# BAYESIAN REFERENCE ANALYSIS

LUC DEMORTIER
*The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA*
*E-mail: luc@fnal.gov*

As a carefully thought-out attempt to develop the objective side of Bayesian inference, reference analysis provides procedures for point and interval estimation, hypothesis testing, and the construction of objective posterior distributions. For physicists, the interest of these procedures lies in their very general applicability, their invariance under reparametrization, their coherence, and their good performance under repeated sampling.

## 1 Introduction

One aspect that distinguishes experimental inference in physics from that in other sciences is the objective randomness of quantum processes. As a result, statistical models for quantum phenomena are exact, supporting a strict frequentist analysis of their measurement. Nevertheless, Caves *et al.*[1] have brilliantly motivated a subjective Bayesian interpretation of quantum probabilities, whose form depends on the information available to the observer but is otherwise fully prescribed by a fundamental law. When dealing with actual measurements however, no fundamental law constrains their analysis, summary and report, so that some other objective method must be found.

Ideally, such a method should be very general, applicable to all kinds of measurements regardless of the number and type of parameters and data involved. It should make use of *all* available information, and coherently so, in the sense that if there is more than one way to extract all relevant information from data, the final result will not depend on the chosen way. The desiderata of generality, exhaustiveness and coherence are satisfied by Bayesian procedures, but that of objectivity is more problematic due to the Bayesian requirement of specifying prior probabilities in terms of degrees of belief. Reference analysis[2], an objective Bayesian method developed over the past twenty-five years, solves this problem by replacing the question "what is our prior degree of belief?" by "what would our posterior degree of belief be, if our prior knowledge had a minimal effect, relative to the data, on the final inference?"

In addition to an objective method for specifying priors, reference analysis provides techniques to summarize posterior distributions in terms of point estimates and intervals, and to test precise hypotheses against vague alternatives, a notoriously subtle problem. All these techniques are based on information theory, and in particular on the central concept of intrinsic discrepancy between two probability distributions. This concept is introduced in section 2 and applied to the definition of reference priors in section 3. Section 4 describes the extraction of intrinsic point and interval estimates from posterior distributions.

Due to space limitations, the development of the paper is rather conceptual, with few details in the calculations. The interested reader is encouraged to consult the references, especially Bernardo[2].

## 2 Intrinsic Discrepancy and Missing Information

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

$$\delta\{p_1, p_2\} \;=\; \min\{\, \kappa\{p_1 \,|\, p_2\},\; \kappa\{p_2 \,|\, p_1\}\,\}\,, \qquad (1)$$

$$\text{where} \quad \kappa\{p_i \,|\, p_j\} \;\equiv\; \int dx\; p_j(x)\;\ln\frac{p_j(x)}{p_i(x)} \qquad (2)$$

is the Kullback-Leibler divergence between $p_i$ and $p_j$. 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. It is invariant under one-to-one transformations of $x$, and is information-additive: the discrepancy for a set of $n$ independent observations is $n$ times the discrepancy for one observation. A simple interpretation of $\delta\{p_1, p_2\}$ is as a measure, in natural information units, of the minimum amount of information that one observation may be expected to provide in order to discriminate between $p_1$ and $p_2$. Another interpretation is as the minimum expected log-likelihood ratio in favor of the probability model that generates the data.

Suppose now that we have a parametric model for some data $x$:

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

and consider the joint probability density of $x$ and $\theta$, $p(x,\theta) = p(x \mid \theta)\, p(\theta)$, where $p(\theta)$ is a prior for $\theta$. Relative to the product of marginals $p(x)p(\theta)$, the joint density captures in some sense the information carried by $x$ about $\theta$. This suggests defining the expected intrinsic information $I\{p(\theta) \mid \mathcal{M}\}$, from one observation of $\mathcal{M}$ about the value of $\theta$ when the prior density is $p(\theta)$, as:

$$I\{p(\theta) \mid \mathcal{M}\} \ = \ \delta\{p(x,\theta),\ p(x)\, p(\theta)\}, \qquad (3)$$

where $p(x) = \int d\theta\ p(x \mid \theta)\, p(\theta)$. According to this definition, the stronger the prior knowledge described by $p(\theta)$, the smaller the information the data may be expected to provide, and vice-versa. In the limit where $p(\theta)$ is a delta function, $I\{p(\theta) \mid \mathcal{M}\} = 0$

