Simple Bayesian Updating by Cross-Entropy Minimization: Application at Leibstadt NPP

Jan Kasperek
Kernkraftwerk Leibstadt AG, Leibstadt, Switzerland, *Jan.t.Kasperek@kkl.ch*

Abstract: A novel method for performing Bayesian updates was developed for applications in Probabilistic Safety Assessment (PSA) at Leibstadt NPP (KKL). The approach is based on the conversion of generic industry priors into conjugate distributions, yielding simple formulas for posterior distribution parameters. While this general idea is established in PSA guidelines (see NUREG/CR-6823 and EPRI TR-1002936), the new approach is based on minimizing cross-entropy between the generic and approximating prior, rather than matching mean and variance (moment matching).

According to the Swiss PSA regulation ENSI-A05, KKL is required to perform Bayesian updates of component failure rates and frequencies of certain initiating events (e.g., internal fires and flooding), combining generic industry data with plant-specific evidence. Most generic data sources used in KKL PSA employ lognormal distributions, which are not conjugate to the Poisson likelihood. Thus, a numerical integration routine for Bayesian updates had previously been developed. The new approach retains the simplicity of moment matching and can easily be implemented in a spreadsheet. It simplifies code maintenance, avoids numerical issues and is more transparent. Unlike moment matching, the resulting posteriors closely resemble those obtained through numerical integration.

To evaluate the new approach, Bayesian updated component failure rates in the KKL PSA model were replaced by those derived from the new method. Results show no significant risk impact: mean component failure rates decreased by an average of 2.5 percent and total core damage frequency increased by 3.7 percent. Posterior credible intervals tend to be wider. The largest deviations from the numerical posteriors occur at low percentiles, but are considerably smaller than with moment matching. A complementary study on synthetic data - covering realistic prior-evidence combinations for KKL PSA applications - confirmed these results.

The use of Gamma distributions also resolves a mathematical issue when generic priors are split into multiple distributions (e.g., to match component boundaries in the PSA model). This also addresses an issue raised during the 2014 IAEA IPSART mission at KKL.

Based on properties of Beta and Gamma distributions, the method can be extended to a variety of prior distribution types as well as probability parameters (Binomial likelihood).

The new method will serve as a basis for the next Bayesian updates of KKL component failure rates and is intended to be used for all future Bayesian updates in KKL PSA.

Keywords: Bayesian, Uncertainty, Component Reliability, Initiating Event Frequency

1. INTRODUCTION

According to the national PSA regulation ENSI-A05 [1], nuclear power plants in Switzerland are required to perform Bayesian updates of component reliability data, frequencies of internal initiating events, as well as internal fire and flooding frequencies.

In many cases, Bayesian updating is conducted numerically, since most generic probability distributions used in KKL PSA are not conjugate to the relevant likelihood functions. Improving on some shortcomings of the techniques used before, a method for numerical integration was developed previously at KKL [2], together with an implementation in the programming language Ruby.

However, there are still drawbacks to numerical methods. Calculations can occasionally fail and may require a significant amount of time to terminate. Maintaining the necessary scripts requires time and programming skills. Moreover, numerical algorithms are essentially “black boxes” to the user. Originally motivated by an overhaul of the KKL component reliability database and the necessary scripts to automate the corresponding report, this prompted an exploration of alternatives to streamline the Bayesian updating process.

To simplify Bayesian updating workflows, conjugate priors offer a practical alternative. In cases where generic industry priors are non-conjugate, a well-known approach is to convert them into conjugate types by matching mean and variance [3, 4, 5]. However, this method is known to produce priors that differ greatly from generic industry data. Accordingly, the resulting updated (posterior) distributions can deviate significantly from the numerically calculated posteriors [3, section 6.2.2.7].

KKL has thus developed a new approach based on approximate minimization of cross-entropy and Kullback-Leibler divergence. Section 2 introduces the new approach to Bayesian updating with a discussion of its properties, including comparisons to moment matching and the numerical method currently implemented at KKL. Application in KKL PSA is discussed in Section 3, and concluding remarks are given in Section 4. Detailed mathematical derivations are found in the Appendix.

2. METHODOLOGY

2.1. Bayesian Updating

In PSA, generic industry data and plant-specific experience are commonly combined using Bayes’ theorem. Let θ be a continuous random variable, directly representing a probability, failure rate, or other frequency. $f(\theta)$ is then the probability density assigned to θ before observing any plant-specific data (prior density). The evidence of interest for PSA applications is mostly of the form “k events in T units of time” or “k events in N trials”, where only k is interpreted as a realization of a random variable K. These observed data are commonly modeled with the Poisson and Binomial distribution, respectively. The probability density of θ after observing plant data (i.e., the posterior density) is then given by

$$f(\theta \mid K = k) = \frac{L(\theta; k) \cdot f(\theta)}{\int_0^\infty L(\theta; k) \cdot f(\theta)\, d\theta} \tag{1}$$

where $L(\theta; k)$ denotes the likelihood function. This is the probability of observing exactly k events within a certain time or number of trials, but interpreted as function of θ with k fixed.

In general, the posterior $f(\theta \mid K = k)$ cannot be written in closed form and must be computed numerically instead. For PSA applications, there are two important exceptions: Combining a Poisson likelihood with a Gamma prior, and a Binomial likelihood (with fixed number of trials N) with a Beta prior. These are called conjugate models, since the posterior distribution is of the same type as the prior.

Specifically, a $Gamma(\alpha, \beta)$ prior (shape-rate parametrization) combined with Poisson likelihood (k events within T time units) yields a $Gamma(\alpha + k, \beta + T)$ posterior. Pairing a $Beta(\varrho, \xi)$ prior with a binomial likelihood yields a $Beta(\varrho + k, \xi + N - k)$ posterior distribution. Thus, using only these combinations for Bayesian updating does not require numerical methods.

