LA-UR-05-2252

Approved for public release; distribution is unlimited.

Title:

Reliability Modeling Using Both System Test and Quality Assurance Data

Author(s):

Christine Anderson-Cook, Todd Graves, Nicolas Hengartner, Richard Klamann, Andrew Koehler, Alyson Wilson
Los Alamos National Laboratory, Los Alamos, NM

Greg Anderson, George Lopez
Marine Corps Program Department, Fallbrook, CA

Submitted to:

Military Operations Research

Los Alamos
NATIONAL LABORATORY

Reliability Modeling Using Both System Test and Quality Assurance Data

Christine Anderson-Cook, Todd Graves, Nicolas Hengartner, Richard Klamann,

Andrew Koehler, Alyson Wilson

Los Alamos National Laboratory, Los Alamos, NM

Greg Anderson, George Lopez

Marine Corps Program Department, Fallbrook, CA

Abstract

A system with several components may undergo full system pass/fail testing as well as quality assurance testing at the single component level. The component tests are informative about system reliability, although they measure different things than the full system tests. We present a Bayesian framework for integrating the two types of test data for better reliability estimates. Our formulation allows the reliability to depend on covariates such as age. One result of the inference is a better understanding of the relationship between component tests and system performance. We illustrate the ideas using a small subsystem of a larger (proprietary) system.

1 Introduction

The U.S. Armed Forces maintain multiple stockpiles of weapons, which need to be monitored, managed and maintained over time. Typically, these weapon systems degrade over time, and it is important to be able to estimate the fraction of units within a given stockpile that will perform as expected at any given time. Current strategies for obtaining an accurate estimate of the reliability rely on full system tests on a sample of units from the stockpile, and using the results of full-system tests to obtain estimates. While this approach is commonly used and does provide the needed estimates, it tends to be an expensive approach to answering this question.

Complex weapon systems are frequently comprised of well-defined subsystems, which in turn are composed of many components. The reliability of such systems can more effectively and efficiently be assessed by using all possible sources of data. In addition to the full system tests, which are typically destructive and hence very expensive, it may be possible to obtain measurements for subsystems or components to verify that key elements are in appropriate working order. With a system model that details how the various elements of the system are interrelated and affect the system performance, this subsystem data can be utilized to increase precision of the estimated system reliability.

As weapons are produced over a number of years, it is common for various components to be upgraded or modified. This results in several variants of the weapon, which need not have the same reliability characteristics. An alternate source of variants would be if several manufacturers produced one of the components of the system.

1.1 The Mini-missile system

Because of the proprietary nature of the full systems for which this methodology was developed, we present our approach for a greatly simplified system. While the full system has many more components and variants, the simplified system both retains all the interesting analytical challenges of the full proprietary system and allows us to present the new approach in a manageable form. We will refer to this system as "Mini-missile:" it is a series system of two components, a launch motor (LM) and a flight motor. The flight motor comes in two variants, FM1 and FM2. The system being in series implies that for a flight to be successful, both the launch motor and the flight motor need to work appropriately. In this paper we will refer to the LM as component 1, the FM1 as component 2, and the FM2 as component 3.

1.2 Data

Two types of testing are performed on the missiles: destructive full-system tests, which involve firing of the round, and quality assurance testing, which involves separate testing of the individual components of the system in a laboratory. We assume that the missile needs to be disassembled in order to perform quality assurance tests. As a result, we will view the quality assurance tests to be destructive as well.

During a test flight, these two components perform sequentially, so that normally in the event of a failed system test, it is sometimes possible to diagnose which motor failed, although we allow the possibility that some system failures cannot be resolved further. Since the motors are connected in series, it will normally be impossible to determine the status of the FM in the event of an LM failure. Hence variants of the systems will be comprised either of components 1 and 2, or of 1 and 3. The result of a full-system test therefore yields a success, failure or undetermined for each of the components, as well as an overall success or failure for the entire system.

Alternatively, these components can undergo quality assurance testing to verify that they meet engineering specifications. To this end, the missiles are disassembled and the various components are tested. FM1's are measured for ignition and total impulse, FM2's are measured for these two plus activation time, and launch motors are measured for these three plus maximum thrust. In total, then, we have nine different quantities measured for Mini-missile. Each of these measurements is recorded on a continuous scale and has prescribed engineering specification limits: they are either to be below, above, or between known operational limits. Both types of tests are destructive, so that we never observe both a component's spec measurements and the exact same component's performance in a flight test.

Finally, for each component test and for each flight test, we have recorded the age of the missile. Our analysis will enable us to identify which components degrade with age and will assess the overall reliability of the full system as a function of age. This will enable us to develop maintenance and retirement plans of items in the stockpile.

The framework discussed in this paper allows reliability to depend on other covariates. In particular, the full missile analysis studied the effects of firing the missile in cold conditions, taking the missile to and/or deploying it in the desert, and different component manufacturers.

