Journal of the Korea Academia-Industrial cooperation Society
Vol. 15, No. 5 pp. 2603-2609, 2014

http://dx.doi.org/10.5762/KAIS.2014.15.5.2603
ISSN 1975-4701 / eISSN 2288-4688

신뢰성 해석을 위한 결합분포함수의 통계모델링

노유정[1*], 이상진[1]
[1]계명대학교 기계자동차공학과

Statistical Modeling of Joint Distribution Functions for Reliability Analysis

Yoojeong Noh[1*] and Sangjin Lee[1]
[1]Department of Mechanical and Automotive Engineering, Keimyung University

요 약 기계시스템의 신뢰성 해석을 위해서는 기계시스템에 성능을 미치는 변수의 확률 분포와 파라미터를 결정하는 통계적 모델링은 반드시 필요하다. 하지만, 신뢰성 해석에서 상당수의 변수는 상관관계가 있음에도 불구하고 독립변수로 취급되거나 실험데이터 수가 부족하다는 이유로 통계 모델에 대한 잘못된 가정을 하는 경우가 많다. 본 연구에서는 베이지안 방법을 이용하여 상관관계를 갖는 데이터의 결합분포함수를 copula를 이용하여 모델링함으로써 적은 수의 데이터로부터 정확한 입력모델을 산정하는 방법을 제안하였으며, 방법의 검증을 위해 다양한 상관계수와 데이터 수에 대해 통계 시뮬레이션을 수행하였다. 그 결과 Bayesian 방법은 상관계수가 낮아 후보함수가 유사하거나 샘플수가 적어 정확한 모델을 산정하기 어려운 경우에도 후보 copula 중 실제 copula와 가장 근사한 후보 copula를 선정하였다. 이러한 근사 후보 copula는 신뢰성 해석 결과 역시 실제 copula 함수를 이용한 신뢰성 해석 결과와 유사한 결과를 가짐을 확인할 수 있으므로 베이지안 방법은 신뢰성 해석을 위해 정확한 통계모델링을 제공함을 알 수 있다.

Abstract Reliability analysis of mechanical systems requires statistical modeling of input random variables such as distribution function types and statistical parameters that affect the performance of the mechanical systems. Some random variables are correlated, but considered as independent variables or wrong assumptions on input random variables have been used. In this paper, joint distributions were modeled using copulas and Bayesian method from limited number of data. To verify the proposed method, statistical simulation tests were carried out for various number of samples and correlation coefficients. As a result, the Bayesian method selected the most probable copula types among candidate copulas even though the candidate copula shapes are similar for low correlations or the number of data is limited. The most probable copulas also yielded similar reliabilities with the true reliability obtained from a true copula, so that it can be concluded that the Bayesian method provides accurate statistical modeling for the reliability analysis.

Key Words : Bayesian method, Copula, Joint cumulative distribution function, Reliability analysis, Statistical modeling

1. 서론

기계시스템의 신뢰성이란 주어진 환경이나 조건에서 특정 기간 동안 요구되는 시스템의 성능을 만족하는 확률을 의미한다. 대부분의 시스템은 여러 가지 작동요건, 재료물성치, 치수 등에서 다양한 불확실성이 존재하므로 이러한 랜덤 변수에 대한 분포함수의 종류와 평균과 분산과 같은 분포함수의 파라미터를 결정하는 통계모델링

본 논문은 한국연구재단의 지원을 받아 수행된 일반연구자지원사업(NRF-2013R1A1A3013500)임을 밝히며, 연구비 지원에 감사드립니다.
*Corresponding Author : Yoojeong Noh(Keimyung Univ.)
Tel: +82-53-580-5264 email: romana79@kmu.ac.kr
Received April 2, 2014 Revised May 7, 2014 Accepted May 8, 2014

은 반드시 필요하다.

이러한 통계 모델링은 신뢰성 해석을 위한 입력 모델로 이용되며 정확한 신뢰성 해석을 위해서 정확한 통계 모델링은 반드시 필요하다.

