Probabilistic modeling of S-N curves

Rémy Fouchereau, Gilles Celeux, Patrick Pamphile

Probabilistic modeling of S-N curves

corresponding author : Rémy Fouchereau[a,c], +33672330781
fouchereau.remy@gmail.com
Gilles Celeux[b], Gilles.Celeux@inria.fr
and Patrick Pamphile[a], patrick.pamphile@math.u-psud.fr

[a]Labo de mathmatiques, Université Paris-Sud 11, 91405 Orsay, France
[b]Select, Inria, 91405 Orsay, France
[c]Safran-Snecma, France

November 22, 2013

Abstract

S-N curve is the main tool to analyze and predict fatigue lifetime of a material, component or structure. But, standard models based on mechanic of rupture theory or standard probabilistic models for analyzing S-N curves could not fit S-N curve on the whole range of cycles without microstructure information. This information is obtained from costly fractography investigation rarely available in the framework of industrial production. On the other hand, statistical models for fatigue lifetime do not need microstructure information but they could not be used to service life predictions because they have no material interpretation. Moreover, fatigue test results are widely scattered, especially for High Cycle Fatigue region where split S-N curves appear. This is the motivation to propose a new probabilistic model. This model is a specific mixture model based on a fracture mechanic approach, and does not require microstructure information. It makes use of the fact that the fatigue lifetime can be regarded as the sum of the crack initiation and propagation lifes. The model parameters are estimated with an EM algorithm for which the maximisation step combines Newton-Raphson optimisation method and Monte Carlo integrations. The resulting model provides a parsimonious representation of S-N curves with parameters easily interpreted by mechanic or material engineers. This model has been applied to simulated and real fatigue test data sets. These numerical experiments highlight its ability to produce a good fit of the S-N curves on the whole range of cycles.

Keywords: S-N curves, Lognormal Distributions, Mixture, Convolution Product, EM algorithm, Quantile Estimation

1 Introduction

A fatigue failure occurs when a component is subject to a repeated stress over a long period of time. Fatigue failures are all the more dangerous since they can occur for stress not higher than the service loading. In aeronautic industry, fatigue is the most common reason of breaking for mechanic parts. Fatigue failure analysis is an important issue for reliability analysis and structures design in many domains such as power generation industry, automotive industry and transportation, construction industry, civilian or military engineering.

Fatigue test is the main tool for analyzing fatigue lifetime of a material, component or structure. A material specimen is subjected to cyclic loading S (stress, strain, amplitude,....) by a testing machine which counts N, the number of cycles until failure. Fatigue test results are then plotted on a S-N curve (*cf.* Figure 1).

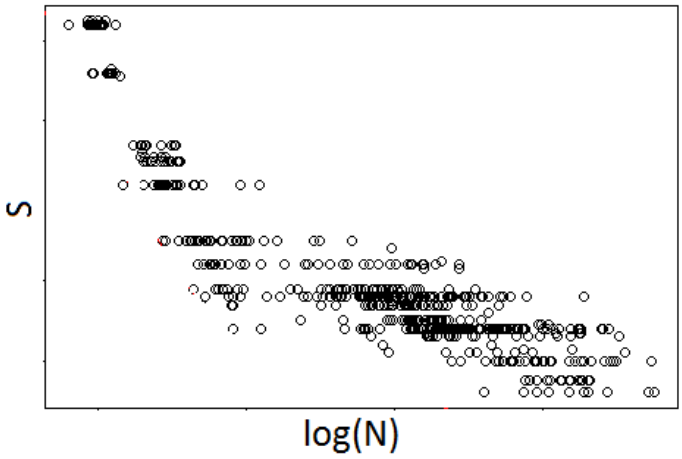


Figure 1: An example of S-N Curve

The resulting lifetime data are highly scattered: the fatigue phenomenon is indeed complex and depends on many effects of mechanical, microstructural, and environmental factors. Furthermore fatigue lifetime database was collected over several years with different material batches.

Many probabilistic models for fatigue lifetime prediction have been proposed. In 1870, Wöhler suggested that the fatigue lifetime N can be expressed as follows:

$$\ln(N) = (a\,S\,+\,b) + \epsilon,$$

where $a\,S\,+\,b$ is the trend fatigue lifetime and ϵ a random noise. Many other, and often more complex, relationships between N and S have been proposed to provide a better fit with test results. For example Pascual and

Meeker [15] proposed the following model:

$$\ln(N) = a - b \times \log(S - S_0) + \epsilon,$$

where S_0 is an unknown random variable that stands for the fatigue limit. For industrial application, data are widely scattered due to variability in the crack mechanism and its synergism with fatigue. However models mentioned above do not deal with crack mechanism, thus they cannot provide a good fit on the overall SN curve.

Moreover, recent fatigue studies reported that for High Cycle Fatigue Region (HCF: $10^4 < N \leq 10^7$), a "duplex SN curve" occurs ([17],[10]). Deterministic and probabilistic approaches have been proposed to take into account this phenomenon.

