Journal of the Korea Academia-Industrial cooperation Society
Vol. 17, No. 5 pp. 645-652, 2016

http://dx.doi.org/10.5762/KAIS.2016.17.5.645
ISSN 1975-4701 / eISSN 2288-4688

통계모델링 방법의 비교 연구

노유정[1*]
[1]부산대학교 기계공학부

A Comparison Study on Statistical Modeling Methods

Yoojeong Noh[1*]
[1]**School of Mechanical Engineering, Pusan National University**

요 약 입력 랜덤 변수(input random variable)의 통계 모델링은 기계시스템의 신뢰성 해석(reliability analysis), 신뢰성 기반 설계(reliability-based design optimization), 해석모델의 통계적 검정(validation) 및 보정(calibration)을 위해 반드시 필요하다. 대표적인 통계모델링 기법에는 Akaike Information Criterion (AIC), AIC correction (AICc), Bayesian Information Criterion, Maximum Likelihood Estimation (MLE), Bayesian 방법 등이 있다. 이러한 방법들은 기본적으로 주어진 데이터로부터 후보 모델의 우도함수값을 이용하여 후보 모델 중 가장 적합한 모델을 선택하는 방법이며, 방법에 따라 데이터 수 혹은 파라미터의 수를 고려하여 모델을 선정한다. 하지만 실제 현장에서 데이터의 통계모델링을 하는 엔지니어는 각 방법의 장단점에 대한 이해가 부족하여 어떤 방법이 정확한 방법인지 몰라 통계모델링 수행 시 어려움이 있다. 본 논문에서는 다양한 통계모델링 방법들을 비교하고 각 방법의 장단점 분석을 통해 가장 적합한 모델링 기법을 제안하고자 한다. 각 방법의 검증을 위해 다양한 모분포를 가정하고 다양한 사이즈의 샘플을 임의로 생성하여 시뮬레이션을 수행하였으며, 실제 공학 데이터를 사용하여 통계모델링 방법의 유효성을 검증하였다.

Abstract The statistical modeling of input random variables is necessary in reliability analysis, reliability-based design optimization, and statistical validation and calibration of analysis models of mechanical systems. In statistical modeling methods, there are the Akaike Information Criterion (AIC), AIC correction (AICc), Bayesian Information Criterion, Maximum Likelihood Estimation (MLE), and Bayesian method. Those methods basically select the best fitted distribution among candidate models by calculating their likelihood function values from a given data set. The number of data or parameters in some methods are considered to identify the distribution types. On the other hand, the engineers in a real field have difficulties in selecting the statistical modeling method to obtain a statistical model of the experimental data because of a lack of knowledge of those methods. In this study, commonly used statistical modeling methods were compared using statistical simulation tests. Their advantages and disadvantages were then analyzed. In the simulation tests, various types of distribution were assumed as populations and the samples were generated randomly from them with different sample sizes. Real engineering data were used to verify each statistical modeling method.

Keywords : AIC, AICc, Bayesian method, BIC, MLE, Statistical modeling

1. 서론

기계시스템의 신뢰성 해석이나 신뢰성 기반 설계를 위해서는 시스템의 성능에 영향을 미치는 입력변수의 통계모델링이 요구된다. 통계모델을 얻기 위해서는 입력변수의 데이터가 필요하지만, 실제 현장에서는 실험 비용

본 논문은 산업통상자원부의 2015년도 산업원천기술개발사업 중 지식서비스 분야의 연구비 지원(10048305)과 2015학년도 부산대학교 신임교수연구 정착금 지원으로 이루어 졌으며, 모든 연구비 지원에 감사드립니다.
*Corresponding Author : Yoojeong Noh (Pusan National Univ.)
Tel: +82-51-510-2308 e-mail: yoonoh@pusan.ac.kr
Received March 22, 2016 Revised (1st April 18, 2016, 2nd April 27, 2016)
Accepted May 12, 2016 Published May 31, 2016

이나 많은 시간이 소요되므로 한정된 데이터를 이용하여 통계모델링을 하게 된다[1-3].

통계모델링방법에는 검정법과 모델선택법 등이 있는데, 검정법은 대상 모델의 적합성을 판단하는 절대적 방법에 해당되며 모델선택법은 여러 개의 후보 모델 중 가장 적합한 모델을 선택하는 상대적 방법에 해당된다. 검정법은 모델선택법에 비해 일반적으로 정확한 모델을 찾는데 많은 데이터가 요구되므로 모델선택법이 더 많이 사용되고 있다.