실험데이터로부터 통계모델링을 하는 방법에는 대표적으로 Goodness-of-fit(GOF) 방법[1]과 Bayesian 방법[2,3]이 있다. 하지만 GOF 방법은 데이터로부터 계산된 파라미터 값에 의존하므로 파라미터값이 정확하지 않은 경우 잘못된 모델이 선정될 수 있다. 반면 Bayesian 방법은 파라미터를 계산하지 않고 샘플을 이용하여 각 함수의 가중치를 계산하여 상대적으로 파라미터에 대한 의존성이 낮다. 이러한 장점으로 인해 Bayesian 방법을 이용한 통계 모델링에 대한 연구가 이루어지고 있으나[2,3], 이는 결합분포 산정을 위한 통계모델링과 이를 이용한 신뢰성 해석결과에 대한 검증은 충분하지 않은 실정이다.

본 연구에서는 코퓰라(copula)를 이용하여 결합분포함수를 모델링하고, 베이지안(Bayesian)방법을 이용하여 데이터로부터 결합분포함수의 종류를 산정한다. 다양한 상관계수와 결합분포함수의 통계시뮬레이션을 통해 방법의 정확성을 검증하고, 다양한 신뢰성 해석 예제를 통해 데이터로부터 산정된 입력모델이 신뢰성 해석 결과에 미치는 영향을 연구하고자 한다.

2. 통계 모델링

2.1 결합분포함수의 모델링

n개의 랜덤변수를 $\boldsymbol{X}=[X_1,\cdots,X_n]$라고 하면, 결합누적분포함수는 아래 수식(1)과 같이 copula $C(\bullet)$를 이용하여 나타낼 수 있다[4].

$$F_{X_1\cdots X_n}(x_1,\cdots,x_n)=C\left(F_{X_1}(x_1),\cdots,F_{X_n}(x_n)|\boldsymbol{\theta}\right) \quad (1)$$

여기서 $F_{X_1\cdots X_n}(x_1,\cdots,x_n)$ 는 랜덤변수의 결합누적분포함수이며, F_{X_i}는 i 번째 랜덤 변수(X_i)의 부분누적분포함수(marginal cumulative distribution function), $\boldsymbol{\theta}$는 각 변수들 간의 상관관계를 나타내는 상관모수(correlation parameter)의 행렬이다.

수식(1)을 각 변수에 대해 미분하면 결합확률밀도함수(joint probability density function)를 유도할 수 있다.

$$f_{X_1\cdots X_n}(x_1,\cdots,x_n)=c\left(F_{X_1}(x_1),\cdots,F_{X_n}(x_n)|\boldsymbol{\theta}\right)\prod_{i=1}^{b}f_{X_i}(x_i) \quad (2)$$

여기서 $c\left(F_{X_1}(x_1),\cdots,F_{X_n}(x_n)|\boldsymbol{\theta}\right)=\dfrac{\partial^n C(z_1,\cdots,z_n)}{\partial z_1\cdots\partial z_n}$ 는 copula의 밀도함수이며, $f_{X_i}(x_i)$는 X_i의 부분확률밀도함수이다.

2.2 베이지안 방법

베이지안 방법은 주어진 샘플이 몇 가지 후보 함수 중 어떠한 copula 함수와 가장 적합한지를 선정하기 위해 사용되었다. 원리는 임의의 후보 함수에 대해 주어진 데이터와 입력 변수에 대해 알려진 부가적인 정보를 이용하여 각 후보 함수와 데이터가 적합하다는 가설을 확률로 계산하여 주어진 데이터와 가장 적합한 결합분포함수를 결정한다[2,3].

$$\Pr(h_k|D,I)=\frac{\Pr(D|h_k,I)\Pr(h|I)}{\Pr(D|I)} \quad (3)$$

여기서 h_k는 k번째 copula 함수에 대한 가설을 의미하며, D는 데이터, I는 사전 정보를 의미한다. 수식 (3)을 가중치(weight)로 표현하면 수식 (4)와 같이 나타낼 수 있다.

$$W_k=\int_{\Omega_k\cap\Lambda^\tau}\prod_{i=1}^{ns}c_k\left(u_i,v_i|g_k^{-1}(\tau)\right)d\tau \quad (4)$$

여기서 u_i, v_i는 랜덤변수 X, Y로부터 얻은 i번째 샘플 x_i, y_i의 누적분포함수값(cumulative distribution function, CDF)인 $F_{X_1}(x_1)$, $F_{X_2}(x_2)$에 해당되며 ns는 샘플수를 의미한다. τ는 상관계수인 Kendall's tau를 의미하며 수식(1)과 (2)에서의 상관모수는 $\theta=g_k^{-1}(\tau)$로 표현되며 상관계수와 모수의 관계식은 각 후보 copula 함수 종류에 의해 결정된다[4]. Λ^τ는 상관계수 τ의 도메인에 대한 사전정보, Ω_k는 각 후보 copula 가 가지는 상관계수 τ의 도메인을 의미한다. 만약 τ에 대한 사전 정보가 없다면 일반적으로 Λ^τ=[-1,1]가 사용된다. 각각의 후보 copula 함수에 대해 가중치를 계산한 후, 이를 정규화된 가중치(normalized weight)로 나타내면 수식 (5)와 같다.