However, generic industry data are often given in the form of non-conjugate distributions, most commonly of the lognormal type. A method to circumvent numerical techniques is to approximate a non-conjugate generic distribution with a conjugate one. A simple and well-known approach to do so is by equating their mean and variance ("moment matching", see [3, 4, 5]). However, this approximation technique can be highly inaccurate, leading to posterior distributions that differ significantly from those calculated with the non-conjugate prior. For an example, see [3, section 6.2.2.7].

This prompted the development of a new method based on the same idea of fitting conjugate priors to non-conjugate generic distributions. The goal is to retain the simplicity of moment matching while obtaining posteriors that are more closely aligned with the results from numerical Bayesian updates.

2.2. Prior Conversion by approximate Cross-Entropy Minimization

An alternative technique to matching mean and variance of probability distributions is to minimize a form of "distance" between them. Instead of matching only "summaries", the complete distributions are taken into account to determine the optimal approximation. A widely adopted measure of such a "distance" with strong foundations in information theory and probability is the Kullback-Leibler divergence [6], which is used in different fields such as Machine Learning [7] and statistical model selection [8]. Related information-theoretic concepts can also be used to construct prior distributions [3]. The conversion of Lognormal distributions into Gamma by minimizing cross-entropy has also successfully been applied in other engineering fields, see [9] for an application in wireless communications.

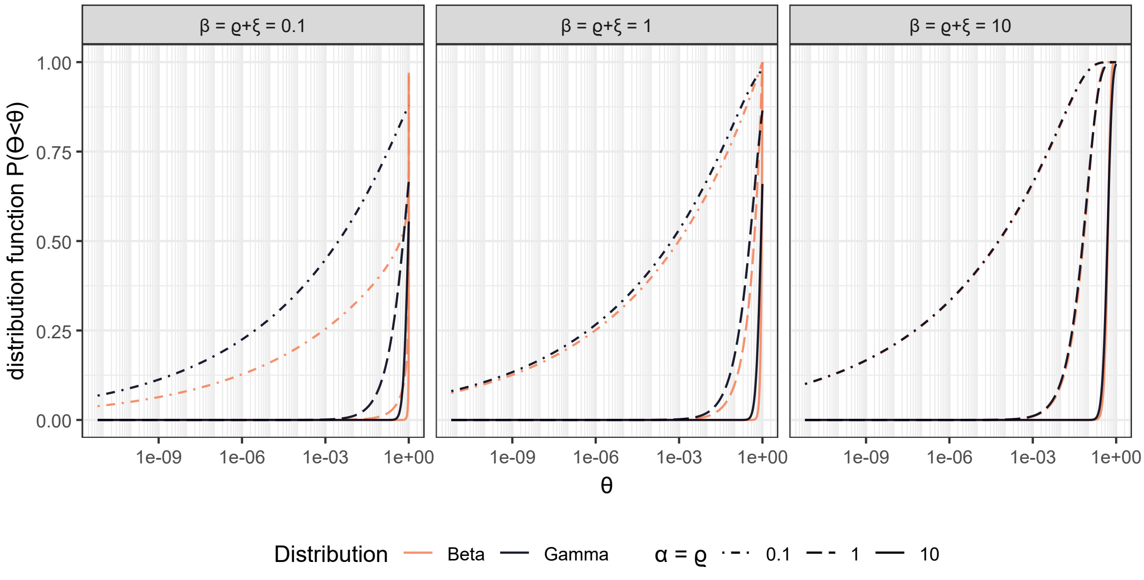
Let f denote the original (generic) and g the desired conjugate prior distribution. Let f and g be continuous distributions and suppose there is no range to which g assigns zero probability, while f assigns non-zero probabilities (i.e., the support of f is a subset of Ω, the support of g). The Kullback-Leibler divergence (KLD) is then

$$\begin{aligned} D_{KL}(f \parallel g) &= \int_\Omega f(\theta) \cdot ln\left(\frac{f(\theta)}{g(\theta)}\right) d\theta \\ &= \int_\Omega f(\theta) \cdot ln(f(\theta))\, d\theta - \int_\Omega f(\theta) \cdot ln(g(\theta))\, d\theta \end{aligned} \tag{2}$$

In the second expression, the first term is the negative of the entropy [10] of the generic distribution while the second integral is called the cross-entropy. Since the generic distribution parameters are fixed values, the entropy is a constant and can thus be ignored in the optimization. Hence, minimization of Kullback-Leibler divergence is equivalent to minimization of cross-entropy for this application.

Note that Kullback-Leibler divergence and cross-entropy are not symmetric and are thus not distances in the usual sense (i.e., $D_{KL}(f \parallel g) \neq D_{KL}(g \parallel f)$). Importantly, $D_{KL}(f \parallel g)$ is infinite (or undefined) when the approximating distribution g is of the Beta type while f is a Gamma or lognormal distribution. In such cases, a workaround based on similarities between Beta and Gamma distributions can be used. Figure 1 shows the well-known result that $Beta(\varrho, \xi)$ distributions quickly approach $Gamma(\varrho, \varrho + \xi)$ distributions with growing ξ and ϱ fixed.

Figure 1: Convergence of Beta to Gamma distribution



Thus, if f assigns positive densities to values > 1, the first step is to find an approximating Gamma distribution, which can then be transformed into a Beta by letting $\varrho = \alpha$ without much additional information loss (provided the rate parameter β is not too small).

