

Assessment of the Reliability of Fault-Tolerant Software: a Bayesian Approach

Bev Littlewood, Peter Popov, Lorenzo Strigini
Centre for Software Reliability, City University, London

Abstract

Fault tolerant systems based on the use of software design diversity may be able to achieve high levels of reliability more cost-effectively than other approaches, such as heroic debugging. Earlier experiments have shown that the reliabilities of multi-version software systems are more reliable than the individual versions. However, it is also clear that the reliability benefits are much worse than would be suggested by naive assumptions of failure independence between the versions. It follows that it is necessary to assess the reliability *actually achieved* in a fault tolerant system. The difficulty here mainly lies in acquiring knowledge of the degree of dependence between the failures processes of the versions. The paper addresses the problem using Bayesian inference. In particular, it considers the problem of choosing a prior distribution to represent the beliefs of an expert assessor. It is shown that this is not easy, and some pitfalls for the unwary are identified.

(Presented at SAFECOMP'2000. Appears in Lecture Notes in Computer Science, Vol. 1943.

This version corresponds to Version 1.2, 31 May, 2000, of the DISPO Technical Report of the same title, PP_DD_TR-05_v1_2)

1. Introduction

Achieving high reliability in software-based systems is not easy. *Assessing* the reliability of a software-based system - in particular, gaining high confidence that the required reliability has been achieved in the case of a critical system - is probably even more difficult.

The use of software design diversity in a fault tolerant architecture has been suggested as one solution to the problem of achieving very high reliability. Whilst the benefits of software diversity do not seem to be as high as those sometimes claimed for hardware redundancy [Knight & Leveson 1986b], there is nevertheless evidence that fault tolerance can deliver levels of reliability higher than could be achieved with a single software version [Knight & Leveson 1986a]. Indeed, there are several examples of apparently successful use of design diversity in engineered systems that are in operational use, e.g. [Voges 1988], [Briere & Traverse 1993], [Kantz & Koza 1995].

The reason that software fault tolerance does not deliver dramatically high reliability is that the failures of different software versions in a fault tolerant system cannot be assumed to be independent [Eckhardt & Lee 1985], [Knight & Leveson 1986b], [Littlewood & Miller 1989]. One reason for this is that human designers and builders of software systems tend to make similar mistakes. This means that even two versions of a program that have been produced using 'independent' development teams will not fail independently - instead there will be a greater tendency for the versions to fail simultaneously, thus decreasing the reliability of a 1-out-of-2 system built from them.

This means that it is imperative that we are able to *assess* the actual reliability that has been achieved in a fault tolerant system: we cannot simply assume that the probability of system failure is the product of the two version failure probabilities. In this paper we address this problem of assessment of reliability using a Bayesian approach. For simplicity we shall restrict ourselves to a demand-based 1-out-of-2 system architecture, i.e. one in which the system fails if and only if both versions of the software fail. Such an architecture has considerable practical interest: for example, it is the architecture employed in many safety systems, such as nuclear reactor protection systems.

The simplest way to assess the reliability of a system - fault tolerant or otherwise - is to observe its failure behaviour in (real or simulated) operation. Quite simple statistical techniques allow the reliability to be estimated from such data [Musa *et al.* 1987], [Brocklehurst *et al.* 1990]. However, it is well-known that such a 'black-box' approach to reliability estimation has severe limitations [Littlewood & Strigini 1993], [Butler & Finelli 1991]. Essentially, it is infeasible to assess very high levels of reliability this way because impracticably long periods of testing would be required.

An important issue, then, is whether knowledge of the fault tolerant nature of a system aids its reliability assessment. Can we use the details of the version failures to obtain greater confidence in the system reliability than would come just from the black-box system data? In this paper we shall investigate a Bayesian approach to questions like this.

2 Bayesian inference

We shall assume that the successive demands are independent, and that the probabilities of failure on demand (*pdfs*) for versions and for the system remain constant.

Consider the case where the system is being treated simply as a black box, so that only *system* failure or success is observed. Denoting the probability of failure on demand for the system as p , the posterior distribution of p after seeing r failures in n demands is:

$$f_p(x|r, n) = \frac{\binom{n}{r} x^r (1-x)^{n-r} f_p(x)}{\int_0^1 \binom{n}{r} x^r (1-x)^{n-r} f_p(x) dx} \quad (1)$$

Here $f_p(\bullet)$ is the prior distribution of p , which represents the assessor's *a priori* beliefs about p , i.e. before seeing the result of the test on n demands.

If complete information on the *versions* comprising the 1-out-of-2 system is available, there are four different possible outcomes for each demand:

Event	Version A	Version B	Number of occurrence in n tests	Probability
α	fails	fails	r_1	P_{AB}
β	fails	succeeds	r_2	P_{AB}^-
γ	succeeds	fails	r_3	P_{AB}^-
δ	succeeds	succeeds	r_4	P_{AB}^-

