The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.4.607
32(4), 607–618

Classical testing based on B-splines in functional linear models

Jihoon Sohn[a] · Eun Ryung Lee[b,1]

[a]Korea Credit Bureau Co., Ltd; [b]Department of Statistics, Sungkyunkwan University

(Received May 18, 2019; Revised June 7, 2019; Accepted June 12, 2019)

Abstract

A new and interesting task in statistics is to effectively analyze functional data that frequently comes from advances in modern science and technology in areas such as meteorology and biomedical sciences. Functional linear regression with scalar response is a popular functional data analysis technique and it is often a common problem to determine a functional association if a functional predictor variable affects the scalar response in the models. Recently, Kong *et al.* (*Journal of Nonparametric Statistics*, **28**, 813–838, 2016) established classical testing methods for this based on functional principal component analysis (of the functional predictor), that is, the resulting eigenfunctions (as a basis). However, the eigenbasis functions are not generally suitable for regression purpose because they are only concerned with the variability of the functional predictor, not the functional association of interest in testing problems. Additionally, eigenfunctions are to be estimated from data so that estimation errors might be involved in the performance of testing procedures. To circumvent these issues, we propose a testing method based on fixed basis such as B-splines and show that it works well via simulations. It is also illustrated via simulated and real data examples that the proposed testing method provides more effective and intuitive results due to the localization properties of B-splines.

Keywords: functional linear regression, functional association test, Wald test, functional principal component analysis, eigenfunctions, B-spline basis

1. 서론

최근 대규모 자료를 수집하고 저장하는 과학 기술이 급격하게 발전함에 따라, 시간, 파장(wavelength) 등과 같은 연속체 변수에 대해 (거의) 연속하게 자료값들이 관측되어 밀집해 기록되는(densely recorded) 형태를 띠는 자료가 등장하고 있다. 이와 같은 (연속체 변수의) 함수 형태를 띠는 자료를 함수형 자료(functional data)라고 부르고 요즘에는 기상학, 계량화학(chemometrics), 생물의학과 같은 많은 응용 분야에서 흔히 발견되어진다. 본 논문의 3.2절에서 분석한 시간에 대해 빽빽하게 관측한 온도곡선이 하나의 함수형 자료라고 할 있다. 이렇게 빽빽하게 얻어진 자료에서 인접하게 얻어진 관측값의 경우 상

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2019R1F1A1062795).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: silverryuee@gmail.com

당히 높은 상관성(correlations)을 가지므로 선형회귀분석(linear regression analysis)와 같은 전통적인 통계분석기법은 작동하지 않는다. 이에 따라, 이러한 함수 형태를 가지는 자료들을 분석하기 위한 새로운 분석 도구과 통계 방법들을 개발하는 것은 최신 통계학의 중요한 문제 중 하나였으며 대규모 단위의 연구가 지난 20년간 수행되어졌다. 지금까지 연구되었던 함수형 자료 분석을 위한 분석기법과 전체 통계방법론들을 간략하게 살펴보고자 하면, Ramsay와 Silverman (1997, 2002), Ramsay 등 (2009) 등을 확인할 수 있다.

함수형 자료는 본질적으로 무한차원(infinite dimension)을 가지고 있기 때문에 회귀, 군집, 분류와 같은 통계 분석을 위해 차원축소가 필요하며, 이 차원축소를 위해 함수형 주성분분석(functional principle component analysis)을 함수형 자료를 분석 하기 위한 유용한 도구로서 매우 광범위하게 사용되는 실정이다. 예를 들어, Ramsay와 Silverman (1997, 2002), Ramsay 등 (2009), Kneip와 Utikal (2001), Müller와 Stadtmüller (2005), Yao 등 (2005), Hall과 Horowitz (2007), Hall 등 (2007), Wang 등 (2016)과 같은 연구에서 확인할 수 있다. 중요한 함수형 자료 분석 주제 중 하나인 회귀분석, 예를 들어, 함수형 선형회귀 모형(functional linear regression models)은 그 중요성과 광범위한 응용성 덕분에 큰 주목을 받고 있다. 회귀 모형에서 설명변수인 함수형 변수가 관심있는 반응변수에 실제로 영향을 미치는지, 즉 함수적인 연관성을 검정하는 문제는 통계적으로도, 실제 응용 분석에서도 모두 의미를 가지는 문제이다. Kong 등 (2016)은 스칼라 값을 가진 반응변수를 위한 함수형 선형회귀에서 함수형 주성분분석의 차원 축소에 기반한 왈트(Wald), 스코어(score), 우도비(likelihood ratio), F-검정과 같은 고전적인 검정방법들을 연구했다.

하지만, 함수형 주성분분석, 구체적으로, Kong 등 (2016)에서 기저(basis)로 활용한 (설명변수의 공분산 작용소의) 고유함수들(eigenfunctions)은 회귀 목적에 적합하지 않을 수 있다. 이는 관심있는 설명변수와 반응변수의 함수적 연관성을 다루는 게 아니라, 설명변수의 변동 (즉, 분산) 만을 고려하기 때문이다. 이러한 단점은 함수형 회귀분석의 회귀계수함수(regression coefficient function) 추정법 연구에서도 고려되어진 적이 있다 (James 등, 2009; Lee와 Park, 2012). 또한, 자료로부터 고유함수들을 추정해야만 하는 단점이 있다. 이러한 이유로, 우리는 검정법에서 고정기저(fixed basis)를 활용하는 것을 제안하고자 하며 수치적으로 좋은 성질들 때문에 비모수 회귀(nonparametric regression) 분야에서 널리 사용되는 고정기저인 B-스플라인(B-splines)을 활용한 왈트 검정법을 논문에서 예로 보여주고자 한다. 본 논문에서 제안한 B-스플라인 기저 근사에 기반한 왈트 검정법은 귀무가설 하에서 유효하게 작동하며, 기존 Kong 등 (2016) 검정법보다 더 우수한 실제 성능과 더 해석하기 쉬운 분석 결과를 주는 것을 모의실험과 실증 분석 연구에서 확인할 수 있었다. 본 논문에서 제안한 함수형 주성분분석의 고유기저 대신 고정기저를 검정법에 활용하는 아이디어는 일반성을 가지므로 Kong 등 (2016)의 다른 검정법들에 쉽게 확장하여 생각할 수 있고, B-스플라인 기저의 장점에서 발생하는 이점을 검정 성능에서 가질 수 있으리라 예상한다.