The focus is thus to find the parameters of a Gamma distribution which minimize the cross-entropy, solving the optimization problem

$$\begin{aligned} &\max_{\alpha,\beta} \int_0^\infty f(\theta) ln(g_{\alpha,\beta}(\theta))\, d\theta \\ &\text{such that} \quad Mean_{generic} = \frac{\alpha}{\beta} \end{aligned} \tag{3}$$

where f can be of different generic distribution types, provided it is continuous and has no support on the interval $(-\infty, 0)$. As demonstrated in the Appendix, the uniquely optimal shape parameter α is found by solving

$$ln(\alpha) - \psi_0(\alpha) \stackrel{!}{=} ln(\mathbb{E}_f(\theta)) - \mathbb{E}_f(ln(\theta)) \tag{4}$$

before calculating $\beta = \alpha/\mathbb{E}_f(\theta)$ to retain the generic mean.

The notation in Equation 4 is:

- $\psi_0(\alpha)$: digamma function, i.e., the first derivative of the log-transformed Gamma function [11]
- $\mathbb{E}_f(\theta)$: mean of the generic distribution
- $\mathbb{E}_f(ln(\theta))$: mean of $ln(\theta)$, according to the generic distribution (i.e., $\int_0^\infty ln(\theta) \cdot f(\theta)\, d\theta$)

Note that solving Equation 4 is asymptotically equivalent to maximum likelihood estimation [7, 8]; estimating Gamma parameters by maximum likelihood consists of solving

$$ln(\alpha) - \psi_0(\alpha) \stackrel{!}{=} ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \frac{1}{n}\sum_{i=1}^{n} ln(x_i) \tag{5}$$

before finding $\beta = \alpha/\bar{x}$ to match the empirical mean $\bar{x}$. The x_i are independent realizations of a random variable with distribution f. When simulating a large number of samples n from a suitable generic distribution, it can thus be observed that Equation 5 converges to Equation 4 with growing n, due to the law of large numbers. This effect can also be seen from the simulation study conducted in [9].

The exact solution to Equation 4 can only be found numerically (e.g., with bisection). However, there exist simple expressions for the bounds of the digamma function which can be used to derive sufficiently accurate approximations for PSA applications. As is shown in the Appendix, the shape parameter α of the approximating Gamma prior is bounded by

$$\frac{1}{2[ln(\mathbb{E}_f(\theta)) - \mathbb{E}_f(ln(\theta))]} < \alpha < \frac{1}{ln(\mathbb{E}_f(\theta)) - \mathbb{E}_f(ln(\theta))} \tag{6}$$

The upper bound can be sharpened with more complicated expressions, but the lower bound appears to be sufficiently close for typical PSA use cases. Consider the approximation of a generic lognormal distribution (by far the most frequent case in KKLPSA), where the bounds are

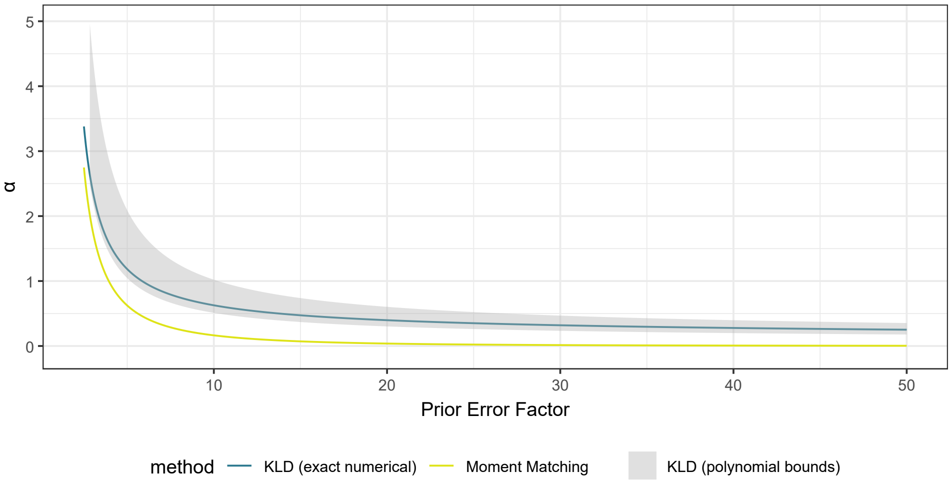
$$\frac{1}{\sigma^2} < \alpha < \frac{2}{\sigma^2} \tag{7}$$

or, in another common parametrization with the error factor EF, defined as the ratio between 95th percentile and median,

$$\frac{2.71}{ln^2(EF)} < \alpha < \frac{5.41}{ln^2(EF)} \tag{8}$$

Figure 2 shows the numerical solution to Equation 4 together with the simple polynomial bounds from Equation 8 as function of the error factor from the generic lognormal distribution. Particularly for smaller values of EF, the lower bound of Equation 8 appears to be a sufficiently accurate approximation to the optimum:

Figure 2: Approximating prior shape parameter as function of generic error factor



