The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.1.109
31(1), 109–122

Time-varying modeling of the composite LN-GPD

Sojin Park[a] · Changryong Baek[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received November 2, 2017; Revised December 1, 2017; Accepted December 15, 2017)

Abstract

The composite lognormal-generalized Pareto distribution (LN-GPD) is a mixture of right-truncated lognormal and GPD for a given threshold value. Scollnik (*Scandinavian Actuarial Journal*, **2007**, 20–33, 2007) shows that the composite LN-GPD is adequate to describe body distribution and heavy-tailedness. This paper considers time-varying modeling of the LN-GPD based on local polynomial maximum likelihood estimation. Time-varying model provides significant detailed information of time dependent data, hence it can be applied to disciplines such as service engineering for staffing and resources management. Our work also extends to Beirlant and Goegebeur (*Journal of Multivariate Analysis*, **89**, 97–118, 2004) in the sense of losing no data by including truncated lognormal distribution. Our proposed method is shown to perform adequately in simulation. Real data application to the service time of the Israel bank call center shows interesting findings on the staffing policy.

Keywords: composite LN-GPD, local polynomial maximum likelihood estimation, call center

1. 서론

현대사회에서 데이터의 양은 점점 더 방대해지고 다양한 특성을 지니고 있다. 특히, 무시할 수 없는 극단값이 포함된 데이터가 다양한 산업군에서 나타나고 있다. 예를 들어, 강수량, 보험, 은행의 콜센터 데이터 등이 있다. 일반적으로 극단값을 포함하고 있는 자료는 꼬리가 두껍고 편향된 분포 형태를 띄며 로그정규분포(lognormal distribution; LN)나 일반화파레토분포(generalized Pareto distribution; GPD)를 사용한다. 하지만 로그정규분포는 두터운 꼬리를 설명하는데 한계가 있으며 GPD분포의 경우 자료를 임계값(threshold)를 기준으로 절삭하기에 데이터의 손실이 있을 뿐만 아니라 몸통(body) 부분의 특징을 반영하지 못한다. 이러한 단점을 극복하기 위해서 Cooray와 Ananda (2005), Scollnik (2007) 등은 로그정규분포와 GPD 분포의 합성(LN-GPD)을 제안하였고 합성된 분포가 실증자료를 더 잘 적합함을 보였다.

예를 들어 Figure 1.1은 Kim 등 (2016)의 실증 분석 결과를 재현한 것으로 임계값 63에 대한 QQ-plot을 나타낸다. 왼쪽 위 그림은 로그정규분포를 이용한 QQ-plot으로 얼핏 보기에는 로그정규분포의 적합도가 좋아 보이지만 확대한 아래 패널에서 살펴보듯이 작은 값들의 경우 적합도가 떨어짐을 알 수 있다.

This research was supported by the Basic Science Research Program from the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1A1A1A05000831).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: crbaek@skku.edu