앞으로, 본 논문은 다음과 같은 구성을 가진다. 2장은 함수형 선형회귀모형과 함수형 연관성을 검정하기 위한 기존 방법 Kong 등 (2016)을 간략하게 리뷰하고 논문에서 제안하는 검정법을 소개한다. 3장에서는 모의실험과 실제 기상자료 분석 예제를 통해 기존 검정법과 우리가 제안한 방법의 실제 성능을 비교하고자 한다.

2. 모형 및 방법론 설명

이 장에서는 본 논문에서 고려하고 있는 함수형 선형회귀모형과 Kong 등 (2016)에서 제안했던 함수형 주성분분석을 활용한 검정 방법을 간략하게 리뷰하고 우리가 제안하는 B-스플라인을 활용한 검정법을 소개하겠다.

2.1. 함수형 선형 회귀 모형

다음과 같은 함수형 선형회귀 모형을 고려하자.

$$Y = \alpha + \int_{\mathcal{I}} X(t)\beta(t)dt + \epsilon. \tag{2.1}$$

반응변수 Y는 실수값을 가지는 확률변수이고 $\epsilon \sim N(0, \sigma^2)$는 기대값이 0, 분산이 σ^2인 정규분포를 따르는 확률 오차항을 뜻한다. 또한, 공변량 $X(t)$와 회귀계수함수 $\beta(t)$는 $L^2(\mathcal{I})$ 공간에 속하는 함수이며, 회귀 모형 (2.1)에서 두 함수의 내적형태로 표현되게 된다. 여기에서 독립이고 동일한 분포를 가지는 $(X(\cdot), Y, \epsilon)$의 카피들 $(X_i(\cdot), Y_i, \epsilon_i)$, $i = 1, 2, \ldots, n$을 고려한다. 이 연구에서 관심있는 것은 함수형 설명변수 $X_i(t)$가 반응변수 Y_i에 영향을 끼치는지 검정하는 문제이고 모형 (2.1)에서는 회귀계수함수 $\beta(t)$가 ($L^2(\mathcal{I})$의 노음에서) 영함수인지와 연관된다. $L^2(\mathcal{I})$의 노음을 $\|\cdot\|_2$라 두자. 다시 말해, 함수형 연관성을 결정하는 문제는 다음과 같은 가설을 검정하는 문제로 변환될 수 있다.

$$H_0 : \|\beta\|_2 = 0, \qquad H_1 : \|\beta\|_2 \neq 0. \tag{2.2}$$

2.2. 함수형 주성분분석을 이용한 검정법

이 절에서는 Kong 등 (2016)이 제안한 기존 검정방법을 소개하겠다. 모형 (2.1)에서 회귀계수함수 β는 무한차원 모수이기 때문에 (유한표본으로부터) 직접 추정할 수 없고 유한차원으로 근사시키는 과정이 필요하다. 이를 위해 Kong 등 (2016)은 $X_i(\cdot)$, $i = 1, \ldots, n$의 함수형 주성분분석을 활용하여 차원 축소를 했다. 함수형 변수 X의 공분산 $K(s,t) = E(X(s), X(t))$은 다음과 같은 스펙트럼 분해를 가진다는 것이 알려져 있다.

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t). \tag{2.3}$$

여기에서 $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$는 $K(s,t)$를 커널로 가지는 적분 작용소(operator)인 공분산 작용소 $\mathcal{K}$의 고유값(eigenvalues)을 나타내고 $\phi_j(\cdot)$, $j = 1, 2, \ldots$은 고유값 λ_j에 대응하는 직교정규(orthonormal)한 고유함수가 된다. 정확하게, 공분산 작용소 $\mathcal{K} : L^2(\mathcal{I}) \to L^2(\mathcal{I})$는 임의의 함수 $f(s) \in L^2(\mathcal{I})$에 대해 $(\mathcal{K}f)(\cdot) = \int K(s, \cdot)f(s)ds$로 정의된다. $E(X)$를 μ라고 하자. 함수형 주성분분석에서 생성된 고유함수 $\phi_j(\cdot)$, $j = 1, 2, \ldots$는 $L^2(\mathcal{I})$ 공간의 기저를 형성하기 때문에 $X - \mu$는

$$X(t) - \mu(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t) \tag{2.4}$$

로 전개(expansion)할 수 있고 이를 흔히 Karhunen-Loëve 전개라고 부른다. 위에서 정의된 $\xi_j = \int_{\mathcal{I}}(X - \mu)\phi_j$는 기대값이 0, 분산이 λ_j인 확률변수이며 이를 주성분점수(Principal Scores)라 부른다. 나아가, 회귀계수함수 $\beta(\cdot)$는 $L^2(\mathcal{I})$공간에 속하기 때문에, $\sum_{j=1}^{\infty} \beta_j \phi_j(t)$와 같은 전개를 가지고 고려하고 있는 가설 (2.2)는

$$H_0 : \beta_j = 0 \text{ for all } 1 \leq j < \infty, \qquad H_1 : \beta_j \neq 0 \text{ for some } j$$

로 표현된다.