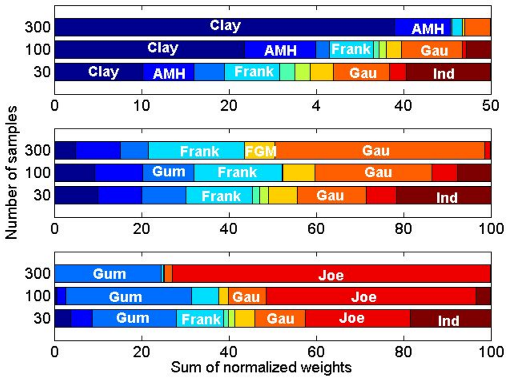
$$w_k=\frac{W_k}{\sum_{i=1}^{Q}W_i} \quad (5)$$

각 후보함수의 가중치를 상대적으로 비교하여 가장 높은 가중치를 갖는 copula가 주어진 데이터를 가장 적

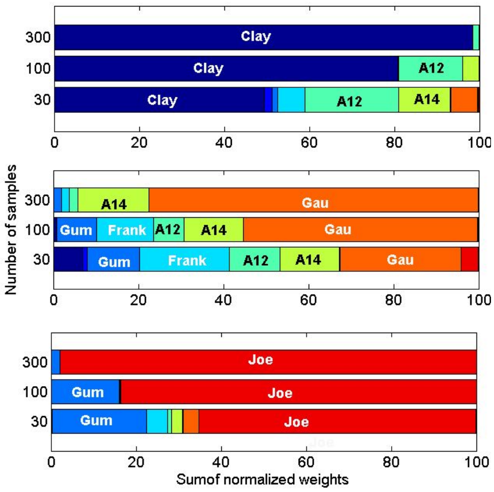
합하게 모델링할 수 있는 결합분포함수로 선정된다.

2.3 통계 시뮬레이션

베이지안 방법이 copula 함수를 얼마나 정확하게 선정하는지 알기 위해 다양한 상관계수(τ=0.2, 0.5, 0.7)와 샘플수(*ns*=30, 100, 300), copula 종류에 대해 통계 시뮬레이션을 수행하였다. 베이지안 방법에서 사용된 후보 copula 함수는 Clayton, AMH(Ali-Mikhail-Haq), Gumbel, Frank, A12, A14, FGM(Farlie - Gumbel - Morgenstern), Gaussian, Joe, Independent copula 가 사용되었다.

