Bayesian Statistics from Subjective Quantum Probabilities to Objective Data Analysis

Luc Demortier
The Rockefeller University

Informal High Energy Physics Seminar

Caltech, April 29, 2008

Bayes versus Frequentism

. . . A long-standing philosophical dispute with important practical implications

- For just about every form of scientific experimentation;
- For the problem of quantum state determination.

Frequentism (1)

Frequentism defines probabilities as relative frequencies in sequences of trials. These are real, objective quantities that can be measured and exist "outside us".

How rigorous is this definition?

- Could we say that probability is a limiting relative frequency in an infinite sequence of trials?
  - Unfortunately infinite sequences are unobservable...
  - Perhaps we could define probability as the limiting relative frequency which *would* be obtained if the sequence of trials *were* extended to infinity.
  - But this is counterfactual: contradicts probability as measurable...
  - Another problem with the infinite sequence definition is that it may not be empirically relevant to finite beings living in a finite region of space-time.

Frequentism (2)

- Can we then define probability in terms of *finite* sequences?
  - This weakens the definition since, in a finite number of trials, *every* sequence has a non-zero probability of occurring.
  - Perhaps we could say that the probability of an event is, For All Practical Purposes, the relative frequency of that event in a sufficiently long sequence of trials. Highly improbable sequences are FAPP impossible.
  - However, sufficiently long sequences are always highly improbable. Where should we draw the line? What matters of course, is which hypothesis about the event probability makes the observed sequence *least* improbable. Absolute probabilities have no meaning.
  - Furthermore, to make inferences from an observed sequence, we must assume that the trials are independent and the event probability is the same in every trial.
  - To say that the probability of an event is the same in every trial requires that probability to be *defined* for every trial.
  - Therefore, we need the concept of a single-case probability.

Bayesianism (1)

Bayesianism makes a strict distinction between propositions and probabilities:

- Propositions are either true or false; their truth value is a fact.
- Probabilities are degrees of belief about the truth of some proposition; they are neither true nor false; they are not propositions.

Bayesian probability:

- is a logical construct rather than a physical reality;
- applies to individual events rather than to ensembles;
- is a statement *not* about what is in fact the case, but about what one can reasonably expect to be the case;
- is epistemic, normative, subjective.

Bayesianism (2)

The theory of probability is a form of extended logic (Jaynes): a process of reasoning by which one extracts uncertain conclusions from limited information.

This process is guided by Bayes' theorem:

$$\pi(\theta \,|\, x) \;=\; \frac{p(x \,|\, \theta)\; \pi(\theta)}{m(x)},$$

where

$$m(x) \;\equiv\; \int_{\Theta} p(x \,|\, \theta)\; \pi(\theta)\; d\theta.$$

Another important tool is marginalization:

$$\pi(\theta \,|\, x) \;=\; \int_{\Lambda} \pi(\theta, \lambda \,|\, x)\; d\lambda.$$

All the basic tools of Bayesian statistics are direct applications of probability theory.

Subjective Quantum Probabilities

Recent research in quantum information science focuses on the question of whether quantum probabilities are objective (frequentist) or subjective (Bayesian).

Part of the motivation for this comes from EPR-style arguments: suppose two systems A and B are prepared in some entangled quantum state and then spatially separated. By measuring one of two observables on A alone, one can immediately write down a new state for B. If one accepts that the "real, objective state of affairs" at B cannot depend on measurements made at A, then the simplest interpretation of the new state for B is that it is a *state of knowledge*.

It is possible to develop this idea of quantum states as states of knowledge in a fully consistent way. There are many aspects to this:

- Subjective probability assignments must follow the standard quantum rule for probabilities (Gleason's theorem).
- The connection between quantum probability and long-term frequency still holds, but is a non-trivial consequence of Gleason's theorem and the concept of maximal information in quantum theory.
- Even quantum certainty (probability-1 predictions for pure states) is always some agent's certainty. Any agent-independent certainty about a measurement outcome would correspond to a pre-existing system property and would be in conflict with locality.

Subjective Quantum Probabilities: Does it Matter?

Aside from providing yet another interpretation of quantum mechanics, do Bayesian quantum probabilities have any practical consequence?

Yes! For example, if vacuum fluctuations are not real events, then we do not need to worry about their effect on the cosmological constant.

E.T. Jaynes showed that spontaneous emission and the Lamb shift can be derived without the need for vacuum fluctuations. He noted that for every differential equation with a non-negative Green's function there is a stochastic problem with the same solution, even though the two problems are physically unrelated.

He also argued (without calculation) that the Casimir effect does not require zero-point energy to reside throughout all space.

On to Data Analysis

In physics data analysis we often need to extract information about a parameter θ about which very little is known a priori. Or perhaps we would like to *pretend* that very little is known for reasons of objectivity. How do we apply Bayes' theorem in this case: how do we construct the prior $\pi(\theta)$?

Although quantum probabilities are constrained by Gleason's theorem, there is no such universal rule available to constrain inferences in data analysis.

Historically, this is the main reason for the development of alternative statistical paradigms: frequentism, likelihood, fiducial probability, objective Bayes, etc. In general, results from these different methods agree on large data samples, but not necessarily on small samples (discovery situations).

For this reason, the CMS Statistics Committee at the LHC recommends data analysts to cross-check their results using three different methods: objective Bayes, frequentism, and likelihood.

Objective Bayesianism

At its most optimistic, objective Bayesianism tries to find a completely coherent objective Bayesian methodology for learning from data. A much more modest view is that it is simply a collection of ad hoc but useful methods to learn from the data. There are in fact several approaches, all of which attempt to construct prior distributions that are minimally informative in some sense:

- Reference analysis (Bernardo and Berger);
- Maximum entropy priors (Jaynes);
- Invariance priors;
- Matching priors;
- . . .

Flat priors tend to be popular in HEP, even though they are hard to justify since they are not invariant under parameter transformations. Furthermore, they sometimes lead to improper posterior distributions and other kinds of misbehavior.

Reference Analysis (1)

Reference analysis is a method to produce inferences that only depend on the model assumed and the data observed. It is meant to provide standards for scientific communication.

In order to be generally and consistently applicable, reference analysis uses the Bayesian paradigm, which immediately raises the question of priors: what kind of prior will produce "objective" inferences?

The primary aim is to obtain posterior distributions that are dominated in some sense by the information contained in the data, but there are additional requirements that may reasonably be considered as necessary properties of any proposed solution:

- *Generality:*
  The procedure should be completely general and should always yield *proper* posteriors.
- *Invariance:*
  If $\phi = \phi(\theta)$, then $\pi(\phi \,|\, x) = \pi(\theta \,|\, x)\, |d\theta/d\phi|$. Furthermore, if $t = t(x)$ is a sufficient statistic, then $\pi(\theta \,|\, x) = \pi(\theta \,|\, t)$.

