Title: Imprecise Probability Methods for Weapons UQ

Author(s): Picard, Richard Roy
Vander Wiel, Scott Alan

Intended for: Report

Issued: 2016-05-13

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Imprecise Probability Methods for Weapons UQ

Rick Picard and Scott Vander Wiel

Abstract

Building on recent work in uncertainty quantification, we examine the use of imprecise probability methods to better characterize expert knowledge and to improve on misleading aspects of Bayesian analysis with informative prior distributions. Quantitative approaches to incorporate uncertainties in weapons certification are subject to rigorous external peer review, and in this regard, certain imprecise probability methods are well established in the literature and attractive. These methods are illustrated using experimental data from LANL detonator impact testing.

1 Introduction

Uncertainty quantification for nuclear weapons work has received much attention over the years, e.g., from external review panels such as the JASONs (Eardley et. al 2004) and NAS (National Academy of Sciences 2009). More recently, shortcomings of Bayesian methods for UQ have been highlighted (Nakhleh, Webster, and Haynes 2015).

In this report, we examine the potential for imprecise probability ("IP") methodology to rectify those shortcomings. IP ideas have their origins in economics (Keynes 1921) and have been refined over the years. Walley's (1991) textbook is a classic reference. We avoid overly broad notions of IPs, such as in the engineering literature (see, e.g., the review article by Beer, Ferson, and Kreinovich 2013, and its 268 references). Instead, we focus on methods that are mature enough to pass muster with external review panels as above. These IP methods have much to offer weapons UQ, while at the same time introducing new challenges in their implementation.

In what follows, experimental data from detonator impact testing is used to illustrate IP methods. This testing shares many qualities of weapons UQ, i.e., performance requirements exist, the formal incorporation of knowledge from subject matter experts is beneficial, there are multiple computational models that give similar results over the observed range of data, and extrapolation to conditions not previously tested is of great interest.

As shown in the examples to follow, Bayesian analyses often produce misleading results in such cases. The precise prior and posterior quantities intrinsic to Bayes theory generally do not allow for accurate quantification of uncertainties in the extrapolated metrics relevant to weapons performance requirements. When multiple physics models exist, the problem is further magnified.

This presentation attempts to overcome the sometimes esoteric nature of IPs and, other than requiring a minimal background in Bayesian methods, is aimed at a general UQ audience.

2 Drop Test Example

The data set consists of 25 detonator impact tests conducted at Los Alamos by LANL Group WX-7. From various heights, a 2.5 kg anvil was dropped on detonators of a specific type. Based on review of the video and/or on the decibel level from the audio, a "go" or "no-go" response was determined for each drop test. Data are given in Table 1.

Table 1. Detonator Impact Testing Data.

Index	Height (cm)	# Tested	# Go	# No-Go
i	h_i	n_i	k_i	$n_i - k_i$
1	45.0	1	0	1
2	50.5	3	1	2
3	57.0	3	2	1
4	64.0	6	2	4
5	71.5	7	5	2
6	80.5	4	3	1
7	90.0	1	1	0

The probability of a "go" response increases monotonically from zero (at height $h = 0$) to one (when the height is sufficiently great). Of interest is the probability of a "go" response as a function of drop height. Several statistical models have been used for such data, and we initially focus on the standard probit model. This model postulates that the probability of a “go” response at drop height h is

$$\Pr(\text{“go”}) \;=\; \Phi\left(\frac{\log h \,-\, \log h_{50}}{\sigma}\right) , \qquad (1)$$

where h_{50} is defined as the drop height whose chance of a “go” is 50%, σ is a scale factor, and $\Phi(\cdot)$ is the cumulative distribution function for a standard Gaussian distribution.

3 Performance Requirements

Performance requirements for detonators are contained in DOE Order O-452.1 (Department of Energy 2015), one of whose goals is to prevent accidents during weapons assembly and disassembly. Among the requirements is that

> “. . . the probability of a premature nuclear explosive detonation must not exceed one in a million (1E-06) per credible nuclear weapon accident or exposure to abnormal environments” (DOE Order O-452.1, p. 7).

The order is intentionally vague on how all possible “credible accidents” or “abnormal environments” can be enumerated. Nonetheless, several accident scenarios are relevant to detonator impact testing (e.g., a worker inadvertently dropping a wrench on a detonator).

Note that the 10^{-6} tail probability in the above performance requirement is not unique to detonators. Nuclear weapons requirements for lifetime premature detonation under normal conditions and for one-point safety, for example (DOE Order O-452.1, p. 7 and p. 13, respectively), also involve extreme tail probabilities.

Performance requirements thus involve two forms of extrapolation. Illustrating with the detonator requirement, the first form extrapolates the physical insult in the drop test to a physical insult of interest − i.e., extrapolating from one situation (the drop of a 2.5 kg anvil on a detonator from a certain height) to another (the drop of a wrench from a different height). Discussion of this form of extrapolation is beyond the scope of this report; for what follows, we assume that an anvil drop from 5 cm is equivalent to dropping a much lighter tool from a height of interest.

The second form of extrapolation is statistical, and involves extrapolating results from the drop heights 45-90 cm in Table 1 to drop heights such as 5 cm, where the probability of a "go" is low. This extrapolation is necessary because a very large number of drop tests would have to be conducted at low drop heights in order to estimate Pr("go") accurately. Because the time/money involved in such an effort would be prohibitive, properties for low-probability drop heights are extrapolated using the predictive model.

Formalizing this approach, the probit model (1) can be inverted to give drop height as a function of the probability of a "go" response,

$$h_{\mathrm{Pr(go)}} \;=\; h_{50} \;\times\; \exp\Big[\, \sigma\, \Phi^{-1}\,(\mathrm{Pr}(\text{“go”}))\,\Big]\;. \tag{2}$$

For Pr("go") = 10^{-6}, the Gaussian quantile $\Phi^{-1}(10^{-6}) = -4.75$, and the 10^{-6} drop height $h_{-6} = h_{50} \times \exp[\,-4.75\,\sigma\,]$ is a known function of the model parameters h_{50} and σ. This functional relation is central to performance assessment.

4 Subject Matter Expertise for the Experiment

The LANL scientist who was to conduct the experiment provided a prior estimate for the 50-50 drop height h_{50} based on work with detonators similar to the type examined here. That prior estimate was 70 cm. His relative uncertainty factor on the estimate was 1.2.

Equivalently, the plus-or-minus one standard deviation interval in log scale has a standard deviation of log(1.2): $h_{50} \in (70/1.2,\ 70\ \times 1.2) \ \Leftrightarrow\ \log h_{50} \in \log 70 \pm [\log 1.2]$.

Expert opinion for the scale factor σ in (1) was elicited through other drop heights besides h_{50}. Subject matter experts are more comfortable contemplating physical quantities like drop heights rather than abstract parameters like σ in a statistical model. Further, a single elicitation on the same physical quantities can be used in conjunction with other models besides the probit (more on this to come). Input on the 10% drop height h_{10} was obtained via the h_{50}/h_{10} ratio, which is directly related to the scale factor σ. The prior estimate for the h_{50}/h_{10} ratio was 2, with a relative uncertainty factor of 1.5.

The values $h_{50} \approx 70$ cm and $h_{50}/h_{10} \approx 2$, together with their uncertainty factors 1.2 and 1.5, are nice round numbers summarizing the expert's best guesses. *It is not the case that "correct" prior values exist*, precise to 100 decimal places of accuracy, or that the expert could arrive at such exact quantities if only he thought long enough about the subject.

5 The Likelihood Function for the Drop Test Data

The first step in a Bayesian or IP data analysis involves defining the likelihood function. Once this has been done, expert opinion can then be quantified into one (for a Bayesian analysis) or many (for an IP analysis) prior distributions on its parameters.

The likelihood function here consists of a binomial probability density function combined with the probit model (1). There were 7 drop heights in the experiment. In the notation of Table 1, the i-th drop height h_i, for $i = 1, \ldots, 7$, involved n_i drops and resulted in k_i "go" responses.

The likelihood function for the data is then

$$\ell(\boldsymbol{n}, \boldsymbol{k} \mid h_{50}, \sigma\,) \;=\; \prod_{i=1}^{7} \frac{n_i!}{k_i!\,(n_i - k_i)!}\;\left(\mathrm{Pr}(\text{ ``go'' })\right)^{k_i}\;\left(1 - \mathrm{Pr}(\text{ ``go'' })\right)^{n_i - k_i}$$

$$= \prod_{i=1}^{7} \frac{n_i!}{k_i!\,(n_i - k_i)!}\;\left(\Phi\left(\frac{\log\,h_i\,-\,\log\,h_{50}}{\sigma}\right)\right)^{k_i}\;\left(1 - \Phi\left(\frac{\log\,h_i\,-\,\log\,h_{50}}{\sigma}\right)\right)^{n_i - k_i}$$

A contour plot of the likelihood function for the experimental data, expressed in terms of the elicited quantities h_{50} and h_{50}/h_{10}, is given in Figure 1.

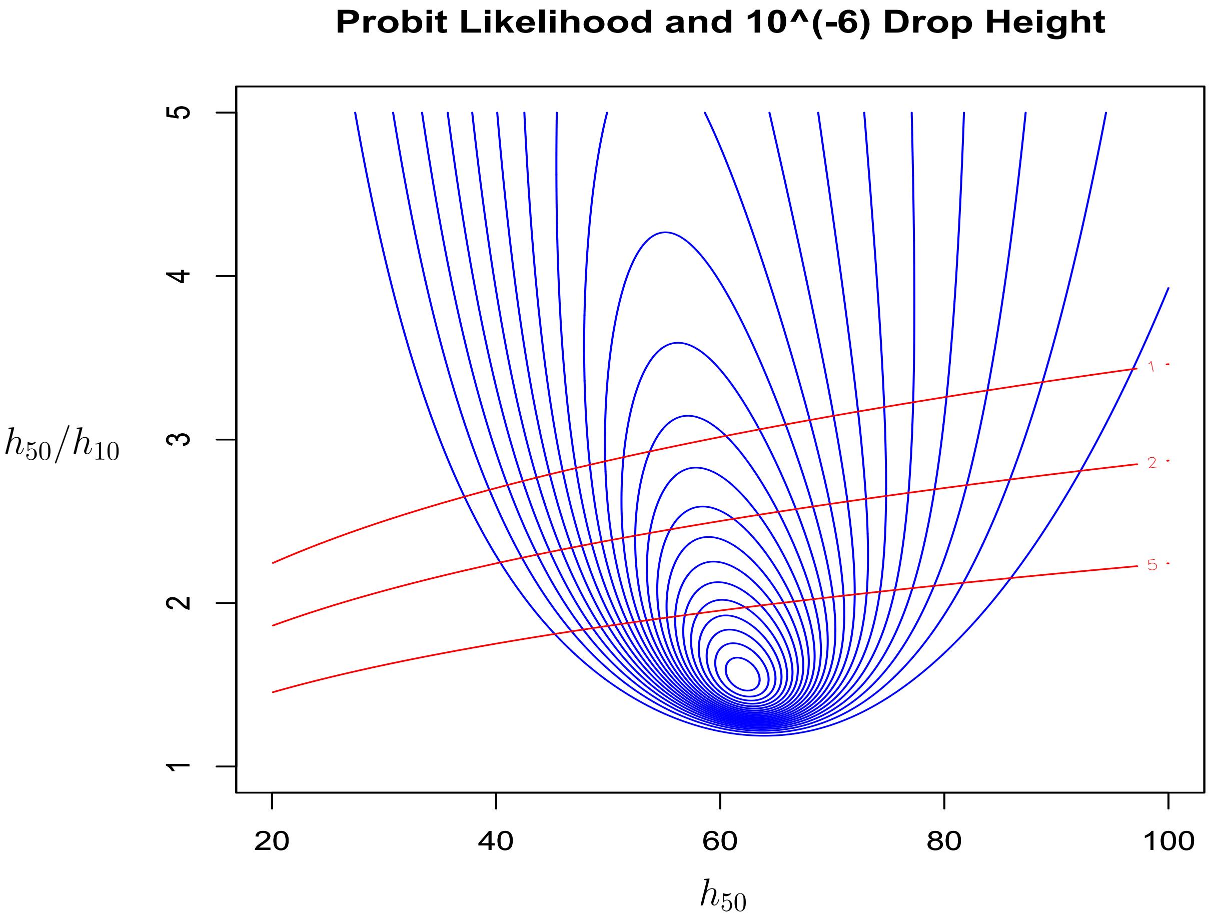

Figure 1: Likelihood (blue) and 10^{-6} Drop Height h_{-6} (red) Contours.

