J. Comput. Struct. Eng. Inst. Korea, 32(3)
pp.173~181, June, 2019
https://doi.org/10.7734/COSEIK.2019.32.3.173

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

다양한 대역폭 선택법에 따른 커널밀도추정의 비교 연구

강 영 진[1] · 노 유 정[2†]

[1]부산대학교 기계기술연구원, [2]부산대학교 기계공학부

Comparison Study of Kernel Density Estimation according to Various Bandwidth Selectors

Young-Jin Kang[1] and Yoojeong Noh[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

To estimate probabilistic distribution function from experimental data, kernel density estimation(KDE) is mostly used in cases when data is insufficient. The estimated distribution using KDE depends on bandwidth selectors that smoothen or overfit a kernel estimator to experimental data. In this study, various bandwidth selectors such as the Silverman's rule of thumb, rule using adaptive estimates, and oversmoothing rule, were compared for accuracy and conservativeness. For this, statistical simulations were carried out using assumed true models including unimodal and multimodal distributions, and, accuracies and conservativeness of estimating distribution functions were compared according to various data. In addition, it was verified how the estimated distributions using KDE with different bandwidth selectors affect reliability analysis results through simple reliability examples.

Keywords : *bandwidth selector, kernel density estimation, multimodal distribution, reliability analysis, statistical modeling, unimodal distribution*

1. 서 론

신뢰성 해석(reliability analysis)과 신뢰성 기반 최적설계(reliability-based design optimization, RBDO) 등의 확률-통계적 해석 및 설계 방법은 분포함수를 가지는 확률변수(random variable)를 입력 값으로 요구한다. 일반적으로 입력변수의 분포를 알지 못하기 때문에 정규분포로 가정해서 해석 및 설계를 수행하지만, 정확한 결과를 산출하기 위해서는 실험 데이터로부터 분포함수를 추정하는 통계모델링(statistical modeling)의 과정이 필요하다. 그렇지만 분포함수의 추정 정확도는 데이터의 개수에 매우 민감하므로 데이터의 개수에 따라 사용하기 적합한 추정 기법들이 다르게 된다.

공학 분야의 변수들은 실험 데이터의 개수가 매우 적기 때문에 분포함수를 추정하기 위해서 모수적인 방법(parametric method)보다 정확하고 신뢰성이 높은 비모수적인 방법(non-parametric method)이 적합하고(Kang *et al.*, 2017) 그 중에서 커널밀도추정(kerenel density estimation, KDE)이 많이 사용되고 있다(Zhang *et al.*, 2014; Jang *et al.*, 2015). KDE에서 분포함수의 추정 정확도에 가장 중요한 인자는 커널함수(kernel function)의 대역폭(bandwidth)이고(Silverman 1986; Wand *et al.*, 1994; Chen 2015), 다양한 대역폭 선택법(bandwidth selector)이 있다. 대역폭 선택법에 따라서 최적의 대역폭이 다르게 계산되고 대역폭이 작을수록 분포함수는 과대적합(overfitting)을 하고 클수록 평활(smoothing)한 분포함수를 추정한다. 대역폭 선택법에 따른 비교 연구가 수행되었지만 대부분 통계적 관점이어서 표본의 개수가 100개 이상이고 표본의 무작위성과 신뢰성 해석의 관점에서 비교 연구는 부족하다(Terrell *et al.*, 1985;

† Corresponding author:
Tel: +82-51-510-2308; E-mail: yoonoh@pusan.ac.kr
Received March 11 2019; Revised April 3 2019;
Accepted May 10 2019

Silverman 1986; Scott 2010). 다양한 방법 중에서 정규분포 기준 법칙(normal reference rule) 중 하나인 Silverman's rule of thumb가 가장 대표적으로 사용되고 있다(Scott 2010; Zhang *et al.*, 2014; Jang *et al.*, 2015).

본 연구에서는 Silverman's rule of thumb와 적응형 추정치(adaptive estimate)를 적용한 rule using adaptive estimate, 그리고 상대적으로 보수적인 추정을 하는 oversmoothing rule의 세 가지 방법을 사용한 KDE의 통계모델링의 결과를 비교하고 분석하였다. 분포함수의 추정 정확도의 특성을 확인하기 위해서 단봉분포(unimodal distribution)와 다봉분포(multimodal distribution)에 대한 통계적 시뮬레이션 후 추정 정확도를 비교하고, 실제 측정 데이터에 대한 분포함수 추정을 통한 추정 정확도 비교와 신뢰성 해석 예제를 통해 각 기법이 신뢰성 해석 결과에 어떻게 영향을 미치는지 비교하고 분석하였다. 최종적으로 통계적 시뮬레이션과 신뢰성 해석을 결과를 토대로 확률변수의 다봉성(multimodality), 분포함수 추정의 정확성과 보수성에 대한 조건에서 적합한 대역폭 선택법을 추천하였다.

2. 커널밀도추정

커널밀도추정(kernel density estimation, KDE)은 데이터만을 이용해서 확률밀도함수(probability density function, PDF)를 추정하는 비모수적 통계모델링(nonparametric statistical modeling)으로서 추정된 커널 추정치(kernel estimate)는 다음과 같다(Silverman, 1986).

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) \tag{1}$$

여기서, x는 확률변수이고 X_i는 데이터, n은 데이터의 개수이다. 그리고 $K(\bullet)$는 커널함수이고 h는 커널함수의 대역폭 또는 평활 매개변수(smoothing parameter)이다. 커널 추정치는 대역폭이 작을수록 데이터에 대해서 과대적합(overfitting)을 하여서 데이터를 잘 따르지만 PDF의 비선형성이 증가하고 꼬리가 짧게 된다. 반면에 대역폭이 클수록 데이터에 대해서 평활(smoothing)해져서 데이터보다 넓은 산포를 가지므로 PDF는 매끄럽고 긴 꼬리를 가지게 된다. KDE의 정확도는 커널함수의 종류와 대역폭의 크기에 의해서 결정되며, 특히 대역폭의 크기가 매우 중요하다(Silverman 1986; Wand *et al.*, 1994; Chen 2015). 그러므로 최적의 대역폭을 선정하는 과정은 KDE에서 추정 정확도에 큰 영향을 준다. 그래서 본 연구에서는 세 가지 대역폭 선정 방법을 KDE에 적용해서 그 결과를 비교하였다.

