The Bayesian Approach to Discovery

James Berger
Duke University, Durham NC, USA

Abstract
The Bayesian approach to discovery is essentially the Bayesian approach to hypothesis testing. This is discussed, through a pedagogical example that illustrates the approach and the differences with non-Bayesian approaches to testing, and through the Bayesian formulation of the generic HEP problem. Ensuing discussion focuses on the potential value of the Bayesian approach in dealing with the highly problematical 'look-elsewhere' effect, and the major challenge to the Bayesian approach, which is the choice of suitable prior distributions for unknown model parameters. A brief discussion of Bayesian unfolding is also included.

1 Introduction

'Discovery' can mean many things, from discovery of a completely anticipated entity such as the Higgs boson, to discovery of completely unanticipated new physics. Bayesian analysis is relevant to both types of discovery, but here we focus primarily on the former, in that it is easier to discuss Bayesian analysis for anticipated events because it (typically) reduces to Bayesian hypothesis testing.

Even the discussion of Bayesian hypothesis testing is, however, rather spotty. We begin with a simple pedagogical example of Bayesian testing, to set the notation and emphasize important distinctions with non-Bayesian approaches. The Bayesian formulation of the generic HEP problem is then introduced, to provide a vehicle for discussion of the look-elsewhere effect (multiple testing in statistical language). We partly focus on this aspect of the Bayesian approach because of its potential for dealing with one of the most troubling issues in discovery in HEP.

A major difficulty in the implementation of Bayesian hypothesis testing is the choice of the needed prior distributions of unknown parameters. This is considerably more problematical than in Bayesian estimation (e.g., in the choice of upper confidence limits), since standard objective priors are not available. A complete discussion of this issue is beyond the scope of this paper, but some of the basic issues involved are discussed.

Finally, an idea concerning unfolding (deconvolution) is discussed, because it also relates to a long-studied Bayesian problem.

2 A pedagogical example of Bayesian testing

As background to later developments, we review the ideas of Bayesian testing, as discussed in [1] through a simple example; we borrow many of the details from that reference.

Suppose the data, X, is the number of events observed in time T that are characteristic of Higgs boson production in an LHC particle collision experiment. The probabilistic model for the data is that X has density

$$\text{Poisson}(x \mid s+b) = \frac{(s+b)^x e^{-(s+b)}}{x!},$$

where s is the mean rate of production of Higgs events in time T in the experiment and b is the (assumed known) mean rate of production of such events from background sources in time T. Two specific values of X and b that we will follow through various analyses are

Case 1: $x = 7$ and $b = 1.2$; *Case 2:* $x = 6$ and $b = 2.2$.

The main purpose of the experiment is supposedly to determine whether or not the Higgs boson exists which, in terms of the probability model for the data, is typically phrased as testing $H_0 : s = 0$ versus $H_1 : s > 0$. Thus H_0 corresponds to 'no Higgs.'

The p**-value** in this example, corresponding to observed data x, is

$$p = P(X \geq x \mid b, s = 0) = \sum_{m=x}^{\infty} \text{Poisson}(m \mid 0 + b)\,.$$

For the two cases,

Case 1: $p = 0.00025$ if $x = 7$ and $b = 1.2$; *Case 2:* $p = 0.025$ if $x = 6$ and $b = 2.2$.

There is general agreement that a small p-value indicates that something unusual has happened, but that the p-value does not have a direct quantitative interpretation as evidence against the null hypothesis. Thus Luc Demortier observed in his talk at the Phystat 07 conference:

> "In any search for new physics, a small p-value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery."

The **Bayes factor** of H_0 to H_1 in our ongoing example is given by

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0 + b)}{\int_0^\infty \text{Poisson}(x \mid s + b)\pi(s)\, ds} = \frac{b^x\, e^{-b}}{\int_0^\infty (s + b)^x\, e^{-(s+b)}\pi(s)\, ds}\,; \tag{1}$$

in the *subjective Bayesian approach*, the prior density, $\pi(s)$, is chosen to reflect the beliefs of the investigators (e.g., it could reflect the standard model predictions for the signal given information about the mass of the Higgs) while, in the *objective Bayesian approach*, it is chosen conventionally and nominally reflects a lack of knowledge concerning s.