1.3 Related work

System reliability is a popular topic, with a number of methods for modeling the system structure, including reliability block diagrams, event trees, fault trees, and Bayesian networks. See Rausand and Hoyland (2004) for more details on any of these approaches.

Many authors have considered assessing the reliability of full systems from both component and system level pass–fail data. In this context, Martz, Waller and Fickas (1988) and Martz and Waller (1990) considered estimating the reliability from binomial data, with the added twist of integrating expert opinion at different levels (which is somewhat different from our problem of integrating two types of data). Graves and Hamada (2004) provide a fully Bayesian analysis of the same data sets. Other authors propose methods for evaluating or bounding moments of the system reliability posterior distribution (Cole (1975), Mastran (1976), Dostal and Iannuzzelli (1977), Mastran and Singpurwalla (1978), Barlow (1985), Natvig and Eide (1987), and Soman and Misra (1993)). These moments can also be used in the beta approximations employed by Martz, Waller and Fickas (1988) and Martz and Waller (1990). Soman and Misra (1993) proposed a distributional approximation based on a maximum entropy principle. Many reliability models do not consider prior expert opinion and data at multiple system levels. Springer and Thompson (1966, 1969) and Tang, Tang and Moskowitz (1994, 1997) provide exact or approximated system reliability distributions obtained by propagating the component posteriors through the system structure. Thompson and Chang (1975), Chang and Thompson (1976), Lampkin and Winterbottom (1983), and Winterbottom (1984) use approximations for exponential lifetimes rather than binomial data.

2 Data and their statistical models

This section introduces our statistical approach and discusses how we combine the disparate pieces of information.

2.1 Statistical models

As with all statistical models for complex systems, specification of the connections between components that form the entire system is an important issue. For our simple model, we assume that both components, the LM and FM, need to work appropriately for the entire system to work. With more complicated systems, block diagrams, fault trees or Bayesian networks are sometimes used to describe the interrelationships between the components of the system. Once this structure is defined, that representation allows for statistical models to be developed and be used to incorporate data to make inferences through the likelihood function. Specifically, the statistical model defines the probability of observing all possible data, with this probability depending on some unknown parameters. For example, the probability of observing k successes in n binomial trials with success probability p is $f(x|p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$ for $k = 0, 1, \ldots, n$. Here p is an unknown parameter. The likelihood function is this same expression considered as a function of the unknown parameter p for a fixed value of the data (k). If multiple independent data sets are analyzed together, their likelihood functions are multiplied.

The unknown parameters are estimated from information contained in the data alone, or in combination with expert knowledge and the observed data. Classical methods focus on estimating the unknown parameters by maximizing the likelihood function given the data observed. Bayesian methods require the encapsulation of the expert knowledge, if any, into a *prior distribution* for the unknown parameters. Once the data become available, the prior distribution is combined with the observed data to become a *posterior distribution* for the unknown parameters. As the amount of observed data increases, the results from Bayesian and classical approaches tend to agree, as the value of expert opinion is dominated by evidence from the data. The Bayesian approach may have substantial advantages when the amount of observed data is small to moderate. In such cases, the knowledge summarized in the prior distribution represents important additional information.

Our model for the Mini-missilemust accommodate both flight data (pass/fail for the entire system and possibly partial pass/fail information for some components using forensics) and continuous component specification data.

2.2 Component test data

We first transform the component test data. They are transformed so that for all of the 9 specification measurements, larger values are always thought to imply higher reliability, and that the stated boundary specification value is always transformed to zero. For example, if the spec states that the measurement should be greater than 10, transform the measurement by subtracting 10. If it should be less than 20, transform by applying the function $-(x-20)$. Finally, if it should be between 10 and 20, apply the function $5 - |x - 15|$.

Denote the ith component test measurement by S_i, for $1 \leq i \leq N_1$, where N_1 is the total number of spec measurements in the data set. Suppose that this ith measurement corresponds to the L_ith spec: for example, $L_i = 1$ means that the measurement is ignition time for a launch motor, while $L_i = 9$ corresponds to the total impulse measurement for an FM2 component. Denote the age of the component in the ith measurement by A_i. We assume that these measurements follow normal linear regression relationships: for $1 \leq i \leq N_1$, assume

$$S_i \sim N(\alpha_{L_i} + \beta_{L_i} A_i, \gamma_{L_i}^2), \tag{1}$$

where $\alpha_1, \ldots, \alpha_9$, $\beta_1, \ldots, \beta_9$, and $\gamma_1, \ldots, \gamma_9$ are all unknown parameters to be estimated using data. (The notation $X \sim N(\mu, \sigma^2)$ means that the random variable X has a normal distribution with mean μ and variance σ^2.) The α's are mean values of the spec measurements at age zero, the β's measure how much these mean values change with age, and the γ's are standard deviations of the spec measurements. In light of the transformations, $\beta_j < 0$ implies age related degradation, whereas $\beta_j > 0$ implies an improvement.