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

## 3 Reference Priors

Let $\mathcal{P}$ be a class of sufficiently regular priors that are compatible with whatever initial information is available about the value of $\theta$. The reference prior is defined to be that prior function $\pi(\theta) = \pi(\theta \mid \mathcal{M}, \mathcal{P})$ which maximizes the missing information about the value of $\theta$ within the class $\mathcal{P}$. The limiting procedure used to define the missing information requires some care in the calculation of $\pi(\theta)$. Formally, one introduces an increasing sequence of subsets $\Theta_i$ of the parameter space $\Theta$, such that $\lim_{i\to\infty} \Theta_i = \Theta$ and $\int_{\Theta_i} \pi(\theta)\ d\theta < \infty$. The reference prior $\pi(\theta)$ is then defined as satisfying:

$$\lim_{k\to\infty} \left[ I\{\pi_i \mid \mathcal{M}^k\} - I\{p_i \mid \mathcal{M}^k\} \right] \ \geq \ 0 \qquad \text{for all } \Theta_i,\ \text{for all } p \in \mathcal{P}, \quad (4)$$

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

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. If the parameter is continuous and one-dimensional, and regularity conditions that guarantee asymptotic normality are satisfied, then the reference prior is Jeffreys' prior:

$$\pi(\theta) \propto i(\theta)^{1/2},$$

$$\text{where} \quad i(\theta) = -\int_{\mathcal{X}} dx\ p(x \mid \theta)\ \frac{\partial^2}{\partial\theta^2} \ln p(x \mid \theta). \quad (5)$$

Note that in the definition of reference priors, 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. A practical advantage of this limiting procedure is that it ensures that reference priors only depend on the asymptotic behavior of the model, thereby greatly simplifying their derivation.

It can be shown that reference priors are independent of sample size, compatible with sufficient statistics (meaning that their form does not depend on whether the model is or is not expressed in terms of sufficient statistics), and consistent under reparametrization (i.e. if $\phi$ is a one-to-one transformation of $\theta$, then their reference posterior densities are related by $\pi(\phi \mid x) = \pi(\theta \mid x)\, |d\theta/d\phi|$).

Finally, it is important to emphasize that 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.

### *3.1 Treatment of Nuisance Parameters*

Suppose the statistical model is $p(x \mid \theta, \lambda)$, with $\theta$ the parameter of interest and $\lambda$ a nuisance parameter. We now need a joint reference prior $\pi(\theta, \lambda)$. The algorithm is sequential:

1. Hold $\theta$ constant and apply the one-parameter reference algorithm to obtain the conditional reference prior $\pi(\lambda \mid \theta)$.
2. Derive the one-parameter integrated model:
$$p(x \mid \theta) \ = \ \int_{\Lambda} d\lambda\ p(x \mid \theta, \lambda)\ \pi(\lambda \mid \theta),$$
where $\Lambda$ is the parameter space for $\lambda$.
3. Apply the one-parameter reference algorithm again, this time to $p(x \mid \theta)$, and obtain the marginal reference prior $\pi(\theta)$.
4. Set $\pi(\theta, \lambda) \ = \ \pi(\lambda \mid \theta)\ \pi(\theta)$.

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

The above algorithm is easily generalized to any number of parameters. However, its sequential character requires that the parameters be ordered. In most applications the order does not affect the result, but there are exceptions. Different orderings may then be used as part of a robustness analysis.

Within a *single* model it is in principle possible to have as many reference priors as there are potential parameters of interest. Indeed, there is no reason for a setup that maximizes the missing information about a parameter $\theta$ to be identical to a setup that maximizes the missing information about a parameter $\eta$, unless $\eta$ is a one-to-one function of $\theta$.

