Bayesian Reference Analysis

Luc Demortier

The Rockefeller University

"Statistical Inference Problems in High Energy Physics and Astronomy"

Banff International Research Station for Mathematical Innovation and Discovery

July 15–20, 2006

Reference analysis is an objective Bayesian methodology that helps to solve problems in the following areas:

1. Calculating reference priors;
2. Estimating parameters;
3. Calculating credible regions;
4. Testing hypotheses;

After a brief introduction, this talk focuses on the last three topics, using a typical example from high energy physics.

Motivation

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 $\pi(\theta \,|\, x) = \pi(\theta \,|\, t)$.
- *Consistent Marginalization* If $p(x \,|\, \theta, \lambda) \to \pi_1(\theta \,|\, x) = \pi_1(\theta \,|\, t)$, $t = t(x)$ and $p(t \,|\, \theta, \lambda) = p(t \,|\, \theta) \to \pi_2(\theta \,|\, t)$, Then $\pi_2(\theta \,|\, t) = \pi_1(\theta \,|\, t)$.
- *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.

Intrinsic Discrepancy (1)

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

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

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:

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

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.

Intrinsic Discrepancy (2)

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

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

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:

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

Properties of the intrinsic discrepancy (continued):

- 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 $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ between two parametric families of distributions does not depend on their parametrizations.
- 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.

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

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

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:

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

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

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)

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 (3)

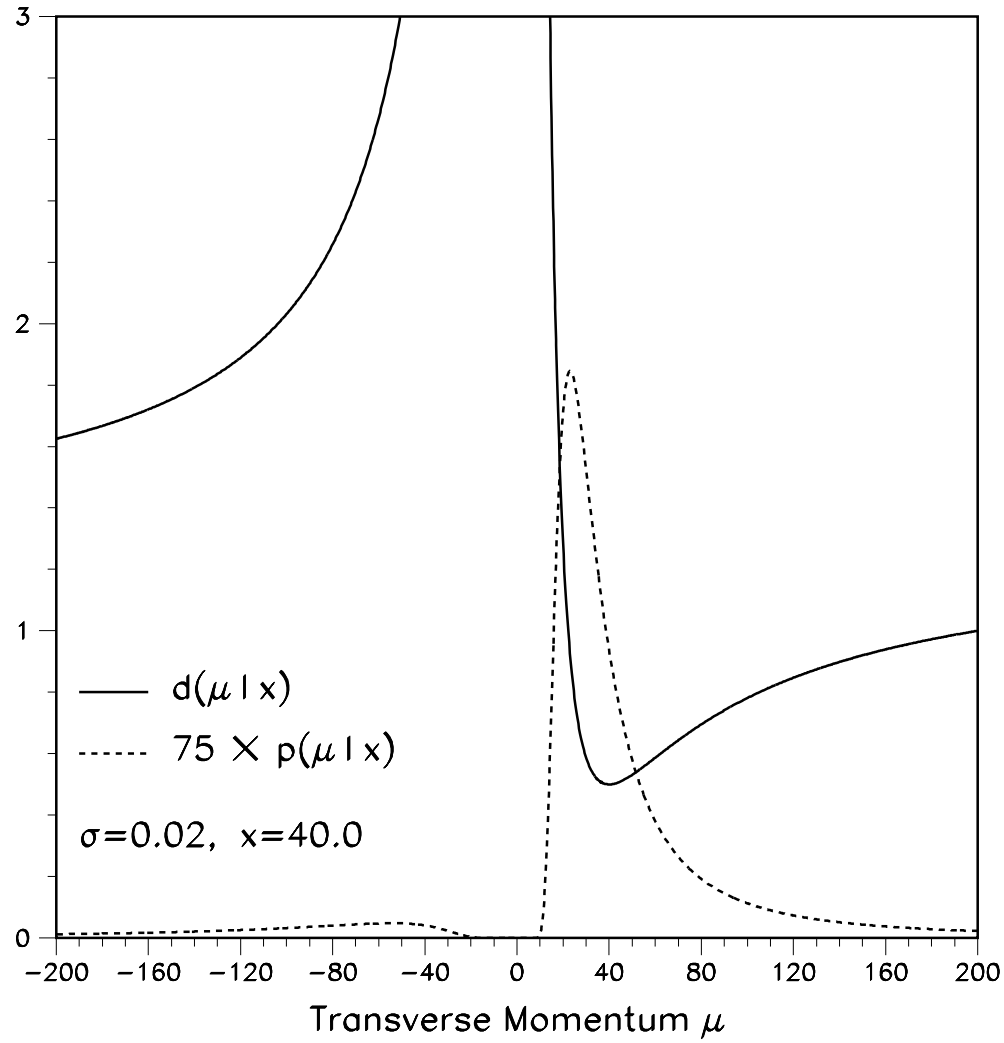

Figure 1: Reference posterior expected intrinsic loss $d(\mu \,|\, x)$ (solid line), and reference posterior density $p(\mu \,|\, x)$ (dashed line) for the problem of measuring transverse momenta in a tracking chamber.

Example: Transverse Momentum Measurement (4)

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:

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

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

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

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

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 $p(x \,|\, \theta_0)$.
2. a_1: Reject H_0 and keep the unrestricted model $p(x \,|\, \theta)$.

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 $p(x \,|\, \theta_0)$ and $p(x \,|\, \theta)$ 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 $d^\star$:
  - $d^* \approx \ln(10) \approx 2.3$: "mild evidence against H_0";
  - $d^* \approx \ln(100) \approx 4.6$: "strong evidence against H_0";
  - $d^* \approx \ln(1000) \approx 6.9$: "very strong evidence against H_0".
- 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.

- Much more information can be found on José Bernardo's web page, `http://www.uv.es/~bernardo/publications.html`.