These linear regressions could be done individually for each spec, and that would provide valuable information about how components age relative to their specifications. However, this only helps us predict whether components will be up to spec. We want to predict the probability that a component of a given age will work properly in a flight test, including estimating the relationship between spec measurement and success in flight. To do this we need to integrate these component test data with flight test data. In order to do this, we need to assume a particular stochastic relationship between falling within the specification values and that aspect of the component working appropriately in a flight test. We estimate the parameters of this relationship based on data. This assumption is necessary since because of the destructive nature of both types of testing, it is not possible for both flight and specification data to be obtained on a single missile.

2.3 Flight test data and the surrogacy assumption

To relate the spec data to flight test data, consider a component with M related specs $S_1, \ldots, S_M$.Write Z for the indicator that the component works in a test. We assume that the probability that, at age A, the component would work properly ($Z = 1$), given $S_1, \ldots, S_M$, is

$$P\{Z = 1|S_1, \ldots, S_M, A\} = \prod_{j=1}^{M} \Phi\left(\frac{S_j - \theta_j}{\sigma_j}\right). \qquad (2)$$

Φ is the standard normal distribution function, $\Phi(x) = \int_{-\infty}^{x}(2\pi)^{-1/2}\exp(-t^2/2)dt$; it has a sigmoidal shape and increases from zero to one. The interpretation of a single Φ term is that the reliability drops below 0.5 when the spec value drops below θ_j, and σ_j parameterizes how steep the decline in reliability is with decreasing values of the spec. We expect that θ_j and σ_j will be related to the published levels that the spec should attain, but we will estimate their values from data in case the published specs are excessively conservative or lenient. Note that age does not appear on the right side of this expression: this is because we are making a "surrogacy assumption" that the only way that component success depends on age is through the spec measurements.

For the flight test data, we do not get to observe the spec measurements, so we want to be able to write the conditional probability $P\{Z = 1|A\}$ of observing a given data pattern given the age A alone. Letting U_i be a synthetic normal random variable with mean θ_i and variance σ_i^2 independent of S_i, and assuming that the spec measurements are independent with joint density function written generically as $p(s_1, \ldots, s_M|A) = \prod_{i=1}^{M} p(s_i|A)$, we calculate

$$P\{Z = 1|A\} \;=\; \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P\{Z = 1|S_1 = s_1, \ldots, S_M = s_M, A\}p(s_1, \ldots, s_M|A)ds_1 \ldots ds_M$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{M} \left\{ \Phi\left(\frac{s_i - \theta_i}{\sigma_i}\right) p(s_i|A) \right\} ds_1 \ldots ds_M \\
&= \prod_{i=1}^{M} \int_{-\infty}^{\infty} \Phi\left(\frac{s_i - \theta_i}{\sigma_i}\right) p(s_i|A) ds_i \\
&= \prod_{i=1}^{M} \int_{-\infty}^{\infty} P\{U_i \leq s_i\} p(s_i|A) ds_i \\
&= \prod_{i=1}^{M} P\{U_i \leq S_i|A\} \\
&= \prod_{i=1}^{M} P\{U_i - S_i \leq 0|A\}.
\end{aligned}
$$

Now, the combination of assumption (2) and the normal linear models for the spec measurements causes a beneficial simplification, and this probability is

$$
\prod_{i=1}^{M} \Phi\left(\frac{\alpha_i + \beta_i A - \theta_i}{\sqrt{\gamma_i^2 + \sigma_i^2}}\right)
$$

since $U_i - S_i \sim N(\theta_i - \alpha_i - \beta_i A, \gamma^2 + \sigma^2)$.

This argument works in exactly the same way if there are more covariates, and it is also possible to work it out algebraically by completing the square. Note that potentially this argument could be used with other distributional families for S_i if one were willing to assume that the Φ function could be replaced by the distribution function of a random variable U_i such that $U_i - S_i$ has a tractable distribution. This result is important because now we can use this expression in a likelihood function for observed flight test data, it is tractable, and it does not contain the unobservable S's.

Although not applicable to the example system described in this paper, the technique can address components with no spec measurements. In this case, we assume that a component of age A works with probability $\Phi(\alpha_i + \beta_i A)$, since if a θ, γ, or σ were to be added to the model, it would become nonidentifiable in the absence of spec data.

Finally, we describe how these probabilities are used in the likelihood function for flight test data. Consider a missile i of variant 1, which has an LM (component 1) and an FM1 (component 2) but no FM2. This missile is of age A_i. Denote the probability of successful operation of component j by p_{ji}, and recall that

$$p_{1i} = \prod_{k=1}^{4} \Phi\left(\frac{\alpha_k + \beta_k A_i - \theta_k}{\sqrt{\gamma_k^2 + \sigma_k^2}}\right), \tag{3}$$

and

$$p_{2i} = \prod_{k=5}^{6} \Phi\left(\frac{\alpha_k + \beta_k A_i - \theta_k}{\sqrt{\gamma_k^2 + \sigma_k^2}}\right). \tag{4}$$

For component 3, the product ranges from 7 to 9. If the ith flight test is a success, the likelihood function contains a term of $p_{1i}p_{2i}$; if the test is a failure due to the flight motor but the launch motor worked, the term is $p_{1i}(1 - p_{2i})$; if the launch motor fails and hence the flight motor is inconclusive, the term is $1 - p_{1i}$, and if the test fails but it is not known which component is responsible, the term is $1 - p_{1i}p_{2i}$. The flight test data part of the likelihood function is the product of all these terms. For a missile of variant 2, p_{3i} is used in place of p_{2i}.