### *3.2 Example: a Cross Section Measurement*

We illustrate the construction of reference priors with a common problem in high energy physics, that of extracting a cross section $\sigma$ from an observed number of events $n$. The latter is assumed to have a Poisson distribution with a mean of the form $b + \epsilon\sigma$, where the sensitivity factor $\epsilon$ and the background $b$ are nuisance parameters. The model is:

$$p(n \,|\, \sigma, \epsilon, b) \;=\; \frac{(b+\epsilon\sigma)^n}{n!}\, e^{-b-\epsilon\sigma}. \qquad (6)$$

Note that $\sigma$, $\epsilon$, and $b$ are not identifiable from a given $n$. This problem is usually addressed by using information from calibration data or simulation studies to form a proper, subjective prior for $\epsilon$ and $b$, say $\pi(\epsilon, b)$. We must therefore find the conditional reference prior $\pi(\sigma \,|\, \epsilon, b)$. If $\epsilon$ and $b$ were exactly known, the reference prior for $\sigma$ would simply be Jeffreys' prior. From the Fisher information for $\sigma$:

$$\Sigma_{\sigma\sigma} \;=\; \mathrm{E}\left[-\frac{\partial^2}{\partial\sigma^2}\ln p(n \,|\, \sigma, \epsilon, b)\right] \;=\; \frac{\epsilon^2}{b+\epsilon\sigma}, \qquad (7)$$

this Jeffreys' prior is calculated to be:

$$\pi_{\mathcal{J}}(\sigma \,|\, \epsilon, b) \;\propto\; \frac{\epsilon}{\sqrt{b+\epsilon\sigma}}. \qquad (8)$$

However, this is *not* the reference prior for this problem, i.e. the prior that would be obtained by strict application of equation (4). Although the $\sigma$ dependence of $\pi_{\mathcal{J}}$ is correct, its $\epsilon$ dependence is not, and this matters because $\pi_{\mathcal{J}}$ is improper and $\epsilon$ is an unknown parameter. As shown in Sun and Berger[3], the correct reference prior is obtained by renormalizing the above prior using a sequence of nested compact sets for $\sigma$. A natural choice for these sets is $[0, u]$, with $u > 0$. Normalizing the above prior over such a set yields:

$$\pi_u(\sigma \,|\, \epsilon, b) \;=\; \frac{\epsilon}{\sqrt{b+\epsilon\sigma}}\;\frac{\mathbf{1}(u \geq \sigma)}{2\sqrt{b+\epsilon u}\;-\;2\sqrt{b}},$$

where $\mathbf{1}(u \geq \sigma)$ is 1 if $u \geq \sigma$ and 0 otherwise. The correct conditional reference prior is then:

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

with $(\sigma_0, \epsilon_0, b_0)$ any fixed point. Although this prior is still improper, its $\epsilon$ dependence is different from that of equation (8).

We can now write down the reference posterior when $\sigma$ is the parameter of interest:

$$\pi(\sigma \,|\, n) \propto \int_0^{\infty} d\epsilon \int_0^{\infty} db\; \frac{(b+\epsilon\sigma)^n\, e^{-b-\epsilon\sigma}}{n!}\;\frac{\sqrt{\epsilon}\,\pi(\epsilon, b)}{\sqrt{b+\epsilon\sigma}}. \qquad (9)$$

An important aspect of reference posteriors is their behavior under repeated sampling. To test this, we calculate an upper limit $U$ on $\sigma$, assuming a product of gamma densities for the subjective prior $\pi(\epsilon, b)$:

$$\pi(\epsilon, b) \;=\; \frac{\tau(\tau\epsilon)^{x-1/2}\, e^{-\tau\epsilon}}{\Gamma(x+1/2)}\;\frac{c(cb)^{y-1/2}\, e^{-cb}}{\Gamma(y+1/2)}. \qquad (10)$$

As we are dealing with a mixture of subjective and objective priors, some care is needed in specifying the ensemble with respect to which the coverage of $U$ is to be calculated. Datta and Sweeting[4] suggest to *average* the coverage with respect to the subjective components of the prior. An example of calculation based on this prescription is shown in Figure 1.