하지만, 실제 자료로부터 $K(\cdot, \cdot)$, $\phi_j(\cdot)$를 계산할 수 없으므로 함수형 주성분분석에서 이들을 추정한다. 함수형 변수의 표본평균은 $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$이고 표본공분산은 $\hat{K}(s,t) = n^{-1}\sum_{i=1}^{n}[X_i(s) -$

$\bar{X}(s)][X_i(t) - \bar{X}(t)]$라 두자. 식 (2.3)과 비슷하게

$$\hat{K}(s,t) = \sum_{j=1}^{\infty} \hat{\lambda}_j \hat{\phi}_j(s)\hat{\phi}_j(t)$$

라 나타낼 수 있고 $(\hat{\lambda}_j, \hat{\phi}_j)$는 $\hat{K}(s,t)$를 커널로 가지는 적분 작용소의 고유값과 이에 대응하는 고유함수의 쌍을 나타낸다. 고유값은 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq 0$ 순서로 나열한 것이고 고유함수는 직교정규 성질을 만족한다. 또한, (추정된) 주성분점수 $\hat{\xi}_{ij} = \int_{\mathcal{I}}(X_i - \bar{X})\hat{\phi}_j$는 $n^{-1}\sum_{i=1}^n \hat{\xi}_{ij} = 0$, $n^{-1}\sum_{i=1}^n \hat{\xi}_{ij}^2 = \hat{\lambda}_j$를 만족한다. 함수형 주성분분석에서는 실제로 관측하지 못하는 ξ대신 $\hat{\xi}$를 활용할 수 있고, 이는 다시 말해 추정된 고유함수 $\hat{\phi}_j$, $j = 1, 2, \ldots$를 ϕ_j대신 기저로 활용하여 회귀계수함수를 추정할 수 있다. Hall과 Hosseini-Nasab (2006), Hall과 Horowitz (2007), Zhu 등 (2014) 과 같은 연구에서 이론적으로 그 차이는 무시할 수 있음(negligible)을 밝혔다. 구체적으로, (n에 의존하는) 정수 k를 선택하고 다음과 같은 최소제곱기준

$$\left(\hat{\beta}_1, \ldots, \hat{\beta}_k\right) = \underset{\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y} - \sum_{j=1}^{k}\beta_j\hat{\xi}_{ij}\right)^2 \tag{2.5}$$

을 이용하고 회귀계수함수는

$$\hat{\beta}(t) = \sum_{j=1}^{k}\hat{\beta}_j\hat{\phi}_j(t)$$

로 추정할 수 있다. $\hat{\xi}_{ij}$를 (i,j) 원소로 가지는 $n \times k$ 행렬을 Ξ라 하고 $Y_i - \bar{Y}$를 i번째 원소로 가지는 n차원 벡터를 $\mathbf{Y}$라고 하자. 그러면, 식 (2.5)에서 구하는 최소제곱추정량은

$$\hat{\boldsymbol{\beta}} \equiv \left(\hat{\beta}_1, \ldots, \hat{\beta}_k\right)^\top = \left(\Xi^\top\Xi\right)^{-1}\Xi^\top\mathbf{Y}$$

로 계산할 수 있다. k는 n에 따라 변하는 조율모수인데 자료로부터 선택해야 한다. Kong 등 (2016)을 비롯한 함수형 주성분 분석 문헌에서는 일반적으로 선택한 k개의 주성분이 원래 변수 $X(t)$의 총 변동(total variation)에서 설명하는 비율인 percentage of variance explained (PVE)라는 기준을 사용한다.

여기에서 무한차원인 β를 k 차원으로 근사, 즉, $\beta \approx \sum_{j=1}^k \beta_j\phi_j$한 것이므로 고려하는 검정가설 (2.2)는

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0, \qquad H_1 : \beta_j \neq 0 \text{ for some } j = 1, 2, \ldots, k \tag{2.6}$$

로 바뀐다. Kong 등 (2016)과 Müller와 Stadmüller (2005)에서는 아래의 왈트 검정통계량

$$W_n = \hat{\boldsymbol{\beta}}^T\left[V\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\hat{\boldsymbol{\beta}}. \tag{2.7}$$

식 (2.6)의 H_0하에서 n이 증가함에 따라 (k가 무한대로 발산할 지라도) 자유도 k를 가지는 χ^2 분포와 점근적 극한이 같다는 것이 알려져 있다. 정확하게, H_0하에서 n이 무한대로 발산함에 따라

$$P(W_n \leq x) - P\left(\chi^2(k) \leq x\right) \to 0$$

임을 보일 수 있다. 여기에서 $V(\hat{\boldsymbol{\beta}})$은 $\hat{\boldsymbol{\beta}}$의 (추정된) $k \times k$ 분산-공분산행렬이며 $(\Xi^\top\Xi)^{-1}\hat{\sigma}^2$로 계산된다 (여기에서, $\hat{\sigma}^2 = \mathbf{Y}^\top(I - H)\mathbf{Y}/(n - k - 1)$, $H = \Xi(\Xi^\top\Xi)^{-1}\Xi^\top$로 정의된다). 따라서, 왈드 검정통계량 W_n이 카이제곱분포의 상위 $\alpha(0 < \alpha < 1)$ 분위수 $\chi^2_\alpha(k)$ 보다 크면 유의수준 α에서 H_0를 기각하고 모형 (2.1)에서 함수형 변수 X가 반응변수 Y에 영향을 미친다고 결론 내린다.