The concept of surrogate variables was introduced by Prentice (1989) and refined by Pepe (1992) in the context of medical studies, where it is also called *non-differential measurement error* by Carroll et al. (1995). Pepe and Fleming (1991) and Wang and Pepe (2000) call the resulting data structure *errors in variables with validation sample* and have considered estimation of θ when both the surrogate A and performance S are discrete random variables taking on finitely many values. As we demonstrated, extensions to continuous performance and surrogate variables is straightforward if a parametric model for the conditional distribution of S given A is available. However, given the complexity of the relationship between the performance and the surrogate variables, such distributions may be hard to specify. In this context, Turnbull and

Jiang (2004) and Jiang and Turnbull (2003) consider estimates of θ derived by a generalization of the method of moments. While their estimators are consistent and asymptotically Gaussian, it is unlikely that they are efficient. Finally, Hengartner (2005) derives an efficient estimate for θ for continuous covariates and surrogates.

3 Bayesian analysis

Gelman *et al* (2003) provides a comprehensive introduction to Bayesian analysis. For the sake of completeness, this section discusses the Bayesian approach to statistics, which quantifies all uncertainties, including uncertainties about parameters, using probability distributions. The Bayesian approach involves combining the likelihood of the data with a *prior distribution* (see §3.1). The prior can either be a convenient way to incorporate information extraneous to the data set, or, as in the present case, can be chosen to have a small effect on the conclusions. Bayesian statistics requires calculation or estimation of integrals (§ 3.2 rather than maximization, and the YADAS software (§3.3) was instrumental in these estimations.

3.1 Prior distributions

To begin a Bayesian analysis, one specifies a probability distribution for the unknown parameters: conceptually, this distribution quantifies the uncertainty about the unknown parameters before collecting the data. Sometimes this distribution is taken to be highly informative in the sense that some possible values are given much lower weight than others and the distribution as a whole is weighted as heavily as a number of data points. It is also common to use noninformative priors, which are flat and/or less influential than data points. An example of a procedure for developing an informative prior is as follows: consider a particular hypothetical spec measurement (say the jth measurement) evaluated at age zero. α_j is the average spec measurement at that age. Suppose that the best guess for this mean is 5, but "it could be as low as 1." Assuming that the normal distribution is an appropriate choice for α_j and interpreting the value of 1 as the fifth percentile of the prior distribution, one can use a $N(5, 2.43^2)$ prior distribution for α_j. Alternatively, using a prior standard deviation larger than 2.43 should be less informative and should have a smaller effect on the conclusions. In our analysis, we took the noninformative approach, using normal priors with mean zero and relatively large variance for the location and slope parameters (α's, β's, and θ's), and using exponential priors with mean one for the scale parameters (σ's and γ's).

3.2 Posterior distributions, and how to estimate them

The information in the raw data is captured by the likelihood which, when multiplied by the prior, becomes proportional to the posterior distribution. While the prior captures the uncertainty about the unknown parameters before we see the data, the posterior distribution embodies the uncertainty after we see the data. Quantities of interest are integrals (expected values) of functions of unknown parameters with respect to the posterior distribution. In nontrivial analyses, these integrals cannot be obtained exactly and must be estimated. Monte Carlo is one approach to solving these complex integrals numerically that is relatively easy to implement. The idea is to draw a sample from the posterior distribution with which the average of a function is calculated. This approach is particularly appealing, because the Monte Carlo sample permits not only the estimation of the integral of any function, but also allows estimation of quantiles and other quantities that are not obviously themselves integrals with respect to the posterior distribution. However, independent sampling from the posterior distribution cannot be done exactly except in special cases. Bayesian statisticians rely on a class of algorithms called Markov chain Monte Carlo (MCMC), popularized by Gelfand and Smith (1990), where the Markov chain term refers to the fact that the probability distribution of the next sample depends on the last sample.

Gibbs samplers are a popular subclass of MCMC algorithms. In a Gibbs sampler, the unknown parameter vector is divided into components that are iteratively updated by drawing new values of the component from its conditional distribution given the current values of the other components. If it is not possible to sample from these conditional distributions, a Gibbs sampler cannot be used. A more general building block of MCMC samplers is the Metropolis-Hastings algorithm, which relies on a "proposal distribution" to draw new values of the unknown parameter, and an acceptance/rejection calculation that either moves to this new value or stays put Algorithms called variously *Metropolis-within-Gibbs* or, more correctly, *variable-at-a-time Metropolis*, divide up the parameter vector as in Gibbs sampling, and replace the sample from a conditional distribution by a Metropolis–Hastings move.

