INL/CON-10-18287
PREPRINT

Minimally Informative Prior Distributions for PSA

PSAM-10

Dana L. Kelly
Robert W. Youngblood
Kurt G. Vedros

June 2010

The INL is a
U.S. Department of Energy
National Laboratory
operated by
Battelle Energy Alliance

Minimally Informative Prior Distributions for PSA

Dana L. Kelly[a1], Robert W. Youngblood[a], and Kurt G. Vedros[a]
[a]Idaho National Laboratory, Idaho Falls, ID USA

Abstract: A salient feature of Bayesian inference is its ability to incorporate information from a variety of sources into the inference model, via the prior distribution (hereafter simply "the prior"). However, over-reliance on old information can lead to priors that dominate new data. Some analysts seek to avoid this by trying to work with a minimally informative prior distribution. Another reason for choosing a minimally informative prior is to avoid the often-voiced criticism of subjectivity in the choice of prior. Minimally informative priors fall into two broad classes: 1) so-called noninformative priors, which attempt to be completely objective, in that the posterior distribution is determined as completely as possible by the observed data, the most well known example in this class being the Jeffreys prior, and 2) priors that are diffuse over the region where the likelihood function is non-negligible, but that incorporate some information about the parameters being estimated, such as a mean value. In this paper, we compare four approaches in the second class, with respect to their practical implications for Bayesian inference in Probabilistic Safety Assessment (PSA). The most commonly used such prior, the so-called constrained noninformative prior, is a special case of the maximum entropy prior. This is formulated as a conjugate distribution for the most commonly encountered aleatory models in PSA, and is correspondingly mathematically convenient; however, it has a relatively light tail and this can cause the posterior mean to be overly influenced by the prior in updates with sparse data. A more informative prior that is capable, in principle, of dealing more effectively with sparse data is a mixture of conjugate priors. A particular diffuse nonconjugate prior, the logistic-normal, is shown to behave similarly for some purposes. Finally, we review the so-called robust prior. Rather than relying on the mathematical abstraction of entropy, as does the constrained noninformative prior, the robust prior places a heavy-tailed Cauchy prior on the canonical parameter of the aleatory model.

Keywords: PRA, Bayesian inference, prior distribution.

1. INTRODUCTION

A salient feature of Bayesian inference is its ability to incorporate information from a variety of sources into the inference model, via the prior distribution (hereafter simply "the prior"). Done properly, Bayesian inference integrates old information and new information into an evidence-based state-of-knowledge distribution. However, if the situation being evaluated is changing with time, then over-reliance on old information in formulating the prior can lead to priors that excessively dominate new data.

Some analysts seek to avoid this by trying to work with a minimally informative (less direct but synonymous terms are *diffuse*, *weak*, and *vague*) prior distribution. Another reason for choosing a minimally informative prior is to avoid the often-voiced criticism of subjectivity in the choice of prior. Minimally informative priors fall into two broad classes: 1) so-called noninformative priors, which attempt to be completely objective, in that the posterior distribution is determined as completely as possible by the observed data. The most well known example in this class is the Jeffreys prior; 2) priors that are diffuse over the region where the likelihood function is non-negligible, but that incorporate some information about the parameters being estimated, such as a mean value. The reader is referred to (1) for a thorough review of prior distributions in the first class. In this paper, we compare four approaches in the second class, with respect to their practical implications for Bayesian inference in PSA. The most commonly used such prior, the so-called constrained noninformative

[1] Dana.Kelly@inl.gov

prior (CNIP) (2), is a special case of the maximum entropy prior, which is discussed by (3) and others. The CNIP is formulated as a conjugate distribution for the most commonly encountered aleatory models in PSA, and is correspondingly mathematically convenient; but it has a relatively light tail, and is correspondingly somewhat unresponsive to updates with sparse data, an issue discussed in (4) in the context of the Mitigating System Performance Index. Other issues with maximum entropy priors are discussed by (5) and (6). A more informative prior that is capable, in principle, of dealing more effectively with sparse data is a mixture of conjugate priors, as discussed by (7) and (8). A particular diffuse nonconjugate prior, the logistic-normal, is shown to behave similarly for some purposes. Finally, we review the so-called robust prior, first described by (5). Rather than relying on the mathematical abstraction of entropy, as does the constrained noninformative prior, the robust prior places a heavy-tailed Cauchy prior on the canonical parameter of the aleatory model.