모델선택법 중에서 가장 많이 사용되는 방법은 AIC, AICc, BIC, MLE, Bayesian 방법이 있다. 모든 모델선택법은 우도함수(Likelihood function)을 이용하여 후보 모델의 상대적 적합성을 판단하는데, 후보 모델의 분포가 데이터 분포와 일치하면 할수록 우도함수 값은 더 커지게 된다. MLE 방법은 우도함수 값 자체만을 비교하는데 반해 AIC는 우도함수 값과 함께 모델의 파라미터 개수(자유도)를 반영하여 Generalized Extreme Value 분포와 같은 자유도가 큰 분포가 지나치게 오버 피팅(Over fitting)하는 경우를 고려하여 모델을 선택한다[4]. AICc는 데이터 수를 고려한 보정된 AIC방법[5]에 해당되며, BIC 역시 우도함수 값, 파라미터 개수, 샘플 수를 고려하여 모델을 선택한다[6]. Bayesian 방법은 우도함수를 파라미터에 대해 적분하여 모델을 선택하는 방법이다[1,2].

다양한 통계모델링 방법에 대한 연구는 오랫동안 진행되어 왔지만 각 방법에 대한 비교 분석이 없어 실제 현장에서 통계모델링이 필요한 엔지니어에게는 어려움이 많다. 그러므로 본 논문에서는 다양한 통계모델링 기법의 장단점을 분석함으로써 현장에서 실험 데이터의 통계모델링을 하는 엔지니어에게 적합한 통계모델링 기법을 제안하고자 한다. 이를 위해 기존의 통계모델링 기법에 대한 이론을 소개하고, 통계 시뮬레이션을 통해 각 방법의 장단점을 비교하였다. 또한, 각 방법의 검증을 위해 SHPH440 재질의 탄성계수 데이터를 사용하여 실험 데이터에 대한 통계모델링을 수행하였으며, 각 방법을 적용한 통계모델링 결과를 분석하여 설명하였다.

2. 본론

2.1 통계모델링 방법

2.1.1 MLE

MLE 방법은 우도함수(L)를 사용하되, 음의 로그 우도함수 값을 통계량으로 정의한다. 가장 적합한 모델은 가장 큰 우도함수 값을 갖게 되므로 우도함수 값이 크면 클수록 음의 로그 우도함수 값은 감소한다. 그러므로 MLE 값이 가장 낮은 값을 가진 후보 모델이 데이터 분포와 가장 적합한 모델로 선택된다. 여기서 각 후보모델의 파라미터는 MLE 방법을 이용하여 최대 우도 값을 갖는 파라미터가 선택된다.

$$MLE = -\ln(L) = -\ln\left(\prod_{i=1}^{n} f_k(x_i|\boldsymbol{\theta})\right) \quad (1)$$

$f_k(x_i|\boldsymbol{\theta})$는 i 번째 데이터 x_i 에서 k 번째 모델의 확률밀도함수(Probability Density Function, PDF), n은 데이터 수이다.

MLE 방법은 후보모델이 모분포보다 자유도가 높을 경우, 후보모델의 우도값이 최대가 되는 다수의 파라미터를 자유롭게 선택할 수 있으므로 파라미터의 개수가 많은 후보모델이 모분포로 선정될 확률이 높으며, 데이터 수가 적은 경우 더욱 그러한 경향은 강해지게 된다. 그러므로 모분포의 자유도가 후보모델의 자유도보다 낮은 경우 잘못된 모분포를 선택할 확률이 높다.

2.1.2 AIC

AIC 방법은 가장 최소의 정보 손실(Information loss)를 갖는 모델이 가장 데이터와 적합한 모델로 선택이 된다. 가장 최소의 정보 손실을 갖는 모델은 가장 낮은 AIC 값을 갖게 되므로 최소의 AIC값을 갖는 모델이 최적의 모델로 선택된다[4].

$$AIC = -2\ln(L) + 2k \quad (2)$$

여기서 k는 후보모델의 파라미터 수이다.

AIC는 첫 번째 항에서 MLE 방법을 사용하지만 두 번째 항에서 파라미터 수를 보정하므로 파라미터 수가 높은 모델의 AIC 값에 페널티(Penalty)를 주어 자유도가 높은 모델의 선택을 회피하도록 해준다. 만약 모분포가 자유도가 높은 경우에는 이러한 패널티로 인해 오히려 모분포가 선택되지 않을 수 있지만, 데이터가 증가할수록 우도함수값이 커져 파라미터 수는 AIC 값에 큰 영향을 미치지 않으므로 모분포를 정확하게 예측할 수 있게 해준다. 하지만 AIC방법은 데이터 수에 대한 보정은 없

으므로 작은 수의 데이터가 주어진 경우 정확도가 낮을 수 있다.