2.3. 제안: B-스플라인을 활용한 검정법

이전 절에서 소개한 식 (2.7)에서 정의된 왈트통계량 검정법은 함수형 주성분분석 결과 발생하는 고유함수 ϕ_j (혹은 그 추정값 $\hat{\phi}_j$)를 기저로 고려한 것이다. 앞서 서술한 것처럼 ϕ_j, 나아가 주성분분석은 설명변수 X의 변동만을 고려한 것일 뿐 X와 Y의 연관성을 고려한 것은 아니다. 이에 따라, 고유기저 대신 고정기저를 이용한 검정법을 고려하고 대표적인 고정기저의 예인 B-스플라인을 활용한 왈트 검정통계량을 제안한다. Eilers와 Marx (1996), Eubank (1988) 등에서 볼 수 있는 것처럼, 스플라인 기저는 복잡한 함수를 수치적으로 잘 근사하는 좋은 성질을 가지고 있기 때문에 비모수 회귀분석에서 많이 사용되는 기저 중 하나로 알려져 있다. 나아가, 본 논문에서 사용하고 있는 B-스플라인은 각 기저함수가 국소적으로 정의되어 있어서 현재 고려하고 있는 검정 문제에서 해석상 이점을 가지고 있다. (추가 분석을 실행하여) 각 계수의 유의성을 식별할 수 있다면, 이에 해당하는 (기저함수의) 정의역의 영역에서 설명변수 $X(t)$이 Y에 영향을 끼친다고 해석할 수 있다. 이것은 설명변수 $X(t)$이 Y에 영향을 미치는 (t의) 영역을 통계적으로 식별하는 것으로 볼 수 있고 이런 맥락에서 James 등 (2009), Lee와 Park (2012)와 연관성을 가지고 있다.

먼저, 모형 (2.1)의 회귀계수함수 $\beta(\cdot)$를 근사하기 위해, 동등하게 나눈 b개의 내적매듭(internal knots)을 가지고 오더 $d+1$인 B-스플라인기저 함수들 $B_1, B_1, \ldots, B_q$를 고려하자. 단, 기저함수의 개수 q는 $b+d+1$과 같다. 회귀함수는 $\beta(t) \approx \sum_{j=1}^{q} b_j B_j(t)$로 근사되고 이를 모형 (2.1)에 대입하면 아래와 같다. 임의의 $1 \le i \le n, 1 \le j \le q$에 대해 η_{ij}를 $\int_{\mathcal{I}}(X_i - \mu)B_j$라 정의하고 $\alpha' = \alpha + \int_{\mathcal{I}} \beta(t)\mu(t)dt$라 두면,

$$\begin{aligned} Y_i &= \alpha' + \int_{\mathcal{I}} (X_i(t) - \mu(t))\beta(t)dt + \epsilon_i \\ &\approx \alpha' + \int_{\mathcal{I}} (X_i(t) - \mu(t)) \left\{ \sum_{j=1}^{q} b_j B_j(t) \right\} dt + \epsilon_i = \alpha' + \sum_{j=1}^{q} b_j \eta_{ij} + \epsilon_i \end{aligned} \tag{2.8}$$

로 표현된다. 이전 절에서 소개한 Kong 등 (2016)과 비슷하게 추정과 검정법을 수립할 수 있다. 함수형 주성분분석의 $\hat{\xi}_{ij}$ 대신 $\hat{\eta}_{ij} = \int_{\mathcal{I}}(X_i - \bar{X})B_j$를 원소로 가진 $n \times q$ 설계행렬(design matrix) D를 고려하고 이를 활용한 최소제곱 추정량

$$\hat{\boldsymbol{b}} \equiv \left(\hat{b}_1, \ldots, \hat{b}_q\right)^\top = \left(\boldsymbol{D}^\top \boldsymbol{D}\right)^{-1} \boldsymbol{D}^\top \mathbf{Y}$$

을 생각한다. 이때, 회귀계수함수 $\beta(\cdot)$의 추정량은 $\sum_{j=1}^{q} \hat{b}_j B_j(\cdot)$와 같다. 본 논문에서 $\hat{\mathbf{b}}$를 고려한 왈드 검정통계량

$$W_n^* = \hat{\mathbf{b}}^T \left[V^* \left(\hat{\mathbf{b}} \right) \right]^{-1} \hat{\mathbf{b}} \tag{2.9}$$

을 검정법에 이용할 것을 제안한다. 이때 $V^*(\hat{\mathbf{b}})$는 (설계행렬이 $\boldsymbol{D}$이고 반응변수 벡터가 $\mathbf{Y}$인) 해당 최소제곱 회귀에서 $\hat{\mathbf{b}}$의 분산-공분산 행렬 추정치와 같다. 본 논문에서 관심있는 귀무가설과 대립가설

$$H_0 : b_1 = b_2 = \cdots = b_q = 0, \qquad H_1 : b_j \neq 0, \quad \text{for some} \;\; j = 1, 2, \ldots, q$$

중 H_0가 사실일 때, 위에서 제안한 왈트 통계량 W_n^*의 분포는 자유도가 q인 카이제곱분포 $\chi^2(q)$로 근사할 수 있음을 보일 수 있다. 정확히, $q = q_n \to \infty$가 느리게 증가하면,

$$P(W_n^* \le x) - P\left(\chi^2(q) \le x\right) \to 0$$

임을 보일 수 있다. 다시 말해, 식 (2.9)부터 계산한 왈드 검정통계량 W_n^*이 귀무분포(null distribution)인 자유도가 q인 카이제곱분포의 상위 α-분위수 $\chi^2_\alpha(q)$ 보다 클 때 귀무가설 H_0를 기각하고 대립가설 H_1을 택한다. 즉, $X(\cdot)$와 Y는 함수적 연관성을 가진다고 결론을 내릴 수 있다.

3. 모의실험 및 실증 예제

본 장에서는 모의실험과 실증 예제를 통해 Kong 등 (2016)에서 제안했던 함수형 주성분분석을 활용한 검정법과 우리가 제안한 B-스플라인을 이용한 검정법의 실제 성능을 비교하고자 한다.

3.1. 모의실험 예제

본 논문에서 수행한 모의실험 연구는 크게 두 부분으로 나누어진다. 먼저 함수적 연관성이 없는 경우, 즉, 귀무가설이 사실일 때, 제안한 검정법의 제 1종 오류를 확인하고, 함수적 연관성이 있는 경우, 즉, 대립가설이 사실일 때, Kong 등 (2016)과 우리가 제안한 방법들의 검정력을 살펴본다.

