The Korean Journal of Applied Statistics
2021, Vol. 34, No. 5, 773–794

DOI: https://doi.org/10.5351/KJAS.2021.34.5.773
Print ISSN1225-066X / Online ISSN 2383-5818

Functional regression approach to traffic analysis

Injoo Lee[a], Young K. Lee[1,a]

[a]Department of Statistics, Kangwon National University

Abstract

Prediction of vehicle traffic volume is very important in planning municipal administration. It may help promote social and economic interests and also prevent traffic congestion costs. Traffic volume as a time-varying trajectory is considered as functional data. In this paper we study three functional regression models that can be used to predict an unseen trajectory of traffic volume based on already observed trajectories. We apply the methods to highway tollgate traffic volume data collected at some tollgates in Seoul, Chuncheon and Gangneung. We compare the prediction errors of the three models to find the best one for each of the three tollgate traffic volumes.

Keywords: auto-covariance, cross-covariance, functional principal components, functional singular components, smooth backfitting

1. 서론

정보통신 기술의 발달로 많은 양의 데이터들을 수집, 저장, 관리하기가 용이해지고 이를 활용한 교통정보 분석 및 활용이 가능해졌다. 다양한 교통 데이터를 이용해 실생활과 관련된 교통 연구들이 활발히 이루어지고 있으며, 수년간 축적된 교통 데이터를 이용하여 교통량 패턴을 분석할 수 있고 교통량 예측을 통해 사회적, 경제적 이익을 창출할 수 있다 (An, 2017).

사회적, 경제적 이익을 창출하는 첫 번째 방법은 교통혼잡비용의 해결책을 마련하는 것이다. 교통혼잡비용은 교통량이 늘어나면서 차량 정체 등 교통혼잡으로 발생하는 사회적 손실 규모를 화폐가치로 환산한 지표로, 1994년 10조 원으로 발표된 이후 2012년을 기점으로 30조 원을 넘어섰으며 2016년 55.9조 원으로 추정되었다. Park (2015)의 연구에 따르면 교통혼잡비용을 해결하기 위해 이미 구축된 시설물을 최대한 활용하는 것이 효율적이며 이를 위해 교통 제어 및 교통 수요 분산을 위한 전략으로 미래 교통상황에 대한 예측 기술 개발의 필요성에 대해 언급하였다. 유입교통량의 증가는 상습정체를 발생시키고 교통사고에 대한 위험도 증가시키므로 실시간 정보에 기반을 둔 사후적인 교통관리 개념에서 벗어나 교통상황 예측 기술 개발을 이용해 사전에 교통량 분산과 정체 해소 등의 교통 환경을 개선함으로써 도로 교통 운영 및 관리를 최적화할 수 있을 것이라 기대된다 (Park, 2015).

두 번째 방법은 교통량 수요예측의 오차를 줄이는 것이다. 교통 기반시설에 대한 투자는 국가 예산의 사용에 대한 의사결정에 매우 중요한 부분을 차지하고 있기 때문에, 적절한 투자가 이루어지기 위해서는 교통수요에 일어나고 있는 변화를 최대한 정확히 감지하여 미래 양상을 예측하는 것이 중요하다 (Yoon 등, 2017).

This Research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2021R1A2C1003920).

[1] Corresponding author: Department of Statistics, Kangwon National University, 1 Gangwondaehakgil, Chuncheon-si, Gangwon-do 24341, Korea. E-mail: youngklee@kangwon.ac.kr

Published 31 October 2021 / journal homepage: http://kjas.or.kr

따라서 교통량 예측 기술 개발을 통해 교통량 수요예측의 정확도를 높인다면 경제적 손실을 줄일 수 있을 것이다.

또한, 영업소 출구 교통량은 해당 지역으로 유입되는 방문객 수와 연관이 있으며 이는 지역 발전에도 영향을 미친다. Sohn과 Kim (2014)의 연구에 따르면 도로공사에서 제공하는 고속도로 영업소 유입 교통량은 관광지로 유입되는 교통량에 대한 가장 신뢰할만한 정보이고 문화체육관광부의 관광지 입장객 수 정보와 강한 상관관계를 가진다고 언급하였다. 이를 통해 관광지의 주말 교통량을 관광객 유입량으로 이용하여 지역 경제 활성화를 위한 방안으로도 활용될 수 있다.

지금까지 교통량 추정에 관한 연구들은 주로 평일 교통량에 중점을 두고 있으며 평일 교통량을 이용하여 주말 통행량을 예측하는 연구는 제대로 이루어지지 않았다. Hunt 등 (2005)의 주말 통행량 연구도 주로 평일과 주말의 단순 통행 특성을 비교하는 데 그치고 있어 구체적인 주말 교통량 예측의 필요성이 대두되고 있다. 특히, Choo 등 (2007)의 연구에서는 국내에서 실시된 주말 통행 실태와 교통량 조사 자료를 이용한 평일 통행량과의 비교분석을 통해 주말 통행 특성을 파악하였으며, 주말 교통량 추정에 관한 연구 활성화의 중요성을 언급하였다. 평일 교통량 패턴은 해당 주의 주말 교통량 패턴을 가늠하는데 중요한 정보를 제공할 수 있기 때문에 평일 교통량 정보를 활용하여 주말 교통량의 패턴을 예측하는 것이 유효한지 살펴보는 것이 필요하다.

본 연구에서는 한국도로공사에서 제공하는 영업소별 교통량 데이터를 활용하여 평일 교통량과 주말 교통량 사이의 관계를 함수적 회귀분석 방법을 통해 파악하고자 한다. 본 연구의 목적은 평일 교통량을 이용해 주말 교통량 패턴을 예측하는 효율적 방법론을 제시하는 것이 목적이고, 평일 교통량은 월요일–목요일, 월요일–금요일, 월요일, 화요일, 수요일, 목요일, 금요일 7가지 경우로 하여 주말 교통량인 토요일과 일요일 패턴을 예측하였다. 함수적 회귀분석 방법으로는 함수가법모형(functional additive model), 특이가법모형(singular additive model), 특이선형모형(singular linear model)의 세 가지 방법을 이용하였으며, 방법론별 예측 결과를 상대적분제곱예측오차(relative integrated squared prediction error)로 비교 평가하였다.

2. 방법론

평일 또는 주말 교통량은 어떤 시구간에서 연속적으로 관측되는 함수데이터(functional data)이다. 함수데이터는 기본적으로 무한차원의 데이터이며 이를 유한차원의 데이터로 차원을 축소한 후 분석할 수 있다. 이는 다변량데이터(multivariate data)의 차원 축소와 궤를 같이 한다. 시구간 $\mathcal{S} \subset \mathbb{R}$에서 정의된 공변과정(covariate process)을 $\{X(s) : s \in \mathcal{S}\}$, 그리고 $\mathcal{T} \subset \mathbb{R}$에서 정의된 반응과정(response process)을 $\{Y(t) : t \in \mathcal{T}\}$라 하고 다음과 같이 제곱적분 가능하다고 하자.

$$\mathrm{E}\left(\|X\|^2\right) = \mathrm{E}\left(\int_{\mathcal{S}} X(s)^2\, ds\right) < \infty, \quad \mathrm{E}\left(\|Y\|^2\right) = \mathrm{E}\left(\int_{\mathcal{T}} Y(t)^2\, dt\right) < \infty.$$

평일교통량을 이용하여 주말교통량을 예측할 때 공변과정은 평일교통량이고 반응과정은 주말교통량이 된다. 이 절에서는 함수데이터의 차원 축소 방법인 함수주성분분석(functional principal component analysis)과 함수특이성분분석(functional singular component analysis)을 우선 소개하기로 한다. 다음으로 평일교통량과 주말교통량 사이의 관계를 설정하는 세가지 모형을 소개하고 이들의 추정을 통한 예측 방법론을 설명하기로 한다.

2.1. 함수주성분분석

함수주성분분석(functional principal component analysis)은 다변량데이터(multivariate data)의 주성분분석(principal component analysis)처럼 함수데이터(functional data)의 주요 변동 요인을 찾아서 이들을 통해 데이터의 변동을 설명하는 것이다. 다변량 회귀분석에서처럼 함수주성분분석을 통해 얻은 유한개의 주성분에 기반하

여 무한차원의 원 데이터를 유한차원으로 근사시키고 이를 함수회귀분석(functional regression analysis)에 활용할 수 있다.

함수주성분분석의 주요 요소는 평균함수(mean function), 자기공분산함수(auto-covariance function) 그리고 자기공분산작용소(auto-covariance operator) 등이다. 공변과정 X와 반응과정 Y의 평균함수들은 각각

$$\mu_X(s) = \mathrm{E}(X(s)), \quad \mu_Y(t) = \mathrm{E}(Y(t))$$

와 같이 정의되며, 이들의 자기공분산함수는 각각

$$C_{XX}(s,s') = \mathrm{E}(X^c(s)X^c(s')), \quad C_{YY}(t,t') = \mathrm{E}(Y^c(t)Y^c(t'))$$

으로 정의된다. 여기서, $X^c(s) = X(s) - \mu_X(s)$이고 $Y^c(t) = Y(t) - \mu_Y(t)$이다. 또한, 이들의 자기공분산작용소 $\mathscr{C}_{XX} : L^2(\mathcal{S}) \to L^2(\mathcal{S})$와 $\mathscr{C}_{YY} : L^2(\mathcal{T}) \to L^2(\mathcal{T})$는 각각

$$\mathscr{C}_{XX}(f)(s) = \int_{\mathcal{S}} C_{XX}(s,s')f(s')\,ds', \quad \mathscr{C}_{YY}(g)(t) = \int_{\mathcal{T}} C_{YY}(t,t')g(t')\,dt' \tag{2.1}$$

와 같이 정의된다. 여기서, $L^2(\mathcal{S})$와 $L^2(\mathcal{T})$는 각각 시구간 $\mathcal{S}$와 $\mathcal{T}$ 위에서 정의된 제곱적분가능한 함수들의 공간을 뜻한다.

이제 $\mathscr{C}_{XX}$의 고유값을 $\lambda_{X,1} \geq \lambda_{X,2} \geq \cdots$라 하고 그에 대응되는 고유함수(eigenfunctions)를 $\phi_{X,j},\ j \geq 1$이라 하자. 즉,

$$\mathscr{C}_{XX}(\phi_{X,j}) = \lambda_{X,j}\phi_{X,j}, \quad j \geq 1.$$

마찬가지로 $\mathscr{C}_{YY}$의 순서고유값과 그에 대응되는 고유함수를 $\lambda_{Y,k}$와 $\phi_{Y,k}$라 하자. 자기공분산함수는 양의 정부호(positive definite)이고 대칭(symmetric)이기 때문에 자기공분산작용소는 양(positive)이고 자기수반(self-adjoint)이다. 따라서, Mercer의 정리를 적용할 수 있다. Mercer의 정리 (Hsing과 Eubank, 2015)에 의하면 정규직교(orthonormal) 고유함수 $\phi_{X,j}$, $\phi_{Y,k}$가 존재하고 공분산함수 C_{XX}와 C_{YY}에 대해 다음과 같은 스펙트럼분해(spectral decomposition)가 성립한다.

$$C_{XX}(s,s') = \sum_{j\geq 1} \lambda_{X,j}\phi_{X,j}(s)\phi_{X,j}(s'), \quad C_{YY}(t,t') = \sum_{k\geq 1} \lambda_{Y,k}\phi_{Y,k}(t)\phi_{Y,k}(t'). \tag{2.2}$$

또한, Karhunen-Loève 정리 (Hsing과 Eubank, 2015)에 따르면 X와 Y는

$$X^c(s) = \sum_{j=1}^{\infty} \xi_{X,j}\phi_{X,j}(s), \quad Y^c(t) = \sum_{k=1}^{\infty} \xi_{Y,k}\phi_{Y,k}(t) \tag{2.3}$$

와 같이 전개될 수 있다. 여기서,

$$\xi_{X,j} = \int_{\mathcal{S}} X^c(s)\phi_{X,j}(s)\,ds, \quad \xi_{Y,k} = \int_{\mathcal{T}} Y^c(t)\phi_{Y,k}(t)\,dt$$

