J. Comput. Struct. Eng. Inst. Korea, 32(1)
pp.55~63, February, 2019
https://doi.org/10.7734/COSEIK.2019.32.1.55

pISSN 1229-3059 eISSN 2287-2302
Computational Structural
Engineering Institute of Korea

다변량 확률분포함수의 추정을 위한 MKDE-ebd 개발

강 영 진[1] · 노 유 정[2†] · 임 오 강[2]

[1]부산대학교 기계기술연구원, [2]부산대학교 기계공학부

Development of MKDE-ebd for Estimation of Multivariate Probabilistic Distribution Functions

Young-Jin Kang[1], Yoojeong Noh[2†] and O-Kaung Lim[2]

[1]*Research Institute of Mechanical Technology, Pusan Nat'l Univ., Busan, 46241, Korea*
[2]*School of Mechanical Engineering, Pusan Nat'l Univ., Busan, 46241, Korea*

Abstract

In engineering problems, many random variables have correlation, and the correlation of input random variables has a great influence on reliability analysis results of the mechanical systems. However, correlated variables are often treated as independent variables or modeled by specific parametric joint distributions due to difficulty in modeling joint distributions. Especially, when there are insufficient correlated data, it becomes more difficult to correctly model the joint distribution. In this study, multivariate kernel density estimation with bounded data is proposed to estimate various types of joint distributions with highly nonlinearity. Since it combines given data with bounded data, which are generated from confidence intervals of uniform distribution parameters for given data, it is less sensitive to data quality and number of data. Thus, it yields conservative statistical modeling and reliability analysis results, and its performance is verified through statistical simulation and engineering examples.

Keywords : *correlated data, multivariate kernel density estimation, multivariate kernel density estimation with estimated bounded data, nonparametric statistical method, relative root mean squared error, reliability analysis*

1. 서 론

공학문제에서 입력변수들은 불확실성을 가지는 경우가 많으며 이러한 입력변수들은 서로 상관관계(correlation)를 가지면서 해석 및 설계의 결과에 영향을 준다. 해석 및 설계에서 서로 상관관계가 있는 불확실성 데이터를 다루기 위해서는 주어진 데이터로부터 결합확률밀도함수(joint probability density function, jPDF) 또는 결합누적분포함수(joint cumulative distribution function, jCDF)를 정의하는 통계적 모델링 과정이 필요하다(Noh *et al*., 2010). 이러한 통계모델링 방법은 사전에 정의된 분포함수와 실험 데이터를 사용해서 분포함수를 추정하는 모수적 통계모델링 방법(parametric statistical modeling method)과 데이터만을 사용해서 추정하는 비모수적 통계모델링 방법(nonparametric statistical modeling method)으로 구분된다(Kang *et al*., 2017; Kang *et al*., 2018a ; Kang *et al*., 2018b). 모수적 방법은 사전에 정의된 분포함수를 사용하여서 추정된 각 변수의 주변 확률함수(marginal probability function)와 코퓰라(copula)와 같이 상관관계를 표현하는 특정 함수의 모델링 결과를 조합하여서 결합확률밀도함수 또는 결합누적분포함수를 추정하는 방법이다(Noh *et al*., 2010 ; Hong *et al*., 2018). 반면에 비모수적 방법은 주변확률분포와 상관관계의 모델링없이 오직 데이터만을 사용하여서 결합확률분포를 추정하는 방법이다.

다변량 커널밀도추정(multivariate kernel density estimation, MKDE)은 대표적인 비모수적 통계모델링 방법으로써 결합분포를 모델링하기 위해서 데이터만을 필요로 하므로

† Corresponding author:
Tel: +82-51-510-2308; E-mail: yoonoh@pusan.ac.kr
Received November 5 2018; Revised November 13 2018;
Accepted November 14 2018

데이터의 개수가 적은 경우 모수적 기법보다 분포함수의 추정 정확도가 우수하고 분포형태의 표현의 자유도가 높아서 다봉분포(multimodal distribution)에 대해서 높은 정확도를 보이는 것으로 확인되었다(Hong *et al.*, 2018). 하지만 MKDE은 확률밀도함수의 꼬리를 짧게 추정하는 경향이 있어서 신뢰성 해석에서 파손확률을 과소추정(underestimation)하는 심각한 문제점이 있는 것으로 확인되었다(Hong *et al.*, 2018).

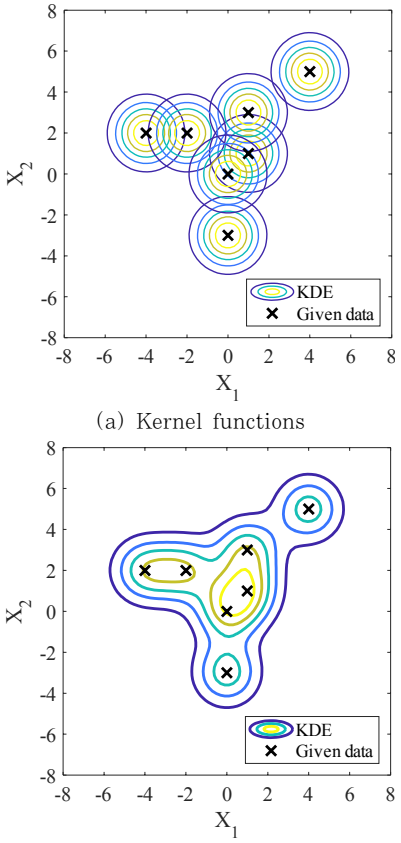
본 논문에서는 MKDE가 신뢰성 해석 시 파손확률을 과소추정하는 문제점을 개선하기 위해서 단변량 데이터를 위한 경계데이터를 이용한 커널밀도추정 방법(Kang *et al.*, 2018a; 2018b)을 다변량 문제로 확장함으로써 데이터의 개수가 적은 경우에는 꼬리가 두껍고 긴(보수적인) 결합확률밀도함수를 추정하는 경계데이터를 이용한 다변량 커널밀도추정(MKDE with estimated bounded data, MKDE-ebd)를 제안하고 통계적 시뮬레이션을 통해 분포함수의 추정 정확도와 신뢰성 해석을 통해 MKDE와 MKDE-ebd의 특성을 비교하였다.

