Banter on Bayes: debating the usefulness of Bayesian approaches to solving practical problems

By Marti A. Hearst
University of California, Berkeley
hearst@sims.berkeley.edu

In this installment of Trends and Controversies, we consider the status of Bayesian statistical methods for the analysis of complex real-world problems.

Most AI problems require a method for representing uncertainty. Currently, there is intense interest in the use of Bayesian reasoning, Bayesian belief networks, and other Bayesian methods for the modeling of uncertainty in many avenues of AI research. For instance, the contributions to the conference on uncertainty in AI focus primarily on Bayesian reasoning methods. Additionally, many researchers consider Bayesian networks the method of choice in the nascent field of data mining. Practioners argue that the Bayesian paradigm for setting up and modeling a causal description of a problem can create more insight than those approaches that compute associations more "blindly." However, Bayesian methods are not without their detractors, and in this issue we hear three different viewpoints on the matter.

David Draper and David Madigan open the piece by reacquainting us with the difference between frequentist and Bayesian statistical methods. They then argue for the understanding that can be gained from a Bayesian analysis of the characteristics of a problem, and present a long list of what they call Bayesian success stories (most of which use the newly popular Markov Chain Monte Carlo methods).

The next essay is by Leo Breiman, who takes direct aim at the Bayesian techniques and declares that he has never seen a true Bayesian success story. He argues that an analyst using Bayesian techniques must focus on the machinery rather than the problem, and declares there are ways to incorporate domain knowledge into a statistical problem that avoid the need for selection of prior probabilities endemic to Bayesian approaches. He also names problems with the MCMC methods for computing posterior distributions.

A moderating voice is heard from Wray Buntine, who declares himself a Bayesian in principle, if not in practice. He argues that practitioners should not restrict themselves to an exclusively Bayesian framework; instead, they should employ what ever theoretical tool best suits the task at hand. Buntine also attempts to tease apart the differences between Bayesian networks, graphical models, and probabilistic models in general, and closes with the suggestion that perhaps the theory's biggest contribution is its ability to unify competing theories within a common language.

The authors were each allowed to see the others' essays and modify their positions accordingly. This conversation is the most contentious of those I've hosted so far; I hope it stimulates further discussion in the field.

The scientific value of Bayesian statistical methods

David Draper, University of Bath
David Madigan , University of Washington

Statistics is the study of uncertainty: how to measure it and what to do about it. Uncertainty itself, as people regard it today, is a fairly new concept in the history of ideas. For example, probability—the branch of mathematics devoted to quantifying uncertainty—dates only from the middle 1600s, brought to life through disputes about how to wager fairly in games of chance.

Since then, two main ways to give meaning to the concept of probability have arisen and, to some extent, fought with each other: the *frequentist* and *Bayesian* approaches:

- In the frequentist definition of probability, you restrict attention to phenomena that are inherently repeatable under identical conditions, and define probability as a limiting proportion in a hypothetical infinite series of such repetitions.
- In the Bayesian approach, you imagine betting with someone about the truth of a *proposition* (which can be anything—not just an aspect of a repeatable process—whose truth value is not yet known), and ask yourself what odds you would need to give or receive to make the bet fair.

Thus, strictly speaking, the frequentist probability (viewed from 1997) that Al Gore will be elected US President in 2000 is undefined (because this situation is sufficiently unique that it defies natural embedding in a repeatable sequence). By contrast, all forms of uncertainty are inherently quantifiable with the Bayesian approach (although there is no guarantee that the answer you get by querying yourself about betting odds will retrospectively be seen by you or others as good).

Bayesian and non-Bayesian contrasts

When you confront these two viewpoints with data, in a setting in which you are trying to draw inferences about one or more unknowns, interesting and useful contrasts emerge.

- When we reason in a frequentist way, the need to tell stories involving repetition causes us to regard the data values before us as only one possible outcome of the supposedly random process of data generation, and we regard the unknowns as constants. Thus, we view the data as random and the unknowns as fixed.
- When we are thinking Bayesianly, we hold constant things we know, including the data values, and we use the machinery of probability to quantify our uncertainty about the unknowns—the data are fixed and the unknowns are random. This makes it easy to explicitly acknowledge the sequential nature of scientific learning: before the current data arrive, we know something about

Bayesian success stories

Recent progress in Bayesian methodology has led to the following successful Bayesian collaborative analyses:

- *Medicine and medical decision-making*: response of CD4 lymphocytes to treatment using ddI or ddC in HIV/AIDS[1] and relating risks of breast cancer to decisions about screening;[2]
- *Civil engineering*: estimation of capital investment in water;[3]
- Pharmaceuticals: fitting physiologically-based models in toxicokinetics[4] and pharmacokinetics;[5]
- Environment: identifying homes with high radon levels,[6] and estimation of the bowhead whale population size;[7]
- Public policy: estimating the incumbency advantage in US congressional elections,[8] and the effects of redistricting in US legislatures;[9]
- Genetics: pedigree analysis and genotypic inference for management and conservation of small animal populations;[10]
- Physiology: modeling of eye movements during visual processing,[11] and mixture modeling applied to post-synaptic potential fluctuations;[12] and
- Business: modeling customer satisfaction data,[13] and analysis of real-estate sales patterns.[14]

We could have just as easily included many other application areas on this list, including epidemiology, climatology, manufacturing, geography, and econometrics. For instance, see *http://www.stat.cmu.edu/meetings/Bayes97/invited.html* and *abs-sub.html* for details on the fourth in a series of workshops on case studies in Bayesian statistics.