2.1.3 AICc

AICc 방법은 데이터 수를 반영하여 보정된 AIC 방법에 해당되며, 데이터 수가 적은 경우 AIC 방법보다 더 정확한 것으로 알려져 있다. AIC와 마찬가지로 최소의 AICc값을 가진 모델이 가장 적합한 모델로 선택된다[5].

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{3}$$

2.1.4 BIC

BIC 방법은 Bayesian 이론에서 우도함수와 사전분포(Prior probability distribution) 이용하여 계산된 사후분포(Posteior probability distribution)를 근사화하여 유도된 통계량이다. 앞의 방법과 유사하게 최소의 BIC값을 갖는 모델이 가장 적합한 모델로 선택되게 된다[6].

$$BIC = -2\ln(L) + k\ln(n) \tag{4}$$

AICc와 BIC 방법은 식 (3)과 (4)의 두 번째 항을 비록 다르게 정의하지만 일반적으로 10개 이상의 샘플이 주어진 경우 로그우도함수 값이 주로 AICc와 BIC 값을 결정하므로 두 방법의 통계모델링 결과는 큰 차이가 없다.

2.1.5 Bayesian method

Bayesian 방법은 Bayesian 이론을 이용하여 각 후보 모델에 대한 가설의 확률을 계산하여 이를 가중치로 표현하는 방법이다. 앞의 방법과는 달리 Bayesian 방법은 각 후보모델의 상대적 가중치를 계산하므로 가중치가 높으면 높을수록 데이터와 가장 적합한 모델을 의미한다[2].

$$\Pr(h_k|D,I) = \frac{\Pr(D|h_k,I)\Pr(h_k|I)}{\Pr(D|I)} \tag{5}$$

여기서 h_k는 k번째 모델의 가설, D는 데이터, I는 사전정보를 의미한다. 식 (5)에서 $\Pr(h_k|D,I)$는 k 번째 모델이 참일 가설의 확률, $\Pr(D|h_k,I)$는 우도함수, $\Pr(h_k|I)$는 k번째 모델의 가설에 대한 사전정보, $\Pr(D|I)$는 정규 상수이다. 식 (5)는 아래 식 (6)과 같이 가중치에 대한 계산으로 표현될 수 있다[2].

$$W_k = \frac{1}{\lambda(\Lambda^\mu)} \int_{\Omega_k^\mu \cap \Lambda^\mu} \prod_{i=1}^{n} f_k(x_i|\boldsymbol{\theta}) d\mu \tag{6}$$

여기서 Λ^μ는 Ω_k^μ는 각 각 평균 μ의 도메인, k번째 모델에서 μ의 도메인에 해당된다.

본래 식 (5)는 각 파라미터에 대해 적분이 필요하지만, 다중 적분은 수치적으로 계산이 어렵고 각 파라미터에 대한 도메인을 결정하기가 번거로우므로 평균에 대한 1차 적분으로 표현할 수 있다.[2] 식 (6)은 다양한 파라미터에 대해 1차 적분으로만 표현하기 때문에 파라미터 수에 대한 보정이 필요하다. 그러므로 본 논문에서는 AIC에서 사용된 파라미터 보정 계수 $2k$ 를 적용하여 아래와 같이 가중치를 계산하였다.

$$W_k' = e^{-k} W_k \tag{7}$$

식 (6)은 식 (2)에서 $AIC = \ln(e^k/L)^2$ 이므로 로그함수 대신 우도함수값에 적용한 결과 이다. (6)에서 계산된 가중치는 각 후보모델에 대한 상대적 가중치 계산을 위해 정규 가중치(normalized weight)로 나타낸다.

$$w_k' = W_k' / \sum_{k=1}^{N_c} W_k' \times 100 \tag{8}$$

N_c는 후보모델의 수이다.

앞의 방법들은 MLE방법을 이용하여 데이터로부터 계산된 파라미터를 사용하여 통계량을 계산하지만, Bayesian 방법은 파라미터에 대한 적분을 수행하므로 파라미터 값에 덜 민감한 결과가 도출된다. 그러므로 특히 데이터 수가 적은 경우, 앞의 방법들보다 통계모델링의 결과가 정확한 경우가 많다.