2. 결합확률분포함수의 추정 방법

2.1 다변량 커널 밀도 추정(MKDE)

MKDE는 특정한 확률분포함수와 상관관계에 대한 정의없이 오직 데이터만을 사용하여 결합확률밀도함수를 추정하는 비모수적 통계모델링 방법이다. MKDE는 분포함수의 형태가 고정되지 않고 데이터에 따라서 유연하게 표현할 수 있기 때문에 분포함수의 표현의 자유도가 매우 높은 기법이다. MKDE는 각 데이터에 동일한 커널함수(kernel function)를 정의하고 모든 커널함수를 최적의 대역폭(bandwidth)에 따라서 합하면 최종적으로 결합확률밀도함수가 추정되게 된다(Silverman, 1986). Fig. 1은 이변량(bivariate) 데이터에서 모든 데이터에서 생성된 커널함수와 MKDE를 사용해서 추정된 결합확률밀도함수의 등고선을 나타낸 것이다. Fig. 1(a)는 모든 주어진 데이터(given data)에서 커널함수를 그린 것이고, (b)는 MKDE를 통해 추정된 확률밀도함수를 나타낸다. Fig. 1(a)의 등고선의 확률밀도함수값의 레벨은 [0.026, 0.053, 0.079, 0.0105, 0.132]이고, (b)의 레벨은 [0.006, 0.011, 0.017, 0.023, 0.0028]이며 등고선의 색이 밝아질수록 높은 값을 나타낸다. MKDE을 사용한 추정된 확률밀도함수의 수식은 다음과 같다(Scott, 2015).

$$f_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|^{1/2}} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}^i}{\mathbf{H}^{1/2}}\right) \tag{1}$$



(a) Kernel functions

(b) PDF using MKDE

Fig. 1 Contours using Kernels and MKDE

여기서, $f_{\mathbf{H}}(\mathbf{x})$는 추정된 결합확률밀도함수이고, $\mathbf{X}^i$는 n개의 주어진 상관 데이터이다. $K(\bullet)$는 커널함수이고, $\mathbf{H}$는 $d \times d$ 대각-대역폭 행렬(diagonal bandwidth matrix)이며, d는 다변량변수의 개수(차원)이다. MKDE는 대역폭과 커널함수 종류의 선택이 분포함수의 추정 정확도에 중요하고, 특히 대역폭의 선정이 더욱 중요하다(Silverman *et al.*, 1986; Chen, 2015). 그러므로 본 논문에서는 일반적으로 가장 많이 사용되고 최적의 대역폭에 대한 수식이 정의된 가우시안 커널(gaussian kernel)을 사용하였다(Silverman *et al.*, 1986; Chen, 2015; Kang *et al.*, 2018a; 2018b). 다변량 가우시안 커널 함수는 다음과 같다.

$$K_g(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right) \tag{2}$$

최적의 대역폭은 Silverman's rule of thumb를 사용하였으며 최적의 대역폭 $\mathbf{H}$의 주 대각성분은 다음과 같다(Silverman *et al.*, 1986; Scott, 2015).

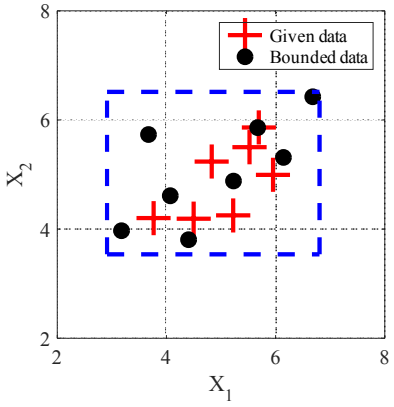
$$h_l^* = \left(\frac{4}{(d+2)n}\right)^{1/(d+4)} \hat{\sigma}_l, \; l = 1,\ldots,d \tag{3}$$

여기서, $\hat{\sigma}_l$은 l번째 변수의 표본에 대한 표준편차이다.

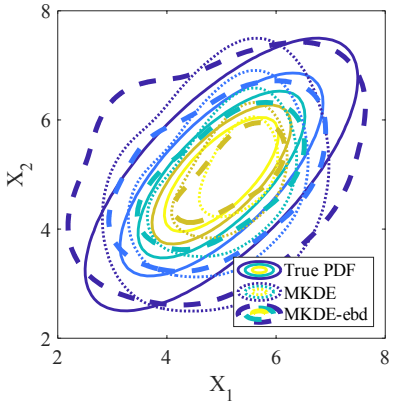
2.2 경계데이터를 이용한 다변량 커널 밀도 추정 (MKDE-ebd)

MKDE는 다양한 분포형태를 표현할 수 있지만, 데이터의 개수가 매우 적은 경우에는 결합확률밀도함수의 꼬리가 짧은 분포함수를 추정하여서 신뢰성 해석에서 파손확률(probability of failure)을 과소추정하는 심각한 문제점이 있다(Hong *et al.*, 2018). 공학적 관점에서 정보가 부족한(데이터가 부족한) 경우에는 파손확률을 보수적으로 과대추정(overestimation)하는 것이 보다 안전한 시스템을 설계할 수 있으므로 데이터가 적은 경우에 대해서 기존의 MKDE보다 결합확률밀도함수를 두껍고 길게(보수적으로) 추정할 수 있는 방법이 필요하다.

MKDE-ebd는 단변량 변수에 대한 KDE-ebd를 다변량 변수로 확장한 것으로, 주어진 데이터에 추정된 경계로부터 추출된 경계데이터를 합하여서 결합확률밀도함수를 추정한다. Fig. 2는 이변량 데이터에 대해서 MKDE-ebd에 의해 추가된 경계데이터와 추정된 결합확률밀도함수의 등고선이다. Fig. 2(a)는 주어진 데이터와 경계데이터를 함께 나타낸 것으로