2.1 Silverman's rule of thumb

최적의 대역폭을 선정하기 위해서 가장 대표적으로 사용되는 방법이 정규분포 기준 법칙으로서 모집단이 정규분포라고 가정하고 모집단과 추정된 PDF의 평균적분제곱오차(mean integrated squared error, MISE)를 최소화하는 대역폭을 선정하는 방법이다. 커널 추정치의 MISE는 다음과 같다.

$$\begin{aligned}\mathrm{MISE}(\hat{f}(x)) &= \int E\left[(\hat{f}(x)-f(x))^2\right]dx \\ &= \int \{E[\hat{f}(x)]-f(x)\}^2dx+\int Var[\hat{f}(x)]dx \\ &= \int Bias[\hat{f}(x)]^2dx+\int Var[\hat{f}(x)]dx\end{aligned} \tag{2}$$

여기서, $\hat{f}(x)$는 커널 추정치이고 $f(x)$는 모집단의 확률밀도함수이다. 그리고 $E[\bullet]$는 기댓값(expectation)이고 $Var[\bullet]$은 분산(variance)이며 $Bias[\bullet]$는 편향(bias)이다.

MISE를 직접 최소화하기는 어려우므로 커널 추정치의 기댓값 $E[\hat{f}(x)]$를 테일러 급수 전개(taylor's series expansion) 시킨 후 나머지 항을 제외하여 식 (1)을 재구성한 점근 평균적분제곱오차(asymptotic mean integrated squared error, AMISE)를 최소화하는 대역폭을 선정하는 방식으로 최적의 대역폭을 유도한다. 2차 커널함수에 대해서 AMISE를 구하면 다음과 같다(Silverman, 1986).

$$\mathrm{AMISE}(\hat{f}(x))=\frac{1}{4}h^4k_2^2\int f''(x)^2dx+n^{-1}h^{-1}\int K(t)^2dt \tag{3}$$

여기서, k_2은 커널함수의 분산이다. 식 (3)을 대역폭 h에 대해서 미분하여 $\frac{d}{dh}$AMISE가 0이 되는 h를 계산하면 최적의 대역폭 h^*가 계산된다(Silverman, 1986).

$$h^*=k_2^{-2/5}\left\{\int K(t)^2dt\right\}^{1/5}\left\{\int f''(x)^2dx\right\}^{-1/5}n^{-1/5} \tag{4}$$

Silverman's rule of thumb는 KDE에서 가장 대표적으로 사용되는 최적 대역폭 선정 방법으로서, 정규분포 기준 법칙에서 커널함수를 가우시안 커널로(gaussian kernel)로 선택해서 유도된다. 모집단이 정규분포를 따른다고 가정하면 식 (4)의 $\int f''(x)^2dx$는 다음과 같이 되고(Silverman, 1986),

$$\int f''(x)^2dx=\hat{\sigma}^{-5}\int \phi''(x)^2dx=\frac{3}{8}\pi^{-1/2}\hat{\sigma}^{-5} \tag{5}$$

커널함수를 가우시안 커널로 선택하면 $k_2=1$, $\int K(t)^2dt = 1/2\sqrt{\pi}$ 이 되므로(Scott, 2015), 식 (4)는 다음과 같이 정리된다.

$$h^* = (2\sqrt{\pi})^{-1/5}\left\{\frac{3}{8}\pi^{-1/2}\right\}^{-1/5}\hat{\sigma}n^{-1/5} \tag{6}$$

식 (6)을 정리하면 다음과 같이 Silverman's rule of thumb을 사용한 최적의 대역폭이 다음과 같이 계산된다(Silverman, 1986; Scott 2010).

$$h_{SRT} = \left(\frac{4}{3}\right)^{1/5}\hat{\sigma}n^{-1/5} \approx 1.0592\hat{\sigma}n^{-1/5} \tag{7}$$

여기서, $\hat{\sigma}$은 표본의 표준편차이고 본 연구에서 표준편차의 강건한 추정치(robust estimate of sample's standard deviation)를 사용하였고 추정치는 다음과 같다(Analytical Methods Committee, 1989).

$$\hat{\sigma} = \frac{\text{Median}(|X_i - \text{Median}(X_i)|)}{0.6745} \tag{8}$$

2.2 Rule using adaptive estimate

왜도와 첨도가 큰 분포에서 정확한 대역폭을 얻기 위해서 식 (7)의 표준편차 $\hat{\sigma}$ 대신에 산포에 대한 적응형 추정치(adaptive estimate)를 사용하고 식 (7)의 계수를 감소시키면 식 (7)은 다음과 같이 변형된다(Silverman, 1986).

$$h_{AE} = 0.9An^{-1/5} \tag{9}$$

$$A = \min(\hat{\sigma}, IQR(x)/1.34) \tag{10}$$

여기서, A는 적응형 추정치이고 $IQR(x)$는 표본의 사분위범위(interquartile range)이다.