There are now four parameters, given in the last column of the table, in the probability model. However, these four probabilities sum to unity, hence there are only three degrees of freedom: any three of these four parameters completely specify the model. If we choose to deal with the first three, the posterior distribution is:

$$f_{P_{AB}, P_{AB}^-, P_{AB}^-}(x, y, z | r_1, r_2, r_3, n) = \frac{L(r_1, r_2, r_3, n | P_{AB}, P_{AB}^-, P_{AB}^-) f_{P_{AB}, P_{AB}^-, P_{AB}^-}(x, y, z | r_1, r_2, r_3, n)}{\iiint_{P_{AB}, P_{AB}^-, P_{AB}^-} L(r_1, r_2, r_3, n | P_{AB}, P_{AB}^-, P_{AB}^-) f_{P_{AB}, P_{AB}^-, P_{AB}^-}(x, y, z | r_1, r_2, r_3, n) dP_{AB} dP_{AB}^- dP_{AB}^-} \quad (2)$$

where

$$L(r_1, r_2, r_3, n | P_{AB}, P_{AB}^-, P_{AB}^-) = \frac{n!}{r_1! r_2! r_3! (n - r_1 - r_2 - r_3)!} P_{AB}^{r_1} P_{AB}^{r_2} P_{AB}^{r_3} (1 - P_{AB} - P_{AB}^- - P_{AB}^-)^{n - r_1 - r_2 - r_3} \quad (3)$$

is the multinomial likelihood of seeing r_1 coincident failures of both versions, r_2 failures of version A only, r_3 failures of version B only, respectively, in n tests, and $f_{P_{AB}, P_{AB}^-, P_{AB}^-}(\bullet, \bullet, \bullet)$ is the prior distribution.

It may sometimes be convenient to use a different parameterisation. For example, we could use the triplet P_A , P_B and P_{AB} , denoting the probabilities of failure of both versions and the probability of coincident failure of the pair (i.e. failure of the system). Trivially,

$$P_A = P_{AB}^- + P_{AB} \quad \text{and} \quad P_B = P_{AB}^- + P_{AB}$$

and a simple change of variable gives the posterior:

$$f_{P_{AB}, P_A, P_B}(x, y, z | r_1, r_2, r_3, n) = \frac{L(r_1, r_2, r_3, n | P_{AB}, P_A, P_B) f_{P_{AB}, P_A, P_B}(x, y, z | r_1, r_2, r_3, n)}{\iiint_{P_{AB}, P_A, P_B} L(r_1, r_2, r_3, n | P_{AB}, P_A, P_B) f_{P_{AB}, P_A, P_B}(x, y, z | r_1, r_2, r_3, n) dP_{AB} dP_A dP_B} \quad (4)$$

where

$$L(r_1, r_2, r_3, n | P_{AB}, P_A, P_B) = \frac{n!}{r_1! r_2! r_3! (n - r_1 - r_2 - r_3)!} P_{AB}^{r_1} (P_A - P_{AB})^{r_2} (P_B - P_{AB})^{r_3} (1 + P_{AB} - P_A - P_B)^{n - r_1 - r_2 - r_3} \quad (5)$$

is the likelihood of the same observation as above, and $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$ is the prior.

From either of (2) and (4) the posterior marginal distribution of the *system* probability of failure can be computed.

The practical advantage of these multivariate posterior distributions over the ‘black-box’ version, (1), is that they take account specifically of any information we might have concerning the individual version reliabilities and the failure dependence between versions. It might be expected that such extra information would provide greater confidence in the reliability (or, conversely, unreliability) of the system: i.e. that the fault tolerant architecture provides benefits in the *assessment* of reliability, as it does in its *achievement*.

Consider, for example, the situation where one or both of the versions have been operating for considerable time in other systems, and thus have built up extensive data on successful and failed demands. They are then bought in as commercial-off-the-shelf (COTS) software components to be used in a novel fault tolerant system. If the assessor believes that these earlier data come from operational environments that are sufficiently similar to that of the new application, they could be incorporated into the prior to express ‘strong’ beliefs about the version reliabilities. In contrast, such earlier individual version histories would tell the assessor little about the *system* reliability, and hence the prior for (1) would have to be a vague one.

In this scenario, the past operational experience of the versions allows the assessor to learn about their individual reliabilities, but not about the dependence between their failure behaviours. Information about this would have to come from testing the versions together on the n operationally representative demands.

Clearly, the role of the prior is crucial in the analysis. Eliciting from an expert assessor his or her prior beliefs can be very difficult, especially when these take the form of a multi-variate distribution. When safety-critical systems are being assessed, in particular, it would be useful to have an assurance that this elicited prior was in some sense a ‘conservative’ one. Unfortunately, it is not clear currently how to do this. In the next section we report on a preliminary study of some classes of prior distributions for this simple 1-out-of-2 model.

3 Prior distributions

3.1 Dirichlet distribution

It is common in Bayesian statistics to use a *conjugate family* of distributions to represent prior belief. This is a parametric family of distributions that has the property for a particular problem (i.e. likelihood function) that if the expert uses a member of the family to represent his/her prior beliefs, then the posterior will automatically also be a member of the family. It can be shown that, when a conjugate family exists (this is not always the case), it is unique.

For the simple example of the binomial likelihood used in (1), the conjugate family is the Beta distribution: if the prior is $B(x, \alpha, \beta)$, and r failures are observed in n demands, the posterior will be $B(x, \alpha + r, \beta + n - r)$.

This result generalises to deal with the case considered above in (2) and (4): when the likelihood is multinomial, the conjugate family is the Dirichlet distribution [Johnson & Kotz 1972]. This takes the form:

$$Dirichlet(\theta_0, \theta_1, \theta_2, \theta_3) = f_{y_1, y_2, y_3}(y_1, y_2, y_3) = \frac{\Gamma\left(\sum_{j=0}^m \theta_j\right)}{\prod_{j=0}^m \Gamma(\theta_j)} \left(1 - \sum_{j=1}^m y_j\right)^{\theta_0 - 1} \prod_{j=1}^m y_j^{\theta_j - 1} \quad (6)$$

If this is used as the prior for the 1-out-of-2 system comprising versions A and B , and the observed data are r_1 coincident failures of both channels, r_2 failures of only version A and r_3 failures of only version B in n demands, the posterior distribution is again a Dirichlet distribution:

$$Dirichlet(\theta_0 + n - r_1 - r_2 - r_3, \theta_1 + r_1, \theta_2 + r_2, \theta_3 + r_3).$$

A property of the Dirichlet distribution is that its marginal distributions are also Dirichlet distributions. In particular, its univariate marginals are Beta distributions; in this sense the Dirichlet can be seen as a multivariate generalisation of the Beta distribution.