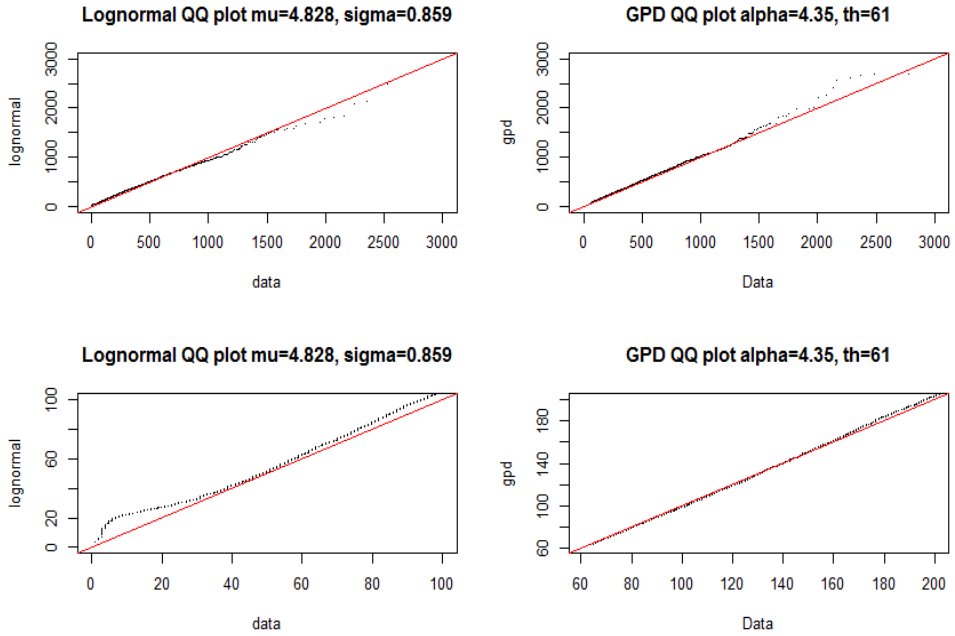

Figure 1.1. QQ plot of lognormal distribution and GPD. GPD = generalized Pareto distribution.

이는 곧 두터운 꼬리를 설명하기 위해서 더 큰 값의 모수들을 추정하여 작은 값들에 대한 적합도가 나빠지기 때문이다. 반면 GPD 분포의 경우 QQ-plot을 확대하여 보더라도 비틀림 없이 꼬리 부분에 대한 적합도가 매우 좋지만 자료에 대한 손실이 있어 로그정규분포 혹은 GPD 분포 단독으로는 자료에 대한 특징을 충분히 설명하기 부족하다. 이는 곧 LN-GPD 합성분포를 사용한다면 이러한 단점을 잘 보완할 수 있음을 보여준다.

본 논문은 선행 연구를 확장하여 시간에 따라 변하는 LN-GPD 평균 모형에 대해서 연구한다. 은행 콜센터의 서비스 시간에 대해서 생각해봤을 때, 실시간으로 평균 서비스 시간을 LN-GPD 분포로 정확히 알아낸다면 이를 토대로 고객의 평균 대기 시간 및 평균 지연 시간을 추정해 내어 콜센터의 운영에 있어서 몇 명의 상담원을 두어야 할지, 경력이 많은 상담원을 어느 시간에 배치하여 서비스의 질을 높일지 등에 대한 전반적인 이해를 돕는데 훨씬 유용한 정보를 제공해주어 시간에 따라 변하는 모형이, 더욱 두터운 꼬리를 적절히 설명해주는 모형에 대한 연구는 반드시 필요하다.

방법론에 있어서 우리가 제안한 방법은 국소다항최대우도추정법으로 효율성이 높은 방법이다. 하지만, LN-GPD 분포의 추정에 있어서 임계값 역시 모수이기 때문에 추정상에 많은 어려움을 겪음을 Kim 등 (2016)이 보고하였다. 이를 극복하기 위해서 Kim 등 (2016)은 2단계에 걸쳐 모수를 추정하는 방법, 즉 임계값을 비모수적인 방법으로 먼저 추정하고 이를 토대로 나머지 모수를 추정하는 방법을 제안하였다. 본 논문에서 고려한 시간에 따라 변하는 LN-GPD 모형의 경우, 국소화된 추정값을 다루기 위해 적은 개수의 자료를 통해 많은 모수를 추정해야 하는 문제에 봉착하게 된다. 또한 적절한 띠너비의 선택이 제안한 방법의 전반적인 성능에 있어서 매우 중요한 역할을 한다. 따라서, 본 논문은 2단계에 걸쳐 주어진 임계값에 대해서 국소다항최대우도추정법으로 시간에 따라 변화하는 LN-GPD 모형을 추정하는 방법에 대해서 소개하고 교차타당성을 통해 띠너비를 선택하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 로그정규분포와 GPD에서 각각 시간에 따라 변하는 모수 추정방법에 대해 소개한다. 이를 확장하여 국소다항최대우도법을 바탕으로 LN-GPD에서의 시간에 따라 변하는 모수 추정 방법을 제안한다. 제 3장에서는 모의실험을 통해 우리가 제안한 방법이 시간에 따라 변하는 모수들을 잘 추정함을 보이고, 제 4장에서는 이스라엘 은행의 콜센터 실증자료 분석을 통해 기존의 로그정규분포만을 이용한 결과와 비교하여 새로운 결과를 보고하고 그 의미를 논의한다. 마지막으로 제 5장에서는 본 논문에서 제시한 모형의 장단점과 적용방안에 대해 논의하고 마무리한다.

