996). It uses no randomization but proceeds by altering the weights on the training set. Its performance in terms of low prediction error is excellent (for details see Breiman, 1998).

ACKNOWLEDGMENTS

Many of my ideas about data modeling were formed in three decades of conversations with my old friend and collaborator, Jerome Friedman. Conversations with Richard Olshen about the Cox model and its use in biostatistics helped me to understand the background. I am also indebted to William Meisel, who headed some of the prediction projects I consulted on and helped me make the transition from probability theory to algorithms, and to Charles Stone for illuminating conversations about the nature of statistics and science. I'm grateful also for the comments of the editor, Leon Gleser, which prompted a major rewrite of the first draft of this manuscript and resulted in a different and better paper.

REFERENCES

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation* **9** 1545–1588.

Arena, C., Sussman, N., Chiang, K., Mazumdar, S., Macina, O. and Li, W. (2000). Bagging Structure-Activity Relationships: A simulation study for assessing misclassification rates. Presented at the Second Indo-U.S. Workshop on Mathematical Chemistry, Duluth, MI. (Available at NSussman@server.ceoh.pitt.edu).

Bickel, P., Ritov, Y. and Stoker, T. (2001). Tailor-made tests for goodness of fit for semiparametric hypotheses. Unpublished manuscript.

Breiman, L. (1996a). The heuristics of instability in model selection. *Ann. Statist.* **24** 2350–2381.

Breiman, L. (1996b). Bagging predictors. *Machine Learning J.* **26** 123–140.

Breiman, L. (1998). Arcing classifiers. Discussion paper, *Ann. Statist.* **26** 801–824.

Breiman. L. (2000). Some infinity theory for tree ensembles. (Available at www.stat.berkeley.edu/technical reports).

Breiman, L. (2001). Random forests. *Machine Learning J.* **45** 5–32.

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations in multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge Univ. Press.

Daniel, C. and Wood, F. (1971). *Fitting equations to data*. Wiley, New York.

Dempster, A. (1998). Logicist statistic 1. Models and Modeling. *Statist. Sci.* **13** 3 248–276.

Diaconis, P. and Efron, B. (1983). Computer intensive methods in statistics. *Scientific American* **248** 116–131.

Domingos, P. (1998). Occam's two razors: the sharp and the blunt. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (R. Agrawal and P. Stolorz, eds.) 37–43. AAAI Press, Menlo Park, CA.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery* **3** 409–425.

Dudoit, S., Fridlyand, J. and Speed, T. (2000). Comparison of discrimination methods for the classification of tumors. (Available at www.stat.berkeley.edu/technical reports).

Freedman, D. (1987). As others see us: a case study in path analysis (with discussion). *J. Ed. Statist.* **12** 101–223.

Freedman, D. (1991). Statistical models and shoe leather. *Sociological Methodology 1991* (with discussion) 291–358.

Freedman, D. (1991). Some issues in the foundations of statistics. *Foundations of Science* **1** 19–83.

Freedman, D. (1994). From association to causation via regression. *Adv. in Appl. Math.* **18** 59–110.

Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference* 148–156. Morgan Kaufmann, San Francisco.

Friedman, J. (1999). Greedy predictive approximation: a gradient boosting machine. Technical report, Dept. Statistics Stanford Univ.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Ann. Statist.* **28** 337–407.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, New York.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Analysis and Machine Intelligence* **20** 832–844.

Landswher, J., Preibon, D. and Shoemaker, A. (1984). Graphical methods for assessing logistic regression models (with discussion). *J. Amer. Statist. Assoc.* **79** 61–83.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Meisel, W. (1972). *Computer-Oriented Approaches to Pattern Recognition*. Academic Press, New York.

Michie, D., Spiegelhalter, D. and Taylor, C. (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York.

Mosteller, F. and Tukey, J. (1977). *Data Analysis and Regression.* Addison-Wesley, Redding, MA.

Mountain, D. and Hsiao, C. (1989). A combined structural and flexible functional approach for modelenery substitution. *J. Amer. Statist. Assoc.* **84** 76–87.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B* **36** 111–147.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Vapnik, V (1998). *Statistical Learning Theory*. Wiley, New York.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Zhang, H. and Singer, B. (1999). *Recursive Partitioning in the Health Sciences.* Springer, New York.