Reference Analysis (2)

- *Consistent Marginalization:*
  Suppose $p(x \,|\, \theta, \lambda) \rightarrow \pi(\theta, \lambda \,|\, x)$, and $\pi_1(\theta \,|\, x) \equiv \int \pi(\theta, \lambda \,|\, x)\, d\lambda = \pi_1(\theta \,|\, t)$, where $t = t(x)$.
  Suppose also that $p(t \,|\, \theta, \lambda) = p(t \,|\, \theta) \rightarrow \pi_2(\theta \,|\, t)$.
  Then, consistent marginalization requires that $\pi_2(\theta \,|\, t) = \pi_1(\theta \,|\, t)$.

- *Consistent sampling properties:*
  The family of posterior distributions $\pi(\theta \,|\, x)$ obtained by repeated sampling from the model $p(x \,|\, \theta, \lambda)$ should concentrate on a region of Θ which contains the true value of θ.

Reference analysis aims to replace the question "What is our prior degree of belief?" with "What would our posterior degree of belief be, if our prior knowledge had a minimal effect, relative to the data, on the final inference?"

Intrinsic Discrepancy (1)

Reference analysis techniques are based on information theory, and in particular on the central concept of intrinsic discrepancy between probability densities:

The intrinsic discrepancy between two probability densities p_1 and p_2 is:

$$\delta\{p_1, p_2\} \;=\; \min\left\{ \int\! dx\; p_1(x)\; \ln\frac{p_1(x)}{p_2(x)},\;\; \int\! dx\; p_2(x)\; \ln\frac{p_2(x)}{p_1(x)} \right\},$$

provided one of the integrals is finite. The intrinsic discrepancy between two parametric models for x,

$$\mathcal{M}_1 = \{p_1(x\,|\,\phi), x \in \mathcal{X}, \phi \in \Phi\} \text{ and } \mathcal{M}_2 = \{p_2(x\,|\,\psi), x \in \mathcal{X}, \psi \in \Psi\},$$

is the minimum intrinsic discrepancy between their elements:

$$\delta\{\mathcal{M}_1, \mathcal{M}_2\} \;=\; \inf_{\phi,\psi}\; \delta\{p_1(x\,|\,\phi), p_2(x\,|\,\psi)\}.$$

Intrinsic Discrepancy (2)

Properties of the intrinsic discrepancy:

- $\delta\{p_1, p_2\}$ is symmetric, non-negative, and vanishes if and only if $p_1(x) = p_2(x)$ almost everywhere.
- $\delta\{p_1, p_2\}$ is invariant under one-to-one transformations of x.
- $\delta\{p_1, p_2\}$ is information-additive: the discrepancy for a set of n independent observations is n times the discrepancy for one observation.
- The intrinsic discrepancy $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ between two parametric families of distributions does not depend on their parametrizations.
- The intrinsic discrepancy $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ is the minimum expected log-likelihood ratio in favor of the model which generates the data.
- The intrinsic discrepancy $\delta\{p_1, p_2\}$ is a measure, in natural information units, of the minimum amount of expected information required to discriminate between p_1 and p_2.

Missing Information

The expected intrinsic information $I\{p(\theta)\,|\,\mathcal{M}\}$ from one observation of

$$\mathcal{M} \equiv \{p(x\,|\,\theta),\ x \in \mathcal{X},\ \theta \in \Theta\}$$

about the value of θ when the prior density is $p(\theta)$, is:

$$I\{p(\theta)\,|\,\mathcal{M}\} \;=\; \delta\{p(x,\theta),\ p(x)\,p(\theta)\},$$

where $p(x,\theta) = p(x\,|\,\theta)\,p(\theta)$ and $p(x) = \int d\theta\; p(x\,|\,\theta)\,p(\theta)$.

The stronger the prior knowledge described by $p(\theta)$, the smaller the information the data may be expected to provide. Conversely, weak initial knowledge about θ corresponds to large expected information from the data.

Consider the intrinsic information about θ, $I\{p(\theta)\,|\,\mathcal{M}^k\}$, which could be expected from making k independent observations from $\mathcal{M}$. As k increases, the true value of θ would become precisely known. Thus, as $k \to \infty$, $I\{p(\theta)\,|\,\mathcal{M}^k\}$ measures the amount of *missing information* about θ which corresponds to the prior $p(\theta)$. For large k one can show that

$$I\{p(\theta)\,|\,\mathcal{M}^k\} \;=\; \mathrm{E}_x\left[\int d\theta\; p(\theta\,|\,x)\;\ln\frac{p(\theta\,|\,x)}{p(\theta)}\right]$$

Reference Priors for One-Parameter Models

Let $\mathcal{P}$ be a class of sufficiently regular priors that are compatible with whatever "objective" initial information one has about the value of θ.

The reference prior is then defined to be that prior function $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ which maximizes the missing information about the value of θ within the class $\mathcal{P}$ of candidate priors.

If the parameter space is finite and discrete, $\Theta = \{\theta_1, \ldots, \theta_m\}$, the missing information is simply the entropy of the prior distribution, $-\sum_{i=1}^{m} p(\theta_i) \ln p(\theta_i)$, and one recovers the prior proposed by Jaynes for this case.

In the continuous case however, $I\{p(\theta) \,|\, \mathcal{M}^k\}$ diverges as $k \to \infty$, and reference priors must be defined with a special limiting procedure:

$\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ is a reference prior for model $\mathcal{M}$ given $\mathcal{P}$ if, for some increasing sequence $\{\Theta_i\}_{i=1}^{\infty}$ with $\lim_{i \to \infty} \Theta_i = \Theta$ and $\int_{\Theta_i} \pi(\theta)\, d\theta < \infty$,

$$\lim_{k \to \infty} \left[I\{\pi_i \,|\, \mathcal{M}^k\} - I\{p_i \,|\, \mathcal{M}^k\} \right] \geq 0 \qquad \text{for all } \Theta_i, \text{ for all } p \in \mathcal{P},$$

where $\pi_i(\theta)$ and $p_i(\theta)$ are the renormalized restrictions of $\pi(\theta)$ and $p(\theta)$ to Θ_i.