본 논문에서 제안한 검정방법을 계산할 때, 내부매듭이 2개를 가진 1차 B-스플라인을 이용했고, R 함수 'bs'를 이용하여 B-스플라인 함수값을 구했다. Kong 등 (2016)의 검정통계량 W_n을 계산할 때, R 패키지 'fdapace'에 있는 함수 'FPCA'를 이용하여 식 (2.5)의 함수형 주성분 점수 $\hat{\xi}_{ij}$를 구했다. 이 때, (뒤에서 정의할 X의 정의역인) $[0, 10]$에서 동일하게 배분된 이산화된 점(discretized points) 300개를 사용했다. 또한, Kong 등 (2016) 검정법에서 주성분 점수의 개수 k에 대해서 두 가지 선택을 고려했다. Kong 등 (2016)의 모의실험에서 고려한 것처럼 PVE가 99%가 되는 $k = 6$과 앞에서 정한 1차 B-스플라인 기저함수의 개수(= 3)와 같은 $k = 3$, 다시 말해, PVE가 80%가 되는 선택을 고려했다. 여기에서 본 논문에서 제안한 검정법은 'Proposed', 그리고 Kong 등 (2016)에서 $k = 3, 6$개의 주성분들을 이용한 검정법은 'Kong3', 'Kong6'으로 각각 부르기로 한다.

모의실험 연구에서 비교를 목적으로 Kong 등 (2016)와 비슷한 설정을 선택했다. 먼저, $L_2(\mathcal{I})$ 공간에 속하는 확률함수(random functions) $X_i(t)$를 생성하기 위해 고유값을 다음과 같이

$$\lambda_1 = 16,\ \lambda_2 = 12,\ \lambda_3 = 8,\ \lambda_4 = 4,\ \lambda_5 = 2,\ \lambda_6 = 1, \quad \text{and} \quad \lambda_k = 0, \quad \text{for } k \geq 7$$

로 설정하고 $[0, 10]$에서 정의된 푸리에 기저함수(Fourier basis functions)를 고유함수로 택했다. 정확히, $0 \leq t \leq 10$에 대해, $\phi_1(t) = \cos(\pi t)/\sqrt{5}$, $\phi_2(t) = \sin(\pi t)/\sqrt{5}$, $\phi_3(t) = \cos(3\pi t)/\sqrt{5}$, $\phi_4(t) = \sin(3\pi t)/\sqrt{5}$, $\phi_5(t) = \cos(5\pi t)/\sqrt{5}$, $\phi_6(t) = \sin(5\pi t)/\sqrt{5}$로 두었다. 이에 따라, ($i$번째 관측값의) j번째 주성분 점수 ξ_{ij}들은 $N(0, \lambda_j)$에 독립적으로 발생시킨 난수이며 함수형 설명변수는 $X_i(\cdot) = \sum_{j\geq 1} \xi_{ij}\phi_j(\cdot)$, $1 \leq i \leq n$이다. 또한, 오차항 ϵ_i, $1 \leq i \leq n$는 $N(0, 1)$에서 독립적으로 발생시킨다. 앞에서 발생한 $X_i(\cdot)$와 ϵ_i들을 가지고 회귀계수함수는 국소적으로 정의된 함수인

$$\beta_c(t) = c \cdot I(t > 8) \tag{3.1}$$

로 고려하고 $\alpha = 0$로 설정한 함수선형회귀모형 (2.1)로부터 반응변수 Y_i, $1 \leq i \leq n$를 생성한다. 이 때, c가 0인 것은 귀무가설이 사실인 상황을 뜻하며, c가 0이 아닌 것은 대립가설이 사실인 경우를 의미한다. 본 모의실험에서 표본수는 $n = 500$, 몬테카를로(Monte Carlo) 표본수는 $M = 1000$으로 정했다.

먼저 이 모의실험 모형에서 $c = 0$인 경우, 본 논문에서 제안한 W_n^*를 활용한 검정법의 실제 성능, 즉, 제 1종 오류 성능을 확인하고자 한다. 유의수준 α의 값이 0.01에서 0.2로 0.01씩 증가시킴에 따라 우리가 제안한 검정법이 총 $M = 1000$ 몬테카를로 반복 중에서 기각한 사건의 비율, 즉, (추정된) 제 1종 오류을 구했다. 계산한 기각 비율들은 Table 3.1에 나타나 있고, 본 논문에서 제안한 검정법은 유효한

Table 3.1. (Estiated) Type I error probability for $\alpha = 0.01, 0.02, \ldots, 0.20$

α	제1종 오류 확률
0.01	0.006
0.02	0.015
0.03	0.032
0.04	0.035
0.05	0.046
0.06	0.054
0.07	0.062
0.08	0.077
0.09	0.083
0.10	0.098
0.11	0.109
0.12	0.115
0.13	0.124
0.14	0.133
0.15	0.142
0.16	0.150
0.17	0.165
0.18	0.180
0.19	0.195
0.20	0.197

제 1종 오류 성능을 실제로 보여주고 있는 것을 확인할 수 있다. 이 모의실험에서 택한 모형은 Kong 등 (2016)의 디자인 1과 같으므로, Kong 등 (2016) 검정법의 제 1종 오류 실제 성능은 해당 논문에서 확인할 수 있다.

다음은, 본 논문에서 고려하는 검정법들의 실제 검정력들을 비교하고자 하기 위해, 0.05에서 0.3까지 0.01씩 증가시키면서 총 26개의 영이 아닌 c 값들을 고려했다. 이 실험에서 유의수준은 $\alpha = 0.05$로 고정시켰다. 이 설정들 하에 고려하는 세 검정방법(Kong3, Kong6, Proposed)들이 총 $M = 1000$ 몬테카를로 샘플 중에서 기각한 사건의 비율, 즉, (추정된) 검정력들을 구했고, 이는 Table 3.2에서 찾을 수 있다. 본 논문에서 제안한 검정법(Proposed)가 기존 검정 방법들(Kong3, Kong6)보다 더 높은 검정력을 가짐을 확인할 수 있다. 현재 설정한 회귀계수함수 $\beta_c(\cdot)$가 전체가 아닌 국소 영역(local region)에서 정의되어 있기 때문에 (기존 방법에서 사용하는 함수형 주성분분석의 고유기저 보다) B-스플라인 기저가 국소적으로 근사를 더 잘 할 수 있고 이에 따라 검정력에서도 더 이점을 가졌으리라 추측한다.