2. CONSTRAINED NONINFORMATIVE PRIOR

The constrained noninformative prior (CNIP) is, as pointed out by (6), a type of maximum entropy prior distribution. Prior to the advent of the CNIP in (2), the most prevalent definition of entropy in PSA was the straightforward extension of the Shannon entropy to the case of a continuous variable:

$$H = -\int \pi(\theta) \log[\pi(\theta)] d\theta \tag{1}$$

The CNI prior uses a definition of entropy due to Jaynes (3), which defines entropy as the negative of the Kullback-Leibler distance between $\pi(\theta)$ and the "natural" noninformative prior:[2]

$$H = -\int \pi(\theta) \log\left[\frac{\pi(\theta)}{\pi_{NI}(\theta)}\right] d\theta \tag{2}$$

There is ambiguity as to what the "natural" noninformative prior should be, and (2) adopted the Jeffreys prior for $\pi_{NI}(\theta)$. The attractiveness of the CNI prior was that it is, like the Jeffreys prior, invariant to reparameterization. However, like the maximum entropy prior under the extended Shannon definition, the CNI prior can fail to exist, even in simple models such as exponential time to failure. Also note that in the case of continuous distributions, entropy under either definition is often negative.

In the setting of a binomial aleatory model, the unknown parameter in the above equations, θ, is equal to p, the probability of failure on demand in each Bernoulli trial. In this case, the CNI prior cannot be written down in closed form, but can be approximated well by a beta distribution with first parameter approximately equal to 0.5, and second parameter determined from the specified mean constraint. The maximum entropy prior under the extended Shannon definition can be written in closed form as a truncated exponential distribution, as given in (6), and for small values of p, it is approximately an exponential distribution with rate equal to the reciprocal of the specified mean constraint.[3] The figure below shows these two maximum entropy prior densities for a mean of 0.001. Note the vertical asymptote at zero, which is characteristic of the CNI prior. The asymptote is inherited from the Jeffreys prior, which is a beta(0.5, 0.5) distribution.

---

[2] Under this definition, the Shannon entropy is the negative of the Kullback-Leibler distance from a uniform distribution. Thus, the maximum entropy prior under the extended Shannon definition is as close as possible (in terms of K-L distance) to a uniform distribution.

[3] Because of space constraints, we treat only the binomial model explicitly. Note that the CNI prior for the Poisson(λt) model is a gamma distribution with shape parameter = 0.5 and rate parameter = 1/(2×mean). Under the extended Shannon definition, the maximum entropy prior is exponential with rate = mean.

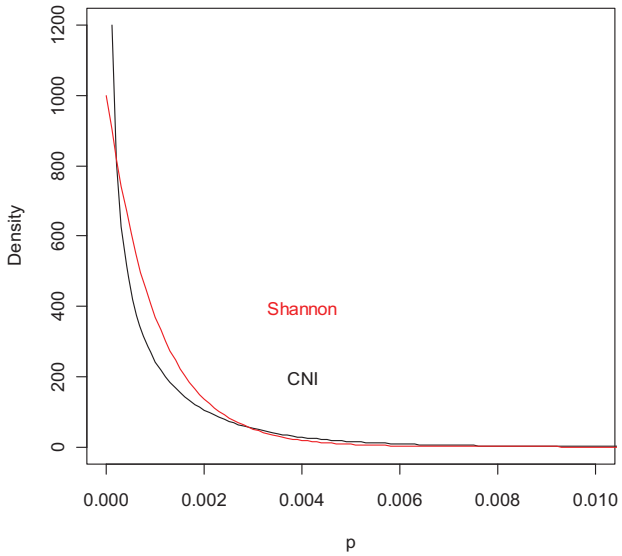

Figure 1 Two maximum entropy densities for p, both with mean = 0.001, CNI prior displays vertical asymptote at p = 0. To specify the maximum entropy prior, one must first specify the definition of entropy.