Some Properties of Reference Priors

- In the definition, the limit $k \to \infty$ is *not* an approximation, but an essential part of the definition, since the reference prior maximizes the *missing* information, which is the expected discrepancy between prior knowledge and *perfect* knowledge.
- Reference priors only depend on the asymptotic behavior of the model, which greatly simplifies their derivation. For example, in one-parameter models and under appropriate regularity conditions to guarantee asymptotic normality, the reference prior is simply Jeffreys' prior:

$$\pi(\theta) \propto i(\theta)^{1/2}, \qquad \text{where} \quad i(\theta) = -\int_{\mathcal{X}} dx\, p(x \,|\, \theta)\, \frac{\partial^2}{\partial \theta^2} \ln p(x \,|\, \theta).$$

- Reference priors are independent of sample size.
- Reference priors are compatible with sufficient statistics and consistent under reparametrization, due to the fact that the expected information is invariant under such transformations.
- Reference priors do not represent subjective belief and should not be interpreted as prior probability distributions. In fact, they are often improper. Only reference *posteriors* have a probability interpretation.

Reference Priors in the Presence of Nuisance Parameters

Suppose the statistical model is $p(x \mid \theta, \lambda)$, where θ is the parameter of interest and λ is a nuisance parameter. We now need a joint reference prior $\pi(\theta, \lambda)$. The algorithm is sequential and based on the decomposition $\pi(\theta, \lambda) \;=\; \pi(\lambda \mid \theta)\, \pi(\theta)$:

1. Apply the one-parameter reference algorithm to obtain the conditional reference prior $\pi(\lambda \mid \theta)$.

2. Derive the one-parameter integrated model:

$$p(x \mid \theta) \;=\; \int_{\Lambda} d\lambda\, p(x \mid \theta, \lambda)\, \pi(\lambda \mid \theta)$$

3. Apply the one-parameter reference algorithm again, this time to $p(x \mid \theta)$, and obtain the marginal reference prior $\pi(\theta)$.

Note that step 2 will not work if $\pi(\lambda \mid \theta)$ is improper ($p(x \mid \theta)$ will not be normalizable). The solution in that case is to introduce a sequence $\{\Lambda_i\}_{i=1}^{\infty}$ of subsets of Λ that converges to Λ and such that $\pi(\lambda \mid \theta)$ is integrable over each Λ_i. The integration at step 2 is then performed over Λ_i instead of Λ. This procedure results in a sequence of posteriors $\{\pi_i(\theta \mid x)\}_{i=1}^{\infty}$ and the desired reference posterior is obtained as the limit of that sequence.

Reference Priors: Comments

- Generalization of the reference algorithm from two to any number of parameters is straightforward.

- Since the algorithm is sequential, it requires that the parameters be ordered, say in order of inferential interest. In most applications it is found that the order does not affect the result, but there are exceptions.

- A direct consequence of this sequential algorithm is that, within a *single* model, it is possible to have as many reference priors as there are possible parameters of interest. This is because a setup that maximizes the missing information about a parameter θ will generally differ from a setup that maximizes the missing information about a parameter η, unless η is a one-to-one function of θ.

- The good news is that using different non-subjective priors for different parameters of interest is the *only* way to avoid the marginalization paradoxes.

Restricted Reference Priors

The definition of reference priors specifies that they must be taken from a class $\mathcal{P}$ of priors that are compatible with whatever initial information is available. If there is no initial information, the class is labeled $\mathcal{P}_0$ and the prior is unrestricted. Initial information can take several forms:

1. Constraints on parameter space.
2. Specified expected values.
   Suppose that the initial information about θ is of the form $\mathsf{E}[g_i(\theta)] = \beta_i$, for appropriately chosen functions $g_i, i = 1, \ldots, m$. It can then be shown that the reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ must be of the form:
   $$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}) \;=\; \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) \; \exp\left\{\sum_{i=1}^{m} \lambda_i \, g_i(\theta)\right\},$$
   where the λ_i's are constants determined by the constraints which define $\mathcal{P}$.
3. Subjective marginal prior.
   Suppose the model depends on two parameters, θ_1 and θ_2, and the subjective marginal $\pi(\theta_1)$ is known. The reference conditional $\pi(\theta_2 \,|\, \theta_1)$ is then proportional to $|\Sigma_{22}(\theta_1, \theta_2)|^{1/2}$, where $\Sigma_{22}(\theta_1, \theta_2)$ is the per observation Fisher information for θ_2, given that θ_1 is held fixed. If the resulting $\pi(\theta_2 \,|\, \theta_1)$ is improper, it must be corrected via a sequence of compact subsets argument in order to preserve the information based interpretation.

Example: a Poisson Process with Uncertain Mean

Consider the likelihood:

$$\mathcal{L}(\sigma, \epsilon, b \,|\, n) \;=\; \frac{(b + \epsilon\sigma)^n}{n!}\, e^{-b-\epsilon\sigma},$$

where the parameter of interest is σ (say a cross section), whereas ϵ (an effective efficiency) and b (a background) are nuisance parameters.

Note that σ, ϵ, and b are not identifiable. This problem is usually addressed by introducing a subjective prior for ϵ and b, say $\pi(\epsilon, b)$.

A common choice of prior for σ is $\pi(\sigma) = 1$ (improper!), the claim being that this is noninformative. . . Whatever one may think of this claim, if the ϵ prior has non-zero density at $\epsilon = 0$ (such as a truncated Gaussian), the posterior will be improper.