2.2 통계시뮬레이션

각 방법의 비교를 위해 다양한 분포를 모분포로 가정하였다. 사용된 분포는 Normal (NORM), Lognormal (LOGN), Weibull(WBL), Gamma(GAM), Extreme value (EV), Logistic(LOG), exponential (EXP), Generalized

extreme value(GEV) 분포를 포함해 8개이다. 통계모델링 방법의 정확한 비교를 위해 8개의 분포는 평균(Mean), 분산(Variance), 왜도(Skewness)와 첨도(Kurtosis)를 달리하였다. 이 중 1개 파라미터를 갖는 분포는 EXP, 3개 파라미터를 갖는 분포는 GEV, 나머지 분포는 2개의 파라미터를 갖는다.

Fig. 1 은 각 모분포의 PDF를 나타내며 각 함수의 파라미터는 분포 이름 옆에 기입되어 있다. 각 모분포는 후보모델로 사용되었으며 각 모분포로부터 n=10, 20, 30, 50, 100개의 샘플을 1,000회 생성한 후 가장 적합한 모델을 선택하였다.

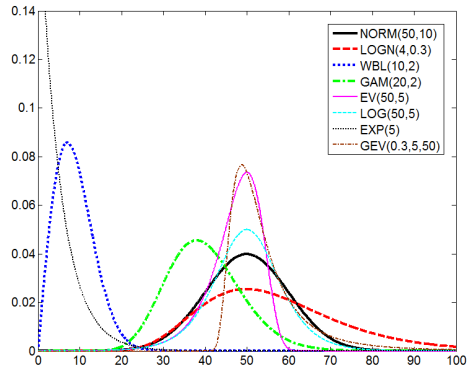

Fig. 1. PDFs of populations

Fig. 2는 NORM과 LOGN 분포가 모분포인 경우 n=10, 20, 30, 50, 100의 샘플을 1,000회 생성하여 모분포를 채택하는 횟수를 채택률(Identification rate)로 나타내었다. 여기서 점과 세모는 각각 NORM과 LOGN이 모분포인 경우를 의미한다.

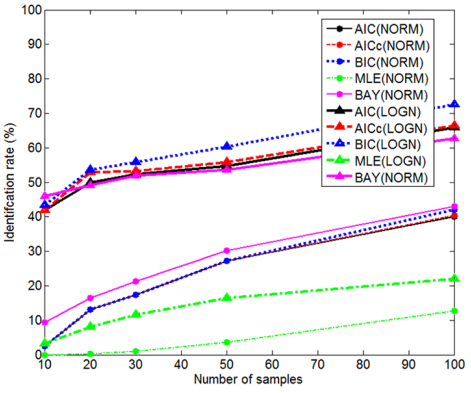

Fig. 2. Identification rates for NORM(50,10) and LOGN(4,0.3)

NORM이 모분포인 경우 Bayesian 방법이 가장 정확했고, AIC, AICc, BIC 방법은 유사한 결과를 보였으며 MLE 방법의 정확도가 가장 낮았다. LOGN이 모분포인 경우 BIC가 가장 높은 채택률을 보였으며, MLE를 제외한 다른 방법의 채택률 역시 유사한 결과를 보였다. 두 가지 경우 모두 샘플 수가 증가함에 따라 모분포의 채택률은 점점 증가하였다.

MLE 방법이 채택률이 낮은 이유는 샘플 수가 적을 경우 모분포의 자유도는 2(파라미터 개수)인데 반해 GEV의 자유도는 3이므로 자유도가 더 높은 분포가 샘플 분포와 더 유사해지므로 NORM보다는 GEV를 모분포로 인식하게 된다. 하지만, AIC, AICc, BIC, Bayesian 방법은 이러한 과도한 적합을 고려해서 보정된 방법이므로 MLE 방법에 비해 모분포의 채택률이 더 높아지게 된다.

NORM이 모분포인 경우 LOGN이 모분포인 경우에 비해 상대적으로 모분포 채택률이 낮다. 이는 다수의 후보모델이 NORM과 유사한 PDF를 갖기 때문이다. Fig. 3을 보면 모분포와 다수의 후보모델의 PDF가 흡사하다는 사실을 알 수 있다. 특히, 샘플 수가 적은 경우 모분포를 정확하게 찾기 어려우므로 유사한 후보 모델이 선택될 확률이 높아지게 된다. 후보모델의 PDF 형상의 유사성은 Bayesian 방법의 정규 가중치 값을 비교해보면 알 수 있다. Fig. 4는 전체 후보모델의 가중치 합이 100이 되도록 각각의 후보함수의 가중치를 1,000회 동안 더해서 10으로 나눈 값을 나타낸 그래프이다.