So, as pointed out by (6), there is an unavoidable arbitrariness in "maximum entropy priors," because there is no clear definition of entropy for a continuous random variable. However, entropy is a measure of uncertainty, and so one would expect a maximum entropy prior, which in a specialized mathematical sense is maximally uncertain, to be minimally "stiff" (maximally responsive) in terms of how it responds to data in Bayesian updating. But does this turn out to be the case?

2.1 Bayesian Updating of CNI Prior

Throughout this paper, we will take as a running example the failure to start of a motor-driven pump, assumed to have a mean failure probability of 0.001. We will assume that there are 50 demands on this pump, and that failures to start are described by a binomial distribution with parameters p and 50. From (2), the beta distribution that approximates the CNI prior has parameters $\alpha = .498$ and $\beta = 498$. The posterior distribution is beta($\alpha + x$, $\beta + 50 - x$), where x is the number of failures observed in 50 demands.

3. MIXTURE PRIOR

Refs. (7) and (8) proposed the use of a mixture of two conjugate prior distributions, one representing performance of a degraded component, the other representing performance of a component in its normal state. Of these two mixture prior models, the one most applicable to the present situation is the "variable-constituent" prior, described in (8). This prior was originally formulated in the context of performance assessment; the presumption is that performance (e.g., fail-to-start probability) can vary with time, and the application of the prior is to assess current performance based on current data. As mentioned in the Introduction, in an application such as this, it is clearly inappropriate to bias the prior towards long-term average performance; such a bias would be a case of old information dominating the new.

For simplicity, the treatment will consider only two performance states: "good" and "degraded." The implementation is straightforward even if more performance states are introduced (e.g., slightly

degraded, average, …), but this is not warranted for purposes of illustration. The mixture prior is then formulated in terms of a probability distribution conditional on being in the good state, a probability distribution conditional on being in the degraded state, and a mixture parameter representing the probability of being in the degraded state:

$$g_{mix}(p) = (1 - \pi)g_{conj}(p\ ;\ \alpha_0, \beta_0) + \pi g_{conj}(p\ ;\ \alpha_1, \beta_1) \quad (1)$$

where π is the probability of being in the degraded state, g_{conj} is the natural conjugate distribution (beta in this case, which is conjugate to the binomial distribution), and α and β are the parameters of the distribution, the subscript "0" denoting the distribution conditional on being in the "good" state and the subscript "1" denoting the parameters of the distribution conditional on being in the degraded state. Thus, the mixture prior contains five parameters: π, α_0, β_0, α_1, and β_1. Data are observed, and then the five parameters are updated by Bayes' theorem to give their posterior values.

The mixture prior is conjugate, in the sense that the posterior distribution has the same form as the prior distribution, differing only in the parameter values. Because the constituent distributions g_i are updated, this model was called the variable-constituent model by (8). Ref. (7) describes a variant, the "fixed-constituent model," in which the prior has the same form, but only the mixture parameter is updated. We will not discuss this variant further here.

The mixture prior was shown by (8) to behave differently from the CNI prior in examples of practical interest. For regions of parameter space explored in (8), the posterior mean expressed as a function of the observed (small) number of failures increased more slowly than the mean from updating the CNI prior, but increased more rapidly than the update of the CNI prior as the number of observed failures increased. This behavior, among other features, makes the mixture prior an interesting candidate for performance assessment applications, but quantification of its five parameters is seen as a significant practical disadvantage.

The parameters of the mixture prior were chosen using the approach described in (7). In particular, the overall mean was taken to be 0.001, and the mean of the degraded state was taken to be a factor of ten worse than the overall mean. The value of the mixing parameter, which is the prior probability of being in the degraded state, was taken to be 0.01. This allows us to calculate the mean of the good state, which is 9.1×10^{-4}. To determine the shape of the beta distribution describing the uncertainty in p for each state, (7) used a CNI prior for the good state, and chose a beta distribution with first parameter equal to 1.5 for the degraded state, allowing the second parameter to be calculated from the known mean of 0.01. The value of 1.5 ensures that the mode of the degraded state will be much greater than mean of the good state. The parameter values of the mixture prior for the example problem are $\pi = 0.01$, $\alpha_0 = 0.498$, $\beta_0 = 547.3$, $\alpha_1 = 1.5$, $\beta_1 = 148.5$.