2.3 Oversmoothing rule

기존의 대역폭 선정 척도와 달리 Terrell과 Scott(1985), Terrell(1990)은 다른 척도를 사용해서 최대 대역폭 원리(maximal smoothing principle)를 제안하고, 이를 사용해서 유도된 대역폭 선정 방법이 oversmoothing rule이다. 식 (4)에서 $\int f''(x)^2dx$의 값을 구하기 위해서 다음과 같은 변분법 문제를 정의한다(Scott, 2015).

$$\min_f \int_{-\infty}^{\infty} f''(x)^2dx,\ s/t \int f = 1 \text{ and } \int x^2 f = 1 \tag{11}$$

식 (11)의 풀면(Terrell *et al.*, 1985; Terrell, 1990) 최솟값은 다음과 같다.

$$\int \left[\left\{f_X^*(x)\right\}''\right]^2 dx = \frac{35}{243\sigma^5} \tag{12}$$

그리고 $\int \left[\left\{f^*(x)\right\}''\right]^2 dx \le \int f''(x)^2dx$의 관계식을 사용하면(Terrell *et al.*, 1985; Terrell, 1990), 식 (12)는 다음과 같이 대역폭의 범위를 가지게 되고,

$$\int \left[\{f(x)\}''\right]^2 dx \ge \frac{35}{243\sigma^5} \tag{13}$$

식 (13)을 식 (4)에 대입하여 정리하면 최대 대역폭 원리에 따른 최적 대역폭의 범위가 정의된다.

$$h^* = \left[\frac{\int K(t)^2dt}{nk_2^2\int f''(x)^2dx}\right]^{1/5} \le \left[\frac{243\sigma^5\int K(t)^2dt}{35nk_2^2}\right]^{1/5} \tag{14}$$

식 (14)에서 상한 값을 대역폭으로 선정하면 Terrell과 Scott이 제안한 oversmoothing rule이고 가우시안 커널을 적용한 oversmoothing rule은 다음과 같다(Scott, 2015).

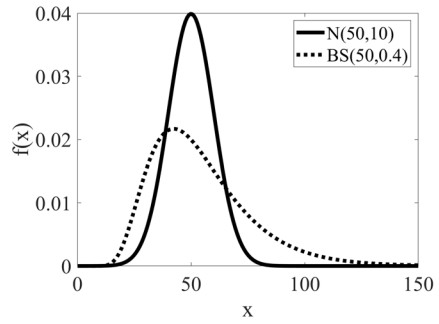
$$h_{OS} = 1.144\hat{\sigma}n^{-1/5} \tag{15}$$

세 가지 선정법에 의한 최적의 대역폭은 공통적으로 표준편차($\hat{\sigma}$)와 데이터의 개수(n)에 따라 변하게 된다. 본 연구에서 대역폭 계산 시, 표준편차는 모두 식 (8)의 표본의 표준편차 계산식을 사용하여서 계산되었고 데이터의 개수는 모두 같으므로 표준편차와 개수와 관한 항의 앞의 계수에 따라서 대역폭이 서로 다르게 된다. 식 (7), (9), (15)을 보면 각 기법의 계수 중에서 rule using adaptive estimate가 가장 작고 oversmoothing rule이 가장 크므로 대역폭의 크기는 rule using adaptive estimate, Silverman's rule of thumb, oversmoothing rule 순으로 커진다.

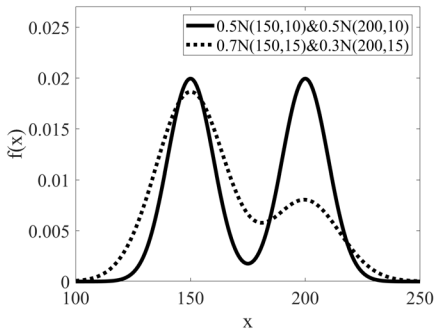
3. 통계적 시뮬레이션

3.1 실제모델 정의와 표본 생성

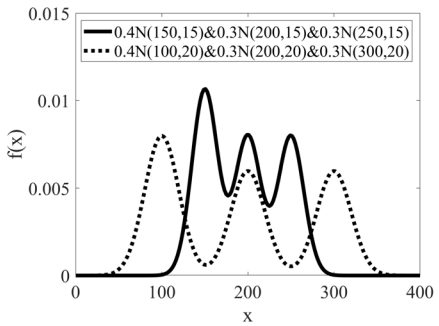
다양한 대역폭 선정 방법과 그 방법에 따른 커널 추정치의 정확도를 비교하기 위해서 2종류의 단봉분포(unimodal distribution), 2종류의 이봉분포(bimodal distribution), 2종류의 삼봉분포(trimodal distribution)를 실제모델로 가정하고 그 모델들로부터 데이터 개수(n)를 5 부터 50까지 증가시켜가면서 표본을 임의로 1,000세트 씩 추출하였다. 그리고 각 데이터 세트에 대해서 세 가지 방법을 사용하여서 PDF를 추정하고 면적척도인 교차면적(intersection area, IA)를 사용하여서 추정된 PDF와 실제 PDF의 일치성을 비교하였다. 교차면적은 두 PDF의 일치성을 나타내는 것으로 0에서 1사이의 값을 가지며 그 값이 0 이면 두 PDF는 완전치 불일치이고 1이면 완전치 일치하는 것을 의미한다(Jung *et al.*, 2017). Fig. 1은 사용된 실제 모델의 PDF를 보여준다. Fig. 1(a)는 대칭인 정규분포(N(50,10))와 비대칭성이 매우 큰 Birnbaum-Saunders 분포(BS(50,0.4)의 PDF를 보여주고, (b)는 이봉형태가 약한 0.7N(150,15) & 0.3N(200,15) 분포와 이봉형태가 매우 심한 0.5N(150,10) & 0.5N(200,10) 분포이고, (c)는 삼봉형태가 약한 0.4N(150,15) & 0.3N(200,15) & 0.3N(250,15)분포와 삼봉형태가 심한 0.4N(100,20) & 0.3N(200,20) & 0.3N(300,20)의 분포를 보여준다.



