

Bayesian Modeling of Population Variability – Practical Guidance and Pitfalls

Dana L. Kelly^{a*}, Corwin L. Atwood^b

^aIdaho National Laboratory, Idaho Falls, ID, USA

^bStatwood Consulting, Silver Spring, MD, USA

Abstract: With the advent of easy-to-use open-source software for Markov chain Monte Carlo (MCMC) simulation, hierarchical Bayesian analysis is gaining in popularity. This paper presents practical guidance for hierarchical Bayes analysis of typical problems in probabilistic safety assessment (PSA). The guidance is related to choosing parameterizations that accelerate convergence of the MCMC sampling and to illustrating the potential sensitivity of the results to the functional form chosen for the first-stage prior. This latter issue has significant ramifications because the mean of the average population variability curve (PVC) from hierarchical Bayes (or the mean of the point estimate distribution from empirical Bayes) can be very sensitive to this choice in cases where variability is large. Numerical examples are provided to illustrate the issues discussed.

Keywords: Bayesian inference, Hierarchical Bayes, Population Variability

1. INTRODUCTION

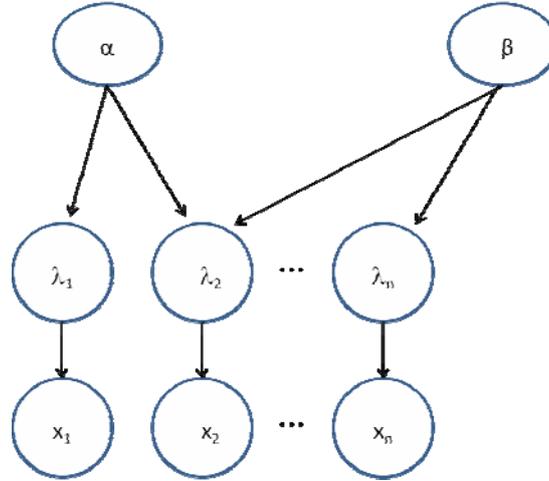
It has long been recognized by the risk assessment community that population variability can be an important consideration in Bayesian inference. The earliest efforts in this direction by the risk assessment community were probably those of Stan Kaplan, who developed his so-called two-stage Bayesian approach to the problem (1) which is a first cousin to modern hierarchical Bayesian analysis. This approach was applied in the risk assessments performed for a number of U. S. utilities. Because of the numerical difficulties associated with the integration required for hierarchical Bayes, most analyses to date have followed either Kaplan's two-stage approach, which reduces the required integration to arithmetic summation by discretizing the distributions, or they have taken the approximate route of parametric empirical Bayes. The empirical Bayes method estimates the distribution representing population variability by choosing a conjugate distributional form (e.g., gamma population variability distribution for Poisson data), which allows the marginal likelihood function to be written in closed form (e.g., as a product of gamma-Poisson distributions). It then finds the parameter values that maximize the logarithm of the marginal likelihood. Thus, integration is replaced by optimization. This approach has been employed widely in analyses performed for the U. S. Nuclear Regulatory Commission, for example (2). With the advent of the freely available and easy-to-use (and now open-source) software package WinBUGS for Markov chain Monte Carlo (MCMC) simulation (3), hierarchical Bayesian analysis is gaining in popularity. This paper presents practical guidance for hierarchical Bayesian analysis of typical PRA problems. The guidance is related to choosing parameterizations that accelerate convergence of the MCMC sampling and to illustrating the potential sensitivity of the results to the functional form chosen for the first-stage prior. This latter issue has significant practical ramifications, because the mean of the average population variability curve (PVC) from hierarchical Bayes (or the mean of the point estimate distribution from empirical Bayes) can be very sensitive to this choice in cases where variability is large. Numerical examples are provided to illustrate the issues discussed.

* Dana.Kelly@inl.gov

2. HIERARCHICAL BAYES

As discussed by Siu and Kelly (4), treatment of variability that can exist among sources of data is important if total uncertainty, including population variability, is to be properly represented by the resulting posterior distribution. We first review the general framework of hierarchical Bayes. We then illustrate this framework with two numerical examples. The problem is perhaps most simply illustrated with a Bayesian network, as shown below. Here, for each source, one observes x_i events in time t_i . A Poisson model is used for each source, and the parameter of interest is then the Poisson rate for each source, denoted λ_i . The rate for each source is assumed to be randomly selected from the population variability distribution, which might be gamma and will itself depend on parameters α and β that specify the distribution.