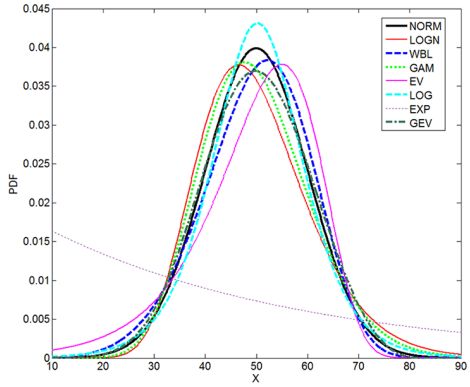

Fig. 3. PDFs of candidate distributions for NORM(50,10)

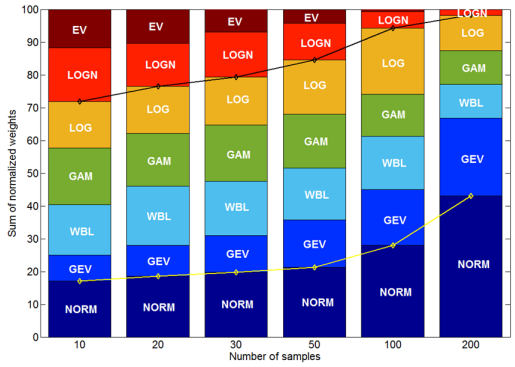

Fig. 4. Sum of normalized weights for NORM(50,10)

샘플 수가 증가하면 후보모델 중 모분포인 NORM과 상이한 분포(EV)의 가중치는 감소하지만 모분포와 유사한 분포들은 높은 가중치를 지속적으로 유지한다. 특히, NORM 분포의 PDF와 가장 유사한 GEV 분포는 샘플 수가 증가함에 따라 오히려 가중치가 증가하는 경향을 보인다. Fig. 3을 보면 모분포와 가장 가까운 모델은 GEV, WBL, GAM, LOG 분포 이므로 NORM 분포를 제외한 4가지 모델의 가중치가 크다. 그러므로 모분포 채택률은 샘플 수가 n=10 인 경우에 10%에 불과하지만 모분포와 유사한 분포인 GEV, WBL, GAM, LOG 분포의 가중치를 더하면 70이상의 가중치 값을 가지며, 채택율의 합 역시 70% 정도를 갖게 된다. Bayesian 방법과 달리 AIC, AICc, BIC, MLE 모두 음의 로그 우도함수 값을 이용하여 각 후보모델의 적합성을 판단하는데, 모분포와 유사한 모델일수록 유사한 로그 우도함수 값을 갖게 된다. 하지만 Bayesian 방법과는 달리 로그 우도함수 값은 정규화 할 수 없어 본 연구에서는 Bayesian 방법을 이용해서 계산된 가중치를 이용하여 각 후보모델의 유사도를 검증하였다.

NORM 분포는 유사한 분포가 많아 모분포의 채택률이 낮았지만, LOGN은 GAM과 GEV와는 유사하지만 이외의 후보모델의 PDF 형상이 다소 상이(Fig. 5)하므로 샘플 수가 적은 경우에도 모분포의 채택률이 상대적으로 높다. LOGN과 유사한 모델인 GAM과 GEV 분포의 가중치를 보면 샘플 수가 증가함에도 불구하고 점점 증가하면서 일정 수준의 가중치를 유지한다는 사실을 알 수 있다.

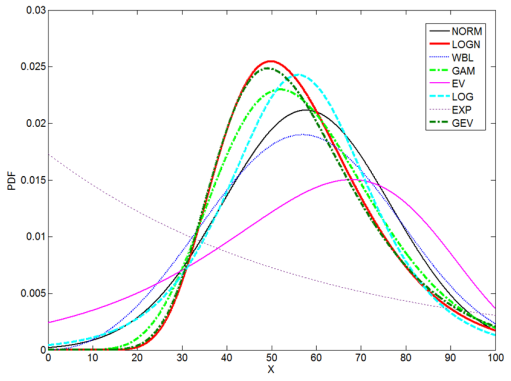

Fig. 5. PDFs of candidate distributions for X~LOGN(4,0.3)

Fig. 2에서 특히 LOGN이 모분포인 경우 BIC의 모분포 채택율이 높다. 그 이유는 GEV가 LOGN과 유사해서 GEV를 모분포로 선택할 확률이 높은데도 불구하고 BIC가 다른 방법에 비해 GEV에 대한 패널티값이 크므로 LOGN을 모분포로 선택하는 빈도가 더 높기 때문이다.