(a) Unimodal distributions



(b) Bimodal distributions

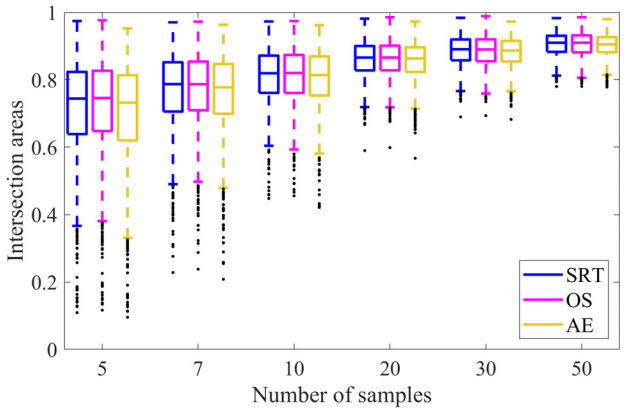


(c) Trimodal distributions

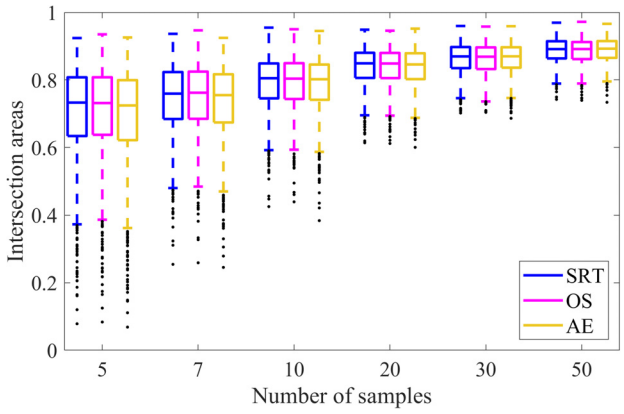
Fig. 1 PDFs of true models

3.2 분포함수의 추정 정확도 비교

생성된 표본에 대해서 3가지 최적 대역폭 선정 방법을 사용하여서 커널밀도함수를 추정하였고, 각 데이터의 개수에서 1,000세트의 표본의 결과를 비교하기 위해서 결과를 상자그림(boxplot)으로 나타내고, 실제모델과 데이터의 개수에 따라 비교하였다. 상자그림은 데이터의 분포를 표현하는 대표적인 방법으로서 상자에서 하단, 중단, 상단의 수평선은 각각 제1사분위수(1st quartile), 제2사분위수(2nd quartile), 제3사분위수(3rd quartile)이다. 그리고 수직선과 점 표시는 전체 데이터의 분포를 나타낸다(Tukey, 1977). Fig. 2~4는 실제모델에 따른 추정된 커널밀도함수와 실제 모델의 PDF의 교차면적을 나타낸 것이다. 여기서 SRT는 Silverman's rule of

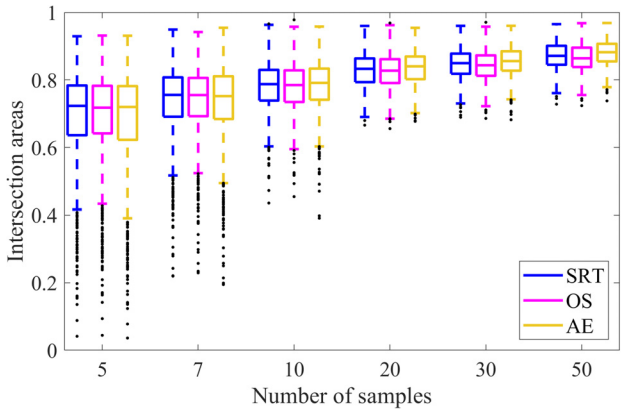


(a) X~N(50,10)

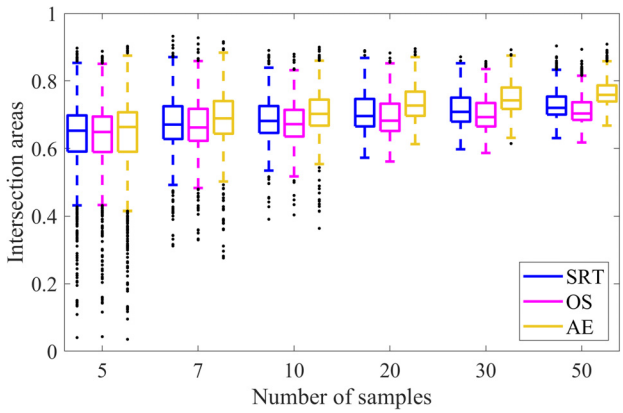


(b) X~BS(50,0.4)

Fig. 2 Intersection areas for unimodal distributions

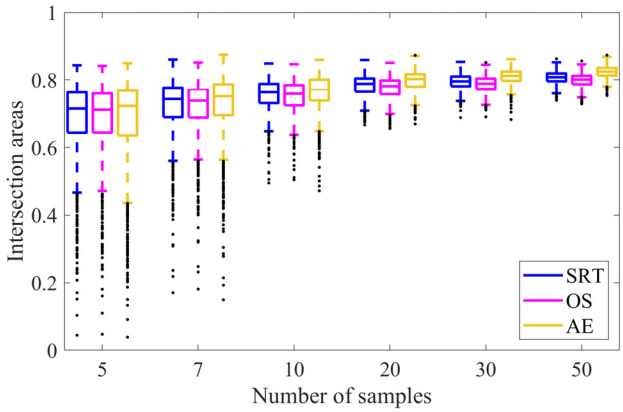


(a) X~0.7N(150,15) & 0.3N(200,15)

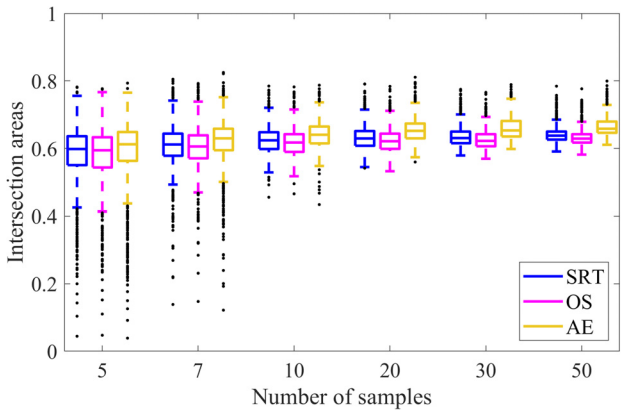


(b) X~0.5N(150,10) & 0.5N(200,10)

Fig. 3 Intersection areas for bimodal distributions



(a) X~0.4N(150,15) & 0.3N(200,15) & 0.3N(250,15)



(b) X~0.4N(100,20) & 0.3N(200,20) & 0.3N(300,20)

Fig. 4 Intersection areas for trimodal distributions

thumb, OS는 Oversmoothing rule, AE는 rule using adaptive estimate를 의미한다.