Deterministic models are based on the fracture mechanic theory. Those models involve microstructure parameters which can explain the S-N curve duality: crack nucleation in surface/subsurface, crack growth rate for small-crack/large-crack, ...(cf.[18],[4]). Unfortunately, those models could not be used without costly microstructure investigation involving a scanning electron microscope. Fractography analysis is not completed for the whole collection of industrial data. In addition, microstructure parameters are difficult to be accurately estimated.

Probabilistic approach make use of competing risk models ([16],[5]) or mixture models ([12], [8]). Components of competing risk, or mixture models, are connected to fatigue mechanisms: crack nucleation surface/subsurface, etc. Yet again, those components are pre-identified through costly microstructure analysis of the material or the fracture. Therefore, those methods are both difficult to be used for industrial production where microstructure analysis are rarely done.

Alternative probabilistic models fit the data without using microstructural or mechanical information. Those models are data dependent and thus cannot be employed for service life predictions.

We propose a mixture model based on the fracture mechanism, and which do not require fractography investigations. This model exploits the fact that the fatigue can be regarded as the sum of the crack initiation life and the crack propagation life. The initiation lifetime, N_i, may be defined as the number of cycles required to form a small crack (of the order of the material grain size). The propagation lifetime, N_p, is the number of cycles required to extend the crack from this small crack size to the critical size at which fracture occurs. Thus, a fatigue test lifetime can be written as $N = N_p$ when a crack appears at the first load or $N = N_p + N_i$ otherwise

(cf. [1]). This behavior leads to the mixture model

$$f_N = \pi(S) f_{N_p} + (1 - \pi(S)) f_{N_i+N_p}, \tag{1}$$

$\pi(S)$ being the probability of having a crack initiation at the first load of S. Since the crack mechanism (propagation with or without initiation) is unknown, the model parameters are estimated through the Expectation-Maximisation algorithm (cf. [6]).
The paper is organized as follows. In Section 2, we propose a mixture model based on a statistical analysis of fatigue data and detail its parametric form. In Section 3, the maximum likelihood estimate of the mixture model parameters is presented through the EM algorithm. Numerical experiments from simulated data and from the collection of real data under study are presented in Section 4. Some concluding remarks are given in Section 5, while technical points are postponed to Appendix A.

2 The Initiation-Propagation Mixture model

The proposed model has been inspired by the analysis of fatigue data from industrial production. Figure 2 represents the QQ-plot of these fatigue data against a lognormal distribution.

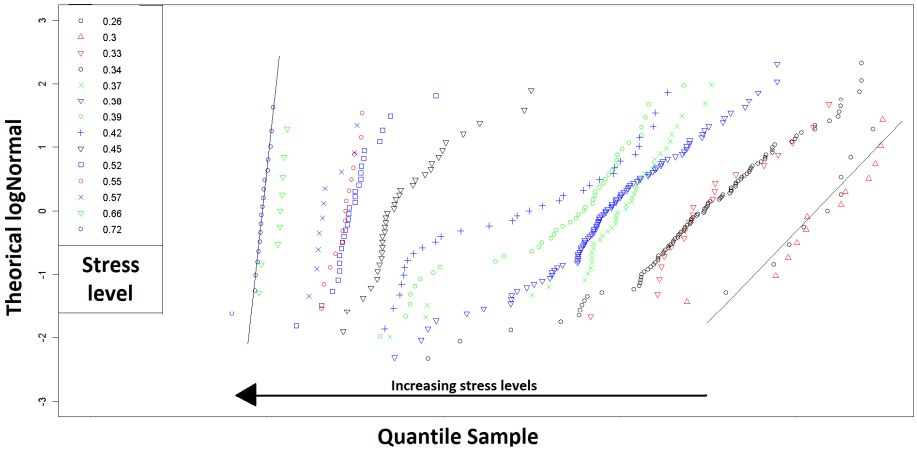


Figure 2: QQ-plots of the empirical distribution of N vs. a log-normal distribution as a function of S

Each symbol represents a different stress level. For high stress, a single lognormal fits well the data: the corresponding QQ-plot is adjusted with a line. The same observation can be made for the lowest stress. For medium stresses, broken-line QQplots indicate an underlying mixture of two distributions. Moreover the proportion of the two mixture components vary with the stress: for high and low levels, a single component is present, while there are two components for medium stress levels.

The question is to understand where these two components come from. If we refer to [1], the lifetime can be decomposed into two components:

- N_i: the crack initiation period,
- N_p: the crack propagation period.

For high stress levels, at the first load, a small crack appears. Thus there is no initiation. On the contrary the propagation lifetimes can be seen at all stress levels. Finally a fatigue lifetime can be written as

$$N = \begin{cases} N_p \text{ if } Z = 1 \\ N_i + N_p \text{ if } Z = 0 \end{cases}$$

where Z denotes the label for an initiation at the first load ($Z = 1$ if the crack starts at the first cycle, $Z = 0$ otherwise).