References

1. A.I. Goldman et al., "Response of CD4 Lymphocytes and Clinical Consequences of Treatment Using ddI or ddC in Patients with Advanced HIV Infection," *J. Acquired Immune Deficiency Syndromes and Human Retrovirology*, Vol. 11, No. 2, 1996, pp. 161–169.
2. G. Parmigiani et al., Modeling Risk of Breast Cancer and Decisions about Genetic Testing," to be published in *Case Studies in Bayesian Statistics*, Vol. 4, B.P. Carlin et al., eds., Springer-Verlag, New York.
3. A. O'Hagan and F.S. Wells, "Use of Prior Information to Estimate Costs in a Wewerage Operation," in *Case Studies in Bayesian Statistics*, C. Gatsonis et al., eds., Springer-Verlag, 1993, pp. 118–163.
4. F.Y. Bois et al., "Population Toxicokinetics of Tetrachloroethylene," *Archives of Toxicology*, Vol. 70, No. 6, 1996, pp. 347–355.
5. J.C. Wakefield, "Bayesian Individualization via Sampling-Based Methods," *J. Pharmacokinetics and Biopharmaceutics*, Vol. 24, No. 1, 1996, pp. 103–131.
6. P.N. Price, A.V. Nero, and A. Gelman, "Bayesian Prediction of Mean Indoor Radon Concentrations for Minnesota Counties," *Health Physics*, Vol. 71, No. 6, 1996, pp. 922–936.
7. G.H. Givens, J.E. Zeh, and A.E. Raftery, "Implementing the Current Management Regime for Aboriginal Subsistence Whaling to Establish a Catch Limit for the Bering-Chukchi-Beaufort Seas Stock of Bowhead Whales," *Report of the Int'l Whaling Commission*, Vol. 46, 1996, pp. 593–600.
8. A. Gelman and G. King, "Estimating Incumbency Advantage Without Bias," *Am. J. Political Science*, Vol. 34, No. 4, 1990, pp. 1142–1164.
9. A. Gelman and G. King, "A Unified Model for Evaluating Electoral Systems and Redistricting Plans," *Am. J. Political Science*, Vol. 38, No. 2, 1994, pp. 514–554.
10. A. Thomas, "Genotypic Inference with the Gibbs Sampler," in *Population Management for Survival and Recovery: Analytical Methods and Strategies in Small-Population Conservation*, J.D. Ballou, M. Gilpin, and T.J. Foose, eds., Columbia Univ. Press, New York, 1995, pp. 261–270.
11. C.R. Genovese and J.A. Sweeney, "Functional Connectivity in the Cortical Circuits Subserving Eye Movements," to be published in *Case Studies in Bayesian Statistics*, Vol. 4, B.P. Carlin et al., eds., Springer-Verlag.
12. D.A. Turner and M. West, "Bayesian Analysis of Mixtures Applied to Postsynaptic Potential Fluctuations," *J. Neuroscience Methods*, Vol. 47, Nos. 1–2, 1993, pp. 1–23.
13. L.A. Clark et al., "Modeling Customer Survey Data," to be published in *Case Studies in Bayesian Statistics*, Vol. 4, B.P. Carlin et al., eds., Springer-Verlag.
14. J.R. Knight, "Modeling Real Estate Sales Patterns Using the Gibbs Sampler, Poster Presentation," to appear in *Fourth Ann. Conf. Case Studies in Bayesian Statistics*, Springer-Verlag, New York, 1997.

the unknowns (people call this *prior* information). Then the data arrive, and our job is to quantify what we now know about the unknowns (the *posterior* information) in light of the data.

As a result of this distinction, Bayesian inferential statements tend to be easier to interpret, and it is often easier with the Bayesian approach to build highly complex models when such complexity is realistic.

As an example of the former assertion, consider the usual confusion over interpreting frequentist interval estimates. A Bayesian 90% probability interval for a mean, for instance, has—for those who accept the assumptions on which it is based—90% probability of containing the unknown mean. But we can interpret a frequentist 90% confidence interval only in relation to a sequence of similar inferences that might be made with other data sets. Of course, most frequentists interpret such intervals in a Bayesian way anyhow!

For a simple example of the second assertion, when medical researchers conduct blood enzyme studies on a representative cross section of a population, they often observe unexplained heterogeneity (which will appear in histograms or density traces as two or more distinct modes). Scientific interest then focuses on how many subpopulations have in effect been mixed together. Such *mixture* analyses with an unknown number of components are vexedly difficult to undertake using frequentist inferential methods, but are straightforward with a Bayesian formulation.[1]

These advantages, however, come at an apparent price: the Bayesian approach requires the quantitative specification of prior information, which can be regarded as either a bug (it can be hard to do) or a feature (if you can do it well, you will get better answers). The view that prior specification is a bug—which tends to go hand in hand with the frequentist approach—has dominated the field of statistics for much of the last 100 years. Another contributing factor to this dominance has been that the computations associated with the Bayesian viewpoint in complicated problems have been prohibitive until very recently.

Success stories

The last decade has seen a dramatic change. The "discovery" by statisticians of a new class of computational tools, *Markov Chain Monte Carlo* methods[2]—which actually amounts to a reinvention of methods that have been in active use in statistical physics and chemistry for 45 years—has opened the floodgates to a host of serious Bayesian applications. The sidebar lists just a few recent Bayesian success stories, most of them made possible by MCMC methods.

An example

We now present a short summary of the Bayesian approach in action in a problem of moderate complexity (the analysis here closely follows that of David Spiegelhalter[3]).