2. 방법론

본 장은 로그-정규분포와 일반화파레토분포에서의 시간에 따라 변하는 모수 추정 방법에 대해서 먼저 소개하고, 최종적으로 LN-GPD 합성 분포에서 시간에 따라 변하는 모수 추정 방법을 제안한다. 본 논문에서는 시간에 대한 공변량 X에 대해서 종속변수 Z를 관측하는 모형으로 자료의 순서쌍을 (X_i, Z_i), $i = 1, \ldots, n$으로 나타낸다. 즉, 조건부 분포 $Z|X = x$의 모형 가정에 의해 조건부 평균값인

$$E(Z|X = x)$$

를 시간에 따라 모수가 부드럽게 변하는 성질을 반영하여 추정하는 방법을 소개하고자 한다.

2.1. 시간에 따라 변하는 로그정규분포 모형

시간 공변량 X에 대해서 조건부 분포 $Z|X = x$가 로그정규분포를 따르는 경우를 로그정규분포에서 시간에 따라 변하는 모형이라 정의한다. $Y = \log(Z)$로 Z에 대한 로그 변환 후, $Y|X = x$는 평균이 $\mu(x)$이고 분산이 $\sigma^2(x)$인 정규 분포를 따른다. 따라서, 자료 (X_i, Y_i), $i = 1, \ldots, n$에 대해서 다음과 같이 표현 할 수 있다.

$$Y_i = \mu(X_i) + \sigma(X_i)\epsilon_i, \quad \epsilon_i|X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1).$$

또한, 로그정규분포의 q-차 적률은

$$E(Z^q|X = x) = \exp\left[q\mu(x) + \frac{q^2\sigma^2(x)}{2}\right]$$

으로 주어지므로 시간에 따라 변하는 $\mu(x)$ 및 $\sigma^2(x)$를 추정하여 로그정규분포의 평균값 $\nu(x) = \exp(\mu(x) + \sigma^2(x)/2)$ 등을 간편하게 계산할 수 있다. 따라서 로그정규분포의 시간에 따라 변하는 모수인 $\mu(x)$ 및 $\sigma(x)^2$를 추정하는 것이 핵심이다. 본 논문은 Shen과 Brown (2006)의 비모수적인 방법으로 국소다항회귀(local polynomial regression)에 기반한 방법을 사용하였으며 다음과 같다.

테일러 전개에 의해서 x_0의 근방에서

$$\mu(x) \approx \sum_{j=0}^{p} \beta_j (x - x_0)^j$$

이므로 국소다항회귀를 이용한 $\mu(\cdot)$의 추정은 커널함수 $K(\cdot)$에 의해 결정되는 가중화된 제곱합을 최소화 시키는 계수값으로 정의된다.

$$\left(\hat{\beta}_0, \ldots, \hat{\beta}_p\right) = \underset{\beta_0, \ldots, \beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1(X_i - x_0) - \cdots - \beta_p(X_i - x_0)^p\right]^2 K_h(X_i - x_0).$$

커널 $K_h(\cdot) = h^{-1}K(\cdot/h)$는 띠너비 $h > 0$를 가지는 커널함수로 본 논문에서는 정규분포를 사용하였다. 따라서, 시간 x_0에서의 추정값은

$$\hat{\mu}(x_0) = \hat{\beta}_0$$

이다.