이며 이들을 함수주성분(Functional principal components, FPCA)이라 부른다. 함수주성분들은 평균이 0이고 서로 무상관(uncorrelated)이며 분산은 해당 고유값이 된다. 즉,

$$\begin{aligned} &\mathrm{E}(\xi_{X,j}) = 0, \quad \mathrm{E}(\xi_{X,j}\xi_{X,j'}) = 0 \text{ for } j \neq j', \quad \mathrm{E}(\xi_{X,j}^2) = \lambda_{X,j}, \\ &\mathrm{E}(\xi_{Y,k}) = 0, \quad \mathrm{E}(\xi_{Y,k}\xi_{Y,k'}) = 0 \text{ for } k \neq k', \quad \mathrm{E}(\xi_{Y,k}^2) = \lambda_{Y,k}. \end{aligned} \tag{2.4}$$

따라서, 함수주성분전개식 (2.3)은 X와 Y를 각각 함수공간 $L^2(\mathcal{S})$와 $L^2(\mathcal{T})$에서 서로 직교이고 무상관인 변동의 요인으로 분해하는 것으로 이해할 수 있다. 다시 말해서,

$$\mathrm{E}\left(\|X^c\|^2\right) = \sum_{j=1}^{\infty} \mathrm{E}\left(\|\xi_{X,j}\phi_{X,j}\|^2\right) = \sum_{j\geq 1} \lambda_{X,j}, \quad \mathrm{E}\left(\|Y^c\|^2\right) = \sum_{k=1}^{\infty} \mathrm{E}\left(\|\xi_{Y,k}\phi_{Y,k}\|^2\right) = \sum_{k\geq 1} \lambda_{Y,k}.$$

위 식과 아래 논의에서 우리는 $L^2(\mathcal{S})$와 $L^2(\mathcal{T})$의 노름(norm)을 구분하지 않고 모두 $\|\cdot\|$으로 표기한다. 확률과정 X와 Y의 변동이 각각의 고유값의 합으로 표현됨은 위 식 (2.2)로부터도 유도할 수 있다. 즉,

$$\mathrm{E}\left(\|X^c\|^2\right) = \int_{\mathcal{S}} C_{XX}(s,s)\,ds = \sum_{j\geq 1} \lambda_{X,j}, \quad \mathrm{E}\left(\|Y^c\|^2\right) = \int_{\mathcal{T}} C_{YY}(t,t)\,dt = \sum_{k\geq 1} \lambda_{Y,k}.$$

만약 처음 M_X와 M_Y개의 X와 Y의 주성분에 대하여,

$$\frac{\sum_{j=1}^{M_X} \lambda_{X,j}}{\sum_{j=1}^{\infty} \lambda_{X,j}} \simeq 1, \quad \frac{\sum_{k=1}^{M_Y} \lambda_{Y,k}}{\sum_{k=1}^{\infty} \lambda_{Y,k}} \simeq 1$$

이면,

$$X^c(s) \simeq \sum_{j=1}^{M_X} \xi_{X,j}\phi_{X,j}(s), \quad Y^c(t) \simeq \sum_{k=1}^{M_Y} \xi_{Y,k}\phi_{Y,k}(t)$$

와 같이 X와 Y 과정을 유한개의 주성분으로 근사시킬 수 있다. 함수주성분분석에 관한 보다 자세한 내용은 Hsing과 Eubank (2015)를 참고하기 바란다. 고유값 $\lambda_{X,j}$와 $\lambda_{Y,k}$는 미지이므로 실제 자료 분석에서는 식 (2.17)의 표본자기공분산함수 $\hat{C}_{XX}$와 $\hat{C}_{YY}$의 함수주성분분석을 통하여 구한 표본고유값 $\hat{\lambda}_{X,j}$와 $\hat{\lambda}_{Y,k}$을 사용하여 선택할 수 있다. 함수회귀모형의 추정 시에는 2.3절에서와 같이 적절한 교차검증기준(cross-validation criterion)에 근거하여 M_X와 M_Y를 선택할 수 있다.

2.2. 함수특이성분분석

앞의 2.1절에서 소개한 함수주성분분석은 X와 Y 각각의 변동을 잘 설명하는 함수주성분을 추출하는 방법으로 두 확률과정의 관계를 잘 설명하는 주요 성분을 추출하는데 효과적이 아닐 수 있다. 공변과정 X와 Y의 함수적 연관성을 설명하고자 하는 함수회귀분석(functional regression analysis)에서는 두 확률과정의 관계를 잘 설명하는 주요 성분을 추출하는 것이 필요하며 여기에는 이 절에서 소개하는 함수특이성분분석(functional singular component analysis; Yang 등, 2011)이 유용하다.

두 확률과정 X와 Y의 함수특이성분분석은 자기공분산함수 대신에 교차공분산함수(cross-covariance function)을 사용한다. 두 확률과정 X와 Y의 교차공분산함수 C_{XY}와 C_{YX}는

$$C_{XY}(s,t) = \mathrm{E}(X^c(s)Y^c(t)) = C_{YX}(t,s)$$

로 정의된다. 교차공분산함수는 자기공분산함수와 달리 대칭(symmetric)함수가 아니다. 즉, $C_{XY}(s,t) \neq C_{XY}(t,s)$ 그리고 $C_{YX}(t,s) \neq C_{YX}(s,t)$일 수 있다. 또한, 자기공분산함수와 달리 양의 정부호가 아니다. 이들의 교차공분산함수로 부터 교차공분산작용소(cross-covariance operators) $\mathscr{C}_{XY} : L^2(\mathcal{T}) \to L^2(\mathcal{S})$와 $\mathscr{C}_{YX} : L^2(\mathcal{S}) \to L^2(\mathcal{T})$는 각각

$$\mathscr{C}_{XY}(g)(s) = \int_{\mathcal{T}} C_{XY}(s,t)g(t)\,dt, \quad \mathscr{C}_{YX}(f)(t) = \int_{\mathcal{S}} C_{YX}(t,s)f(s)\,ds \tag{2.5}$$

로 정의된다. 교차공분산작용소는 교차공분산함수가 양의 정부호가 아니기 때문에 양의 작용소(positive operator)가 될 수 없다. 또한, 교차공분산작용소는 자기수반이 아니다. 따라서, 자기공분산함수와 자기공분산작용소의 경우처럼 Mercer의 정리를 적용할 수 없다.

교차공분산작용소 $\mathscr{C}_{YX}$는 $\mathscr{C}_{YX}$의 수반(adjoint)이다. 즉,

$$\langle \mathscr{C}_{YX}(f), g\rangle = \int_{\mathcal{T}} \mathscr{C}_{YX}(f)(t)\cdot g(t)\,dt = \int_{\mathcal{S}} f(s)\cdot \mathscr{C}_{XY}(g)(s)\,ds = \langle f, \mathscr{C}_{XY}(g)\rangle.$$

노름의 경우와 마찬가지로 위 식과 아래 논의에서 $L^2(\mathcal{S})$와 $L^2(\mathcal{T})$의 내적(inner product)은 구분하지 않고 모두 $\langle\cdot,\cdot\rangle$로 표기한다. 교차공분산작용소 $\mathscr{C}_{YX}$와 $\mathscr{C}_{YX}$로부터 이들의 합성작용소(compound operators) $\mathscr{C}_{XYX} : L^2(\mathcal{S}) \to L^2(\mathcal{S})$와 $\mathscr{C}_{YXY} : L^2(\mathcal{T}) \to L^2(\mathcal{T})$를 다음과 같이 정의하자.

$$\begin{aligned}
\mathscr{C}_{XYX}(f)(s) &= (\mathscr{C}_{XY}\circ\mathscr{C}_{YX}(f))(s) = \int_{\mathcal{S}} C_{XY}(s,t)\cdot\mathscr{C}_{YX}(f)(t)\,dt = \int_{\mathcal{S}}\int_{\mathcal{T}} C_{XY}(s,t)C_{YX}(t,s')f(s')\,ds'\,dt,\\
\mathscr{C}_{YXY}(g)(t) &= (\mathscr{C}_{YX}\circ\mathscr{C}_{XY}(g))(t) = \int_{\mathcal{T}} C_{YX}(t,s)\cdot\mathscr{C}_{XY}(g)(s)\,ds = \int_{\mathcal{S}}\int_{\mathcal{T}} C_{YX}(t,s)C_{XY}(s,t')g(t')\,ds\,dt'. \qquad (2.6)
\end{aligned}$$

그러면, $\mathscr{C}_{XYX}$와 $\mathscr{C}_{YXY}$은 자기수반이며 각각 다음에 정의하는 양의 정부호 이원함수(bivariate functions) C_{XYX}와 C_{YXY}를 커널(kernel)로 갖는 양의 작용소가 된다.

$$\begin{aligned}
\mathscr{C}_{XYX}(f)(s) &= \int_{\mathcal{S}} C_{XYX}(s,s')f(s')\,ds', \quad C_{XYX}(s,s') = \int_{\mathcal{T}} C_{XY}(s,t)C_{YX}(t,s')\,dt,\\
\mathscr{C}_{YXY}(g)(t) &= \int_{\mathcal{T}} C_{YXY}(t,t')g(t')\,dt', \quad C_{YXY}(t,t') = \int_{\mathcal{S}} C_{YX}(t,s)C_{XY}(s,t')\,ds.
\end{aligned}$$

또한, $\mathscr{C}_{XYX}$와 $\mathscr{C}_{YXY}$은 Hilbert-Schmidt 작용소이다 (Kato, 1995).

Hilbert-Schmidt 작용소에 대한 스팩트럼정리에 의하면 $\mathscr{C}_{XYX}$와 $\mathscr{C}_{YXY}$은 고유값 $\sigma_1^2 \ge \sigma_2^2 \ge \cdots \ge 0$을 공유하며 각각 고유함수 $\psi_{X,1}, \psi_{X,2}, \ldots,$ 와 $\psi_{Y,1}, \psi_{Y,2}, \ldots,$을 가진다. 여기서 고유함수 짝 $(\psi_{X,1}, \psi_{Y,1})$는 $\|u\| = \|v\| = 1$의 제약조건 하에서 $\langle u, X^c\rangle$와 $\langle v, Y^c\rangle$의 공분산을 최대로 하는 (u,v)이고 그 최대값이 바로 σ_1이 된다. 그리고 $j \ge 2$에 대하여 고유함수 짝 $(\psi_{X,j}, \psi_{Y,j})$는 $\|u\| = \|v\| = 1$과 $\langle u, \psi_{X,k}\rangle = \langle v, \psi_{Y,k}\rangle = 0$, $k \le j-1$의 제약조건 하에서 $\langle u, X^c\rangle$와 $\langle v, Y^c\rangle$의 공분산을 최대로 하는 (u,v)이고 그 최대값이 바로 σ_j이다. 즉, 고유함수 짝들 $(\psi_{X,j}, \psi_{Y,j})$는 확률과정 X와 Y의 상관성을 잘 설명하는 성분을 $\langle\psi_{X,j}, X^c\rangle$와 $\langle\psi_{Y,j}, Y^c\rangle$의 공분산을 통하여 크기 순으로 차례대로 추출한 것이다. 이들 고유함수 $(\psi_{X,j}$와 $\psi_{Y,j})$를 각각 X와 Y의 특이함수(singular functions)라 하고 고유값의 제곱근인 σ_j를 특이값(singular values)라 부른다. 또한,

$$\zeta_{X,j} := \langle\psi_{X,j}, X^c\rangle = \int_{\mathcal{S}} \psi_{X,j}(s)X^c(s)\,ds, \quad \zeta_{Y,j} := \langle\psi_{Y,j}, Y^c\rangle = \int_{\mathcal{T}} \psi_{Y,j}(t)Y^c(t)\,dt$$

를 각각 X와 Y의 함수특이성분(functional singular components)라 한다. 이들 함수특이성분들은 X와 Y의 상관성을 잘 대표하는 확률변수로서 다음을 만족한다.

$$\begin{aligned}
&\mathrm{E}(\zeta_{X,j}) = 0,\ j \ge 1 \quad \text{and} \quad \mathrm{E}(\zeta_{Y,k}) = 0,\ k \ge 1,\\
&|\mathrm{E}(\zeta_{X,j}\zeta_{Y,j})| = \sigma_j,\ j \ge 1 \quad \text{and} \quad \mathrm{E}(\zeta_{X,j}\zeta_{Y,k}) = 0,\ j \ne k. \qquad (2.7)
\end{aligned}$$

여기서 주목할 점은 함수주성분과 달리 서로 다른 X의 함수특이성분, 그리고 서로 다른 Y의 특이성분들은 무상관이 아니라는 것이다. 즉, $j \ne k$에 대하여 $\mathrm{E}(\zeta_{X,j}\zeta_{X,k}) \ne 0$이고 $\mathrm{E}(\zeta_{Y,j}\zeta_{Y,k}) \ne 0$일 수 있다.