(a) Data



(b) Contour of PDF

Fig. 2 Data and PDF using MKDE-ebd

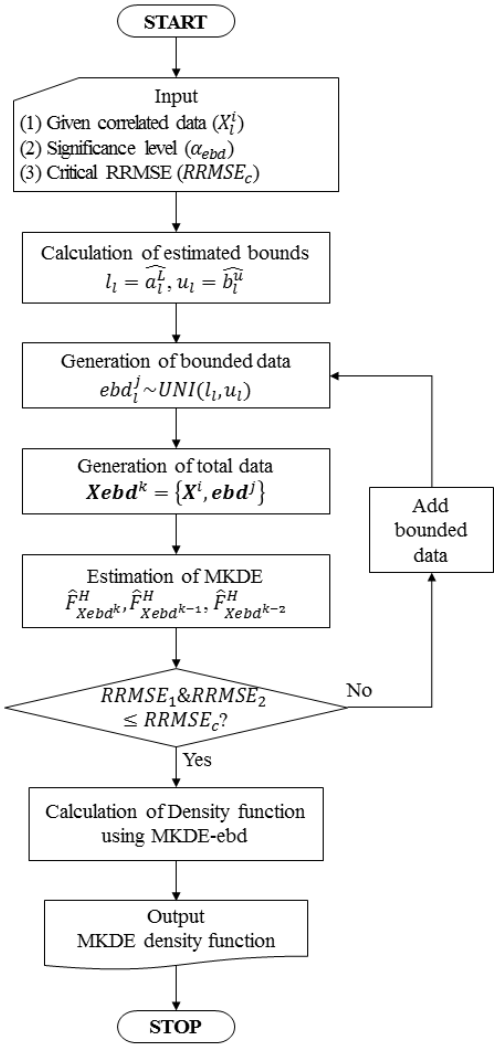


Fig. 3 Flowchart of MKDE-ebd

경계데이터는 쇄선으로 표시된 추정된 경계영역 내에서 생성되고 전체 데이터에 대해서 결합확률밀도함수를 추정하면 (b)의 MKDE-ebd와 같이 기존 MKDE보다 꼬리가 긴 보수적인 분포함수가 추정된다. Fig. 2(b)에서 True PDF, MKDE, MKDE-ebd의 등고선 레벨은 모두 [0.01, 0.04, 0.07, 0.1, 0.13]이다.

Fig. 3은 MKDE-ebd의 흐름도(flowchart)를 보여주는 것으로 먼저 l번째 변수의 주어진 데이터를 사용하여 l번째 변수에 대해 균일분포(uniform distribution)로 구간추정(interval estimation)을 한다. l번째 변수의 추정된 구간은 다음과 같다.

$$\{l_l, u_l\} = \left[\hat{b}_l - \frac{\hat{b}_l - \hat{a}_l}{\alpha_{ebd}^{1/n}}, \hat{a}_l + \frac{\hat{b}_l - \hat{a}_l}{\alpha_{ebd}^{1/n}}\right] \quad (4)$$

여기서, $\hat{a}_l$과 $\hat{b}_l$은 각각 균일분포함수에 대한 점 추정치(point

estimate)의 하한-상한 값이고 α_{ebd}는 구간추정에서 유의수준(significance level)이며, l_l과 u_l은 구간추정에 의한 l번째 변수의 하한-상한 경곗값이 된다.

각 변수의 구간추정을 통해 결정된 경계를 다차원으로 결합하게 되면 Fig. 2(a)처럼 경계영역이 생성되고 이 경계영역 내에서 결측자료(missing data) 또는 부족한 데이터를 보완하기 위한 경계데이터(bounded data)를 추가적으로 생성하게 된다. 경계데이터는 균일분포함수를 따르기 때문에 개수가 너무 많으면 모집단의 분포와 다르게 되고 너무 적으면 보완효과가 거의 없게 된다. 그러므로 생성된 경계로부터 적절한 경계데이터의 개수를 결정하는 과정이 수행되어야 한다. 단변량 변수의 경우에는 경계데이터의 개수를 결정할 때 면적척도(area metric)인 교차면적(intersection area)을 사용하였지만(Kang *et al*., 2018a ; Kang *et al*., 2018b), 다변량 변수에서는 차원이 증가할수록 교차면적의 사용은 매우 비효율적이기 때문에 MKDE-ebd에서는 경계데이터의 생성 및 추가를 위해서 상대 평균 제곱근 오차(relative root mean squared error, RRMSE)를 사용하여서 결합누적분포함수의 차이를 비교한다. 결합누적분포함수 값의 RRMSE는 다음과 같이 구해진다(Li *et al*., 2013).

$$RRMSE=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(F_i^1-F_i^2\right)^2}}{\frac{1}{n}\sum_{i=1}^{n}F_i^1}\times 100 \tag{5}$$

여기서, F_i^1는 기준이 되는 결합누적분포함수 값이고, F_i^2는 비교되는 결합누적분포함수 값이다. RRMSE값이 10%보다 작으면 정확도가 매우 우수한 것이고, 10~20%이면 우수, 20~30%이면 적당함, 30%보다 크면 부정확하다고 판단된다(Jamieson *et al*., 1991 ; Li *et al*., 2013).

먼저 주어진 데이터만 사용해서($i=n$, $j=0$, $k=i+j$) k, k-1, k-2개의 데이터에 대해서 수식 (1)을 사용해서 각 개수에서 결합확률밀도함수를 계산하고 이를 적분하여서 결합누적분포함수 $\hat{F}^H_{Xebd^k}$, $\hat{F}^H_{Xebd^{k-1}}$, $\hat{F}^H_{Xebd^{k-2}}$를 계산한다. 그리고 $\hat{F}^H_{Xebd^k}$와 $\hat{F}^H_{Xebd^{k-1}}$, $\hat{F}^H_{Xebd^k}$와 $\hat{F}^H_{Xebd^{k-2}}$의 RRMSE, 즉 $RRMSE_1$과 $RRMSE_2$를 계산하고 두 값이 모두 임계 RRMSE(critical RRMSE, $RRMSE_c$)보다 작거나 같으면 경계데이터가 추가되더라도 분포함수의 추정에 영향을 주지 않기 때문에 추가적인 경계데이터를 필요로 하지 않으므로 경계데이터의 생성을 종료한다. 반대의 경우에는 추가적인 경계데이터를 필요로 하므로 경계데이터를 1개씩 증가시면서 위 과정을 반복한다. 경계데이터의 개수를 결정하기 위한 알고리즘에 대해서는 Kang 등(2018a; 2018b)의 논문에 상세히 설명되어 있다.

경계데이터의 생성이 완료되면 주어진 데이터와 경계데이터를 합친 전체 데이터에 대해서 다변량 커널밀도추정을 수행하고 MKDE-ebd에 의해서 추정된 결합확률밀도함수는 다음과 같다.