The likelihood function is maximized at roughly the centroid of the smallest blue contour in Figure 1. Maximum likelihood estimates, based on the data alone, are $h_{50} = 62$ and $h_{50}/h_{10} = 1.55$. As expected, these data-based estimates are not far from the expert's prior estimates $h_{50} \approx 70$ and $h_{50}/h_{10} \approx 2$ when judged by the expert's stated uncertainties.

Also superimposed on Figure 1 are contours for the 10^{-6} drop height, as related to the performance requirement in Section 3. Recall that this drop height is

$$h_{-6} \;=\; h_{50} \;\times\; exp\,[\,-4.75\,\sigma\,]\;\;.$$

Values $(h_{50}, h_{50}/h_{10})$ corresponding to the 10^{-6} drop height are displayed in the plot. The red curve labelled '5' denotes the subset of parameter space where $h_{-6} = 5$ cm. For reference, contours for $h_{-6} = 1$ cm and $h_{-6} = 2$ cm are also included.

Ideally, there would be high confidence that the actual $(h_{50}, h_{50}/h_{10})$ lies below the red contour for 5 cm. If so, there is high confidence that the performance requirement is met.

6 The Conventional Bayesian Paradigm

The most widely used approach to combine subject matter expert knowledge with experimental data is the Bayesian paradigm, which force-fits the expert's knowledge into a single, specific ("precise") prior distribution. The force-fitting does not accurately capture expert opinion: per Section 4, actual opinion about h_{50} and h_{50}/h_{10} isn't nearly so precise. Nonetheless, a mandatory compliance edict of Bayesian doctrine demands such a force-fitting, and decrees a precise bivariate prior for $(h_{50}, h_{50}/h_{10})$. That prior distribution is then used to obtain posterior results via Markov chain Monte Carlo (MCMC) simulation.

The performance requirement here is that 5 cm is a safe drop height for a certain accident scenario, where the word “safe” means that there is no more than a 10^{-6} chance of a “go” response for an anvil drop of 5 cm. In IP-like notation, let the set

$$A = \{h_{50}, h_{50}/h_{10} \mid h_{-6} > 5 \text{ cm } \}$$

be the region of parameter space where the 10^{-6} drop height exceeds 5 cm. In Figure 1, the set A consists of the region below the red contour $h_{-6} = 5$ cm. The goal is to estimate the probability $\Pr(\ (h_{50}, h_{50}/h_{10}) \in A\ )$ that the performance requirement is met.

The Bayes approach begins with a prior distribution on $(h_{50}, h_{50}/h_{10})$ guided by expert input. Log Gaussian priors for the 50-50 drop height $h_{50} > 0$ are typically used, here

$$\log h_{50} \ \sim\ N(\log 70, [\log 1.2]^2\,)\ ,$$

where the value 70 is the expert’s prior estimate for the 50-50 drop height, and the variance $[\log 1.2]^2$ follows from the relative uncertainty factor 1.2.

The prior distribution for the scale factor σ is derived from the expert input on the h_{50}/h_{10} ratio. Using the drop height relation for the 10% drop height h_{10},

$$h_{10} \ = \ h_{50} \ \times \ \exp[\,\sigma\,\Phi^{-1}(0.10)\,] \ = \ h_{50} \ \times \ \exp[\,-1.28\,\sigma\,]\ ,$$

it follows from this equation that

$$\sigma \ = \ \frac{1}{1.28} \log\,(h_{50}\,/\,h_{10})\ .$$

To make physical sense, it is also necessary that $h_{50}/h_{10} > 1$, or that $h_{10}/h_{50} \in (0, 1)$.

For quantities within (0, 1), the most common Bayesian prior is the beta distribution. By varying its parameters, the beta density function can take on a wide variety of shapes (uniform, linearly increasing or decreasing, unimodal, symmetric, skewed, U-shaped, etc.)

as warranted by the situation. In IP applications, the imprecise beta distribution (e.g., Walley 1996; Walley, Gurrin, and Burton 1996) is commonly used.

The beta distribution has two parameters, denoted α and β. The ratio α/β determines the mean of the distribution through the relation $\alpha/\beta = \text{mean}/(1 - \text{mean})$. Magnitudes of α and β determine the standard deviation.

The prior estimate $h_{50} \,/\, h_{10} \approx 2$ corresponds to a beta distribution with mean value $E[\, h_{10} \,/\, h_{50} \,] \approx 1/2$, implying that $\alpha \approx \beta$. An uncertainty factor 1.5 implies the interval

$$h_{50}/h_{10} \;\in\; (\; 2/1.5,\; 2 \times 1.5 \;) \Leftrightarrow\; h_{10}/h_{50} \;\in\; \left(\frac{1}{2 \times 1.5},\; \frac{1.5}{2}\right) \;=\; (1/3, 3/4)\;.$$

Without using more sophisticated elicitation methods (Yu, Shih, and Moore 2008), the half width of this interval, (3/4 - 1/3) / 2 ≈ 0.21, is equated to one standard deviation for the beta prior. Adding the constraint $\alpha \approx \beta$ gives the parameter values $\alpha \approx \beta \approx 2.4$, or

$$h_{10}/h_{50} \;\sim\; \text{Be}\;(2.4, 2.4)\;.$$

Independently coupling this beta prior for h_{10}/h_{50} with the log Gaussian prior for h_{50} produces the "nominal" prior distribution. Combining with the probit likelihood function and data in Table 1, Bayes analysis simulates an MCMC sample $\{(h_{50}, \sigma)_j \;;\; j = 1, \ldots, N\}$ from the posterior. For each element of the sample, its corresponding 10^{-6} drop height $(h_{-6})_j = (h_{50})_j \;\times\; exp[\, -4.75\, \sigma_j]$, and the set $\{(h_{-6})_j\}$ is used for probability assessment. An MCMC simulation (10^6 samples) gives $\text{Pr}_{\text{posterior}}(A) \;=\; 0.45216$.

The use of five decimal places emphasizes that the posterior probability could indeed be determined to arbitrary accuracy by running the MCMC simulation until eternity. Quoting MCMC sampling error as if it were the sole source of uncertainty in the estimate, as is often done, only reinforces a false sense of precision. Such precision is misleadingly illusory, as shown shortly. The illusion is, however, an immediate consequence of the Bayesian paradigm, which requires precise prior distributions that lead to precise posterior distributions, and then to precise quantities such as $\Pr_{\text{posterior}}(A)$.

Pretentious accuracy aside, the posterior probability 0.45216 is too small to provide high confidence in meeting the performance requirement. The degree of confidence is modest for three reasons. First, uncertainties in the prior distribution are too large to allow for high confidence based on prior information alone. Next, the data set is limited, consisting of only 25 go/no-go tests. And finally, the extrapolation is considerable, from the 45-90 cm range of drop heights in the experimental data to a 5 cm drop height and 10^{-6} tail probability of interest $-$ and this level of extrapolation propagates to greater uncertainty.

Were additional data obtained, confidence would improve (assuming, of course, that the additional data were consistent with safe operation). Because the design of additional detonator impact tests goes beyond the scope of this report, it is not pursued further.

7 Imprecise Probabilities

Bayes methods are open to the severe criticism that subjective prior information is not well characterized by the force-fitting process required to produce a precise prior distribution. For weapons UQ (Nakhleh, Webster, and Haynes 2015, p. 6), "in many, or most, cases of practical interest, the information is insufficient to constrain the analyst to a single (prior)." For the detonator data in particular, many prior distributions are consistent with expert opinion, and each such prior gives a different posterior probability $\Pr_{\text{posterior}}(A)$.

The first task of IP methods is to formally identify the set $\mathcal{P}$ of prior distributions consistent with the expert's beliefs. Once done, the plausible range of $\Pr_{\text{posterior}}(A)$ values is then determined. In economics (e.g., Weatherson 2002), the set $\mathcal{P}$ is sometimes called the representor. The maximum value of $\Pr_{\text{posterior}}(A)$ over the prior distributions in $\mathcal{P}$ is called the upper probability of $\Pr_{\text{posterior}}(A)$ and is denoted $\overline{\Pr}(A)$. Similarly, the minimum value of $\Pr_{\text{posterior}}(A)$ over $\mathcal{P}$ is the lower probability of $\Pr_{\text{posterior}}(A)$ and is denoted $\underline{\Pr}(A)$.

By computing the upper and lower probabilities $\overline{\Pr}(A)$ and $\underline{\Pr}(A)$, conclusions are reached that better reflect the beliefs of the subject matter expert(s). The difference

$$\Delta(A) \quad = \quad \overline{\Pr}(A) - \underline{\Pr}(A)$$

is called the *imprecision* of the event A.

IP methodology is often (e.g., Walley 1991, p. 44) attributed to the economist John Maynard Keynes and subsequent work by others. In the early 1900s, before the advent of R. A. Fisher and modern frequentist statistics, Keynes (1921) espoused a number of ideas related to imprecise probabilities. The subject has since received considerable attention, e.g., Walley's (1991) textbook, and much other related literature.

The attractiveness of IPs stems from their (far) more accurate characterization of subjective prior information than is incorporated in standard Bayesian methods. Despite this attractiveness, IPs have not been widely used because

1) in several situations, bottom-line IP conclusions do not differ much from those of a conventional Bayesian analysis,

2) the IP paradigm postulates an already-given set $\mathcal{P}$ of priors consistent with expert opinion, when in fact considerable effort is required to construct the set $\mathcal{P}$, and

3) in realistic applications, there are computational challenges in carrying out the maximization and minimization needed to determine upper and lower probabilities.

These issues are discussed in turn.

Several situations exist where Bayes analyses are robust against specification of the prior, and imprecisions $\Delta(A)$ are small enough that IPs are unnecessary. The most common such case is where all plausible priors are noninformative (i.e., the subject matter experts simply aren't very expert). At the other end of the spectrum, where experts are so certain of parameter values that it is not cost-effective to obtain additional data, imprecisions are also small. Thirdly, in cases where there is an overabundance of relevant data, and the effect of the prior distribution is swamped by the data in a Bayesian updating, imprecisions are again minimal. None of these situations apply to weapons UQ, and IPs are attractive.

The second issue concerns construction of the set $\mathcal{P}$ of plausible priors. Similar to most presentations of Bayesian analysis, where the precise prior distribution is treated as already given, the same is true for the set $\mathcal{P}$ in an IP analysis. Unfortunately, there is not a natural black-and-white distinction between priors that are "consistent with" expert opinion versus priors that aren't, and it can be problematic where to draw the line between the two.