A reasonable objective (though proper) prior here is the *intrinsic prior* $\pi^I(s) = b(s + b)^{-2}$ (see [1]). For this prior, the Bayes factor is given by

$$B_{01} = \frac{b^x\, e^{-b}}{\int_0^\infty (s + b)^x\, e^{-(s+b)} b(s + b)^{-2}\, ds} = \frac{b^{(x-1)}\, e^{-b}}{\Gamma(x - 1, b)}\,,$$

where Γ is the incomplete gamma function. The result for the two cases is

Case 1: $B_{01} = 0.0075$ (recall $p = 0.00025$); *Case 2:* $B_{01} = 0.26$ (recall $p = 0.025$)

The objective choice of prior probabilities of the hypotheses is $\Pr(H_0) = \Pr(H_1) = 0.5$, in which case

$$\Pr(H_0 \mid x) = \frac{B_{01}}{1 + B_{01}}\,.$$

For the two cases in the example,

Case 1: $\Pr(H_0 \mid x) = 0.0075$ (recall $p = 0.00025$); *Case 2:* $\Pr(H_0 \mid x) = 0.21$ (recall $p = 0.025$).

In addition to the uncertainty in the hypotheses, there is also uncertainty in s, given that H_1 were true. The complete posterior distribution is thus determined by

- $\Pr(H_0 \mid x)$, the posterior probability of the null hypothesis;
- $\pi(s \mid x, H_1)$, the posterior distribution of s under H_1.

For Case 1 in the example, Figure 1 presents these two parts of the full posterior distribution. One way of thinking of this is that the vertical bar gives the probability that one has just observed noise, while the density part says where s is likely to be if there is a discovery.

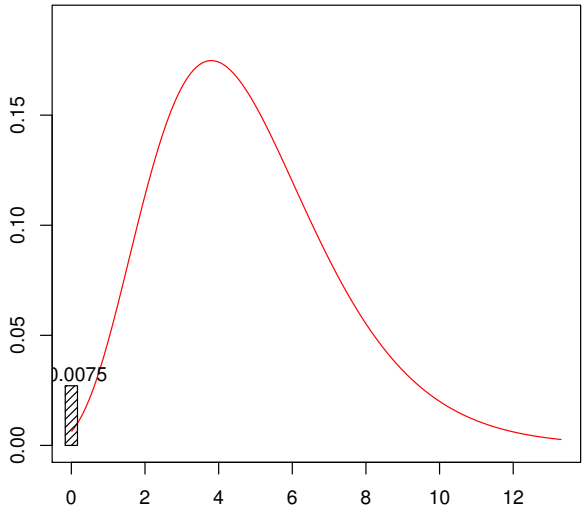

Fig. 1: For Case 1, $\Pr(H_0 \mid x)$ (the vertical bar), and the posterior density for s given $x = 7$ and H_1.

A useful summary of the complete posterior is $\Pr(H_0 \mid x)$ and C, a (say) 95% posterior confidence interval for s under H_1. For the two cases, and with C chosen to be an equal-tailed 95% posterior confidence interval (i.e., omitting 2.5% of the posterior mass on the left and the right)

Case 1: $\Pr(H_0 \mid x) = 0.0075$ and $C = (1.0, 10.5)$; *Case 2:* $\Pr(H_0 \mid x) = 0.21$ and $C = (0.2, 8.2)$.

C could, alternatively, be chosen to be a one-sided confidence bound, if desired.

Note that confidence intervals alone are *not* a satisfactory inferential summary. In Case 2, for instance, the 95% confidence interval does not include 0, and so many mistakenly believe that one can accordingly reject $H_0 : s = 0$. But, the full posterior distribution also has a probability of 0.21 that $s = 0$, which would hardly imply a confident rejection.