4. NONCONJUGATE DIFFUSE PRIORS

Both the CNI prior and the mixture of conjugate priors discussed above are mathematically convenient in that no numerical integration is required to obtain the posterior distribution, which can be written down in closed form. Bayesian inference with these priors can be performed in a spreadsheet, with the only complication being the distribution percentiles, which must be calculated numerically. But even this is no longer a problem, as the required inverse cumulative distribution functions are included in common spreadsheets. However, diffuse conjugate priors, despite their mathematical convenience, have relatively light tails, which can influence the posterior mean of an update with sparse data. In this section we examine two nonconjugate priors with heavier tails. Bayesian inference must be done numerically with these priors, but this is no longer a limitation given the availability of modern numerical tools, implemented in open-source software.

4.1 Logistic-Normal Prior

Historically, the lognormal distribution has been used widely in PRA to represent parameter uncertainty. The error factor of the lognormal distribution, defined commonly as the ratio of the 95th percentile to the median (50th percentile), measures how diffuse the distribution is, with an error factor of 10 being indicative of a distribution that is quite diffuse (factor of 100 between 5th and 95th percentiles). The range of the lognormal distribution is unbounded, however, making it sometimes a problematic choice for representing uncertainty in a binomial parameter p, since p, being a probability, must lie between 0 and 1. For modeling a probability, then, it is useful to consider the *logistic-normal* distribution, which has features similar to the lognormal distribution (e.g., heavy tail), but is constrained to range from 0 to 1, as desired for a probability. It was suggested by (9) as an alternative to the beta conjugate prior for p.

The logarithm of a lognormally distributed variable has a normal distribution, hence the name. Correspondingly, in the case of the logistic-normal distribution, the analogous relationship is that $\log[p/(1-p)]$ has a normal distribution with mean μ and variance σ^2. The logistic-normal density function is given by

$$g(p) = \frac{1}{\sqrt{2\pi}\sigma p(1-p)} \exp\left\{ -\frac{\left[\log\left(\frac{p}{1-p}\right) - \mu\right]^2}{2\sigma^2} \right\} \quad -\infty < \mu < \infty, \sigma > 0 \qquad (3)$$

The median is given by $exp(\mu)/[1 + exp(\mu)]$, and the 95th percentile by $exp(\mu + 1.645\sigma)/[1 + exp((\mu + 1.645\sigma)]$. Unfortunately, the mean and variance of p cannot be written down in closed form and must be found via numerical integration.

For the present comparison, we chose a logistic-normal distribution that has the same mean and 95th percentile as the CNI prior.[4] By construction, this choice preserves the mean value, which is often taken as a point estimate of an overall industry average value.[5] The 95th percentile is a commonly used upper percentile value in PRA. Thus, both the CNI prior and the nonconjugate prior will have 5% area above the 95th percentile, but the distribution of this 5% will be different. These equalities give two equations in two unknowns for the logistic-normal parameters μ and σ, which can be solved numerically. The solution of these equations and the Bayesian update of the logistic-normal prior were performed using the R package (10). For our example problem, with a mean of 0.001 and 95th percentile of 0.0031, we find $\mu = -7.7$ and $\sigma = 1.3$.

4.2 Robust (Logistic-Cauchy) Prior

The robust prior is a nonconjugate distribution placed on $\theta = \log[p/(1-p)]$.[6] In that sense it is similar to the logistic-normal distribution, which placed a normal(μ, σ^2) distribution on θ. However, the robust prior places a distribution with heavier tails on θ, namely a Cauchy distribution, which is equivalent to a Student-t distribution with 1 degree of freedom. The Cauchy density function (on the θ scale) is given by

---

[4] For the Poisson aleatory model, the analogous choice would be a lognormal prior with mean and 95th percentile set equal to those of the gamma CNI prior.