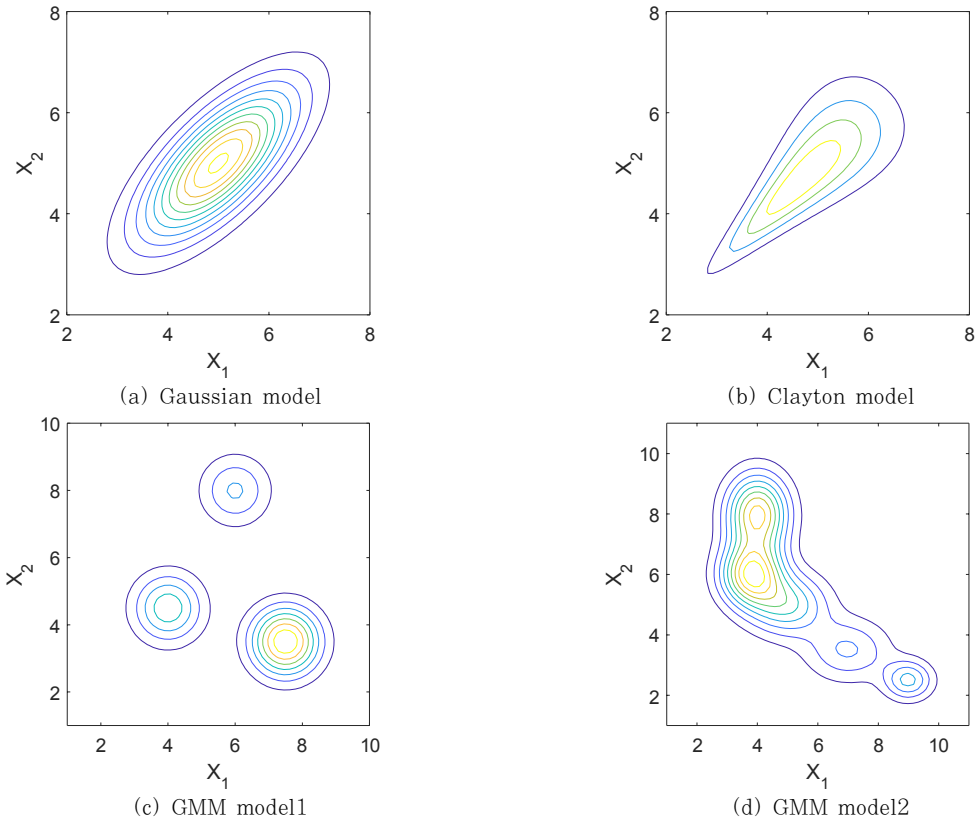
$$f^{ebd}_{\boldsymbol{H}}(\mathbf{x})=\frac{1}{n|\mathbf{H}^k|^{1/2}}\sum_{k=1}^{n+m}K\left(\frac{\mathbf{x}-\mathbf{Xebd}^k}{\mathbf{H}^{k^{1/2}}}\right) \tag{6}$$

여기서, $\mathbf{Xebd}^k=\{\mathbf{X}^i,\mathbf{ebd}^j\}_{k=1}^{n+m}$ 주어진 데이터 $\mathbf{X}^i$와 경계데이터 $\mathbf{ebd}^j$가 합쳐진 전체 상관 데이터이고 $\mathbf{H}^k$는 전체 상관 데이터에 대해서 계산된 최적의 대역폭 행렬이다.

3. 통계적 시뮬레이션

MKDE와 MKDE-ebd를 사용한 분포함수의 추정 정확도를 비교하기 위해서 표본 데이터로부터 분포함수를 추정하는 통계적 시뮬레이션을 수행하였다. 시뮬레이션에 앞서 4종류의 모집단을 정의하였다. Fig. 4는 정의된 실제 모델의 결합확률밀도함수의 등고선을 보여준다. 첫째, (a)는 X_1과 X_2의 주변확률분포가 모두 평균이 5이고 표준편차가 1인 정규분포를 가지고 X_1과 X_2가 가우시안 코퓰라(gaussian copula)를 따르며 등고선의 레벨은 0.02부터 0.22까지 0.02씩 증가하면서 나타내었다. 둘째, (b)의 주변확률분포는 (a)와 같지만 X_1과 X_2의 상관관계가 클레이튼 코퓰라(clayton copula)를 따르며 레벨은 0.05부터 0.2까지 0.05씩 증가시키면서 나타내었다. (a)와 (b) 모두 상관계수인 켄달 타우(kendall's tau, τ)는 0.5이다. 셋째, (c)는 3개의 모드(mode)를 가지는 가우시안 혼합모델(gaussian mixture model, GMM)이고, GMM의 평균은 각각 〔4 4.5〕, 〔6 8〕, 〔7.5 3.5〕이고 공분산은 모두 〔0.5 0; 0 0.5〕이며 레벨은 0.02부터 0.14까지 0.02씩 증가하면서 나타내었다. 넷째, (d)는 5개의 모드를 가지는 가우시안 혼합모델이고, 평균은 각각 〔4 8〕, 〔3.8 6〕, 〔5 5〕, 〔7 3.5〕, 〔9 2.5〕이고 공분산은 〔0.5 0; 0 0.8〕, 〔0.5 0; 0 0.5〕 〔0.8 0; 0 0.5〕, 〔0.9 0; 0 0.5〕, 〔0.3 0; 0 0.2〕이며 레벨은 0.01부터 0.09까지 0.01씩 증가시키면서 나타내었다.

통계적 시뮬레이션을 수행하기 위해서 정의된 모집단으로부터 데이터의 개수(n)를 5개에서 50개까지 증가시키면서 각 1000개의 표본 데이터 세트를 생성하였다. 그리고 각 표본에 대해서 MKDE와 MKDE-ebd를 사용해서 결합누적분포함수를 추정하고 실제 모집단의 결합누적분포함수와의 RRMSE를



(a) Gaussian model (b) Clayton model

(c) GMM model1 (d) GMM model2

Fig. 4 True models

계산하여서 분포함수의 추정 정확도를 평가하였다. MKDE-ebd를 수행하기 위해서는 입력 값으로 유의수준(α_{ebd})과 임계 RRMSE($RRMSE_c$)을 요구한다. α_{ebd}와 $RRMSE_c$는 경계데이터의 영역과 개수를 결정하는 주요인자 이므로 α_{ebd}를 0.05, 0.1, 0.2, $RRMSE_c$를 3, 5, 10%로 변화시켜 가면서 테스트를 하였고, 그 결과 $\alpha_{ebd}=0.1$과 $RRMSE_c=5\%$가 추정 정확도와 보수적인 특성을 가장 합리적으로 절충하였으므로 그 값을 본 논문의 통계적 시뮬레이션과 신뢰성해석에서 모두 사용하였다. α_{ebd}를 증가시키면 추정된 경계가 좁아져서 보수적인 경향이 감소되고, 반대의 경우 경계가 넓어져서 보수적이 경향이 증가한다. 또한 $RRMSE_c$의 값을 증가하면 경계데이터의 개수가 감소하여서 보수적인 경향이 감소되고, 반대의 경우 경계데이터의 개수가 증가하여 보수적이 경향이 증가한다.