한편, Hilbert-Schmidt 정리 (Yang 등, 2011)에 의하면 다음이 성립한다.

$$X^c(s) = \sum_{j:\sigma_j>0} \zeta_{X,j}\psi_{X,j}(s) + \nu_X(s), \quad Y^c(t) = \sum_{k:\sigma_k>0} \zeta_{Y,k}\psi_{Y,k}(t) + \nu_Y(t). \qquad (2.8)$$

위 전개식이 함수주성분분석의 전개식 (2.3)과 다른 점은 잔여과정(remainder processes)인 ν_X와 ν_Y을 포함한다는 사실이다. 이들 잔여과정들은 X와 Y의 교차공분산으로 설명되지 않은 부분이며, 결과적으로 특이함수들의 집합체인 $(\psi_{X,j} : \sigma_j > 0, j \geq 1)$와 $(\psi_{Y,k} : \sigma_k > 0, k \geq 1)$이 각각 $L^2(\mathcal{S})$와 $L^2(\mathcal{T})$의 완비기저(complete basis)가 아님을 의미한다. 이들 잔여과정들은 다음을 만족시킨다.

$$\int_{\mathcal{S}}\int_{\mathcal{T}} (\mathrm{E}(\nu_X(s)\nu_Y(t)))^2\, ds\, dt = 0,\ \int_{\mathcal{S}} \left(\mathrm{E}(\zeta_{Y,j}\nu_X(s))\right)^2\, ds = 0,\ \int_{\mathcal{T}} \left(\mathrm{E}(\zeta_{X,j}\nu_Y(t))\right)^2\, dt = 0, \quad j \geq 1. \tag{2.9}$$

즉, ν_X와 ν_Y가 무상관이고, ν_X는 Y의 특이성분 $\zeta_{Y,j}$들과 그리고 ν_Y는 X의 특이성분 $\zeta_{X,j}$들과 교차 무상관이다. 서로 다른 X의 함수특이성분, 그리고 서로 다른 Y의 특이성분들은 무상관이 아니라는 점, 그리고 함수특이성분전개식 (2.8)이 잔여과정 ν_X와 ν_Y를 포함한다는 사실은 함수특이성분을 사용한 함수회귀분석의 방법론이 함수주성분을 사용할 때와 차별화 되어야 함을 의미한다. 자세한 내용은 2.3절과 2.5절에서 논의하기로 한다.

2.3. 함수가법모형

반응과정 Y의 함수주성분전개식 (2.3)으로부터

$$\mathrm{E}(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{\infty} \mathrm{E}(\xi_{Y,k}|X) \cdot \phi_{Y,k}(t) \tag{2.10}$$

이다. 함수가법모형(functional additive model)은 Müller와 Yao (2008)에 의해 처음 도입되었다. 이 모형은 각 $k \geq 1$에 대하여 $\mathrm{E}(\xi_{Y,k}|X)$을 X의 함수주성분 $\xi_{X,j}$의 가법함수 형태로 모델링하는 것이다. 즉, 다음을 가정한다.

$$\mathrm{E}(\xi_{Y,k}|X) = f_{k,1}(\xi_{X,1}) + \cdots + f_{k,M}(\xi_{X,M}), \quad k \geq 1. \tag{2.11}$$

위 모형은 X의 모든 주성분을 고려하지 않고 적절한 M을 선택하여 처음 M개의 주성분만을 사용하였다. 이는 실제 적용에서 피할 수 없는 과정이다. 마찬가지로 가법모형 (2.11)을 식 (2.10)에 대입하여 회귀함수 $\mathrm{E}(Y|X = \cdot)$을 추정할 때에도 Y의 모든 주성분이 아닌 적절한 개수의 주성분 $\xi_{Y,k}$ 이후는 절단시킨다. 이 때 모형에 포함시킬 Y의 주성분의 개수와 X의 주성분의 개수는 다를 수 있으나 이 논문에서는 같은 값으로 택하는 경우를 생각한다. 단, M은 미리 정해진 것이 아니라 미지의 값으로 두고 데이터가 선택하도록 한다.

본 논문에서 제시하는 함수가법모형은 다음과 같다.

$$\mathrm{E}(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{M} (f_{k,1}(\xi_{X,1}) + \cdots + f_{k,M}(\xi_{X,M})) \cdot \phi_{Y,k}(t). \tag{2.12}$$

위 모형에서 성분함수 $f_{k,j}, 1 \leq j \leq M$들은 식별가능(identifiable)하지 않다. 반응과정 Y의 함수주성분이 $\mathrm{E}(\xi_{Y,k}) = 0$을 만족시키므로 각 성분함수에 대하여 제약조건

$$\mathrm{E}(f_{k,j}(\xi_{X,j})) = 0, \quad 1 \leq j \leq M, \quad 1 \leq k \leq M \tag{2.13}$$

을 주면 모든 성분함수는 식별가능해진다. 성분함수 $f_{k,j}$의 추정을 위해 식 (2.11)로부터 다음이 성립함에 주목한다.

$$\mathrm{E}(\xi_{Y,k}|\xi_{X,j}) = f_{k,j}(\xi_{X,j}) + \sum_{j'=1,\neq j} \mathrm{E}\left(f_{k,j'}(\xi_{X,j'})|\xi_{X,j}\right). \tag{2.14}$$

그런데, 식 (2.4)으로부터 $j \neq j'$에 대하여 $\xi_{X,j}$와 $\xi_{X,j'}$은 무상관이다. 본 논문에서는 무상관보다 더 강한 조건인 독립(independence)를 가정한다. 실제로 X가 가우시안과정(Gaussian process)이면 $j \neq j'$에 대하여 $\xi_{X,j}$와 $\xi_{X,j'}$은 서로 독립이다. 이러한 독립 가정 하에서는

$$\mathrm{E}\left(f_{k,j'}(\xi_{X,j'})|\xi_{X,j}\right) = \mathrm{E}\left(f_{k,j'}(\xi_{X,j'})\right) = 0$$

이 성립하여 성분함수 $f_{k,j}$는 하나의 공변량이 있는 것처럼 일변량(univariate)회귀함수의 비모수추정법으로 추정할 수 있다. 예를 들면, 나다라야-왓슨(Nadaraya-Watson)추정법을 사용하여 $f_{k,j}$를 다음과 같이 추정할 수 있다.

$$\tilde{f}_{k,j}(u) = \frac{\sum_{i=1}^{n} K_h(\xi_{X,ij} - u)\xi_{Y,ik}}{\sum_{i=1}^{n} K_h(\xi_{X,ij} - u)}. \tag{2.15}$$

여기서, $K_h(v) := h^{-1}K(h^{-1}v)$는 콤팩트(compact) 토대(support)를 가지면서 $K \geq 0$, $\int K = 1$인 대칭함수로서 기저커널(baseline kernel)이고, $h > 0$은 추정량의 평활도와 편의(bias)를 결정하는 밴드폭(bandwidth)이다. 기저커널의 선택은 추정량의 성능에 크게 영향을 미치지 않는 것으로 알려져 있다. 본 논문의 3장 교통량 분석에서는 Epanechnikov 커널

$$K(u) = \frac{3}{4}\left(1 - u^2\right) I(|u| \leq 1)$$

을 사용하였다. 또한, $\xi_{X,ij}$와 $\xi_{Y,ik}$는 각각 i번째 관측된 과정 X_i와 Y_i에 대한 함수주성분이다. 즉,

$$\xi_{X,ij} = \int_{\mathcal{S}} X_i^c(s)\phi_{X,j}(s)\,ds, \quad \xi_{Y,ik} = \int_{\mathcal{T}} Y_i^c(t)\phi_{Y,k}(t)\,dt. \tag{2.16}$$

나다라야-왓슨(Nadaraya-Watson)추정량은 $\xi_{X,ij}$와 $\xi_{Y,ik}$의 값을 필요로 한다. 그러나, 이들은 위 식 (2.16)에서 보는 바와 같이 미지의 고유함수 $\phi_{X,j}$와 $\phi_{Y,k}$에 의존하고 있다. 실제 적용 가능한 나다라야-왓슨 추정량의 정의는 다음과 같이 자기공분산함수 C_{XX}와 C_{YY}를 추정하는 것으로부터 시작한다.

$$\begin{aligned}
\hat{C}_{XX}(s,s') &= n^{-1}\sum_{i=1}^{n}\left(X_i(s) - \bar{X}(s)\right)\left(X_i(s') - \bar{X}(s')\right), \\
\hat{C}_{YY}(t,t') &= n^{-1}\sum_{i=1}^{n}\left(Y_i(t) - \bar{Y}(t)\right)\left(Y_i(t') - \bar{Y}(t')\right).
\end{aligned} \tag{2.17}$$

여기서 $\bar{X}(s) = n^{-1}\sum_{i=1}^{n} X_i(s)$이고 $\bar{Y}(t) = n^{-1}\sum_{i=1}^{n} Y_i(t)$이다. 위 자기공분산함수의 추정량으로부터 자기공분산 작용소의 추정량 $\hat{\mathscr{C}}_{XX}$와 $\hat{\mathscr{C}}_{YY}$를 식 (2.1)에서 처럼 정의하고 이들의 스펙트럼분해로부터 고유함수 $\phi_{X,j}$와 $\phi_{Y,k}$의 추정량 $\hat{\phi}_{X,j}$와 $\hat{\phi}_{Y,k}$을 구한다. 함수주성분 $\xi_{X,ij}$와 $\xi_{Y,ik}$의 추정량은 다음과 같이 정의된다.

$$\hat{\xi}_{X,ij} = \int_{\mathcal{S}}\left(X_i(s) - \bar{X}(s)\right)\hat{\phi}_{X,j}(s)\,ds, \quad \hat{\xi}_{Y,ik} = \int_{\mathcal{T}}\left(Y_i(t) - \bar{Y}(t)\right)\hat{\phi}_{Y,k}(t)\,dt.$$

실제 사용 가능한 성분함수 $f_{k,j}$의 나다라야-왓슨 추정량은 식 (2.15)에서 $\xi_{X,ij}$와 $\xi_{Y,ik}$ 대신에 $\hat{\xi}_{X,ij}$와 $\hat{\xi}_{Y,ik}$을 대입하여 다음과 같이 정의한다.

$$\hat{f}_{k,j}(u) = \frac{\sum_{i=1}^{n} K_h\left(\hat{\xi}_{X,ij} - u\right)\hat{\xi}_{Y,ik}}{\sum_{i=1}^{n} K_h\left(\hat{\xi}_{X,ij} - u\right)}.$$

또한, 함수가법모형 (2.12)의 추정은 μ_Y 대신에 $\bar{Y}$를 대입하고 $f_{k,j}$ 대신에 $\hat{f}_{k,j}$를 대입하면 된다. 즉,

$$\hat{\mathrm{E}}(Y(t)|X) = \bar{Y}(t) + \sum_{k=1}^{M}\left(\hat{f}_{k,1}(\xi_{X,1}) + \cdots + \hat{f}_{k,M}(\xi_{X,M})\right)\cdot\hat{\phi}_{Y,k}(t). \tag{2.18}$$

위 추정모형을 활용하여 미래의 값 Y_{new}에 대응하는 X_{new}을 알 때, 다음과 같이 Y_{new}를 예측할 수 있다.

$$\hat{Y}_{\text{new}}(t) = \bar{Y}(t) + \sum_{k=1}^{M}\left(\hat{f}_{k,1}(\xi_{X,1,\text{new}}) + \cdots + \hat{f}_{k,M}(\xi_{X,M,\text{new}})\right)\cdot\hat{\phi}_{Y,k}(t). \tag{2.19}$$

여기서,

$$\xi_{X,j,\text{new}} = \int_{\mathcal{S}} \left(X_{\text{new}}(s) - \bar{X}(s)\right) \hat{\phi}_{X,j}(s)\, ds$$

이며, $\bar{X}$, $\bar{Y}$, $\hat{f}_{k,j}$, $\hat{\phi}_{X,j}$, $\hat{\phi}_{Y,k}$, $1 \le j,k \le M$ 등은 모두 $\{(X_i, Y_i) : 1 \le i \le n\}$을 이용하여 구한다.