Figure 1: Bayes network illustrating hierarchical Bayes formulation of population variability



In the more general case, not necessarily Poisson, we use θ for the parameter of interest instead of λ , and we write φ instead of the vector $(\alpha, \beta)^T$. Hierarchical Bayes is so-named because it utilizes hierarchical or multistage prior distributions. In the hierarchical Bayes framework, the prior distribution for the parameter of interest, denoted by $\pi(\theta)$ is written as

$$\pi(\theta) = \int_{\Phi} \pi_1(\theta | \varphi) \pi_2(\varphi) d\varphi \quad (1)$$

In Eq. (1), $\pi_1(\theta | \varphi)$ is the first-stage prior, representing the population variability in θ for a given value of φ , where φ is typically a vector, and $\pi_2(\varphi)$, called the hyperprior, is the distribution representing the uncertainty in φ , whose components are called hyperparameters. The first-stage prior, $\pi_1(\theta | \varphi)$, is usually assumed to be of a particular parametric form, such as a gamma or lognormal distribution. It is also typical, although not necessary, to use a diffuse hyperprior for φ . Nothing limits the analysis to two stages, although the use of more than two stages has been rare in applications.

Continuing with the above Poisson example, each source has intensity λ_i , so θ is the value taken on by some random variable λ . To model variability in λ among sources, we will use a gamma first-stage prior, with parameters α and β . We choose independent, diffuse hyperpriors for α and β , so $\pi_2(\varphi)$ becomes the product of these diffuse hyperpriors. The posterior distribution of λ_i , will be given by an average of the posterior distribution for λ_i , conditional upon α and β (a gamma distribution), weighted by the joint posterior distribution for α and β . Thus, the posterior distribution for λ_i is a continuous mixture of gamma distributions:

$$\pi(\lambda_i | x, t) = \iint \pi_1(\lambda_i | x, t, \alpha, \beta) \pi_2(\alpha, \beta | x, t) d\alpha d\beta \quad (2)$$

The predictive distribution for λ , representing source-to-source variability, will be a similar mixture of gamma distributions.

3. ILLUSTRATIVE EXAMPLE – CONVERGENCE PROBLEMS

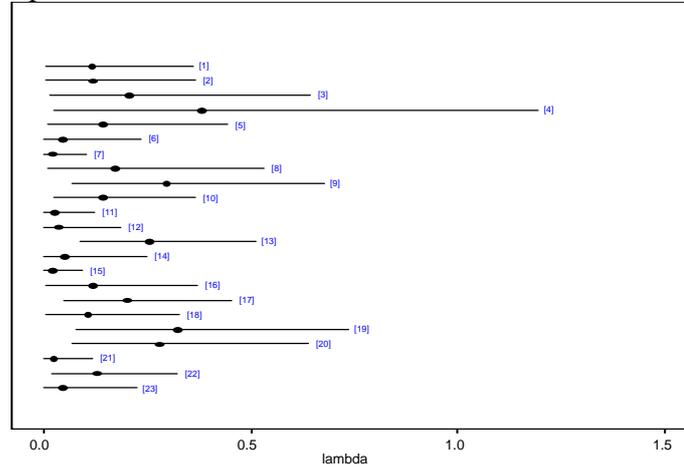
Consider the following 23 sources of data for loss of offsite power, taken from (5)

Table 1: Example Data for Loss of Offsite Power

Events	Exposure time (yr)	Events	Exposure Time (yr)
1	13.054	5	21.5
1	12.77	0	10.075
1	7.22	0	26.32
1	3.944	1	12.54
1	10.548	3	17.5
0	10.704	1	14.3
0	24	3	10.89
1	8.76	3	12.5
3	11.79	0	21.38
2	17.5	2	19.65
0	20.03	0	11.34
0	13.39		

Let us assume that the failure count for each source is Poisson-distributed with rate λ_i . The figure below shows a side-by-side plot of the 95% credibility intervals for λ for each of these sources, based on a Bayesian update of a Jeffreys prior (see (4) for details), and illustrates the variability in the sources. However, most of the intervals overlap, suggesting that the source-to-source variability is not large, and the overall PVC will not be too highly skewed.