분산 모수인 $\sigma^2(\cdot)$의 추정은 2단계로 차분에 기초한 분산 추정치이다. 첫 번째로, 데이터 전처리 작업을 위해 $\{X_i, Y_i\}$, $i = 1, \ldots, n$을 다음과 같이 두 개의 그룹으로 분리한다. 이 때, 순서쌍 $\{X_i, Y_i\}$는 시간변수 X_i의 크기순으로 정렬한 자료이다.

$$(X_{2i-1}, Y_{2i-1}; X_{2i}, Y_{2i}), \quad i = 1, \ldots, \left[\frac{n}{2}\right].$$

분산 함수 $\sigma^2(\cdot)$를 추정을 위해 유사 잔차 제곱(squared pseudo-residual)을

$$D_{2i} = \frac{(Y_{2i} - Y_{2i-1})^2}{2}$$

라고 하면 적절한 가정하에서 유사 잔차 제곱은 근사적으로 $\chi^2(1)\sigma^2(X_{2i})$를 따르고 $E(D_{2i}|X_{2i}) = \sigma^2(X_{2i})$이고 $\mathrm{Var}(D_{2i}|X_{2i}) = 2\sigma^4(X_{2i})$가 됨을 보일 수 있다. 유사 잔차 제곱을 이용한 자세한 추정 방법 및 이론적 성질은 Brown과 Levine (2007)을 참고하기 바란다. 따라서 자료쌍 (X_{2i}, D_{2i}), $i = 1, \ldots, [n/2]$에 대해서 다음의 모형 가정을 통해서

$$D_{2i} = \sigma^2(X_{2i}) + \sqrt{2}\sigma^2(X_{2i})\varepsilon_{2i}, \quad \varepsilon_{2i}|X_{2i} \overset{\text{iid}}{\sim} \mathcal{N}(0,1),$$

$i = 1, \ldots, [n/2]$인 선형 회귀 방정식을 얻어낼 수 있다. 여기서 주의할 점은 X_{2i}가 주어질 경우에 D_{2i}는 각각 서로 겹치지 않은 자료에서 생성되었으므로 독립인 자료들이다. 따라서 국소다항회귀를 이용하여 손쉽게 추정할 수 있다.

x_0 근방에서의 테일러 전개를 이용하며, $\sigma^2(\cdot)$는 음수가 될수 없으므로,

$$\sigma^2(x) \approx \exp\left(\sum_{j=0}^{p} \alpha_j (x - x_0)^j\right)$$

임을 이용하면

$$(\hat{\alpha}_0, \ldots, \hat{\alpha}_p) = \underset{\alpha_0, \ldots, \alpha_p}{\operatorname{argmin}} \sum_{i=1}^{\left[\frac{n}{2}\right]} \left[D_{2i} - \exp\left(\sum_{j=0}^{p} \alpha_j (X_{2i} - x_0)^j\right)\right]^2 K_h(X_{2i} - x_0)$$

으로 추정 가능하고

$$\sigma^2(x_0) = \exp(\hat{\alpha}_0)$$

이다. 따라서 로그정규분포의 평균은

$$\hat{\nu}(x_0) = \exp\left(\hat{\beta}_0 + \frac{e^{\hat{\alpha}_0}}{2}\right)$$

으로 주어진다.

커널을 이용한 비모수적인 방법에서는 띠너비의 선택이 매우 중요하다. Shen과 Brown (2006)에서는 자료에 적응하는 방법인 다중 교차 타당성(multifold cross-validation)을 이용한 방법을 제안하였

다. K차 교차 타당성이란, 데이터를 K개의 그룹으로 동일하게 나눈 뒤 $(K-1)$개의 그룹을 훈련 집합(training set)으로 두어 모수를 추정한 다음에 이 모수를 이용하여 나머지 선택되지 않은 그룹인 검증 집합(test set)의 값을 추정하여 그 오차를 최소화하는 띠너비를 선택하는 방법이다. 예를 들어 $\mu(x)$의 추정을 생각하자. 검증 집합에 속한 시간 x_i에 상응하는 Y_i의 추정값은 i번째 데이터가 속하지 않은 훈련 집합으로부터 주어진 띠너비 h에 대해서 얻어진 모수의 추정치, $\hat{\beta}_j^{-K(i)}$, $j=0,\ldots,p$이고 $\hat{\mu}(x_i)=\beta_0^{-K(i)}$로 주어진다. 따라서 제곱 오차를 최소화하는 띠너비는

$$\hat{h}=\underset{h}{\operatorname{argmin}}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0^{-K(i)}\right)^2$$

으로 주어진다. K는 일반적으로 5나 10을 사용하는데, 여기서는 5차 교차 타당성을 사용하였다.