The third issue concerns computation. Recall the nature of Bayesian statistics before the advent of MCMC: pre-1990s computational tools did not exist for Bayesians to solve many real problems, which combined with other factors to greatly limit the practicality of Bayes/IP methods (Efron 1986). In today's Bayes analyses, it is common to simulate MCMC samples from the posterior, as in the previous section. For weapons UQ, it is not computationally feasible to embed lengthy MCMC simulations within an optimization routine to maximize/minimize probabilities $\Pr_{\text{posterior}}(A)$ over $\mathcal{P}$. We describe computational tricks to avoid lengthy MCMC simulations later, so that IP computing is often practical.

To be fair, there has been grudging acknowledgement within parts of the Bayes community that force-fitting expert opinion into a single precise prior distribution is indeed misleading. One response to this acknowledgement is to impose hierarchical prior distributions on prior parameters. Here, an informative hyperprior density $p_{\phi}(\boldsymbol{\phi})$ on the prior parameters $\boldsymbol{\phi}$ is elicited directly or indirectly (see Oakley and O'Hagan 2007 for a novel example of using elicited physical quantities to derive such a hyperprior).

An advantage of hierarchical models is that they can provide insights to the data, e.g., by examining the distribution of $\Pr_{\text{posterior}}(A \mid \boldsymbol{\phi})$ induced by $\boldsymbol{\phi} \sim p_{\phi}(\boldsymbol{\phi})$ and extracting a credible interval. Although this approach is conceptually straightforward, it quickly runs into practical difficulties. Each value $\boldsymbol{\phi}$ indexes a precise prior, so that a $\Pr_{\text{posterior}}(A \mid \boldsymbol{\phi})$ evaluation requires a separate MCMC run for each $\boldsymbol{\phi}$. Depending on the probability content of the desired credible interval, hundreds/thousands of such MCMC runs are required to obtain reasonable estimates of interval endpoints. And when the set A corresponds to a rare event that must be simulated directly (see Section 10), each individual evaluation entails considerable computation, rendering the approach computationally infeasible for many realistic cases. The appeal of hierarchical modeling is further limited by the fact that experts who don't believe in precise priors also won't believe in precise hyperpriors.

Another response to the problems of force-fitting a single precise prior is Bayesian sensitivity analysis, sometimes called robust Bayesian analysis (e.g., Berger 1990, 1994). Typical Bayesian sensitivity analysis entails informally considering a small handful of alternate prior distributions and assessing their corresponding posteriors. Such informality is nothing more than an ad hoc, oversimplified version of IPs. Walley (1991, p. 107) argues that “most of the theory presented in this book can be regarded as a formalization of Bayesian sensitivity analysis, although we would not advocate that interpretation.”

Ad hockery aside, the main issue with Bayesian sensitivity analysis is that few Bayesians actually implement it. One reason is that they “tend to be aggressive and optimistic with their modeling assumptions” (Efron 2005, p. 1), and are thus less likely to closely scrutinize those assumptions. Further, sensitivity analysis pushes some Bayesians out of their comfort zone. It is philosophically incoherent for a Bayesian purist to assign two different prior probabilities to the same event, as only one such value defines his fair bet on the event. The approach also has potential for non-Bayesian updating of the prior (i.e., updating by other than a myopic application of Bayes theorem), especially when the data warrant non-Bayesian updating (e.g., Dawid 1982). Pragmatic complications involving IPs also matter: more work is required for IPs than for a nominal Bayes turn-the-crank approach, data analysts may be unfamiliar with cases where sensitivity analysis is most needed, or may be unfamiliar with the elicitation/analytic methods to carry it out, etc.

Whatever the reason(s), and despite a modest literature on sensitivity analysis as well as Bayesian versions of model goodness-of-fit diagnostics and outlier detection, technical reports on Bayesian UQ for weapons applications generally do not involve sensitivity analysis (for the record, an exception being Vander Wiel and Gore 2015).

8 IPs for the Drop Test Example

The primary challenge in implementing IP methods is to define the set $\mathcal{P}$ of prior distributions that are consistent with expert opinion. This process is inherently subjective, inherently arbitrary, and there is no magic recipe for it. One common approach to constructing $\mathcal{P}$ is parameter-centric, and another is distribution-centric; either approach can begin (or not) with a nominal prior distribution.

Illustrating with the nominal Gaussian prior distribution, $\log h_{50} \sim N(\log 70, [\log 1.2]^2)$, a parameter-centric approach identifies parameters (m, s) other than $(m_0, s_0) = (70, 1.2)$ that are consistent with prior knowledge. A distribution-centric approach identifies distributions other than $N(\log m_0, [\log s_0]^2)$, such as $N(\log m, [\log s]^2)$, consistent with that knowledge. These two approaches can give different results.

Subject matter experts tend to be more comfortable with parameter-centric plausible regions. The expert's frame of reference is the original elicitation – of $(m, s) \approx (70, 1.2)$ – and a parameter-centric region extends that line of thinking. These regions can be simple, and can sometimes be constructed sequentially.

For the drop test data, suppose the 20% relative error is believed with high confidence to be accurate to within a factor of two, from 10% to 40%, so that uncertainty factors between 1.1 and 1.4 are consistent with prior beliefs. Upon mimicking a 95% confidence interval for approximately Gaussian distributions, a maximal-width interval for m is

$$\log m \in \log 70 \ \pm\ 2\ \times [\log 1.4]\ .$$

One definition of parameters (m, s) consistent with expert input is

a) $s \in [1.1, 1.4]$ is within the range of factors consistent with prior beliefs, and

b) all conditional-on-s intervals $\log\ m\ \pm\ 2 \times [\log\ s]$ lie entirely within the maximal interval $\log 70\ \pm\ 2\ \times [\log 1.4]$.

Figure 2 displays a gridded version of this region of parameter space.

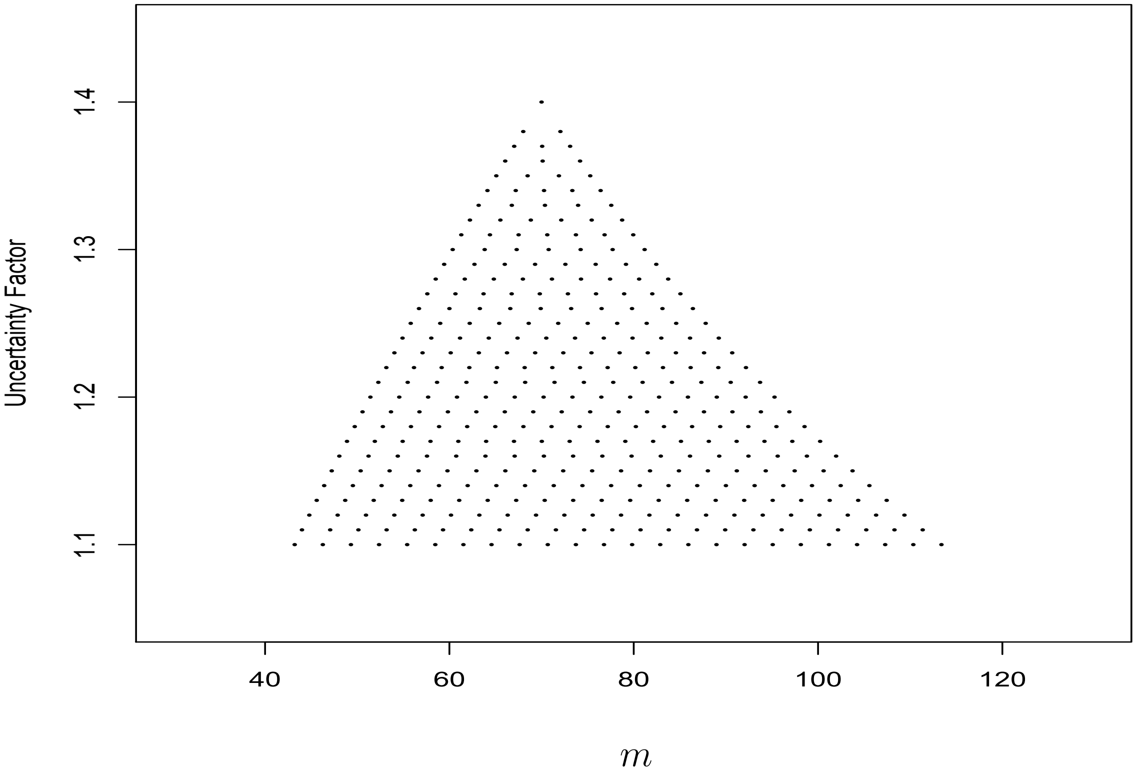

Figure 2: Parameter-Centric Region for (m, s).

As always the case with IP formulations, a prior distribution consistent with expert input can be virtually indistinguishable from a prior distribution inconsistent with expert input, i.e., as when two priors straddle the boundary of $\mathcal{P}$ in Figure 2. Such a precise boundary between plausible/implausible priors is clearly unrealistic, although IP advocates counter that, while imperfect, it is a huge improvement on the Bayesian purist's single-point representor set $\mathcal{P}$. The fiction of a precise boundary for $\mathcal{P}$ is an essential component of an IP analysis because it is needed for the maximization/minimization required to obtain the upper and lower probabilities $\overline{\Pr}(A)$ and $\underline{\Pr}(A)$.

A different approach to constructing $\mathcal{P}$ follows from the nominal prior distribution $N(\log m_0, [\log s_0]^2)$. Recall that $\mathcal{P}$ consists of *distributions* consistent with expert input −

the input here being idealized by the nominal Gaussian prior. As such, distribution-centric plausible regions conform more faithfully to the IP protocol than do parameter-centric regions, even though the latter come more naturally for subject matter experts.

Several options exist for constructing distribution-centric regions. Formal distance measures, such as Kolmogorov-Smirnov or Kullback-Leibler, quantify "how far apart" two probability distributions are. Such a measure could define the prior distributions consistent with expert opinion, e.g., $\mathcal{P}$ could consist of all distributions (Gaussian or otherwise) within a prescribed distance of $N(\log m_0, [\log s_0]^2)$. Similarly, ϵ-contaminated priors, whose probability density functions are equal to a weighted sum of the nominal prior density (weighted by $1 - \epsilon$, where $\epsilon > 0$ is a small number) and alternative densities (weighted by ϵ), can also be used to define $\mathcal{P}$. Different approaches generate plausible regions of different shapes.

Because formal distance measures are esoteric and awkward to work with, a more intuitive approach to distribution-centric regions is based on quantiles (e.g., Berger 1990). Avoiding extreme quantiles, for which expert input can be unreliable, the nominal plus-or-minus one standard deviation interval, for example, is such that

$$\mathrm{Pr}_{m_0,s_0}\left(\log h_{50} \ < \ \log 70 \ - \ \log 1.2\right) \ = \ \Phi(-1) \ = \ 0.1587 \ \equiv \ p_0^- \ , \qquad \text{and}$$

$$\mathrm{Pr}_{m_0,s_0}\left(\log h_{50} \ < \ \log 70 \ + \ \log 1.2\right) \ = \ \Phi(+1) \ = \ 0.8413 \ \equiv \ p_0^+ \ .$$

In actuality, the precise values 0.1587 and 0.8413 are approximations. Suppose the expert deems any probability values $p^- \in [0.05, 0.25]$ and $p^+ \in [0.75, 0.95]$ to be consistent with the precise values $p_0^- = 0.1587$ and $p_0^+ = 0.8413$. Then, in terms of Gaussian priors, any pair of values (m, s) such that

$$\mathrm{Pr}_{m,s}\left(\log h_{50} \ < \ \log 70 \ - \ \log 1.2\right) \ \in \ [0.05, 0.25] \ , \qquad \text{and}$$

$$\Pr_{m,s}\,(\log h_{50}\ <\ \log 70 + \log 1.2)\ \in\ [0.75, 0.95]$$

defines a Gaussian prior distribution on log h_{50} that is consistent with expert input. That is, the 2-D region $p^- \in [0.05, 0.25]$ and $p^+ \in [0.75, 0.95]$ maps to a 2-D region for (m, s). A gridded version of this region is displayed in Figure 3, plotted on the same scale as the parameter-centric region in Figure 2.

