The Korean Communications
in Statistics Vol. 7, No. 3, 2000
pp. 811-816

Data-Driven Smooth Goodness of Fit Test by Nonparametric Function Estimation[1)]

Jongtae Kim[2)]

Abstract

The purpose of this paper is to study of data-driven smoothing goodness of fit test, when the hypothesis is complete. The smoothing goodness of fit test statistic by nonparametric function estimation techniques is proposed in this paper. The results of simulation studies for the powers of show that the proposed test statistic compares well to other.

Keywords : smooth goodness of fit, nonparametric function estimation, Fourier estimation.

1. 서 론

Neyman(1937)이 평활검정통계량을 소개한 이후, 최근 문헌들에서 자료 유추된(data-driven) 평활(smooth)검정통계량에 관한 논문들이 연이어 소개가 되어지고 있다. 그 예들로서 Bickel과 Ritov(1992), Eubank와 Hart(1992), Eubank와 Hart(1993), Eubank와 LaRiccia(1992), Ledwina (1994), Fan(1996), Hart(1997), Kallenberg와 Ledwina(1997), Kim(2000)등 많은 연구 논문들이 있다. 자료 유추된 평활 적합도검정의 가장 핵심적인 문제는 평활검정통계량이 가지는 요소들의 개수를 어떻게 가장 정확하게 추정해 내는가 하는 것이다. 그 이유는 정확한 검정통계량이 가질 수 있는 요소의 개수 추정은 평활검정통계량의 검정력에 많은 영향력을 미치기 때문이다. 실제로 1990년대 이전의 많은 평활검정에 관한 문헌들에서는 요소들의 개수 값들에 대하여 정확한 추정이 없이 임의로 사용하여 발표한 문헌들이 거의 대부분이었다.

다음절에서는 Kim (2000)이 제시한 비모수 함수추정의 방법을 이용하여 평활검정통계량의 요소들의 개수를 비모수 함수추정의 기법을 이용하여 추정하는 방법을 소개하고 가설이 복합가설일 때 자료 유추된 평활검정통계량을 제시하였다.

3절에서는 가설이 복합가설일 경우, 물론 다른 모형에 대하여서도 충분히 다룰 수 있지만, 본 연구에서 귀무가설이 정규성을 위한 가설검정의 문제에 대하여서 만 다루었다. 모의실험을 이용하여 제안된 검정통계량에 대한 기각값들을 제시하였고, 그리고 귀무가설에서의 유의수준에 제안된 통계량이 얼마나 잘 접근 하는가하는 일치성을 보였다. 그리고 다른 분포들에 대하여 검정력들을

1) This research was supported in part by the Taegu University Research Grant, 2000
2) Associate Professor, Department of Statistics, Taegu University. E-mail: jtkim@taegu.ac.kr

비교하였다. 4절에는 제안된 통계량의 문제점과 결론을 제시하였다.

2. 자료 유추된 평활 적합도 검정

확률변수 $X_1, X_2, \ldots, X_n$이 독립이고 동일한 밀도함수 $f(x)$로 분포되어져 있다. 관심 있는 귀무가설은 다음과 같다.

$$H_0 : f(x) \in \{f_0(x;\beta) : \beta \in B\},$$

여기서 $B \subset R^q$ 이고 $\{f_0(x;\beta) : \beta \in B\}$는 주어진 밀도함수들의 집합이고 f_0의 분포함수는 F_0이다. G를 $U = F_0(x;\beta)$의 분포함수로 정의하자. 그러면 $G(u) = F(F_0^{-1}(u))$는 다음과 같은 밀도함수 g를 가진다.

$$g = f(F_0^{-1}(u)) / f_0(F_0^{-1}(u)).$$

일반적인 g의 코사인 퓨리에 시리즈 급수와 계수 a_j는 각각 다음과 같이 표현된다.

$$g(u) = 1 + \sum_{j=1}^{\infty} a_j \sqrt{2} \cos(j\pi u),$$

$$a_j = \int_0^1 g(u) \sqrt{2} \cos(j\pi u)\, dx.$$

이때 표본 퓨리에 코사인 급수의 계수 a_j 에 수렴하는 추정량

$$a_{jn} = \frac{1}{n} \sum_{j=1}^{n} \sqrt{2} \cos(j\pi u)$$

을 이용하여 미지함수 g 에 대한 퓨리에 코사인 급수 추정량을 다음과 같이 구한다.