Poisson Process with Uncertain Signal Efficiency

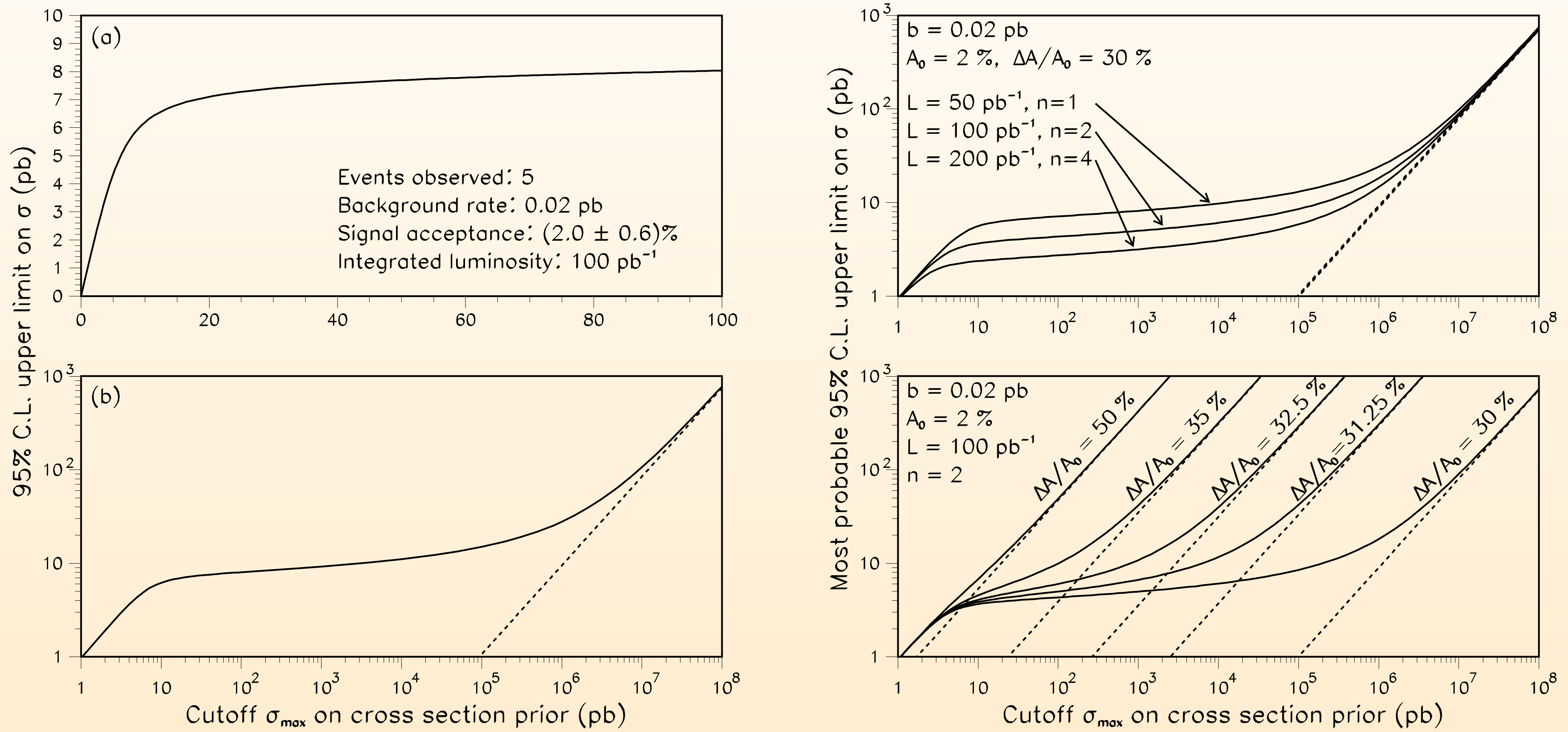

Figure 1: Bayesian upper limits at the 95% credibility level on a signal cross section σ, as a function of the cutoff $\sigma_{\max}$ on the flat prior for σ. The signal efficiency has a truncated Gaussian prior.

Reference Prior for the Poisson Problem

Assume we are given a subjective prior $\pi(\epsilon, b)$. We must therefore find the conditional reference prior $\pi(\sigma \,|\, \epsilon, b)$. As described before, we start by calculating the Fisher information for σ given that ϵ and b are held fixed:

$$\Sigma_{\sigma\sigma} \;=\; \mathrm{E}\left[-\frac{\partial^2}{\partial\sigma^2}\ln\mathcal{L}\right] \;=\; \frac{\epsilon^2}{b+\epsilon\sigma},$$

which would suggest:

$$\pi(\sigma \,|\, \epsilon, b) \;\propto\; \frac{\epsilon}{\sqrt{b+\epsilon\sigma}}.$$

This prior is improper however, requiring that it be renormalized using a sequence of nested compact sets. We take these sets to be of the form $[0, u]$, with $u > 0$. Normalizing the above prior over such a set yields:

$$\pi_u(\sigma \,|\, \epsilon, b) \;=\; \frac{\epsilon}{\sqrt{b+\epsilon\sigma}}\;\frac{\theta(u-\sigma)}{2\sqrt{b+\epsilon u}\;-\;2\sqrt{b}}.$$

The correct conditional reference prior is then given by:

$$\pi(\sigma \,|\, \epsilon, b) \;=\; \lim_{u\to\infty}\frac{\pi_u(\sigma \,|\, \epsilon, b)}{\pi_u(\sigma_0 \,|\, \epsilon_0, b_0)} \;\propto\; \sqrt{\frac{\epsilon}{b+\epsilon\sigma}},$$

with $(\sigma_0, \epsilon_0, b_0)$ any fixed point.

Reference Posterior for the Poisson Problem

To fix ideas, let us consider a product of gamma densities for the subjective prior $\pi(\epsilon, b)$:

$$\pi(\epsilon, b) \;=\; \frac{\tau(\tau\epsilon)^{x-1/2}\, e^{-\tau\epsilon}}{\Gamma(x+1/2)}\, \frac{c(cb)^{y-1/2}\, e^{-cb}}{\Gamma(y+1/2)}.$$

The σ-reference posterior is then:

$$\pi(\sigma\,|\,n) \;\propto\; \int_0^\infty d\epsilon \int_0^\infty db\, \frac{(b+\epsilon\sigma)^{n-1/2}\, e^{-b-\epsilon\sigma}}{n!}\, \frac{\sqrt{\tau}(\tau\epsilon)^{x}\, e^{-\tau\epsilon}}{\Gamma(x+1/2)}\, \frac{c(cb)^{y-1/2}\, e^{-cb}}{\Gamma(y+1/2)}.$$

The integrals may seem daunting, but there is a simple Monte Carlo algorithm to generate σ values from the posterior:

1. Generate $\epsilon \sim \mathrm{Gamma}(x, 1/\tau)$;
2. Generate $b \sim \mathrm{Gamma}(y+1/2, 1/c)$;
3. Generate $t \sim \mathrm{Gamma}(n+1/2, 1)$;
4. If $t < b$, go back to step 2;
5. Set $\sigma = (t-b)/\epsilon$;

where $\mathrm{Gamma}(z\,|\,\alpha, \beta) \equiv z^{\alpha-1}e^{-z/\beta}/\,\Gamma(\alpha)\beta^{\alpha}$.

Repeated Sampling Properties

The Poisson problem just considered involves both subjective and objective priors, which complicates the checking of repeated sampling properties. There are three possible ways to proceed:

1. Full Frequentist Ensemble
   If the nuisance priors are posteriors from actual subsidiary measurements, one can calculate the coverage with respect to an ensemble in which all the parameters are kept fixed, while the observations from both primary and subsidiary measurements are fluctuated. In the Poisson example, the gamma priors can be derived as reference posteriors from Poisson measurements, allowing this type of coverage to be checked.
2. Restricted Frequentist Ensemble
   More often, the nuisance priors incorporate information from simulation studies, theoretical beliefs, etc., precluding a fully frequentist interpretation. The only proper frequentist way to calculate coverage in this case is to keep all the parameters fixed while fluctuating the observation from the primary measurement.
3. Bayesian Averaged Frequentist Ensemble
   Respect the Bayesian interpretation of the subjective priors, and average the coverage over them.