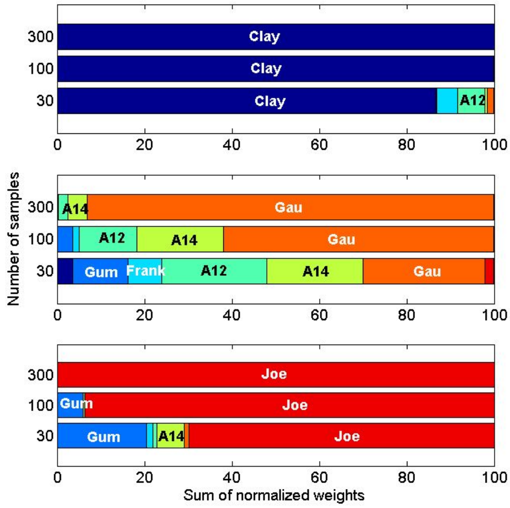


[Fig. 1] Sum of normalized weights for τ=0.2

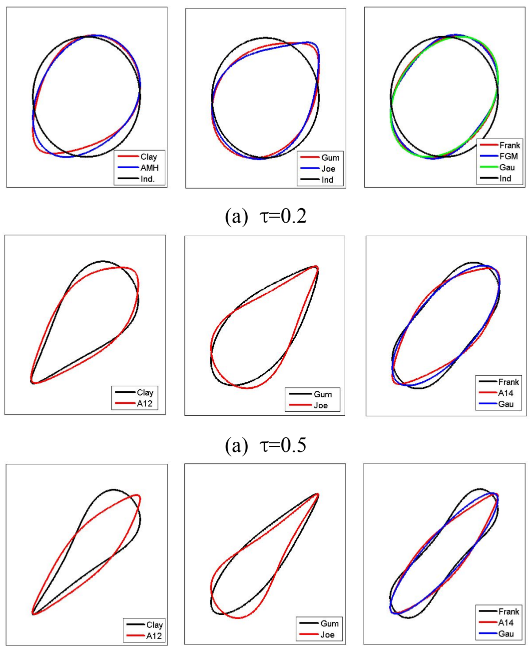


[Fig. 2] Sum of normalized weights for τ=0.5

Fig. 1 - Fig. 3은 Clayton(상단), Gaussian(중간), Joe copula(하단)가 실제(true) copula라고 가정하고 τ=0.2, 0.5, 0.7에서 샘플 수가 30, 100, 300에 대해 100개의 데이터 셋(set)에 대한 정규가중치를 계산한 결과이다. 각각의 데이터 셋이 정규화된 가중치의 합이 1이므로 100개의 데이터 셋의 가중치 합은 100이 된다.



[Fig. 3] Sum of normalized weights for τ=0.7



(a) τ=0.2

(a) τ=0.5

(c) τ=0.7

[Fig. 4] Joint PDF contours for τ=0.2, 0.5, 0.7

τ=0.2 인 경우 Clayton copula는 AMH copula와 유사

한 결합분포함수 형태를 가지므로 두 개의 copula의 가중치는 유사하다. 마찬가지로 Gumbel copula와 Joe copula, 그리고 Frank copula, FGM copula, Gaussian copula 역시 유사한 결합분포함수 모양으로 인해 비슷한 가중치를 갖게 된다. 수식 (4)에서도 확인할 수 있듯이 샘플 수가 늘어나면 늘어날수록 다른 후보 copula 함수에 비해 샘플에서의 실제 copula 밀도함수의 값은 높으므로 실제 copula의 가중치 값은 다른 후보 copula 에 비해 더 높게 나타나는 것을 알 수 있다.

2.4.1 수학예제

copula 형태에 따른 신뢰성 해석 결과 비교를 위해 아래와 같은 한계상태함수(limit state function)를 이용하였다.[5]

$$G(X,Y)=0.7361+(A-6)^2+(A-6)^3 -0.6(A-6)^4+B \tag{6}$$

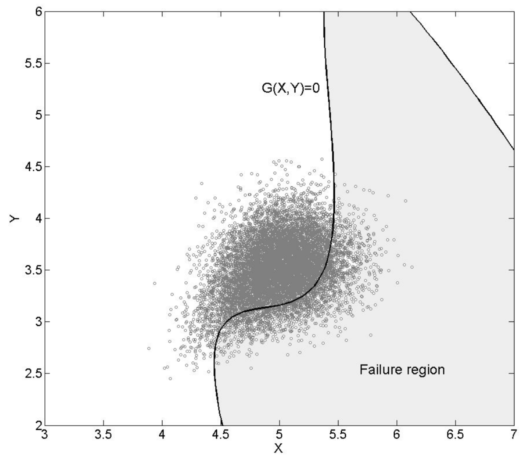
여기서 $\begin{Bmatrix} A \\ B \end{Bmatrix} = \begin{bmatrix} 0.9063 & 0.4226 \\ 0.4226 & -0.9063 \end{bmatrix} \begin{Bmatrix} X \\ Y \end{Bmatrix}$ 이며, 두 변수는 정규분포 $X \sim N(5,0.3), Y \sim N(3.5,0.3)$ 를 따른다고 가정하였다. τ=0.2, 0.5, 0.7에 대해 10개의 copula로부터 샘플을 무작위로 추출하였으며 MCS(Monte Carol Simulation)을 이용하여 한계상태함수가 0보다 작은 경우에 대한 확률을 신뢰도로 계산하였다.

$$P_r = P[G(X,Y)>0] = \frac{1}{N}\sum_{i=1}^{N} I_i \tag{7}$$

여기서 I_i는 indicator function이며 $z \geq 0$이면 $I_i=0$, $z<0$이면 $I_i=0$ 이다. N은 MCS의 샘플수이며, 본 논문에서는 100,000개의 샘플을 사용하였다.

