Comments on Lectures by Jim Berger and Susie Bayarri, Part 2

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu
www.ams.ucsc.edu/∼draper

CBMS-MUM
Santa Cruz CA

26 July 2012

The Big Picture

The ingredients in a problem involving Bayesian inference, prediction and/or decision-making:

• θ, something unknown and of interest to You (a generic person wishing to reason SENSIBLY in the presence of uncertainty) — for all finite-dimensional unknowns, θ can be expressed as a vector $(\theta_1, \ldots, \theta_k)$ of real numbers for integer $1 \leq k < \infty$;

• An information source (data set) y that You judge to be RELEVANT to decreasing Your uncertainty about θ — y can always be expressed as a vector $(y_1, \ldots, y_n)$ of real numbers for integer $1 \leq n < \infty$;

• A set $\mathcal{B}$ of true-false propositions, all judged by You to be true, summarizing background information about the context of the problem and the design of the data-gathering process.

Your job: to synthesize all RELEVANT information about θ, to produce

(I) ACCURATE inferences about θ,

(II) ACCURATE predictions of future data y^*, and

The Big Picture (continued)

(III) GOOD decisions about what actions to take in the face of Your uncertainty about θ.

Theory (de Finetti, RT Cox) says that the best way to do this is via

(a) (inference, prediction) conditional probability distributions of the form $p(\theta|\mathcal{B})$, $p(y|\theta\,\mathcal{B})$, and $p(\theta|y\,\mathcal{B})$ and

(b) (decision) a set $\mathcal{A}$ of possible actions and a utility function $U(a,\theta)$ (taking real values, with big utilities better than small ones by convention) quantifying the costs and benefits of choosing action a when the unknown is really θ.

The existence of y dichotomizes Your total information about θ into two sources: internal and external to y.

$p(y|\theta\,\mathcal{B})$, commonly called Your sampling distribution or likelihood, quantifies Your information about θ internal to y.

$p(\theta|\mathcal{B})$, commonly called Your prior distribution, quantifies Your information about θ external to y.

The Big Picture (continued)

(I) A theorem of Bayes and Laplace then says that $p(\theta|y\,\mathcal{B})$, commonly referred to as Your posterior distribution, satisfies

$$p(\theta|y\,\mathcal{B}) \propto p(\theta|\mathcal{B})\,p(y|\theta\,\mathcal{B})\,; \tag{1}$$

this solves the inference problem.

(II) Calculation reveals that

$$p(y^*|y\,\mathcal{B}) = \int_\Theta p(y^*|\theta\,y\,\mathcal{B})\,p(\theta|y\,\mathcal{B})\,d\theta\,, \tag{2}$$

where Θ is the set of possible θ values; this solves the prediction problem.

(III) A theorem of Ramsay says that the optimal decision maximizes expected utility, where the expectation is over Your posterior distribution:

$$a^* = \operatorname*{argmax}_{a\in\mathcal{A}} E_{(\theta|y\,\mathcal{B})}\,U(a,\theta) = \operatorname*{argmax}_{a\in\mathcal{A}} \int_\Theta U(a,\theta)\,p(\theta|y\,\mathcal{B})\,d\theta\,; \tag{3}$$

this solves the decision problem (as long as "You" represents a single individual or a collective of individuals that agree either on the utility function or the posterior distribution (Kadane)).

The Bayesian Specification Problem

To recap, You have to specify two things for inference and prediction — Your prior $p(\theta|\mathcal{B})$ and Your likelihood $p(y|\theta\,\mathcal{B})$ — and two more things for decision — Your action space $\mathcal{A}$ and Your utility function $U(a,\theta)$ — and equations (1–3) provide good solutions to inferential, predictive and/or decision-making problems in science, policy and business.

Interestingly, the de Finetti-RT Cox theory is largely silent on HOW to specify these ingredients WELL; this entire conference has been about this specification problem (focusing mainly so far on Your prior and Your likelihood, and in fact mainly on Your prior).

As a profession we do not yet have fully satisfying solutions to the specification problem; progress on this front would move us toward a Theory of Applied Statistics, which we need and do not yet have (see, e.g., the copy of my talk at the AFM Smith conference in Crete in June 2011, on my web page).

We're still at the stage of articulating principles, which sometimes lead to axioms and then theorems that suggest ways to arrive at GOOD specifications.

The Calibration Principle

There is not yet universal agreement on the principles; Jim and Susie have advocated some this week, and I'd like to mention a few others.

I'll call $\{p(\theta|\mathcal{B}), p(y|\theta\,\mathcal{B}), \mathcal{A}, U(a,\theta)\}$ Your model M for Your uncertainty about θ; the basic problem here is that You're uncertain about θ, but You're also uncertain about how to specify Your uncertainty about θ (this is model uncertainty).