Would the Dirichlet family of distributions be suitable to represent an expert assessor’s prior beliefs about the joint failure behaviour of the software versions making up a 1-out-of-2 system? Certainly there is a great advantage to be gained by restricting the elicitation exercise within such a parametric family: the expert only has to find the four parameters of the distribution that capture his/her beliefs. Whilst this is by no means a trivial task, it is much easier than trying to describe in an unconstrained way a complete tri-variate distribution.

We now consider how, when the prior is a Dirichlet distribution, posterior beliefs about system reliability based on complete version failure data differ from those using only ‘black-box’ system data. We might expect there to be a benefit in taking account of the more extensive data of the former case.

To make the two cases comparable, we need to assume that the expert's prior belief about the *system reliability* are the same in the two cases. That is, the prior $f_{P_{AB}}(\bullet)$ to be used in the black-box inference is the marginal distribution obtained from the more general Dirichlet prior by integrating out two of the variables, P_{AB}^- and $P_{A\bar{B}}^-$:

$$f_{P_{AB}}(\bullet) = \iint_{P_{AB}^-, P_{A\bar{B}}^-} f_{P_{AB}, P_{AB}^-, P_{A\bar{B}}^-}(\bullet, P_{AB}^-, P_{A\bar{B}}^-) dP_{AB}^- dP_{A\bar{B}}^- = B(\bullet, \theta_1, \theta_0 + \theta_2 + \theta_3) \quad (7)$$

For testing results $(n : r_1, r_2, r_3)$ the black-box approach takes account only of the number of *system* failures; the posterior distribution of the probability of system failure is thus:

$$f_{P_{AB}}(\bullet | n, r_1) = B(\bullet, \theta_1 + r_1, \theta_0 + \theta_2 + \theta_3 + n - r_1) \quad (8)$$

The white-box result, taking account of the joint version failure behaviour, is obtained from the joint posterior, *Dirichlet* $(\theta_0 + n - r_1 - r_2 - r_3, \theta_1 + r_1, \theta_2 + r_2, \theta_3 + r_3)$, by integrating out P_{AB}^- and $P_{A\bar{B}}^-$ to obtain the marginal posterior of the probability of system failure:

$$\begin{aligned} f_{P_{AB}}(\bullet | n, r_1, r_2, r_3) &= \iint_{P_{AB}^-, P_{A\bar{B}}^-} \text{Dirichlet}(\theta_0, \theta_1, \theta_2, \theta_3 | n, r_1, r_2, r_3) dP_{AB}^- dP_{A\bar{B}}^- = \\ &= \iint_{P_{AB}^-, P_{A\bar{B}}^-} \text{Dirichlet}(\theta_0 + n - r_1 - r_2 - r_3, \theta_1 + r_1, \theta_2 + r_2, \theta_3 + r_3 | n, r_1, r_2, r_3) dP_{AB}^- dP_{A\bar{B}}^- = \\ &= B(\bullet, \theta_1 + r_1, \theta_0 + n - r_1 - r_2 - r_3 + \theta_2 + r_2 + \theta_3 + r_3) = \\ &= B(\bullet, \theta_1 + r_1, \theta_0 + \theta_2 + \theta_3 + n - r_1) \end{aligned} \quad (9)$$

Somewhat surprisingly, this is completely identical to the 'black-box' result, (8). In other words, whatever the detailed failure behaviour of the versions, there is no benefit from taking this extra information into account in assessing the reliability of the system. The posterior belief about the system reliability will be exactly the same as if only the *system* failures and successes had been recorded.

This result is disappointing. Whilst fault tolerance based upon software design diversity may help in the *achievement* of high reliability, it brings no advantage for the *assessment* of the reliability that has actually been achieved.

Of course, the result depends crucially upon expressing prior beliefs in the form of the conjugate Dirichlet family of distributions. Whilst restricting an assessor's expression of belief to this family is a great convenience, for the reasons stated earlier, it could be that an assessor's actual beliefs do not fit into this family. For instance, Dirichlet distribution implies negative correlation between each pair of variates, P_{AB} , $P_{A\bar{B}}$ and $P_{\bar{A}B}$, which may be against the assessor's intuition. Negative correlation between P_{AB} and $P_{A\bar{B}}$, for instance, means that seeing A fail and B succeed makes the failure of both channels a less likely event. The assessor, on the contrary, may see in the failure of channel A evidence of increased risk of system failure and, therefore, may wish to reduce his confidence in system reliability. This is impossible with Dirichlet distribution and in such cases Dirichlet distribution, despite its convenience, can not be recommended as a reasonable prior.

3.2 Prior with known failure probabilities of versions

Consider now the plausible situation in which there is a very great deal of data from past operational use for each version, so that each probability of failure on demand can be estimated with great accuracy. Here the uncertainty - or the ignorance of the assessor - almost entirely concerns the *joint* failure behaviour of the versions. In this section we shall approximate to this situation by assuming that version probabilities of failure on demand are known *with certainty*.

The prior in this case can be written:

$$f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet) = f_{P_{AB}}(\bullet | P_A, P_B) f_{P_A, P_B}(P_A, P_B) \quad (10)$$

where the second term on the right is a Dirac delta function at the point $(P_A = P_{Atrue}, P_B = P_{Btrue})$, representing the known version failure probabilities.