Coverage of Reference Bayes Poisson Upper Limits

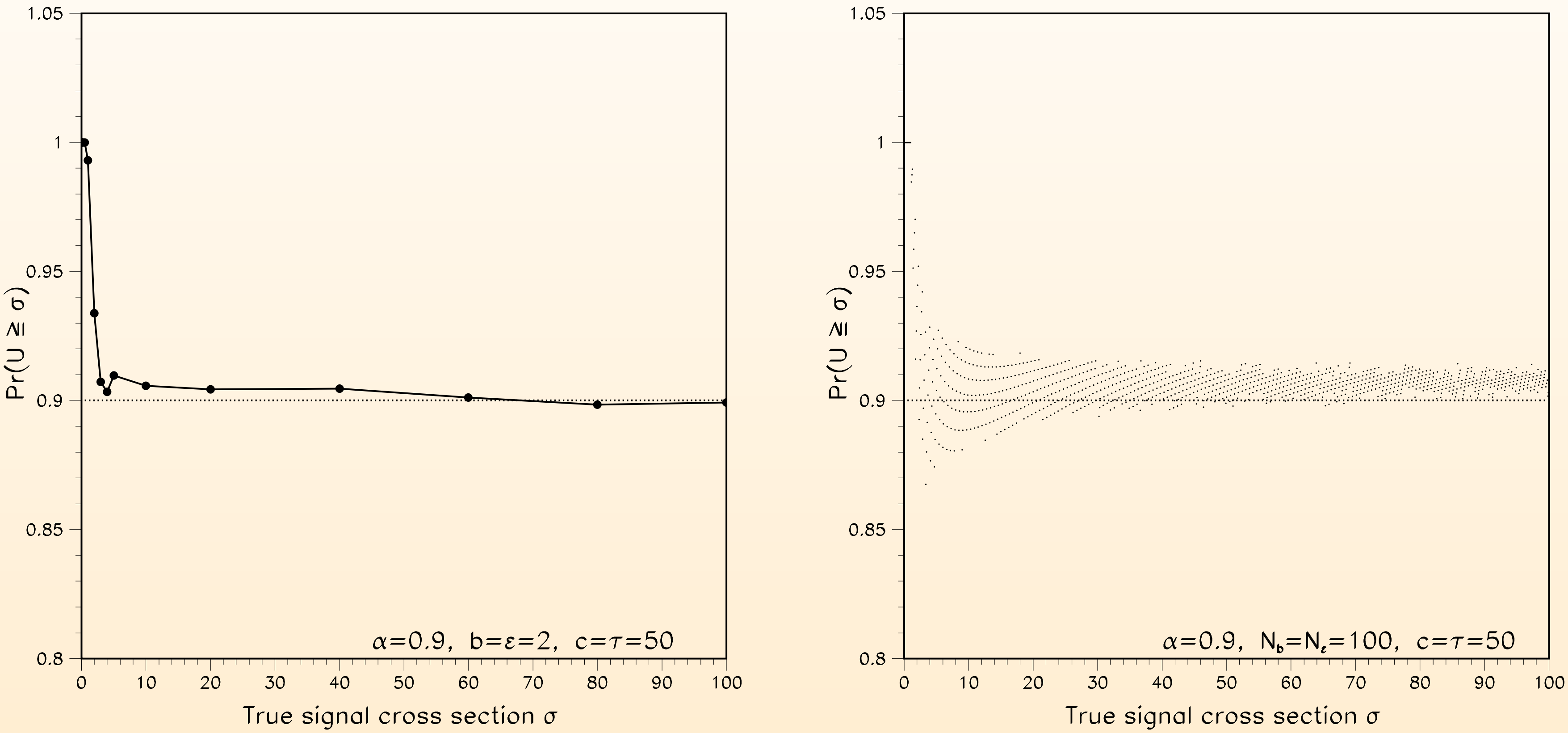

Figure 2: Coverage of 90% credibility level reference Bayes upper limits on a signal cross section σ, as a function of the true value of that cross section. The coverage calculation was done according to a full frequentist ensemble (left) and to a Bayesian averaged frequentist ensemble (right).

Intrinsic Estimation and Intrinsic Credible Regions (1)

It is well known and nevertheless always worth repeating that the Bayesian outcome of a problem of inference is precisely the full posterior distribution for the parameter of interest.

However, it is often useful and sometimes even necessary to *summarize* the posterior distribution by providing a measure of location and quoting regions of given posterior probability content.

The typical Bayesian approach formulates point estimation as a decision problem. Suppose that $\hat{\theta}$ is an estimate of the parameter θ, whose true value θ_t is unknown. One specifies a loss function $\ell(\hat{\theta}, \theta_t)$, which measures the consequence of using the model $p(x \,|\, \hat{\theta})$ instead of the true model $p(x \,|\, \theta_t)$. The Bayes estimator $\theta_b = \theta_b(x)$ of the parameter θ minimizes the corresponding posterior loss:

$$\theta_b(x) \;=\; \arg\min_{\hat{\theta}\in\Theta} \int_{\Theta} d\theta \; \ell(\hat{\theta}, \theta)\; p(\theta \,|\, x).$$

Some conventional loss functions are:

1. <u>*Squared error loss:*</u> $\ell(\hat{\theta}, \theta_t) = (\hat{\theta} - \theta_t)^2 \quad \Rightarrow \quad \theta_b$ is the *posterior mean*.
2. <u>*Zero-one loss:*</u> $\ell(\hat{\theta}, \theta_t) = 1 - \mathrm{I}_{[\theta_t - \epsilon, \theta_t + \epsilon]}(\hat{\theta}) \quad \Rightarrow \quad \theta_b$ is the *posterior mode*.
3. <u>*Absolute error loss:*</u> $\ell(\hat{\theta}, \theta_t) = |\hat{\theta} - \theta_t| \quad \Rightarrow \quad \theta_b$ is the *posterior median*.

Intrinsic Estimation and Intrinsic Credible Regions (2)

In physics, interest usually focuses on the actual mechanism that governs the data. Therefore we need a point estimate that is invariant under one-to-one transformations of the parameter and/or the data (including reduction to sufficient statistics). Fortunately, we have already encountered a loss function that will deliver such an estimate: the intrinsic discrepancy!

The intrinsic discrepancy between two probability densities p_1 and p_2 is:

$$\delta\{p_1, p_2\} \;=\; \min\left\{ \int dx\; p_1(x)\;\ln\frac{p_1(x)}{p_2(x)},\;\; \int dx\; p_2(x)\;\ln\frac{p_2(x)}{p_1(x)} \right\},$$

provided one of the integrals is finite. The intrinsic discrepancy between two parametric models for x,
$\mathcal{M}_1 = \{p_1(x\,|\,\phi), x \in \mathcal{X}, \phi \in \Phi\}$ and $\mathcal{M}_2 = \{p_2(x\,|\,\psi), x \in \mathcal{X}, \psi \in \Psi\}$,
is the minimum intrinsic discrepancy between their elements:

$$\delta\{\mathcal{M}_1, \mathcal{M}_2\} \;=\; \inf_{\phi,\psi}\; \delta\{p_1(x\,|\,\phi), p_2(x\,|\,\psi)\}.$$

This suggests setting $\ell(\hat{\theta}, \theta_t) = \delta\{\hat{\theta}, \theta_t\} \equiv \delta\{p(x\,|\,\hat{\theta}), p(x\,|\,\theta_t)\}$.