3.2. 실증 예제: 대한민국 기상청 자료

이 절에서는 실제 대한민국의 기상자료를 분석한 실증 예제를 제시하고자 한다. 분석한 자료는 대한민국 각 지역(총 89개)에서 얻은 2013-2015년 동안 관측한 일별(평균)온도와 강수량 자료이며, 이들은 기상자료개방포털사이트(https://data.kma.go.kr)에서 이용가능하다. 각 지역에서 (총 3년동안 평균한) 일별기온 곡선들을 설명변수로, (총 3년동안 평균한) 연간평균 강수량을 반응변수로 고려하여 기온이 강수량에 영향력을 끼치는지 살펴보고자 했다. Figure 3.1에서 볼 수 있는 원자료인 일별온도 관측값들은 매우 구불구불한(wiggly) 모양을 띄고 있으므로, Ramsay 등 (2009)에서 제시한 과정을 따라 65개의 푸리에 기저를 이용하여 평탄화(smoothing)하는 전처리 방법을 적용했고 그 결과 얻어진 평탄화된 일별온도곡선을 $X_i(\cdot)$, $i = 1, 2, \ldots, 89$로 택했다. 89개 지역의 평탄화된 일별온도곡선은 Figure 3.2를 참

Table 3.2. Powers of the testing methods for $c = 0.05, 0.06, \ldots, 0.30$ when $\beta_c(t) = c \cdot I(t > 8)$ is the (true) regression coefficient function

c	Kong3	Kong6	Proposed
0.05	0.052	0.053	0.055
0.06	0.057	0.057	0.058
0.07	0.063	0.061	0.065
0.08	0.072	0.065	0.071
0.09	0.084	0.070	0.078
0.10	0.091	0.081	0.087
0.11	0.100	0.092	0.106
0.12	0.115	0.103	0.127
0.13	0.128	0.117	0.141
0.14	0.144	0.130	0.156
0.15	0.157	0.151	0.170
0.16	0.174	0.164	0.190
0.17	0.191	0.178	0.217
0.18	0.215	0.191	0.247
0.19	0.237	0.217	0.270
0.20	0.271	0.241	0.302
0.21	0.298	0.264	0.325
0.22	0.319	0.278	0.359
0.23	0.346	0.297	0.386
0.24	0.374	0.322	0.409
0.25	0.393	0.352	0.450
0.26	0.436	0.388	0.483
0.27	0.464	0.407	0.521
0.28	0.499	0.441	0.545
0.29	0.532	0.468	0.573
0.30	0.572	0.493	0.613

고한다.

먼저 Kong 등 (2016)의 기존 검정법을 고려하자. Table 3.3은 주성분 개수 k에 따라 해당 PVE 값들을 보여준다. 개수를 $k = 1, 2, 3$으로 선택했을 때, 모든 경우에서 기존 검정법의 p-값들은 0.001보다 작았으며 일별 온도와 연간강수량 사이에 함수적 연관성이 있다고 결론을 내릴 수 있었다. 다음으로, 우리가 제안한 방법을 적용해 보겠다. 여기에서 B-스플라인의 내부매듭의 개수와 차수는 일반화 교차타당선(generalized cross-validation) 기준인

$$\text{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i - \sum_{j=1}^{q}\hat{b}_j(q)\,\eta_{ij}}{1 - q/N}\right\}^2$$

를 최소화하는 값으로 선택했다. 최종적으로 10개의 내부매듭을 가지는 1차 B-스플라인이 선택되어 우리 분석에 이용되었다. 이에 따라 이전 절에서 제안한 W_n^*를 계산하여 검정했을 때 해당 p-값은 매우 작게(< 0.0001) 나왔고 우리가 제안한 방법 역시 온도가 강수량 사이에 영향을 끼친다는 같은 결론을 내린다.

Figures 3.3은 R 패키지 'fda'에서 제공하는 함수형 주성분분석 (PVE 80, 90, 95%)과 B-스플라인을 활용한 회귀계수함수 추정치 $\hat{\beta}(\cdot)$와 95% 점별 신뢰구간(pointwise confidence interval)을 보여준다. 추

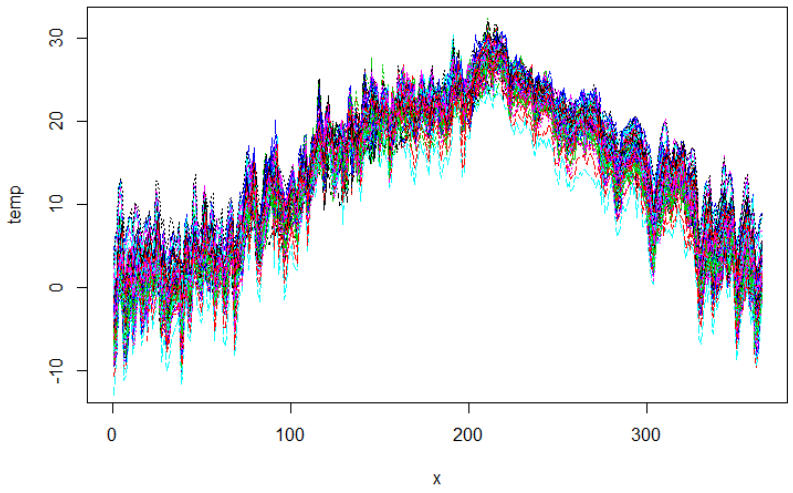


Figure 3.1. Observed daily temperatures for 89 regions in Korea.