In most industrial fatigue databases, the values of the Z are not available. From figure 2, it is reasonable to assume that N_i and N_p follow lognormal distributions.

$$f_{N_i}(n, s) = \frac{1}{n\,\sigma_i\,\sqrt{2\pi}}\,\exp\left(-\frac{[\ln(n) - (\alpha_i\,s + \beta_i)]^2}{2\sigma_i^2}\right); \tag{2}$$

$$f_{N_p}(n, s) = \frac{1}{n\,\sigma_p\,\sqrt{2\pi}}\,\exp\left(-\frac{[\ln(n) - (\alpha_p\,s + \beta_p)]^2}{2\sigma_p^2}\right), \tag{3}$$

s being the stress level of the test. The resulting fatigue lifetime is assumed to be a mixture of the propagation lifetime, N_p, and the total lifetime $N_i + N_p$. Its density function is

$$f_N = \pi(s)\,f_{N_p} \;+\; (1 - \pi(s))f_{N_i+N_p}, \tag{4}$$

π being the probability that a crack initiation occurs at the first load.

The term $N_i + N_p$, represents the standard fatigue behavior, whereas the first term N_p represents an unusual fatigue behavior of "pure propagation". The proportion of each component varies with s, and it is to assumed that π is linked with s by a logistic regression

$$\pi(s) = \frac{e^{\alpha+\beta s}}{1 + e^{\alpha+\beta s}}.$$

Obviously, it is important to prove the identifiability of this model which ensures that a unique optimal set of (4) parameters can be fitted on fatigue data. This is done in Appendix A.

3 Estimation of the model parameters

Since we do not known if an crack initiation took place on first load or after, this model is a missing structure data model. Thus the parameters of model (4) are estimated with the Expectation-Maximisation algorithm (EM, see [6]). The maximization step of the EM algorithm is here difficult, since the expectation of the complete log-likelihood knowing the fatigue data involves a complex non linear function of the parameters. It requires an integral computation through Monte Carlo simulations. In addition the maximization step is achieved by using the Newton-Raphson algorithm.
Let $\theta = (\theta_p, \theta_i, \alpha, \beta)$ be the vector of parameters of the model (4), with $\theta_p = (a_p, b_p, \sigma_p)$ and $\theta_i = (a_i, b_i, \sigma_i)$. The algorithm EM is a two step algorithm maximizing the observed likelihood:

$$L(N, S\,;\,\theta) = \prod_{k}^{m} f_{(N,S)}(n_k, s_k\,;\,\theta). \tag{5}$$

knowing the data (N, S) and a current value of the parameters. The EM algorithm is making use of the complete likelihood

$$L(N, S, Z\,;\,\theta) = \prod_{k}^{m} f_{(N,S,Z)}(n_k, s_k, z_k\,;\,\theta), \tag{6}$$

where Z denotes the missing origin of the crack: $Z = 1$ if the crack starts at the first load, and $Z = 0$ otherwise. Then

$$f_{(N,S,Z)}(n, s, z\,;\,\theta) = (\pi(s)[f_{N_i}(n, s)])^z \times ((1 - \pi(s))[f_{N_i+N_p}(n, s)])^{(1-z)}.$$

Since the indicator variable Z is not observed, the complete likelihood $L(N, S, Z\,;\,\theta)$ cannot be maximized directly. The Expectation step consists of computing the expected value of the completed likelihood knowing the data (N, S) and a current value of the parameters [6].

The EM algorithm Starting from a vector parameter $\theta^{(0)} = (\theta_p^{(0)}, \theta_i^{(0)}; \alpha^{(0)}, \beta^{(0)})$, this algorithm iterates the E and M steps.

1. E step: computation of the expected complete loglikelihood knowing a current parameter $\theta^{(j)}$:

$$\begin{aligned} Q(\theta|\theta^{(j)}) &= \mathbf{E}\left[\ln L(N, S, Z\,;\,\theta)\,|(Z|N\,;\,\theta^{(j)})\right] \\ &= \sum_{k}^{m}\big[\hat{t}_k \ln(f_{N_p}(n_k, s_k\,;\theta_p)) \\ &+(1 - \hat{t}_k)\ln(f_{N_i+N_p}(n_k, s_k\,;\,\theta))\big]; \end{aligned} \tag{7}$$

$\hat{t}_k$ being the conditional distribution of Z:

$$\begin{aligned}\hat{t}_k &= \mathbf{E}[Z|N,S;\theta^{(j)}] \\ &= \mathbf{P}(Z=1|N=n_k,S=s_k\,;\,\theta=\theta^{(j)}) \\ &= \frac{\pi^{(r)}(s_k)f_{N_p}(n_k;s_k,\theta_p^{(r)})}{\sum\limits_{l=(i,p)}\pi_l(s_k)\phi_l(n_k;s_k,\theta^{(r)})};\end{aligned}$$

where ϕ_l stands for f_{N_p} if $l=1$, and $f_{N_i} * f_{N_p}$ otherwise.

2. M step: maximization of $Q(\theta|\theta^{(j)})$

$$\hat{\theta}^{(j+1)} = \arg\max_{\theta} Q(\theta|\theta^{(j)}).$$

It can be decomposed in two independent maximisations

- $\max\limits_{\alpha,\beta}\sum\limits_{k=1}^{N}\hat{t}_k\ln(\pi(s_k))+(1-\hat{t}_k)\ln(1-\pi(s_k))$;
- $\max\limits_{\theta}\sum\limits_{k=1}^{N}\hat{t}_k\ln(f_{N_p}(n_k,s_k,\theta_p))+(1-\hat{t}_k)\ln(f_{N_i}*f_{N_p}(n_k;s_k,\theta))$;

which can be achieved with a Newton-Raphson algorithm.