Intrinsic Estimation and Intrinsic Credible Regions (3)

Let $\{p(x \,|\, \theta), x \in \mathcal{X}, \theta \in \Theta\}$ be a family of probability models for some observable data x. The intrinsic estimator minimizes the reference posterior expectation of the intrinsic discrepancy:

$$\theta^{\star}(x) \;=\; \arg\,\min_{\hat{\theta}\in\Theta}\; d(\hat{\theta} \,|\, x) \;=\; \arg\,\min_{\hat{\theta}\in\Theta}\; \int_{\Theta} d\theta\;\delta\{\hat{\theta}, \theta\}\;\pi_{\delta}(\theta \,|\, x),$$

where $\pi_{\delta}(\theta \,|\, x)$ is the reference posterior when the intrinsic discrepancy is the parameter of interest.

An intrinsic α-credible region is a subset $R^{\star}_{\alpha}$ of the parameter space Θ such that:

$$(i) \qquad \int_{R^{\star}_{\alpha}} d\theta\;\pi(\theta \,|\, x) \;=\; \alpha;$$

$$(ii) \qquad \text{For all } \theta_i \in R^{\star}_{\alpha} \text{ and } \theta_j \notin R^{\star}_{\alpha}, \quad d(\theta_i \,|\, x) \leq d(\theta_j \,|\, x).$$

Although the concepts of intrinsic estimator and credible region have been defined here for *reference* problems, they can also be used in situations where proper prior information is available.

Example: Transverse Momentum Measurement (1)

Consider the measurement of the transverse momentum of particles in a tracking chamber immersed in a magnetic field. The probability density is (approximately) Gaussian in the inverse of the transverse momentum:

$$p(x \,|\, \mu) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{1/x - 1/\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\,\sigma\,x^2},$$

where x is the measured signed p_T, μ is the true signed p_T, and σ is a function of the magnetic field strength and the chamber resolution.

It is easy to verify that a naive Bayesian analysis yields unreasonable results. To begin with, "non-informative" priors such as $\pi(\mu) \propto 1$ or $\pi(\mu) \propto 1/\mu$ lead to improper posteriors. The next choice, $\pi(\mu) \propto 1/\mu^2$, does lead to a proper posterior, but the resulting HPD Bayes estimate of μ is bounded from above, regardless of the measured value x! Similarly, HPD intervals always exclude μ values above a certain threshold, with the consequence that their coverage drops to zero above that threshold.

One would think that a reference analysis of this problem will yield a more satisfactory solution due to its invariance properties.

Example: Transverse Momentum Measurement (2)

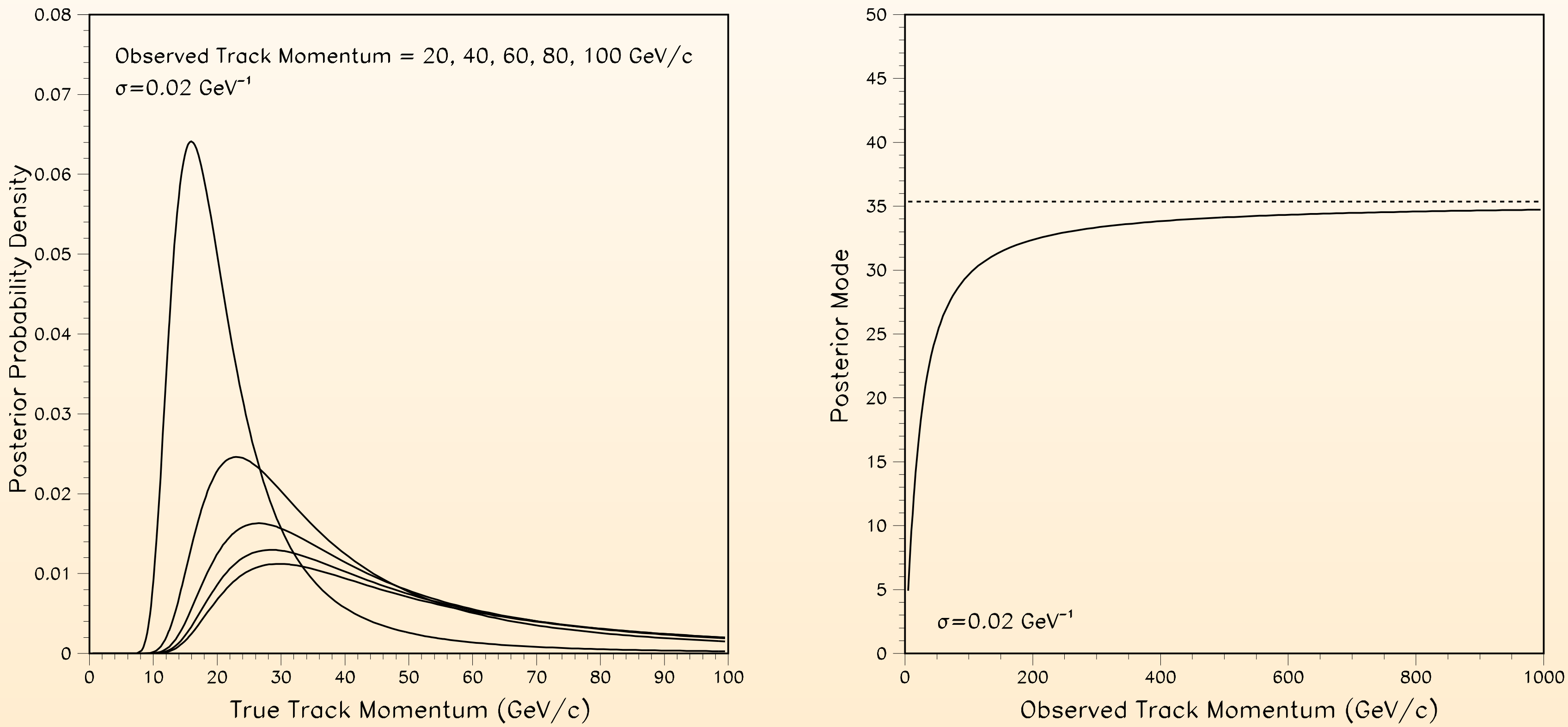

Figure 3: Left: posterior densities for $1/\mu^2$ prior; Right: posterior mode versus observed track momentum.