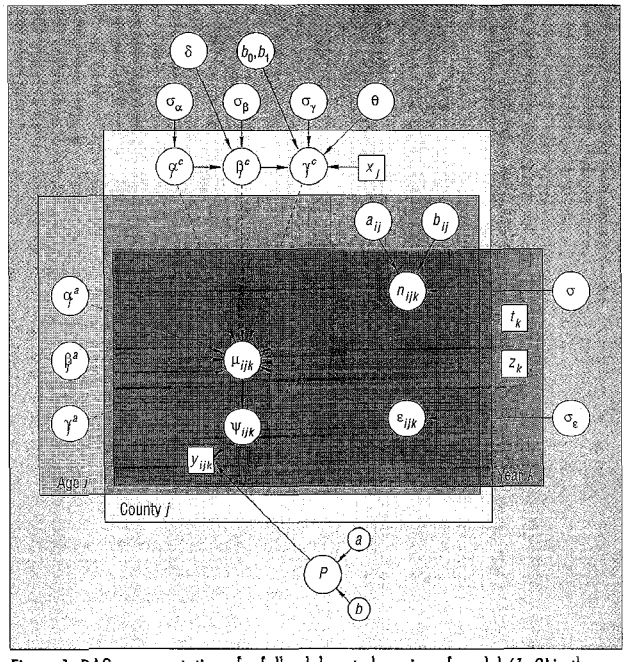


Figure 1. DAG representation of a fully elaborated version of model *(1, 2)* in the abortion example.

In the US, the anti-abortion campaign of the late 1980s and early 1990s generated much publicity and occasionally became violent. Sociologists, epidemiologists, and medical researchers have begun to study this campaign's effect on access to reproductive services and on pregnancy terminations. Interest lies mainly in modeling the incidence rates of pregnancy terminations and their changes over time at the county level, with particular attention focusing on possible changes in response to the anti-abortion campaign. Available data include the observed number y_{ijk} of reported terminations in age group i, county j, and year k (across five age groups, 38 US counties, and the years 1983–94), together with the appropriate population denominators n_{ijk} and a selection of county-level predictor variables x_j, including the estimated number of clinics providing reproductive services in 1983 and in 1993.

A natural model for count data such as y_{ijk} involves regarding the counts as realizations of Poisson random variables,

$$y_{ijk} \sim P(\mu_{ijk}), \tag{1}$$

where the mean structure μ_{ijk} is driven by age group, district, and year. Preliminary analysis suggests the structure

$$\log \mu_{ijk} = \log n_{ijk} + \alpha_i^a + \alpha_j^c + (\beta_i^a + \beta_j^c)t_k + (\gamma_i^a + \gamma_j^c)\, z_k, \tag{2}$$

where z_k is an indicator variable for after-1990 versus before and α_i^a, β_i^a, and γ_i^a are the age effect on the intercept, the change per year, and the gradient after 1990. However, several potential deficiencies in this model are apparent, suggesting the following elaborations:

- *Over-dispersion*: The theoretical mean and variance of the Poisson distribution are equal, but with data such as these, sample variances often greatly exceed sample means. Thus, the model typically needs to be generalized to allow for this extra variability, which we can consider as arising from the omission of additional unobserved predictor variables. Adding a Gaussian error term to Equation 2 can accomplish this generalization.
- *Hierarchical modeling*: The standard Bayesian modeling of the county-level coefficients α_j^c, β_j^c, and γ_j^c would involve specifying prior information—in the form of probability distributions—for each of them (114 separate distributions). However, we might obtain more accurate inferences and predictions by considering the counties as *exchangeable* (similar), which lets us pool our knowledge across them. This is accomplished mathematically by regarding, for instance, the 38 α_j^c's as having arisen from a single (Gaussian) distribution.
- *Relationships among the coefficients*: Prior experience suggests that it is unrealistic to consider the county-level coefficients α_j^c, β_j^c, and γ_j^c as independent, so we need to treat the three distributions just described (one for each of α, β, and γ) as correlated.
- *Underreporting*: Much is known about abortion underreporting rates, which is a form of missing data. This information must be incorporated into the model in a way that fully fleshes out the resulting uncertainty.

Figure 1 summarizes the resulting model, in the form of a *directed acyclic graph (DAG)*.[4] In its treatment of unknowns (the α_j^c) as random and the data values (y_{ijk}) as fixed—and, more fundamentally, in its modeling of the relationships among the unknowns—this figure is fundamentally Bayesian. The recent advances in MCMC computation make calculations based on the figure straightforward. Moreover, complicated functions of the unknowns, such as the true relative rankings in abortion rates among the counties, are available trivially, together with measures of uncertainty (difficult to achieve correctly with non-Bayesian analyses) that fully reflect what is known and what is not. Ranking summaries of this kind in other similar studies have often revealed that, despite previous attempts to base policies on rank-based *league tables*, so much uncertainty about the true rankings exists that such attempts are meaningless.

You can make all of the calculations in this example with an easy-to-use MCMC package, BUGS (Bayesian Inference Using Gibbs Sampling), available free (with extensive documentation) at *http://www.mrc-bsu.cam.ac.uk*.

Discussion

Bayes is not a panacea: prior distributions can be difficult to specify, and computations can be slow. But, with the advent of MCMC, the list of statistical models (and summaries based on them) that many analysts have found can be more satisfyingly attacked with Bayesian methods and outlook is growing daily. This list includes missing data, causal inference, predictive distributions, image analysis, dynamic linear modeling of time series, longitudinal data analysis, and hierarchical modeling.