Fig. 2는 실제분포가 대칭 또는 비대칭 단봉분포를 가질 때 분포함수의 추정에 의한 교차면적을 나타낸 것이다. $n \leq 10$인 경우에는 OS의 정확도가 가장 높고 AE가 정확도가 가장 낮지만 차이가 거의 없고, $n \geq 20$에서는 세 가지 방법의 차이가 거의 없다. 단봉분포는 PDF의 비선형성이 심하지 않기 때문에 추정된 PDF가 대역폭의 변화에 둔감하므로 세 가지 방법에 의한 교차면적의 차이는 작게 된다.

Fig. 3은 이봉분포에 대한 분포함수의 추정 정확도를 나타낸 것이다. 여기서는 (a)는 실제 이봉분포의 형태가 약한 분포이고 (b)는 심한 분포이다. (a)를 보면, AE의 정확도가 가장 높고 OS의 정확도가 가장 낮으며, 데이터의 개수가 증가함에 따라서 세 가지 방법의 차이는 뚜렷해진다. 세 가지 방법의 차이가 단봉분포에 비해서는 분명하지만 분포함수의 추정의 정확도의 차이가 그다지 크지 않았다. (b)를 보면, 전체적인 경향은 (a)와 유사하지만 전체적인 교차면적이 (a)에 비해 매우 낮고, 세 가지 기법에 따른 교차면적의 차이가 커진 것을 볼 수 있다. (b)의 경우 실제분포의 이봉형태가 매우 심하므로 이봉성(bimodality)을 정확하게 표현하기 위해서는 더욱 많은 데이터를 필요로 하므로 같은 데이터의 개수에서는 단봉분포 또는 이봉성이 약한 분포에 비해 분포함수의 추정 정확도가 낮아지게 된다. 하지만 이봉성이 증가하면서 세 가지 기법에 따른 커널밀도함수의 추정 정확도의 차이는 증가하게 된다. 실제 모델이 이봉분포일지라도 데이터의 개수가 적거나 이봉성이 약하면 단봉분포처럼 세 가지 선정법에 따른 교차면적의 차이는 작지만, 데이터의 개수가 증가하거나 이봉성이 커지면 세 방법에 따른 교차면적의 차이가 뚜렷해진다.

Fig. 4는 삼봉분포에 대한 분포함수의 추정 정확도를 나타낸 것이다. 여기서 (a)는 실제 삼봉분포의 형태가 약해서 PDF의 전체적인 형태는 단봉분포와 유사한 분포이고 (b)는 PDF의 형태가 완전한 삼봉분포를 보인다. 삼봉분포의 형태가 약하고 심한 경우 모두 이봉분포의 결과처럼 AE의 정확도가 가장 높고 OS의 정확도가 가장 낮으며, n이 증가하면서 세 가지 기법의 차이는 커진다.

세 가지 대역폭 선정 방법에 따른 통계적 시뮬레이션의 결과를 정리하면, 단봉분포의 경우 OS가 가장 높은 분포함수의 추정 정확도를 보이고 다봉성이 증가할수록 AE의 정확도가 가장

높았다. OS는 세 가지 방법 중에서 가장 대역폭을 넓게 선정하여서 상대적으로 비선형성이 가장 낮고 꼬리가 긴 PDF를 추정한다. 이러한 특성은 단봉분포에서는 데이터의 개수와 품질에 강건하게 하는 효과를 유도하여서 정확도를 높이지만, 다봉분포에서는 PDF의 비선형성을 잘 표현하지 못해서 정확도를 낮추는 효과를 유발한다. AE는 OS와 반대로 가장 대역폭을 좁게 선정하여서 비선형성이 가장 심하고 짧은 꼬리를 가지는 PDF를 추정한다. 그러므로 AE는 다봉성이 증가할수록 다른 방법에 비해서 추정 정확도가 빠르게 증가하게 된다. 세 가지 방법 모두 Fig. 2(a)와 (b)의 대칭, 비대칭 단봉분포, Fig. 3(a)와 Fig. 4(a)의 다봉성인 약한 다봉분포에서는 높은 분포함수 추정 정확도를 보였지만, 세 기법모두 다봉분포에서는 PDF의 평활 정도가 데이터의 분포보다 커서 Fig. 3(b)와 Fig. 4(b)의 다봉분포에서는 상대적으로 낮은 정확도를 보였다.

4. 수치 예제

세 가지 대역폭 선정법을 사용해서 추정된 커널밀도함수를 사용함에 따른 결과를 비교하기 위해서 이봉분포을 가지는 온도 데이터에 대해 분포함수를 추정하여 추정 정확도를 비교하고, 입력변수의 모델링에 따른 출력 값의 비교하기 위해서 두 종류의 신뢰성 해석 예제에서 입력변수의 분포함수를 추정하고 추정된 분포함수를 성능함수에 사용하여서 파손확률(probability of failure)을 예측하여 비교하였다.