Example: Transverse Momentum Measurement (3)

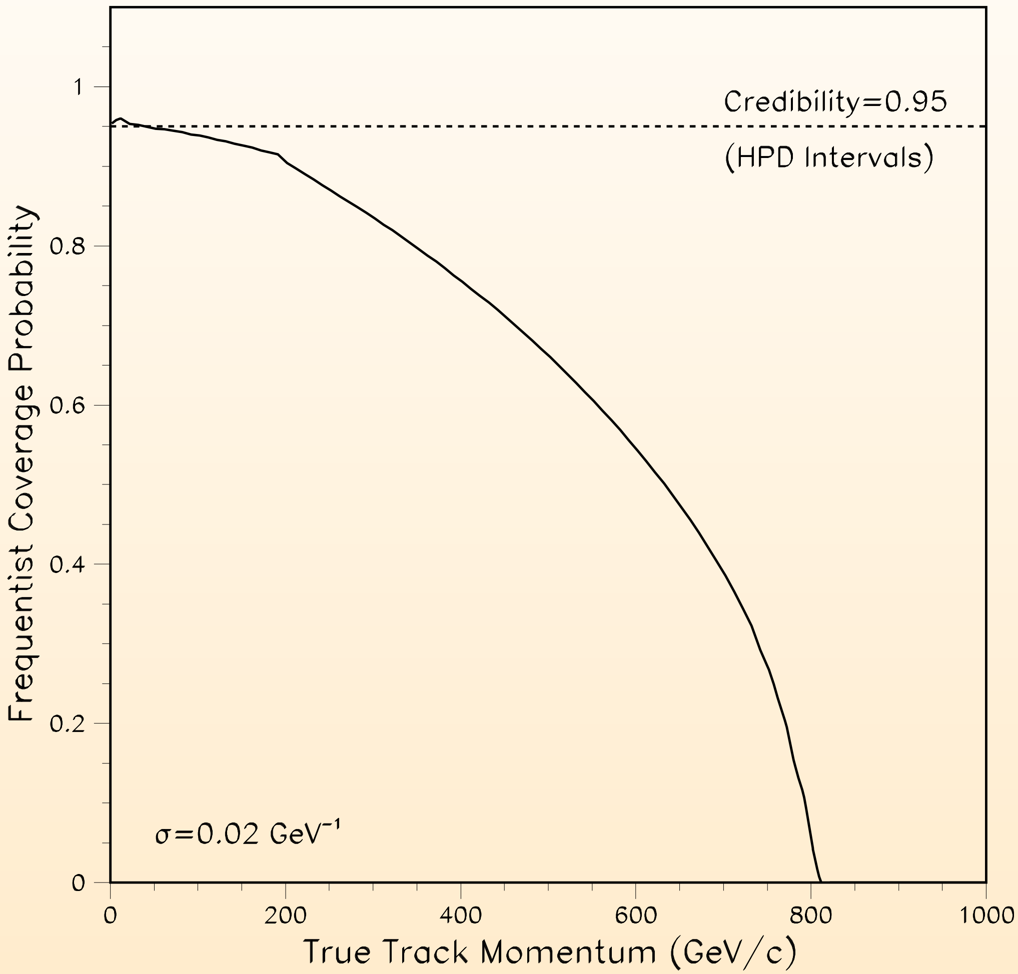

Figure 4: Coverage probability of Highest Posterior Density intervals as a function of true track momentum.

Example: Transverse Momentum Measurement (4)

Fortunately, a reference analysis of this problem can be done entirely analytically:

1. Intrinsic discrepancy:

$$\delta\{\hat{\mu}, \mu\} \;=\; \frac{1}{2}\left(\frac{1/\mu - 1/\hat{\mu}}{\sigma}\right)^2.$$

2. Reference prior when μ is the quantity of interest: $\pi(\mu) \propto 1/\mu^2$.
3. Reference prior when δ is the quantity of interest. Since δ is a piecewise one-to-one function of μ, this reference prior is also $1/\mu^2$.
4. Reference posterior:

$$p(\mu \,|\, x) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{1/x - 1/\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\,\sigma\,\mu^2}.$$

5. Reference posterior expected intrinsic loss:

$$d(\hat{\mu} \,|\, x) \;=\; \frac{1}{2} + \frac{1}{2}\left(\frac{1/x - 1/\hat{\mu}}{\sigma}\right)^2.$$

Example: Transverse Momentum Measurement (5)