We noted at the beginning of this essay that the two leading concepts of probability have fought with each other over the past three centuries. This doesn't have to be: there is no reason why you have to pick one or the other and stick slavishly to it. Wray Buntine's essay in this installment of Trends & Controversies seems to suggest one kind of compromise: let Bayesian principles guide your methodology development, but retain the freedom to implement your methods in non-Bayesian ways (in problems where Bayesian computation is still prohibitive). We prefer a different fusion of Bayes and non-Bayes: many practical Bayesians working on serious applied problems use Bayesian reasoning when formulating their

inferences and predictions, and frequentist reasoning when evaluating their quality—for instance, by keeping score on the accuracy of their predictions for data not used in the modeling process. It seems odd to tie one probabilistic hand behind your back before you even begin, so to speak, as analysts such as Leo Breiman seem to advocate in their opposition to Bayes.

Because this discussion is intended to be a partially interactive format and we find ourselves less in disagreement with Buntine than Breiman, some remarks responding directly to the latter's essay would perhaps be useful.

Breiman has looked for Bayesian applications in the Current Index to Statistics, which is dominated by journals in which statisticians talk to each other. If he had gone one or two layers farther out into the real world—to the actual applied journals in AIDS research and political science, for instance—he would have found more Bayesian work in the trenches. For example, the Inspec database, which covers the physical sciences, electrical engineering, and computer science, returns over 350 entries for 1996–7 in response to the query "`(Bayes or Bayesian) and data`," many involving real problems.

Breiman asks how to figure out what the "right" prior knowledge is, a question that seems to hang up many non-Bayesians to a much greater extent than actual Bayesian applied experience would justify. Here are a few general responses:

- *Sensitivity analysis is crucial.* If you are not sure how to quantify the substantive (expert) knowledge available prior to the current data collection, try several ways that look plausible and see if they lead to substantially the same conclusions. (This is, after all, how Bayesians and non-Bayesians alike specify the part of the statistical model—the *likelihood*—that permits inferences to be drawn from the data information, when no randomization was employed in the data-gathering.)
- *Tune the prior to get good predictive calibration.* Good statistical models—for Bayesians, this includes both the prior and the likelihood—should make good predictions. So, in the absence of strong substantive guidance, one approach to prior specification is to find priors that produce good out-of-sample predictive performance.
- *Vague prior distributions.* Standard non-Bayesian methods based only on the likelihood portion of the model have two common defects. First, they are often driven by approximations that work well with large amounts of data, leaving open the question of their validity in small samples. Second, they can produce poorly calibrated inferences about unknowns of principal interest, in the presence of a large number of unknowns of less direct interest. Bayesian and non-Bayesian analysts alike are finding it increasingly useful to employ the Bayesian machinery with prior distributions that convey little or no prior information (*vague* priors), to produce inferential and predictive procedures with good frequentist calibration properties.

It seems odd that Breiman claims the recent idea of "perturbing and combining predictors" in pattern recognition as a solely frequentist success story, when the research area of *Bayesian model averaging*—which has been active since the late 1980s, and which can trace its roots to papers in the 1960s—has already demonstrated the value of combining predictors in both theory and applications. Buntine supports this with his remarks about support-vector machines and averaging over multiple models.

The example we outlined earlier demonstrates that Bayesian statistics is far more than frequentist statistics with a prior distribution—this seems to be what Breiman and Buntine are thinking of when they say that Bayes "puts another layer of machinery between the problem ... and the problem-solver." We contest this view strongly. We have found that Bayesian methods and outlook provide a modeling framework that is liberating in its ability to represent real-world complexity.

We are not claiming that Bayesian methods are vital to the successful solution of the problems discussed here—non-Bayesian methods might have also succeeded in some or all cases. What we *are* claiming is that—with the advent of MCMC—the field of Bayesian applied statistics is for real: as Andrew Gelman has noted, "Bayesian techniques are out there, they are being used to solve real problems, and the success of these methods can be evaluated based on their results."

Acknowledgments

We are grateful to Brad Carlin, Andrew Gelman, Adrian Raftery, and David Spiegelhalter for discussions, comments, and references.

References

1. S. Richardson and P.J. Green, "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with Discussion)," *J. Royal Statistical Soc., Series B*, Vol. 59, No. 4, 1997, pp. 731–792.
2. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, 1996.
3. D.J. Spiegelhalter, "Bayesian Graphical Modeling: a Case Study in Monitoring Health Outcomes," to be published in *Applied Statistics*.
4. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco, 1988.

No Bayesians in foxholes

Leo Breiman
University of California, Berkeley

In World War II, there was a saying, "there are no atheists in foxholes." The implication was that on the front lines and under pressure, soldiers needed someone to pray to. The implication in my title is that when big, real, tough problems need to be solved, there are no Bayesians.

For decades, the pages of various statistical journals have been littered with theological arguments on the virtues of the Bayesian approach versus frequentist approaches. I have no intention of continuing the debate on this level. My approach is pragmatic: which approach works best when dealing with real data in solving complex problems?

Hardly a better mousetrap

The Current Index of Statistics lists all statistics articles published since 1960 by author, title, and key words. The CIS includes articles from a multitude of journals in various fields—medical statistics, reliability, environmental, econometrics, and business management, as well as all of the statistics journals. Searching under anything that contained the word "data" in 1995–1996 produced almost 700 listings. Only eight of these mentioned Bayes or Bayesian, either in the title or key words. Of these eight, only three appeared to apply

a Bayesian analysis to data sets, and in these, there were only two or three parameters to be estimated.

I spent 13 years as a full-time consultant and continue to consult in many fields today—air-pollution prediction, analysis of highway traffic, the classification of radar returns, speech recognition, and stock-market prediction, among others. Never once, either in my work with others or in anyone else's published work in the fields in which I consulted, did I encounter the application of Bayesian methodology to real data.