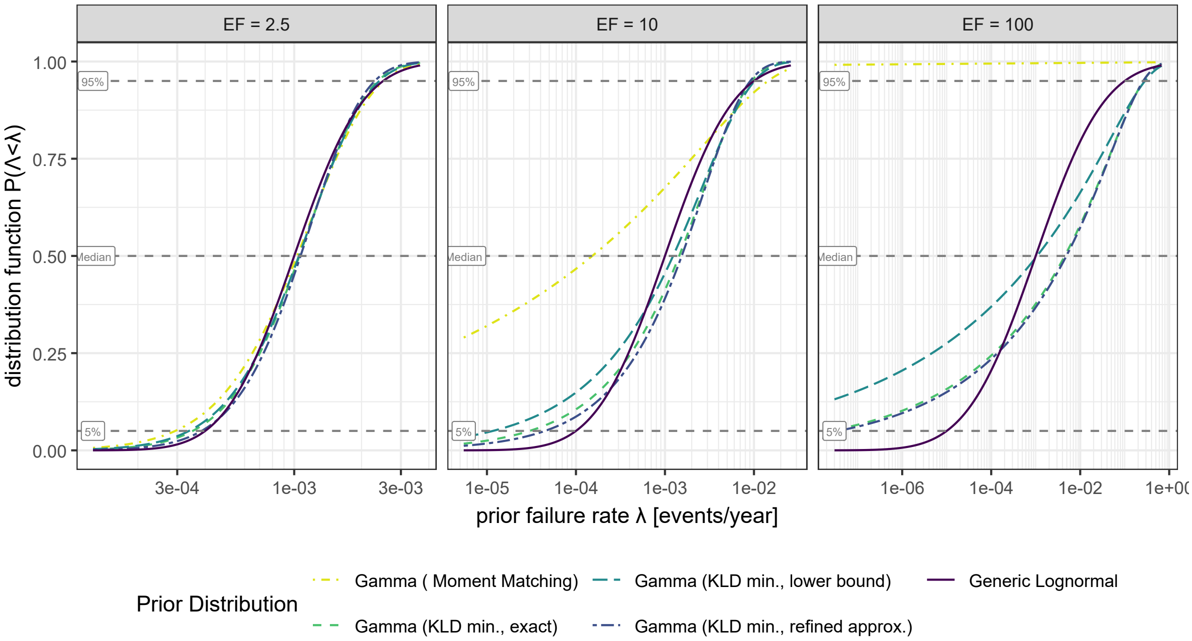
It also becomes clear that the approximating α obtained by matching mean and variance is far away from the information-theoretic optimum.

Refinements based on the polynomial bounds may be considered, especially for large error factors. A simple way to do so is by a function which converges to the lower bound as $EF \to 1$ and converges to the upper bound as $EF \to \infty$. It could also be stipulated that when the cross-entropy between generic distribution and Gamma is equal between the two bounds, α should be their arithmetic mean, $1.5/\sigma^2$. A relatively simple function which fulfills these conditions is given by

$$\alpha \approx \frac{5.41ln(EF) + 11.1}{ln^3(EF) + 4.1ln^2(EF)} \tag{9}$$

To assess the accuracy of Gamma approximations to typical lognormal generic distributions, Figure 3 shows results of different methods for ranges of error factors typically encountered in PSA applications (although error factors greater than 30 are very rare in KKL PSA). The generic mean has no influence on the quality of the approximation and was thus chosen to match the example given in [3, sec. 6.2.2.7]. Therefore, the example with $EF = 10$ reproduces [3, Fig. 6.18]. The Gamma prior obtained from moment matching quickly diverges from the generic distribution when the error factor increases, while both approximate (lower bound) and exact cross-entropy minimization yield distributions which stay close to the generic lognormal one. For large error factors (greater than 10 or 30), Equation 9 also clearly outperforms the lower bound from Equation 8. All Gamma approximations have the tendency to assign higher probabilities to small values, especially when the error factor is large.

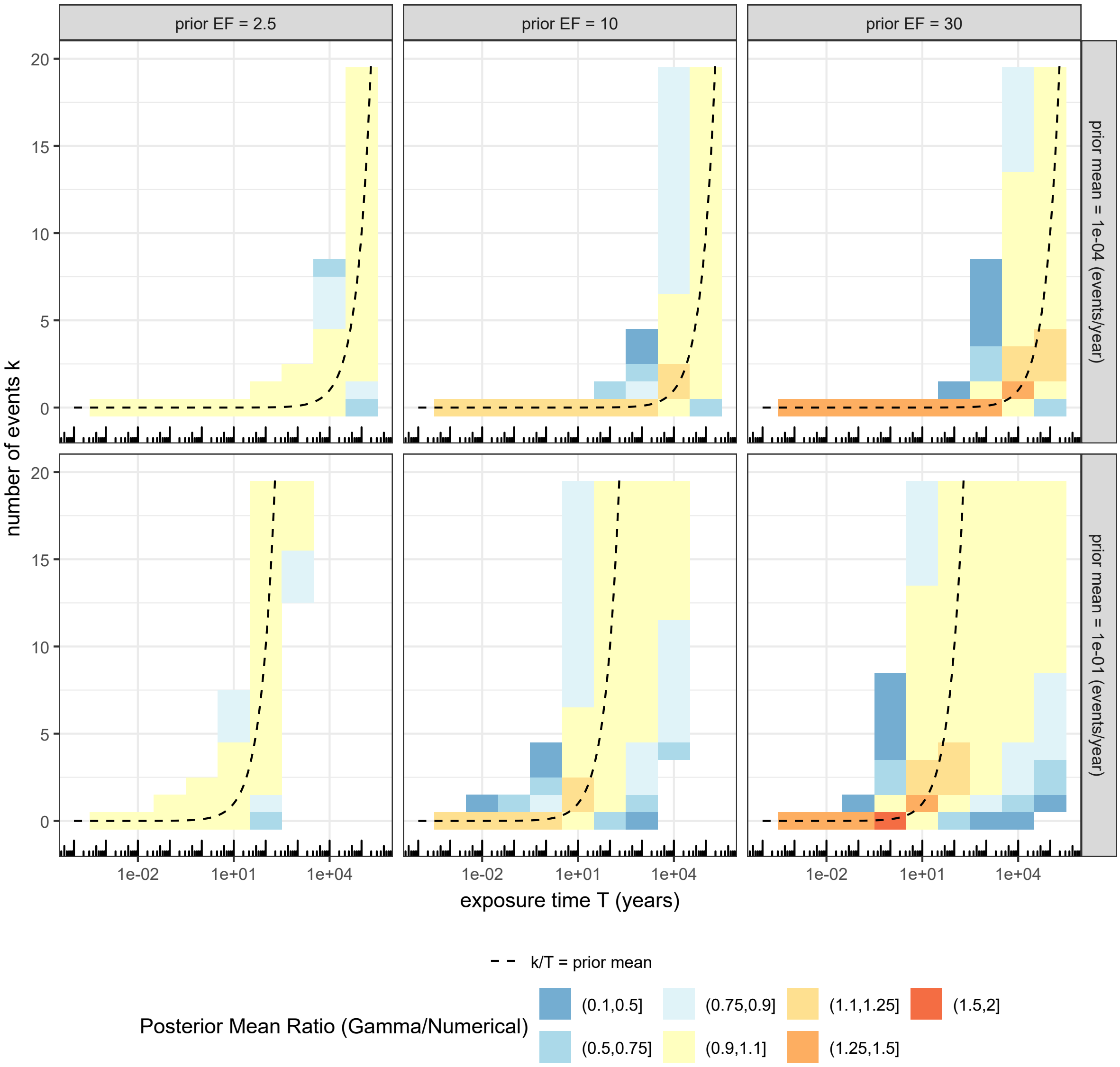
Figure 3: Comparison of Gamma approximations to generic Lognormal distributions