(Calibration Principle) In model specification, it helps to know something about how often the methods You're using to choose one model over another get the right answer, and this can be ascertained by creating simulation environments (structurally similar to the one You presently find Yourself in, scientifically) in which You know what the right answer is and seeing how often Your methods recover known truth.

The reasoning behind the Calibration Principle is as follows:

(axiom) You want to help positively advance the course of science, and repeatedly getting the wrong answer runs counter to this desire.

Well-Calibrated Bayes Versus "Objective" Bayes

(remark) There's nothing in the Bayesian paradigm to prevent You from making one or more of the following mistakes — (a) choosing $p(y|\theta\,\mathcal{B})$ BADLY; (b) inserting {strong information about θ external to y} into the modeling process that turns out after the fact to have been out of step with reality; (c) choosing a BAD action space; (d) choosing a BAD utility function — and repeatedly doing this violates the axiom above.

Jim and Susie talk a lot about "objective Bayesian modeling"; but

(A) I believe that there's no such thing as "objective Bayesian modeling" (dictionary definition of objective (adjective): not influenced by personal interpretations; but see the use of the word JUDGE back on page 2);

(B) I think that what Jim and Susie mean by "objective Bayesian modeling" is actually well-calibrated Bayesian modeling — paying attention to how often You get the right scientific answer; and

(C) I'm convinced that words matter (the terms we use frame our discourse, so we should choose them carefully).

The Modeling-As-Decision Principle

Personally I subscribe to the Calibration Principle, and I'm a fan of well-calibrated Bayesian modeling (see, e.g., Berger J (2006) The case for objective Bayesian analysis. Bayesian Analysis 1 385–402, and my comment on that paper in the same issue).

Next principle — in dealing with model uncertainty, we find ourselves repeatedly asking the following questions:

$\boxed{Q_1}$: Is model M_1 better than M_2?

$\boxed{Q_2}$: Is M_1 good enough (to stop looking for a better model)?

These questions sound fundamental but are not: better for what purpose? Good enough for what purpose?

This implies (see, Bernardo and Smith; Draper; Key et al.) a

Modeling-As-Decision Principle: Making clear the purpose to which the modeling will be put transforms model specification into a decision problem, which should be solved by maximizing expected utility with a utility function tailored to the specific problem under study.

Bayes Factors and Log Scores

Some examples of this may be found (e.g., Fouskakis and Draper, 2008 JASA: variable selection in generalized linear models under cost constraints), but this is hard work; there's a powerful desire for generic model-comparison methods whose utility structure may provide a decent approximation to problem-specific utility elicitation.

Two such methods are Bayes factors (Jim and Susie have focused entirely on this approach, but there are other model-comparison methods out there) and log scores.

As Jim noted, Bayes factors have a utility justification based on (a) a 0–1 utility function and (b) maximizing posterior model probabilities with equal prior model probabilities; as Jim noted, these are not always the best choices.

Log scores derive from a

Prediction Principle: Good models make good predictions, and bad models make bad predictions; that's one scientifically important way You know a model is good or bad.

Log Scores

In the simplest case, in which $y = (y_1, \ldots, y_n)$ are real values modeled exchangeably, two versions of log scores are based on a cross-validation (CV) concept (Ibrahim and Laud; Gelfand),

$$LS_{CV}(M_j|y\,\mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|y_{-i}\,M_j\,\mathcal{B})\,, \tag{4}$$

in which y_{-i} is the y vector with observation i omitted, and a full-sample (FS) concept (Draper and Krnjajić),

$$LS_{FS}(M_j|y\,\mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|y\,M_j\,\mathcal{B}) \tag{5}$$

(LS_{FS} uses the data twice, but only mildly so for small n and negligibly so for moderate to large n; surprisingly, Draper and Krnjajić (submitted) have shown that LS_{FS} can have better small-sample model discrimination ability than LS_{CV} (in addition to being much faster to approximate in a stable and accurate way)).