2.2. 시간에 따라 변하는 GPD 모형

주어진 시간에 대한 조건부 분포 $Z|X=x$가 꼬리 지수(tail index) $\gamma(x)$, 임계값 $\theta(x)$ 및 $\tau(x)$ 모수를 가지는 일반화파레토를 따를 때, 우리는 시간에 따라 변하는 GPD 모형이라 부른다. 일반화파레토 분포의 꼬리지수는 모든 실수값이 가능하나 본 연구에서는 두터운 꼬리(heavy-tailed)를 가지는 모형에 관심이 있으므로 $\gamma(x)$가 양수값을 가지는 경우로 한정한다. 양수값을 가지는 $\gamma(x)$에 대해서 조건부 누적확률 분포 함수는

$$F(y)=1-\left(1+\gamma(x)\frac{y-\theta(x)}{\tau(x)}\right)^{-\frac{1}{\gamma(x)}},\quad y\geq\theta(x),\ \tau(x)>0,\ \gamma(x)>0$$

으로 주어지며, 조건부 평균 $E(Z|X=x)$은

$$\nu(x)=\theta(x)+\frac{\tau(x)}{1-\gamma(x)}$$

이다. 따라서 적절한 모수들의 추정을 통해 조건부 평균을 추정할 수 있다. Beirlant와 Goegebeur (2004)는 주어진 임계값에 대해서 국소다항최대우도(local polynomial maximum likelihood) 방법에 기반하여 시간에 따라 변하는 모수 $\gamma(x)$, $\tau(x)$의 추정법을 다음과 같이 제안하였다. 테일러 급수에 의해서 x_0 근방에서

$$\gamma(x)\approx\sum_{j=0}^{p}\beta_{1j}(x-x_0)^j,\quad \tau(x)\approx\exp\left(\sum_{j=0}^{p}\beta_{2j}(x-x_0)^j\right)$$

임을 이용하면 조건부 확률밀도함수는

$$g(z;\theta(x),\gamma(x),\tau(x))=\frac{1}{\tau(x)}\left(1+\gamma(x)\frac{z-\theta(x)}{\tau(x)}\right)^{-\frac{1}{\gamma(x)}-1}$$

로 주어진다. 따라서 주어진 임계값 $\theta(X_i)$에 대해서 국소다항최대우도 추정값은

$$\left(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2\right)=\underset{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2}{\operatorname{argmax}}\frac{1}{n}\sum_{i=1}^{n}\log g(Z_i;\theta(X_i),\gamma(X_i),\tau(X_i))K_h(X_i-x_0)\tag{2.1}$$

으로 주어진다. 여기에서 $\boldsymbol{\beta}_1=(\beta_{10},\ldots,\beta_{1p})$, $\boldsymbol{\beta}_2=(\beta_{20},\ldots,\beta_{2p})$는 $\gamma(x)$ 및 $\tau(x)$의 모수이며 K_h는 정규분포 커널함수이다. 따라서 조건부 평균은

$$\hat{\nu}(x_0)=\theta(x_0)+\frac{\exp\left(\hat{\beta}_{20}\right)}{1-\hat{\beta}_{10}}$$

이다.

로그정규분포의 경우와 마찬가지로 적절한 띠너비 h의 선택이 추정치의 성능을 크게 좌우한다. 본 논문에서는 다중교차타당성을 바탕으로 선택하였으며 음의로그밀도함수(negative loglikelihood)를 손실함수로 사용하였다. 즉 K차 교차 타당성에서 훈련표본을 통해 얻어진 모수에 대해서

$$\hat{h} = \underset{h}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} -\log g\left(Z_i; \theta(X_i), \hat{\gamma}^{-K(i)}(X_i), \hat{\tau}^{-K(i)}(X_i)\right)$$

이며 $\hat{\gamma}^{-K(i)}(X_i)$, $\hat{\tau}^{-K(i)}(X_i)$는 X_i가 속하지 않은 훈련표본을 이용한 모수 추정값이다. 본 논문에서는 임계값 $\theta(X_i)$에 대해서는 시간에 의존하지 않고 상수로 주어진다고 가정하였다.