To systematically test and compare the currently implemented with the newly developed approach, synthetic data were generated to conduct Bayesian updates with. These data consist of Lognormal priors with different means and error factors, as well as combinations of event counts and exposure times. The ranges were chosen to represent realistic cases for KKL PSA. The number of events k was limited to 20, since with increasing plant-specific experience, the influence of the prior vanishes and the results of different methods will converge. To exclude extremely unrealistic combinations of priors and plant-specific evidence, the prior predictive criterion described in [3, section 6.2.3.5] was used to filter out cases with a very high discrepancy between prior and evidence (using a cutoff probability of 0.001).

The resulting posterior mean ratios between the new (lower bound on α) and the numerical integration technique are shown in Figure 4.

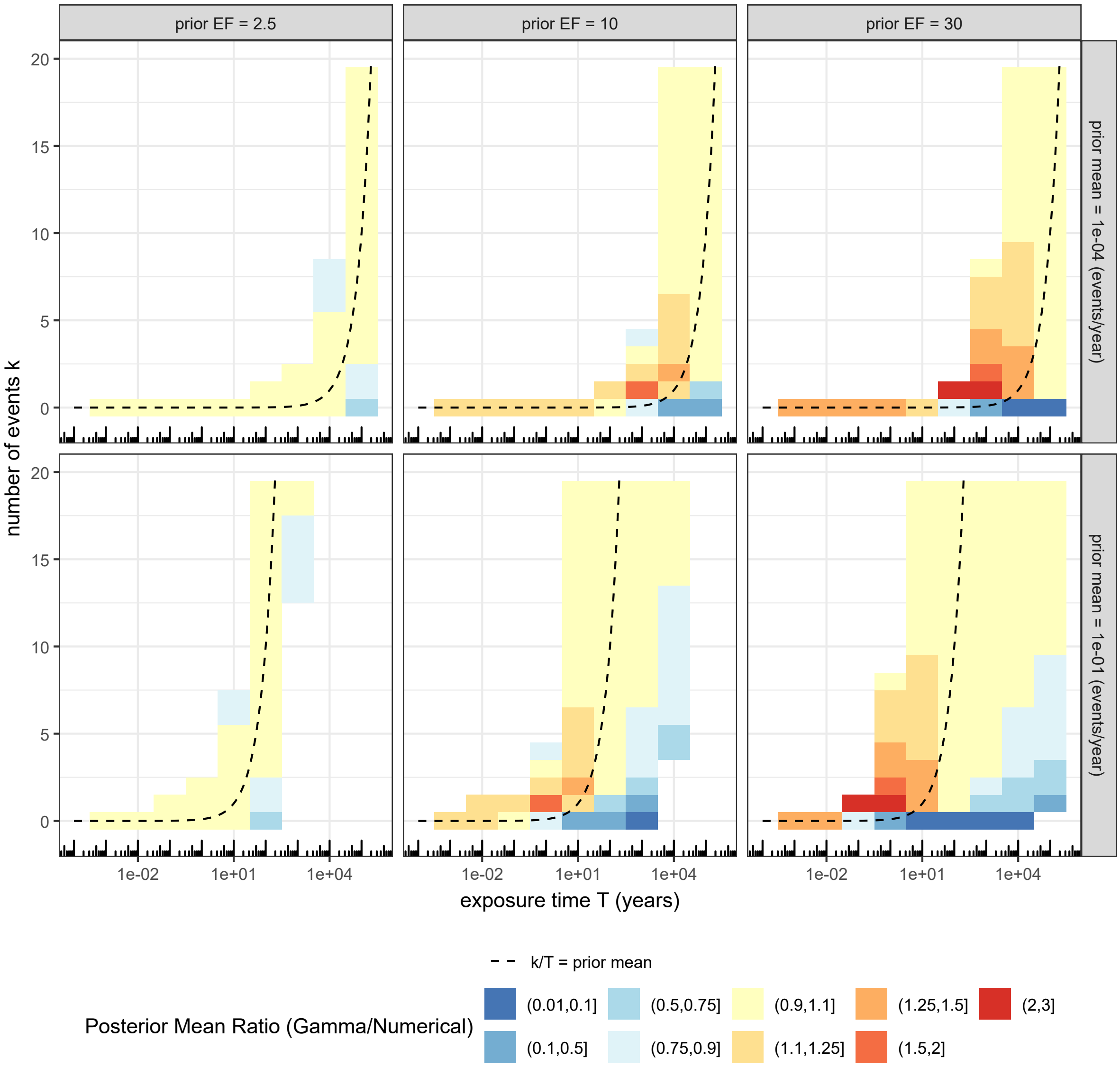
Figure 4: Comparison of Posterior Means: Gamma (KLD, Lower Bound) vs Numerical