3.3 The YADAS package

YADAS (Graves 2003a,b) is a highly extensible software system for creating MCMC algorithms. It is written in Java and its source code is freely available at `yadas.lanl.gov`, together with examples and documentation. The major strengths of YADAS are its flexibility in expressing new statistical models, and its versatility in the MCMC algorithms that can be applied to analyze the data. The first of these strengths was particularly important in the Mini-missile problem, as the model that integrated component test data with ambiguous system test data was new. The problem featured unknown parameters with high posterior correlation, which typically causes poorly performing MCMC algorithms due to inefficiencies in moving the parameters one at a time. YADAS facilitates the modification of poorly performing algorithms through the addition of steps that simultaneously move correlated parameters.

4 Results

In this section we present the results of the analysis. First, we discuss some alternative models, one of which requires slightly richer data than are actually available, and compare the estimated reliability functions for the system as a function of age. Next, we dig down to the component level (even below that, to the component property level) to investigate which properties of which components appear to age. Then, we explore a side benefit of this style of modeling: the study of how the various component test measurements appear to affect component reliability. All estimated posterior distributions reported in this paper are based on 10,000 MCMC iterations, preceded by a burn-in period of at least 200 iterations.

4.1 Comparison with alternative models

Here we discuss other models that could be used to analyze the data and how their results differ from our approach. We display reliability estimates and uncertainty estimates for the four analysis approaches, and for both variants, in Figure 1. The reliability estimates are posterior means of the probability that the system works, and the uncertainty intervals, shown using dashed-dotted lines, are 5th and 95th posterior percentiles of this probability.

1. A natural approach is to fit a logistic regression to the system pass-fail data, using age as the covariate, with the two variants treated separately. In other words, the probability of system success for a system of variant j and age A is

$$\frac{\exp(\alpha_j + \beta_j A)}{1 + \exp(\alpha_j + \beta_j A)}.$$

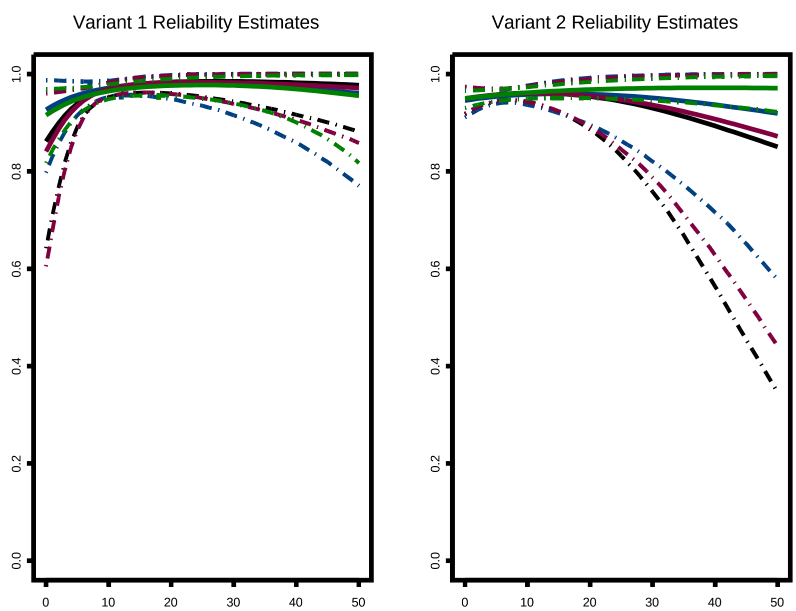

Figure 1: Posterior means and 5% and 95% confidence bounds for reliability, for four models and two variants (left: variant 1, right: variant 2). Blue: logistic regression at system level; black: probit regression at component level, with ambiguous failures, system tests only; red: same as black but with full data for the ambiguous failures imputed; and green: the model described in Section 2, which integrates component test data.

We fit Bayesian logistic regressions with fairly noninformative prior distributions for the α's and β's. This model can potentially be adequate for these data: in principle, it can fail to fit the data if different components age at different rates. This is a much greater danger for complex systems than for Mini-missile, which has just two components per variant. Also, it uses the data inefficiently: tests of one variant should be relevant to estimating the reliability of the other variant, and the component test data are not used at all. The simplicity of the model can contribute to small estimated uncertainties, but these could be misleading if there is lack of fit.