The Bayesian error probabilities given in the previous section differed from the corresponding p-values by factors of 30 and 10 in the two cases, respectively. It might be tempting to say that there is something wrong with the Bayesian analysis, but even a pure likelihood analysis reveals the same effect. In particular (following [2]), note that a lower bound on the Bayes factor over all possible priors can be found by choosing $\pi(s)$ to be a point mass at $\hat{s}$ (the maximum likelihood estimate), yielding

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0 + b)}{\int_0^\infty \text{Poisson}(x \mid s + b)\pi(s)\, ds} \geq \frac{\text{Poisson}(x \mid 0 + b)}{\text{Poisson}(x \mid \hat{s} + b)} = \min\{1, \left(\frac{b}{x}\right)^x e^{x-b}\}\,. \qquad (2)$$

In 'likelihood language,' this says that, for the given data, the likelihood of H_0 relative to the likelihood of H_1 is at least the bound on the right hand side of (2). For the two cases, this bound is

Case 1: $B_{01} \geq 0.0014$ (recall $p = 0.00025$); *Case 2:* $B_{01} \geq 0.11$ (recall $p = 0.025$),

so that a serious discrepancy remains even when the prior is eliminated. This is partly due to the fact that the p-value is based on the probability of the tail area of the distribution, rather than the probability of the actual data. For further discussion of the discrepancy (and problems with interpretation of p-values) see [1]. It is also shown there how the objective Bayesian posterior probabilities are also the optimal conditional frequentist error probabilities, so that both Bayesian and frequentist philosophies support the conclusion that p-values cannot in any sense be viewed as actual error rates.

3 A Bayesian formulation of the basic HEP problem

Following [3], a more complete model for the basic HEP problem can be summarized as follows. Let N be the observed Poisson number of events. The events are independent and each has characteristics ('marks' in the Poisson process world) X_i, $i = 1, \ldots, N$. These events could arise from only a background Poisson process, having mean b and density $f_b(x)$ (hypothesis H_0). In addition to background, there could also be an additive signal Poisson process with mean s and density $f_s(x)$ (hypothesis H_1). It is then of interest to test

H_0 : N has mean b, and the X_i have density $f_b(x) > 0$, versus

H_1 : N has mean $b + s$, and the X_i have density $(\gamma f_b(x) + (1-\gamma) f_s(x))$, where $\gamma = \frac{b}{(b+s)}$.

We will consider the case where $f_b(x)$ and $f_s(x)$ are known but b and s are unknown.

The Bayes factor of H_1 to H_0 for priors $\pi_0(b)$ and $\pi_1(b, s) = \pi_0(b)\pi_1(s \mid b)$ is

$$\begin{aligned} B_{10} &= \frac{\int_0^\infty \int_0^\infty (b+s)^N e^{-(b+s)} \prod_{i=1}^N [\gamma f_b(x_i) + (1-\gamma) f_s(x_i)] \pi_1(b, s)\, ds db}{\int_0^\infty b^N e^{-b} \prod_{i=1}^N [f_b(x_i)]\, \pi_0(b) db} \\ &= \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^N \left[1 + \frac{s f_s(x_i)}{b f_b(x_i)}\right] \pi_0(b)\pi_1(s \mid b)\, ds db}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db}\,. \end{aligned} \tag{3}$$

Note that, if b is known, this becomes

$$B_{10} = \int_0^\infty e^{-s} \prod_{i=1}^N \left[1 + \frac{s f_s(x_i)}{b f_b(x_i)}\right] \pi_1(s \mid b)\, ds\,.$$

In the absence of (or desire not to utilize) subjective priors, recommended objective choices are the intrinsic priors $\pi_0^I(b) = b^{-1/2}$ and $\pi_1^I(s \mid b) = b(s+b)^{-2}$. The latter is (necessarily) proper, and is justified in the same fashion as the choice in the simpler problem given in Section 2. Since b occurs in both models, it is allowable (and desirable) to utilize the standard objective prior for a Poisson mean, which is $b^{-1/2}$; see Section 5 for discussion.