2.3. 시간에 따라 변하는 LN-GPD 모형

합성 분포 함수는 주어진 임계점 θ와 표준화 상수(normalizing constant) 및 확률밀도함수 f_1, f_2에 대해서 밀도함수를 다음과 같이 이어 붙인 분포이다.

$$f(z) = \begin{cases} cf_1(z), & 0 < z \le \theta, \\ cf_2(z), & \theta < z. \end{cases} \tag{2.2}$$

로그정규분포와 GPD 분포의 합성 모형인 LN-GPD은 Scollnik (2007), Nadarajah와 Bakar (2014) 등에서 연구되었으며, 확률밀도함수는

$$\begin{aligned} f\left(z; \mu, \sigma^2, \gamma, \tau\right) &= r f_1\left(z|\mu, \sigma^2, \theta\right) \mathbf{1}_{\{z \le \theta\}} + (1-r) f_2(z|\alpha, \tau, \theta) \mathbf{1}_{\{z > \theta\}}, \\ f_1\left(z|\mu, \sigma^2, \theta\right) &= \frac{1}{\Phi\left(\frac{\log\theta - \mu}{\sigma}\right)} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2}, \\ f_2(z|\gamma, \tau, \theta) &= \frac{1}{\tau}\left(1 + \gamma\frac{z-\theta}{\tau}\right)^{-\frac{1}{\gamma}-1}, \quad \tau > 0,\ \gamma \in (0,1) \end{aligned} \tag{2.3}$$

으로 나타낼 수 있다. 즉, $f_1(x)$는 로그정규분포의 밀도 함수이고, $f_2(x)$는 일반화 파레토 분포의 밀도 함수이다. 하지만 LN-GPD 모델은 두 분포의 혼합(mixture)이 아니라 두 분포를 임계점을 기준으로 이어 붙인 분포이므로 임계점 θ를 기준으로 그 보다 작은 값은 절삭된 로그정규분포(truncated log-normal distribution)을 따르고 임계점 보다 큰 값은 일반화 파레토 분포를 따르는 분포이다. Scollnik (2007)은 밀도 함수가 임계점에서 연속적이고 미분 가능 할 수 있는 분포가 되도록 제안하였다. 하지만, 본 논문에서는 부드럽게 이어 붙인 경우를 생각하지 않고 일반적인 두 분포 함수의 합성에 대해서 고려하였다. 또한, 임계점 θ의 경우 시간에 따라 변하지 않은 모형으로 가정하였다. 그 이유는 Kim 등 (2016)에서 밝혔듯이 부드럽게 이어 붙인 경우 LN-GPD 모형의 임계점을 포함한 모수들의 추정이 수치적으로 매우 불안정하기 때문이다.

본 논문은 국소다항최대우도법을 이용한 시간에 따라 변하는 모수의 추정을 다음과 같이 제안한다. x_0 근방에서의 테일러 급수 전개에 따라

$$\begin{aligned} \mu(x) &\approx \sum_{j=0}^{p} \beta_{1j}(X_i - x_0)^j, \quad \sigma(x) \approx \exp\left(\sum_{j=0}^{p} \beta_{2j}(X_i - x_0)^j\right), \\ \gamma(x) &\approx \sum_{j=0}^{p} \beta_{3j}(X_i - x_0)^j, \quad \tau(x) \approx \exp\left(\sum_{j=0}^{p} \beta_{4j}(X_i - x_0)^j\right) \end{aligned}$$

으로 근사시킬 수 있다. 모수 벡터 $\boldsymbol{\beta}_1 = (\beta_{10}, \ldots, \beta_{1p}), \ldots, \boldsymbol{\beta}_4 = (\beta_{40}, \ldots, \beta_{4p})$에 대해서 커널가중우도함수 식 $L_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$은 다음과 같다.