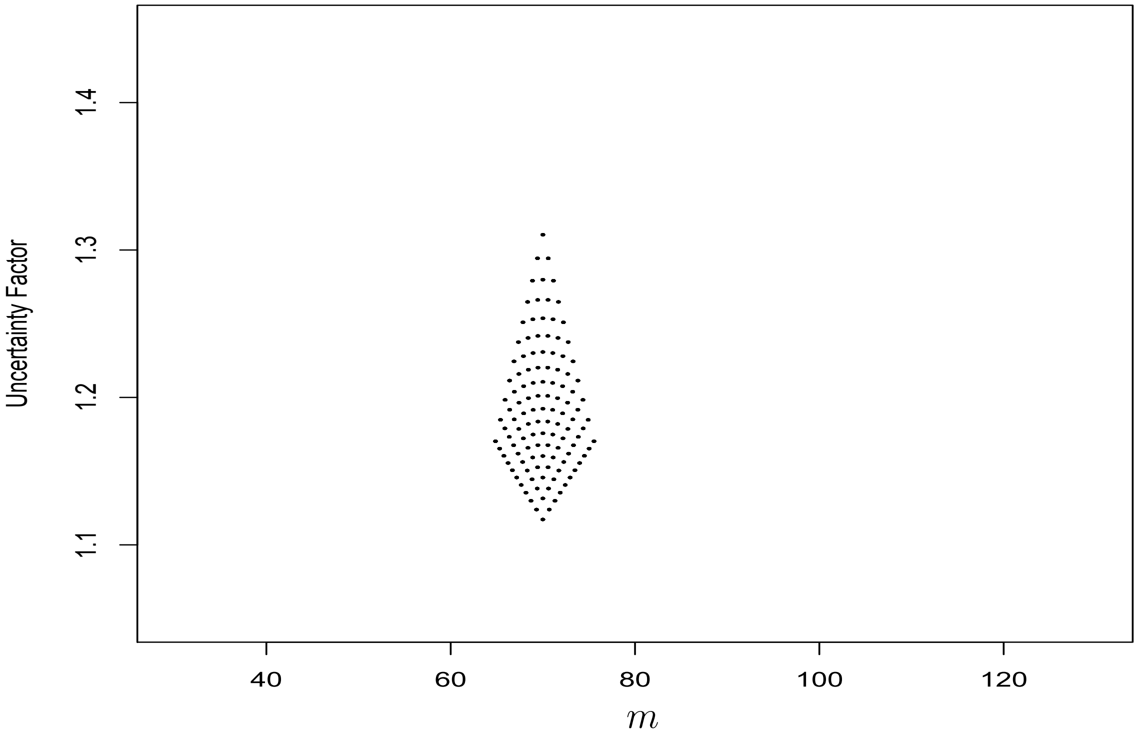

Figure 3: Distribution-Centric Region for (m, s).

Upon comparing Figures 2 and 3, a substantial difference exists between the parameter-centric and distribution-centric plausible regions. In one sense, the difference is not surprising, in that the regions were constructed via different mechanisms. Nonetheless, imprecisions are strongly affected, in that maximizing/minimizing posterior probabilities over the much larger parameter-centric region would be different than doing the same over the smaller distribution-centric region.

A related issue involves miscalibrated experts. For a variety of reasons (e.g., Fischhoff, Slovic, and Lichtenstein 1982, and much related research), experts are often overconfident, and understate the uncertainties between their estimates and reality. Relative to IPs, this means that representor sets $\mathcal{P}$ can be smaller than desired, and IP bounds too narrow. This is not the fault of IP methodology, of course, but it underscores the need for careful thought in defining which prior distributions are truly “consistent with” expert knowledge.

A distribution-centric region for the h_{50}/h_{10} ratio follows similarly. The nominal mean for the h_{10}/h_{50} ratio is 1/2, and the nominal standard deviation is 0.21, leading to the nominal probabilities for $\alpha_0 = \beta_0 = 2.4$

$$\mathrm{Pr}_{\alpha_0,\beta_0}\left(h_{10}/h_{50} \ < \ 0.50 - 0.21\right) \ = \ 0.1816 \ \equiv \ p_0^- \ , \qquad \text{and}$$

$$\mathrm{Pr}_{\alpha_0,\beta_0}\left(h_{10}/h_{50} \ < \ 0.50 + 0.21\right) \ = \ 0.8184 \ \equiv \ p_0^+ \ .$$

Using the same bounds as for h_{50}, and finding beta parameters α and β such that

$$\mathrm{Pr}_{\alpha,\beta}\left(h_{10}/h_{50} \ < \ 0.50 - 0.21\right) \ \in \ [0.05, 0.25] \ , \qquad \text{and}$$

$$\mathrm{Pr}_{\alpha,\beta}\left(h_{10}/h_{50} \ < \ 0.50 + 0.21\right) \ \in \ [0.75, 0.95]$$

leads to the region displayed in Figure 4.

The asymmetry in Figure 4 is an artifact of plotting in elicitation space. That is, the process of converting the elicited $(h_{50}/h_{10}, UF) = (2, 1.5)$ to beta distribution parameters (α, β) can be reversed to convert a region in (α, β) space into the corresponding region in elicitation space. Were the region in Figure 4 to be plotted in terms of the mean and standard deviation for the underlying beta distributions, the plot would be similar to Figure 3.

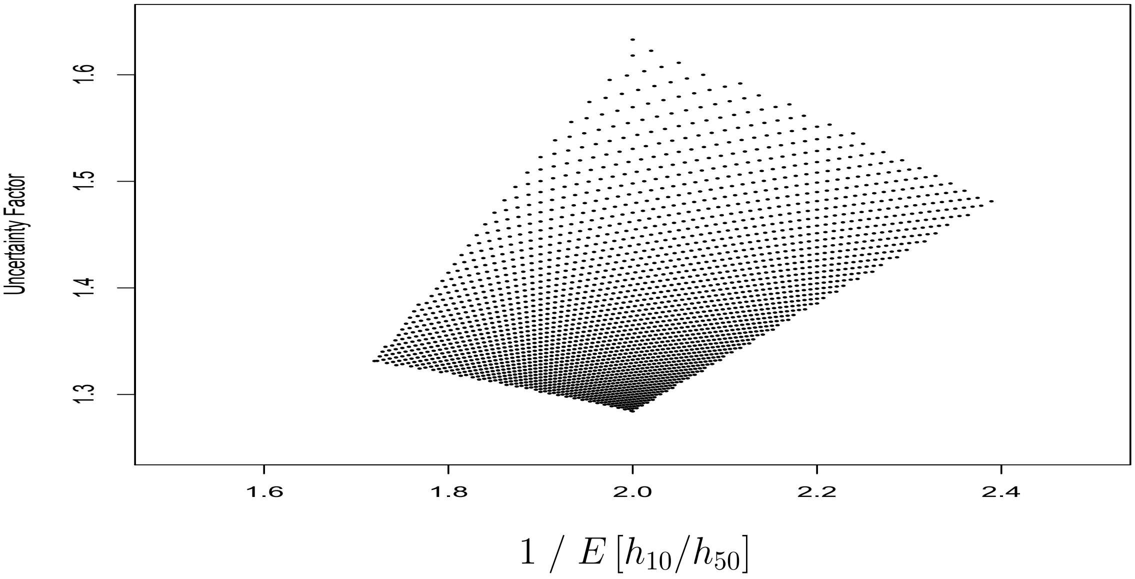

Figure 4: Region of Plausible h_{50}/h_{10} Parameter Values.

We illustrate IP methods using the distribution-centric set $\mathcal{P}$ of prior distributions formed by combining the regions in Figures 3 and 4. The four vertices in Figure 3 are

$$(m\,,\,s) \;=\; (64.9, 1.17), (70, 1.12), (70, 1.31), \text{and } (75.5, 1.17)\;.$$

The four vertices in Figure 4 are

$$(1 \ / \ E\left[h_{10}/h_{50}\right]\,,\,\mathrm{UF}) \;=\; (1.72, 1.33), (2, 1.28), (2, 1.63), \text{and } (2.39, 1.48)\;.$$

As noted, the four vertices in Figure 4 are symmetric in the first two moments of the corresponding beta distribution for h_{10}/h_{50}, e.g., the prior mean 1/2 for the beta distribution is midway between the values 1/1.72 and 1/2.39.

Table 2 summarizes posterior probabilities (based on 100K MCMC samples per case) that the 10^{-6} drop height exceeds 5 cm for the 16 combinations of the h_{50} and h_{50}/h_{10} vertices in $\mathcal{P}$.

Table 2. Probit Posterior Probabilities Pr(A)

$(1\ /\ E\,[h_{10}/h_{50}]$, UF)	(m, s)			
	(64.9, 1.17)	(70, 1.12)	(70, 1.31)	(75.5, 1.17)
(1.72, 1.33)	.63	.65	.59	.62
(2, 1.28)	.49	53	.46	.50
(2, 1.63)	.43	.45	.38	.40
(2.39, 1.48)	.33	.36	.30	.32

From Section 6, the nominal prior led to $\Pr_{\text{posterior}}(A) = 0.45216$. Maximum and minimum values of $\Pr_{\text{posterior}}(A)$ over the 16 priors in Table 2 are $\overline{\Pr}(A) = 0.65$ and $\underline{\Pr}(A) = 0.30$. The non-probabilistic IP bounds [0.30, 0.65] span a factor-of-2 range and allow for a much better interpretation of the data than the precise value 0.45216 alone. It is also interesting that IP bounds for the posterior mean of h_{50} are [63, 68], indicating that non-extrapolated quantities are less sensitive to prior specification than are extrapolated ones.

Despite theoretical work on characterizing extreme values for certain types of representor sets $\mathcal{P}$ and certain posterior quantities (e.g., Sivaganesan and Berger 1989), these results are limited. In particular, we do not have a mathematical proof that the maximium/minimum over $\mathcal{P}$ of $\Pr_{\text{posterior}}(A)$ occur at vertices of the plausible region, although there are intuitive reasons to believe so. All Gaussian/beta prior distributions in $\mathcal{P}$ are unimodal, well behaved, and change smoothly over $\mathcal{P}$; the likelihood function (Figure 1) is smooth as well. The quantity of interest, $\mathrm{Pr}_{\mathrm{posterior}}(A)$, is "ratio linear" (Berger 1990, p. 309) and its corresponding integral is over a simple region of parameter space.

If there were a large number of prior distributions for which $\mathrm{Pr}_{\mathrm{posterior}}(A)$ were computed, and the dimension of the parameter space were small, as for textbook problems, it is useful to generate contour plots of $\mathrm{Pr}_{\mathrm{posterior}}(A)$ as a function of prior parameter values. This can help confirm, to the extent possible, that maximum/minimum values have been found. When posterior quantities are produced by MCMC simulations, it is not practical to consider a large number of priors in $\mathcal{P}$. In that case, it is useful to generate MCMC samples on a space filling design over the plausible region (parallel computing helps here).