Note that ignoring the densities f_s and f_b and basing the answer solely on N (as in Section 2) is equivalent to assuming that $f_s \equiv f_b$. It is thus not the case that the simpler analysis simply utilizes less information; it could actually be misleading.

4 Controlling for look-elsewhere effects

A major strength of Bayesian analysis is that it easily (and often automatically) adjusts for look-elsewhere effects. This is illustrated using (3), for a situation of multiple hypothesis testing. The interesting issues involving multiple cuts are then briefly discussed. Finally, the contrasting difficulty of frequentist adjustment for look-elsewhere effects is illustrated.

4.1 Multiple Hypotheses

Suppose N_j of the X_i are in bin B_j, $j = 1, \ldots, m$, and that we assume we have only densities $f_s(B_j)$ and $f_b(B_j)$. Then

$$B_{10} = \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^m \left[1 + \frac{s f_s(B_j)}{b f_b(B_j)}\right]^{N_j} \pi_0(b)\pi_1(s \mid b)\, ds db}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db}\,.$$

Suppose, in addition, that $f_s(B_j)$ gives probability one to some unknown bin B, with each bin being equally likely. This is equivalent to saying that we are testing the mutually exclusive hypotheses:

H_j : *signal is only in bin* B_j, with the hypotheses having equal prior probability. Then

$$\begin{aligned} B_{10} &= \frac{E^B\left[\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^m \left[1+\frac{sf_s(B)}{bf_b(B_j)}\right]^{N_j} \pi_0(b)\pi_1(s \mid b)\, dsdb\right]}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db} \\ &= \frac{\frac{1}{m}\sum_{j=1}^m \int_0^\infty \int_0^\infty b^N e^{-(b+s)} \left[1+\frac{s}{bf_b(B_j)}\right]^{N_j} \pi_0(b)\pi_1(s \mid b)\, dsdb}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db}\,. \end{aligned}$$

The point here is that the marginal likelihood of hypothesis H_i gets automatically down-weighted by $1/m$, the 'cost' of looking in m different bins. There is no need to make any adjustment for this 'looking elsewhere;' it happens automatically as part of the Bayesian analysis.

4.2 Multiple Cuts

The situation involving multiple cuts is interesting, in that Bayesian analysis does not readily apply. A cut really just produces a subset of the overall data, and there is no natural Bayesian way to separately analyze different subsets of data and combine the analyses for an overall conclusion.

If each of m cuts produces data $X_j, j = 1, \ldots, m$, one could legitimately consider the joint distribution of $(X_1, \ldots, X_m)$, and perform a Bayesian analysis. But this is typically not possible because of the difficulty of determining the dependence between the X_j. Of course, if the cuts were such that the X_j could be considered independent, combined analysis is possible: simply multiply the individual cut likelihoods and proceed.

It seems that use of cuts is a practical necessity, but it is worth noting that they are not inherently needed in Bayesian analysis. Suppose, for instance, that $f_s(x) = 0$ for $x \in \Omega^c$, so that it seems tempting to cut on Ω as this is the only region that can possibly contain the signal. But is this necessary? Note that, in this situation, (3) becomes (with the last expression below following from algebra, not probability logic)

$$\begin{aligned} B_{10} &= \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{\{i:x_i\in\Omega\}} \left[1+\frac{sf_s(x_i)}{bf_b(x_i)}\right] \pi_0(b)\pi_1(s \mid b)\, dsdb}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db} \\ &= \int_0^\infty B_{10}(x_\Omega \mid b)\pi_0(b \mid x_\Omega, x_{\Omega^c}) db\,, \end{aligned}$$

where x_Ω (x_{Ω^c}) is the data in Ω (in Ω^c). Thus the Bayesian analysis indeed uses the cut data to find the Bayes factor given b, as would be expected, but it uses all the data to learn more about b, which is clearly desirable (unless, of course, b were different over Ω and Ω^c).