[5] Except for the Jeffreys prior, which has a mean of 0.5 in the case of the binomial distribution, all of the prior distributions considered in this paper have the same mean value, but differ in their tail behavior and correspondingly in the responsiveness of the posterior mean to new data.

[6] For the Poisson aleatory model, the Cauchy prior would be placed on $\theta = \log\lambda$.

$$g(\theta) = \frac{1}{\pi\sigma\left[1+\left(\frac{\theta-\mu}{\sigma}\right)^2\right]} \quad -\infty < \mu < \infty, \sigma > 0 \tag{4}$$

The Cauchy distribution is symmetric about μ, like the normal distribution, but unlike the normal distribution, the mean, variance, and higher moments do not exist; σ is a scale parameter, but it is not the standard deviation, which does not exist. The Cauchy cumulative distribution function, which can be used to obtain percentiles, is given by

$$G(\theta) = \frac{1}{\pi}\tan^{-1}\left(\frac{\theta-\mu}{\sigma}\right) + \frac{1}{2} \tag{5}$$

We follow the approach of (11) in finding the Cauchy parameters, starting with the parameters of the beta distribution used to approximate the CNI prior for *p*. If α and β are the parameters of the beta distribution used to approximate the CNI prior, then the Cauchy location and scale parameters will be given by[7]

$$\begin{aligned} \mu &= \psi(\alpha) - \psi(\beta) \\ \sigma &= \sqrt{\psi'(\alpha) + \psi'(\beta)} \end{aligned} \tag{6}$$

In this equation, ψ(•) is the digamma function and ψ'(•) is the trigamma function, the first and second derivatives of the logarithm of the gamma function. For our example problem, with α = 0.497 and β = 621, we find μ = -8.2 and σ = 2.2.

The robust logistic-Cauchy prior acts somewhat like a switch in Bayesian updating. If the data are consonant with the prior, then the prior influences the analysis (albeit weakly). However, when the prior and data are in conflict, the robust prior acts more like a uniform distribution, allowing the data to drive the result. For the mathematical version of this description, see (11)

5 SUMMARY

Figure 2 compares the posterior means from updating the diffuse priors we have examined with binomially distributed failure data. This figure shows the "switching" behavior of the robust logistic-Cauchy prior clearly. For 0 or 1 failures, the data are consistent with the prior, and the prior moderates the influence of the data on the posterior mean. However, for 2 or more failures (in 50 demands), the behavior switches, and the posterior mean is driven by the observed data, and the slope of the graph is about equal to that obtained from updating the Jeffreys prior, which allows the data to dominate the posterior mean.

For 3 or fewer failures in 50 demands, the mixture prior and logistic-normal prior produce similar posterior means. For more than 3 failures, the logistic-normal prior allows the data to dominate the posterior mean ("switching," if you will, from the prior to the data, much like the logistic-Cauchy prior), while the mixture prior approaches the result one would obtain from updating the degraded-state prior, which in this case is a beta(1.5, 148.5) distribution. With 10 failures in 50 demands, the posterior mean from updating this distribution would be 0.0575, the same value to 4 significant figures obtained with the mixture prior with this distribution representing the degraded state. Thus, there is a switching behavior in the case of the mixture prior, with the switch being data-driven, but the transition is from the normal to the degraded state.

---

[7] For the Poisson aleatory model, the Cauchy parameters will be given by $\mu = \psi(\alpha) - \log\beta$ and $\sigma = \sqrt{\psi'(\alpha)}$.

Figure 2 clearly shows the relatively strong and persistent influence of the CNI prior on the posterior mean, even as the number of failures grows quite large. Even if the failure fraction reached 50% (25 failures), the posterior mean from updating the CNI prior would only be 0.05, a factor of 10 less than the update of the Jeffreys prior. The reason for this behavior is the relatively light tail of the CNI prior, which places vanishingly small probability on very large values of p; the 99.9th percentile is only about 0.01. In contrast, the logistic-normal prior has a 99.9th percentile of about 0.03, and the 99.9th percentile with the robust logistic-Cauchy prior is effectively at 1.0, the same as the Jeffreys prior.