More specifically, speech recognition is a difficult and important prediction problem. Many approaches have been tried—large neural nets, hidden Markov chains, decision trees, and so on. But there are no working Bayesian speech-recognition systems. The same holds for handwritten character recognition and many other less well-known but difficult high-dimensional problems. In the reinforced learning field, there are no strong Bayesian chess-playing or backgammon algorithms—but there are effective ad hoc frequentist-based programs.

Over the years, I have discussed solving a large variety of prediction problems with hundreds of researchers in many fields. With very few exceptions, even the avowed Bayesians confess that when faced with data and a problem—for instance, using a data set of 15,000 examples, find an accurate algorithm that recognizes 34 different patterns using 128 inputs—they will think frequentist.

Thousands of smart people are working in various statistical fields—in pattern recognition, neural nets, machine learning, and reinforced learning, for example. Why do so few use a Bayesian analysis when faced with applications involving real data? It's not that the Bayes approach is new. It has been around almost since the beginnings of statistics and has been the subject of thousands of theoretical papers. It's quite respectable, and numbers of eminent theoretical statisticians are devoted Bayesians. But if it's a better mousetrap, why is the path to its doorstep so overgrown?

Bayesians say that in the past, the extreme difficulty in computing complex posteriors prevented more widespread use of Bayesian methods. There has been a recent flurry of interest in the machine-learning/neural-net community because Markov Chain Monte Carlo methods might offer an effective method of using computer-generated random tracks to approximate posterior distributions. They hope that MCMC methods for computing posterior distributions will lead to effective Bayesian work on complex problems.

I am dubious for several reasons: MCMC has drawbacks—it is a slow, compute-intensive method with no known means of judging convergence. If we apply it to situations with more than, say, 1,000 parameters (typical in complex problems), the difficulty is compounded because the posterior distribution is generally concentrated on a lower-dimensional subspace, and the random paths might spend most of their orbit outside this subspace.

No matter how you select priors, they might not be appropriate for the problem. In high-dimensional problems, to decrease the dimensionality of the prior distribution to manageable size, we make simplifying assumptions that set many parameters to be equal but of a size governed by a *hyperparameter*. For instance, in linear regression, we could assume that all the coefficients are normally and independently distributed with mean zero and common variance. Then the common variance is a hyperparameter and is given its own prior.

This leads to what is known in linear regression as *ridge regression*. This method, which has been widely tested, has proven itself in some cases, but fails in situations when some of the coefficients are large and others small. A Bayesian would say that the wrong prior knowledge had been used, but this raises the perennial question: how do you know what the right prior knowledge is?

For instance, some nonlinear-prediction methods have recently been moderately successful in fiscal market prediction. I recall a workshop some years ago at which a well-known Bayesian claimed that the way to do prediction in the stock market was to put priors on it. I was rendered speechless by this assertion. But one of the principals in a successful fiscal market prediction company shot to his feet and emphatically rejected the idea as being totally hopeless.

But the biggest reason that Bayesian methods have not been used more is that they put another layer of machinery between the problem to be solved and the problem solver. Given that there is no evidence that a Bayesian approach produces solutions superior to those gotten by a non-Bayesian methods, problem solvers clearly prefer approaches that get them closest to the problem in the simplest way.

Papers on current Bayesian applications focus primarily on the machinery—on the selection of the priors and on how the MCMC was run, for example—and pay less attention to the shape of the problem and nature of the data. In higher-dimensional problems with, say, over a few dozen parameters, researchers typically do not select priors by expert opinion about the distribution of the parameters. Rather, they make selections in terms of technical considerations that have little to do with the problem's context.

The Bayesian claim that priors are the only (or best) way to incorporate domain knowledge into the algorithms is simply not true. Domain knowledge is often incorporated into the structure of the method used. For instance, in speech recognition, some of the most accurate algorithms consist of neural nets whose architectures were explicitly designed for the speech-recognition context. In handwritten digit recognition, one of the most accurate algorithms uses nearest-neighbor classification with a distance that is locally invariant to things such as rotations, translations, and thickness.

Adventuresome tinkering

Incorporating domain knowledge into the structure of a statistical procedure is a much more immediate and intuitively appealing approach than setting up the Bayes machinery. It lets you strike at the heart of the problem and focuses your attention on what is important, not on technical aspects of machinery. Many people, including me, approach problems by trying this, trying that, tinkering here, tinkering there, seeing what works and what doesn't. Adventuresome tinkering goes painfully slowly if you are running MCMC on a high-dimensional parameter space.

If you are trying to solve a high-dimensional problem using neural nets, there are generally hundreds or thousands of parameters. Bayesian machinery treats this situation by piling hyperparameters onto hyperparameters onto parameters, then running a long MCMC to evaluate the posterior. As Christopher Bishop points out in his sympathetic overview of Bayesian methods, even the claim that the Bayesian approach picks the model of appropriate complexity for the data does not hold up in terms of picking the best predictive model.[1]

Problem solvers have little motivation, then, to use Bayesian methods. What

makes many skeptical about Bayesian methodology is the dearth of impressive success stories that could provide motivation. All it would take to convince me are some major success stories in complex, high-dimensional problems where the Bayesian approach wins big compared to any frequentist approach.

The last few years have seen major successes in terms of methods that give improved pattern-recognition prediction accuracy on a large variety of data sets. One such is the concept of support-vector machines originated by Vladimir Vapnik.[2] Another comes from the idea of perturbing and combining predictors. Both are based firmly on frequentist ideas. It is not at all clear if they can be understood, derived, or successfully implemented in a Bayesian framework.

My message to the Bayesians is that I can be convinced, but it's going to be a hard sell.

Other viewpoints

Wray Buntine's discussion is quite interesting. Although it looks like we would take similar approaches when faced with live problems growling at us, we base our conceptualizations on different frameworks. The proof of the pudding might be how we would teach a graduate course in "using complex data to solve problems." I suspect that differences would surface in the first week, but they might simply be different paths to the same goal.