The sequence $\theta^{(1)}$, $\theta^{(2)}$,... generated by EM is expected to converge toward a local maximum of the observed-data likelihood $L(N,S\,;\,\theta)$ under fairly general conditions (cf. [6]).

The solution provided by EM could be highly dependent of the initial parameter values. It appears not for the initiation-propagation mixture model and the following initial values provide satisfactory parameter estimates.

- $\pi^{(0)}=0.5$;
- $\theta^{(0)}=(\theta_p^{(0)},\theta_i^{(0)})$ has been derived from a clusterwise linear regression on the log-lifetimes (cf. [7]). For LCF region($N<10^4$) $N\simeq N_p$, whereas for VHCF region $N\simeq N_i$. Then clusterwise regression provides quickly an honest initial estimation of the model parameters and the EM algorithm converges a sensible local maximum.

The maximization of $Q(\theta|\theta^{(j)})$ requires a Newton-Raphson algorithm combined with a Monte Carlo algorithm to evaluate the density of $f_{N_i+N_p}$ which is now described.

The convolution product

$$f_{N_i+N_p}(n)=\int_0^n f_{N_i}(n-x)f_{N_p}(x)dx \tag{8}$$

used in (7) does not lead to a closed form expression. Using a Gauss-Legendre Quadrature would not be efficient here, because for some x values the function to be integrated is highly picked. This is why a Monte Carlo approximation has been used and is now described. The integral can be written

$$\begin{aligned} f_{N_i+N_p}(n) &= \int_0^n f_{N_i}(n-x) f_{N_p}(x) dx && (9) \\ &= \mathsf{E}_{N_p}[f_{N_i}(n-N_p)^+] && (10) \\ &= \mathsf{E}_{N_i}[f_{N_p}(n-N_i)^+]. && (11) \end{aligned}$$

It is then approximated by importance sampling using either (10), either (11), according to $\hat{\sigma}(N_p) < \hat{\sigma}(N_i)$ or not. Without loss of generality assuming that $\hat{\sigma}(N_p) < \hat{\sigma}(N_i)$ the procedure is as follows:

1. b independents replications of f_{N_p} are simulated: $n_{p_j},\ j = 1, \cdots, b$. In practice, $b = 1000$ appears to provide satisfactory results.

2. using (10), $f_{N_i+N_p}(n)$ is estimated by

$$\hat{f}_{N_i+N_p}(n) = \frac{1}{b} \sum_{j=1}^{b} f_{n_i}(n - n_{p_j})^+.$$

Dealing with censored data Often fatigue data are right censored, the likelihood includes also the probability $\mathbb{P}(N > c)$. Since this probability is difficult to compute, two alternative solutions are proposed:

a. using an asymptotic approximation (*cf* [2]), we have

$$\mathbb{P}(N > c) \sim \frac{\pi_2 \sigma_i e^{-\frac{1}{2}\frac{(log(c)-\mu_i)^2}{\sigma_i^2}}}{\sqrt{2\pi}(log(c) - \mu_i)}.$$

But an asymptotic approximation is not realistic when censoring occurs at low stress levels.

b. simulating the data over the censorship c, the EM algorithm is replaced by the Stochastic EM algorithm (*cf* [3]). But, simulate values greater than c, could take a huge computation time.

In order to circumvent the mentioned numerical problems, we make use of the following hybrid heuristics which provides good results in practice.

- Each censored data is assumed to arises from the component with density $f_{N_i+N_p}(n)$ and thus $t_{k1} = 0$.

- Before each M-step, for each censored data, simulate at most one hundred times the two lognormal distributions and sum them:
  - if one of the 100 resulting values is greater than the censorship, consider this simulated value as an observed value in the M step;
  - otherwise, the censorship is considered to belong to the distribution tail and the asymptotic approximation above can be used.

4 Numerical experiments

4.1 Simulated data experiments

Table 1 reports the errors between the estimates and the actual values of the model (4) for 50 lifetime data sets with size $m = 200$ simulated according to this model with $\theta = (a_p = 22, b_p = -10, \sigma_p = .2, a_i = 33, b_i = -30, \sigma_i = 1, \alpha = 15, \beta = -35)$. The results are satisfactory. The estimation of π is less accurate. This is probably due to the fact that it is difficult to assign some observations to a mixture component.

parameters	α	β
simulation parameters	15	-35
mean over 50 simulations	17.5	-40.9
standard deviation over 50 simulations	4.9	11