Figure 2: Side-by-side plot of 95% intervals for 23 data sources in Table 1



Let us first take a conjugate first-stage prior, which in this case is a gamma distribution. Assuming independence among the sources, the likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{23} \frac{\Gamma(\alpha + x_i)}{x_i! \Gamma(\alpha)} \left(\frac{t_i}{\beta} \right)^{x_i} \left(1 + \frac{t_i}{\beta} \right)^{-(\alpha + x_i)} \quad (3)$$

As described in (4), parametric empirical Bayes employs numerical optimization techniques to find values of α and β that maximize the likelihood given by Eq. (3). The values of α and β that maximize

the likelihood for the data given in Table 1 are 2.1 and 23 yr. Using these point estimates of α and β in the first-stage gamma prior that describes source-to-source variability, the average rate is 0.09/yr, with a 90% credible interval of (0.02, 0.2).

Turning now to hierarchical Bayesian analysis using WinBUGS, we must specify hyperpriors for α and β . Let us use independent, diffuse distributions, so that the data drive the results. A gamma distribution with both parameters very small gives a distribution that is essentially flat over the positive real axis. With both parameters equal to 10^{-4} , the 5th percentile is zero, effectively, and the 95th percentile is approximately 10^{-220} . Thus, there is a very sharp vertical asymptote at zero, and the density is essentially flat (and approximately equal to zero) for values greater than zero.

Following the guidance provided by (6), we run multiple chains in order to monitor convergence, both graphically and quantitatively, using the BGR convergence diagnostic calculated by WinBUGS (7). WinBUGS cannot generate initial values from the diffuse gamma hyperpriors used for α and β , so initial values must be supplied. We use the point estimates from the empirical Bayes analysis above, and pick values around these, as these approximate the mode of the joint posterior distribution for α and β .

Figure 3 shows the history of 100,000 samples for α . The two chains are not well mixed, indicating potential convergence problems. The BGR plot is shown in Figure 4. If the chains have converged to the posterior distribution, the red line should be about 1.0, and all three lines should be stable. This is clearly not the case, so we have not converged after 100,000 samples.

Figure 3: History of 100,000 iterations for α , showing failure of chains to mix

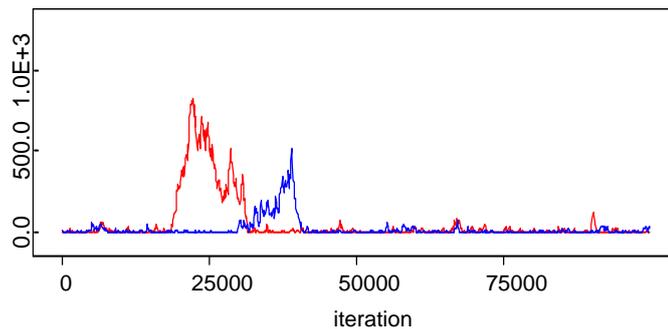
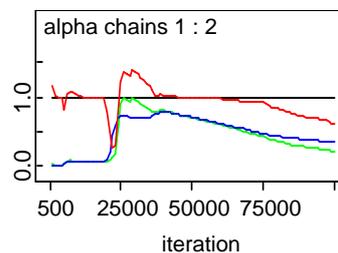


Figure 4: BGR diagnostic for first 100,000 iterations for α , illustrating failure to converge



Running the chains for an additional 100,000 iterations gives results similar to those above. Even after more than 10^6 iterations, this behavior was still observed, and the estimates of the two chains differed significantly.

Highly correlated parameters can be an obstacle to convergence, as noted by (8), among others. In this case, the rank correlation between α and β in the joint posterior distribution is estimated by WinBUGS to be 0.98. We can ameliorate this problem by reparameterizing the gamma first-stage prior in terms of the mean and coefficient of variation, which we would expect to be nearly

independent. For a gamma distribution, the mean is given by α/β . The coefficient of variation is the ratio of the standard deviation to the mean, which for the gamma distribution will be $\alpha^{-0.5}$. We place independent, diffuse gamma hyperpriors on these parameters. Initial values are again chosen using the results from the empirical Bayes analysis above.

As the figures below illustrate, we now appear to have convergence within the first couple of thousand iterations (because of space limitations, we only show plots for one parameter; convergence should be checked for all parameters in the model). To be safe, we discard the first 10,000 iterations and do not use these samples in estimating parameter values.