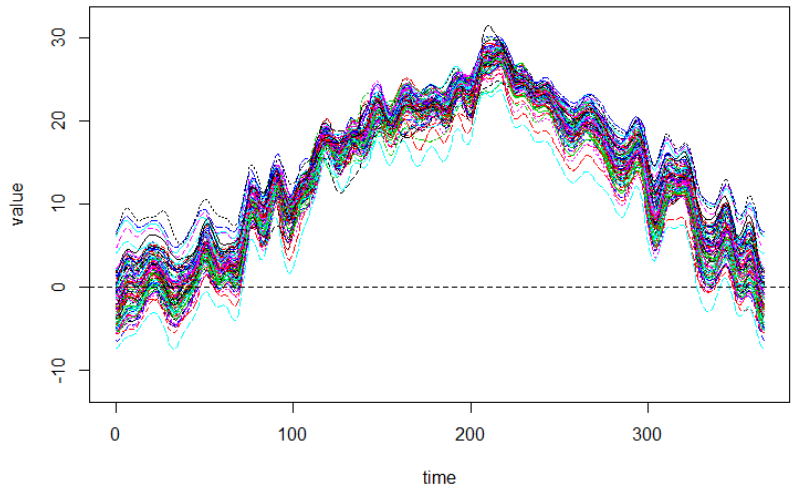


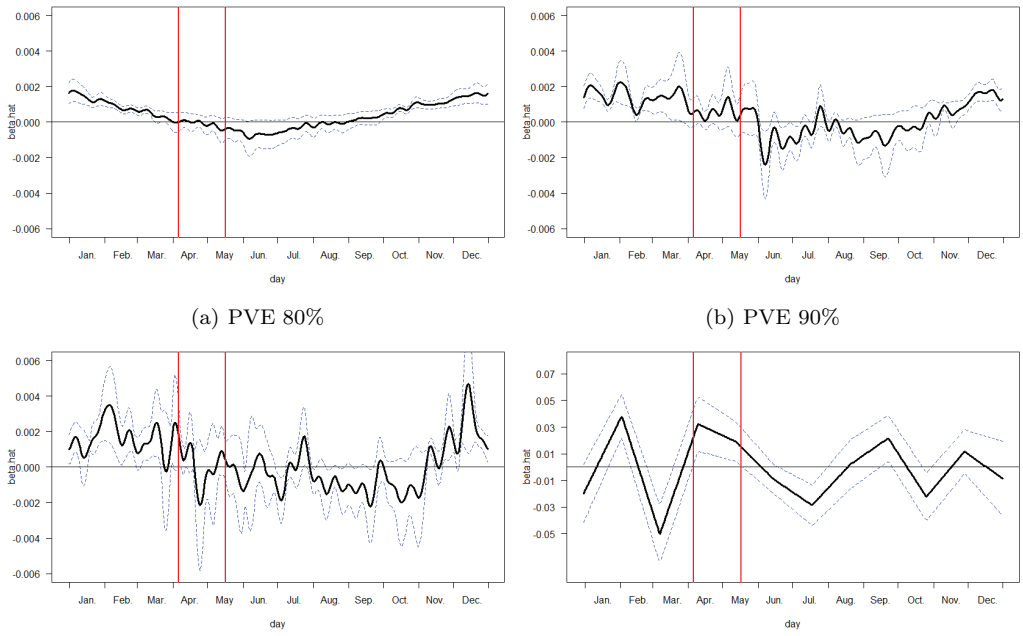
Figure 3.2. Smoothed temperature curves (right) for 89 regions in Korea.

Table 3.3. PVE when $k = 1, 2, 3$ is chosen in the functional PCA result with daily temperature data

주성분 개수	PVE
1	81.98%
2	93.29%
3	96.60%

PVE = percentage of variance explained; PCA = principle component analysis.

가적으로, 모든 그림들에 4월 6일과 5월 17일에 해당하는 시점들이 빨간 선으로 표시되었다. 4월 6일부터 5월 17일까지 구간에서 함수형 주성분분석의 회귀함수에 대한 점별 신뢰구간들은 모두 0을 포함하고 B-스플라인을 활용한 점별 신뢰구간들은 0을 포함하지 않음을 볼 수 있다. 즉, (해당 패키지에서 제공하는) 회귀계수함수 추정치와 점별 신뢰구간 정보만으로는 해당 구간 4월 6일–5월 17일에서 일별 온



(a) PVE 80% (b) PVE 90%

(c) PVE 95% (d) B-spline basis

Figure 3.3. Estimated regression coefficient function $\hat{\beta}$. PVE = percentage of variance explained.

도와 강수량의 연관성이 있는지 결론을 내릴 수 없고 추후 통계 분석이 필요함을 알 수 있다. 하나의 해결책은 본 논문에서 제안한 검정법에서 b_4에 대한 유의성을 검정하는 것이다. 더 나아가, B-스플라인의 모든 계수 $b_1, b_2, \ldots, b_{11}$가 0은 아니라는 대립 가설을 택했을 때 다음 과정으로 어떤 b_j가 0이 아닌지 다중가설검정(multiple hypothesis testing)을 하는 것은 어떤 국소지역에서 일별 온도와 강수량이 연관이 있는지 통계적으로 식별하기 위해 필요하다. 이러한 맥락에서 함수형 회귀분석에서 다중가설검정을 위한 통계방법론을 설립하는 건 실제적 의미를 가지며 이를 추후 연구과제로 남긴다.