$$L_n(\beta_1, \beta_2) = \frac{1}{n}\sum_{i=1}^{n} \log f(Z_i; \mu(X_i), \sigma(X_i), \gamma(X_i), \tau(X_i)) K_h(X_i - x_0) \tag{2.4}$$

따라서, 모수들의 추정값은

$$\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\beta}}_4\right) = \underset{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_4}{\operatorname{argmin}} L_n\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4\right) \tag{2.5}$$

이다. 주어진 임계점 θ에 대해서 합성분포함수는 절삭된 로그정규분포와 GPD 함수의 조합이므로 합성분포함수를 통해 모든 모수를 한꺼번에 추정하는 것보다는 각각 추정하는 방법이 수치적으로 훨씬 간단하다. 즉,

$$\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2\right) = \underset{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \log f_1(Z_i; \mu(X_i), \sigma(X_i)) \mathbf{1}_{\{Z_i \le \theta\}} K_{h1}(X_i - x_0), \tag{2.6}$$

$$\left(\hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\beta}}_4\right) = \underset{\boldsymbol{\beta}_3, \boldsymbol{\beta}_4}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \log f_2(Z_i; \gamma(X_i), \tau(X_i)) \mathbf{1}_{\{Z_i > \theta\}} K_{h2}(X_i - x_0) \tag{2.7}$$

을 통해 보다 빠르게 모수를 추정할 수 있다. 한편으로는 절삭된 로그정규분포와 GPD에 대해서 서로 다른 띠너비 값을 사용할 수 있기에 보다 정확한 추정값을 기대할 수 있다. K차 교차 타당성에 의해서

$$\hat{h}_1 = \underset{h}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} -\log f_1\left(Z_i; \hat{\mu}(X_i)^{-K(i)}, \hat{\sigma}(X_i)^{-K(i)}\right) \mathbf{1}_{\{Z_i \le \theta\}},$$

$$\hat{h}_2 = \underset{h}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} -\log f_2(Z_i; \hat{\gamma}\left(X_i\right)^{-K(i)}, \hat{\tau}(X_i)^{-K(i)}\right) \mathbf{1}_{\{Z_i > \theta\}}$$

으로 띠너비를 구할 수 있다.

또한, LN-GPD 모형에서의 평균값 역시

$$\begin{aligned}\nu(x_0) &= E(Z|X = x_0) = E(Z|X = x_0, Z \le \theta)P(Z \le \theta) + E(Z|X = x_0, Z > \theta)P(Z > \theta) \\ &= \log\left(\mu(x_0) - \frac{\phi\left(\frac{\theta - \mu(x_0)}{\sigma(x_0)}\right)}{\Phi\left(\frac{\theta - \mu(x_0)}{\sigma(x_0)}\right)}\right) P(Z \le \theta) + \left(\theta + \frac{\tau(x_0)}{1 - \gamma(x_0)}\right) P(Z > \theta)\end{aligned}$$

으로 각각 표준정규분포의 밀도함수 및 누적분포함수인 ϕ, Φ에 결정되므로

$$P(Z \le \theta) \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Z_i \le \theta\}}$$

와 모수 추정값 (2.6)–(2.7)을 대입하여

$$\hat{\nu}(x_0) = \log\left(\hat{\beta}_{10} - \frac{\phi\left(\left(\theta - \hat{\beta}_{10}\right)\right) / \exp\left(\hat{\beta}_{20}\right)}{\Phi\left(\left(\theta - \hat{\beta}_{10}\right)\right) / \exp\left(\hat{\beta}_{20}\right)}\right) \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Z_i \le \theta\}} + \left(\theta + \frac{\exp\left(\hat{\beta}_{40}\right)}{1 - \hat{\beta}_{30}}\right) \sum_{i=1}^{n} \mathbf{1}_{\{Z_i > \theta\}}$$

으로 구할 수 있다.