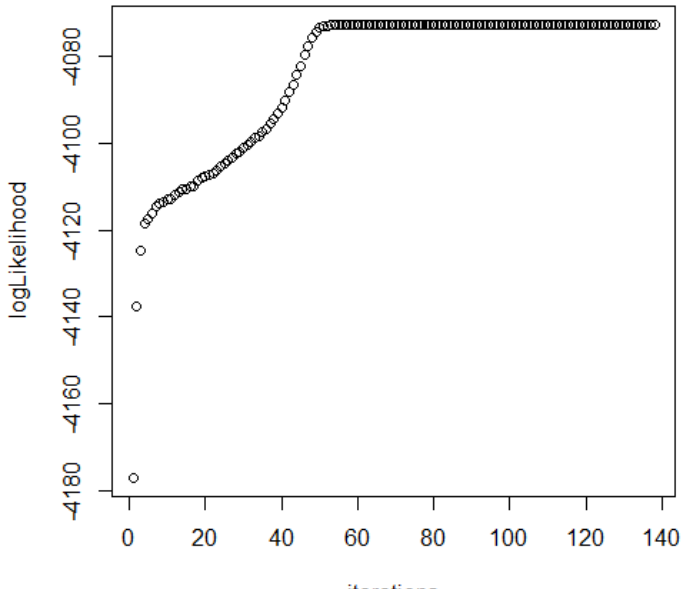
parameters	α_i	β_i	σ_i
simulation parameters	22	-10	.2
mean over 50 simulations	21.9	-9.99	.20
standard deviation over 50 simulations	.11	.16	.01

parameters	α_p	β_p	σ_p
simulation parameters	33	-30	1
mean over 50 simulations	33.1	-30.3	.99
standard deviation over 50 simulations	.54	1.71	.08

Table 1: Results of the parameters estimation for the simulated data

Figures (3) and (4) display the behavior of the EM algorithm. EM does not produce an increasing likelihood sequence as expected. This is due to the Monte Carlo approximation. Therefore, in order to get honest parameter estimates, we recommend to compute the mean of the estimated values on the last 50 iterations of EM after a burn-in period of about 100 EM iterations.

Moreover, it is important to note that the Monte Carlo integration is highly CPU time consuming.



Fig

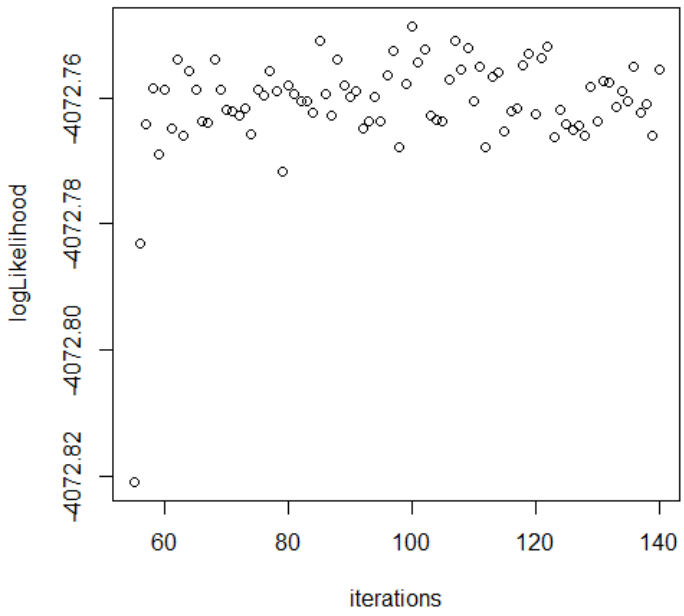


Figure 4: Zoom on the behavior of the likelihood along the last EM iterations

4.2 S-N fitting on real data

All fatigue tests are carried out with constant amplitude loading and temperature. Test specimens came from a unique superalloy with approximatively 600 points for the whole curve.

After estimating the initiation-propagation mixture model, each data can be assigned to one of the mixture components according to the following Maximum A Posteriori (MAP) rule:

$$\begin{cases} \text{component 1 if } P(Z=1) > 0.5 (\text{cross}) \\ \text{component 2 otherwise (circles)} \end{cases}$$

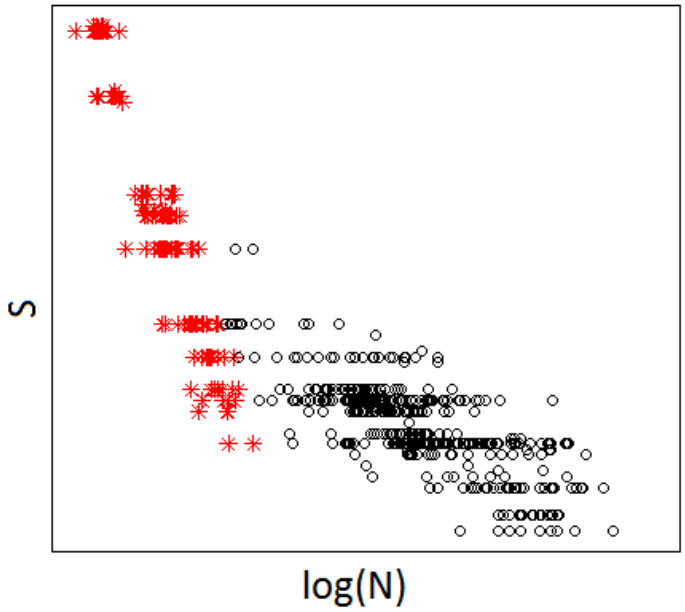


Figure 5: Classification of the failure times in the propagation component (cross) or the initiation-propagation component (circles)

According to Figure 5, the classification is as good as expected. High stress are all classify in the first component ("pure propagation"). For low levels, the model makes a reasonable distinction between high and low lifetimes. This prove that the initiation-propagation mixture model is efficient to retrieve the failure origins without fractography investigations.