Log scores have a direct utility basis in which the utility function rewards predictive accuracy, and — if, thinking calibratively, You posit

Model Comparison

an underlying data-generating model M_{DG} — $LS_{FS}(M_j|y\,\mathcal{B})$ also has a nice interpretation as an approximation to the Kullback-Leibler divergence between M_{DG} and $p(\cdot|y\,M_j\,\mathcal{B})$, in which M_{DG} is approximated by the empirical CDF:

$$\begin{aligned} KL[M_{DG}||p(\cdot|y\,M_j\,\mathcal{B})] &= E_{M_{DG}} \log M_{DG} - E_{M_{DG}} \log p(\cdot|y\,M_j\,\mathcal{B}) \\ &\doteq E_{M_{DG}} \log M_{DG} - LS_{FS}(M_j|y\,\mathcal{B})\,; \end{aligned} \tag{6}$$

the first term on the right side of (6) is constant in $p(\cdot|y\,M_j\,\mathcal{B})$, so minimizing $KL[M_{DG}||p(\cdot|y\,M_j\,\mathcal{B})]$ is approximately the same as maximizing LS_{FS}.

So now we have two model-comparison methods (Bayes factors and log scores); which is BETTER?

Actually there are many more than two in routine use, not all of them Bayesian; here I'll contrast the behavior of five of them:

$\{$Bayes factors, BIC$\}$ versus $\{$AIC, DIC, $LS_{FS}\}$

BIC is a special case of Bayes factors that employs a Laplace-type approximation to the log marginal likelihood with a

Model Comparison (continued)

unit-information prior; BIC, AIC and DIC all involve an explicit trade-off between model lack of fit and model complexity, differing in how they measure these two quantities.

From a calibration point of view, one model-comparison method is better than another in contrasting M_1 with M_2 when its repeated-sampling rates of identification of the actual data-generating model M_{DG} are better than those of its competitors under a variety of relevant scenarios.

Toy Example 1: $y = (y_1, \ldots, y_n)$, y_i non-negative integer:

$M_1 =$ Geometric(γ_1) likelihood ($\gamma_1 = \theta_1$) with a conjugate prior on θ_1;

$M_2 =$ Poisson(γ_2) likelihood ($\gamma_2 = \theta_2$) with a conjugate prior on θ_2.

Here $M_{DG} =$ Geometric(γ_1) with $\gamma_1 = 3$ would be a special case of M_1; note that M_1 and M_2 have the same degree of complexity (have parameter vectors of equal dimension).

False-Positive and False-Negative Model Comparison Errors

Toy Example 2: $y = (y_1, \ldots, y_n)$, y_i real:

$M_1 = N(\gamma_1)$ likelihood $(\gamma_1 = (\mu_1, \sigma_1^2), \mu_1 = 0)$ with a conjugate prior on σ_1^2;

$M_2 = N(\gamma_2)$ likelihood $(\gamma_2 = (\mu_2, \sigma_2^2))$ with a conjugate prior on (μ_2, σ_2^2).

Note here that M_2 is more complicated (has a parameter vector of higher dimension) than M_1.

When comparing M_1 and M_2, let's agree to say that $\{$choosing M_2 when M_{DG} is a special case of $M_1\}$ is a false-positive mistake, and $\{$choosing M_1 when M_{DG} is a special case of $M_2\}$ is a false-negative mistake.

Many people estimate these false-positive and false-negative error rates asymptotically, and call a model-comparison method asymptotically consistent at model M_j if the repeated-sampling rate at which the method chooses M_j goes to 1 for all γ_j as $\{n \to \infty$ while M_{DG} remains fixed at $M_j(\gamma_j)\}$.

However, this form of asymptotics is irrelevant to actual applied practice, for three reasons:

Irrelevance of Usual Asymptotics to Applied Practice

• When Your data set has (e.g.) $n = 82$, You don't care what happens when $n = \infty$, and

• In actual applied practice, as n increases You'll need to consider more and more complicated parametric models to have a hope of "keeping up with reality," so it's unrealistic to keep M_{DG} fixed as n grows.

• If (as is often true) Your data-gathering process takes time, it can be actively dangerous to pretend that the data continue to be conditionally IID from the same model, because this implies that the process You're observing is stationary and time-homogeneous, and this assumption can be laughable.

In contrast to an opinion Jim expressed, in my view it's usually far more useful to make calculations or run simulations (at analysis time) for YOUR (finite) value of n or (at design time) over a range of finite sample sizes that's realistic to Your actual experiment, comparing false-positive and false-negative error rates by varying M_{DG} appropriately.

When You do this (see, e.g., the talk (on my web page) I gave at the SBIES meeting earlier this year), You find the following:

{Bayes factors, BIC} Versus {AIC, DIC, log scores}

- None of {Bayes factors, BIC, AIC, DIC, LS_{FS}} uniformly dominates the others simultaneously on false-positive and false-negative performance.

- {Bayes factors with sensible priors, BIC} behave similarly with respect to false-positive and false-negative rates.

- {AIC, DIC, log scores} behave similarly.

- {Bayes factors, BIC} behave differently from {AIC, DIC, log scores}.