4.1 온도의 분포함수 추정

실제 데이터의 분포함수 추정의 정확도를 비교하기 위해서 시스템의 재료물성 또는 작업환경의 중요 외부인자인 온도 데이터에 대해서 시뮬레이션을 수행하였다(Lee *et al.*, 2011). 사용된 데이터는 서울을 2007년도 시간별 온도 데이터이고 모집단은 총 8,760개의 데이터로 구성된다(KMA, 2019). 분포함수 추정을 위해서 표본 5~50개를 무작위로 모집단으로부터 추출하였고, 각 표본의 개수에서 1,000 세트의 표본을 생성하였다. Fig. 5는 모집단의 히스토그램으로 이봉분포을 가지는 것을 확인할 수 있다(Lee *et al.*, 2011). 분포함수의 정확도를 비교하기 위해서 각 표본 집단에 세 가지 대역폭 선정법을 사용해서 분포함수를 추정하고 모집단과 교차면적을 계산하여 Fig. 6에 나타내었다. 온도 분포의 추정 정확도는 Fig. 3의 이봉분포 시뮬레이션 처럼 AE의 정확도가 가장 높고 OS의 정확도가 가장 낮으며, 표본의 크기가 증가함에 따라서 세 방법의 차이는 뚜렷해졌다. 하지만 Fig. 3과 달리 온도 데이터는 측정오차와 같은 불확실성을 최소화하기 위해서 측정되었고, 총

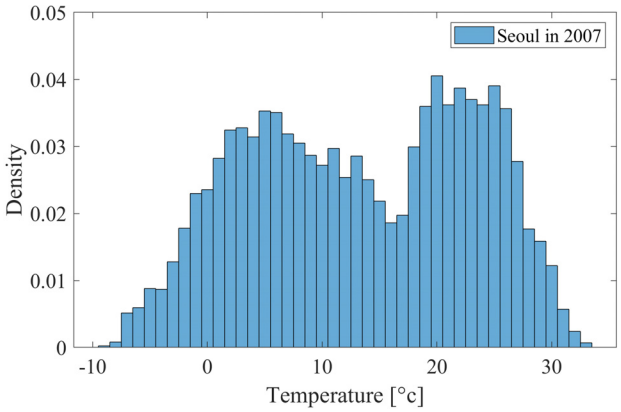


Fig. 5 Histogram of temperature in Seoul in 2007

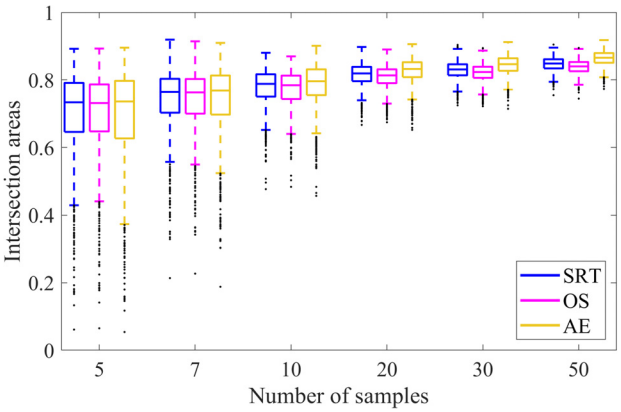


Fig. 6 Intersection areas for temperature data

8,760개의 데이터로 부터 표본을 생성하였기 때문에 표본의 무작위성이 낮아서 상자그림의 변동 폭이 Fig. 3보다 상대적으로 좁으면서 정확도의 변동성이 낮은 것을 볼 수 있다. 그리고 온도 분포는 Fig. 3(a)보다 이봉성이 뚜렷하고 (b)보다는 약해서 추정 정확도의 수렴속도가 (a)와 (b)의 사이 값을 보인다. 특히 표본의 크기가 증가함에 따라서 온도 분포의 비대칭형 이봉형태 때문에 (a)와 수렴속도 차이는 더욱 커졌다.

4.2 I-beam

신뢰성 해석을 위해서 먼저 실제 확률변수로부터 각 데이터의 개수에서 1,000개의 표본 세트를 생성하고 각 표본에 대해서 세 가지 방법을 사용하여서 KDE를 분포함수를 추정하고 이를 성능함수(performance function)에 사용하여서 파손확률을 예측하였다. 신뢰성 해석 예제로서 입력-확률변수가 단봉분포인 I-beam 문제와 다봉분포인 Bimodal ratio 문제를 사용하였다. 두 문제 모두 신뢰성 해석을 위해서 몬테카를로 시뮬레이션(monte-carlo simulation)을 수행하였고, 이 때 MCS 표본 수는 10^6개로 하였다.

I-beam 예제는 Fig. 7처럼 단순지지보의 임의의 지점에

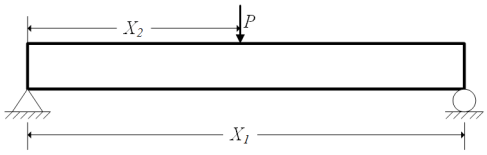


Fig. 7 I-beam

Table 1 Input variables of I-beam

Variables	Symbol	Dist.	Mean/Std.	CV(%)
Length(mm)	X_1	Normal	3000/25	0.83
Length(mm)	X_2	Normal	1830/25	1.37
Load(N)	P	-	27000	-
Section modulus(mm^3)	S	-	0.0822×59^3	-
Limited stress(MPa)	σ_c	-	1170	-

하중이 작용할 때 파손확률을 예측하는 문제이다(Ah *et al.*, 2009). Table 1은 I-beam의 입력변수를 나타내고, P, S, σ_c는 확정론적 변수(deterministic variable)로서 각각 하중, 단면계수, 한계응력이고 X_1과 X_2는 확률변수로서 각각 보의 길이와 보의 왼쪽 끝단으로부터 하중 점까지의 거리를 나타낸다. 확률변수에 따른 I-beam의 성능함수는 다음과 같다.

$$g(\mathbf{X})=\sigma_c-\frac{P\times X_2\times(X_1-X_2)}{S\times X_1} \tag{16}$$

여기서, 첫 번째 항은 I-beam의 한계응력이고 두 번째 항은 발생하는 최대 응력으로 한계응력보다 크면 보가 파손된다. 즉, I-beam의 파손확률은 $P_F=P[g(\mathbf{X})<0]$로 정의된다.