Figure 6 displays the QQ-plot of the empirical distribution of the data *versus* the fitted mixture model (continuous lines). Some remarks are in order:

- the model adequacy is pretty good when the propagation component is prevailing (high stress);
- the mixing proportion estimate is also good, as the broken line on the empirical distribution is apparent on the fitted mixture distribution;
- for medium and low stress the scatter is huge. If a lognormal distribution well fit the propagation, the multiple causes of crack initiations

are badly taken into account with a single lognormal distribution.

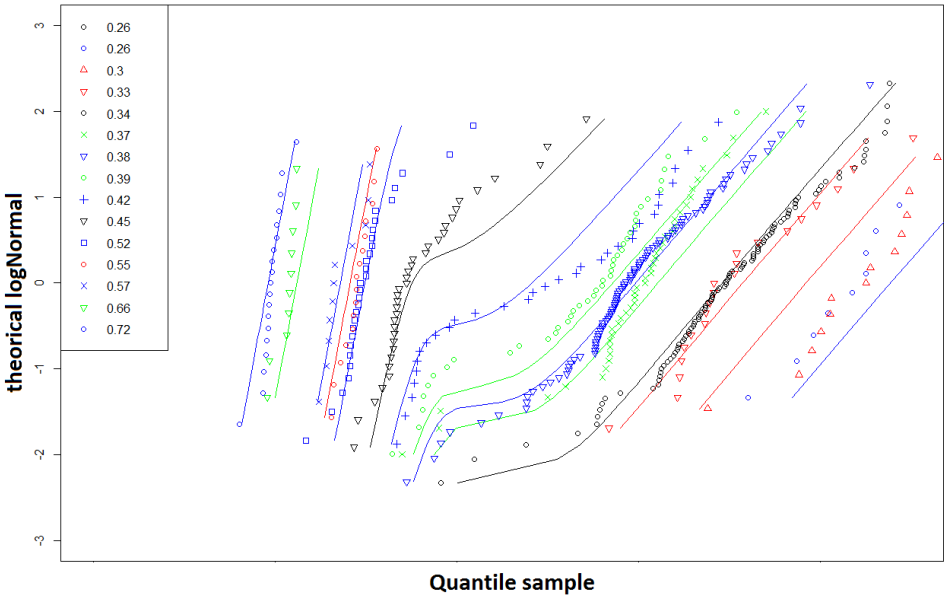


Figure 6: QQ-plot with model approximation

An important question to be answered when modeling fatigue lifetime is to provide good estimates of extreme quantiles on the whole range of stress values. Using the estimated initiation-propagation mixture model, a 0.1%-quantile has been designed from the simulation of 10^6 lifetimes according to the estimated distribution. The resulting quantile is displayed in figure (7). It appears to be quite satisfactory since one of the 600 observed data is beyond this 0.1%-quantile. It means that the initiation-propoagation mixture model provides a useful estimation of the fatigue lifetime on the whole range of stress values.

5 Conclusion

Fatigue tests and the resulting S-N curve, are the main basic tools for analyzing and predicting fatigue lifetime of a material, component, or structure. S-N curves are in general widely scattered, and therefore service lifetime predictions are difficult. For instance, a "duplex behavior" appears in the S-N curves of High Cycle Fatigue region. This "duplex behavior" can sometimes be characterized by costly fractography investigations. In an industrial framework, fractography investigations are not completed after each test, and standard prediction models cannot be used. We have proposed a mixture model based on the fracture mechanism, without fractography information. This is a latent structure model and its parameters are estimated

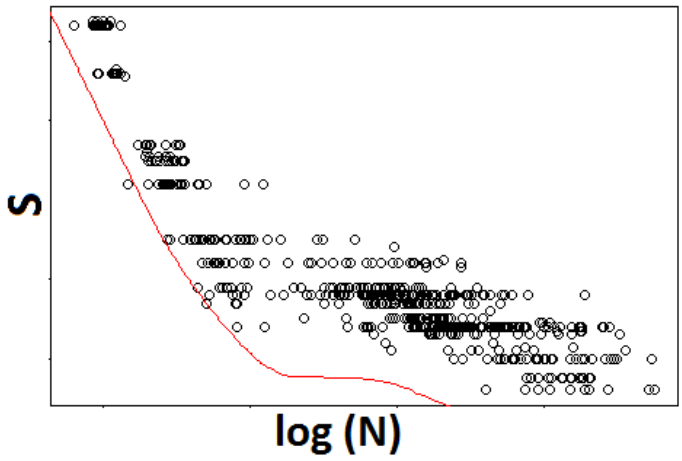


Figure 7: The curve is the 0.1%-quantile provided by the initiation-propagation mixture model

through an EM algorithm. Since this so-called initiation-propagation model is based on a fracture mechanic approach, this parsimonious model is easily interpreted by material or mechanic engineers. Numerical experiments on simulated and real fatigue data sets show that the initiation-propagation mixture model fit well fatigue lifetime data on the whole range of cycles. In particular, the "duplex" phenomenon on High Cycle Fatigue region is properly identified and it is possible to assign the lifetimes to the "initiation-propagation" or the "propagation" components using a the maximum a posteriori classification rule with a high level of confidence. Morevoer this model an honest estimation of extremes quantiles of the lifetimes distributions. Thus, it could be expected to be an efficient tool to set off alarms. And, this model should also be a relevant diagnostic tool for material elaboration.