모형 추정에 사용할 주성분의 개수 M을 선택하는 방법은 여러가지가 있을 수 있으나 여기에서는 3장의 교통량 데이터 분석에 사용한 5겹 교차검증(five-fold cross-validation)을 소개하기로 한다. 훈련자료 $\{(X_i, Y_i) : 1 \le i \le n\}$를 5등분하여 각각을 $\{(X_i, Y_i) : i \in I_r\}$, $1 \le r \le 5$ 이라 하자. 주성분의 개수 M을 사용하고 r번째 폴드(fold)를 제외한 훈련자료를 이용하여 모형을 추정하고 이를 바탕으로 Y_i, $i \in I_r$를 예측한 값을 $\hat{Y}_{i,M}^{(-r)}$이라 하자. 주성분의 개수 M의 선택을 위해 최소화하는 5겹 교차검증기준은 다음과 같다.

$$CV(M) = \frac{1}{n} \sum_{r=1}^{5} \sum_{i \in I_r} \left(Y_i - \hat{Y}_{i,M}^{(-r)}\right)^2. \tag{2.20}$$

2.4. 특이가법모형

특이가법모형(singular additive model)은 기본적으로 함수가법모형과 유사하나 함수주성분 대신에 특이성분을 활용한 모형이다. 이 모형은 Park 등 (2018)에 의해 처음 도입되었다. 모형의 도입을 위해 교차공분산작용소 $\mathscr{C}_{XY}$와 $\mathscr{C}_{YX}$가 유한랭크(finite rank)라면, 즉 $\sigma_j = 0$, $j \ge M+1$이라면 함수특이성분전개식 (2.8)은

$$X(s) = \mu_X(s) + \sum_{j=1}^{M} \zeta_{X,j} \psi_{X,j}(s) + \nu_X(s), \quad Y(t) = \mu_Y(t) + \sum_{k=1}^{M} \zeta_{Y,k} \psi_{Y,k}(t) + \nu_Y(t) \tag{2.21}$$

으로 축소된다. 따라서,

$$\mathrm{E}(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{M} \mathrm{E}(\zeta_{Y,k}|X) \cdot \psi_{Y,k}(t) + \mathrm{E}(\nu_Y(t)|X) \tag{2.22}$$

을 얻는다. 식 (2.2)의 논의에 의하면 위 식 (2.22)에서 ν_Y는 X의 특이성분인 $\zeta_{X,j}$들 그리고 ν_X와 무상관이므로 X와 무상관이라 할 수 있다. 본 논문에서는 무상관보다 강한 독립을 가정한다. 이 가정 하에서는 $\mathrm{E}(\nu_Y(t)|X) = \mathrm{E}(\nu_Y(t)) = 0$이다. 즉,

$$\mathrm{E}(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{M} \mathrm{E}(\zeta_{Y,k}|X) \cdot \psi_{Y,k}(t) \tag{2.23}$$

를 얻는다. 또한, X의 특이성분전개식 (2.21)에서 등장하는 ν_X는 관측할 수도 없고 교차공분산분석으로부터 추정할 수도 없는 항으로 변수내오차 문제(errors-in-variables problem)에서 측정오차(measurement error)와 같은 속성을 가진다. 따라서, 변수내오차 문제에서처럼 $\mathrm{E}(\zeta_{Y,k}|X)$을 $\mathrm{E}(\zeta_{Y,k}|\zeta_{X,1}, \ldots, \zeta_{X,M})$로 대체하는 모형을 생각할 수 있다. 이때, $\mathrm{E}(\zeta_{Y,k}|\zeta_{X,1}, \ldots, \zeta_{X,M})$는 함수가법모형에서와 같이 특이성분 $\zeta_{X,j}$의 가법함수 형태로 모델링한다고 하자. 그러면 다음의 특이가법모형을 얻는다.

$$\mathrm{E}(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{M} \left(g_{k,1}(\zeta_{X,1}) + \cdots + g_{k,M}(\zeta_{X,M})\right) \cdot \psi_{Y,k}(t). \tag{2.24}$$

위 특이가법모형의 성분함수 $g_{k,j}$, $1 \le j \le M$ 들도 식별가능(identifiable)하지 않다. 반응과정 Y의 특이성분이 $\mathrm{E}(\xi_{Y,k}) = 0$을 만족시키므로 각 성분함수에 대하여 제약조건

$$\mathrm{E}(g_{k,j}(\zeta_{X,j})) = 0, \quad 1 \le j \le M, \quad 1 \le k \le M \tag{2.25}$$

을 주면 모든 성분함수는 식별가능해진다. 성분함수 $g_{k,j}$의 추정은 함수가법모형에서처럼 주변회귀(marginal regression)법을 적용할 수 없다. 왜냐하면 $j \neq j'$에 대하여 $\zeta_{X,j}$와 $\zeta_{X,j'}$이 서로 상관(correlated)되어 있어,

$$\mathrm{E}(\zeta_{Y,k}|\zeta_{X,j}) = g_{k,j}(\zeta_{X,j}) + \sum_{j'=1,\neq j} \mathrm{E}(g_{kj'}(\zeta_{X,j'})|\zeta_{X,j})$$

에서 $\mathrm{E}(g_{kj'}(\zeta_{X,j'})|\zeta_{X,j}) = 0$이 될 수 없기 때문이다. 따라서, 각 $k \geq 1$에 대하여 특이성분 $\zeta_{Y,k}$을 반응변수로 하는 가법모형

$$\mathrm{E}(\zeta_{Y,k}|\zeta_{X,1}, \ldots, \zeta_{X,M}) = g_{k,1}(\zeta_{X,1}) + \cdots + g_{k,M}(\zeta_{X,M}) \tag{2.26}$$

의 추정은 평활역적합(smooth backfitting; Mammen 등, 1999)을 적용한다. 아래에 이 기법을 간략히 소개한다.

함수가법모형의 추정에서처럼 실제 적용 가능한 모형의 추정을 위해 우선 교차공분산함수 C_{XY}와 C_{YX}를 다음과 같이 추정한다.

$$\hat{C}_{XY}(s,t) = n^{-1} \sum_{i=1}^{n} \Big(X_i(s) - \bar{X}(s)\Big)\Big(Y_i(t) - \bar{Y}(t)\Big) = \hat{C}_{YX}(t,s).$$

위 교차공분산함수의 추정량으로부터 교차공분산작용소의 추정량 $\hat{\mathscr{C}}_{XY}$와 $\hat{\mathscr{C}}_{YX}$를 식 (2.5)에서처럼 정의하고 이들의 합성작용소 $\hat{\mathscr{C}}_{XYX}$와 $\hat{\mathscr{C}}_{YXY}$도 식 (2.6)에서처럼 정의한다. 이들의 스펙트럼분해로부터 고유함수 $\psi_{X,j}$와 $\psi_{Y,k}$의 추정량 $\hat{\psi}_{X,j}$와 $\hat{\psi}_{Y,k}$을 구한다. 다음으로 특이성분 $\zeta_{X,ij}$와 $\zeta_{Y,ik}$의 추정량도 다음의 식을 통해 구한다.

$$\hat{\zeta}_{X,ij} = \int_{\mathcal{S}} \Big(X_i(s) - \bar{X}(s)\Big)\hat{\psi}_{X,j}(s)\,ds, \quad \hat{\zeta}_{Y,ik} = \int_{\mathcal{T}} \Big(Y_i(t) - \bar{Y}(t)\Big)\hat{\psi}_{Y,k}(t)\,dt.$$

이제 $\zeta_{Y,ik}$를 반응변수로 하는 가법모형 (2.26)을 $\hat{\zeta}_{Y,ik}$와 $\hat{\zeta}_{X,i1}, \hat{\zeta}_{X,i2}, \ldots, \hat{\zeta}_{X,iM}$을 이용하여 적합(fitting)시키는 방법을 기술하기로 한다. 다차원 변수 $(\zeta_{X,1}, \ldots, \zeta_{X,M})$가 값을 취하는 영역은 일반성을 잃지 않고 $[0,1]^M$으로 둘 수 있다. 영역 $[0,1]^M$에서 임의의 점을 $\mathbf{u} = (u_1, \ldots, u_M)$이라 하자. 주어진 $1 \leq k \leq M$에 대하여 가법함수 $m_k(\mathbf{u}) := g_{k,1}(u_1) + \cdots + g_{k,M}(u_M)$의 추정량은 다음의 연립적분방정식의 해로 정의한다.

$$\hat{g}_{k,j}(u_j) = \tilde{g}_{kj}(u_j) - \sum_{j' \neq j}^{M} \int \hat{g}_{k,j'}\left(u_{j'}\right) \frac{\hat{p}_{j,j'}\left(u_j, u_{j'}\right)}{\hat{p}_j\left(u_j\right)} du_{j'}, \quad 1 \leq j \leq M. \tag{2.27}$$

여기서,

$$\tilde{g}_{kj}(u_j) = \frac{\sum_{i=1}^{n} K_h\left(u_j, \hat{\zeta}_{X,ij}\right)\hat{\zeta}_{Y,ik}}{\sum_{i=1}^{n} K_h\left(u_j, \hat{\zeta}_{X,ij}\right)},$$

$$\hat{p}_j(u_j) = n^{-1} \sum_{i=1}^{n} K_h\left(u_j, \hat{\zeta}_{X,ij}\right),$$

$$\hat{p}_{j,j'}(u_j, u_{j'}) = n^{-1} \sum_{i=1}^{n} K_h\left(u_j, \hat{\zeta}_{X,ij}\right) K_h\left(u_{j'}, \hat{\zeta}_{X,ij'}\right)$$

이고 커널가중치함수 $K_h(u,v)$는 모든 $v \in [0,1]$에 대하여 $\int_0^1 K_h(u,v)\,du = 1$이 되도록 다음과 같이 표준화한다.

$$K_h(u,v) = \frac{K_h(u-v)}{\int_0^1 K_h(u'-v)}\,du', \quad u, v \in [0,1].$$

밴드폭 h와 기저커널 K에 대한 매우 약한 조건에서 그리고 힐버트공간 값을 가지는 반응변수에 대하여 위 연립방정식 (2.27)의 해가 유일하게 존재함은 Jeon과 Park (2020)이 증명하였다. 이 때 해라고 함은 가법함수 $\hat{m}_k(\mathbf{u}) := \hat{g}_{k,1}(u_1) + \cdots + \hat{g}_{k,M}(u_M)$를 의미하며, 개별 성분함수의 투플(tuple)인 $(\hat{g}_{k,1}, \ldots, \hat{g}_{k,M})$이 아님을 밝혀 둔다. 개별 성분함수의 추정량을 구하기 위해서는 $\int_0^1 \hat{g}_{k,j}(u_j)\hat{p}_j(u_j)\,du_j = 0$와 같은 제약 조건이 필요하다. 본 논문에서는 개별 성분함수의 추정보다는 가법함수 m_k 자체의 추정을 통해 미관측 반응과정을 예측하는 것이 목적이기 때문에 이런 제약 조건을 주지 않는다. 실제로 평활역적합방정식(smooth backfitting equation)의 해는 반복알고리즘을 통해 이루어지며 자세한 것은 Jeon과 Park (2020)을 참고하기 바란다. 모든 $1 \le k \le M$에 대하여 가법함수 $m_k(\mathbf{u})$의 추정량 $\hat{m}_k(\mathbf{u})$를 얻으면 식 (2.24)의 특이가법모형은 다음과 같이 추정할 수 있다.

$$\hat{\mathrm{E}}(Y(t)|X) = \bar{Y}(t) + \sum_{k=1}^{M} \hat{m}_k(\zeta_{X,1}, \ldots, \zeta_{X,M}) \cdot \hat{\psi}_{Y,k}(t). \tag{2.28}$$

함수가법모형에서처럼 위 추정모형을 활용하여 미래의 값 Y_{new}에 대응하는 X_{new}을 알 때, 다음과 같이 Y_{new}를 예측할 수 있다.

$$\hat{Y}_{\text{new}}(t) = \bar{Y}(t) + \sum_{k=1}^{M} \hat{m}_k(\zeta_{X,1,\text{new}}, \ldots, \zeta_{X,M,\text{new}}) \cdot \hat{\psi}_{Y,k}(t). \tag{2.29}$$

여기서,

$$\zeta_{X,j,\text{new}} = \int_{\mathcal{S}} \big(X_{\text{new}}(s) - \bar{X}(s)\big)\hat{\psi}_{X,j}(s)\,ds$$

이며, $\bar{X}$, $\bar{Y}$, $\hat{m}_k$, $\hat{\psi}_{X,j}$, $\hat{\psi}_{Y,k}$, $1 \le j,k \le M$ 등은 모두 $\{(X_i, Y_i) : 1 \le i \le n\}$을 이용하여 구한다. 특이가법모형의 적용에서도 함수가법모형처럼 식 (2.20)와 같은 5겹 교차검증기준으로 M을 선택할 수 있다.