David Draper and David Madigan give a more traditional Bayesian viewpoint. They begin with a presentation of Bayesian ideas reminiscent of many past theology discussions in statistical journals. I will not deal with these because my orientation is how results are gotten in practice. So I comment only on the parts of their essay that are relevant to this goal.

They do not get the meaning of "success story." In my view, a success story does mean that someone has managed to do a Bayesian analysis of a data set. Simply listing some Bayesian analyses in different fields does not cut it. A success story is a tough problem on which numbers of people have worked where a Bayesian approach has done demonstrably better than any other approach. For instance, they cite no tough, complex prediction problem where a Bayesian analysis has produced significantly more accurate test-set results than anything else tried.

In my experience, Bayesian analyses often are demonstration projects to show that a Bayesian analysis could be carried out. Rarely, if ever, is there any comparison to a simpler frequentist approach. But, I emphasize again, problem solvers need to be pragmatic and nonideological. If I encounter a situation where a Bayesian approach seems useful, I will try it.

In answer to my doubts regarding the selection of priors, Draper and Madigan have the following advice:

- Try several priors and see if they lead to substantially the same conclusion: start with one prior, run the MCMC, pull another prior out of the hat, run MCMC, and keep going until satisfaction sets in.
- Tune the prior to get good predictive results: start with one prior, run MCMC, compute the test-set error. Try another prior, run MCMC, and see what the new test-set error is. Keep going until you achieve a low test-set error.
- On small data sets that they assert frequentists can't handle, a method that works well (they say) is to use vague priors: priors that contain no prior information about the parameters. We have to take their word on this because they give no examples or citations. But this seems to contradict the Bayesian ideology.

No matter how you put it, their advice requires a lot of machinery, computing time, experience, and good instincts to get it right.

Two minor points:

- They note that Bayesian model averaging goes back to the 1960s, so that the success of "perturb and combine" methods is not really a solely frequentist success story. This is like saying that because Leonardo da Vinci thought of the possibility of human flight, the credit for flight does not solely belong to Orville and Wilbur Wright. Bayesian model averaging has been around for awhile but was never taken up seriously because it only modestly improved accuracy while requiring a lot of machinery. Recent frequentist methods such as bagging (Breiman[3]) and boosting (Freund and Schapire[4]) dramatically improve accuracy while requiring only trifling amounts of machinery. These methods are not simple model averaging.
- I was taken aback by their dismissal of

David Draper is a reader in the Statistics Group of the School of Mathematical Sciences at the University of Bath. His research interests include Bayesian inference and prediction; model uncertainty and exploration of the mapping from statistical assumptions to conclusions; the theory of data analysis; hierarchical modeling, causal inference, and MCMC methods; and applications of statistical methods to the social, medical, and biological sciences. He received his BSc in mathematics at the University of North Carolina, Chapel Hill, and PhD in statistics from the University of California, Berkeley. Contact him at the Statistics Group, School of Mathematical Sciences, Univ. of Bath, Claverton Down, Bath BA2 7AY, U.K.; d.draper@maths.bath.ac.uk; http://www.bath.ac.uk/~masdd/.

David Madigan is an associate professor in the Department of Statistics at the University of Washington and also the founder of Talaria Inc., a Seattle-based software company specializing in medical education. His research interests include statistical graphics, knowledge-based systems, computer-assisted learning, information retrieval, and hypermedia. He received his BA in mathematical sciences and his PhD in statistics from Trinity College, Dublin. He is an associate editor of the *Journal of the Royal Statistical Society* and the *Journal of Computational and Graphical Statistics*. Contact him at the Dept. of Statistics, Box 354322, Univ. of Washington, Seattle, WA 98195-4322; madigan@stat.washington.edu; http://bayes.stat.washington.edu.

Leo Breiman is the director of the Statistical Computing Facility at the University of California, Berkeley, where he is also Professor Emeritus of Statistics. His research interests include computationally intensive multivariate analysis, including the use of nonlinear methods for pattern recognition and prediction in high-dimensional spaces. He received his PhD in mathematics from the University of California, Berkeley.He is coauthor of *Classification and Regression Trees*, which introduced the concepts embodied in the CART program. Contact him at the Statistical Computing Facility, Evans Hall, Univ. of California, Berkeley, CA. 94720; leo@stat.berkeley.edu; http://www.stat.berkeley.edu/users/breiman/.

Wray Buntine is Research Engineer with the CAD group in the Electrical Engineering and Computer Science Department at the University of California, Berkeley, and Vice President of R&D at Ultimode Systems. His research interests include probabilistic machine-learning methods and theory, data mining, graphical models, and inductive logic programming. He received a PhD in computer science from the University of Technology, Sydney. He is on the editorial boards of *The International Journal of Knowledge Discovery and Data Mining* and the *Journal of New Generation Computing*. Contact him at 555 Bryant St. #186, Palo Alto, Calif. 94301; wray@ultimode.com; http://www.ultimode.com/~wray.

the journals listed in CIS as "statisticians talking to each other." Is their implication that the many good applied statisticians in a wide variety of fields are universally wrong in not using Bayesian methods? An equally valid reading might be that the longer Bayesian methods are known to researchers in an area, the less they are used.

References

1. C. Bishop, *Neural Nets for Pattern Recognition*, Claredon Press, Oxford, U.K., 1995.
2. V. Vapnik, *The Nature of Statistical Learning Theory,* Springer, New York, 1995.
3. L. Breiman, "Bagging Predictors," *Machine Learning,* Vol. 26, No. 2, 1996, pp. 123–140.
4. Y. Freund and R. Schapire, "Experiments with a New Boosting Algorithm," *Machine Learning: Proc. 13th Int'l Conf.,* 1996, pp. 148–156.