4.3 The difficulty in frequentist control of the look-elsewhere effect

To indicate the difference between the Bayesian and frequentist approaches to controlling look-elsewhere effects, consider the simple multiple testing scenario of testing $H_{0i} : \mu_i = 0$ versus $H_{1i} : \mu_i > 0$, $i = 1, ..., m$, based on data $X_i, i = 1, \ldots, m$, that are normally distributed with mean μ_i, variance 1, and correlation ρ. Furthermore, suppose we know that there is at most one signal.

If $\rho = 0$, one can just do the individual tests at level α/m (Bonferroni) to obtain an overall error probability of α. If $\rho > 0$, however, the situation is more difficult. One natural way to proceed would be to choose the overall decision rule "declare μ_i to be a signal if X_i is the largest value and $X_i > K$," and then compute the corresponding frequentist type I error probability

$$\alpha = \Pr(\max_i X_i > K \mid \mu_1 = \ldots = \mu_m = 0) = E^Z\left[1 - \Phi\left(\frac{K - \sqrt{\rho}Z}{\sqrt{1-\rho}}\right)^m\right],$$

where Φ is the standard normal cdf and Z is a standard normal random variable.

This gives (essentially) the Bonferroni correction when $\rho = 0$, but can be shown to converge to $1 - \Phi[K]$ as $\rho \to 1$, which is the type I error that would result from a single test. Thus the needed frequentist control for multiple testing ranges from the drastic Bonferroni correction to none, depending on the correlations among the data.

In contrast, the Bayesian adjustment for multiple testing does not depend on correlations among the data, and occurs only through the choice of prior probabilities of hypotheses [4]. In the above scenario, for instance, one might assign prior probability 1/2 to no signal, and assign each of the possible alternative hypotheses prior probability of $1/(2m)$ (recall that we are assuming that at most one alternative is true.). The ensuing Bayesian analysis correctly controls for the look-elsewhere effect, regardless of the data distribution.

5 Remarks on choice of priors

A primary difficulty in the Bayesian approach to discovery is the difficulty in choosing prior distributions. Of course, if subjective (or evidence-based) priors are available – and if use of such priors is viewed as appropriate – there is no problem. Typically, however, such priors are not available for, at least, many parameters in the likelihood, and default choices are needed. A brief overview of the situation is given here; more extensive discussions and references can be found in [5, 6].

5.1 Testing versus estimation priors

Unfortunately, the situation in testing (discovery) is quite different from the situation in estimation (e.g., setting confidence limits). For the latter problem, excellent objective priors are available, such as reference priors [7, 8]. These priors are typically improper, which is not a problem for estimation but is often a problem for testing. In (1) for instance, suppose we were to consider the improper prior $\pi(s) = c/\sqrt{s}$, where c is a constant; this is the standard reference prior for estimation. Note that there is no natural choice of c, since the prior is improper and, since the choice of c is irrelevant to estimation, one typically sees the choice $c = 1$. In (1), however, use of $\pi(s) = c/\sqrt{s}$ yields

$$B_{01}(x) = \frac{b^x \, e^{-b}}{c \int_0^\infty (s+b)^x \, e^{-(s+b)} s^{-1/2} \, ds} \,,$$

which is arbitrary since c is arbitrary. In general, parameters that occur in one hypothesis – but not the other – require proper priors.

5.2 Choosing priors for "common parameters"

Often parameters occur in both the likelihoods in the numerator and denominator of a Bayes factors, as does b in (3). Then it is possible (and usually desirable) to use the estimation default priors, e.g. $c/\sqrt{b}$, since c will now cancel in the numerator and denominator.

The difficulty here is in ensuring that the parameters are actually the same in both likelihoods. When the parameters have physical meaning, such as b, this is not an issue. Also, when parameters have what is called the same group invariance structure, one can use the right-Haar priors for those parameters [9].