2.5. 특이선형모형

함수선형모형(functional linear model)은 원래 $\mathrm{E}(Y^c(t)|X)$를 $\int_{\mathcal{S}} \beta(s,t)X^c(s)\,ds$로 모델링하는 것이다. 이 절에서는 특이성분을 이용하여 이 모형을 근사시키는 접근 방법을 생각해 본다. 먼저 $\tilde{X}(s) := \sum_{j=1}^{M} \zeta_{X,j}\psi_{X,j}(s)$라 하자. 특이가법모형에서처럼 $X^c(s) = \tilde{X}(s) + \nu_X$에서 ν_X를 측정오차로 간주하여 $\mathrm{E}(Y^c(t)|X)$를 $\mathrm{E}(Y^c(t)|\tilde{X})$로 대체한 다음의 모형을 생각해 보자.

$$\mathrm{E}\left(Y^c(t)|\tilde{X}\right) = \int_{\mathcal{S}} \beta(s,t)\tilde{X}(s)\,ds. \tag{2.30}$$

또한, 2.4절에서처럼 $\mathscr{C}_{XY}$와 $\mathscr{C}_{YX}$가 랭크 M이라 하고 ν_Y가 X와 독립이라고 가정하면 식 (2.21) Y의 특이성분전개로부터 다음을 얻는다.

$$\mathrm{E}\left(Y^c(t)|\tilde{X}\right) = \sum_{k=1}^{M} \mathrm{E}\left(\zeta_{Y,k}|\tilde{X}\right) \cdot \psi_{Y,k}(t). \tag{2.31}$$

식 (2.31)에서 $\mathrm{E}(\zeta_{Y,k}|\tilde{X})$는 특이성분의 정의, 식 (2.30) 그리고 $X^c(s) = \sum_{j=1}^{M} \zeta_{X,j}\psi_{X,j}(s) + \nu_X(s)$에 의해 다음과 같이 표현할 수 있다.

$$\begin{aligned}
\mathrm{E}(\zeta_{Y,k}|\tilde{X}) &= \int_{\mathcal{T}} \mathrm{E}\left(Y^c(t)|\tilde{X}\right)\psi_{Y,k}(t)\,dt \\
&= \int_{\mathcal{T}}\int_{\mathcal{S}} \beta(s,t)\tilde{X}(s)\,\psi_{Y,k}(t)\,ds\,dt \\
&= \sum_{j=1}^{M} \zeta_{X,j} \cdot \int_{\mathcal{T}}\int_{\mathcal{S}} \beta(s,t)\psi_{X,j}(s)\psi_{Y,k}(t)\,ds\,dt \\
&\overset{\text{let}}{=} \sum_{j=1}^{M} \beta_{j,k} \cdot \zeta_{X,j}.
\end{aligned}$$

위 표현을 식 (2.31)에 대입하면 다음의 모형을 얻게 되고 이를 특이선형모형(singular linear model)이라 하자.

$$\mathrm{E}(Y^c(t)|\tilde{X}) = \sum_{j=1}^{M}\sum_{k=1}^{M} \beta_{j,k} \cdot \zeta_{X,j} \cdot \psi_{Y,k}(t). \tag{2.32}$$

이제 특이선형모형 (2.32)에서 $\beta_{j,k}$를 추정하는 방법을 생각해 보자. 이를 위해

$$\boldsymbol{\beta}_k = (\beta_{1,k}, \ldots, \beta_{M,k})^\top, \quad \boldsymbol{\zeta}_X = (\zeta_{X,1}, \ldots, \zeta_{X,M})^\top$$

라 하자. 그러면 $\mathrm{E}(\zeta_{Y,k}|\tilde{X}) = \sum_{j=1}^{M} \beta_{j,k}\zeta_{X,j} = \boldsymbol{\zeta}_X^\top\boldsymbol{\beta}_k$이고, 따라서 $\boldsymbol{\beta}_k$는 $\mathbb{R}^M$에서 $\mathrm{E}(\zeta_{Y,k} - \boldsymbol{\zeta}_X^\top\mathbf{b}_k)^2$을 최소로 하는 M 차원의 벡터 $\mathbf{b}_k$이다. 즉,

$$\boldsymbol{\beta}_k = \left(\mathrm{E}\left(\boldsymbol{\zeta}_X\boldsymbol{\zeta}_X^\top\right)\right)^{-1}\mathrm{E}(\zeta_{Y,k}\boldsymbol{\zeta}_X).$$

한편, 식 (2.7)에 의하면 $j \neq k$에 대하여 $\mathrm{E}(\zeta_{X,j}\zeta_{Y,k}) = 0$이므로 M차원의 벡터 $\mathrm{E}(\zeta_{Y,k}\boldsymbol{\zeta}_X)$의 원소 중 k번째 원소를 제외한 모든 원소는 0이 된다. 따라서, $\mathbf{D} = \mathrm{E}\left(\boldsymbol{\zeta}_X\boldsymbol{\zeta}_X^\top\right)$이라 두면

$$\boldsymbol{\beta}_k = \left(k^{th} \text{ column of } \mathbf{D}^{-1}\right) \cdot \mathrm{E}(\zeta_{X,k}\zeta_{Y,k}) \tag{2.33}$$

이다. 위 식에서 $\mathbf{D}$와 $s_k := \mathrm{E}(\zeta_{X,k}\zeta_{Y,k})$는 각각

$$\begin{aligned}
\hat{\mathbf{D}} &:= \left(\hat{D}_{j,j'}\right),\ \hat{D}_{j,j'} = n^{-1}\sum_{i=1}^{n} \hat{\zeta}_{X,ij} \cdot \hat{\zeta}_{X,ij'}, \\
\hat{s}_k &:= n^{-1}\sum_{i=1}^{n} \hat{\zeta}_{X,ik} \cdot \hat{\zeta}_{Y,ik}
\end{aligned}$$

으로 추정한다. 즉,

$$\hat{\boldsymbol{\beta}}_k = \left(k^{th} \text{ column of } \hat{\mathbf{D}}^{-1}\right) \cdot \hat{s}_k.$$

이와 같이 추정된 $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{1,k}, \ldots, \hat{\beta}_{M,k})^\top$, $k \geq 1$ 을 이용하여 특이선형모형 (2.32)을 다음과 같이 추정한다.

$$\hat{\mathrm{E}}\left(Y^c(t)|\tilde{X}\right) = \sum_{j=1}^{M}\sum_{k=1}^{M} \hat{\beta}_{j,k} \cdot \zeta_{X,j} \cdot \hat{\psi}_{Y,k}(t).$$

함수가법모형과 특이가법모형에서처럼 위 추정모형을 활용하여 미래의 값 Y_{new}에 대응하는 X_{new}을 알 때, 다음과 같이 Y_{new}를 예측할 수 있다.

$$\hat{Y}_{\text{new}}(t) = \bar{Y}(t) + \sum_{j=1}^{M}\sum_{k=1}^{M} \hat{\beta}_{j,k} \cdot \zeta_{X,j,\text{new}} \cdot \hat{\psi}_{Y,k}(t) \tag{2.34}$$

여기서,

$$\zeta_{X,j,\text{new}} = \int_{\mathcal{S}} \left(X_{\text{new}}(s) - \bar{X}(s)\right) \hat{\psi}_{X,j}(s)\, ds$$

이며, $\bar{X}$, $\bar{Y}$, $\hat{\beta}_{j,k}$, $\hat{\psi}_{X,j}$, $\hat{\psi}_{Y,k}, 1 \le j,k \le M$ 등은 모두 $\{(X_i, Y_i) : 1 \le i \le n\}$을 이용하여 구한다. 특이선형모형의 적용에서도 함수가법모형처럼 식 (2.20)와 같은 5겹 교차검증기준으로 M을 선택할 수 있다.

3. 분석

교통량 예측은 효율적인 고속도로 설계 및 유지 관리를 수행하는데 도움이 되며, 관광지역의 관광객 유입량에 대한 정보로 활용될 수 있다. 이 절에서는 평일 교통량을 이용하여 해당 주의 주말교통량을 예측하는 문제에 2절에서 소개한 방법들을 적용한 결과를 살펴보기로 한다. 이 분석에 활용한 데이터는 2019년 1월7일부터 2021년 1월 3일까지의 105개 주(week)에 대해 0시부터 24시까지 1시간 간격으로 측정된 고속도록 영업소 출구 교통량이다. 고속도로 영업소 출구 교통량 데이터는 한국도로공사 공공데이터 포털(http://data.ex.co.kr)에서 수집하였다. 전국 고속도록 영업소 중 서울, 춘천, 강릉 3개의 영업소에 대해 분석하였고, 평일(월–목, 월–금, 월, 화, 수, 목, 금)과 주말(토,일) 14개 쌍에 대해 평일에 관측된 교통량을 공변과정 X로 주말 교통량을 반응과정 Y로 하여 3개의 모형을 적합시켰다. 105주 중에서 평일이 공휴일인 주와 이상값이 포함된 주는 분석에서 제외하였다. 결과적으로 서울 영업소의 경우 총 97주, 춘천과 강릉영업소는 98주의 교통량 데이터가 분석에 사용되었다. 함수주성분분석과 함수특이성분분석의 수행을 위해 R프로그래밍 언어의 'fdapace' 패키지에서 제공하는 FPCA함수(http://rdrr.io/cran/fdapace/src/R/FPCA.R)와 FSVD함수(http://rdrr.io/cran/fdapace/src/R/RSVD.R)를 사용하였다.

예측 성능 비교를 위해 전체자료를 훈련자료(training data)와 시험자료(test data)로 나누었다. 2절에서 소개한 모형 각각을 $Y = f(X)$라 두고 훈련자료를 이용하여 추정한 모형을 $\hat{f}$이라 하자. 각 모형의 예측성능은 시험자료의 공변과정 X_{test}를 추정모형에 대입한 $\hat{f}(X_{\text{test}})$와 실제 관측된 시험자료의 반응과정 Y_{test}를 비교하여 측정하였다. 보다 구체적으로 설명하기 위해 훈련자료 군을 $\{(X_i, Y_i) : 1 \le i \le n\}$이라 하고 시험자료를 $\{(X_i^{\text{test}}, Y_i^{\text{test}}) : 1 \le i \le N\}$이라 하자. 함수가법모형의 경우에는 먼저 훈련자료를 이용하여 $\bar{X}$, $\bar{Y}$, $\hat{f}_{k,j}$, $\hat{\phi}_{X,j}$, $\hat{\phi}_{Y,k}, 1 \le j,k \le M$ 등을 구하였다. 이를 이용하여 i번째 시험자료 관측 짝 $(X_i^{\text{test}}, Y_i^{\text{test}})$에 대하여 식 (2.19)에서와 같이

$$\text{FAM}: \quad \hat{Y}_i^{\text{test}}(t) = \bar{Y}(t) + \sum_{k=1}^{M} \left(\hat{f}_{k,1}(\xi_{X_i^{\text{test}},1}) + \cdots + \hat{f}_{k,M}(\xi_{X_i^{\text{test}},M})\right) \cdot \hat{\phi}_{Y,k}(t) \tag{3.1}$$

을 구하여 $Y_i^{\text{test}}(t)$와 비교하였다. 여기서,

$$\xi_{X_i^{\text{test}},j} = \int_{\mathcal{S}} \left(X_i^{\text{test}}(s) - \bar{X}(s)\right) \cdot \hat{\phi}_{X,j}(s)\ ds$$

이다. 특이가법모형의 경우에도 훈련자료를 이용하여 $\bar{X}$, $\bar{Y}$, $\hat{m}_k$, $\hat{\psi}_{X,j}$, $\hat{\psi}_{Y,k}, 1 \le j,k \le M$ 등을 얻은 다음,

$$\text{SAM}: \hat{Y}_i^{\text{test}}(t) = \bar{Y}(t) + \sum_{k=1}^{M} \hat{m}_k \left(\zeta_{X_i^{\text{test}},1}, \ldots, \zeta_{X_i^{\text{test}},M}\right) \cdot \hat{\psi}_{Y,k}(t) \tag{3.2}$$

와 $Y_i^{\text{test}}(t)$를 비교하였다. 여기서도

$$\zeta_{X_i^{\text{test}},j} = \int_{\mathcal{S}} \left(X_i^{\text{test}}(s) - \bar{X}(s)\right) \cdot \hat{\psi}_{X,j}(s)\, ds$$

이다. 마지막으로 특이선형모형의 경우에는 훈련자료를 이용하여 $\bar{X}$, $\bar{Y}$, $\hat{\beta}_{j,k}$, $\hat{\psi}_{X,j}$, $\hat{\psi}_{Y,k}$, $1 \le j,k \le M$ 등을 계산하고

$$\text{SLM}: \quad \hat{Y}_i^{\text{test}}(t) = \bar{Y}(t) + \sum_{j=1}^{M}\sum_{k=1}^{M} \hat{\beta}_{j,k} \cdot \zeta_{X_i^{\text{test}},j} \cdot \hat{\psi}_{Y,k}(t) \tag{3.3}$$

와 $Y_i^{\text{test}}(t)$를 비교하였다.