Fig. 5는 한계상태함수와 Clayton copula로부터 10,000개의 샘플 생성하여 함께 나타낸 그림이다. $G(X,Y)<0$은 실패영역(failure region)에 해당되며, 실패영역 이외의 영역에 있는 샘플의 수를 계산하여 신뢰도를 계산하게 된다. 그림에서 보는 바와 같이 copula 종류와 상관계수값에 따라 copula 형태가 달라지고, 이에 따라 신뢰도값 역시 달라지므로(Table 1) copula 종류 선정은 중요하다는 사실을 알 수 있다. 여기서 AMH의 상관계수의 도메인은 [-0.182, 0.333]이므로 τ= 0.5, 0.7에 대한 신뢰도 계산은 불가능하다.

앞서 설명한 바와 같이 copula 형태가 유사한 낮은 상관계수에 대해서는 전반적으로 신뢰도값이 유사하며, 상관계수가 높을수록 각 copula 종류에 따른 신뢰도값 역시 차이가 크다. 상관계수가 낮을수록 정확한 copula 종류를 선정하기는 어려우나 유사한 결합분포함수 형태로 인해 신뢰도 역시 유사하므로 copula 종류 선정 오류로 인한 신뢰도 해석 결과의 오류는 크지 않음을 알 수 있다. 반면 상관계수가 높은 경우 copula 종류에 따른 신뢰도값의 차이가 크므로 정확한 copula 종류를 선정해야한다. 하지만 통계시뮬레이션 결과에서도 알 수 있듯이 높은 상관계수의 경우 적은 데이터 수에도 copula 종류 선정이 쉬우므로 실험데이터만 주어진다면 결합분포함수에 대한 정확한 통계 모델링 가능하며, 결과적으로 신뢰도 결과 역시 정확한 결과를 얻을 수 있다.



[Fig. 5] Mathematical Example

[Table 1] Reliabilities of the mathematical example

Copulas	τ=0.2	τ=0.5	τ=0.7
Clayton	0.8126	0.8766	0.9171
AMH	0.8067	-	-
Gumbel	0.8016	0.8462	0.8820
Frank	0.8046	0.8452	0.8751
A12	0.8226	0.8657	0.9024
A14	0.8288	0.8603	0.8911
FGM	0.8013	0.8235	0.8193
Gaussian	0.8023	0.8490	0.8921
Joe	0.7989	0.8374	0.8622
Independent	0.7826	0.7819	0.7806

하지만, 극히 적은 수의 샘플(20개 이하)의 경우 인지론적 불확실성(epistemic uncertainty)이 존재하므로 기존의 통계 모델링을 할 경우 통계모델을 선정하는데 어려움이 있다. 잘못된 통계모델이 선택될 경우, 신뢰성 해석 결과 역시 믿을 수 없으므로 향후 연구 계획은 인지론적 불확실성을 갖는 데이터로부터 신뢰도 있는 신뢰성 해석의 결과를 얻기 위한 통계모델링 기법 개발을 수행할 예정이다.

2.4.2 공학예제

피로에 의한 파괴는 기계적인 파괴의 70~80%를 차지하고 있으며, 대부분의 피로파괴는 반복 하중으로 인해 수명의 예측이 어렵고 자동차나 항공기, 발전소의 기계부품 설계에 있어 중요한 관심사이다. 하지만, 피로 수명에 영향을 주는 다양한 인자의 정확한 확률적 특성을 고려하지 않고 피로 수명을 예측하는 경우가 많아 실제 피로 수명과는 상당한 차이가 있다.

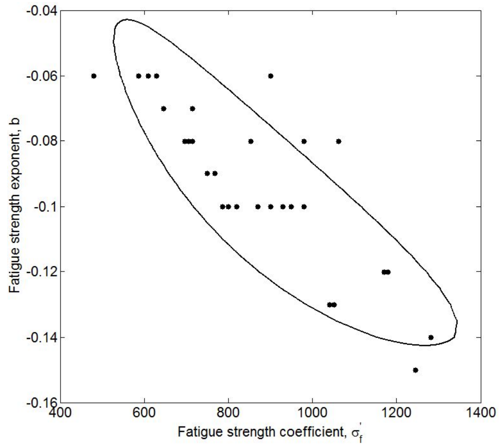
본 예제에서는 950X Steel 재질에 대해 중요한 피로재료물성치인 피로강성계수($\sigma_f^{'}$)와 지수(b), 피로연성계수($\varepsilon_f^{'}$)와 지수(c)의 29 셋의 실험데이터를 이용하여 통계모델링을 수행하고, 이를 이용하여 탄성효과와 소성효과가 동일하게 나타나는 천이 피로 수명을 확률적으로 예측하였다.