For variant 1, the posterior means (standard deviations) for α and β are 3.67 (0.41) and 0.072 (0.087). For variant 2, these quantities are 3.07 (0.20) and 0.026 (0.056). The positive means for the β's suggest slightly improved reliability with age, but since these means are small compared with their standard deviations, the evidence of this improvement is quite weak. (Reliability that improves with age is possible; it could reflect a superior production process in the past that later deteriorated, but if this were the case it would very questionable to conclude that reliability will continue to improve in the future. We stress that we don't have strong evidence that this is happening.) The estimates for reliability as a function of age are in blue in Figure 1: there is a small improvement over time, and uncertainty is much larger at ages far from the bulk of the data. The oldest missile tested was age 25, and for each variant, the oldest missile that failed was about fifteen years old. A substantial number of missiles of variant 1 older than fifteen years were tested, but few for variant 2. The median ages of missiles that went through system test were 10.5 years for variant 1 and 4.5 years for variant 2.

2. The next analysis would be natural if the data were available: this is an analysis

of a synthetic data set based on the actual data set, but where we have replaced the system tests that failed but where we don't know which components failed, with more informative test results. These imputed data were created based on estimated reliabilities from model (3) below. We then fit three probit regressions for the age-dependent reliability of each component. More informative results should imply an improvement in uncertainty, but only eight tests were changed, so this improvement should be small. We are still ignoring the component test data. The model assumes that the success probability for component j at age A_i is $\Phi(\alpha_j + \beta_j A_i)$. Posterior means (standard deviations) for α_1 and β_1 are 2.14 (0.10) and 0.014 (0.024), for component 2 they are 1.96 (0.20) and 0.10 (0.06), and for component 3 1.99 (0.11) and 0.004 (0.03). Still, reliabilities are not decreasing with age, and component 2 seems to be most likely to be improving, although this is still very tentative. Reliability and uncertainty curves for this methodology are in red in Figure 1.

3. Very similar to approach (2), but suitable for the data we have, is the model for the system test data discussed in Section 2, omitting the component test data. Because we don't analyze the component data here, the parameters in the probit regression for component success probability can be reduced: the probability of correct functioning of component j in test i, where the age is A_i, is again $p_{ji} = \Phi(\alpha_j + \beta_j A_i)$, but we depart from approach 2 by using the correct likelihood for the observed result of the test (for example, the probability of observing that at least one of components 1 and 2 failed in test i is $1 - p_{1i} p_{2i}$). Here the posterior means (standard deviations) are 2.12 (0.10) and 0.023 (0.025) for α_1 and β_1, 2.00 (0.22) and 0.084 (0.064) for α_2 and β_2, and 2.01 (0.12) and -0.002 (0.033) for α_3 and β_3. Note that all the standard deviations have increased a small amount relative to analysis (2), as would be expected. Reliability estimates

for this model are shown in black in Figure 1. Since there is little evidence that any component is aging at all, let alone that the components are aging at different rates, it is not surprising that model (1) would fit the Mini-missile data adequately. Such is not the case for the full system, however.

4. The fourth model is the one described in Section 2, and it is shown in green in Figure 1. This model incorporates the component tests, so it has the potential to increase the precision of the inference. One cost of this is the introduction of several more parameters: this model has 45 (an $\alpha, \beta, \theta, \gamma$, and σ for each of the nine component test types) while approach (3) had just six. This increased complexity should ensure that the data are fit adequately, but if this is more complexity than is needed, the estimates will be more uncertain than necessary. Figure 1 shows that the confidence bounds for reliability as a function of age are in general tighter for this model, especially for variant 2. In particular, the use of the component test data has greatly eased our concerns about aging of Variant 2, as can be seen by the much higher lower confidence bounds for the green model. The second type of flight motor (in variant 2) has considerably more component tests than the first type, about four times as many component tests per spec, so it in unsurprising that these data were much more powerful for variant 2.

4.2 Aging

Our analysis permits precise study of the aging behavior of the system. Not only are some components allowed to age while others do not, we are even capable of identifying which aspect of a component appears to age. Since we have one set of unknown parameters for each type of component test, the parameter function $\beta/\sqrt{\gamma^2+\sigma^2}$ for a given spec measurement measures the effect of one additional year