4. 결론

본 논문은 스칼라 값을 가지는 반응변수를 가진 함수선형회귀모형에서의 검정 문제를 고려하고 있다. Kong 등 (2016)에서 제안한 검정법을 형성할 때 사용하는 함수형 주성분분석의 고유기저가 함수적 연관성을 전혀 고려하지 못해 회귀분석 목적에 적합하지 못할 수 있다는 한계점에 착안하여 고정기저 중의 하나인 B-스플라인 기저를 활용한 검정법을 제안했다. B-스플라인은 비모수 회귀분석에서 널리 사용되는 기저 중 하나로 특히, 비모수 회귀함수가 국소적으로 성질이 변할 때에도 잘 적응가능한(adaptive) 것으로 알려져 있다. 이러한 좋은 성질 때문에 앞에서 실제 수행했던 모의실험과 실증예제에서 우리가 제안한 검정법이 기존의 함수적 주성분분석의 고유기저를 사용한 Kong 등 (2016) 검정법보다 더 우수한 검정력 성능과 좀 더 해석하기 쉬운 분석 결과를 줄 수 있음을 확인했다. 본 논문에서 제안한 고정기저를 검정법에 활용하는 아이디어는 일반성을 가지므로 Kong 등 (2016)의 F, 스코어, 우도비와 같은 다른 검정법들에 쉽게 확장 가능하다. 뿐만 아니라, 우리가 본 논문에서 왈트 검정에서 확인했던 수치적, 실제적 이점들을 다른 검정법에서도 얻을 수 있을 거라 예상한다. 나아가, 우리가 제안한 검정법이

기각했을 때 다음 단계로 국소적으로 어떤 지역에서 설명변수와 반응변수의 함수적 연관성을 가지는지 식별하는 것이 실제 필요할 수 있으며, 이러한 목적을 위해 적절한 통계방법론을 설립하는 것은 추후 연구가 필요한 중요한 과제라 생각한다.

References

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties, *Statistical Science*, **11**, 89–121.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression, *Annals of Statistics*, **35**, 70–91.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society*, **B68**, 109–126.

Hall, P., Lee, Y. K., and Park, B. U. (2007). A method for projecting functional data onto a low-dimensional space, *Journal of Computational and Graphical Statistics*, **16** 799–812.

James, G., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable, *Annals of Statistics*, **37**, 2083–2108.

Kneip, A. and Utikal, K. J. (2001). Inference for density families using functional principal component analysis, *Journal of the American Statistical Association*, **96**, 519–532.

Kong, D., Staicu, A.-M., and Maity, A. (2016). Classical testing in functional linear models, *Journal of Nonparametric Statistics*, **28**, 813–838.

Lee, E. R. and Park, B. U. (2012). Sparse estimation in functional linear regression, *Journal of Multivariate Analysis*, **105**, 1–17.

Müller, H. G. and Stadtmüller (2005). Generalized functional linear models, *Annals of Statistics*, **33**, 774–805.

Ramsay, J. and Silverman, B. W. (1997). *Functional Data Analysis*, Springer.

Ramsay, J. and Silverman, B. W. (2002). *Applied Functional Data Analysis*, Springer.

Ramsay, J., Hooker, G., and Silverman, B. W. (2009). *Functional Data Analysis with R and MATLAB*, Springer.

Wang, J. L., Chiou, J., and Müller, H. G. (2016). Functional data analysis, *Annual Review of Statistics and Its Application*, **3**, 257–295.

Yao, F., Müller, H., and Wang, J. (2005). Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association*, **100**, 577–590.

Zhu, H., Yao, F., and Zhang, H. H. (2014). Structured functional additive regression in reproducing kernel Hilbert spaces, *Journal of the Royal Statistical Society*, **B76**, 581–603.

함수형 선형모형에서의 B-스플라인에 기초한 검정

손지훈[a] · 이은령[b,1]

[a]코리아크레딧뷰로(주), [b]성균관대학교 통계학과

(2019년 5월 18일 접수, 2019년 6월 7일 수정, 2019년 6월 12일 채택)

요 약

현대 과학기술의 발전으로 인해 함수 형태의 자료(functional data)는 기상학, 생물의학과 다양한 분야에서 발생하고 있으며 이러한 자료를 분석하는 것은 새롭고 흥미로운 통계과제라 할 수 있다. 스칼라 반응변수를 가진 함수형 선형회귀 모형(functional linear regression models with scalar response)은 널리 사용되는 함수형 자료 분석 기법 중의 하나라 할 수 있고 이 회귀 모형에서 함수형 자료 (설명변수) 가 스칼라 반응변수에 영향력을 미치는지 검정하는 것은 중요한 문제라 할 수 있다. 최근, Kong 등은 함수형 주성분분석(functional principle component analysis)에 의한 차원 축소, 즉, 함수형 주성분분석 결과 얻어지는 고유함수(eigenfunctions)를 활용한 검정방법을 제안했다. 하지만, 그 고유함수들은 검정문제에서 관심사인 함수형 설명변수와 스칼라 반응변수의 연관성이 아니라 함수형 설명변수의 변동만을 고려하기 때문에 회귀문제에 사용하기에 일반적으로 적합한 기저가 아니다. 게다가, 자료로부터 추정하여야 하기 때문에 이 불필요한 추정오차가 검정 절차 성능에 포함될 가능성이 있다. 이러한 단점을 피하기 위해 본 논문에서는 기존의 고유기저함수가 아닌 고정기저(fixed basis)인 B-스플라인(B-splines) 함수를 활용한 검정 방법을 제안한고 모의실험을 통해 검정방법이 잘 작동한다는 것을 보여준다. 또한, 제안한 검정 방법은 B-스플라인의 국소화 성질 때문에 때론 효율적이고 직관적인 결과를 제공하는데 이를 모의실험과 실증자료 분석을 통해 보여줄 것이다.

주요용어: 함수형 선형 회귀, 함수형 연관성 검정, 왈트 검정, 함수형 주성분분석, 고유함수기저, B-스플라인 기저

본 연구는 한국 연구재단의 지원을 받아 수행한 연구임 (No. NRF-2019R1F1A1062795).

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: silverryuee@gmail.com