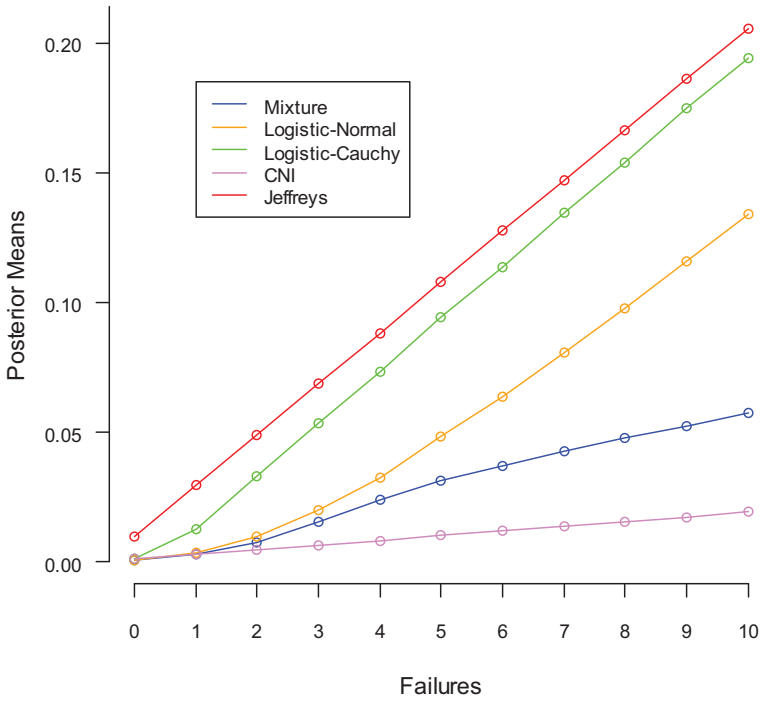

Figure 2 Posterior means from updating various diffuse priors with binomial failure data (out of 50 demands)

Acknowledgements

This paper was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

This manuscript has been authored by Battelle Energy Alliance, LLC under Contract No. DE-AC07-05ID14517 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

1. Kass, Robert E. and Wasserman, Larry, "The Selection of Prior Distributions by Formal Rules," September 1996, *Journal of the American Statistical Association*, Vol. 91, pp. 1343-1370.
2. Atwood, Corwin L, "Constrained Noninformative Priors in Risk Assessment," 1996, *Reliability Engineering and System Safety*, Vol. 53, pp. 37-46.
3. Jaynes, E. T., "Prior Probabilities," 1968, *IEEE Transactions on Systems Science and Cybernetics*.
4. Dube, D. A., et al. *Independent Verification of the Mitigating Systems Performance Index (MSPI) Results for the Pilot Plants.* U. S. Nuclear Regulatory Commission. Washington, D.C. 2005, NUREG-1816.
5. Berger, James O., *Statistical Decision Theory and Bayesian Analysis, Second Edition.* Springer, 1985.
6. Siu, Nathan O. and Kelly, Dana L., "Bayesian Parameter Estimation in Probabilistic Risk Assessment," 1998, *Reliability Engineering and System Safety*, pp. 89-116.
7. Youngblood, Robert W. and Atwood, Corwin L., "Mixture Prior for Bayesian Performance Monitoring 1: Fixed-Constituent Model," 2005, *Reliability Enginnering and System Safety*, Vol. 89, pp. 151-163.
8. Atwood, Corwin L. and Youngblood, Robert W., "Mixture Priors for Bayesian Performance Monitoring 2: Variable-Constituent Model," 2005, *Reliability Engineering and System Safety*, Vol. 89, pp. 164-176.
9. Atwood, C., et al,. *Handbook of Parameter Estimation for Probabilistic Risk Assessment.* U. S. Nuclear Regulatory Commission, 2003. NUREG/CR-6823.
10. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria : R Foundation for Statistical Computing, 2009. ISBN 3-900051-07-0.
11. Jairo, A., et al. *A Case for Robust Bayesian Priors with Applications to Binary Clinical Trials.* Department of Biostatistics, MD Anderson Cancer Center. 2008. Working Paper. 44.