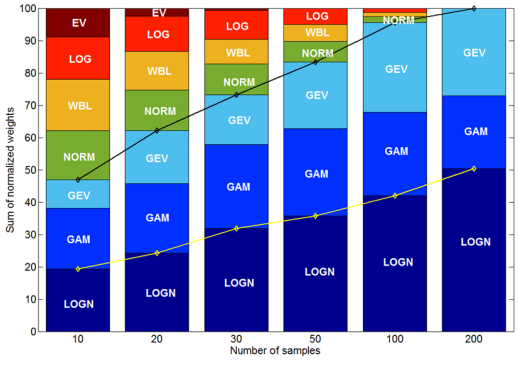

Fig. 6. Sum of normalized weights for X~LOGN(4,0.3)

비록 유사한 후보모델이 적합한 모델로 선택된다고 하더라도 이로 인한 신뢰성 해석 결과와 설계 결과 역시 모분포를 사용한 결과와 유사[1,2]하므로 유사한 모델을 해석과 설계에서 사용된다고 해서 문제가 되지 않는다.

마찬가지로 WBL과 GAM 분포의 모분포 채택률을 보면 AIC, AICc, BIC, Bayesian 방법은 유사한 채택률을 가지지만 MLE는 그에 비해 낮은 채택률을 보인다는 사실을 확인할 수 있다. Bayesian방법은 WBL 이 모분포인 경우 가장 높은 모분포 채택률을 가지며, GAM 이 모분포인 경우 MLE를 제외한 방법들은 유사한 모분포 채택률을 가진다. 특히 GAM 이 모분포인 경우 WBL에 비해 전반적으로 낮은 모분포 채택률을 얻는다. 이러한

경향 역시 NORM과 동일한 이유로 WBL에 비해 GAM 분포의 PDF와 유사한 후보 모델이 다수 있으므로 발생한 것이라고 볼 수 있다.

특히 WBL이 모분포일 때 Bayesian 방법이 다른 방법에 비해 모분포 채택률이 높은 이유는 왜도가 높은 비대칭 분포는 대칭인 분포에 비해 고차의 모멘트계산이 필요하므로 적은 샘플 수로 모분포를 찾아내기가 쉽지 않다. 분포의 비대칭성이 클수록 샘플의 평균값은 모분포의 평균값과는 다른 경우가 많아 우도 값을 기반으로 하는 AIC, AICc, BIC, MLE 방법은 잘못된 모델을 선택하는 경우가 다수 있다. 반면 Bayesian은 우도함수를 평균에 대해 적분하므로 샘플의 평균값이 모분포의 평균값과 다르더라도 다른 방법에 비해 더욱 정확하게 모분포를 찾을 수 있다.

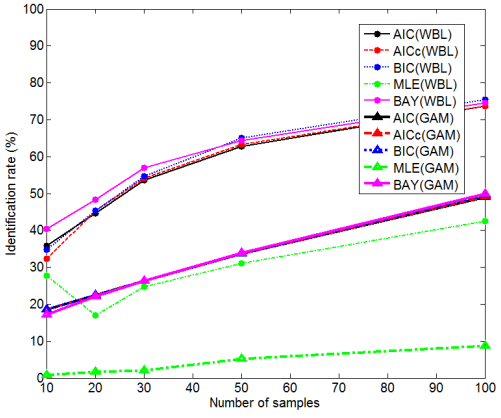

Fig. 7. Identification rate for WBL and GAM

EV가 모분포인 경우 Bayesian, BIC, AICc, AIC 순서로 높은 모분포 채택률을 가진다. EV가 모분포인 경우에도 Bayesian 방법의 정확도가 높은데, 그 이유 역시 모분포의 비대칭성으로 인해 파라미터에 대한 적분을 사용하는 Bayesian 방법이 주어진 샘플로부터 계산된 파라미터를 사용하는 AIC, AICc, BIC, MLE 방법에 비해 파라미터에 덜 민감하기 때문이다. LOG 가 모분포일 경우 NORM과 GAM 의 모분포 채택률 결과와 유사하게 분포의 대칭성으로 인해 MLE를 제외한 모든 방법이 유사한 모분포 채택률을 가지게 된다.