Figure 5: History of 10,000 iterations for coefficient of variation, illustrating convergence

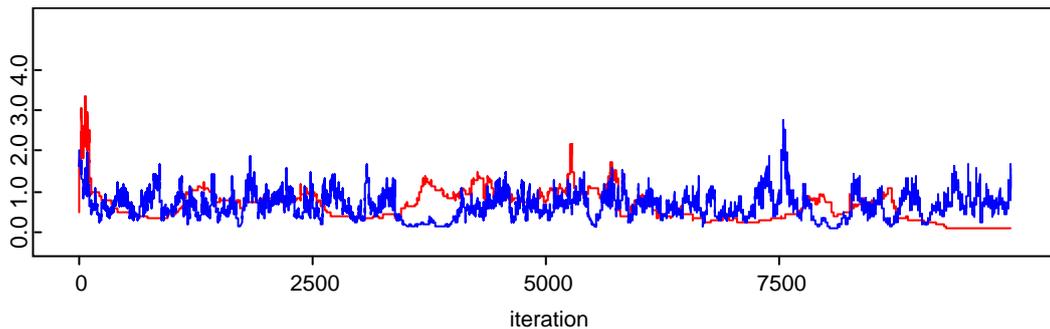
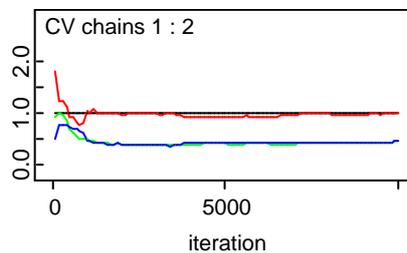
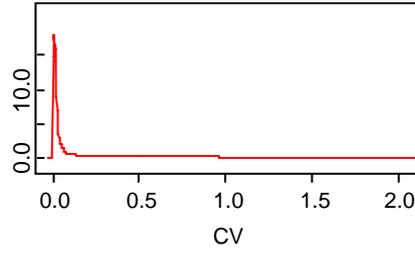


Figure 6: BGR diagnostic for 10,000 iterations for coefficient of variation, illustrating convergence



Another 100,000 iterations were used to estimate the desired parameters. The average PVC is obtained in WinBUGS by sampling α and β from the joint posterior distribution, and then generating a new value of λ from the first-stage gamma distribution, using the values of α and β just drawn. This predictive distribution has mean 0.09/yr, with a 90% credibility interval of (0.02, 0.20). These results match those from the empirical Bayes analysis above; in general, the results are not necessarily so similar. In particular, credible intervals from the hierarchical Bayesian analysis tend to be wider, because of the inclusion of uncertainty about the hyperparameters α and β . This uncertainty is sometimes included in empirical Bayes analyses through asymptotic approximations, although to simplify the exposition we did not do it in this example. For this example, the uncertainty in the hyperparameters is relatively small. The figure below shows the marginal posterior distribution for the coefficient of variation. Note that the density is concentrated at small values, indicating there is little source-to-source variability in the data.

Figure 7: Marginal posterior distribution for coefficient of variation



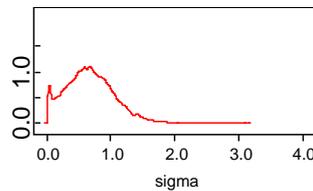
We now examine the impact of choosing a different functional form for the first-stage prior describing the variability in λ from source to source. A common alternative is the lognormal distribution:

$$\pi_1(\lambda | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma\lambda} \exp\left[-\frac{(\log \lambda - \mu)^2}{2\sigma^2}\right], -\infty < \mu < \infty, \sigma > 0 \quad (4)$$

Care must be taken in the choice of hyperpriors in order to avoid numerical difficulties. In particular, the diffuse gamma distribution used above is problematic as a hyperprior for σ . This issue is discussed at length in (9), where a uniform distribution with finite bounds is recommended as an alternative. A value of 1.4 for σ corresponds to a lognormal error factor of 10, and therefore to a factor of 100 between the 95th and 5th percentiles. Therefore, a uniform(0, 5) hyperprior for σ should be quite diffuse in practice. We will check the marginal posterior distribution for σ to ensure that the upper tail has not been truncated, a sign that a wider prior distribution is needed.

With a uniform(0, 5) hyperprior on σ , and a flat hyperprior for μ , convergence was achieved in the first 1,000 iterations. Running another 10,000 samples to estimate parameter values, we find the mean of the average PVC to be 0.10. The 90% credible interval is (0.02, 0.24), slightly wider than with the gamma first-stage prior above. The marginal posterior density of σ is shown below. The right tail is not truncated, indicating that the uniform(0, 5) hyperprior was sufficiently diffuse. Note again the “spike” at 0 in this distribution, analogous to the one in Figure 7 above. As above, this is an indication that there is little source-to-source variability in the data.