The prior marginal distribution of the system *pdf* is

$$\begin{aligned} f_{P_{AB}}(\bullet) &= \iint_{P_A, P_B} f_{P_{AB}}(\bullet | P_A, P_B) f_{P_A, P_B}(P_A, P_B) dP_A dP_B = \\ &= f_{P_{AB}}(\bullet | P_{Atrue}, P_{Btrue}) \end{aligned}$$

In other words, the uncertainty in the prior concerns only the (marginal) probability of system failure, as we would expect.

The posterior marginal distribution of system pdf is:

$$\begin{aligned}
 f_{P_{AB}}(\bullet | n, r_1, r_2, r_3) = & \\
 & \frac{\iint_{P_A, P_B} (P_{AB})^{r_1} (P_A - P_{AB})^{r_2} (P_B - P_{AB})^{r_3} (1 - P_A - P_B + P_{AB})^{n-r_1-r_2-r_3} f_{P_{AB}}(\bullet | P_A, P_B) f_{P_A, P_B}(P_A, P_B) dP_A dP_B}{\iiint_{P_{AB}, P_A, P_B} f_{P_{AB}}(\bullet | P_A, P_B) f_{P_A, P_B}(P_A, P_B) dP_{AB} dP_A dP_B} = (11) \\
 & \frac{(P_{AB})^{r_1} (P_{Atrue} - P_{AB})^{r_2} (P_{Btrue} - P_{AB})^{r_3} (1 - P_{Atrue} - P_{Btrue} + P_{AB})^{n-r_1-r_2-r_3} f_{P_{AB}}(\bullet | P_{Atrue}, P_{Btrue})}{\int_{P_{AB}} (P_{AB})^{r_1} (P_{Atrue} - P_{AB})^{r_2} (P_{Btrue} - P_{AB})^{r_3} (1 - P_{Atrue} - P_{Btrue} + P_{AB})^{n-r_1-r_2-r_3} f_{P_{AB}}(\bullet | P_{Atrue}, P_{Btrue}) dP_{AB}}
 \end{aligned}$$

We illustrate this set-up with a few numerical examples. In each case we shall assume that $P_{Atrue} = 0.001$, $P_{Btrue} = 0.0005$. Clearly, a 1-out-of-2 system will be at least as reliable as the more reliable of the two versions, so the prior distribution of the probability of simultaneous version failure on demand is zero outside $[0, 0.0005]$ in this case. We shall consider below two examples of this distribution: a uniform prior, Beta(x, 1, 1), and a Beta(x, 10, 10) both constrained to lie on this interval.

Table 1

Uniform prior Pab Pa,Pb		Percentiles					
		10%	50%	75%	90%	95%	99%
	Prior	0.00005	0.00025	0.000375	0.00045	0.000475	0.000495
	Black Box	0.000011	0.00007	0.000137	0.000225	0.000286	0.00041
$r_1=0$	White Box (version failures)	0.000008	0.00005	0.000095	0.000148	0.00018	0.000245
	White Box (no version failures)	0.000268	0.00042	0.000462	0.00048	0.000485	0.00049
$r_1 = 1$	Black Box	0.000045	0.000155	0.000246	0.000342	0.000396	0.000465
	White Box (version failures)	0.00004	0.00012	0.000179	0.000238	0.000271	0.00033
$r_1 = 3$	Black Box	0.00015	0.0003	0.000384	0.000443	0.000465	0.000485
	White Box (version failures)	0.000165	0.000283	0.000343	0.00039	0.000413	0.000448
$r_1 = 5$	Black Box	0.000235	0.000375	0.000435	0.00047	0.00048	0.00049
	White Box (version failures)	0.000345	0.00044	0.000469	0.000482	0.000485	0.000489
Non-uniform prior Pab Pa,Pb		Percentiles					
		10%	50%	75%	90%	95%	99%
	Prior	0.000175	0.000245	0.000283	0.000317	0.000335	0.000368
$r_1=0$	Black Box	0.000146	0.000215	0.000253	0.000286	0.000306	0.000343
	White Box (version failures)	0.00013	0.000188	0.00022	0.00025	0.000269	0.0003
	White Box (no version failures)	0.000205	0.000278	0.000313	0.000345	0.00036	0.00039
$r_1 = 1$	Black Box	0.000161	0.000228	0.000265	0.0003	0.000318	0.000353
	White Box (version failures)	0.00015	0.00021	0.000244	0.000275	0.000291	0.000325
$r_1 = 3$	Black Box	0.000185	0.000251	0.000287	0.00032	0.000336	0.00037
	White Box (version failures)	0.000195	0.000255	0.00029	0.00032	0.000335	0.000365
$r_1 = 5$	Black Box	0.000205	0.000271	0.000305	0.000335	0.000353	0.00038
	White Box (version failures)	0.00024	0.000304	0.000335	0.00036	0.000375	0.000402

Table 1: Two groups of results are summarised: with uniform prior and non-uniform prior, Pab|Pa,Pb=Beta(x,10,10) on the interval $[0, 0.0005]$. The percentiles illustrate the cumulative distribution $P(\theta \leq X) = Y$, where X are the values shown in the table and Y are the chosen percentiles, 10%, 50%, 75%, 90%, 95% and 99%. Rows labelled 'Black box' represent the percentiles, calculated with the black-box inference, those labelled 'White box' show the percentiles calculated for a posterior derived with (11). '(no version failures)' and '(version failures)' refer to two different observations, in which no individual failures of channels and individual channel failures were observed, respectively.

We make no claims for 'plausibility' for these choices of prior. However, it should be noted that each is quite pessimistic: both priors, for example, have mean 0.00025, so it suggests a prior belief that about half channel B failures will also result in channel A failure.

We assume that $n=10,000$ demands are executed in an operational test environment. In each of the examples in Table 1, one case is that where the observed numbers of channel A and channel B failures take their (marginal) expected values, i.e. 10 and 5 respectively. The other case is the extreme one where there are no single channel failures.