$$g_m(u) = 1 + \sum_{j=1}^{m} a_{jn} \sqrt{2} \cos(j\pi u) .$$

이때 m 은 $1 \le m \le n$ 을 만족하는 상수이다. Eubank와 Lariccia (1992)는 피어슨 phi-squared 거리척도 $\phi^2 = \int_0^1 (g(u)-1)^2 du$를 이용하여 Neyman의 평활검정을 연구하였다. 그들의 연구에서 나타나 있듯이 귀무가설 H_0의 검정은 $\phi^2 = 0$ 혹은 $a_j = 0$의 검정과 동일함을 알 수 있다. 본 연구에서는 적합한 차원의 개수 m을 추정해 내기 위하여 다음과 같은 평균적분제곱오차의 척도를 사용한다.

$$\begin{aligned} R(m) = E\{L(m)\} &= E\{\int_0^1 (g_m(u) - g(u))^2 du\} \\ &= -H(m) + \sum_{j=1}^{\infty} a_j^2. \end{aligned}$$

여기서 $-H(m) = \sum_{j=1}^{m} (a_j^2 - \sigma_j^2)$과 $\sigma_j^2 = \mathrm{var}(a_{jn}) = n^{-1}(1 + a_{2j}/\sqrt{2} - a_j^2)$이다. $R(m)$의 마지막 부분 $\sum_{j=1}^{\infty} a_j^2$는 m에 의존하지 않으므로 $H(m)$을 최대로 하는 m은 $R(m)$을 최소로 할 것

이다. 적절한 m을 찾기 위하여 $H(m)$에 대한 불편추정량 $\hat{H}(m)$은 다음과 같다.

$$\hat{H}(m)=\begin{cases} 0, & m=0, \\ \sum_{j=1}^{m}(a_{jn}^2-2\hat{\sigma}_j^2), & m>0. \end{cases}$$

여기서 $\hat{\sigma}_j^2=(1+a_{2jn}/\sqrt{2}-a_{jn}^2)/(n-1)$은 σ_j^2의 불편추정량이다.

우도비 검정통계량에 유추된 Neyman의 평활통계량과 동일한 검정통계량으로서 Eubank와 Lariccia(1992), Kim(2000)은 다음과 같은 검정통계량을 제안하였다.

$$T_{mn}=\int_0^1(g_m(u)-1)^2du=n\sum_{j=1}^{m}a_{jn}^2 .$$

국소대립(local alternatives)분포들에 대한 밀도함수 $g_n(u)=1+b(n)\delta(u)$에 대하여 $n\to\infty$에 따라서 $m\to\infty$이고, $m^{1/4}/\sqrt{n}\to 0$일 때,

$$Z_{mn}=\frac{T_{mn}-m}{\sqrt{2m}} \xrightarrow{\ \mathrm{d}\ } N(\|\delta\|^2/\sqrt{2},1).$$

이때 $b(n)=m^{1/4}/\sqrt{n}$이다. 그러므로 Z_{mn}은 일치검정통계량이다. (참조 Eubank와 Lariccia(1992), Kim(2000).)

3. 모의실험을 이용한 검정력 비교

귀무가설 H_0의 분포로는 여러 가지 분포를 사용할 수 있지만 본 논문에서는 가장 많은 관심의 대상이 되는 다음의 정규성 검정문제만을 다루었다. 귀무가설

$$H_0: f(x)\in\{f_0(x;\beta): \beta\in B\}$$

에 대하여 $f_0(x)=(\sqrt{2\pi}\sigma)^{-1}\exp\{-\frac{1}{2}(x-\mu)^2/\sigma^2\}$이고, $\mu\in R$과 $\sigma>0$을 가지는 $\beta=(\mu,\sigma)$이다. F_0는 정규확률밀도함수 f_0의 분포함수이다. 그러면 코사인 퓨리에급수 계수의 불편추정량

$$a_{jn}=\frac{1}{n}\sum_{j=1}^{n}\sqrt{2}\cos(j\pi F_0(x))$$

에 대하여 평활검정통계량은 $T_{mn}=n\sum_{j=1}^{m}a_{jn}^2$에 대하여 다음과 같이 표현된다.