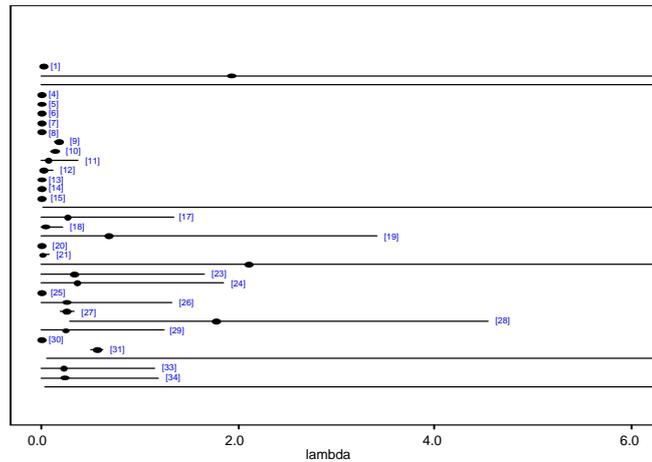
Figure 5: Marginal posterior distribution for σ , with “spike” at 0, an indication of little source-to-source variability



4. ILLUSTRATIVE EXAMPLE – CHOICE OF FIRST-STAGE PRIOR

In (10), the authors present a hierarchical Bayesian analysis of failure rate data for digital instrumentation and control equipment. Figure 9 illustrates the extremely large variability in the sources the authors examined (many intervals do not overlap), and suggests that a highly skewed distribution will be required to adequately model the source-to-source variability.

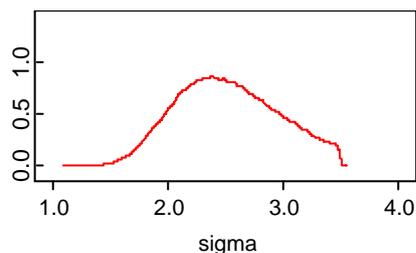
Figure 9: Side-by-side plot of 95% intervals for data sources in [10]. Dots are posterior means from update of Jeffreys prior



With a gamma first-stage prior, and diffuse hyperpriors on α and β , WinBUGS converges within 1,000 iterations, with no reparameterization needed. An additional 50,000 iterations gave an average PVC with a mean value of 0.09, a median of 0.011, and a 90% credible interval of $(6.65 \times 10^{-8}, 0.43)$. The posterior mean of the gamma shape parameter, α , was 0.24. This leads to a sharp vertical asymptote at zero, with an excessive amount of probability mass near zero, which is the only way a gamma first-stage prior can accommodate such extremely large source-to-source variability.

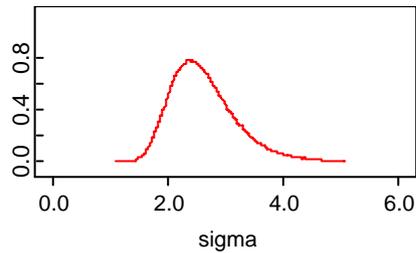
The lognormal distribution is perhaps a better choice to model variability that ranges over orders of magnitude. The authors of (10) recommend a lognormal distribution in place of the gamma distribution; however, they based their hyperpriors for μ and σ on the data, using a uniform(-7, -0.1) distribution for μ , and a uniform(1, 3.5) distribution for σ . With these choices, the average PVC has a mean of 0.29, a median of 0.007, and a 90% interval of $(8.6 \times 10^{-5}, 0.49)$. Figure 6 shows the marginal posterior distribution for σ , showing the truncation of the upper tail caused by the overly narrow data-based hyperprior.

Figure 6: Marginal posterior distribution for σ illustrating truncation of upper tail



The mean of the lognormal distribution is $\exp(\mu + \sigma^2/2)$. Thus, the mean is very sensitive to σ , the median, given by e^μ , is unaffected by σ . If we replace the data-based hyperpriors used in (10) with more diffuse distributions, we expect the mean to increase. We replaced the hyperprior for μ with a flat prior over the real axis, and the distribution for σ with a more diffuse uniform(0, 5) distribution. With these more diffuse hyperpriors, the mean of the average PVC increased to 1.1, with a 90% credible interval of $(6.3 \times 10^{-5}, 0.55)$. Note that the mean, which lies well to the right of the 95th percentile, is no longer a representative value. The median, as expected, was robust against this change, remaining at 0.007. The marginal posterior distribution for σ is shown below. Note the lack of truncation in the upper tail.