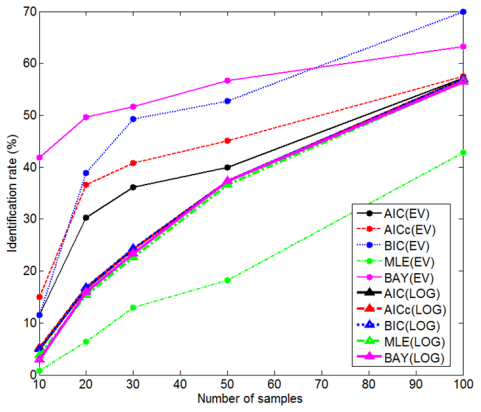

Fig. 8. Identification rate for EV and LOG

모분포가 EXP인 경우(Fig. 9) EXP의 왜도는 상당히 크므로 앞의 EV 경우처럼 Bayesian 방법은 n=10일 때 모분포의 채택률이 100%에 가까운 반면 나머지 방법들은 40-70%의 채택률을 가진다. GEV가 모분포인 경우 Bayesian 방법은 높은 자유도를 갖는 GEV에 대한 페널티로 인해 다른 방법에 비해 모분포 채택률이 다소 낮다. 하지만, 샘플수가 증가함에 따라 AIC, AICc, BIC 의 채택률과 유사한 결과를 가지게 된다.

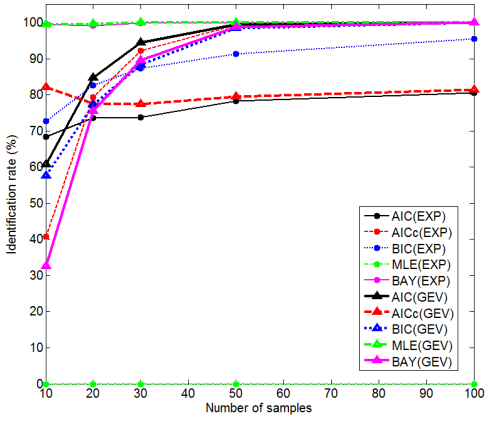

Fig. 9. Identification rate for EXP and GEV

2.3 공학예제

구조용 강의 탄성계수는 기계 시스템의 해석 및 설계에 있어 가장 기본적인 재료 물성에 해당된다. 일반적으로 강 재질의 탄성계수의 대푯값은 주어져 있지만 실제 회사에서 실험 데이터를 공유하고 있지 않으므로 분포 종류와 파라미터에 대한 정보는 얻기 어렵다. 회사에서

는 통계적 정보를 얻기 위해 재료 시험을 하지만 시험 비용이나 많은 시간 소요로 인해 제한된 데이터를 사용하는 경우가 많다. 본 연구에서는 자동차용 구조용 강인 SAPH440 재질의 탄성계수 데이터 22개[7]에 대해 모델 선택법을 적용하여 통계모델링을 수행하였다.

통계모델링 결과 평균값은 208.06 GPa, 변동계수는 2.3% 이며 각각의 모델 선택법을 이용하여 얻은 통계모델링 결과는 Table 2에 나타내었다. 통계 모델링 결과 데이터와 가장 적합도가 높은 순위는 LOG, LOGN, GAM, NORM 이다. 일반적으로 영의 계수는 NORM이나 LOGN 분포를 가장 많이 사용하는데[7], 본 논문에서 사용된 데이터는 평균 주위에 데이터가 밀집되어있고 대칭적인 분포 형태를 가지므로 LOG 분포가 가장 적합한 모델로 선택되었다. 22개의 데이터의 히스토그램과 LOG, LOGN, GAM, NORM 분포의 PDF를 비교해 보면, 4개의 모델 모두 데이터 분포과 유사하며 EV, WBL, EXP 분포가 데이터 분포와는 상이하다는 사실을 확인할 수 있다.

Table 1. Results of statistical modeling

Dist.	AIC	AICc	BIC	MLE	BAY
NORM	134.42	135.05	136.60	65.21	0.168
LOGN	134.17	134.80	136.35	65.08	0.191
WBL	142.41	143.05	144.60	69.21	0.0
GAM	134.22	134.85	136.41	65.11	0.184
EV	143.35	143.98	145.53	69.11	0.0
LOG	**132.22**	**132.85**	**134.40**	**64.11**	**0.386**
EXP	280.87	281.07	281.96	139.43	0.0
GEV	136.21	137.54	139.48	65.10	0.069