피로강성계수와 지수, 연성계수와 지수의 데이터는 D. Socie의 피로기술노트[6]로부터 얻었으며, 베이지안방법을 이용하여 각 후보 copula에 대한 가중치를 계산하였다. 피로강성계수와 연성계수의 부분분포함수는 로그정규함수(lognormal distribution)를, 피로강성지수와 연성지수는 정규함수(normal distribution)을 따르는 것으로 알려져 있으며, 각 분포함수의 파라미터는 데이터로부터 직접 계산하였다. 각각의 부분분포함수를 이용하여 각각의 후보 copula에 대한 가중치를 계산한 결과, 피로강성계수와 지수의 결합분포함수로는 Gaussian copula, 피로연성계수와 지수의 결합분포함수로는 Frank copula가 선정되었다. 다른 후보에 대한 가중치는 모두 0에 가까우므로 Table 2에서는 생략되었다.

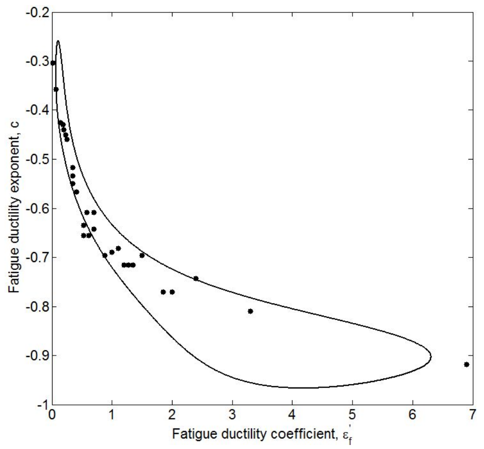
실험데이터와 선정된 copula를 이용하여 결합분포함수의 등고선과 실제 실험데이터를 그래프로 함께 나타내보면 Fig. 6과 같다. 그림에서 보는 바와 같이 선정된 copula는 실제 실험데이터가 흩어진 형태와 흡사한 PDF 등고선을 갖는다.

[Table 2] Normalized weights of candidate copulas

Copulas	$\sigma_f^{'}$ & b	$\varepsilon_f^{'}$ & c
Frank	0.3341	0.9991
Gaussian	0.6658	0.0009



(a) Fatigue strength coefficient and exponent



(b) Fatigue ductility coefficient and exponent

[Fig. 6] Paired data for 950X Steel and PDF contours of Gaussian copula and Frank copula

선정된 모델을 이용하여 피로 변형률 진폭에 기여하는 탄성효과와 소성효과가 동일할 경우의 피로수명인 천이피로수명을 예측해 보자. 아래 수식은 천이 피로 수명을 나타내는 수식이다[7].

$$2N_f = \left(\frac{E\varepsilon_f^{'}}{\sigma_f^{'}} \right)^{1/(b-c)} \tag{8}$$

여기서 $2N_f$가 천이 피로 수명이며, E는 영의 계수이다. 950X Steel의 경우 190~210GPa의 영의계수를 가지므로 E=190GPa을 사용하였다.

임계 천이 피로수명(N_c)을 각각 7,000 사이클과 8,000 사이클로 선정하면 각각의 허용 피로 수명보다 높을 확률 계산한 결과 아래와 같은 결과를 얻을 수 있다.

[Table 3] Reliabilities of the fatigue example

Cases	N_c=7,000	N_c=8,000
Gau-Frank	0.878	0.830
Frank-Frank	0.806	0.768
Ind-Ind	0.596	0.577

만약 선정된 copula를 이용하여 계산한 경우 7,000과 8,000 사이클의 임계 피로 수명에 대해 0.878, 0.830의 신뢰도를 얻을 수 있지만, Gaussian copula 대신 Frank copula를 적용할 경우 확률은 감소하며, 각각의 변수가 독립변수라고 가정할 경우(Independent copula) 확률은 더욱더 감소한다. 만약 실제 피로강성계수와 지수의 결합분포함수가 선정된 Gaussian copula로 모델링할 수 있다면, 두 번째로 가중치가 높은 Frank 를 사용할 경우 낮은 확률을 가지므로 더 보수적인 피로수명예측이 가능하며, 독립변수로 가정할 경우 피로수명은 지나치게 보수적인 결과를 가질 수 있다. 그러므로 정확한 피로수명예측을 위해서는 정확한 통계 모델링이 필요하다는 사실을 다시 한 번 확인할 수 있다.