In most realistic combinations of generic parameters and plant-specific evidence, the posterior means deviate by less than 10%. When the generic error factor is large and few plant-specific events have been observed, the new method tends to deliver more pessimistic results (unless the exposure time is very large). When there is a large discrepancy between prior mean failure rate and the plant-specific observed failure rate, the new method tends to be more optimistic. Note however that the evaluated combinations were filtered very leniently; a common cutoff probability is 0.05, not 0.001 [3]. In common practice, many regions in which the new method delivers more optimistic posterior mean values would be considered serious discrepancies between generic data and evidence.

Figure 5 shows the same comparison with Gamma posteriors obtained from moment matching. Larger deviations between the posterior means are more common when using the moment matching approach, especially when little plant-specific experience is available.

Figure 5: Comparison of Posterior Means: Gamma (Moment Matching) vs Numerical



Finally, the prior conversion for selected non-conjugate combinations of generic distribution types and likelihoods is summarized in Table 1. When the likelihood is Poisson, the third column gives the lower bound on the Gamma shape parameter α from Equation 8. For the Binomial likelihood, it shows the approximate lower bound on the first shape parameter ϱ of the Beta distribution (the unusual notation was chosen to avoid confusion with the Gamma distribution). The last column provides important notes on the combination of prior and likelihood. "Support mismatch" means that the generic distribution is defined over a different range than the parameter to be updated (i.e., $(0, \infty)$ for a probability). Some combinations of generic distribution and likelihood imply a mismatch in units (e.g., a Beta prior on the rate parameter of the Poisson distribution). Still, such combinations can occur in practice. In KKL PSA component reliability for instance, some generic Beta distributions (failure probabilities on demand) are converted into failure rates to match the time-based reliability modeling of the plant (counting exposure times instead of demands). Furthermore, such combinations can be useful for purposes other than Bayesian updating. For instance, widely used generic Error Probabilities in Human Reliability Analysis assign lognormal uncertainty distributions [12]. To avoid additional numerical calculations by PSA software to limit the lognormal to the range $(0, 1)$, the entry for the Lognormal & Binomial case can be used instead to convert the uncertainty distribution into a proper probability without significantly altering its shape.

Table 1: Conversion of Common Generic Distributions into Conjugate Priors

Generic Prior	Likelihood	(First) Shape Parameter, Lower Bound	Remark
$Beta(\varrho, \xi)$	Poisson	ϱ	mismatch of support & units
$Gamma(\alpha, \beta)$	Binomial	α	mismatch of support & units
$LogNormal(Mean,\ EF)$	Poisson	$2.71/ln^2(EF)$	for large EF, see Equation 9
$LogNormal(Mean,\ EF)$	Binomial	$2.71/ln^2(EF)$	mismatch of support & units
$LogUniform(a, b),\ 0 < a < b \leq 1$	Binomial	$1/(2ln(b-a) - 2ln(ln(b/a)) - ln(a) - ln(b))$	mismatch of support
$LogUniform(a, b),\ 0 < a < b$	Poisson	$1/(2ln(b-a) - 2ln(ln(b/a)) - ln(a) - ln(b))$	mismatch of support

The second parameter of the conjugate prior is then found by matching the generic mean and, if necessary, multiplying it by a scalar for conversion into the correct units. Bayesian updating is then conducted with the well-known formulae for conjugate priors.

3. APPLICATION AT KKL

To assess the effects of applying the new method in KKL PSA, a Bayesian update with the whole component reliability database was performed with both the current and the new method, and mean core damage frequency was calculated using the most recent version of the fully integrated KKL PSA model to assess whether introducing the new method has a significant effect on model results.

Posterior mean failure rates calculated with the new approach were, on average, 2.5% lower. The median ratio between posterior means was equal to 1, so 50% of mean component failure rates increased and 50% decreased. 5% of failure rates decreased by more than 60%, and 5% increased by more than 45%. The largest relative changes in the posterior distributions were observed at lower quantiles, where the 5th percentile decreased, on average, by more than half and only 5% increased by more than 2%. The change in posterior 95th percentiles was less pronounced, with an average increase of 14%.

Mean core damage frequency increased by 3.7% when using component failure rates calculated with the new method. This is likely because posterior means for the most important components slightly increased. Overall, replacing the numerical integration method by the new one did not significantly alter the PSA results. Since the prior distributions and evidence are similar to other Bayesian updates, this is expected to apply to other KKL PSA use cases as well. The new method is thus intended for use in all future Bayesian updates to facilitate simplified and robust workflows, accessible to all analysts.