1000개의 데이터 세트에 대한 정확도를 평가하기 위해서 상자그림(boxplot)을 사용해서 RRMSE 값의 산포 특성을 비교하였다. 상자그림은 산포를 가지는 데이터의 통계적 특성을 보여주는 대표적인 방법으로서 제1사분위수(1^{st} quartile, Q_1)와 제3사분위수(3^{rd} quartile, Q_3)를 사용해서 데이터의 분포를 개략적으로 표현해 준다(Tukey, 1977). 상자그림에서 상자의 하한-상한 수평선이 Q_1과 Q_3이고 중앙의 수평선은 중앙값(median or 2^{nd} quartile, Q_2)이다. 그리고 상자 아래-위로 연장된 점선은 이상치(outlier)가 아닌 데이터를 나타내며 이는 본 논문에서 대부분의 데이터(97~99%)를 포함하고 이 범위를 벗어난 점은 이상치이다. 이상치가 아닌 데이터의 범위를 계산하는 다양한 기준이 있으나 대부분 〔$Q_1-1.5\times IQR$, $Q_3+1.5\times IQR$〕을 기준으로 사용하고(Frigge *et al.*, 1989) 본 논문의 상자그림도 모두 같은 기준으로 표시되었다. 여기서 IQR은 사분위수범위(interquartile range)로서 Q_3-Q_1이다.

Fig. 5는 각 1000개의 표본 세트와 데이터의 개수에 따른 RRMSE 값의 분포를 나타낸 것이다. Fig. 5(a)의 가우시안 모델을 보면 $n\le 10$인 경우에는 경계데이터의 효과로 결합확률밀도함수의 꼬리가 길고 두꺼워 지면서 MKDE-ebd가 MKDE보다 RRMSE값이 크지만 $n\ge 20$부터는 경계데이터가가 거의 생성되지 않거나 추정된 분포함수에 영향을 주지 못해서 두 기법의 RRMSE 값이 거의 같아진다. Fig. 5(b)의 클레이튼 모델의 경우에도 가우시안 모델처럼 $n\le 10$에서는 MKDE의 RRMSE값이 MKDE-ebd보다 작고 $n\ge 20$에서는 두 기법의 RRMSE값이 거의 같아진다. Fig. 5(c)의 GMM 모델1의 경우에는 $n\le 10$에서 RRMSE의 크기는 MKDE가 더욱 작지만 산포는 MKDE-ebd가 더욱 좁다. 왜냐하면 GMM 모델1의 경우 분포함수의 비선형성이 매우 크기 때문에 데이터의 품질에 매우 민감한 MKDE의 추정 정확도가 산포를 크게