$$Z_{mn}=\frac{T_{mn}-m}{\sqrt{2m}} .$$

이때 m의 값은 2절에서 소개한 $\hat{H}(m)$을 가장 최대로하는 m을 선택한다. 제안된 평활 검정통계량 Z_{mn}과 D'Agostino와 Stephens(1986)에 소개된 기존의 검정통계량들, Kolmogorov-Smirnov (D), Kuiper (V), Cramer von Mises (W^2), Watson (U^2), Finkelstein과 Schefer (S)과의 검정력을 비교 분석한다. <표 1>은 귀무가설 H_0의 표준정규분포에 대한 평활검정통계량의 기각값으로, 표본의 크기 $n=10,20,\ldots,50$에 대해 모의실험 반복 수를 100,000번 실행시켜 얻은 결과이다.

<표 1> 정규성 검정을 위한 평활검정통계량 Z_{mn}의 기각값

		α		
		0.05	0.025	0.01
n	10	.863	1.137	1.372
	20	.963	1.311	1.707
	30	.978	1.334	1.753
	40	.988	1.350	1.784
	50	1.004	1.377	1.823

다음 <표 2>는 표본의 크기 $n=50$일 때, <표 1>의 기각값이 평균 $\mu=-2,0,2,3$와 분산 $\sigma=0.5,1,2$을 가지는 정규분포의 난수를 10,000 발생시킨 후 μ와 σ의 추정량을 이용하여 표준정규분포로 정규화 시킨 다음 유의수준 $\alpha=0.05$에 얼마나 잘 적합하는지를 모의실험 하였다. 다음의 <표 2>에서 보듯이 Z_{mn} 통계량의 정규성에 대한 기각값들의 검정력이 유의수준에 잘 적합됨을 알 수 있다.

<표 2> 유의수준 $\alpha=0.05$에 대한 정규분포에서의 검정력 비교

μ	-2.0			0.0			2.0			3.0		
σ	0.5	1.0	2.0	0.5	1.0	2.0	0.5	1.0	2.0	0.5	1.0	2.0
power	.05	.049	.05	.05	.049	.051	.05	.049	.05	.05	.05	.049

<표 3>은 제시된 평활검정통계량의 검정력과 기존의 검정통계량들의 검정력을 비교하기 위하여 모수 a와 b를 가지는 2모수 와이블분포(Weibull distribution)를 대립 모형으로 하여 표본의 크기 $n=50$, 유의수준 $\alpha=0.05$에 대하여 반복 횟수를 10,000번으로 하여 모의실험의 결과이다.

<표 3> 유의수준 $\alpha=0.05$에 대한 와이블분포에서의 검정력 비교

a	b	Z_{mn}	D	V	W^2	U^2	S
1.0	1.0	.992	.957	.986	.992	.978	.993
1.2	1.0	.927	.843	.914	.929	.900	.937
1.0	2.0	.989	.959	.989	.993	.983	.994
1.2	2.0	.921	.815	.896	.912	.876	.932
1.0	3.0	.989	.954	.989	.987	.974	.989
1.2	3.0	.925	.839	.915	.939	.901	.941

<표 4>은 제시된 평활검정통계량의 검정력과 기존의 검정통계량들의 검정력을 비교하기 위하여 지수분포(exponential distribution)를 대립 모형으로 하여 표본의 크기 $n=50$, 유의수준

$\alpha=0.05$에 대하여 반복 횟수를 10,000번으로 하여 모의실험을 한 결과이다.

<표 4> 유의수준 $\alpha=0.05$에 대한 지수분포에서의 검정력 비교

b	Z_{mn}	D	V	W^2	U^2	S
0.1	.992	.939	.992	.992	.985	.992
1.0	.991	.954	.987	.988	.981	.989
10.0	.990	.939	.983	.982	.970	.983

<표 5>은 제시된 평활검정통계량의 검정력과 기존의 검정통계량들의 검정력을 비교하기 위하여 모수 a와 b를 가지는 감마 분포(gamma distribution)를 대립 모형으로 하여 표본의 크기 $n=50$, 유의수준 $\alpha=0.05$에 대하여 반복 횟수를 10,000번으로 하여 모의실험을 한 결과이다.