In each case our main interest is in how our assessment of the system reliability based upon the full information, r_1, r_2, r_3 , differs from the assessment based only upon the black-box evidence, r_1 , alone.

Discussion

Table 1 ($r_1=0$) shows the increased confidence that comes when extensive testing reveals *no system failures*. The black-box posterior belief in the system *pdf* is everywhere better than the prior belief: the two distributions are stochastically ordered. More importantly, the posterior belief 'White box (version failures)' based on version failures, but no system failures, is better than the black-box posterior: these distributions are also stochastically ordered. Here the extra information of version failure behaviour has allowed greater confidence to be placed in the system reliability compared with what could be claimed merely from the system behaviour alone.

The result is in accord with intuition. Seeing no system failures in 10,000 demands, when there have been 10 channel *A* failures and 5 channel *B* failures suggests that there is some *negative correlation* between failures of the channels: even if the channels were failing *independently* we would expect to see some common failures (the expected number of coincident failures, conditional on 10 *A*s and 5 *B*s, is 2.5).

The rows where ($r_1 \neq 0$) show what happens when there are system failures (common version failures), with the same numbers of channel failures (10 *A*s, 5 *B*s). As would be expected, the more system failures there are on test, the more pessimistic are the posterior beliefs about system reliability. More interesting, however, is the relationship between the black-box and white-box results. Consider ($r_1=5$). These rows of Table 1 represent the most extreme case where all channel *B* failures are also channel *A* failures. This would suggest strongly that there is *positive correlation* between the version failures. Here the black-box posterior belief would give results that are too optimistic compared with those based on the complete failure data: the white-box and black-box distributions are stochastically ordered.

These results show that knowledge of the detailed version failure data can strongly influence belief about the reliability of a 1-out-of-2 fault tolerant system, compared with the case where only the results of system tests are known. In particular, it may be possible to exploit evidence of negative correlation of version failures from test to sustain stronger claims for reliability than could be made just from knowledge of system behaviour.

As a cautionary note, when $r_1=0$ the posterior 'White box (no version failures)', show the posterior distribution of system *pdf* for the case where there have been no failures of either version (and hence no system failures). At first glance these results are surprising. Whilst the test has shown no failures - in particular no system failures - in 10,000 demands, we nevertheless believe the system reliability to be worse than it was *a priori*. How can the observation of such 'good news' make us lose confidence in the system?

The reason for this apparent paradox lies in the constraints on the parameters of the model that are imposed by assuming the versions reliabilities are known with certainty. Consider Table2:

Table 2

	A fails	A succeeds	
B fails	θ	$P_B - \theta$	P_B
B succeeds	$P_A - \theta$	$1 - P_A - P_B + \theta$	$1 - P_B$
	P_A	$1 - P_A$	1

There is only one unknown parameter, θ , the system *pdf*, which appears in all the cells above representing the four possible outcomes of a test. If we observe no failures in the test, this makes us believe that the entry in the (*A* succeeds, *B* succeeds) cell, $1 - P_A - P_B + \theta$, is large. Since P_A, P_B are known, this makes us believe that θ is large.

Of course, it could be argued that observing no version failures in 10,000 demands, with the known version *pdf*s 0.001, 0.0005, is extremely improbable - i.e. observing this result in practice is essentially impossible. This does not remove the difficulty, however: it can be shown that *whatever the value of n*, the 'no failures' posterior will be more pessimistic than the prior. In fact there is stochastic ordering here, both between the prior and the posteriors, and among the posteriors for different *n*.

The practical conclusion seems to be that if this model is to be used the number of demands in test should be great enough to ensure that at least some version failures are observed.

3.3 Other plausible priors

The previous two examples show some danger of using "simple" priors. What other "plausible" choices for priors do we have? Among the plethora of choices we will address only one - prior in which the uncertainties about P_A and P_B are independent. This formally can be expressed as:

$$f_{P_A, P_B}(\bullet, \bullet) = f_{P_A}(\bullet) f_{P_B}(\bullet) \tag{12}$$

The assumption we make can be spelled out as: "Even if I were told the value of P_A , this knowledge would not change my uncertainty about P_B (and vice versa). This assumption is not necessarily the only one plausible. It may also be plausible to believe in a particular form of dependence between uncertainties, represented

by $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$. If this is the case, $f_{P_A, P_B}(\bullet, \bullet) \neq f_{P_A}(\bullet)f_{P_B}(\bullet)$, then the dependence should be quantified, which makes the definition of the prior much more difficult¹.

Notice that (12) is not equivalent to assuming independence between the failures of the two channels, which is well known to be unreasonable both empirically, e.g. [Knight & Leveson 1986b], and theoretically, e.g. [Eckhardt & Lee 1985], [Littlewood & Miller 1989]. In fact, (12) says *nothing* about the probability of coincident failure, P_{AB} . The uncertainty about the values of P_{AB} are represented by the conditional distribution, $f_{P_{AB}}(\bullet | P_A = y, P_B = z)$. Certainty here can be expressed as:

$$f_{P_{AB}}(\bullet | P_A = y, P_B = z) = \delta(Kyz) \quad (13)$$

where $K \in [0, \max(y, z)]$. Independence of failures is represented by the special case $K=1$. Indeed, in this case the probability of system failure with certainty will be equal to the product of the probabilities of channel failures, $P_A = y, P_B = z$.

The assumption of certainty is extremely strong - it does not change whatever we observe from a system in operation, and is, therefore, implausible for a Bayesian inference, using which we expect to *learn* something about the system from observing it in operation. Therefore, (13) should be avoided in defining a plausible prior.