In other contexts, however, one has to be careful. In variable selection in regression for instance, the meaning of regression coefficients is highly affected by which other coefficients are in the model, and use of a common prior may then be inappropriate. This is usually dealt with by orthogonalizing the parameters (making a transformation so that the partial Fisher information is zero) – see [10].

5.3 Basics of choosing priors for non-common parameters

This is the difficult situation, and there are no ready answers. We have already seen that improper priors cannot be used. Even worse is the (all-too-common) use of vague proper priors.

Example: Suppose X has the normal density $\phi(x \mid \mu, 1)$ with mean μ and variance 1. It is desired to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ with a Uniform$(-c, c)$ prior for μ where c is large, so that one has a 'vague proper prior.' The Bayes factor is

$$B_{01}(c) = \frac{\phi(x \mid 0, 1)}{\int_{-c}^{c} \phi(x \mid \mu, 1)(2c)^{-1} d\mu} \approx \frac{2c\, \phi(x \mid 0, 1)}{\int_{-\infty}^{\infty} \phi(x \mid \mu, 1) d\mu}$$

for large c, which depends dramatically on the choice of c. One might try specifying a plausible range $c_1 \leq c \leq c_2$ and look at $B_{01}(c)$ over this range, hoping the the conclusion is robust, but this will typically not be the case.

It is interesting to obtain some feel for the sensitivity of Bayes factors to the choice of prior. In this example, consider three prior distributions:

- $\pi^u(\mu)$ is Uniform$(0, 10)$, corresponding, say, to a known upper limit on μ.
- $\pi^e(\mu)$ is Normal$(4, 1)$, an evidence-based prior arising from a previous experiment.
- $\pi^t(\mu)$ is a point mass at 4, the prediction of a new theory.

Table 1 gives the posterior probabilities of the null hypothesis that result from these priors for various values of x (equivalent to the number of σ from zero), assuming that the prior probability of the null hypothesis is 1/2. Indeed the posterior probabilities are quite sensitive to the choice of prior. The view of Bayesians, however, is that this sensitivity is an unavoidable fact of life; the three priors correspond to quite different types of prior knowledge and that this knowledge can have a pronounced effect should not be surprising. The corresponding p-values are also given in Table 1, to emphasize the fact that, while the Bayesian answers are sensitive to the prior, they have much more in common with each other than with the p-value.

Table 1: Posterior probabilities of H_0, given various data x from a $N(\mu, 1)$ distribution, assuming prior probability of 1/2 for H_0 and use of the uniform, normal and point mass priors for μ, as well as the corresponding p-values.

	$\pi^u(\mu)$	$\pi^e(\mu)$	$\pi^t(\mu)$	p-value
$x = 2$	0.54	0.34	0.50	$p = 0.025$
$x = 4$	1.3×10^{-3}	4.7×10^{-4}	3.4×10^{-4}	$p = 3.1 \times 10^{-5}$
$x = 6$	6.0×10^{-8}	5.8×10^{-8}	1.1×10^{-7}	$p = 1.0 \times 10^{-9}$

5.4 Various proposed default priors for non-common parameters

There is a long history concerning suggestions for priors for non-common parameters in hypothesis testing and model selection. We give a brief description of the most commonly used methods here. Much more extensive discussions of this history can be found in [5, 6]. A focus of much of this work is to ensure that priors are appropriately balanced between the hypotheses, often called *predictive matching*.

5.4.1 Fractional Bayes factors

Fractional priors [11] use a fraction γ of the likelihood $L(s)$ as the prior, i.e., $\pi(s) = L(s)^\gamma / \int L(s)^\gamma ds$, with the remaining part of the likelihood, $L(s)^{1-\gamma}$, being treated as the likelihood for the Bayes factor computation. For testing between two models, with likelihoods $L_1(s_1)$ and $L_0(s_0)$, this leads to the Bayes factor

$$B_{10} = \frac{\int L_1(s_1)^{1-\gamma_1} \pi_1(s_1) ds_1}{\int L_0(s_0)^{1-\gamma_0} \pi_0(s_0) ds_0} = \frac{\int L_1(s_1) ds_1 \int L_0(s_0)^{\gamma_0} ds_0}{\int L_0(s_0) ds_0 \int L_1(s_1)^{\gamma_1} ds_1}\,.$$