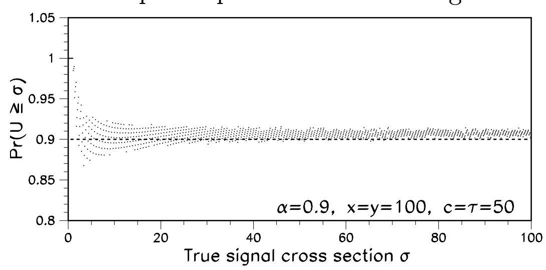


Figure 1. Coverage of 90% credibility level reference Bayes upper limits on a signal cross section $\sigma$, as a function of the true value of that cross section.

The coverage appears to converge asymptotically towards the credibility level. Although this behaviour is typical of all sufficiently regular priors, in many cases the convergence is faster when a reference prior is used.

## 4 Intrinsic Estimation and Testing

It is well known 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 and interval estimation as decision problems. Suppose that $\hat{\theta}$ is an estimate of the parameter $\theta$, whose true value $\theta_t$ 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 $\theta_b$ of $\theta$ minimizes the corresponding posterior loss:

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

In physics, interest usually focuses on the actual mechanism that governs the data. Therefore we need point and interval estimates that are invariant under one-to-one transformations of the parameter and the data (including reduction to sufficient statistics). A loss function that will deliver such an estimate is the intrinsic discrepancy: $\ell(\hat{\theta}, \theta_t) = \delta\{p(x \,|\, \hat{\theta}), p(x \,|\, \theta_t)\}$. Its reference posterior expectation is:

$$d(\hat{\theta} \,|\, x) \;=\; \int_\Theta d\theta \; \delta\{p(x \,|\, \hat{\theta}), p(x \,|\, \theta)\} \; \pi_\delta(\theta \,|\, x), \quad (11)$$

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

The *intrinsic estimator* of $\theta$ minimizes $d(\hat{\theta} \,|\, x)$:

$$\theta^\star(x) \;=\; \arg \min_{\hat{\theta} \in \Theta} d(\hat{\theta} \,|\, x), \quad (12)$$

and an intrinsic $\alpha$-credible region for $\theta$ is a subset $R^\star_\alpha$ of the parameter space $\Theta$ such that:

$$\int_{R^\star_\alpha} d\theta \; \pi(\theta \,|\, x) \;=\; \alpha, \quad \text{and}$$
$$\text{for all } \theta \in R^\star_\alpha, \; \theta' \notin R^\star_\alpha : d(\theta \,|\, x) \leq d(\theta' \,|\, x). \quad (13)$$

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, subjective prior information is available.

Finally, in hypothesis testing, a typical problem is to decide whether a precise value $\theta_0$ may be used as a "proxy" for the unknown value of $\theta$. The reference approach is to use $d(\theta_0 \,|\, x)$ from equation (11), with $\theta_0$ replacing $\hat{\theta}$, as an intrinsic test statistic. Its magnitude is a direct measure of the evidence against the null hypothesis $\theta = \theta_0$.

## 5 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.

## References


1. Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack, "Quantum probabilities as Bayesian probabilities," Phys. Rev. **A** 65, 022305 (2002).
2. José M. Bernardo, "Reference Analysis," Handbook of Statistics **25** (D. Dipak and C.R. Rao, eds.) Amsterdam: Elsevier, 2005. See also `http://www.uv.es/~bernardo/publications.html`.
3. D. Sun and J. O. Berger "Reference priors with partial information," Biometrika **85**, 55 (1998).
4. Gauri Sankar Datta and Trevor J. Sweeting, "Probability matching priors," Research Report No. 252, Department of Statistical Science, University College London (March 2005); also at `http://www.ucl.ac.uk/Stats/research/Resrprts/psfiles/rr252.pdf`.