Since, by definition, the system can not be less reliable than the more reliable channel, once the channel failure probabilities are known, $(P_A = y, P_B = z)$, P_{AB} can take non-zero values within $[0, \min(P_A, P_B)]$ only.

"Indifference" between the possible values of P_{AB} can be a plausible prior. In this case:

$$f_{P_{AB}}(\bullet | P_A = y, P_B = z) = \frac{1}{\min(y, z)}.$$

The expected value $E[P_{AB} | P_A, P_B] = \frac{1}{2} \min(y, z) > yz$ for realistically small y and z . In other words, "indifference" represents a (reasonable) belief that the system is worse than if the channels failed independently.

Plausible may be any other $f_{P_{AB}}(\bullet | P_A = y, P_B = z)$ defined on $[0, \min(P_A, P_B)]$, too. Selecting a particular shape for this conditional distribution allows us to represent different beliefs about the probability of system failure including channels' failures being negatively correlated, in which case $f_{P_{AB}}(\bullet | P_A = y, P_B = z)$ should be parameterised in such a way that:

$$E[P_{AB} | P_A = y, P_B = z] < yz$$

Defining the prior through conditional distributions, $f_{P_{AB}}(\bullet | P_A, P_B)$, clearly poses difficulties, especially if we want to avoid "indifference". These come from the fact that the conditional distributions should be defined for every possible pair of values of the channels probabilities of failure, (P_A, P_B) , which may be hard.

Another difficulty comes from the fact that in the general case of prior, $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$, the inference may only be possible numerically. Inferring the posterior with high precision may be prohibitively expensive in terms of computational efforts required and the solution is not always feasible. The known methods of calculating multiple definite integrals, for example, are not guaranteed to converge. Therefore, it is important to reduce the amount of calculations by cutting away those intervals of integration which have a negligible effect on the calculations. In our particular case we can truncate the tail of the $f_{P_A, P_B}(\bullet, \bullet)$ beyond some known P_{Amax} and P_{Bmax} which remain reasonable upper bounds on P_A and P_B even for the posteriors. In some cases these may be possible to determine. For instance, an organisation may have records from previous similar single channel software systems and know some θ which can confidently be assumed to be an upper bound on the probabilities of failure of newly developed software. Then θ can be used as an upper bound on P_A and P_B of the newly created two-channel system. The definition of the marginal distributions $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$ thus should be on the interval $[0, \theta]$ (instead on $[0, 1]$).

In summary, the following *algorithm* of defining a prior $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$ can be suggested:

¹ There is good reason for assuming independence. Increasing the amount of testing (observation) of the system leads to $f_{P_A, P_B}(\bullet, \bullet)$ becoming asymptotically a Dirac function. Similarly, $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$ will both become Dirac functions. Formally, in this case, $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$ can be treated as independent. The point is, whatever we justify as a reasonable dependence between $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$, this will disappear as the time for observing the system increases. The question is, of course, if such asymptotic 'independence' is a sufficient argument for us to assume independence between prior $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$.

1. Justify independence between $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$, if you can.
2. If this can be done:
 - 2.1. Define $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$. For instance, use Bayesian inference on channels, [Miller *et al.* 1992], [Littlewood & Wright 1997].
 - 2.2. Use (12) to define the joint distribution $f_{P_A, P_B}(\bullet, \bullet)$.
3. If (1) is unreasonable - define the dependence between $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$.
4. Truncate $f_{P_A, P_B}(\bullet, \bullet)$ beyond upper bounds P_{Amax} and P_{Bmax} , if these can be defined.
5. Justify a particular shape for $f_{P_{AB}}(\bullet | P_A = y, P_B = z)$. Try 'indifference' first.
6. The full prior is defined by (10).

3.4 Priors allowing conservative claims for system reliability

Here we show that a prior based on the assumption of P_A and P_B being known with certainty can be used as conservative in many cases, and is therefore useful despite the problems described in section 3.2.

Clearly for every prior $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$ with marginal distribution of the probability of system failure, $f_{P_{AB}}(\bullet)$, for which upper bounds on the probabilities of channel failures, P_{amax} and P_{bmax} , are defined, a new prior could be defined, $f_{P_{AB}, P_A, P_B}^*(\bullet, \bullet, \bullet) = f_{P_{AB}}(\bullet) \delta(P_{Amax}, P_{Bmax})$, where $\delta(P_{Amax}, P_{Bmax})$ is a Dirac function taking a non-zero value only at (P_{Amax}, P_{Bmax}) and zero elsewhere. Hence, $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$ and $f_{P_{AB}, P_A, P_B}^*(\bullet, \bullet, \bullet)$ are *comparable* in the sense that the uncertainty about *system reliability* with both priors is the same, represented by $f_{P_{AB}}(\bullet)$.

Now we want to compare the posterior marginal distributions, $f_{P_{AB}}(\bullet | n, r_1, r_2, r_3)$ and $f_{P_{AB}}^*(\bullet | n, r_1, r_2, r_3)$, derived with the comparable priors, $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$ and $f_{P_{AB}, P_A, P_B}^*(\bullet, \bullet, \bullet)$, for the same observation ($n : r_1, r_2, r_3$). We illustrate the relationship between the two posterior distributions in Table 3 in which their percentiles are shown. The percentiles of the prior and the black-box posteriors are also included.