예측성능의 비교를 위해 상대적분제곱예측오차값을 계산하였고, 주성분과 특이성분의 개수(M)와 밴드폭(h)의 선택은 훈련자료 내에서 2.3절에서 소개한 5겹 교차검증(5 fold cross-validation)을 통해 진행하였다. 상대적분제곱예측오차(relative integrated square prediction error, RISPE)는 다음과 같다.

$$\text{RISPE}_i = \frac{\int_{\mathcal{T}} \left[\hat{Y}_i^{\text{test}}(t) - Y_i^{\text{test}}(t)\right]^2 dt}{\int_{\mathcal{T}} \left(Y_i^{\text{test}}(t)\right)^2}, \quad i = 1, \ldots, N.$$

예측성능은 이들 RISPE 값들의 표본평균인

$$\text{Mean}\left(= \overline{\text{RISPE}}\right) = \frac{1}{N}\sum_{i=1}^{N} \text{RISPE}_i$$

으로 비교하였고, 위 표본평균의 변동성을 나타내는 지표로서 몬테카를로(Monte Carlo) 표본오차

$$\text{SE} = N^{-\frac{1}{2}} \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\text{RISPE}_i - \overline{\text{RISPE}}\right)^2}$$

를 사용하였다. 위 표본오차는 훈련자료를 통해 추정한 예측모형이 고정되었을 때 시험자료의 변동성으로부터 유발되는 표본평균 Mean의 변동성 만을 고려한 것임을 밝혀 둔다. 이 분석에서는 평일 교통량이 주말 교통량의 예측에 유효한지를 검토하기 위해, 위 세가지 방법 외에 훈련자료에서 계산한 $\bar{Y}(t)$를 시험자료의 $Y_i^{\text{test}}(t)$와도 비교하였다. 이 방법은 공변과정 X의 정보를 전혀 사용하지 않은 것으로

$$\text{RISPE}_i = \frac{\int_{\mathcal{T}} \left[\bar{Y}(t) - Y_i^{\text{test}}(t)\right]^2 dt}{\int_{\mathcal{T}} \left(Y_i^{\text{test}}(t)\right)^2}, \quad i = 1, \ldots, N$$

의 평균을 계산하여 FAM, SAM, SLM의 상대적분제곱예측오차값의 평균과 비교함으로써 공변과정이 반응과정의 변화를 잘 예측하는데 유용한지 여부를 판단할 수 있다.

3.1. 서울영업소

서울 영업소는 경기도 성남시 분당고 대왕판교로 240(궁내동)에 위치한 경부고속도로 상의 요금소로 교통량이 상당히 많은 수도권 지역의 주요 영업소이다. 이상값이 존재하는 1개 주를 제거하고, 훈련자료는 2019년 1월7일부터 2020년 8월 30일까지 초기 70개 주(2019년 3월5일 0시 241만대 이상값이 포함된 1개의 주 제외), 시험자료(test set)는 2020년 8월31일부터 2021년 1월 3일까지 17개 주로 하여 분석하였다.

Table 1은 상대적분제곱예측오차의 평균 및 표본오차와 함께 FAM, SAM, SLM 모형 추정시 사용한 주성분 또는 특이성분의 개수 M 그리고 FAM과 SAM모형의 추정시 사용한 밴드폭 h를 나타낸다. 여기서, M

Table 1: Mean of RISPE for SeoulTG

		FAM			SAM			SLM		
X	Y	h	M	Mean (SE)	h	M	Mean (SE)	M	Mean (SE)	$\bar{Y}$
월~목		0.32	2	0.0120 (0.0042)	0.56	5	0.0145 (0.0052)	2	0.0097 (0.0035)	
월~금		0.36	2	0.0118 (0.0046)	0.61	4	0.0140 (0.0053)	3	0.0092(0.0034)	
월		0.18	3	0.0132 (0.0047)	0.29	3	0.0147 (0.0053)	2	0.0112 (0.0038)	
화	토	0.18	3	0.0137 (0.0051)	0.24	2	0.0141 (0.0052)	3	0.0116 (0.0048)	0.0184
수		0.12	2	0.0111 (0.0039)	0.20	4	0.0129 (0.0044)	2	0.0098 (0.0030)	
목		0.83	5	0.0182 (0.0067)	0.88	4	0.0179 (0.0065)	4	0.0151 (0.0063)	
금		0.33	5	0.0258 (0.0122)	0.48	4	0.0279 (0.0138)	4	**0.0082 (0.0021)**	
월~목		0.32	4	0.0111 (0.0026)	0.53	2	0.0142 (0.0037)	5	0.0113 (0.0027)	
월~금		0.30	2	0.0122 (0.0039)	0.61	2	0.0141 (0.0037)	4	0.0168 (0.0054)	
월		0.14	2	0.0126 (0.0031)	0.30	2	0.0174 (0.0045)	3	0.0137 (0.0049)	
화	일	0.17	4	0.0127 (0.0030)	0.28	2	0.0150 (0.0038)	3	0.0110 (0.0027)	0.0201
수		0.10	2	**0.0103 (0.0023)**	0.20	4	**0.0119 (0.0029)**	3	0.0105 (0.0024)	
목		0.85	2	0.0198 (0.0052)	0.83	2	0.0196 (0.0051)	5	0.0150 (0.0047)	
금		0.33	5	0.0413 (0.0195)	0.55	3	0.0189 (0.0049)	3	0.0163 (0.0047)	

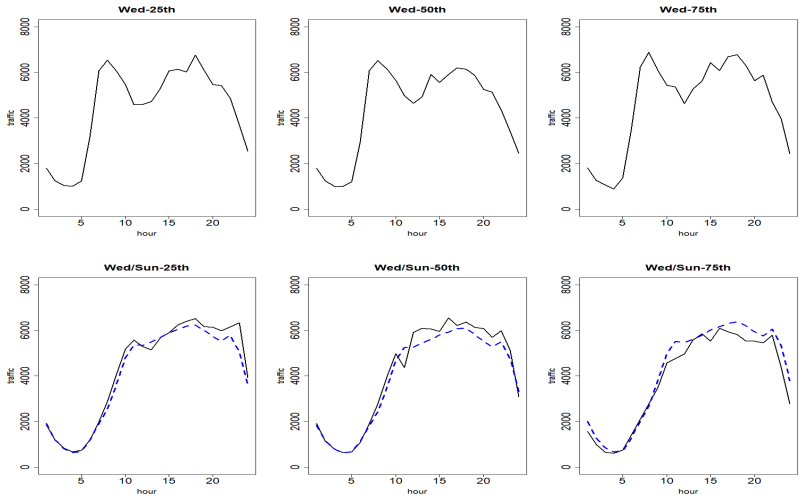

Figure 1: *FAM results of SeoulTG for the* 25^{th}, 50^{th} *and* 75^{th} *percentiles of RISPE. The upper panels depict the traffic on the predictor processes, the lower panels the corresponding observed (black solid curves) and estimated (blue dashed curves) traffic on response process.*

과 h는 2.3절의 식 (2.20)와 같은 5겹 교차검정기준으로 선택한 것이다. 우선 $\bar{Y}$와의 예측오차 비교를 통해 평일 교통량이 주말 교통량을 예측하는데 유효함을 확인할 수 있다. SLM의 평균 오차는 금요일 교통량으로 토요일 교통량을 예측한 경우에 0.0082로 가장 작게 나타났으며, FAM과 SAM의 평균 오차는 수요일 교통량으로 일요일 교통량을 예측한 경우에 각 0.0103, 0.0119로 가장 작게 나타났다. 이 분석 결과로부터 토요일 교통량을 예측하고자 할 때는 금요일의 교통량을 공변과정으로 하는 SLM이 가장 성능이 우수하고, 일요일 교통량을 예측하고자 할 때는 수요일의 교통량을 공변과정으로 하는 FAM이 가장 성능이 우수함을 알 수 있다. 전반적으로 SLM이 가장 성능이 우수함을 보여주는 결과이다.

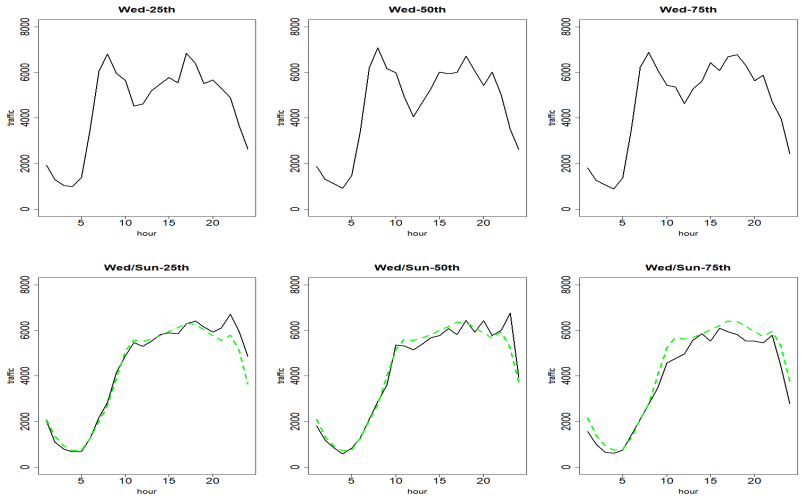

Figure 2: *SAM results of SeoulTG for the* 25^{th}, 50^{th} *and* 75^{th} *percentiles of RISPE. The upper panels depict the traffic on the predictor processes, the lower panels the corresponding observed (black solid curves) and estimated (green dashed curves) traffic on response process.*

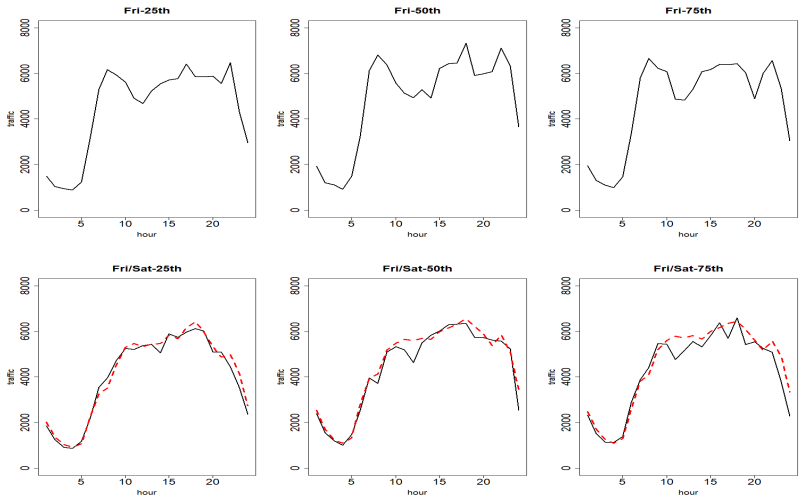

Figure 3: *SLM results of SeoulTG for the* 25^{th}, 50^{th} *and* 75^{th} *percentiles of RISPE. The upper panels depict the traffic on the predictor processes, the lower panels the corresponding observed (black solid curves) and estimated (red dashed curves) traffic on response process.*

Figure 1은 FAM의 평균 오차가 가장 작은 경우인 수요일/일요일에 대한 그래프로, 위쪽 패널은 RISPE_i의 크기 기준으로 25, 50, 70 백분위수에 해당되는 i(3개 주) 의 공변과정 X_i^{test}(수요일 교통량)을 나타내고, 아래쪽 패널은 반응과정 Y_i^{test}(일요일 교통량: black solid curves)와 FAM에 의한 예측 교통량 $\hat{Y}_i^{\text{test}}$(dashed curves)를 나타내고 있다. Figure 2와 Figure 3은 각각 SAM의 수요일/일요일과 SLM의 금요일/토요일에 대한 결과 그래프이다.