Note that the differences between the posteriors are only partly due to the changes in Bayesian updating methodology itself, but also due to an additional preprocessing step applied to the generic distributions. In KKL PSA, component boundaries are often more granular than the boundaries in generic data, such that multiple KKL PSA components are contained in one generic failure rate. In such cases, generic failure rates are split between subcomponents using a fixed ratio. While this is trivial for the prior mean, deriving the corresponding prior parameters is not, since sums of multiple lognormally distributed random variables do not yield another lognormal distribution. As was also noted by a PSA expert during the 2014 IPSART mission at KKL, such splitting of generic distributions is possible in the Gamma case, assuming the failure rates of subcomponents λ_i are independent and have the same rate parameter β:

$$\begin{aligned} \lambda_i &\sim Gamma(\alpha_i,\ \beta) \\ \implies \sum_{i=1}^{n} \lambda_i &\sim Gamma\left(\sum_{i=1}^{n} \alpha_i,\ \beta\right) \end{aligned} \tag{10}$$

where $i = \{1, ..., n\}$ is an index for the subcomponents. An exactly split prior is thus obtained by multiplying the shape parameter α of the total generic failure rate λ with the split factor r_i. This is an additional advantage of converting other distribution types (most notably lognormal) into Gamma. The largest deviations between the new and currently implemented Bayesian updating techniques occurred in cases where generic failure rates were split, so especially the changes in posterior quantiles may be less extreme when factoring out the effect of preprocessing many generic distributions.

4. CONCLUSION

A new approach to Bayesian updating for PSA applications has been developed at KKL. The method consists of converting generic industry priors into conjugate distributions by approximately minimizing cross-entropy and Kullback-Leibler divergence and leveraging relationships between Beta and Gamma distributions. It circumvents complex and potentially unstable numerical calculations without significantly changing the results. Since it produces posterior distributions that more closely resemble those obtained from numerical calculations, the new approach presents an improvement over the similar and well-known method of matching mean and variance to obtain conjugate priors. This is demonstrated through applications to both real plant reliability data and synthetic data covering realistic ranges for KKL PSA applications. Due to the simplified workflow, transparency, and independence from any specific programming language, the methodology is already employed for component reliability data and ad-hoc analyses, and will likely be used for all future Bayesian updates in KKL PSA, subject to regulatory acceptance.

References

[1] Swiss Federal Nuclear Safety Inspectorate (ENSI). "*Probabilistic Safety Analysis (PSA): Quality and Scope*", ENSI-A05/e, (2019).

[2] A. Ayoub, V. Ariu, and O. Nusbaumer. "*A fast and robust Bayesian update of components failure rates in a nuclear power plant*", Progress in Nuclear Energy, 118, pp. 1030–67, (2020).

[3] U.S. Nuclear Regulatory Commission. "*Handbook of Parameter Estimation for Probabilistic Risk Assessment*", NUREG/CR-6823, (2003).

[4] U.S. Nuclear Regulatory Commission. "*Applying Statistics*", NUREG-1475, (2011).

[5] Electrical Power Research Institute. "*Reliability and Preventive Maintenance: Balancing Risk and Reliability*", EPRI TR-1002936, (2002).

[6] S. Kullback and R. A. Leibler. "*On Information and Sufficiency*", The Annals of Mathematical Statistics, 22 [1], pp. 79–86, (1951).

[7] I. Goodfellow, Y. Bengio, and A. Courville, "*Deep Learning*", MIT Press, 2016, Cambridge.

[8] H. Akaike. "*Information Theory and an Extension of the Maximum Likelihood Principle*", in *Proceedings of the Second International Symposium on Information Theory*, pp. 267–281, (1973).

[9] S. Dang et al. "*On Optimality of Gamma Approximation for Lognormal Shadowing Models*", IEEE Antennas and Wireless Propagation Letters, 22 [5], pp. 1099–1103, (2023).

[10] C. E. Shannon. "*A Mathematical Theory of Communication*", Bell System Technical Journal, 27, pp. 379–423, (1948).

[11] M. Abramowitz and I. A. Stegun, "*Handbook of Mathematical Functions*", Dover, 1972, New York.

[12] U.S Nuclear Regulatory Commission. "*Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Appplications*", NUREG/CR-1278, SAND80-0200, (1983).

[13] A. Erdélyi et al., "*Higher Transcendental Functions*", McGraw-Hill, 1953, New York.

APPENDIX: MATHEMATICAL DERIVATION

Prerequisites

The following functions are referenced in the derivations that follow:

Hurwitz Zeta Function [13, eq. 1.10.1 and 1.10.7]:

$$\begin{aligned}\zeta(s,a) &= \sum_{n=0}^{\infty}\frac{1}{(n+a)^s} \\ &= \frac{1}{2a^s} + \frac{a^{1-s}}{s-1} + 2\int_0^{\infty}\frac{sin(s\,tan^{-1}(t/a))}{(a^2+t^2)^{s/2}(exp(\pi 2t)-1)}dt \quad ,\forall\, s>1\end{aligned} \tag{11}$$

Polygamma function [11, eq. 6.4.1, 6.3.21 and 6.4.10]:

$$\psi_n(z) = \frac{d^{n+1}ln(\Gamma(z))}{dz^{n+1}} \quad , n\in\{0,1,2,...\}$$

$$\psi_n(z) = \begin{cases} ln(z) - \dfrac{1}{2z} - 2\displaystyle\int_0^{\infty}\frac{t}{(t^2+z^2)[exp(2\cdot\pi t)-1]}dt & \text{for } n=0 \\ (-1)^{n+1}\cdot n!\cdot\zeta(n+1,z) & \text{for } n>0\end{cases} \tag{12}$$

Deriving Gamma Prior Parameters

Plugging the Gamma log-density

$$ln(g_{\alpha,\beta}(\theta)) = \alpha ln(\beta) - ln(\Gamma(\alpha)) + (\alpha-1)ln(\theta) - \beta\theta$$

into the objective from Equation 3,

$$\int_0^{\infty} f(\theta)ln(g_{\alpha,\beta}(\theta))d\theta = (\alpha ln(\beta) - ln(\Gamma(\alpha)))\int_0^{\infty} f(\theta)d\theta + (\alpha-1)\int_0^{\infty} ln(\theta)f(\theta)d\theta - \beta\int_0^{\infty}\theta f(\theta)d\theta \tag{13}$$

Analyzing the integrals on the right-hand side:

1. $f(\theta)$ must be a probability density on $\omega\subseteq(0,\infty)$, so the first integral equals 1.
2. By definition, the second integral is the expected value of the logarithm of the generic distribution, $\mathbb{E}_f(ln(\theta))$.
3. The third integral is the mean of the generic prior distribution, $\mathbb{E}_f(\theta)$.