$f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$ used in the examples was defined with the following additional assumptions:

- $f_{P_A, P_B}(\bullet, \bullet) = f_{P_A}(\bullet) f_{P_B}(\bullet)$. The marginal distributions $f_{P_A}(\bullet)$ and $f_{P_B}(\bullet)$ were assumed to be Beta distributions, $f_{P_A}(\bullet) = \text{Beta}(x, 20, 10)$ and $f_{P_B}(\bullet) = \text{Beta}(x, 20, 20)$ within the interval $[0, 0.01]$, respectively. $P_{Amax} = P_{Bmax} = 0.01$.
- "indifference" between the possible values of P_{AB} , i.e.:

$$f_{P_{AB}}(\bullet | P_A, P_B) = \frac{1}{\min(P_A, P_B)} \text{ within } [0, \min(P_A, P_B)] \text{ and } 0 \text{ elsewhere.}$$

The system was subjected to 4000 tests, and the failures of the system, channel A and B, represented by r_1, r_2 and r_3 , respectively, are shown in the table. The selected examples cover a range of interesting testing results: no failure, no system failure but channel failures, system failure only, a combination of system and channel failures.

The percentiles reveal stochastic ordering of the posteriors, $f_{P_{AB}}(\bullet | data)$ and $f_{P_{AB}}^*(\bullet | data)$ whatever the observations. The ordering is in favour of $f_{P_{AB}}(\bullet | data)$: the probability that the system reliability will be better than any reliability target will be greater with $f_{P_{AB}}(\bullet | data)$ than with $f_{P_{AB}}^*(\bullet | data)$. In other words, if we use $f_{P_{AB}, P_A, P_B}^*(\bullet, \bullet, \bullet)$ instead of $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$, the confidence that we will be prepared to place on a claim that system reliability is better than a predefined reliability target will be lower than if the claim would be based on using $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$, i.e. the predictions based on $f_{P_{AB}, P_A, P_B}^*(\bullet, \bullet, \bullet)$ will be conservative.

This observation, if universally true, is a good news for everyone willing to use Bayesian inference for conservative reliability assessment of a two-channel system. It suggests a relatively easy way of avoiding the difficulty in defining the full $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$. In order for us to be conservative, we need to define

$f_{P_{AB}, P_A, P_B}^*(\bullet, \bullet, \bullet)$ only which is not much more difficult than defining the marginal probability of system failure, $f_{P_{AB}}(\bullet)$. The only extra parameters it requires are the upper bounds P_{Amax} and P_{Bmax} , which, as stated above, seems feasible a task.

The data presented in Table 2 illustrate only a small part of the experiments we carried out with different priors and testing results. The observations we had were consistent with those presented: when $f_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$ was used as a prior the posterior probability of system failure was stochastically smaller than the posterior obtained with $f^*_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$. Proof that the ordering between the two posteriors will always be in place (no matter the type of the prior and the testing results) is yet to be found. This is beyond the scope of this paper.

Table 3

Percentiles		10%	50%	75%	90%	95%	99%
Prior		0.00025	0.00225	0.0035	0.00435	0.00485	0.0056
$r_1 = 0, r_2 = 0$	$f_{P_{AB}}(\bullet n, r_1, r_2, r_3)$	0.00278	0.003525	0.00406	0.00445	0.00473	0.00124
$r_3 = 0$	$f^*_{P_{AB}}(\bullet n, r_1, r_2, r_3)$	0.0055	0.00635	0.0067	0.00705	0.00725	0.0076
Black-box posterior		0	0	0.000125	0.00035	0.0005	0.001
$r_1 = 1, r_2 = 0$	$f_{P_{AB}}(\bullet n, r_1, r_2, r_3)$	0.0029	0.00372	0.0042	0.00455	0.0048	0.00525
$r_3 = 0$	$f^*_{P_{AB}}(\bullet n, r_1, r_2, r_3)$	0.0055	0.00638	0.0067	0.00685	0.00735	0.00763
Black-box posterior		0	0.00033	0.00058	0.00092	0.00115	0.00173
$r_1 = 1, r_2 = 24$	$f_{P_{AB}}(\bullet n, r_1, r_2, r_3)$	0	0.0001	0.00035	0.00062	0.00076	0.001245
$r_3 = 20$	$f^*_{P_{AB}}(\bullet n, r_1, r_2, r_3)$	0.00035	0.00123	0.00178	0.00235	0.00275	0.0033
$r_1 = 0, r_2 = 20$	$f_{P_{AB}}(\bullet n, r_1, r_2, r_3)$	0	0	0.00022	0.00049	0.00067	0.00112
$r_3 = 15$	$f^*_{P_{AB}}(\bullet n, r_1, r_2, r_3)$	0.00015	0.00135	0.00224	0.0029	0.00331	0.004

Table 3: Percentiles of the prior marginal distribution $f_{P_{AB}}(\bullet)$ and the following three posterior distributions:

$f_{P_{AB}}(\bullet | n, r_1, r_2, r_3)$, $f^*_{P_{AB}}(\bullet | n, r_1, r_2, r_3)$ and black-box posterior, are shown.

The usefulness of the conservative prior $f^*_{P_{AB}, P_A, P_B}(\bullet, \bullet, \bullet)$ seems *limited*. Indeed, for the important special case of testing which does not reveal any failure ($r_1 = 0, r_2 = 0, r_3 = 0$), the conservative result is too conservative and hence not very useful: the posterior will be more pessimistic than the prior, which is due to the phenomenon explained in section 3.2. This fact reiterates the main point of this paper: prior elicitation is difficult and there does not seem to exist easy ways out of this difficulty.