One space filling design for the detonator impact data consists of 32 additional MCMC runs, 16 of which are on the boundary of $\mathcal{P}$ and the other 16 are in the interior. The boundary points are the midpoints along the (slightly nonlinear) edges between each pair of adjacent vertices, and the interior points are the midpoints of the nominal prior parameters and each vertex. For the h_{10}/h_{50} ratio, midpoints are obtained in the space of the mean and standard deviation. Results are tabled in the appendix and plotted in Figure 5.

In Figure 5, $\mathrm{Pr}_{\mathrm{posterior}}(A)$ is plotted against the most important prior parameter, the h_{50}/h_{10} ratio (technically, the reciprocal of $E\left[h_{10}/h_{50}\right]$). A clear trend exists, with lower h_{50}/h_{10} values yielding higher posterior probabilities. Other prior parameters also matter, but not to the same extent. This result is not surprising, given the h_{-6} contours in Figure 1.

Results from the space filling design provide a soft confirmation that the IP bounds are good ones. Nonetheless, for most realistic problems, there is no way to guarantee that the true $\overline{\mathrm{Pr}}(A)$ and $\underline{\mathrm{Pr}}(A)$ have in fact been obtained.

One other issue is worthy of note. In the drop test example, posterior distributions

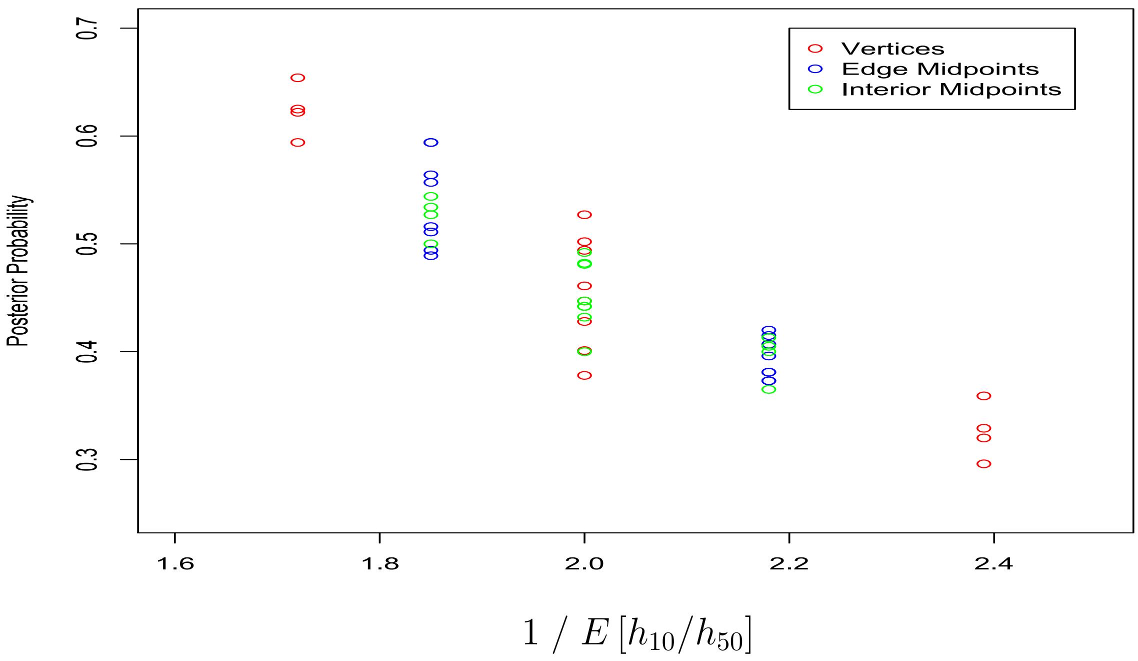

Figure 5: Posterior Probabilities from Space Filling Design.

are "smaller" than the prior distributions. That is, uncertainties from the data and prior combined are smaller than those from the prior alone. Although this situation is typical, it need not always occur. The term *dilation* refers to the case where posterior imprecisions are larger than prior imprecisions (e.g., Seidenfeld and Wasserman 1993).

One common cause of dilation is surprise data – e.g., the expert felt fairly certain of a situation initially, but surprising experimental results left the expert feeling less confident after seeing the results than he thought he was beforehand. A second cause of dilation occurs in prior-by-committee situations regarding complex systems, where interest lies in overall system reliability, but experts disagree at the subsystem and component levels. IP analyses better quantify dilation effects than does conventional Bayesian reliability.

9 Computational Models

To this point, attention has focused on calibrating the probit model (1) to the drop test results, and combining with expert opinion to extrapolate tail probability predictions outside the range of the data. Besides the probit model, several other sigmoid-like models, such as the Arrenhius model for chemical kinetics, are also used in conjunction with monotonically increasing phenomena in (0,1). For present purposes, it suffices to consider the two-parameter Weibull model (Meeker and Escobar 1998, Eq. 4.7). This model postulates that the probability of a "go" response at height h is

$$\Pr(\text{ “go” }) \;=\; 1 \;-\; \exp\left\{-\exp\left[\frac{\log h \;-\; \mu}{\sigma_{\mathrm{W}}}\right]\right\} \;, \tag{3}$$

where $(\mu, \sigma_{\mathrm{W}})$ are model parameters (the notation σ_{W} used to distinguish the Weibull scale factor from the probit scale factor σ). Similar to the probit model, prior information on the 50-50 drop height h_{50} and the h_{50}/h_{10} ratio can be translated directly into prior information on Weibull model parameters. See the appendix for details.

Using the nominal prior distributions on h_{50} and h_{50}/h_{10} in Section 6, posterior means for μ and σ_{W} can be substituted into the Weibull probability relation (3) to provide a curve-fit approximation to Pr("go") as a function of drop height. The same can be done for the probit model (1), and overlaying the two curves gives Figure 6.

There is no practical difference between the probit and Weibull model curve fits over the 45-90 cm range of the experimental data. The strong similarity is not surprising, in that both computational models are based on the same prior information and same data. When the models are extrapolated, however, substantial differences exist.

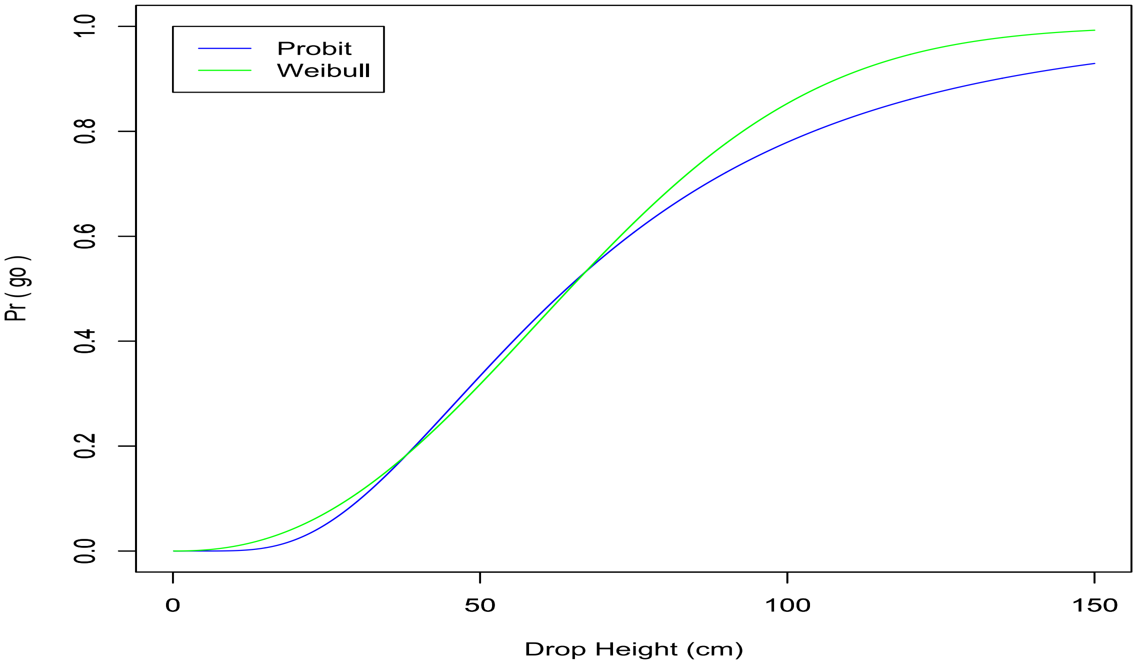

Figure 6: Overlay of Probit and Weibull Data Fits.

The (green) Weibull curve extrapolates well above the (blue) probit curve for high/low drop heights. In particular, the Weibull model's estimated probability of a "go" at a 5 cm drop height is well above that for the probit model. This greatly affects the estimated 10^{-6} drop height. For example, a 1M MCMC sample yields the Weibull posterior probability $\mathrm{Pr}^{\mathrm{W}}_{\mathrm{posterior}}(A) = 0.035125$ (five significant digits to re-emphasize the pretentious accuracy in lengthy MCMC samples).

Weibull IP bounds over the 16 combinations of vertices in Table 2 are [.01, .07], and do not even overlap with those of the probit model, [.30, .65]. Interpretation of the Weibull-versus-probit difference is non-probabilistic. In contrast to Bayesian model averaging, the subject matter expert does not assign a prior probability that each model is "correct" in a physical sense. Indeed (analogous to Newtonian mechanics, for example) both models are known to be "wrong" but are still useful for curve fitting, interpolation, and extrapolation.

In IP analyses, incorporation of multiple, reasonable computational models in the representor set $\mathcal{P}$ is important, and can reveal the large uncertainties that often accompany extrapolation. Were the extrapolation less severe than the 10^{-6} tail probability for a 5 cm drop height, uncertainties would be reduced. Consider Figures 7-9, which examine extrapolation in "both dimensions," where the first dimension extrapolates to drop heights outside the 45-90 cm range of the data, and the second dimension extrapolates to tail probabilities such as the 10^{-6} probability for the performance requirement.

In Figure 7, the probit IP bounds [0.30, 0.65] and Weibull IP bounds [0.01, 0.07] are displayed at drop height $h = 5$ cm. To the right side of the plot, there is no practical difference, in that both computational models agree that there is essentially no chance that the 10^{-6} drop height exceeds 30 cm. Model-to-model differences here increase with the degree of extrapolation in drop height.

Extrapolation in tail probability is also important. Consider Figure 8 and the 10% drop height, which is a much less severe extrapolation than the 10^{-6} drop height. The models agree that it is almost certain the 10% drop height exceeds 5 cm. Moreover, comparing the nominal values and IP bounds for both models across the range of drop heights, the minimal extrapolation in tail probability leads to model-to-model differences that are comparatively minor; compare the blue probit bounds to the green Weibull bounds in Figure 8. For both models, however, IP bounds are wide enough to matter for several drop heights.

As the extrapolation in tail probability becomes more substantial, the situation changes.