I-beam의 파손확률을 예측하기 위해서 세 가지 방법으로 추정된 커널 확률밀도함수(kernel PDF)를 사용해서 몬테카를로 시뮬레이션을 수행하고 데이터의 개수와 각 데이터의 개수에서 1,000개의 표본 세트에 대해서 예측된 파손확률을 상자그림으로 나타내면 Fig. 8과 같다. Fig. 8의 범례에서 P_F^{Exact}는 실제 확률변수를 사용한 파손확률이고 세 가지 기법에 따른 결과를 비교하기 위한 기준으로서 사용하였다. 데이터의 개수가 증가하면서 세 가지 기법으로 예측된 파손확률의 상자그림은 좁아지면서 실제 파손확률로 수렴하였다. 결과에서 상자의 높이는 세 가지 기법 모두 비슷하지만, AE가 실제 값에 가장 유사하고 OS가 가장 차이가 컸으며, SRT는 AE와 OS의 사이의 값을 예측하였다.

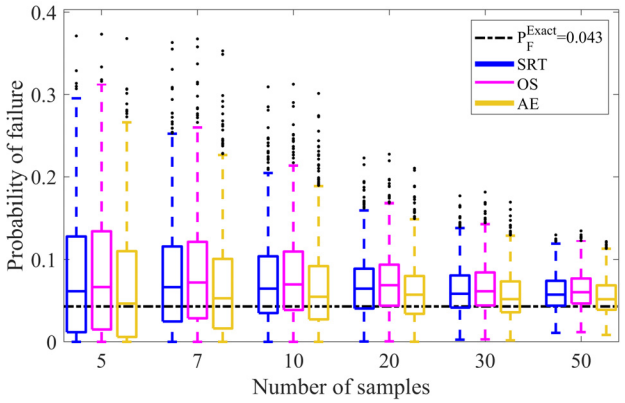


Fig. 8 Probability of failure according to the number of samples for I-beam example

4.3 Bimodal ratio

Bimodal ratio 예제는 기존의 lognormal ratio 문제(Eldred *et al.*, 2007)에서 두 확률변수의 분포함수를 이봉분포로 변환한 문제이다. Table 2는 bimodal ratio 문제의 입력변수를 나타내며, 여기서 R은 확정론적 변수로서 한계기준값이고, X_1과 X_2는 확률변수로서 두 개의 정규분포가 결합된 이봉확률분포이며 PDF는 Fig. 9에 나타내었다. Bimodal ratio의 성능함수는 다음과 같다.

Table 2 Input variables of bimodal ratio

Variables	Dist.	Mean/Std.	CV(%)
X_1	0.5N(150,10) & 0.5N(200,10)	175/26.93	15.39
X_2	0.5N(150,10) & 0.5N(200,10)	175/26.93	15.39
R	-	0.7/-	-

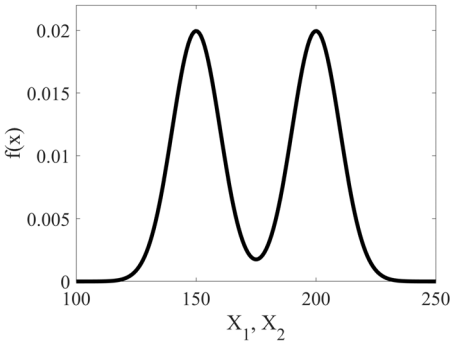


Fig. 9 Random variables of bimodal ratio example

$$g(\mathbf{X})=\frac{X_1}{X_2}-R \tag{17}$$

I-beam 문제와 같이 파손확률을 계산해서 Fig. 10에 나타내었다. Fig. 10의 예측된 파손확률을 보면, I-beam 문제와 같이 세 가지 기법을 사용한 경우 모두 데이터의 개수가 증가함에 따라서 실제 파손확률에 수렴하지만, 확률변수의 이봉성으로 인해 분포함수의 추정 정확도가 낮아져서 수렴속도가 I-beam 문제보다 느리다. 세 가지 기법을 비교해 보면, I-beam처럼 AE가 실제 값에 가장 가깝고 OS가 가장 멀었지만 AE와 OS의 차이는 I-beam 문제보다 크다.

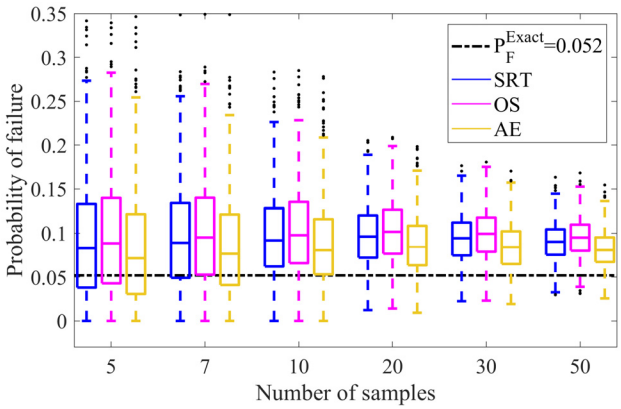


Fig. 10 Probability of failure according to the number of samples for Bimodal ratio example

결과를 정리하면, OS는 두껍고 긴 꼬리를 가지는 PDF를 추정하여서 예측된 파손확률이 실제 파손확률보다 높게 분포되어서 과소추정(underestimation)의 비율이 가장 낮으면서 가장 보수적인 결과를 보였다. 반면에 AE는 가장 짧은 꼬리의 PDF를 추정하여서 예측된 파손확률이 실제 파손확률을 주위로 분포되어서 가장 실제 값에 가까운 결과를 보인다. 하지만 AE는 과소추정의 비율이 가장 높은 한계점을 가진다. 그리고 세 가지 기법모두 데이터의 증가하면서 실제 파손확률에 수렴을 상대적으로 빠르게 하지만 다봉분포의 경우에는 분포함수의 과대평활화 때문에 파손확률 또한 보수적인 성향을 계속 가진다.