The last two cases presented in Table 2 with testing results ($r_1 = 1, r_2 = 24, r_3 = 20$) and ($r_1 = 0, r_2 = 20, r_3 = 15$), respectively, illustrate the interplay between the black-box and the white-box inferences. In both cases there is ordering between the full white-box and the black-box posteriors but the ordering is in reverse order. In the case with a single system failure the black-box posterior is more pessimistic than the full white-box posterior, while in the case with no system failure the black-box posterior gives more optimistic prediction about system reliability. In the case ($r_1 = 1, r_2 = 24, r_3 = 20$) we have evidence of negative correlation between failures of channels. The expected number of system failures under the assumption of independence is 1.4 in 4000 tests, while we only observed 1. In the case ($r_1 = 0, r_2 = 20, r_3 = 15$) even though no system failure is observed the evidence of negative correlation is weaker (lower number of individual failures is observed). In result, the white-box prediction is worse than the black-box one.

In summary, using the black-box inference for predicting system reliability may lead to overestimating or underestimating the system reliability.

4. Conclusions

1. We described the mathematically correct way of doing Bayesian inference of the reliability of a 1-out-of-2 on-demand system using the testing results. This development, hardly innovative, nevertheless is important because there exist attempts in the literature to use Bayesian inference with unjustified simplified priors. The interested community can be misled by such attempts to believe that Bayesian inference can be easily applied. We showed that even plausible simplifying assumptions used in defining the prior can have counterintuitive consequences.
2. We analysed in depth some priors which are convenient for the inference in the case of a two channel system, which may be of interest for the assessors of software safety, and showed their limitations, which are new results.
3. We illustrated that a conservative prior for the two channel system can be defined with relative ease if upper bounds on channel probabilities of failure are known with certainty. The level of conservatism, however, can be too great, in particular in the important case when no failures are revealed by the testing.

Open problems:

1. Elicitation of the full prior which can be trusted is still an open problem.
2. Formal proof of the conservatism of the prior of section 3.4 is needed.
3. For the case of no failures during the testing *reducing the level of conservatism* and making it useful upper bound on system reliability will make the Bayesian inference much more attractive for the practical software reliability/safety assessment.

Acknowledgement

This work was supported partially by British Energy Generation (UK) Ltd. under the 'Diverse Software Project' (DISPO) and by EPSRC under the 'Diversity In Safety Critical Software' (DISCS) project. Authors want to thank Martin Newby for helpful discussions.

References

- [Briere & Traverse 1993] D. Briere and P. Traverse, "Airbus A320/A330/A340 Electrical Flight Controls - A Family Of Fault-Tolerant Systems", in *23rd International Symposium on Fault-Tolerant Computing (FTCS-23)*, Toulouse, France, 22 - 24, pp.616-623, IEEE Computer Society Press, 1993.
- [Brocklehurst *et al.* 1990] S. Brocklehurst, P. Y. Chan, B. Littlewood and J. Snell, "Recalibrating software reliability models", *IEEE Transactions on Software Engineering*, SE-16 (4), pp.458-470, 1990.
- [Butler & Finelli 1991] R. W. Butler and G. B. Finelli, "The Infeasibility of Experimental Quantification of Life-Critical Software Reliability", in *ACM SIGSOFT '91 Conference on Software for Critical Systems*, in *ACM SIGSOFT Software Eng. Notes*, Vol. 16 (5), New Orleans, Louisiana, pp.66-76, 1991.
- [Eckhardt & Lee 1985] D. E. Eckhardt and L. D. Lee, "A theoretical basis for the analysis of multiversion software subject to coincident errors", *IEEE Transactions on Software Engineering*, SE-11 (12), pp.1511-1517, 1985.
- [Johnson & Kotz 1972] N. L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions*, Wiley Series in Probability and Mathematical Statistics, 4, 333p., John Wiley and Sons, Inc, 1972.
- [Kantz & Koza 1995] H. Kantz and C. Koza, "The ELEKTRA Railway Signalling-System: Field Experience with an Actively Replicated System with Diversity", in *25th IEEE Annual International Symposium on Fault-Tolerant Computing (FTCS-25)*, Pasadena, California, pp.453-458, IEEE Computer Society Press, 1995.
- [Knight & Leveson 1986a] J. C. Knight and N. G. Leveson, "An empirical study of failure probabilities in multi-version software", in *16th International Symposium on Fault-Tolerant Computing (FTCS-16)*, Vienna, Austria, pp.165-170, IEEE Computer Society Press, 1986a.
- [Knight & Leveson 1986b] J. C. Knight and N. G. Leveson, "An Experimental Evaluation of the Assumption of Independence in Multi-Version Programming", *IEEE Transactions on Software Engineering*, SE-12 (1), pp.96-109, 1986b.
- [Littlewood & Miller 1989] B. Littlewood and D. R. Miller, "Conceptual Modelling of Coincident Failures in Multi-Version Software", *IEEE Transactions on Software Engineering*, SE-15 (12), pp.1596-1614, 1989.
- [Littlewood & Strigini 1993] B. Littlewood and L. Strigini, "Validation of Ultra-High Dependability for Software-based Systems", *Communications of the ACM*, 36 (11), pp.69-80, 1993.
- [Littlewood & Wright 1997] B. Littlewood and D. Wright, "Some conservative stopping rules for the operational testing of safety-critical software", *IEEE Transactions on Software Engineering*, 23 (11), pp.673-683, 1997.
- [Miller *et al.* 1992] K. W. Miller, L. J. Morell, R. E. Noonan, S. K. Park, D. M. Nicol, B. W. Murrill and J. M. Voas, "Estimating the Probability of Failure When Testing Reveals No Failures", *IEEE Transactions on Software Engineering*, 18 (1), pp.33-43, 1992.
- [Musa *et al.* 1987] J. D. Musa, A. Iannino and K. Okumoto, *Software Reliability: Measurement, Prediction, Application*, 621p., McGraw-Hill International Editions, 1987.
- [Voges 1988] U. Voges (Ed.), *Software diversity in computerized control systems*, Dependable Computing and Fault-Tolerance series, 2, Springer-Verlag, Wien, 1988.