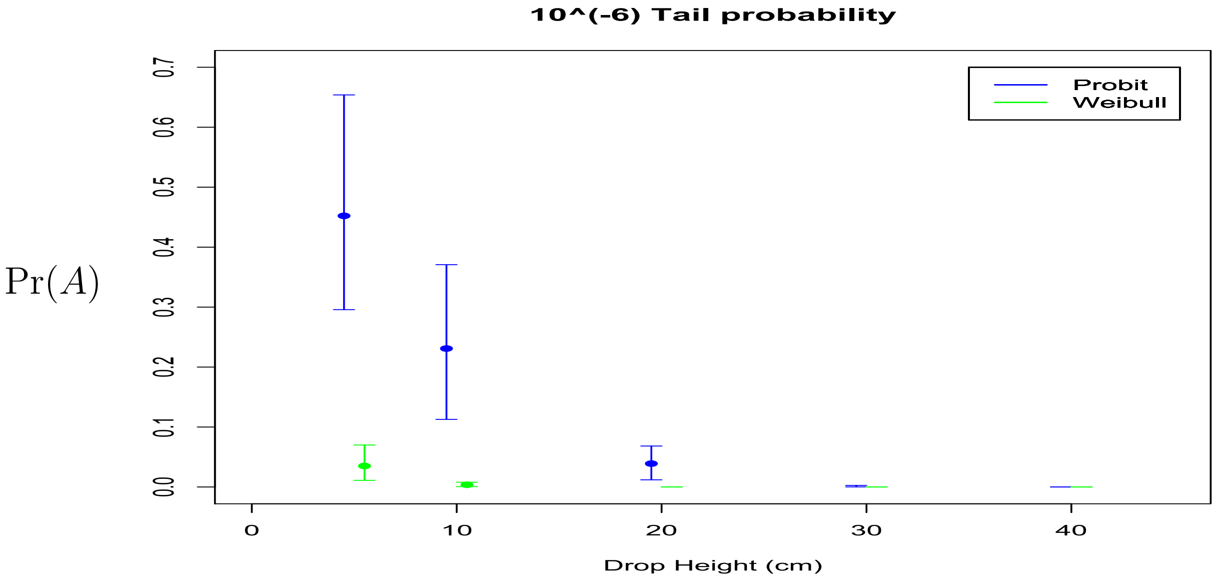

Figure 7: IP Bounds.

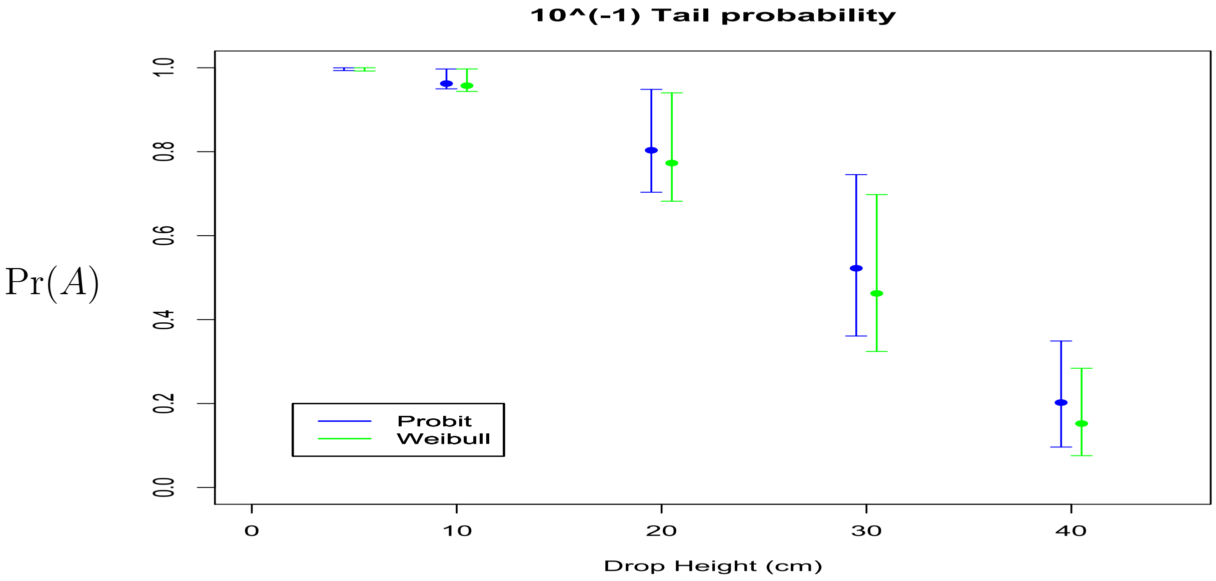

Figure 8: IP Bounds.

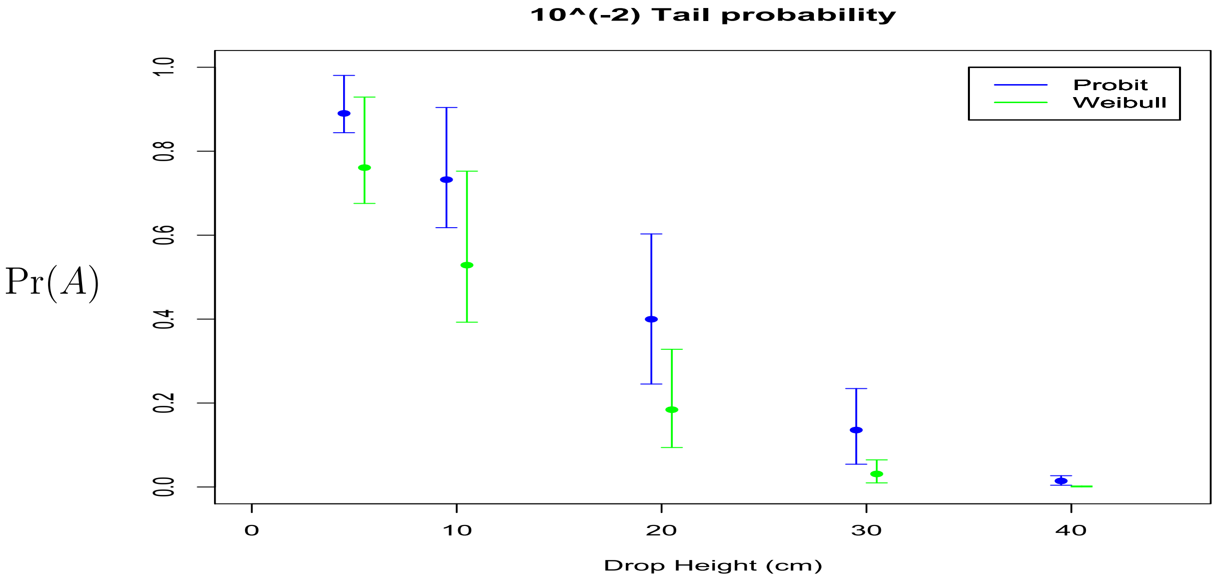

Figure 9: IP Bounds.

Consider Figure 9, examining 1% drop heights, which are intermediate to the 10^{-6} drop height in Figure 7 and the 10% drop height in Figure 8. Model-to-model differences become more apparent in comparison to those for the 10% drop heights. Those differences become still larger with the extrapolation in tail probability, eventually leading to the huge differences in Figure 7 for the 10^{-6} performance requirement at 5 cm.

In a sense, the probit-versus-Weibull model extrapolations are reassuring. If the degree of extrapolation is modest, the different computational models give comparable results. On the other hand, if the two models are each extrapolated to the moon, as per a 10^{-6} tail probability for a 5 cm drop height, the IP bounds range from 0.01 (the lower probability for the Weibull model) to 0.65 (the upper probability for the probit model).

Finally, other artifacts of Bayes theory, such as credible intervals and decision analysis, are also impacted by artificially precise notions of precision. Displays of Bayesian credible intervals (superficially similar to Figures 7-9) can be misleading. IP versions of credible intervals exist, and can be much wider than nominal credible intervals because they account for imprecision in the nominally "precise" prior. Similarly, classical Bayes decision theory requires precise priors and precise utility functions; IP alternatives (e.g., Walley 1991, Sec. 3.9) are more realistic.

10 Computational Methods via Direct Simulation

The probit and Weibull models for the drop test data are convenient for illustrating IP concepts, for displaying representor sets $\mathcal{P}$ of prior distributions, graphing 10^{-6} drop height curves, and so on. More importantly, the 10^{-6} rare events are known functions of parameters, and each step of the MCMC provides a separate estimate of the 10^{-6} drop height.

IP applications to weapons UQ would not be so nice computationally. A more common situation arises when using computer models that directly simulate the physical phenomena of interest. Return to the accident scenario of a worker inadvertently dropping a wrench on a detonator, and consider a hypothetical computer code that

a) captures through certain code inputs the accident scenario (i.e., describing the many ways a wrench could be dropped on a detonator),

b) captures through other code inputs the detonator-to-detonator differences that result from manufacturing variation, and

c) calculates the detonator physics for the incident (which, as in the drop test example, vary from one computational model to another).

Were the computer code to faithfully simulate the accident scenario for a safe drop height $-$ say, with a 10^{-6} "go" probability $-$ then only one in a million ordinary code runs, on the average, would result in a "go" outcome. Unless code runs could be done quickly, assessing the accident scenario would be impractical even if only a nominal prior distribution were considered. Carrying out an IP analysis over many prior distributions across $\mathcal{P}$ would be hopeless.

The resolution to this situation involves *not* simulating the accident scenario directly, but instead to use importance sampling. Importance sampling is a concept that dates back to the 1950s, and is in widespread use at LANL, a simple example being the use of the MCNP code to assess a radiation shielding application, where a very small portion of particles penetrate the shielding.

Statistical methods to address rare events for complex computer codes are presented in Picard and Williams (2013), and can be extended to IP analyses in straightforward fashion.

There are computational tricks involving importance sampling to improve efficiency for rare events by orders of magnitude relative to direct simulation. There are other tricks to obtain results for multiple prior distributions as part searching the space $\mathcal{P}$ without re-running MCMCs from scratch, further improving efficiency. Providing details would amount to a separate report in itself, though basic items are sketched in the appendix.

Because there is no actual computer model to directly simulate the accident scenario (i.e., a wrench inadvertently dropped from a height of interest), the above discussion is purely conceptual. The potential for huge efficiency gains involving rare events is considerable, however.

11 Summary

The fundamental premise of Bayesian methodology is that expert opinion is accurately quantified by a single, precise prior distribution. When this premise is false, (mis)application of Bayes methods can produce posterior quantities whose apparent precisions are misleading, as shown in the detonator case study. In such situations, IPs are a good alternative.

Overriding conclusions relative to the use of IP methods for weapons UQ are many:

1) There do exist situations where the conventional Bayes approach is adequate, especially when prior information is limited. For the informative priors and extrapolated quantities common to weapons UQ, however, IP bounds are often wide enough to expose the shortcomings of Bayesian methods.

2) IP bounds are not free. Construction of the set $\mathcal{P}$ of priors requires careful thought. There is no simple recipe to follow, and deciding where to draw the line between priors that are consistent with expert beliefs and those that aren't is problematic.

3) In realistic applications, true upper and lower probabilities $\overline{\mathrm{Pr}}(A)$ and $\underline{\mathrm{Pr}}(A)$ are never obtained exactly. Posterior quantities are estimated with MCMC sampling error, and when MCMC runs are time consuming, it is not possible to fully explore the set $\mathcal{P}$. Computational tricks involving importance sampling are helpful here.

4) Lastly, neither the work involved in 2) nor the uncertainties in 3) justify being misled by a Bayesian analysis. Benefits of IP analyses in more realistically representing expert beliefs $-$ and quantifying uncertainties that follow from them $-$ are substantial.