<표 5> 유의수준 $\alpha=0.05$에 대한 감마분포에서의 검정력 비교

a	b	Z_{mn}	D	V	W^2	U^2	S
1.5	1.0	.923	.818	.903	.935	.902	.940
2.0	1.0	.803	.676	.758	.830	.768	.884
1.5	2.0	.922	.807	.891	.927	.886	.930
2.0	2.0	.802	.692	.765	.840	.777	.852
1.5	3.0	.921	.817	.900	.929	.893	.936
2.0	3.0	.803	.666	.732	.821	.767	.843
1.5	4.0	.922	.798	.884	.913	.871	.932
2.0	4.0	.807	.676	.752	.826	.762	.841

4. 결론과 문제점

3절의 <표 3>에서 <표 5>까지의 모의실험의 결과에서 제안된 Z_{mn} 검정통계량은 기존의 검정통계량들, Kolmogorov-Smirnov (D), Kuiper (V), Cramer von Mises (W^2), Watson (U^2), Finkelstein과 Schefer (S)과의 검정력들 보다 우수하거나 거의 같음을 알 수 있다.보다 세밀히 살펴보면 제안된 검정통계량 Z_{mn}은 Cramer von Mises (W^2)와 Finkelstein과 Schefer (S)의 검정통계량들의 검정력들과 경쟁이 되어지고 그 외 Kolmogorov-Smirnov (D), Kuiper (V), Watson (U^2)의 검정통계량들의 검정력 보다는 검정력 값들이 높음을 알 수 있다.

그러나 본 연구의 초기의 생각은 분명 평활검정통계량이 기존의 검정 통계량인 Cramer von Mises (W^2)와 Finkelstein과 Schefer (S)의 검정통계량들 보다 검정력 측면에서 훨씬 월등히 좋을 것이라는 기대감에서 출발하였다. 그 이유는 평활 계수 m의 값이 잘 추정되어 질 수 있다는 확신에 기초한 것이고, 또한 Neyman 평활통계량의 우수성을 인지하고 있었기 때문이었다. 그러나

결과론적으로 Z_{mn}통계량은 우수한 통계량이기는 하지만 모수분포의 검정에 있어서 기존의 통계량들의 검정력 보다 월등히 뛰어나지는 못하였다. 이러한 문제점을 해결하기 위해서는 Kallenberg와 Ledwina(1997)의 방법에 대한 연구를 필요로 한다.

참고문헌

[1] Bickel, P.J. and Ritov, Y. (1992) *Testing Goodness of Fit: A New Approach,* in Non-parametric Statistics and Related Topics, ed. A.K.Md.E. Saleh, Amsterdam: North-Holland, 51-57.

[2] D'Agostino R.B. and Stephens M.A. (1986). *Goodness-of-Fit Techniques,* Marcel Dekker, Inc, New York.

[3] Eubank, R.L. and Hart, J.D. (1992). Testing Goodness-of-Fit in Regression via Order Selection Criteria, *The Annals of Statistics,* Vol 20, 1412-1425.

[4] Eubank, R.L. and Hart, J.D. (1993). Commonality of Cusum, von Neumann and Smoothing-Based Goodness-of-Fit Tests, *Biometrika,* Vol 80, 89-98.

[5] Eubank, R, and LaRiccia, V. N. (1992)., Asymptotic Comparison of Cramer-von Mises and Nonparametric Function Estimation Techniques for Testing Goodness-of-Fit, *Annals of Statistics,* Vol 20, 2071 - 2086.

[6] Fan, J (1996). Test significance Based on Wavelet Thresholding and Neyman's Truncation, *Journal of American Statistical Association,* Vol. 91, 674-688.

[7] Hart, J.D. (1997). *Nonparamtric Smoothing and Lack-of-Fit Tests,* Springer-Verlag, New York.

[8] Kallenberg, W.C.M. and Ledwina, T. (1997). Data-Dirven Smooth Tests When the Hypothesis Is Complete, *Journal of American Statistical Association,* Vol. 92, 1904-1104.

[9] Kim, J.T. (2000). Testing Goodness-of-Fit via Order Section Criteria, *Journal of American Statistical Association,* Vol.95, 829-835.

[10] Ledwina, T. (1994). Data-Driven Version of Neyman's Smooth Test of Fit, *Journal of American Statistical Association,* Vol. 89, 1000-1005.

[11] Neyman, J. (1937). Smooth Tests for Goodness of Fit, *Skand. Aktuar.* Vol. 20, 150-199.