Figure 11: Marginal posterior distribution for σ with more diffuse hyperprior illustrating no truncation in upper tail



5. CONCLUSIONS

We have illustrated the usefulness of choosing, for some problems, a parameterization in which the parameters are approximately independent as an aid in enhancing convergence. In the case of the gamma first-stage prior, this was accomplished by using the mean and coefficient of variation as the parameters in place of α and β . When there is significant variability in the data, as there is in our second example, the gamma first-stage prior is forced to have a small shape parameter (α) in order to accommodate such large variability. This helps to lessen the positive correlation between α and β , allowing quick convergence of the MCMC sampling to the joint posterior distribution without reparameterizing. When the variability in the data is smaller, α can take on larger values, and the correlation can be higher, leading to very slow convergence and a need for reparameterization.

We have also illustrated how the choice of first-stage priors can have a significant impact on the average results when source-to-source variability is very large. A conjugate first-stage prior can only accommodate large variability by having a shape parameter $\ll 1$. This introduces a sharp vertical asymptote at zero, and the excessive probability mass near zero causes the mean value and lower percentiles to be too low.

A lognormal first-stage prior avoids this problem, but because the mean is very sensitive to σ , the mean of the average PVC may not be very robust. One must be careful in choosing a hyperprior for σ to avoid truncating the upper tail, as this will lead to a mean value that is too small. In particular, the hyperprior should not be based on the observed data. With a diffuse hyperprior and realistic amounts of data, the marginal posterior distribution for σ may still be quite broad, and this can cause the mean value to be unrepresentatively high, well beyond the 95th percentile of the average PVC in our second example.

The median of the average PVC is a more robust measure of central tendency. With a gamma first-stage prior in our second example, the median was found to be 0.01. With a lognormal first-stage prior, the median was 0.007, about the same, despite the much larger mean value (1.1 versus 0.09). When modeling extremely large source-to-source variability, as in the second example, the median may be a more representative value than the mean. Note, also, the relative robustness of the 95th percentile: 0.43 with the gamma first-stage prior and 0.55 with the lognormal. Because of the difference in shape near zero, the 5th percentile is not at all robust.

References

- [1] Kaplan, Stan, *On a Two-Stage Bayesian Procedure for Determining Failure Rates*. 1983, IEEE Transactions on Power Apparatus and Systems, pp. 195-262.
- [2] Eide, Steven L. et al. *Industry-Average Performance for Components and Initiating Events at U. S. Commercial Nuclear Power Plants*. U. S. Nuclear Regulatory Commission, NUREG/CR-6928, 2007..

- [3] Lunn, David J., et al. *WinBUGS - A Bayesian Modelling Framework: Concepts, Structure, and Extensibility*. Statistics and Computing, **10** (2000), pp. 325-337.
- [4] Siu, Nathan O. and Kelly, Dana L., *Bayesian Parameter Estimation in Probabilistic Risk Assessment*. Reliability Engineering and System Safety, **62** (1998), pp. 89-116.
- [5] Atwood, C. L., et al. *Evaluation of Loss of Offsite Power Events at Nuclear Power Plants: 1980-1996*. U. S. Nuclear Regulatory Commission, NUREG/CR-5496, 1997.
- [6] Gelman, Andrew. Inference and Monitoring Convergence. [book auth.] W. R. Gilks and et al. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996, pp. 131-143.
- [7] Brooks, S. P. and Gelman, A., *General Methods for Monitoring Convergence of Iterative Simulations*. Journal of Computational and Graphical Statistics, **7**, No. 4, (1998), pp. 434-455.
- [8] Gelman, Andrew, et al. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, 2004.
- [9] Gelman, Andrew. , *Prior Distributions for Variance Parameters in Hierarchical Models.*, Bayesian Analysis, Vol. 1, No. 3, (2006),pp. 515-533.
- [10] Yue, Meng and Chu, Tsong-Lun. *Estimation of Failure Rates of Digital Components Using a Hierarchical Bayesian Method*. New Orleans : 2006. International Conference on Probabilistic Safety Assessment and Management.