Bayesian in principle, but not always in practice

Wray Buntine EECS, UC Berkeley and Ultimode Systems

Before considering the whys and wherefores of Bayesian reasoning in practice, let's first consider the different uses of the term and the role that Bayesian reasoning should play for the scientific community.

Probabilistic reasoning or modeling, together with its partner, decision theory, is a key theoretical tool for addressing uncertainty in intelligent systems of all kinds. Probabilistic reasoning comes in many shapes and sizes, and does not necessarily imply the use of Bayesian methods proper, as I'll discuss. However, probabilistic methods are undoubtedly the most important family of tools in the general modeling of intelligent systems, including for machine learning, neural networks, speech recognition, natural language, and, increasingly, theoretical computer science, integrated circuits and computer architectures, robotics, and vision. Fuzzy logic—one supposed replacement for probabilistic modeling—is at best an important engineering tool that provides a methodology for techniques such as *control by interpolation from cases* and other neat tricks. Consequently, I would expect that all members of this Trends & Controversies discussion—Leo Breiman, David Draper, David Madigan, and I—would be strong proponents of probabilistic modeling.

For the statistician, Bayesian methods incorporate the use of *prior probabilities*, which include, for instance, weighting schemes for judging neural networks before seeing any data and the use of an expert's subjective opinion on different outcomes, again before having seen any data. (My Website contains a detailed tutorial and reference list on prior probabilities.[1]) Breiman's critique is largely about Bayesian methods in this statistical sense. Draper and Madigan mention some of the more creative statistical uses of Bayesian methods.

Confusingly, Bayesian classifers—now the darling of the machine-learning community and coming to the corporate world through SGI's MindSet program for knowledge discovery—are probabilistic models. They rarely incorporate Bayesian methods in the accepted statistical sense and, in many cases, have no need to.

Equally confusingly, Bayesian networks, as used by the uncertainty in artificial intelligence (UAI) community, are likewise not Bayesian in the accepted statistical sense. As a form of probabilistic model, they are a subset of the more general and rich family of graphical models. Unfortunately, many in the neural networks, expert systems, and machine-learning communities see these graphical models merely as an alternative to a feed-forward network, rule-based system, decision tree, or linear regression. What an undersell! The UAI community has done itself a great disservice by not representing graphical models in their full glory. They are a powerful representational tool for probabilistic modeling in general that encompasses both the functional and axiomatic properties of issues such as independence, problem decomposition, and mapping knowledge. This is fundamental stuff that all students of intelligence and computation should learn early in their graduate training. Graphical models are also a beautiful match for Bayesian methods when modeling a new problem. (See my Website for a fuller discussion of statistical computation via graphical models.[2])

The statistician's view

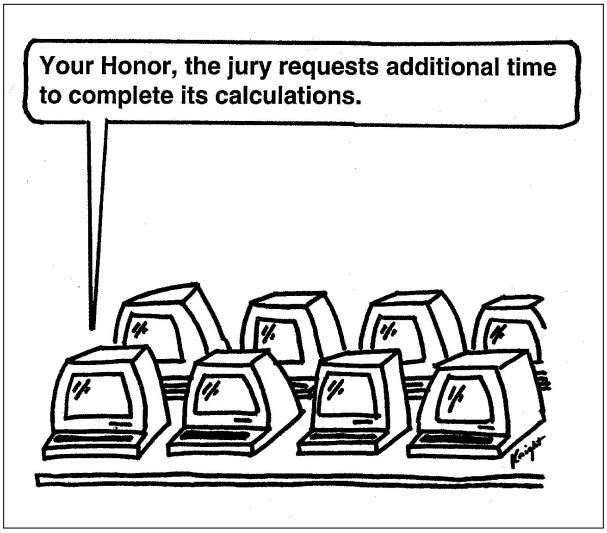
Having considered these other kinds of Bayesians, what then of Bayesian reasoning in the statistician's sense? Thankfully, in many parts of the academic community, the long Bayesian versus non-Bayesian wars have ended. Bayesian methods are now a prominent theory in areas of economics and statistics. They provide a coherent theoretical framework that, given certain qualifications, must be a central theory of statistics and, more generally, of intelligent systems. This claim rests on the theory of rationality, which states, more or less, that *a single agent acting under uncertainty but with infinite computational resources should behave in a manner consistent with Bayesian decision theory.*

Many of the so-called problems and paradoxes people find with Bayesian methods arise from mistakes or poor modeling. The inherent consistency of Bayesian methods means that if you are dealing with a single agent, and if you are ignoring tractability, producing a paradox is almost impossible. (My online tutorial lists some notable bloopers by distinguished professors.)

Bayesian theory allows ample room for flexibility. Vladimir Vapnik's support-vector machines, which have achieved considerable practical success, are a recent shining example of the principle of rationality and thus of Bayesian decision theory. You do not have to be a card-carrying Bayesian to act in agreement with these principles. You only have to act in accord with Bayesian decision theory. Academics love to invent their own paradigms: machine learning, for instance, has a half dozen competing paradigms at last count. Competing paradigms are the stuff of which tenure and journals are made. A great academic industry has arisen for converting algorithms and methods from Bayesian to MDL, from maximum likelihood to Bayesian, and from PAC to Bayesian, for example. (Yes from my IJCAI'89 paper, I am guilty—as are many others! Good tutorial articles here can be found at Jon Oliver's web site at Monash University.) However, the principle of rationality merely says that our behavior—the predictions we make or the actions we take—should be approximately the same as those of some Bayesian methods. Support-vector machines have an elegant explanation in Bayesian theory; hence, they agree with it as far as necessary.