3. 결론

본 논문에서는 상관관계를 갖는 변수의 데이터로부터 결합분포함수를 결정하기 위해 copula를 이용하여 결합분포함수를 모델링 하였으며, 베이지안 방법을 이용하여 copula 종류를 결정하였다. 상관계수와 copula 종류, 샘플 수에 따른 통계시뮬레이션을 수행한 결과 적은 수의 데이터가 있는 경우에도 베이지안 방법은 실제 copula를 정확하게 산정하였다. 간단한 수학예제와 공학예제의 신뢰성 해석을 통해 통계모델링의 오류가 신뢰성 해석에 미치는 영향을 관찰하였으며, 정확한 신뢰성 해석 결과를 얻기 위해서는 정확한 통계 모델링이 필요하다는 사실을 확인할 수 있었다. 본 연구의 결과는 극히 적은 수의 데이터를 갖는 랜덤변수의 통계모델링 기법 개발의 기초연구 자료로 이용될 예정이며, 향후 인지론적 불확실성을 갖는 신뢰성 해석문제에서 결과의 신뢰도를 보장할 수 있는 통계 모델링 기법을 연구할 예정이다.

References

[1] C. Genest, A.C. Favre, "Everything you always wanted to know about copula modeling but were afraid to ask". Journal of Hydrologic Engineering, Vol. 12, No.4, pp. 347--368, 2007.
DOI: http://dx.doi.org/10.1061/(ASCE)1084-0699(2007)12:4(347)

[2] D. Huard, G. Évin, A.C. Favre, "Bayesian Copula Selection", *Computational Statistics & Data Analysis*, Vol. 51, pp. 809-822, 2006.
DOI: http://dx.doi.org/10.1016/j.csda.2005.08.010

[4] Y. Noh, I. Lee, K.K. Choi, "Identification of Marginal and Joint CDFs Using Bayesian Method for RBDO", Structural Multidisciplinary Optimization, Vol.40, p. 35-51, 2010.
DOI: http://dx.doi.org/10.1007/s00158-009-0385-1

[5] R.B. Nelseon, *An Introduction to Copulas, Springer*, New York, 1999.
DOI: http://dx.doi.org/10.1007/978-1-4757-3076-0

[6] I. Lee, K.K. Choi, L. Du, D. Gorsich, "Inverse analysis method using MPP-based dimension reduction for reliability-based design optimization of nonlinear and multi-dimensional systems" Computational Methods Applied Mechanical Engineering, Vol. 198, pp. 14-27, 2008.
DOI: http://dx.doi.org/10.1016/j.cma.2008.03.004

[7] D. Socie, Probabilistic Statistical Simulations Technical Background, eFatigue LLC, 2008, https://www.efatigue.com/probabilistic/background/statsim.html#Cor, April, 2014.

[8] G. Nam, B. Lee, “Evaluation of Chassis Component Reliability Considering Variation of Fatigue Data”, Journal of Korean Society for Precision Engineering, Vol. 24, No. 2, pp.110-117, 2007.

노 유 정(Yoojeong Noh) **[정회원]**

- 2001년 2월 : 부산대학교 기계공학부 (공학사)
- 2003년 2월 : 한국과학기술원 기계공학과 (공학석사)
- 2009년 12월 : Univ. of Iowa 기계공학과 (공학박사)
- 2010년 2월 ~ 2010년 12월 : 부산대학교 박사후연구원
- 2010년 12월 ~ 2011년 8월 : 한국기계연구원 선임연구원
- 2011년 9월 ~ 현재 : 계명대학교 기계자동차공학과 조교수

<관심분야>
신뢰성해석, 신뢰성 기반 최적설계, 통계모델링

이 상 진(Sangjin Lee) **[준회원]**

- 2014년 2월 : 계명대학교 기계자동차공학과 (공학사)
- 2014년 3월 ~ 현재 : 계명대학교 기계공학과 (대학원)

<관심분야>
전산해석, 신뢰성 기반 최적설계, 통계모델링