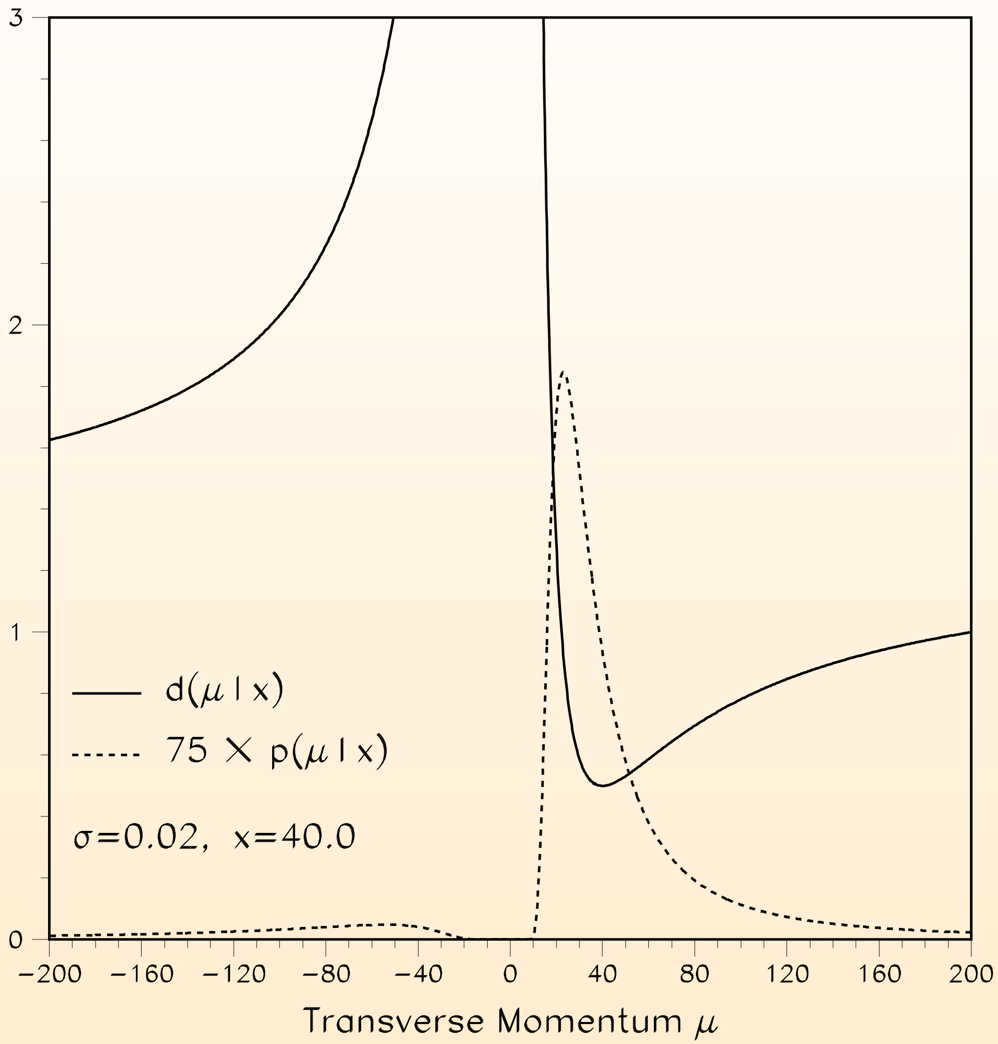

Figure 5: Reference posterior expected intrinsic loss $d(\mu \,|\, x)$ (solid line), and reference posterior density $p(\mu \,|\, x)$ (dashed line) for the problem of measuring transverse momenta in a tracking chamber.

Example: Transverse Momentum Measurement (6)

The results of the reference analysis are as follows:

- The intrinsic estimate of μ, i.e. the value of μ that minimizes the reference posterior expected intrinsic loss, is $\mu^\star = x$.

- Minimum reference posterior expected intrinsic loss intervals have the form:

$$\text{If } d < \frac{1}{2} + \frac{1}{2\sigma^2 x^2}: \qquad \left[\frac{x}{1+\sigma x\sqrt{2d-1}}, \frac{x}{1-\sigma x\sqrt{2d-1}}\right],$$

$$\text{If } d = \frac{1}{2} + \frac{1}{2\sigma^2 x^2} \text{ and } x \geq 0: \qquad \left[\frac{x}{2}, +\infty\right],$$

$$\text{If } d = \frac{1}{2} + \frac{1}{2\sigma^2 x^2} \text{ and } x < 0: \qquad \left[-\infty, \frac{x}{2}\right],$$

$$\text{If } d > \frac{1}{2} + \frac{1}{2\sigma^2 x^2}: \qquad \left[-\infty, \tfrac{x}{1-\sigma x\sqrt{2d-1}}\right] \cup \left[\tfrac{x}{1+\sigma x\sqrt{2d-1}}, +\infty\right],$$

where d is determined by the requirement of a specified posterior probability content. Note that $\mu^\star$ is contained in all the intrinsic intervals.

Reference Analysis and Hypothesis Testing (1)

The usual Bayesian approach to hypothesis testing is based on *Bayes factors*. Unfortunately this approach tends to fail when one is testing a precise null hypothesis ($H_0 : \theta = \theta_0$) against a "vague" alternative ($H_1 : \theta \neq \theta_0$) (cfr. Lindley's paradox).

Reference analysis provides a solution to this problem by recasting it as a decision problem with two possible actions:

1. a_0: Accept H_0 and work with $p(x \mid \theta_0)$.
2. a_1: Reject H_0 and keep the unrestricted model $p(x \mid \theta)$.

The consequence of each action can be described by a loss function $\ell(a_i, \theta)$, but actually, only the *loss difference* $\Delta\ell(\theta) = \ell(a_0, \theta) - \ell(a_1, \theta)$, which measures the advantage of rejecting H_0 as a function of θ, needs to be specified. Reference analysis uses the intrinsic discrepancy between the distributions $p(x \mid \theta_0)$ and $p(x \mid \theta)$ to define this loss difference:

$$\Delta\ell(\theta) \;=\; \delta\{\theta_0, \theta\} - d^\star,$$

where $d^\star$ is a positive constant measuring the advantage of being able to work with the simpler model when it is true.

Reference Analysis and Hypothesis Testing (2)

Given available data x, the *Bayesian reference criterion* (BRC) rejects H_0 if the reference posterior expected intrinsic loss exceeds a critical value $d^\star$, i.e. if:

$$d(\theta_0 \,|\, x) \;=\; \int_\Theta d\theta \; \delta\{\theta_0, \theta\} \; \pi_\delta(\theta \,|\, x) \;>\; d^\star.$$

Properties of the BRC:

- As the sample size increases, the expected value of $d(\theta_0 \,|\, x)$ under sampling tends to one when H_0 is true, and tends to infinity otherwise;
- The interpretation of the intrinsic discrepancy in terms of the minimum posterior expected likelihood ratio in favor of the true model provides a direct calibration of the required critical value $d^\star$:

$$\begin{aligned}
d^* &\approx \ln(10) &&\approx 2.3 : &&\text{“mild evidence against } H_0\text{”;}\\
d^* &\approx \ln(100) &&\approx 4.6 : &&\text{“strong evidence against } H_0\text{”;}\\
d^* &\approx \ln(1000) &&\approx 6.9 : &&\text{“very strong evidence against } H_0\text{”.}
\end{aligned}$$

- In contrast with frequentist hypothesis testing, the statistic d is measured on an absolute scale which remains valid for any sample size and any dimensionality.

Summary

- Noninformative priors have been studied for a long time and most of them have been found defective in more than one way. Reference analysis arose from this study as the only *general* method that produces priors that have the required *invariance* properties, deal successfully with the *marginalization* paradoxes, and have consistent *sampling* properties.

- Reference priors should not be interpreted as probability distributions expressing subjective degree of belief; instead, they help answer the question of what could be said about the quantity of interest if one's prior knowledge were dominated by the data.

- Reference analysis also provides methods for summarizing the posterior density of a measurement. Intrinsic point estimates, credible intervals, and hypothesis tests have invariance properties that are essential for *scientific* inference.

- There exist numerical algorithms to compute reference priors, and the CMS statistics committee hopes to implement one of these for general use.

Some References

D.M. Appleby, "Probabilities are single-case, or nothing," arXiv:quant-ph/0408058v1 (8 Aug 2004).

Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack, "Quantum probabilities as Bayesian probabilities," Phys. Rev. **A** 65, 022305 (2002).

Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack, "Subjective probability and quantum certainty," arXiv:quant-ph/0608190v2 (26 Jan 2007).

E.T. Jaynes, "Probability in quantum theory," `http://bayes.wustl.edu/etj/articles/prob.in.qm.pdf` (1990).

José M. Bernardo, "Reference analysis," `http://www.uv.es/~bernardo/RefAna.pdf` (2005).