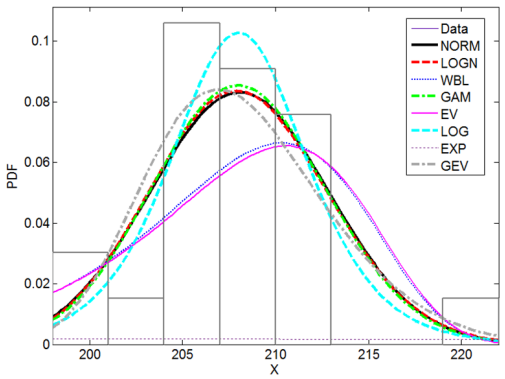

Fig. 10. Histogram of data and PDFs of candidate models

3. 결론

본 논문은 입력 변수의 통계모델링 방법 중 가장 많이 사용되고 있는 AIC, AICc, BIC, MLE, Bayesian 방법에 대해 이론적으로 연구하고 각 방법의 비교 분석을 위해 통계 시뮬레이션을 수행하였으며 결과는 다음과 같이 요약할 수 있다.

1) MLE 는 자유도가 높은 GEV 분포를 과도하게 피팅하여 다양한 종류의 모분포에 대해 모분포 채택률이 현저히 낮았지만 AIC, AICc, BIC, Bayesian 방법은 비교적 높은 모분포 채택률을 가졌다.
2) 모분포가 대칭인 NORM, LOG, GAM 분포인 경우 MLE를 제외한 4가지 방법의 정확도는 유사하였으나 왜도를 갖는 비대칭적 분포인 EV, WBL, EXP 분포의 경우 Bayesian 방법이 나머지 방법에 비해 높은 모분포 채택률을 가졌다.

그러므로 데이터 분포의 대칭성이 있다면, 어떤 방법을 사용해도 유사한 통계모델링 결과를 얻지만, 비대칭적인 분포인 경우 Bayesian 방법이 가장 추천할만한 통계모델링 방법이라 할 수 있다.

References

[1] Y. Noh, and S. Lee “Statistical Modeling of Joint Distribution Functions for Reliability Analysis”, Journal of the Korean Academia-Industrial cooperation Society, Vol. 15, No. 5, 2014, pp. 2603-2609.
DOI: http://dx.doi.org/10.5762/KAIS.2014.15.5.2603

[2] Y. Noh, K. K. Choi, I. Lee, “Identification of Marginal and Joint CDFs Using the Bayesian Method for RBDO”, Structural and Multidisciplinary Optimization, Vol. 40, No.1, 2010, pp.35-51.
DOI: http://dx.doi.org/10.1007/s00158-009-0385-1

[3] S.-Y. Baik, B.-G., Lee, “A Study on Reliability Design of Fracture Mechanics Method Using FEM”, Journal of the Korea Academia-Industrial cooperation Society, Vol. 16, No. 7, pp.4398-4404, 2015.

[4] H. Akaike, “A New Look at the Statistical Model Identification”, IEEE Transactions on Automatic Control, Vol.19, No.6, pp. 716-723, 1974.
DOI: http://dx.doi.org/10.1109/TAC.1974.1100705

[5] N. Sugiura, Further Analysts of the Data by Akaike's Information Criterion and the Finite Corrections”, Communications in Statistics-Theory and Methods, Vol. 7, No. 1, 1978, pp. 13-26.
DOI: http://dx.doi.org/10.1080/03610927808827599

[6] D. F. Findley, “Counter examples to Parsimony and BIC”, Annals of the Institute of Statistical Mathematics, Vol. 43, No. 3, 1991, pp. 505-514.
DOI: http://dx.doi.org/10.1007/BF00053369

[7] J. P. Piper, The Fatigue Properties of Formable Hot Rolled Strip Steels, Ph.D. Thesis, Univ. of Woolongong, 1993.

[8] P. Hess, D. Bruchman, I. Asskkaf, B. M. Ayyub, “Uncertainties in Material Strength, Geometric, and Load Variables”, Naval Engineers Journal, Vol. 114, No. 2, 2002, pp. 139-166.
DOI: http://dx.doi.org/10.1111/j.1559-3584.2002.tb00128.x

노 유 정(Yoojeong Noh) **[정회원]**

- 2009년 12월 : Univ.ofIowa 기계공학과 (공학박사)
- 2010년 12월 ~ 2011년 8월 : 한국기계연구원 선임연구원
- 2011년 9월 ~ 2015년 2월 : 계명대학교 기계자동차공학과 조교수
- 2015년 3월 ~ 현재 : 부산대학교 기계공학부 조교수

<관심분야>
신뢰성 기반 최적 설계, 통계모델링