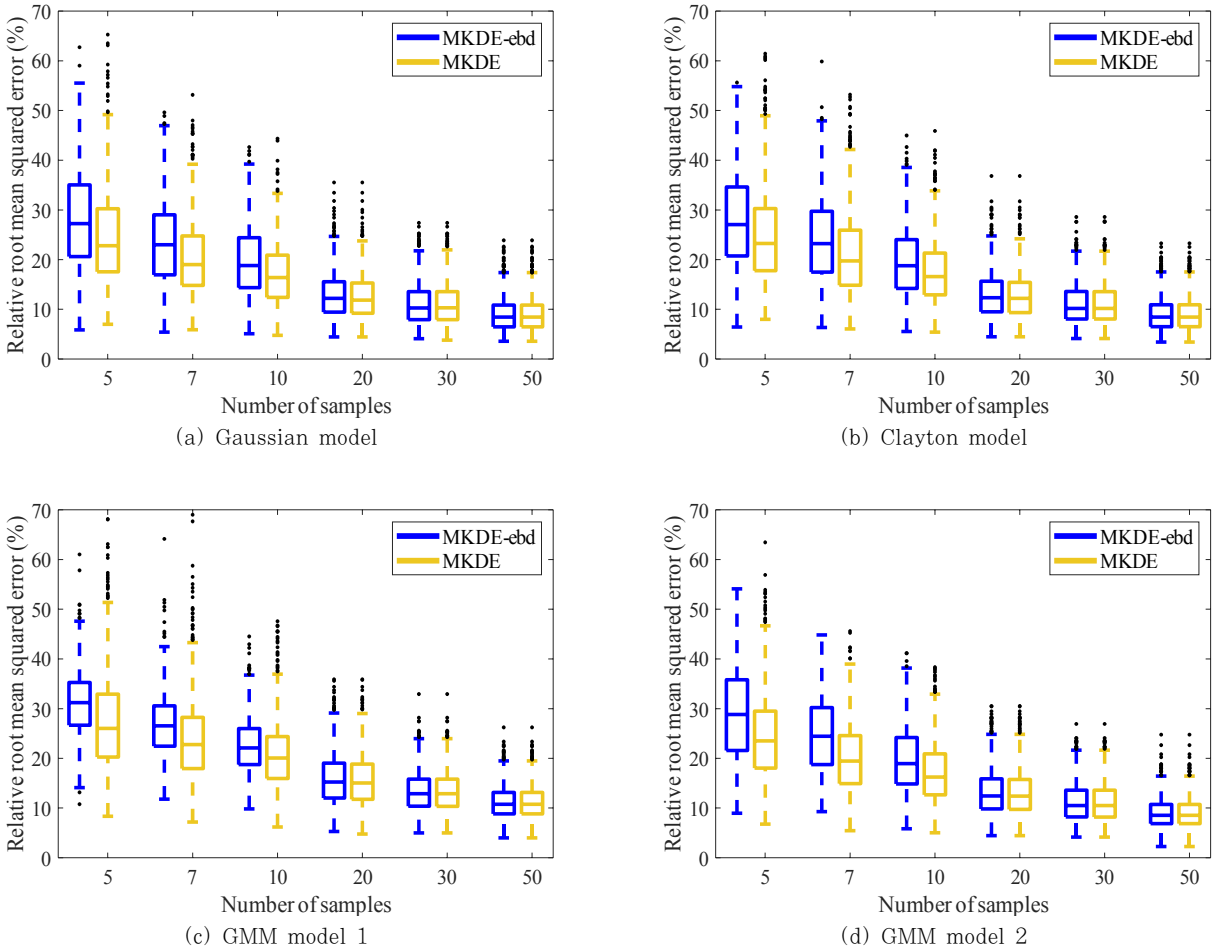


Fig. 5 RRMSE according to the number of samples

가지기 때문이다. 하지만 데이터의 개수가 증가하면서 데이터의 품질이 강건해지면 MKDE의 RRMSE 산포는 점차 감소하여 MKDE-ebd와 유사해진다. $n \geq 20$에서는 가우시안, 클레이튼 모델처럼 두 기법의 RRMSE값은 거의 같아진다. Fig. 5(d)의 GMM 모델2의 경우에는 GMM 모델1처럼 가우시안 혼합모델이지만 분포함수의 비선형성이 심하지 않기 때문에 가우시안, 클레이튼 모델과 유사한 결과를 보였다.

4. 신뢰성 해석 예제

MKDE와 MKDE-ebd를 사용해서 입력-확률변수의 확률분포함수를 모델링하고 모델링된 결합누적분포함수를 사용하여서 신뢰성 해석을 수행하고 두 기법이 파손확률 예측에 어떻게 영향을 주는지 비교하였다. 데이터의 개수와 품질에 의한 영향을 확인하기 위해서 정의된 모집단의 결함분포함수로부터 표본 데이터의 개수를 5개부터 50개까지 증가시켜 가면서 각 200개의 데이터 세트를 무작위로 생성하였다. 각 데이터 세트에 대해서 MKDE와 MKDE-ebd로 결합분포함수를 추정하고 추정된 분포로부터 10000개의 표본을 생성하여서 몬테카를로 시뮬레이션을 통해 파손확률을 예측하였다. 본 논문에서는 상관관계가 있는 입력변수를 포함한 외팔보와 2부재 트러스 문제의 파손확률을 예측하였다.

첫째, 외팔보 예제는 수평-수직하중을 동시에 받는 구조물의 처짐에 관한 신뢰성 해석 문제이다(Eldred *et al.*, 2007; Hong *et al.*, 2018). Fig. 6은 외팔보 구조물을 보여주고 Table 1은 외팔보의 입력변수를 나타낸다. 여기서 외팔보의 길이(L), 폭(w), 두께(t)는 결정론적 변수(deterministic variable)로서 각각 2.54m, 0.063m, 0.0889m이고 탄성계수(E), 수평하중(P_x), 수직하중(P_y)는 확률변수(random variable)로서 주변확률함수는 모두 정규분포이고, 수평-수직하중은 가우시안 코퓰라(τ=0.5)를 따르는 상관관계를 가진다. 외팔보의 처짐에 관한 성능함수는 다음과 같다.

$$g_D(\mathbf{x}) = \frac{4L^3}{Ewt}\sqrt{\left(\frac{P_x}{w^2}\right)^2 + \left(\frac{P_y}{t^2}\right)^2} - D_0 \tag{7}$$

여기서, 앞쪽 항은 외팔보의 최대 처짐이고 D_0는 외팔보 처짐의

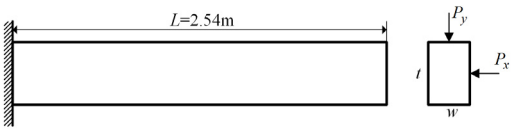


Fig. 6 Cantilever beam

Table 1 Input variables of cantilever beam

Variables	Symbol	Dist.	Value	CV
Length(m)	L	-	2.54	-
Width(m)	w	-	0.0635	-
Thickness(m)	t	-	0.0889	-
Young' modulus(GPa)	E	Normal	13.885	0.05
Horizontal load(N)	P_X	Normal	2224.11	0.2
Vertical load(N)	P_Y	Noraml	4448.22	0.1

(P_X and P_Y are correlated as Gaussian copula with τ=0.5)

허용량(displacement tolerance)으로 본 논문에서 0.0546m로 선정하였다. 외팔보의 최대 처짐이 D_0보다 크게 되면 파손이 발생하므로 파손확률은 다음과 같이 정의된다.

$$P_F = P[g_D(\mathbf{x}) > 0] \tag{8}$$

Fig. 7은 데이터의 개수에 따른 두 기법을 사용하여서 예측된 파손확률의 상자그림을 보여준다. Fig. 7에서 P_F^{Exact}(일점쇄선)는 모집단의 결합분포함수를 사용해서 계산된 실제 파손확률이고, P-box(파선)는 1000개의 표본 데이터에 대해 확률경계접근법(probability bounds approach, p-box theory)을 적용하여서 계산된 하한(lower)-상한(upper) 파손확률이다(Verma *et al.*, 2010). P_F^{Exact}와 P-box의 하한-상한 값은 각각 10.38, 8.52, 12.45%로 예측되었고, P_F^{Exact}와 P-box 값은 데이터 개수의 증가에 따라서 예측된 파손확률의 수렴성을 보기위해서 사용되었다. 즉 예측된 파손확률이 P_F^{Exact}에 가깝고 상하한 P-box 사이로 수렴할수록 정확한 파손확률을 예측하는 것이다.