Table 2: Mean of RISPE for ChuncheonTG

		FAM			SAM			SLM		
X	Y	h	M	Mean(SE)	h	M	Mean(SE)	M	Mean(SE)	$\bar{Y}$
월~목		0.08	5	0.0280 (0.0092)	0.10	4	0.0324 (0.0102)	2	0.0415 (0.0185)	
월~금		0.07	5	0.0778 (0.0444)	0.15	3	0.0319 (0.0108)	2	0.0349 (0.0144)	
월		0.04	5	0.0346 (0.0084)	0.05	3	0.0264 (0.0061)	5	0.0242 (0.0051)	
화	토	0.03	5	0.0272 (0.0073)	0.09	3	0.0270 (0.0071)	3	0.0256 (0.0067)	0.0319
수		0.03	5	**0.0244 (0.0070)**	0.05	4	**0.0238 (0.0065)**	2	**0.0228 (0.0063)**	
목		0.05	5	0.0485 (0.0175)	0.10	2	0.0356 (0.0128)	2	0.0616 (0.0279)	
금		0.10	5	0.0468 (0.0202)	0.15	5	0.0255 (0.0053)	3	0.0270 (0.0078)	
월~목		0.05	5	0.0616 (0.0215)	0.13	4	0.0424 (0.0125)	4	0.0545 (0.0205)	
월~금		0.06	5	0.0949 (0.0474)	0.15	5	0.0415 (0.0126)	4	0.0563 (0.0207)	
월		0.04	5	0.0536 (0.0130)	0.05	5	0.0463 (0.0115)	4	0.0442 (0.0125)	
화	일	0.03	2	0.0407 (0.0106)	0.09	2	0.0421 (0.0119)	2	0.0413 (0.0109)	0.0478
수		0.03	5	0.0369 (0.0108)	0.05	5	0.0377 (0.0111)	2	0.0364 (0.0106)	
목		0.05	5	0.0606 (0.0206)	0.09	2	0.0515 (0.0176)	3	0.0785 (0.0331)	
금		0.10	5	0.0637 (0.0233)	0.15	5	0.0419 (0.0118)	4	0.0369 (0.0117)	

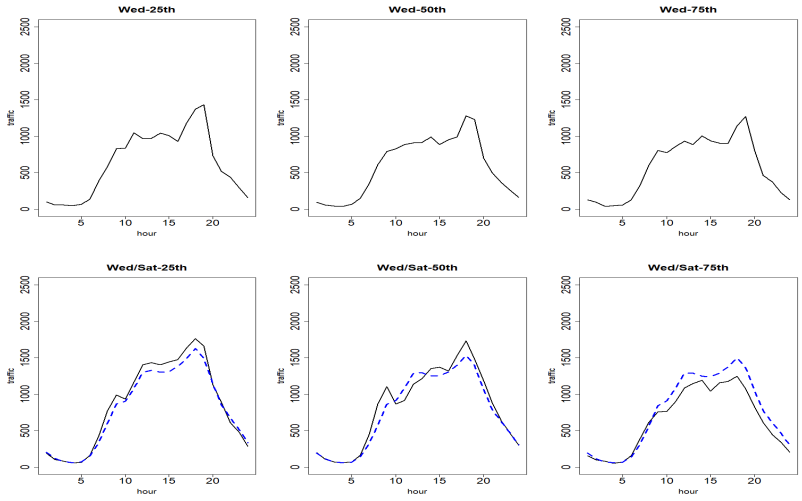

Figure 4: *FAM results of ChuncheonTG for the 25th, 50th and 75th percentiles of RISPE. The upper panels depict the traffic on the predictor processes, the lower panels the corresponding observed (black solid curves) and estimated (blue dashed curves) traffic on response process.*

3.2. 춘천영업소

춘천영업소는 강원도 춘천시 동내면 고은길 88에 위치한 중앙고속도로로 상의 요금소로 춘천 시내로 나가는 가장 빠른 관광 지역의 주요 영업소이다. 훈련자료는 2019년 1월 7일부터 2020년 8월 16일까지 70개 주, 시험자료는 2020년 8월 24일부터 2021년 1월 3일까지 18개 주로 하여 분석하였다.

Table 2는 상대적분제곱예측오차의 평균 및 표본오차와 함께 FAM, SAM, SLM 모형 추정시 사용한 주성분 또는 특이성분의 개수 M 그리고 FAM과 SAM모형의 추정시 사용한 밴드폭 h를 나타낸다. 여기서, M과 h

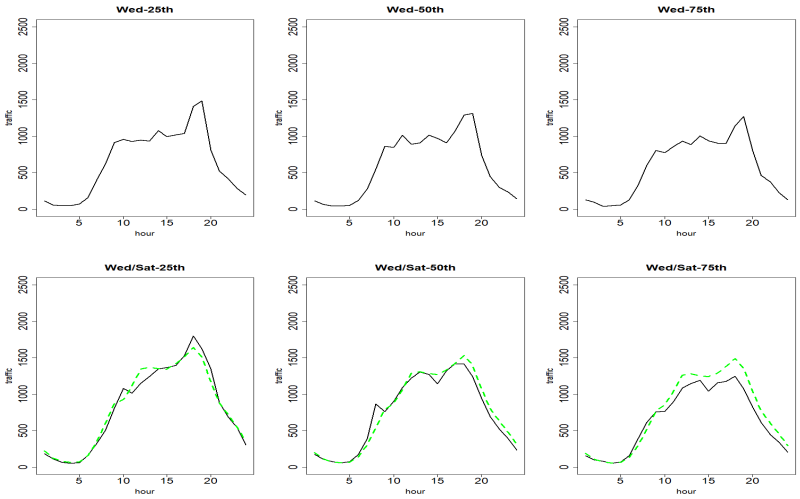

Figure 5: *SAM results of ChuncheonTG for the 25th, 50th and 75th percentiles of RISPE. The upper panels depict the traffic on the predictor processes, the lower panels the corresponding observed (black solid curves) and estimated (green dashed curves) traffic on response process.*

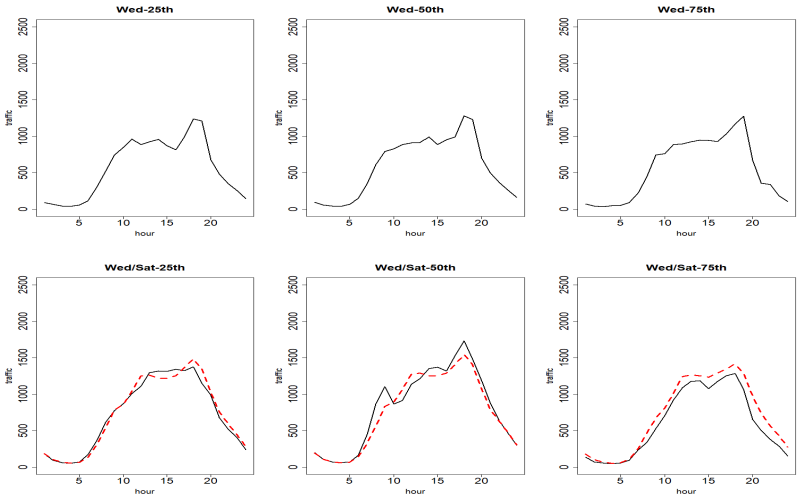

Figure 6: *SLM results of ChuncheonTG for the 25th, 50th and 75th percentiles of RISPE. The upper panels depict the traffic on the predictor processes, the lower panels the corresponding observed (black solid curves) and estimated (red dashed curves) traffic on response process.*

는 2.3절의 식 (2.20)와 같은 5겹 교차검정기준으로 선택한 것이다. FAM, SAM, SLM 모두 수요일 교통량으로 토요일 교통량을 예측한 경우에 상대적분제곱예측오차의 평균이 각각 0.0244, 0.0238, 0.0228로 가장 작게 나타났다. 토요일 교통량을 예측하고자 할 때는 수요일의 교통량을 공변과정으로 하는 SLM이 가장 성능이 우수하고, 일요일 교통량을 예측하고자 할 때도 수요일의 교통량을 공변과정으로 하는 SLM이 가장 성능이 우수함을 알 수 있다. 춘천영업소의 경우에도 SLM이 전반적으로 가장 성능이 우수함을 보여주고 있다. 여

Table 3: Mean of RISPE for GangneungTG

		FAM			SAM			SLM		
X	Y	h	M	Mean(SE)	h	M	Mean(SE)	M	Mean(SE)	$\bar{Y}$
월~목		0.08	5	0.0428 (0.0162)	0.27	5	0.0445 (0.0156)	4	0.0649 (0.0291)	
월~금		0.13	5	0.0386 (0.0132)	0.28	5	0.0433 (0.0148)	5	0.0399 (0.0116)	
월		0.16	5	0.0448 (0.0153)	0.16	3	0.0448 (0.0153)	2	0.0365 (0.0121)	
화	토	0.08	2	0.0458 (0.0152)	0.11	2	0.0440 (0.0146)	5	0.0462 (0.0124)	0.0469
수		0.06	3	0.0445 (0.0152)	0.13	5	0.0441 (0.0152)	3	0.0452 (0.0159)	
목		0.08	3	0.0450 (0.0172)	0.17	4	0.0469 (0.0167)	2	0.0729 (0.0361)	
금		0.05	2	0.0858 (0.0450)	0.19	2	0.0426 (0.0141)	5	0.0506 (0.0203)	
월~목		0.08	5	**0.0340 (0.0102)**	0.27	4	0.0427 (0.0128)	2	0.0393 (0.0124)	
월~금		0.14	5	0.0357 (0.0096)	0.29	5	0.0432 (0.0121)	2	0.0375 (0.0106)	
월		0.06	3	0.0423 (0.0111)	0.15	2	0.0442 (0.0128)	2	0.0471 (0.0121)	
화	일	0.03	3	0.0457 (0.0136)	0.09	3	0.0417 (0.0121)	2	0.0484 (0.0145)	0.0463
수		0.05	4	0.0367 (0.0111)	0.13	3	**0.0411 (0.0123)**	5	**0.0237 (0.0085)**	
목		0.07	4	0.0390 (0.0114)	0.17	5	0.0450 (0.0133)	2	0.0481 (0.0187)	
금		0.05	2	0.1106 (0.0515)	0.19	4	0.0428 (0.0121)	5	0.0350 (0.0111)	

기에서도 가장 성능이 우수한 모형과 $\bar{Y}$의 예측오차를 비교해보면 평일 교통량이 주말 교통량을 예측하는데 유효함을 확인할 수 있다.

Figure 4는 FAM의 평균 오차가 가장 작은 경우인 수요일/토요일에 대한 그래프로, 위쪽 패널은 RISPE$_i$의 크기 기준으로 25, 50, 70 백분위수에 해당되는 i(3개 주)의 공변과정 X_i^{test}(수요일 교통량)을 나타내고, 아래쪽 패널은 반응과정 Y_i^{test}(일요일 교통량: black solid curves)와 FAM에 의한 예측 교통량 $\hat{Y}_i^{\text{test}}$(dashed curves)를 나타내고 있다. Figure 5와 Figure 6은 각각 SAM과 SLM에 대한 결과 그래프이다.

3.3. 강릉영업소

강릉영업소는 강원도 강릉시 성산면 동해고속도로로 348에 위치한 동해고속도로로 상의 요금소로 강릉 시내로 나가는 가장 빠른 관광 지역의 주요 영업소이다. 훈련자료는 2019년 1월 7일부터 2020년 8월 16일까지 70개 주, 시험자료는 2020년 8월 24일부터 2021년 1월 3일까지 18개주로 하여 분석하였다.

Table 3은 상대적분제곱예측오차의 평균 및 표본오차와 함께 FAM, SAM, SLM 모형 추정시 사용한 주성분 또는 특이성분의 개수 M 그리고 FAM과 SAM모형의 추정시 사용한 밴드폭 h를 나타낸다. 여기서, M과 h는 2.3절의 식 (2.20)와 같은 5겹 교차검정기준으로 선택한 것이다. SLM은 수요일 교통량으로 일요일 교통량을 예측한 경우에 상대적분제곱예측오차값이 0.0237로 가장 작게 나타났으며, FAM은 월요일–목요일 교통량으로 일요일 교통량을 예측한 경우에 상대적분제곱예측오차의 평균이 0.0340으로 가장 작게 나타났고, SAM은 수요일 교통량으로 일요일 교통량을 예측한 경우가 0.0411로 가장 작게 나타났다. 토요일 교통량을 예측하고자 할 때는 월요일의 교통량을 공변과정으로 하는 SLM이 가장 성능이 우수하고, 일요일 교통량을 예측하고자 할 때도 수요일의 교통량을 공변과정으로 하는 SLM이 가장 성능이 우수함을 알 수 있다. 강릉영업소의 경우에도 SLM이 전반적으로 가장 성능이 우수함을 알 수 있다. 여기에서도 가장 성능이 우수한 모형과 $\bar{Y}$의 예측오차를 비교해보면 평일 교통량이 주말 교통량을 예측하는데 유효함을 확인할 수 있다.