This is often computationally attractive, and usually does correspond (at least asymptotically) to a real Bayesian analysis if the γ_i are chosen appropriately; the typically recommended choice is $\gamma_i = n/p_i$, where n is the sample size of the data, and p_i is the dimension of the parameter s_i. For discussion of the strengths and weaknesses of this approach, see [5].

5.4.2 Intrinsic priors

Intrinsic priors (see [5] for discussion and earlier references) are generated in a bootstrap fashion using either subsets of the data or artificial data. They are very widely applicable and have excellent Bayesian properties, but can be computationally intensive.

One of the methods of deriving an intrinsic prior is through the *expected posterior* prior construction [12]. For the situation where the data consists of i.i.d. observations from a density $f(x \mid s)$, and for testing $H_0 : s = s_0$ versus $H_1 : s \neq s_0$, the construction is as follows:

- let $\pi^O(s)$ be a good estimation objective prior, so that $\pi^O(s \mid \boldsymbol{x}) = [\prod_{i=1}^n f(x_i \mid s)]\pi^O(s)/m^O(\boldsymbol{x})$ is the resulting posterior, where $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $m^O(\boldsymbol{x}) = \int [\prod_{i=1}^n f(x_i \mid s)]\pi^O(s)\, ds$;
- then the intrinsic prior is $\pi^I(s) = \int \pi^O(s \mid \boldsymbol{x}^*)[\prod_{i=1}^q f(x_i \mid s_0)]\, d\boldsymbol{x}^*$, with $\boldsymbol{x}^* = (x_1, \ldots, x_q)$ being (unobserved) data of the minimal sample size q such that $m^O(\boldsymbol{x}^*) < \infty$.

Note that this will be a proper (not vague proper) prior.

The idea behind this prior is that, if one were handed the data $\boldsymbol{x}^*$ but allowed to use it only for prior construction, one would happily compute $\pi^O(s \mid \boldsymbol{x}^*)$ and use this proper prior to conduct the test. We don't have $\boldsymbol{x}^*$ available, but we could simulate $\boldsymbol{x}^*$ from the null model, and compute the resulting 'average' prior. This is the method used to derive the intrinsic prior $\pi_1^I(s \mid b)$ in Section 1.

5.4.3 Conventional priors

For common specific situations, proper conventional priors have been proposed. For instance, in testing involving the normal linear model, numerous proper default priors have been proposed that depend only on the design matrix of the model being considered. The most popular of these conventional priors are the Zellner-Siow priors [13], which were developed following ideas of [10]. These priors have some very nice properties. In particular, they result in answers that

- are invariant to scale changes in covariates (i.e., the units of measurement used);
- are consistent (i.e., the true model will be selected as $n \to \infty$), if the true model is among those being considered;
- are information-consistent (i.e., will reject the null model as the associated t or F statistics $\to \infty$).

For description and further discussion, see [5].

5.4.4 Approximations

Because of the difficulty in choosing prior distributions and in computing Bayes factors, use of approximations such as BIC (Bayes information criterion [14]) is often considered; BIC neither requires specification of priors nor integration over the likelihood. The accuracy of the approximation is, however, mixed, at best. The approximation does capture part of the influence of prior distribution in a generic way, thus allowing for an Ockham's Razor effect of preferring the more parsimonious of two models that equally well explain the data. But BIC basically ignores constants in the Bayes factor that arise from the priors, and these constants can be arbitrarily large or small, so the approach is by no means a cure-all.