Possible improvements, under study, concern the use of information from the manufacturing process to get more accurate lifetime estimation for medium level of stress.

Acknowledgment

The research was supported by SNECMA, subsidiary of the SAFRAN Group. The authors are particularly indebted with Catherine Duveau for her helpful suggestions and comments.

References

[1] Alexandre, F. (2004) Aspects probabilistes et microstructuraux de l'amorçage des fissures de fatigue dans l'alliage INCO 718. Phd thesis. Ecole Normale Supérieure des Mines de Paris.

[2] Asmussen, S. and Rojas-Nandayapa, L. (2005), Sums of dependent log-normal random variables: asymptotics and simulation. Research report, Thiele Centre for Applied Matematics in Natural Science

[3] Celeux, G. and Diebolt, J. (1985) The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational Statistics Quaterly*, 2, 73-85.

[4] Chan, K.W. (2010), Role of microsture in fatigue crack initiation. *International Journal of Fatigue*, 32, 1428-1447.

[5] Crowder, M. (1999) Classical Competing Risks. *Harlow, England: Addison-Wesley*

[6] Dempster, A.P., Laird N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society.*

[7] DeSarbo, W.S. and Cron, W.L. (1988) A maximum likelihood methodology for clusterwise linear regression. *International Journal of Classification*, 5, 249-282.

[8] Harlow, D.G., Wei, R., Sakai,T. and Oguma, N. (2006), Crack growth based probability modeling of S-N response for high strength stell. *International Journal of Fatigue*, 28 ,1479-1485.

[9] Hanaki, S., Yamashita, M., Uchida, H. and Zako, (2010) M. On stochastic evaluation of SN data based on fatigue strength distribution. *International Journal of Fatigue*, 32, 605-609.

[10] Jha, S.K. and Ravi Chandran, K.S. (2003), An unusual fatigue phenomenon: duality of the SN fatigue curve in the b-titanium alloy Ti10V2Fe3Al. *Scripta Materialia*, 48, 1207-1212.

[11] Jha, S.K., Caton, M.J. and Larsen, J.M. (2007), A new paradigm of fatigue variability behavior and implications for life prediction. *Material Science and Engineering :A*, 468-470, 23-32.

[12] Jha, S.K., Larsen, J.M. and A.H. Rosenberg (2009), Toward a physics-based description of fatigue behaviour in probabilistic life-prediction. *Engineering Fracture Mechanics*, 76, 681-694.

[13] McLachlan, G. and Peel, D. (2000), Finite mixture models, Wiley.

[14] Nelson , W. (1982), Applied life data analysis, Wiley.

[15] Pascual, F.G. and Meeker, W.Q. (1999), Estimating Fatigue Curves with the Random Fatigue-Limit Model. *Technometrics*, 41, 277-290.

[16] Ravi Chandran, K.S., Chang, P. and Cashman, G.T. (2010) Competing failure modes and complex SN curves in fatigue of structural materials *International Journal of Fatigue*, 32, 482-491.

[17] Sakai, T., Nakayasu, H. and Nishikawa, I. (2005) Establishment of JSMS standard regression method of SN curves for metallic materials. *Safety and Reliability of Engineering Systems and Structures*, 643

[18] Shiozawa, K., Murai, M., Shimatani, Y. and Yoshimoto, T. (2010) Transition of fatigue failure mode of NiCrMo low-alloy steel in very high cycle regime. *International Journal of Fatigue*, 32, 541-550.

Appendix A

This appendix is to devoted to prove the identifiability of the model

$$f_N = \pi(s)\, f_{N_p} \ + \ (1-\pi(s)) f_{N_i+N_p},$$

under the realistic assumption for material lifetimes that $\sigma_i > \sigma_p$. It is difficult to prove the identifiability in a single exercise and the proof is done step by step.

1. At first, it is assumed that the mixing proportion π and the stress level s are fixed, and the uniqueness of the position and dispersion parameters μ and σ are proved for the two lognormal distributions.

2. Then by allowing the stress level s to vary, it is proved that the regression parameters a and b used to define μ are unique for the two lognormal distributions. Similarly the uniqueness of the parameters α and β, involved in the logistic regression model for the proportion π, are proved.