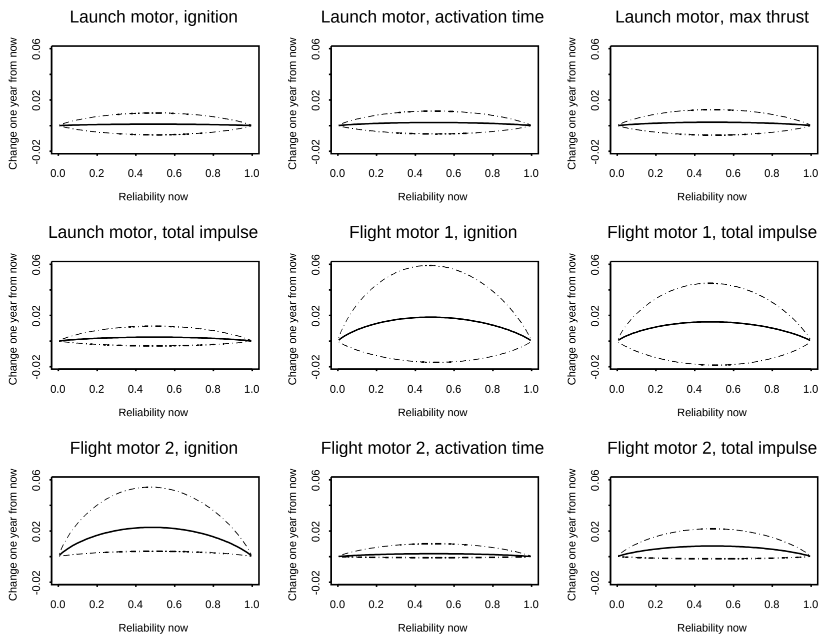

Figure 2: Illustration of estimated aging effects for the spec measurements. For each possible current reliability, we show posterior means and intervals for the change in reliability between now and next year.

of age on the inverse probit of the reliability. Figure 2 plots means, 5th, and 95th percentiles of the posterior distribution of $\Phi(\Phi^{-1}(p+\beta/\sqrt{\gamma^2+\sigma^2})-p$ as a function of p; this shows the change in reliability that will happen with one more year of age as a function of the current reliability. The seventh plot, ignition time for flight motor 2, is the only one with the lower confidence bounds above zero, suggesting improvement with age, although it is possible that this improvement is so small as to be negligible. Most plots are open to both the possibilities of improvement or worsening with age. The full system analysis identified strong evidence that some components worsen to some degree with age.

4.3 Relationship between specs and component success

One assumption in our model is that if spec measurement j would have been measured to be s, the probability that it would have caused a failure in a system test is

$$\Phi(\{s - \theta_j\}/\sigma_j).$$

We obtain a joint posterior distribution for θ_j and σ_j, so we can estimate the relationship between each spec measurement and system test failure using this expression. It should be noted that no data directly report on this relationship: θ_j and σ_j can only be estimated using the system test data, for which s is not observed. There are other parameters $\alpha_j, \beta_j, \gamma_j$ in the probability of success, and estimates of these can be obtained from component test data. As a result of the indirectness of our information about this relationship, it should be unsurprising that our uncertainties about these curves can be quite large, especially for values of s where data are sparse.

Estimates of this relationship are shown in Figure 3. The solid curve in the jth plot is the posterior mean of $\Phi(\{s - \theta_j\}/\sigma_j)$ as a function of s, and the dotted curves are the 5th and 95th percentiles of this quantity. The small solid curve in the corner is a density estimate of the spec data distribution, normalized to have a maximum value of 0.4. The vertical lines are at the point(s) of the published spec value: blue lines are used for the spec value under cold conditions when it differs from the normal condition spec. The plots suggest that the specs for tests 5 and 9 may be too conservative. It is possible that the specs for other tests are too lenient, but in no case is the information anything close to conclusive (in all cases the upper confidence bound for reliability at the spec value is essentially 1).

In our approach, we do assume that we know the direction of the relationship.

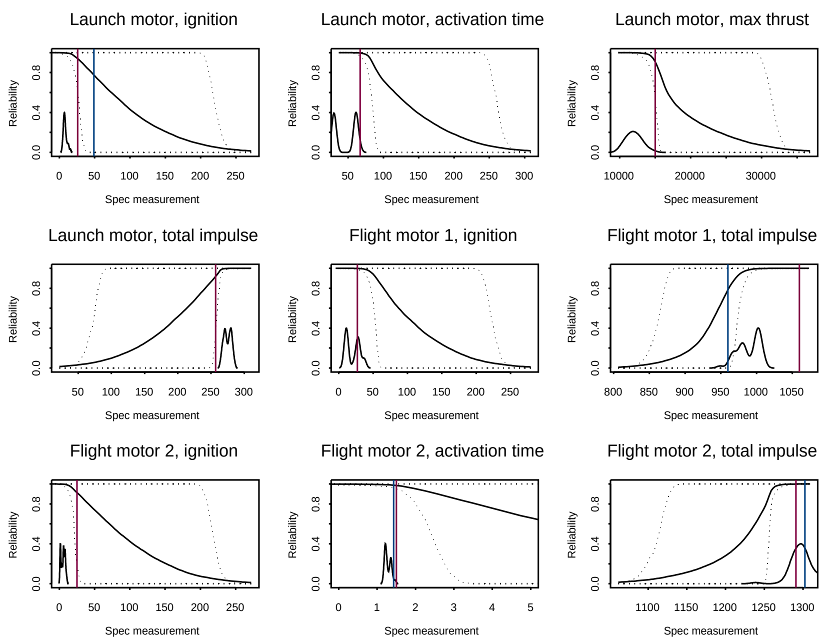

Figure 3: Estimated reliability curves as a function of spec measurement. The solid curves are posterior means, and the dotted curves are 5% and 95% confidence intervals. A density estimate for the observed spec data is also shown, normalized so that its maximum value is 0.4. The red vertical line is the published spec value under normal conditions, and, if present the blue vertical line is the spec under cold conditions.