- {AIC, DIC, log scores} tend to make more false-positive mistakes than {Bayes factors, BIC} when M_2 is more complicated than M_1, and under those same conditions {Bayes factors, BIC} tend to make more false-negative mistakes than {AIC, DIC, log scores}.

- To choose a model-comparison method well in Your problem, You need to think about the real-world consequences of false-positive and false-negative mistakes; in other words, choosing a model-comparison method is itself a decision problem!

Example of a common inferential setting in which false-negative mistakes are worse than false-positive errors:

False-Negatives Can Be Worse Than False-Positives in Inference

$y = (y_1, \ldots, y_n)$, y_i non-negative integer; fixed-effects (M_1) versus random-effects (M_2) Poisson modeling:

$$M_1: \left\{ \begin{array}{ccc} (\lambda|\mathcal{B}) & \sim & p(\lambda|\mathcal{B}) \\ (y_i|\lambda,\mathcal{B}) & \overset{\text{IID}}{\sim} & \text{Poisson}(\lambda) \end{array} \right\}; \tag{7}$$

$$M_2: \left\{ \begin{array}{ccc} (\beta_0, \sigma^2|\mathcal{B}) & \sim & p(\beta_0, \sigma^2|\mathcal{B}) \\ (y_i|\lambda_i,\mathcal{B}) & \overset{\text{indep}}{\sim} & \text{Poisson}(\lambda_i) \\ \log(\lambda_i) & = & \beta_0 + e_i \\ (e_i|\sigma^2,\mathcal{B}) & \overset{\text{IID}}{\sim} & N(0,\sigma^2) \end{array} \right\}; \tag{8}$$

M_1 is of course a special case of M_2 with $\left(\sigma^2 = 0, \lambda = e^{\beta_0}\right)$.

Claiming that (a special case of) M_1 is M_{DG} when actually the data came from (a special case of) M_2 (a false-negative error) has far more serious inferential consequences in models like this than making a false-positive mistake: when You conclude that $\sigma^2 = 0$ and actually $\sigma^2 > 0$, Your uncertainty bands for λ will be narrower than they should be (and this effect can be large when n is small).

Both False-Positives and False-Negative Matter in Multiplicity Settings Too

In problems involving finding signal(s) when searching among many noisy channels, let's agree to call {claiming a signal is there when it's really not} a false-positive error and {failing to find a signal when it's really there} a false-negative mistake.

Jim has shown us some wonderful work in which (a) the behavior of Bayes factors when testing sharp-null hypotheses and (b) careful choice of prior distributions can greatly help control false-positive errors.

The point I'd like to make here is that in many of these problems, false-negative mistakes matter too, and Bayesian decision theory is the best way to trade the two kinds of errors off against each other to find an optimal compromise.

Example: When looking for new drugs that are better than existing treatments for (e.g.) hypertension (or new genes that may regulate a given disease), failing to find a new "killer" drug (or gene) when it's out there to be found can be financially disastrous for a drug company (business), and this sort of mistake should worry the FDA too (policy),

The Decision-Versus-Inference Principle

because patients will receive sub-optimal care during the period of time between when drug company 1 fails to find the "killer" drug (or gene) and drug company 2 finds it.

With students at UCSC and investigators at the Swiss drug company Roche, I'm working on this approach now (talks and papers expected to be available fairly soon), and I bet others are doing so too.

This is an example of a general pattern (see, e.g., my AFM Smith meeting talk mentioned earlier, on my web page), which is summarized in the

Decision-Versus-Inference Principle: It's helpful to avoid using inferential methods to try to solve problems that really have a decision-making character: the implicit utility structure in standard inferential methods (e.g., hypothesis testing) is often far from optimal.

Three additional discussion questions for Jim and Susie:

- If I modify your variable-selection methods by using BIC instead of {Bayes factors with your current favorite variable-selection prior},

Additional Discussion Questions

holding the rest of your approach constant, how will my BIC method compare with yours in terms of false positive and false negative rates?

• Do you find yourself sometimes recommending the use of different priors for estimation and for testing in the same problem, and if so how do you reconcile the differences?

• How do we reconcile the Calibration Principle with our desire to use strong contextual prior information when it's available?

Note added after the discussion: I strongly object to Jim's characterization of log scores as un-Bayesian; they arise as the Bayesian solution to the model-comparison problem treated decision-theoretically, with a utility function rewarding models that predict observables well (Jim would disagree with this statement, but I don't find the argument on which his disagreement is based compelling).

I encourage You to give log scores a try (they're much easier to calculate than Bayes factors) in simulations in which You vary M_{DG} and see how various model-comparison methods make the false-positive/false-negative trade-offs.