Inference with Censored Degradation Data

Yang Yang, Vijay Nair
Department of Statistics, University of Michigan, Ann Arbor

Abstract
In applications with few or no failures, survival time provides limited information because of the inevitable right censoring issue. Only a sample with little failed subjects harms the analysis. Degradation data, on the other hand, focus on the degradation process which eventually will lead to the failure. By observing the degradation path of the object, detailed measurements are available over time to predict failure and indicate possible future events. Therefore, degradation data are nowadays popularly used in applications. However, one issue in practice is that data are only available at discrete time points and severe censoring and missing data will complicate the analysis. To estimate the degradation path from censored sample, we suggest an EM algorithm with exact solution under normal increment setting. With the imputed degradation path, we are able to predict the remaining service time and indicates risk factors by doing simple ANOVA.

Introduction
Degradation data are the multiple measurements recorded over time to describe the physical deterioration of the object. For instance, Michigan Department of Transportation uses Distress Index (DI) values to reflect the amount of distress present on the road pavement. DI values help evaluate the road pavement performance over pavement age. The higher the DI value is, the more severe is the damage. The rehabilitation is necessary when DI passes 50. The estimation of the degradation path is essential to predict the remaining service time the product has before it hits the critical threshold. Besides, the significant difference present in different subgroups of the sample may indicate important risk factors. All in all, to suggest reasonable degradation model for the data is extremely valuable for the product reliability improvement.

Real data are rarely observed continuously over time but recorded at specific discrete time points. More often than not, different subjects may not share the same observation times across the sample. Therefore, instead of having a uniform panel data, the degradation data lose some information here and there. Table 1 shows a subsample of recorded road pavements DI values. Most records are not complete and suffering from left, interval and right censoring. Taking into account that censoring is a common problem for the analysis of degradation data, the suggested model should be able to handle this issue. However, nonparametric estimator seem not be able to handle this problem well. Nonparametric estimation by interpolating and smoothing suffer from the bad quality of the data. The measurement error makes the estimate wiggly and the informative censoring causes obvious biased estimation. (1) Linear interpolating missing data [Black] (2) Average over time [Green] (3) Locally-weighted regression estimate with 2/3 data used for smoothing [Red] These nonparametric estimates of the degradation path will cause serious inference bias. Therefore, we here adopt some reasonable assumptions and estimate the path parametrically.

Gaussian Increment Model
Instead of investigating the dependent variable, t, = 1, . . . , T, like the DI series in Table 1, it is usually easier to consider simple parametric models for the change of X, i.e., = − − . Out of the measurement error, X can be negative and we make independent normal assumption for these changes:

\[ X_t = N(\theta_k, \sigma^2) \text{ for } k = 1, \ldots, T; \]
\[ X_0, X_1, \ldots, X_{T-1} \text{ i.i.d. } \]

Since some Xs are censored, the same problem exists for . To complete the missing information in the sample, EM algorithm is proposed.

E-step: Impute the expectation of X, denoted as \( \hat{X}_k \)
(1) left-censored: \( \hat{X}_k = \theta_k + \frac{1}{\sum_{j=1}^{K-1} \theta_j} \sum_{j=K}^{\infty} \theta_j, \quad k = 1, \ldots, K \) (2) interval-censored: \( \hat{X}_k = \theta_k + \frac{1}{\sum_{j=L}^{K-1} \theta_j} \sum_{j=L}^{K} \theta_j, \quad k = L, \ldots, R \)
(3) right-censored: \( \hat{X}_k = \theta_k + \frac{1}{\sum_{j=L}^{K-1} \theta_j} \sum_{j=L}^{K} \theta_j, \quad k = L, \ldots, R \)

M-step: Denote \( \hat{X}^{(n)} \) the values \( \hat{X}_k \) for \( n \) iterations:

\[ \cdot \text{ both available otherwise} \]

Assuming a homogenous sample, we have, the maximum likelihood estimate (MLE) of the mean vector \( \theta \) is \( \frac{1}{n} \sum_{i=1}^{n} X^{(i)-1} \).

M-step: \( \hat{X}^{(n)} \) the values \( \hat{X}_k \) for \( n \) iterations:

\[ \cdot \text{ both available otherwise} \]

Assuming a homogenous sample, we have, the maximum likelihood estimate (MLE) of the mean vector \( \theta \) is \( \frac{1}{n} \sum_{i=1}^{n} X^{(i)-1} \).

Road Pavement Data
The following analysis is for the degradation data from MDOT pavement project. The response \( Y_t \) is the cumulative DI value mentioned in the introduction section. Because the road projects are very different for the observing pavement projects, there are a large proportion of values missing in the sample. Figure 3 shows the result from fitting an independent Gaussian increment model with unequal variance, so satisfying (6) and (7). The left plot gives the estimate of mean vector \( \theta \) of one sub-sample, while the right one displays the imputed degradation paths of the sub-sample.

![Figure 3: Mean and Imputed individual degradation Path](image)

With the estimate above, the remaining service life of the pavement road can be predicted by picking the time when the degradation path crosses 90. For this sub-sample, the failure times are roughly all after 20 years since the pavements were constructed. Other than predicting the remaining service life, we also investigate the covariate effect. For instance, the US highway pavements are mainly separated by the made of its Course Aggregate materials, from either BESlag, Quarry, Crushed Concrete or SDDGR. Perform the analysis for each aggregate subgroup, and adopt functional one-way ANOVA to see if there is significant impact from coarse aggregate based on the imputed degradation paths. From the result in Table 2, we see that other than age 12, coarse aggregate type does make a difference with respect to pavement degradation. Besides, from the time line, BESlag obvious has the worse case scenario, while Crushed Concrete shows best performance overall.

![Figure 2: Nonparametric Estimates](image)

![Table 2: Functional one-way ANOVA for Course Aggregate Factor](image)

Future Work: Out of the severe censoring issue of degradation data, some parametric assumptions are adopted for the sake of analyzation. Here we used gaussian in parameters with different variance over time for pavement data. This approach has not integrated the time effect into the picture yet. Measurement error will cause less problem under a more parametric setting. Besides, random effect should be included to account for unobserved factor effects in the sub-sample in our future study.

![Table 1: Censoring Issue of Degradation Data: “x” value missing](image)

![Table 2: Functional one-way ANOVA for Course Aggregate Factor](image)