5 Conclusions

We described an analysis that makes use of both system test data and component acceptance testing data to obtain more precise reliability estimates. Other benefits of our approach include a better understanding of the effects of aging on reliability and the potential for improving specification limits for component acceptance testing to make them more directly relevant to system success.

References

[1] Barlow, R. E. (1985), "Combining Component and System Information in System Reliability Calculation," in *Probabilistic Methods in the Mechanics of Solids and Structures*, Eds. S. Eggwertz and N.L. Lind, Springer-Verlag, Berlin, 375-383.

[2] Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). Measurement error in nonlinear models. Chapman and Hall, London.

[3] Chang, E. Y. and Thompson, W. E. (1976), "Bayes Analysis of Reliability of Complex Systems," *Operations Research*, **24**, 156-168.

[4] Cole, P. V. Z. (1975), "A Bayesian Reliability Assessment of Complex Systems for Binomial Sampling," *IEEE Transactions on Reliability*, **R-24**, 114-117.

[5] Dostal, R. G. and Iannuzzelli, L. M. (1977), "Confidence Limits for System Reliability When Testing Takes Place at the Component Level," in *The Theory and Applications of Reliability* (Vol. 2), Academic Press, New York, 531-552.

[6] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**:398-409.

[7] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*. CRC Press, London.

[8] Graves, T. L. and Hamada, M. S. (2004), "Combining Multi-Level Data to Assess System Reliability and Allocate Resources Optimally: Bayesian Methods and Computation." Los Alamos National Laboratory Technical Report, LA-UR-04-7157.

[9] Jiang, W. and Turnbull, B.W. (2003). The indirect method: Robust Inference based on intermediate statistics. *Technical Report No 1377, School of Operations Research, Cornell University, Ithaca, New York* http://www.orie.cornell.edu/trlist/trlist.html.

[10] Jiang, W. and Turnbull, B.W. (2004). The indirect method: Inference based on intermediate statistics – A synthesis and examples. *Statistical Science* **19**:239–263.

[11] Lampkin, H. and Winterbottom, A. (1983), "Approximate Bayesian Intervals for the Reliability of Series Systems from Mixed Subsystem Test Data,", *Naval Research Logistics*, **30**, 313-317.

[12] Martz, H. F. and Waller, R. A. (1990), "Bayesian Reliability Analysis of Complex Series/Parallel Systems of Binomial Subsystems and Components," *Technometrics*, **32**, 407-416.

[13] Martz, H. F. Waller, R. A., and Fickas, E. T. (1988), "Bayesian Reliability Analysis of Series Systems of Binomial Subsystems and Components," *Technometrics*, **30**, 143-154.

[14] Mastran, D. V. (1976), "Incorporating Component and System Test Data Into the Same Assessment: A Bayesian Approach," *Operations Research*, **24**, 491-499.

[15] Mastran, D. V. and Singpurwalla, N. (1978), "A Bayesian Estimation of the Reliability of Coherent Structures," *Operations Research*, **26**, 663-672.

[16] Natvig, B. and Eide, H. (1987), "Bayesian Estimation of System Reliability," *Scandinavian Journal of Statistics*, **14**, 319-327.

[17] Pepe, M.S, (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*, **75**, 237–249.

[18] Pepe, M.S. and Fleming, T.R. (1991). A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* **86**: 108–113.

[19] Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, **8**, 431 – 440.

[20] Rausand, M. and Hoyland, A. (2004), *System Reliability Theory: Models, Statistical Methods, and Applications*, New York: Wiley.

[21] Soman, K. P. and Misra, K. B. (1993), "On Bayesian Estimation of System Reliability," *Microelectronics and Reliability*, **33**, 1455-1459.

[22] Springer, M. D. and Thompson, W. E. (1966), "Bayesian Confidence Limits for the Product of N Binomial Parameters," *Biometrika*, **53**, 611-613.

[23] Springer, M. D. and Thompson, W. E. (1969), "Bayesian Confidence Limits for System Reliability," in *Proceedings of the 1969 Annual Reliability and Maintainability Symposium*, Institute of Electrical and Electronics Engineers, New York, 515-523.

[24] Tang, J., Tang, K., and Moskowitz, H. (1994), "Bayes Credibility Intervals for Reliability of Series Systems with Very Reliable Components," *IEEE Transactions on Reliability*, **43**, 132-137.

[25] Tang, J., Tang, K., and Moskowitz, H. (1997), "Exact Bayesian Estimation of System Reliability from Component Test Data," *Naval Research Logistics*, **44**, 127-146.

[26] Thompson, W. E. and Chang, E. Y. (1975), "Bayes Confidence Limits for Reliability of Redundant Systems," *Technometrics*, **17**, 89-93.

[27] Wang, C.Y. and Pepe, M.S. (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society Series B–Statistical Methodology*, *62*:509–524.

[28] Winterbottom, A. (1984), "The Interval Estimation of System Reliability from Component Test Data," *Operations Research*, **32**, 628-640.