6 A comment on unfolding

The workshop also had a focus on unfolding (deconvolution), which has historically also been a central problem in Bayesian statistics. If one has a density $f(x \mid s)$ of data x, given an unknown parameter s, and a prior distribution $\pi(s)$ for s, the predictive (or marginal) distribution of x is then

$$m_\pi(x) = \int f(x \mid s)\pi(s)\, ds\,.$$

Often one is in a situation of having an estimate $\hat{m}(x)$ for the predictive distribution, and the goal is then to find $\pi(s)$ such that $m_\pi(x)$ is as close as possible to $\hat{m}(x)$, the unfolding problem.

In [15], a very interesting algorithm for attacking the problem is presented. Start with any initial $\pi_0(s)$ that has support everywhere. Then iteratively compute

$$\pi_l(s) = \int \pi_{l-1}(s \mid x)\hat{m}(x)dx\,, \quad \text{where } \pi_{l-1}(s \mid x) = \frac{\pi_{l-1}(s)f(x \mid s)}{\int \pi_{l-1}(s)f(x \mid s)ds}\,.$$

Theorem: $\pi^*(s) = \lim_{l\to\infty} \pi_l(s)$ is the density for which $m_\pi(x)$ is as close as possible to $\hat{m}(x)$ in Kullback-Leibler divergence.

Here is a potentially interesting implementation of this algorithm using particle filtering:

- Represent $\pi_l(s)$ by a collection of *particles* $\{s_i\}$ with weights $\{w_i^{(l)}\}$. (Initialize with a random sample $\{s_i\}$ from $\pi_0(s)$, so the initial weights are equal.)
- Then $\pi_l(s \mid x)$ would be the same collection of particles but with modified weights

$$w_i^{(l)}(x) = \frac{w_i^{(l-1)} f(x \mid s_i)}{\sum_j w_j^{(l-1)} f(x \mid s_j)}\,,$$

  and $\pi_{l+1}(s)$ would be the same collection of particles but with weights

$$w_i^{(l+1)} = \int w_i^{(l)}(x)\hat{m}(x)dx\,.$$

- As one progresses one will need to add new particles adapting to the evolving density, but there are likely techniques in particle filtering for doing this.

Acknowledgements

This work was supported by NSF Grants AST-0507481, DMS-0757549-001, and DMS-1007773.

References

[1] Berger, J. (2008). A comparison of testing methodologies. In *Proceedings of the PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics,* **CERN 2008-001,** 8–19.

[2] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

[3] Lockhart, R. (2010). Banff Challenge 2 – statistician's language. Available at http://www.birs.ca/workshops/2010/10w5068/10w5068BanffChallenge2.pdf.

[4] Scott, J. and Berger, J. (2010). Bayes and Empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics* **38**, 2587–2619.

[5] Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison (with Discussion). In *Model Selection*, P. Lahiri, ed., Institute of Mathematical Statistics Lecture Notes – Monograph Series, volume 38, Beachwood Ohio, 135–207.

[6] Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.*, **103**, 410–423.

[7] Bernardo, J. M. (2005). Reference analysis. Handbook of Statistics 25 (D. K. Dey and C. R. Rao eds.). Amsterdam: Elsevier, 17–90.

[8] Demortier, L. (2005). Bayesian reference analysis for particle physics. *PHYSTAT05 Proceedings on "Statistical Problems in Particle Physics, Astrophysics and Cosmolgy"*. Imperial College Press.

[9] Berger, J., Pericchi, L, and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā*, **A 60**, 307–321.

[10] Jeffreys, H. (1961). *Theory of Probability*, London: Oxford University Press.

[11] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *Journal of the Royal Statistical Society*, Ser. B, **57**, 99–138.

[12] Pérez, J.M. and Berger, J. (2002). Expected posterior prior distributions for model selection. *Biometrika*, **89**, 491–512.

[13] Zellner, A. and Siow, A. (1980). Posterior Odds Ratios for Selected Regression Hypotheses. In Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain) (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. University of Valencia.

[14] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

[15] Shyamalkumar, N.D. (1996). Cyclic I_0 projections and its applications to statistics. Technical Report 96-24, Department of Statistics, Purdue University.