1: Assuming s and π fixed, under the assumption $\sigma_i > \sigma_p$, we have the asymptotic equivalence for the tail distribution of model (4), see [2]:

$$\mathbb{P}(N > x) \sim \frac{\pi_2 \sigma_i e^{-\frac{1}{2}\frac{(\log(x)-\mu_i)^2}{\sigma_i^2}}}{\sqrt{2\pi}(\log(x)-\mu_i)}.$$

Thus, if the distribution of N has two parameterizations $(\pi_1, \pi_2 = 1-\pi_1, \mu_i, \sigma_i, \mu_p, \sigma_p)$ and $(\pi_1', \pi_2' = 1-\pi_1', \mu_i', \sigma_i', \mu_p', \sigma_p')$, we have

$$\lim_{x\to\infty} \left(\frac{\frac{\pi_2\sigma_i}{\sqrt{2\pi}(\log(x)-\mu_i)} e^{-\frac{1}{2}\frac{(\log(x)-\mu_i)^2}{\sigma_i^2}}}{\frac{\pi_2'\sigma_i'}{\sqrt{2\pi}(\log(x)-\mu_i')} e^{-\frac{1}{2}\frac{(\log(x)-\mu_i')^2}{\sigma_i'^2}}} \right) = 1.$$

That is

$$\lim_{x\to\infty} \left(\frac{\pi_2\sigma_i}{\pi_2'\sigma_i'} \frac{\log(x)-\mu_i}{\log(x)-\mu_i'} e^{-\frac{1}{2}\log^2(x)(\frac{1}{\sigma_i^2}-\frac{1}{\sigma_i'^2})-\frac{1}{2}\log(x)(-\frac{2\mu_i}{\sigma_i^2}+\frac{2\mu_i'}{\sigma_i'^2})-\frac{1}{2}(\frac{\mu_i^2}{\sigma_i^2}-\frac{\mu_i'^2}{\sigma_i'^2})} \right) = 1$$

Since $\lim\limits_{x\to\infty} \left(\frac{\log(x)-\mu_i}{\log(x)-\mu_i'}\right) = 1$, it leads to

$$\lim_{x\to\infty} \left(\frac{\pi_2\sigma_i}{\pi_2'\sigma_i'} e^{-\frac{1}{2}\log^2(x)(\frac{1}{\sigma_i^2}-\frac{1}{\sigma_i'^2})-\frac{1}{2}\log(x)(-\frac{2\mu_i}{\sigma_i^2}+\frac{2\mu_i'}{\sigma_i'^2})-\frac{1}{2}(\frac{\mu_i^2}{\sigma_i^2}-\frac{\mu_i'^2}{\sigma_i'^2})} \right) = 1.$$

A necessary condition to ensure this equation is that the term in $\log^2(x)$ is zero, thus $-\frac{1}{2}(\frac{1}{\sigma_i}-\frac{1}{\sigma'_i}) = 0 \Rightarrow \sigma_i = \sigma'_i$. Then, the term in $\log(x)$ has to be zero too and thus $\mu_i = \mu'_i$. And, finally, it leads to $\lim_{x\to\infty}(\frac{\pi_2}{\pi'_2}) = 1$ and thus $\pi_2 = \pi'_2$.

To show the identifiability of the propogation parameter, we use the Laplace transformation $L(f_N) = \int_0^{+\infty} e^{st} f_N(t)\,\mathrm{d}t$ of f_N. We have

$$L(f_N) = \pi\, L(f_{N_p}) \;+\; (1-\pi)L(f_{N_i})L(f_{N_p}),$$

$$L(f_N) = L(f_{N_p})(\pi + (1-\pi)L(f_{N_i})).$$

We have already shown that $\theta_i = \theta'_i$, thus

$$L(f_{N_p}(n;\theta_p))(\pi + (1-\pi)L(f_{N_i})) = L(f_{N_p}(n;\theta'_p))(\pi + (1-\pi)L(f_{N_i})),$$

which implies $L(f_{N_p}(n;\theta_p)) = L(f_{N_p}(n;\theta'_p))$. If two variables have the same Laplace transform, then they have the same distribution: $f_{N_p}(n;\theta_p) = f_{N_p}(n;\theta'_p)$. Finally $\theta_p = \theta'_p$ since the lognormal distribution is identifiable.

2: For any stress level s, we have by definition of the model (4):

$$f_N = \pi(s)\, f_{N_p}(n;s,\mu_i(s),\sigma_i) \;+\; (1-\pi(s))f_{N_i+N_p}(n;s,\mu_i(s),\mu_p(s),\sigma_i,\sigma_p),$$

with $\pi(s) = \frac{e^{\alpha+\beta s}}{1+e^{\alpha+\beta s}}$, $\mu_i(s) = a_i + b_i s$, and $\mu_p(s) = a_p + b_p s$.
For any s, we have proved that $a_p + b_p s = a'_p + b'_p s$ and $a_i + b_i s = a'_i + b'_i s$. It implies straightforwardly that $a_p = a'_p$, $b_p = b'_p$, $a_i = a'_i$ and $b_i = b'_i$.
Similarly, for any s, we have $\frac{e^{\alpha+\beta s}}{1+e^{\alpha+\beta s}} = \frac{e^{\alpha'+\beta' s}}{1+e^{\alpha'+\beta' s}}$ and thus $\alpha = \alpha'$ and $\beta = \beta'$
□