Due to the qualifications applied to the justification for rationality, you should be aware of the limitations on applying Bayesian methods. For instance:

- If the network is the computer, as Sun Microsystems would have us believe, then some computing is not done by a single agent but by a bunch of them.



- When the US Surgeon General makes statements on behalf of the US Government for the US public based on statistical information, she is not acting as a single agent but as a representative of many.
- Computational resources are not infinite, as any practitioner of MCMC methods will readily attest.

Bayesian decision theory certainly has its limitations, but its advantages can often overcome these.

From theory to practice

But that is theory. What of practice? Breiman summarizes the issues precisely: "But the biggest reason that Bayesian methods have not been used more is that they put another layer of machinery between the problem to be solved and the problem solver." I agree completely. Moreover, if renowned scientists can stumble over the complexity of Bayesian modeling (as many do, in public, on a regular basis), how can we expect our fresh graduates heading off to industry to comprehend and apply the same techniques?

Let's consider an analogy with quantum theory. Quantum theory might be an important basis of modern physics, but the pool shark at the local billiards emporium no more needs quantum theory to improve his shot than does the Great White Shark to improve his golf game. And engineers get by with good old Newtonian physics—does that invalidate quantum physics? No, certainly not. Is quantum theory the only true theory of physics, and all others imposters? Certainly not. Can other theories produce good results with much less intellectual effort or lead to superior results? Certainly.

While Bayesian decision theory is inherently more useful than quantum physics, it is not the only framework on which practitioners should rely, nor is it the only theory that we should use. I employ a number of skills in my practice: clever optimization, frequentist methods when I believe sufficient data exists, software-engineering skills, and minimum encoding methods if the particular Bayesian approach seems too impenetrable, among others. I only pull out my Bayesian-methods tool when I need it. Others might use alternative methods, such as minimum description length or one of the other competing frameworks, which often give similar results to a Bayesian method. For a good motivating example see the work by Rohan Baxter and Jon Oliver.[3] Who cares which statistical framework they use so long as they get good results? I often find Bayesian methods easier to use primarily because of my real analysis background. Others, especially computer scientists, prefer coding techniques. From a hundred feet, we agree on what needs to be done. That's a good thing.

The true Bayesian is not fanatical but understands the principles of rationality and its shortcomings. In practice, the true Bayesian could use any one of a number of techniques and theories. In matters of theory, I am a Bayesian; in practice, I'll use whatever works. Not infrequently, I find the signposts that Bayesian theory leaves planted all over a domain lead me to results faster than my colleagues who use other approaches.

For instance, in investigating a new problem, it's important to understand key prior knowledge from which you can gain advantage. Unfortunately, only certain individuals can design quality multidimensional prior probabilities, an issue I discuss in detail in the online tutorial. In practice, it's a good bet instead to encode your prior probabilities into your model's structures and transformations.

As a practical Bayesian then, I practice and preach exactly what Leo Breiman does: "Incorporating domain knowledge into the structure of a statistical procedure is a much more immediate ... approach than setting up the [full] Bayes machinery." This is one of many important principles that Bayesian theory can provide the practitioner. Averaging over multiple models is another way to improve the performance of predictive models such as neural networks, something I first realized in 1987, along with a few other pioneers of the multiple models field in machine learning.

More advocates coming

Neither the fanatical Bayesian nor the fanatical anti-Bayesian fully appreciates Bayesian theory and its implications. Perhaps the theory's biggest contribution is its ability to unify competing theories and provide a common language for them. Those with paradoxes about Bayesian theory invariably get egg on their face. But for practitioners, all this doesn't mean much because too few applied Bayesians are developing the software and methodology necessary to make the approach more readily usable by the engineer in the field. As Brieman says, you don't want to put excess machinery between the problem and the problem solver. So, while I firmly believe that all our PhD students need a heavy dose of Bayesian methods early in their graduate training, I believe practitioners of intelligent systems need more friendly tools and applied methodologies that incorporate Bayesian reasoning behind the scenes. Fortunately, the language of graphical models provides a perfect basis for building a probabilistically oriented programming language and for automating many of the steps in Bayesian reasoning, thus making the whole processes easier and more efficient. Many of us believe that Bayesian reasoning will see even more advocates outside the growing core of statisticians as these computational capabilities are realized and as probabilistic methods infiltrate computer science proper.

References

1. W. Buntine, "Prior Probabilites," NATO Workshop on Learning in Graphical Models; http://www.ultimode.com/~wray/refs.html#talks.
2. W. Buntine, "Comptation with the Exponential Family and Graphical Models," NATO Workshop on Learning in Graphical Models; http://www.ultimode.com/~wray/refs.html#talks.
3. R. Baxter and J. Oliver, "The Kindest Cut: Minimum Message Length Segmentation," *Proc. Seventh Int'l Workshop Algorithmic Learning Theory, LCNS,* Vol. 1160, Springer-Verlag, Berlin, 1996.

Wray Buntine is Research Engineer with the CAD group in the Electrical Engineering and Computer Science Department at the University of California, Berkeley, and Vice President of R&D at Ultimode Systems. His research interests include probabilistic machine-learning methods and theory, data mining, graphical models, and inductive logic programming. He received a PhD in computer science from the University of Technology, Sydney. He is on the editorial boards of *The International Journal of Knowledge Discovery and Data Mining* and the *Journal of New Generation Computing*. Contact him at 555 Bryant St. #186, Palo Alto, Calif. 94301; wray@ultimode.com; http://www.ultimode.com/~wray.