Appendix: The Weibull Model

Several models exist for analyzing drop test data. The probit model was described in Section 2. Another model that is commonly used is the two-parameter Weibull, which postulates that the probability of a "go" response at height h is

$$\Pr(\text{ “go” }) \;=\; 1 \;-\; \exp\left\{-\exp\left[\frac{\log h \;-\; \mu}{\sigma_{\mathrm{W}}}\right]\right\} \;,$$

where μ and σ_{W} are parameters to be estimated from the prior and the data.

As with the probit model, the Weibull model can be inverted to give drop height as a function of the probability of a "go" response:

$$h_{\Pr(\mathrm{go})} \;=\; \exp\left\{\mu \;+\; z_{\Pr(\mathrm{go})}\,\sigma_{\mathrm{W}}\right\} \qquad \text{for} \quad z_{\Pr(\mathrm{go})} \;=\; \log\left[-\log\left(1-\Pr(\text{ “go” })\right)\right] .$$

Parameterizing the Weibull model in a way that is consistent with the expert input, the model can be expressed in terms of the 50-50 drop height h_{50}. For $\Pr(\text{ “go” }) = 0.5$, $z_{50} = -0.37$, and

$$h_{50} \;=\; \exp\{\mu + z_{50}\,\sigma_{\mathrm{W}}\} \qquad \Rightarrow \qquad \mu \;=\; \log h_{50} - z_{50}\,\sigma_{\mathrm{W}} \;.$$

Continuing, the h_{50}/h_{10} ratio is

$$\begin{aligned}\frac{h_{50}}{h_{10}} \;&=\; \frac{\exp\{\;\mu \;+\; z_{50}\,\sigma_{W}\;\}}{\exp\{\;\mu \;+\; z_{10}\,\sigma_{\mathrm{W}}\;\}} \\ &=\; \exp\{\,(z_{50} \;-\; z_{10}\,)\,\sigma_{\mathrm{W}}\} \;,\end{aligned}$$

which solves to

$$\sigma_{\mathrm{W}} \;=\; \frac{\log\;(h_{50}/h_{10})}{z_{50} - z_{10}} .$$

As with the probit model, the scale factor σ_W differs from the log h_{50}/h_{10} ratio by a multiplying constant, here $z_{50} - z_{10} = 1.88$.

Appendix: Space Filling Design Results

MCMC results (100K samples per case) for the space filling design are given in the tables below. As expected, the range of values in these tables, from 0.37 to 0.59, lies within the range [0.30, 0.65] for the values from the vertices in Table 1 of the text.

Table A-1. Probit Posterior Probabilities Pr(A) for Edge Midpoints

$(1/E[h_{10}/h_{50}]$, UF)	(m, s)			
	(67.4, 1.14)	(67.4, 1.24)	(72.8, 1.14)	(72.8, 1.24)
(1.85, 1.31)	.59	.56	.59	.56
(1.85, 1.46)	.51	.49	.52	.49
(2.18, 1.37)	.42	.38	.41	.38
(2.18, 1.57)	.42	.37	.40	.37

Table A-2. Probit Posterior Probabilities Pr(A) for Interior Midpoints

$(1/E[h_{10}/h_{50}]$, UF)	(m, s)			
	(67.4, 1.18)	(70, 1.31)	(70, 1.16)	(72.8, 1.18)
(1.85, 1.41)	0.53	0.50	0.54	0.53
(2, 1.39)	0.48	0.44	0.49	0.48
(2, 1.57)	0.44	0.40	0.45	0.43
(2.18, 1.49)	0.41	0.37	0.41	0.40

Appendix: Importance Sampling

Unfortunately, describing detailed practical computational methods for handling rare events requires introducing some notation. To that end, return to the hypothetical computer code discussed in Section 10, which directly simulates the physical process of interest. Let

a) $\boldsymbol{x}$ denote inputs to the hypothetical code describing the accident scenario, e.g., specifying how a wrench could fall on a detonator and how a particular detonator may differ from others in the same manufacturing lot,

b) $\boldsymbol{\theta}$ denote uncertain inputs to the code such as physics constants in a computational model that are calibrated to experimental data,

c) $\eta(\boldsymbol{x}, \boldsymbol{\theta})$ denote the code output of interest, the output being whether the detonator response was a "go" or a "no-go" (in actuality, such codes provide many outputs of interest, but for simplicity, we focus here on only one output),

d) the set $\mathcal{A} = \{(\boldsymbol{x}, \boldsymbol{\theta}) \,|\, \eta(\boldsymbol{x}, \boldsymbol{\theta}) = \text{"go"}\}$ denote inputs $(\boldsymbol{x}, \boldsymbol{\theta})$ for which the code gives a "go" output, and

e) $1_{\mathcal{A}}(\boldsymbol{x}, \boldsymbol{\theta})$, denote the indicator function for the set $\mathcal{A}$, i.e., the indicator function equals 1 for inputs in $\mathcal{A}$ and equals zero otherwise.

For $f(\boldsymbol{x})$ the probability distribution describing the accident scenario and $\pi(\boldsymbol{\theta})$ a single (precise) posterior distribution for the calibration parameters, the posterior probability of $\mathcal{A}$ can be written in integral form

$$\Pr_{\text{posterior}}(\mathcal{A}) \;=\; \int\!\!\int 1_{\mathcal{A}}(\boldsymbol{x}, \boldsymbol{\theta}) \; f(\boldsymbol{x}) \; \pi(\boldsymbol{\theta}) \; d\boldsymbol{x} \; d\boldsymbol{\theta} \; . \tag{4}$$

The next step is to explicitly incorporate the prior distribution. For $\boldsymbol{y}$ the calibration data, $\ell(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ the likelihood function for the data, and $p(\boldsymbol{\theta})$ the precise prior distribution, the posterior density is

$$\pi(\boldsymbol{\theta}) \;=\; \frac{\ell(\boldsymbol{y} \,|\, \boldsymbol{\theta}) \; p(\boldsymbol{\theta})}{\int \; \ell(\boldsymbol{y} \,|\, \boldsymbol{\theta}) \; p(\boldsymbol{\theta}) \; d\boldsymbol{\theta}} \;\equiv\; \frac{\ell(\boldsymbol{y} \,|\, \boldsymbol{\theta}) \; p(\boldsymbol{\theta})}{L^{(p)}\,(\boldsymbol{y})} \;,$$

where $L^{(p)}\,(\boldsymbol{y})$ is the so-called marginal likelihood for the prior $p(\boldsymbol{\theta})$. The posterior probability is then

$$\begin{aligned} \mathrm{Pr}_{\text{posterior}}(\mathcal{A}) &= \int\int 1_{\mathcal{A}}(\boldsymbol{x},\boldsymbol{\theta})\; f(\boldsymbol{x})\; \pi(\boldsymbol{\theta})\; d\boldsymbol{x}\; d\boldsymbol{\theta} \\ &= \frac{\int\int 1_{\mathcal{A}}(\boldsymbol{x},\boldsymbol{\theta})\; f(\boldsymbol{x})\; \ell(\boldsymbol{y}\,|\,\boldsymbol{\theta})\; p(\boldsymbol{\theta})\; d\boldsymbol{x}\; d\boldsymbol{\theta}}{L^{(p)}\,(\boldsymbol{y})} \end{aligned} \tag{5}$$

The plan is to estimate separately the numerator and denominator of (5).

The denominator is simpler to deal with, and is considered first. $L^{(p)}\,(\boldsymbol{y})$ is the normalizing constant for the posterior distribution $\pi(\boldsymbol{\theta})$. In isolated cases such as conjugate priors, the integral for $L^{(p)}\,(\boldsymbol{y})$ can be solved analytically; in other isolated cases such as for low-dimensional $\boldsymbol{\theta}$, it can be computed using numerical integration. For most realistic cases, simulation estimates must be obtained, and importance sampling is a good option.

Write

$$\begin{aligned} L^{(p)}\,(\boldsymbol{y}) &\equiv \int \ell(\boldsymbol{y}\,|\,\boldsymbol{\theta})\; p(\boldsymbol{\theta})\; d\boldsymbol{\theta} \\ &= \int \ell(\boldsymbol{y}\,|\,\boldsymbol{\theta})\; \frac{p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\;\; g(\boldsymbol{\theta})\; d\boldsymbol{\theta}\;, \end{aligned}$$

where the second equation is obtained upon multiplying and dividing by an importance density $g(\,\boldsymbol{\theta}\,)$ to be selected. Parameter values $\{\boldsymbol{\theta}_j\,;\; j = 1,\ldots,N\}$ are simulated from $g(\boldsymbol{\theta})$, and the estimated marginal likelihood follows directly from this integral,

$$\widehat{L}^{(p)}\,(\boldsymbol{y}) \;=\; \frac{1}{N}\sum_{j=1}^{N} \frac{\ell(\,\boldsymbol{y}\,|\,\boldsymbol{\theta}_j\,)\; p(\,\boldsymbol{\theta}_j\,)}{g(\boldsymbol{\theta}_j)}\;. \tag{6}$$

The variance of this estimate is

$$\begin{aligned} Var_g\left[\,\widehat{L}^{(p)}(\boldsymbol{y})\right] &= \frac{1}{N}\,E_g\left\{\left[\frac{\ell(\,\boldsymbol{y}\,|\,\boldsymbol{\theta}\,)\,p(\,\boldsymbol{\theta}\,)}{g(\boldsymbol{\theta})} \;-\; L^{(p)}\,(\boldsymbol{y})\right]^2\right\} \\ = \frac{\left[\,L^{(p)}(\boldsymbol{y})\right]^2}{N}\; &E_g\left\{\left[\frac{\ell(\,\boldsymbol{y}\,|\,\boldsymbol{\theta}\,)\,p(\,\boldsymbol{\theta}\,)}{g(\boldsymbol{\theta})\;L^{(p)}\,(\boldsymbol{y})} \;-\; 1\right]^2\right\} \end{aligned}$$

$$= \frac{\left[L^{(p)}(\boldsymbol{y}) \right]^2}{N} \int \left\{ \left[\frac{\ell(\,\boldsymbol{y}\,|\,\boldsymbol{\theta}\,)\; p(\,\boldsymbol{\theta}\,)}{g(\boldsymbol{\theta})\; L^{(p)}\,(\boldsymbol{y})} \; - \; 1 \right]^2 \right\} g(\boldsymbol{\theta})\; d\boldsymbol{\theta}$$

$$= \frac{\left[L^{(p)}(\boldsymbol{y}) \right]^2}{N} \int \frac{\left[\ell(\,\boldsymbol{y}\,|\,\boldsymbol{\theta}\,)\; p(\,\boldsymbol{\theta}\,)\,/\,L^{(p)}\,(\boldsymbol{y}) \; - \; g(\boldsymbol{\theta}) \right]^2}{g(\boldsymbol{\theta})}\, d\boldsymbol{\theta}$$

$$= \frac{\left[L^{(p)}\,(\boldsymbol{y}) \right]^2}{N} \int \frac{[\,\pi(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\,]^2}{g(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\;, \tag{7}$$

for $\pi(\boldsymbol{\theta}) \;=\; \ell(\,\boldsymbol{y}\,|\,\boldsymbol{\theta}\,)\;\; p(\,\boldsymbol{\theta}\,)\,/\,L^{(p)}\,(\boldsymbol{y})$ the true posterior. Note that (7) is essentially the χ^2 distance between the importance density $g(\boldsymbol{\theta})$ and the actual posterior $\pi(\boldsymbol{\theta})$. This means that an importance distribution approximating the posterior will yield good results. A useful approach here is to use the MCMC sample from the posterior to construct an estimated posterior distribution $\widehat{g}(\boldsymbol{\theta})$ and to estimate $L^{(p)}(\boldsymbol{y})$ via (6) using $\widehat{g}(\boldsymbol{\theta})$.