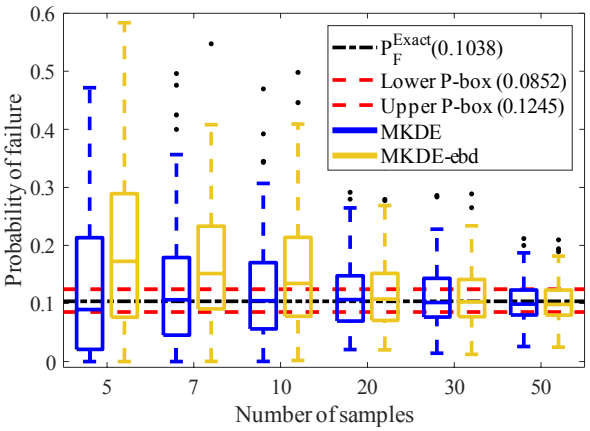


Fig. 7 Probability of failure according to the number of samples for the cantilever beam

MKDE에 의한 파손확률의 중앙값은 데이터의 개수에 상관없이 P_F^{Exact}에 가깝고 파손확률의 산포는 데이터의 개수가 증가함에 따라서 감소하였다. MKDE-ebd의 산포도 MKDE와 같이 데이터의 개수가 증가함에 따라서 감소하지만 $n \leq 10$에서 중앙값은 MKDE보다 높은 파손확률을 산출하면서 P_F^{Exact}에서 좀 더 멀리 위치하였다. 중앙값만을 보면 MKDE가 더욱 정확해 보이지만 $n \leq 10$에서 MKDE는 P_F^{Exact}보다 작은 파손확률을 예측하는 과소추정이 50% 이상이며 심지어 하한 P-box의 값보다 작은 값을 예측하는 비율이 매우 높은 것을 볼 수 있다. MKDE-ebd는 데이터의 개수가 적을 때는 P_F^{Exact}보다 파손확률을 크게 예측하며 보수적 추정을 하다가 데이터의 개수가 증가하면서 보수적인 경향은 점차 감소한다. 결론적으로 MKDE는 데이터의 개수가 적은 경우에는 과소추정의 특성이 매우 심해서 과소설계를 유발할 수 있지만 MKDE-ebd는 정보가 적은 경우에는 보수적인 추정을 하고 정보가 증가함에 따라서 정확한 추정을 하는 보다 안전한 통계모델링 방법임을 확인할 수 있다.

둘째, 2부재 트러스 문제는 두 부재가 회전 조인트로 결합된 시스템 구조물이고 조인트에 수직-수평하중을 함께 받고 있다(Park *et al.*, 2015; Hong *et al.*, 2018). Fig. 8은 2부재 트러스 문제를 보여주고 Table 2는 2부재 트러스 문제의 입력변수를 보여준다. 여기서 각도(α), 부재 1과 2의 면적(A_1, A_2)은 결정론적 변수로서 각각 45도, 0.35m^2, 1m^{2}이고 부재

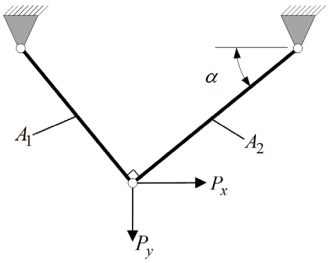


Fig. 8 Two-member truss

Table 2 Input variables of two-member truss

Variables	Symbol	Dist.	Value	CV
Angle(°)	α	-	45	-
Area 1(m^2)	A_1	-	0.35	-
Area 2(m^2)	A_2	-	1	-
Ultimate strength 1(MPa)	σ_{u1}	Normal	250	0.05
Ultimate strength 2(MPa)	σ_{u2}	Normal	250	0.05
Horizontal load(kN)	P_X	Normal	50	0.05
Vertical load(kN)	P_Y	Normal	50	0.3

(P_X and P_Y are correlated as Clayton copula with τ=0.5)

1과 2의 극한강도(σ_{u_1}, σ_{u_2}), 수평하중(P_x), 수직하중(P_y)은 확률변수로서 주변확률함수는 정규분포이며, 수평-수직하중은 클레이튼 코퓰라(τ=0.5)를 따르는 상관관계를 가진다. 부재 1, 2의 성능함수는 각각 다음과 같다.

$$g_1(\mathbf{x}_1)=\frac{1}{2}\left(\frac{P_y}{\cos\alpha}+\frac{P_x}{\sin\alpha}\right)-A_1\sigma_{u_1} \tag{9}$$

$$g_2(\mathbf{x}_2)=\frac{1}{2}\left(\frac{P_y}{\cos\alpha}-\frac{P_x}{\sin\alpha}\right)-A_2\sigma_{u_2} \tag{10}$$

여기서, 각 부재는 각 성능함수가 0보다 큰 경우 파괴되고 2부재 트러스는 직렬형 시스템(series system)이므로 파손확률은 다음과 같다.

$$P_F^{SYS}=P[g_1(\mathbf{x}_1)>0]+P[g_2(\mathbf{x}_2)>0] - P[g_1(\mathbf{x}_1)>0]\cdot P[g_2(\mathbf{x}_2)>0] \tag{11}$$

Fig. 9는 데이터의 개수에 따른 파손확률을 보여주며 P_F^{Exact}와 P-box의 하한-상한 값은 각각 9.61, 6.22, 11.6%로 외팔보 문제에 비해서 약간 낮은 파손확률을 보였다. 외팔보처럼 MKDE는 과소 추정하는 경향이 있고 MKDE-ebd는 보수적인 추정을 하면서, 데이터의 개수가 증가하면서 실제 파손확률에 수렴하였다. 하지만 2부재 트러스의 경우 데이터의 개수 증가에 따른 두 기법의 수렴속도가 외팔보 문제보다 느렸으며 MKDE-ebd는 좀 더 보수적인 경향이 심하였다. 그 이유는 2부재 트러스 문제는 수직하중의 변동계수(coefficient of variation, CV)가 30%로서 상당히 큰 편이기 때문에 두 기법 모두 데이터 개수의 증가에 따른 분포함수 추정 정확도의 증가률이 상대적으로 낮기 때문에 예측된 파손확률의 수렴속도도 감소하게 된다.