Figure 7은 FAM의 평균 오차가 가장 작은 경우인 월요일–목요일/일요일에 대한 그래프로, 위쪽 패널은 RISPE$_i$의 크기 기준으로 25, 50, 70 백분위수에 해당되는 i(3개 주)의 공변과정 X_i^{test}(월요일–목요일 교통량)을 나타내고, 아래쪽 패널은 반응과정 Y_i^{test}(일요일 교통량: black solid curves)과 FAM에 의한 예측 교통량

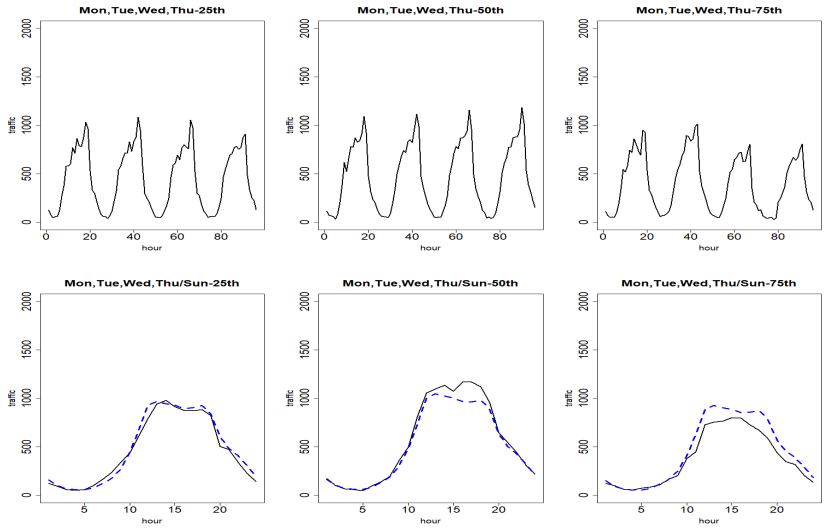

Figure 7: *FAM results of GangneungTG for the* 25^{th}, 50^{th} *and* 75^{th} *percentiles of RISPE. The upper panels depict the traffic on the predictor processes, the lower panels the corresponding observed (black solid curves) and estimated (blue dashed curves) traffic on response process.*

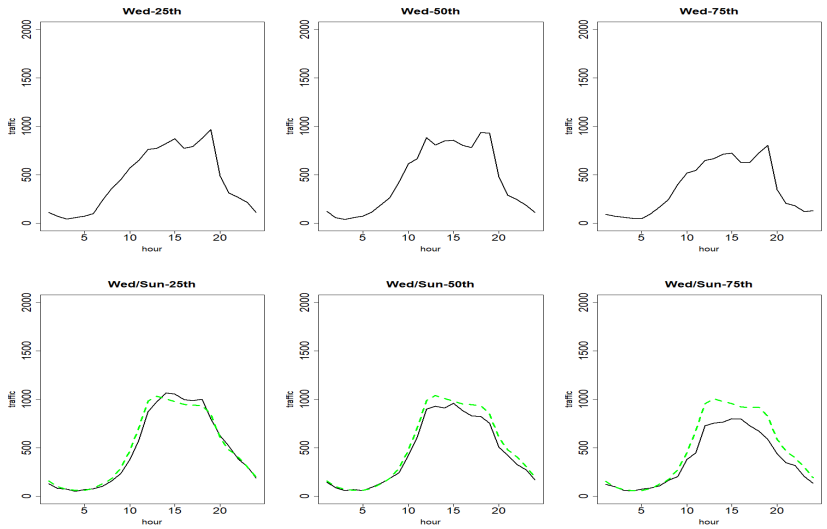

Figure 8: *SAM results of GangneungTG for the* 25^{th}, 50^{th} *and* 75^{th} *percentiles of RISPE. The upper panels depict the traffic on the predictor processes, the lower panels the corresponding observed (black solid curves) and estimated (green dashed curves) traffic on response process.*

$\hat{Y}_i^{\text{test}}$(dashed curves)를 나타내고 있다. Figure 8과 Figure 9는 각각 SAM과 SLM의 수요일/일요일에 대한 결과 그래프이다.

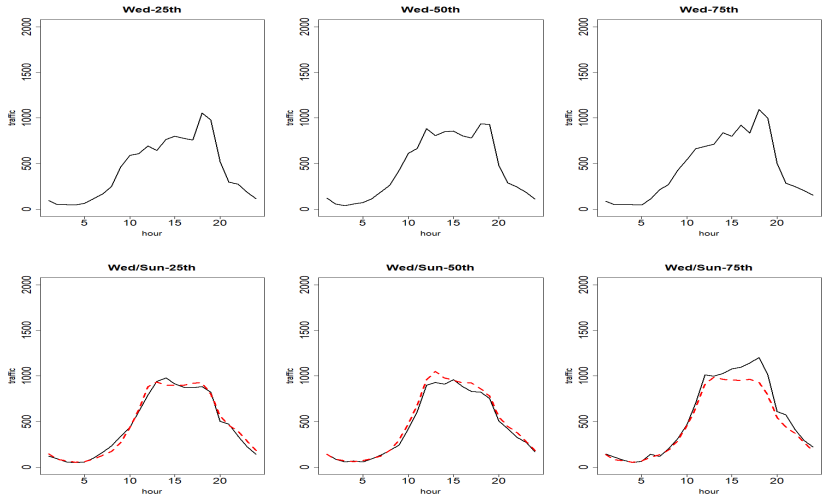

Figure 9: *SLM results of GangneungTG for the25th, 50th and 75th percentiles of RISPE. The upper panels depict the traffic on the predictor processes, the lower panels the corresponding observed (black solid curves) and estimated (red dashed curves) traffic on response process.*

4. 결론

본 연구에서는 한국도로공사에서 제공하는 0시에서 24시까지 1시간 간격으로 측정된 고속도로 영업소 출구 교통량 데이터를 이용해 평일 교통량으로 주말 교통량의 시간대별 패턴을 예측하였다. 공휴일인 평일이 있는 주와 이상치가 있는 주를 제거하여 분석하였으며, 서울, 춘천, 강릉 영업소를 대상으로 하였다. 함수가법모형, 특이가법모형, 특이선형모형의 세가지 방법을 이용하여 주말 교통량 궤적을 예측하였으며, 예측 성능은 상대적분제곱예측오차(RISPE)를 기반으로 비교하였다. 세 모형의 추정에서 조정모수인 주성분(principal component)과 특이성분(singular component)의 개수 그리고 밴드폭(bandwidth)은 5겹 교차검증(5-fold cross-validation)을 이용해 선택하였다.

분석 결과 평일 교통량이 주말 교통량을 예측하는데 유효함을 확인하였다. 세 영업소 모두 예측성능 면에서 특이선형모형이 가장 우수하였고, 세가지 방법 모두 예측한 교통량이 실제 교통량과 유사한 패턴을 보이는 등 높은 정확도의 예측 성능을 가짐을 알 수 있었다. 일반적으로 모수적 방법은 비모수적 방법에 비해 추정량의 분산은 작지만 데이터의 속성이 설정된 모형으로부터 멀어질수록 모형 편의가 증가하여 추정량 분산의 감소량을 능가할 수 있다. 본 연구의 분석 결과를 놓고 볼 때 특이선형모형이 주말교통량과 주중 교통량의 실제 함수적 관계를 잘 설명하기 때문에 특이선형모형 기반의 예측이 가장 우수한 성능을 보인 것으로 생각한다.

세 영업소 출구 교통량 데이터에 특이선형모형을 적용한 결과는 다음과 같다. 서울 영업소의 경우, 토요일 교통량 예측에는 금요일의 교통량을 이용하는 것이 가장 성능이 우수하였고 일요일 교통량 예측에는 수요일 교통량을 이용하는 것이 가장 성능이 좋았다. 춘천 영업소의 경우에는, 토요일과 일요일 교통량 예측 모두에 수요일 교통량을 공변과정으로 사용하는 것이 가장 성능이 우수하였다. 강릉영업소의 경우, 토요일 교통량에 예측에 월요일 교통량을 사용하는 것이 가장 좋았으며 일요일 교통량 예측에는 수요일 교통량을 공변과정으로 하는 것이 가장 좋았다.

평일 교통량 패턴을 살펴보면 수도권 지역의 서울 영업소는 '오전 7시–8시경'과 '오후 4–5시경'에 피크가 나타나며 이는 수도권 지역으로 출퇴근하는 교통량이 많다는 것을 알 수 있다. 또한, 관광 지역인 춘천과 강

릉 영업소는 '오후 12시–2시경'과 '오후 5시–6시경'에 피크가 나타나며 이는 점심시간의 이동과 타지역에서 일하고 퇴근하는 교통량이 많다는 것을 알 수 있고 수도권 지역과 달리 출근하는 교통량은 많지 않았다. 주말 교통량 패턴을 살펴보면 세 영업소 모두 '오후 12시–2시경', '오후 4시–6시경'에 피크가 나타났으며, 이는 점심과 저녁 시간의 이동으로 인한 교통량 증가로 보인다.

References

An SY (2017). *A study on the optimal traffic flow by highway section*. Master's thesis, Inje University.

Choo S, Lee SM, Park YI, and Yun J (2007). *A study on methods for constructing weekend origin/destination travel*, The Korea Transport Institute.

Hsing T and Eubank R (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley.

Hunt JD, McMillan P, Stefan K, and Atkins D (2005). Nature of weekend travel by urban households, *2005 Annual Conference of the Transportation Association of Canada*.

Jeon JM and Park BU (2020). Additive regression with Hilbertian responses, *The Annals of Statistics*, **48**, 2671–2697.

Kato T (1995). *Perturbation theory for linear operators*. Berlin: Springer-Verlag.

Mammen E, Linton O, and Nielsen JP (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *The Annals of Statistics*, **27**, 1443–1490.

Müller HG and Yao F (2008). Functional additive models, *Journal of the American Statistical Association*, **103**, 1534–1544.

Park BU, Chen CJ, Tao W, and Müller HG (2018). Singular additive models for function to function regression, *Statistica Sinica*, **28**, 2497–2520.

Park MJ (2015). Traffic prediction technology development using big data, *Monthly KOTI Magagine on Transport*, 42–46.

Sohn C and Kim GH (2014). Influences of weather on the inbound traffic volume of a tourist destination, *The Korea spatial planning review*, 99–111.

Yang W, Müller HG, and Stadtmüller U (2011). Functional singular component analysis, *Journal of the Royal Statistical Society: Series B Statistical Methodology*, **73**, 303–324.

Yoon SY, Lee CY, Kim HJ, Yook DH, and Kim SR (2017). A study on usability of big data to enhance reliability of regional travel demand forecasting, *Korea Research Institute for Human Settlements*.

Received August 30, 2021; Revised September 6, 2021; Accepted September 7, 2021

함수회귀분석을 통한 교통량 예측

이인주[a], 이영경[1,a]

[a]강원대학교 통계학과

요 약

교통량 예측은 지방 행정의 의사결정에 매우 중요한 정보를 제공한다. 교통량 예측을 통해 교통혼잡비용을 줄이고 지역경제를 활성화 함으로써 사회적, 경제적 이익을 창출할 수 있다. 교통량은 미지의 확률적 규칙하에서 시간의 흐름에 따라 궤적을 가지며 변화하는 함수데이터의 일종이다. 본 논문에서는 세 가지 함수회귀모형을 이용하여 과거에 관측된 교통량 궤적을 기반으로 미래의 관측되지 않은 교통량 궤적을 예측하는 방법을 제시한다. 본 논문에서 소개하는 세가지 방법은 전국 고속도로 영업소 중 서울, 춘천, 강릉 세 개 영업소에서 수집된 고속도로 영업소 데이터에 적용한다. 각 영업소 별로 세가지 방법의 예측오차를 비교함으로써 영업소별 최적 교통량 예측모형을 찾는다.

주요용어: 자기공분산, 교차공분산, 함수주성분, 함수특이성분, 평활역적합

이 논문은 한국연구재단 중견연구사업의 지원을 받아 수행된 연구임(NRF-2021R1A2C1003920)

[1]교신저자: (24341) 강원도 춘천시 강원대학길 1, 강원대학교 통계학과. E-mail: youngklee@kangwon.ac.kr