Thus,

$$\begin{aligned}\int_0^{\infty} f(\theta)ln(g_{\alpha,\beta}(\theta))d\theta &= -h(f,g_{\alpha,\beta}) \\ &= \alpha\cdot ln(\beta)\ -\ ln(\Gamma(\alpha)) +\ (\alpha-1)\cdot\mathbb{E}_f(ln(\theta)) - \beta\cdot\mathbb{E}_f(\theta)\end{aligned} \tag{14}$$

Plugging in the constraint from Equation 3 (i.e., $\beta = \alpha/\mathbb{E}_f(\theta)$),

$$-h(f,g_\alpha) = \alpha\cdot[ln(\alpha) - ln(\mathbb{E}_f(\theta)) - 1]\ -\ ln(\Gamma(\alpha)) +\ (\alpha-1)\cdot\mathbb{E}_f(ln(\theta)) \tag{15}$$

Since the constraint is now included and α is the only remaining unknown, solving the optimization problem from Equation 3 is achieved by setting the first derivative of Equation 15 to 0. The first two derivatives are

$$\begin{aligned} \Big[-h(f,g_\alpha)\Big]' &= ln(\alpha) - \psi_0(\alpha) - ln(\mathbb{E}_f(\theta)) + \mathbb{E}_f(ln(\theta)) \\ \Big[-h(f,g_\alpha)\Big]'' &= \frac{1}{\alpha} - \psi_1(\alpha) \end{aligned} \tag{16}$$

The second derivative is negative for all $\alpha > 0$, which follows from plugging in the second case of Equation 12 for $\psi_1(\alpha)$ and verifying that the integral term from Equation 11 must be positive:

$$\begin{aligned} \psi_1(\alpha) &= \zeta(2,\alpha) \\ &= \frac{1}{2\alpha^2} + \frac{1}{\alpha} + 2\int_0^\infty \frac{sin(2\cdot tan^{-1}(t/\alpha))}{(\alpha^2+t^2)(exp(\pi 2t)-1)}dt \\ &> \frac{1}{\alpha}, \quad \forall\, \alpha > 0 \end{aligned} \tag{17}$$

This ensures that setting the first derivative to 0 yields an α which minimizes the cross-entropy. Setting the first derivative to 0 and rearranging yields the objective from Equation 4:

$$ln(\alpha) - \psi_0(\alpha) \overset{!}{=} ln(\mathbb{E}_f(\theta)) - \mathbb{E}_f(ln(\theta))$$

Approximations of α can be derived by considering simple bounds of $\psi_0(\alpha)$, which are quickly derived from plugging in the representation for ψ_0 given by Equation 12 into the objective:

$$\frac{1}{2\alpha} + 2\int_0^\infty \frac{t}{(t^2+\alpha^2)[exp(2\cdot\pi t)-1]}dt \overset{!}{=} ln(\mathbb{E}_f(\theta)) - \mathbb{E}_f(ln(\theta)) \tag{18}$$

The integral term is positive and converges to 0 for growing α. Hence, a simple lower bound is 0. To obtain a simple upper bound of the integral, observe that

$$\begin{aligned} & \frac{t}{(t^2+\alpha^2)[exp(2\pi t)-1]} < \frac{t}{(t^2+\alpha^2)2\pi t} \quad , \forall\, t > 0 \\ \implies & 2\int_0^\infty \frac{t}{(t^2+\alpha^2)[exp(2\pi t)-1]}dt < \frac{1}{\pi}\int_0^\infty \frac{1}{t^2+\alpha^2}dt \\ & \qquad = \frac{1}{\pi}\left[\frac{1}{\alpha}tan^{-1}(t/\alpha)\right]_0^\infty \\ & \qquad = \frac{1}{2\alpha} \end{aligned}$$

Thus, bounds for the shape parameter α are obtained as

$$\frac{1}{2\alpha} < ln(\mathbb{E}_f(\theta)) - \mathbb{E}_f(ln(\theta)) < \frac{1}{\alpha} \tag{19}$$

from which Equation 6 follows. Tighter bounds can be found, but these will involve more complicated expressions and/or transcendental equations (which again require numerical solutions).

When the generic distribution is of the Beta type, the derivation can be simplified given that ξ is usually large and ϱ relatively small. In this way, the density is mainly concentrated over the small event probabilities, as is the typical case in PSA. Plugging the mean of the generic Beta distribution and mean of $ln(\theta)$ (given by $\psi_0(\varrho) - \psi_0(\varrho+\xi)$) into Equation 4 yields

$$ln(\alpha) - \psi_0(\alpha) \overset{!}{=} ln(\varrho) - ln(\varrho+\xi) + \psi_0(\varrho+\xi) - \psi_0(\varrho) \tag{20}$$

It then follows from Equation 12 that $ln(\varrho+\xi) \approx \psi_0(\varrho+\xi)$. Hence,

$$ln(\alpha) - \psi_0(\alpha) \approx ln(\varrho) - \psi_0(\varrho) \tag{21}$$

which leads to $\alpha \approx \varrho$.