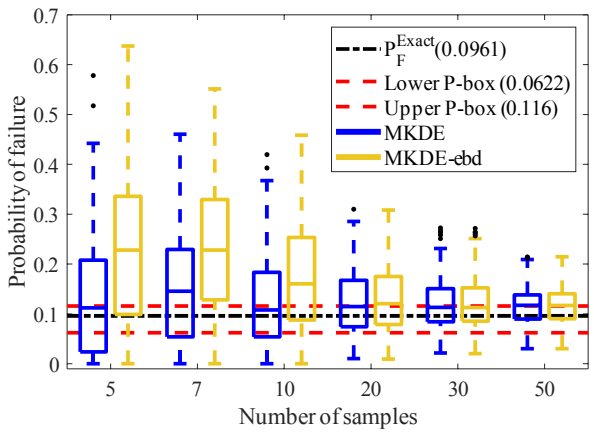


Fig. 9 Probability of failure according to the number of samples for the two-member truss

또한 MKDE-ebd는 데이터의 산포가 큰 경우에는 경계데이터의 영역을 넓게 선정하기 때문에 보수적인 경향이 더욱 두드러지게 된다(Kang *et al.*, 2018a). 2부재 트러스 문제는 외팔보와 수렴속도와 보수적 추정의 정도에서 약간의 차이가 있었지만 여전히 데이터의 개수가 적은 경우 MKDE는 과소추정이 매우 심하였지만 MKDE-ebd는 보수적인 추정을 하다가 점차적으로 실제 파손확률에 수렴하였다.

5. 결 론

본 논문은 결합분포함수의 통계모델링을 위한 경계데이터를 이용한 다변량 커널 밀도 추정(MKDE-ebd) 방법을 제안하였다. 기존의 MKDE는 데이터 수와 질에 민감한 특징을 가지는 것에 반해 본 논문에서 제안한 MKDE-ebd는 주어진 데이터와 경계데이터를 결합하여 결합분포함수의 꼬리부분까지 모델링함으로써 신뢰성 해석에서 파손확률 추정 시 MKDE 보다 더욱 보수적인 결과를 도출할 수 있으므로 더욱 안정적이고 신뢰도 높은 신뢰성 해석이 가능하다. MKDE-ebd 방법은 copula를 이용한 모수적 결합분포 모델링 방법에 비해 비선형성을 가진 다양한 결합분포 모델링이 가능하며 적은 수의 데이터만으로도 보수적인 모델링 결과를 도출할 수 있고, copula에 비해 2차원 이상으로 확장성이 크다는 장점을 가지고 있으므로 공학문제에서 다양한 형태의 결합분포함수 모델링에 사용가능하다.

본 연구에서는 신뢰도 높은 신뢰성 해석 결과 도출을 위한 다변량 통계모델링 방법을 제안하였으나, 추후 제안된 기법을 보완하여 데이터 수가 증가할 경우 정확도를 더욱 높일 수 있는 다변량 통계모델링 기법에 대한 연구를 수행할 예정이다.

감사의 글

이 논문은 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

References

Noh, Y., Choi, K.K., Lee, I. (2010) Identification of Marginal and Joint CDFs using Baeysian Method for RBDO, *Struct. Multidisc. Optim.*, 40(1), pp.35~51.

Chen, S. (2015) Optimal Bandwidth Selection for Kernel Density Functionals Estimation, *J. Probab. Stat.*, pp.1~21.

Eldred, M.S., Agarwal, H., Perez, V.M., Wojtkiewicz Jr, S.F., Renaud, J.E. (2007) Investigation of Reliability Method Formulations in DAKOTA/UQ,

Struct. & Infrastruct. Eng., 3(3), pp.199~213.

Frigge, M., Hoaglin, D.C., Iglewicz, B. (1989) Some Implementations of the Boxplot, *The American Statistician*, 43(1), pp.50~54.

Hong, J., Kang, Y.J., Lim, O.K., Noh, Y. (2018) Comparison of Multivariate Statistical Modeling Methods for Limited Correlated Data, *Trans. Korean Soc. Mech. Eng. A*, 42(5), pp.445~453.

Jamieson, P.D., Porter, J.R., Wilson, D.R. (1991) A Test of the Computer Simulation Model ARCWHEAT1 on Wheat Crops Grown in New Zealand, *Field Crops Res.*, 27(4), pp.337~350.

Kang, Y.J., Hong J., Lim, O.K., Noh, Y. (2017) Reliability Analysis Using Parametric and Nonparametric Input Modeling Methods, *J. Comput. Struct. Eng. Inst. Korea*, 30(1), pp.87~94.

Kang, Y.J., Noh, Y., Lim, O.K. (2018a) Kernel Density Estimation with Bounded Data, *Struct. Multidisc. Optim.*, 57(1), pp.95~113.

Kang, Y.J. (2018b) *Development of Integrated Statistical Modeling Method for Reliability Analysis*, Ph.D. Dissertation, Pusan National University.

Li, M.F., Tang, X.P., Wu, W., Liu, H.B. (2013) General Models for Estimating Daily Global Solar Radiation for Different Solar Radiation Zones in Mainland China, *Energy Convers. & Manag.*, 70, pp.139~148.

Park, C., Kim, N.H., Haftka, R.T. (2015) The Effect of Ignoring Dependence Between Failure Modes on Evaluating System Reliability, *Struct. Multidisc. Optim.*, 52(2), pp.251~268.

Scott, D.W. (2015) *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New Jersey, p.416.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, 26, CRC Press, London.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Pearson, New York, p.23.

Verma, A.K, Srividya, A, Karanki, D.R. (2010) *Reliability and Safety Engineering*, Springer, London, p.571.

요 지

공학문제에서 많은 확률 변수들은 상관성을 가지고 있고, 입력변수의 상관성은 기계시스템의 통계적 성능 분석 결과에 큰 영향을 미친다. 하지만, 상관 변수들은 결합분포함수를 모델링하기 어렵다는 이유로 종종 독립변수로 취급되거나 특정한 모수적 모델로 표현되는 경우가 많으며, 특히 데이터가 적은 경우 결합분포함수를 정확히 모델링하는데 더 큰 어려움이 있다. 본 연구에서 개발된 경계데이터를 이용한 다변량 커널밀도추정은 비선형성을 갖는 다양한 형태의 다변량 확률 분포 추정을 위해 개발되었다. 다변량 커널밀도추정은 주어진 데이터와 균등분포함수의 파라미터의 신뢰구간으로부터 생성된 경계데이터를 결합하여 데이터의 질과 수에 덜 민감하다. 따라서 제안된 방법은 보수적인 통계모델링과 신뢰성 해석 결과를 도출할 수 있으며, 통계시뮬레이션과 공학예제를 통해 그 성능을 검증하였다.

핵심용어 : 경계데이터를 이용한 다변량 커널밀도추정, 다변량 커널밀도추정, 비모수적 통계모델링, 상관 데이터, 상대 평균제곱근 오차, 신뢰성 해석