Dealing with the numerator of the rare event probability (5) is less straightforward. This integral can be written

$$\int \int 1_{\mathcal{A}}(\boldsymbol{x},\boldsymbol{\theta})\; f(\boldsymbol{x})\; \ell(\boldsymbol{y}\,|\,\boldsymbol{\theta})\; p(\boldsymbol{\theta})\; d\boldsymbol{x}\; d\boldsymbol{\theta}$$
$$= \int \int 1_{\mathcal{A}}(\boldsymbol{x},\boldsymbol{\theta})\; \ell(\boldsymbol{y}\,|\,\boldsymbol{\theta})\; \frac{f(\boldsymbol{x})\; p(\boldsymbol{\theta})}{g(\boldsymbol{x},\boldsymbol{\theta})}\; g(\boldsymbol{x},\boldsymbol{\theta})\; d\boldsymbol{x}\; d\boldsymbol{\theta}\;,$$

where again $g(\boldsymbol{x},\boldsymbol{\theta})$ is an importance distribution to be chosen. Although there are occasional advantages to using the same importance density for the denominator of (5) as for the numerator (e.g., O'Neill 2009), especially when non-extrapolated quantities are involved, this is a poor approach for rare events and other extrapolations.

Upon simulating $\{(\boldsymbol{x}_j,\boldsymbol{\theta}_j)\}$ from the importance distribution $g(\boldsymbol{x},\boldsymbol{\theta})$, the importance sampling estimate of the numerator is

$$\frac{1}{N} \sum_{j=1}^{N} 1_{\mathcal{A}}(\boldsymbol{x}_j,\boldsymbol{\theta}_j)\; \ell(\boldsymbol{y}\,|\,\boldsymbol{\theta}_j)\; \frac{f(\boldsymbol{x}_j)\; p(\boldsymbol{\theta}_j)}{g(\boldsymbol{x}_j,\boldsymbol{\theta}_j)}\;.$$

Stated in English, for each code run with a "go" output, the likelihood value $\ell(\boldsymbol{y} \mid \boldsymbol{\theta}_j)$ is multiplied by its importance weight $f(\boldsymbol{x}_j)\, p(\boldsymbol{\theta}_j) \;/\; g(\boldsymbol{x}_j, \boldsymbol{\theta}_j)$. Upon adding these quantities, the overall estimate of the numerator is obtained.

In estimating the numerator, it is inefficient to waste 99.9999% of code runs on the portion of the input space outside of $\mathcal{A}$, where no rare events occur. This is exactly what happens when inputs $(\boldsymbol{x}, \boldsymbol{\theta})$ are sampled from the nominal distribution $f(\boldsymbol{x})\, \pi(\boldsymbol{\theta})$. Instead, it is better to sample inputs $(\boldsymbol{x}, \boldsymbol{\theta})$ from a distribution highlighting the portion of the space that is important for rare events, similar to the biasing tools in MCNP.

The optimal importance distribution $g(\boldsymbol{x}, \boldsymbol{\theta})$ here can be characterized. Were a large MCMC sample $\{\boldsymbol{\theta}_j\}$ to exist from the posterior $\pi(\boldsymbol{\theta})$, and coupled with a large sample $\{\boldsymbol{x}_j\}$ from $f(\boldsymbol{x})$, the subset of inputs $\{(\boldsymbol{x}_j, \boldsymbol{\theta}_j) \in \mathcal{A}\}$ would represent a random sample from the optimal importance distribution $g(\boldsymbol{x}, \boldsymbol{\theta})$. Its corresponding probability density function is proportional to $f(\boldsymbol{x})\, \pi(\boldsymbol{\theta})$ for $(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{A}$ *only.* Because the set $\mathcal{A}$ is not known in practical problems, a good strategy is to begin importance sampling with an initial importance distribution $g_0(\boldsymbol{x}, \boldsymbol{\theta})$ based on subject matter knowledge. Results from this initial simulation can be used to adaptively improve the importance distribution, leading to $g_1(\boldsymbol{x}, \boldsymbol{\theta})$ and, if necessary, to $g_2(\boldsymbol{x}, \boldsymbol{\theta})$, and so on. Details of this adaptive procedure are beyond the scope of this report, but can be found elsewhere (Picard and Williams 2013).

Variance reduction factors to estimate 10^{-6} rare event probabilities, equivalent to the reduced number of importance-sampled code runs required (versus "ordinary" code runs based on samples from the nominal posterior $\pi(\boldsymbol{\theta})$) often range over several orders of magnitude, making the approach very attractive.

Appendix: Post Processing for Other Priors

Return to the marginal likelihood estimate for the prior $p(\boldsymbol{\theta})$,

$$\widehat{L}^{(p)}(\boldsymbol{y}) \quad = \quad \frac{1}{N} \sum_{j=1}^{N} \frac{\ell(\,\boldsymbol{y} \,|\, \boldsymbol{\theta}_j\,)\; p(\,\boldsymbol{\theta}_j\,)}{g(\boldsymbol{\theta}_j)} \;.$$

Consider storing the $\{\boldsymbol{\theta}_j\}$ values that were simulated from the importance distribution $g(\boldsymbol{\theta})$ as part of obtaining $\widehat{L}^{(p)}(\boldsymbol{y})$, as well as storing the likelihood values $\{\ell(\,\boldsymbol{y} \,|\, \boldsymbol{\theta}_j\,)\}$ and the density evaluations $\{g(\boldsymbol{\theta}_j)\}$.

Were another prior density $\tilde{p}(\boldsymbol{\theta}) \in \mathcal{P}$ to be of interest in an IP analysis, its corresponding marginal likelihood estimate $\widehat{L}^{(\tilde{p})}$ is

$$\widehat{L}^{(\tilde{p})}(\boldsymbol{y}) \quad = \quad \frac{1}{N} \sum_{j=1}^{N} \frac{\ell(\,\boldsymbol{y} \,|\, \boldsymbol{\theta}_j\,)\; \tilde{p}(\,\boldsymbol{\theta}_j\,)}{g(\boldsymbol{\theta}_j)} \;.$$

This estimate can be obtained upon evaluating the density function $\tilde{p}(\,\boldsymbol{\theta}\,)$ for the already-stored $\{\boldsymbol{\theta}_j\}$ and combining those values with the other stored quantities from the importance sampling for $p(\,\boldsymbol{\theta}\,)$. There is no need to simulate the estimate $\widehat{L}^{(\tilde{p})}(\boldsymbol{y})$ from scratch.

This computational trick greatly aids in an IP analysis, but there are limits to its effectiveness. If the importance density $g(\boldsymbol{\theta})$ is a good one for the prior $p(\boldsymbol{\theta})$, it will also be a good one for densities $\tilde{p}(\boldsymbol{\theta})$ that are "not too far" from $p(\boldsymbol{\theta})$. Efficiency drops as the distance increases. By computing the standard deviation of the quantities $\{\ell(\,\boldsymbol{y} \,|\, \boldsymbol{\theta}_j\,)\; \tilde{p}(\,\boldsymbol{\theta}_j\,)/g(\boldsymbol{\theta}_j)\}$, the efficiency can be monitored and kept to desired levels. There is no need to run time-consuming space filling designs from scratch, as with ordinary MCMC.

The same computational trick works for the estimate of the numerator

$$\frac{1}{N} \sum_{j=1}^{N} 1_{\mathcal{A}}(\boldsymbol{x}_j, \boldsymbol{\theta}_j)\; \ell(\boldsymbol{y} \,|\, \boldsymbol{\theta}_j)\; \frac{f(\boldsymbol{x}_j)\; p(\boldsymbol{\theta}_j)}{g(\boldsymbol{x}_j, \boldsymbol{\theta}_j)} \;,$$

where quantities from the importance sampling with $p(\boldsymbol{\theta})$ can be re-used for other priors $\tilde{p}(\boldsymbol{\theta})$. Note that alternate accident scenarios, as reflected by alternate probability distributions $\tilde{f}(\boldsymbol{x})$, can also be investigated in this fashion.

References

Beer, M., Ferson, S., and Kreinovich, V. (2013), "Imprecise Probabilities in Engineering Analyses," *Mechanical Sysytems and Signal Processing*, 37, 4-29.

Berger, J. O. (1990), "Robust Bayesian Analysis: Sensitivity to the Prior," *Journal of Statistical Planning and Inference*, 25, 303-328.

Berger, J. O. (1994), "An Overview of Robust Bayesian Analysis," *Test*, 3, 5-124.

Dawid, A. P. (1982), "The Well-Calibrated Bayesian," *Journal of the American Statistical Association*, 77, 605-610.

Department of Energy (2015), "DOE Order O 452.1," from www.directives.doe.gov.

Eardley, D. et. al. (2004), "Quantification of Margins and Uncertainties," JASON Report JSR-04-330, MITRE Corporation.

Efron, B. (1986), "Why Isn't Everyone a Bayesian?" *American Statistician*, 40, 1-11.

Efron, B. (2005), "Bayesians, Frequentists, and Scientists," *Journal of the American Statistical Association*, 100, 1-5.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1982), "Lay Foibles and Expert Fables in Judgments About Risks," *American Statistician*, 36, 240-255.

Keynes, J. M. (1921), *A Treatise on Probability*, Macmillan: London.

Meeker, W. Q. and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New

York: Wiley.

Nakhleh, C. W., Webster, R. B., and Haynes, D. A. (2015), "Quantification of Margins and Uncertainties Using Imprecise Probabilities," Los Alamos National Laboratory Technical Report LA-UR-15-20764.

National Academy of Sciences (2009), "Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile," National Academies Press.

Oakley, J. E. and O'Hagan, A. (2007), "Uncertainty in Prior Elicitations: A Nonparametric Approach," *Biometrika*, 94, 427–441.

O'Neill, B. (2009), "Importance Sampling for Bayesian Sensitivity Analysis," *International Journal of Approximate Reasoning*, 50, 270-278.

Picard, R. and Williams, B. (2013), "Rare Event Estimation for Computer Models," *American Statistician*, 67, 22-32.

Seidenfeld, T. and Wasserman, L. (1993), "Dilation for Sets of Probabilities," *Annals of Statistics* 21, 1139-1154.

Sivaganesan, S. and Berger, J. O. (1989), "Ranges of Posterior Measures for Priors with Unimodal Contaminations," *Annals of Statistics*, 17. 868-889.

Vander Wiel, S. and Gore, R. (2015), "Prediction Uncertainty for Total Yield in the Secondary Validation Suite," Los Alamos National Laboratory Technical Report LA-CP 15-00832.

Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, London: Chapman and Hall.

Walley, P. (1996), "Inferences from Multinomial Data: Learning About a Bag of Marbles," *Journal of the Royal Statistical Society*, Ser. B, 58, 3-73.

Walley, P., Gurrin, L., and Burton, P. (1996), "Analysis of Clinical Data Using Imprecise Prior Probabilities," *The Statistician*, 45, 457-485.

Weatherson, B. (2002), "Keynes, Uncertainty, and Interest Rates," *Cambridge Journal of Economics*, 26, 47-62.

Yu, Y., Shih, W. J., and Moore, D. M. (2008), "Elicitation of a Beta Prior for Bayesian Inference in Clinical Trials," *Biometrical Journal*, 50, 212-223.