5. 결 론

본 논문은 커널밀도추정의 정확도에 가장 중요한 인자인 대역폭을 결정하는 세 가지 선정법을 사용하여서 단봉·다봉분포에 따른 분포함수의 추정 정확도를 비교·분석하였고, 단봉-확률변수를 가지는 I-beam 문제와 이봉-확률변수를 가지는 bimodal ratio 문제에 대해서 세 가지 대역폭 선정 기법에 따른 신뢰성 해석을 수행 후 비교·분석하였고, 결과는 다음과 같다.

(1) 확률변수가 단봉분포임을 알고 있는 경우, 다른 기법에 비해서 두꺼운 꼬리를 가지는 PDF를 예측하는 over-smoothing rule을 사용한 커널밀도추정의 정확도가 가장 정확하므로 OS 기법을 적용한 커널밀도추정을 추천한다.

(2) 확률변수가 다봉분포임을 알고 있는 경우, 다봉분포에서 꼬리가 뚜꺼워지는 효과가 가장 완화된 AE를 사용한 커널밀도추정의 정확도가 데이터의 개수에 상관없이 가장 정확하므로 AE 기법을 적용한 커널밀도추정을 추천한다.

(3) 확률변수가 단봉/다봉분포인지 모르는 경우, 만약 시스템의 불확실성이 커서 보수적인 해석 및 설계가 필요한 경우에는 OS의 사용을 추천하고, 반대로 불확실성이 작은 경우에는 AE의 사용을 추천한다.

(4) 세 가지 대역폭 선정방법 모두 대칭 또는 비대칭에 상관없이 단봉분포의 경우 제한된 데이터에 대해서 상대적으로 높은 분포함수 추정 정확도를 보이지만 다봉분포에서는 수렴속도가 정확도가 매우 낮고, 신뢰성 해석에서도 과도하게 보수적인 결과를 산출하므로 다봉분포를 가지며 보수적인 설계가 필요한 경우에만 세 가지 방법을 사용할 것을 추천한다.

본 연구는 데이터의 개수가 적은 경우에 중점을 두고 세 가지 대역폭 선정방법을 비교하였으나, 다봉분포와 같은 데이터의 개수가 증가한 경우에 대해서는 세 가지 방법의 한계점을 확인할 수 있었다. 그러므로 추후 단봉/다봉분포의 구분하는 기법의 연구와 다봉분포와 같이 데이터의 많은 경우에 대한 대역폭 선정방법을 연구가 추가적으로 필요하다.

감사의 글

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2018R1A6A3A01012213).

References

An, D., Won, J., Kim, E., Choi, J. (2009) Reliability Analysis under Input Variable and Metamodel Uncertainty using Simulation Method based on Bayesian Approach, *Trans. Korean Soc. Mech. Eng. A*, 33(10), pp.1163~1170.

Analytical Methods Committee (1989) Robust Statistics-how not to Reject Outliers, Part 1. Basic Concepts, *Analyst*, 114(12), pp.1693~1697.

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.

Jang, J., Cho, S.G., Lee, S.J., Kim, K.S., Kim, J.M., Hong, J.P., Lee, T.H. (2015) Reliability-based Robust Design Optimization with Kernel Density Estimation for Electric Power Steering Motor Considering Manufacturing Uncertainties, *IEEE Transactions on Magnetics*, 51(3), pp.1~4.

Jung, J.H., Kang, Y.J., Lim, O.K., Noh, Y. (2017) A New Method to Determine the Number of Experimental Data using Statistical Modeling Methods, *J. Mech. Sci. & Technol.*, 31(6), pp.2901~2910.

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.

Korea Meteorological Administration (KMA) https://data.kma.go.kr (accessed Mar., 6, 2019)

Lee, D., Hwang, I.S. (2011) Analysis on the Dynamic Characteristics of a Rubber Mount Considering Temperature and Material Uncertainties, *J. Comput. Struct. Eng. Inst. Korea*, 24(4), pp.383~389.

Scott, D.W. (2010) Scott's Rule, *Wiley Interdiscip. Rev.: Comput. Stat.*, 2(4), pp.497~502

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

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

Terrell, G.R., Scott, D.W. (1985) Oversmoothed Nonparametric Density Estimates, *J. Am. Stat. Assoc.*, 80(389), pp.209~214.

Terrell, G.R. (1990) The Maximal Smoothing Principle in Density Estimation, *J. Am. Stat. Assoc.*, 85(410), pp.470~477.

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

Wand, M.P., Jones, M.C. (1994) *Kernel smoothing*, CRC press, London.

Zhang, F., Liu, Y., Chen, C., Li, Y.F., Huang, H.Z. (2014) Fault Diagnosis of Rotating Machinery based on Kernel Density Estimation and Kullback-Leibler Divergence, *J. Mech. Sci. & Technol.*, 28(11), pp.4441~4454.

요 지

제한된 실험 데이터로부터 확률분포함수를 추정하기 위해서 KDE가 많이 사용되고 있다. KDE에 의한 분포함수는 대역폭 선택법에 따라서 실험 데이터에 대해 평활하거나 과대적합된 커널 추정치를 생성한다. 본 연구에서는 Silverman's rule of thumb, rule using adaptive estimate, oversmoothing rule을 사용해서 각 방법에 따른 정확성과 보수적인 성향을 비교하였다. 비교를 위해서 단봉분포와 다봉분포를 가지는 실제 모델을 가정하고 통계적 시뮬레이션을 수행한 다음 다양한 데이터의 개수에 따른 추정된 분포함수의 정확도와 보수성을 비교하였다. 또한, 간단한 신뢰성 예제를 통해 대역폭 선택법에 따른 KDE의 추정된 분포가 신뢰성 해석 결과에 어떻게 영향을 미치는지 확인하였다.

핵심용어 : 다봉분포, 단봉분포, 대역폭 선택법, 신뢰성 해석, 커널밀도추정, 통계 모델링