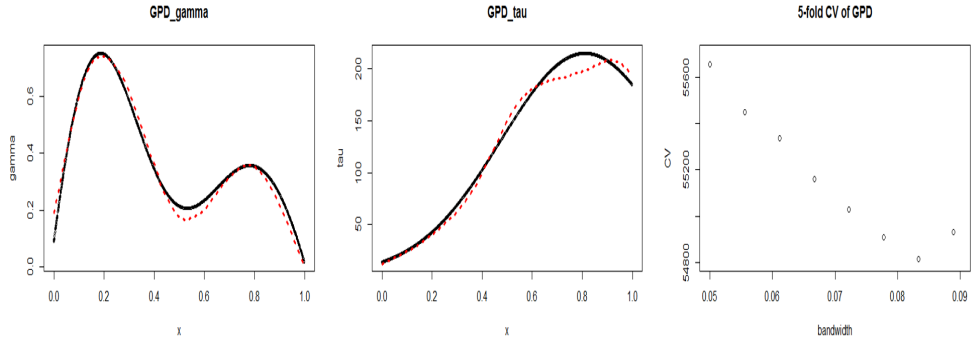

Figure 3.1. $\gamma(x)$, $\tau(x)$, and 5-fold cross-validation ($h = 0.082$). GPD = generalized Pareto distribution; CV = coefficient of variation.

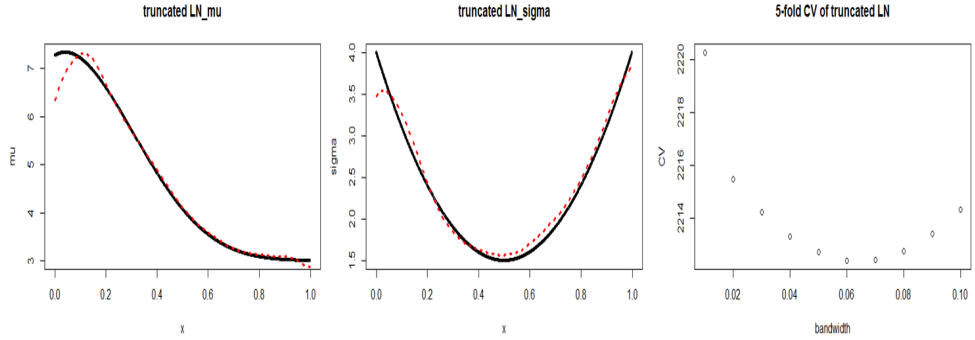

Figure 3.2. $\mu(x)$, $\sigma(x)$, and 5-fold cross-validation ($h = 0.06$). LN = lognormal distribution; CV = coefficient of variation.

3. 모의실험

본 장에서는 2.3절에서 소개한 국소다항최대우도법을 이용하여 LN-GPD에서 시간에 따라 변하는 모수들의 함수를 추정할 때, 그 성능을 모의실험을 통해 확인하고자 한다. 본 모의실험에서는 다음의 LN-GPD 모수 함수들을 임의로 생성하였다. 여기서 $\{X_i\}$는 iid $U(0,1)$을 따르며, 임계값 $\theta = 10$, 표본수 $n = 5000$으로 하였다. 생성한 모형은 다음과 같다. 먼저 GPD의 모수는

$$\gamma(x) = 3 + 6(x + 0.3)e^{-8x^2} + 2(x + 0.3)e^{-4(x-0.7)^2} - 4.8,$$
$$\tau(x) = 3 + 6(x + 0.3)e^{-8x^2} + 200(x + 0.3)e^{-4(x-0.7)^2}$$

이고 절삭된 로그정규분포의 경우

$$\mu(x) = 3 + 6(x + 0.3)e^{-8x^2} + 2(x + 0.3)e^{-4(x-0.7)^3},$$
$$\sigma(x) = 1.5 + 10(x - 0.5)^2$$

을 사용하였다.

위와 같이 생성된 자료에 대해서 2.3절에서 소개한 주어진 임계값에 대해서 그 보다 작은 값들은 절삭된 로그정규분포를 큰 값들은 GPD분포를 이용하여 추정한 방법을 적용하였다. 다항함수의 차수 $p = 1$로

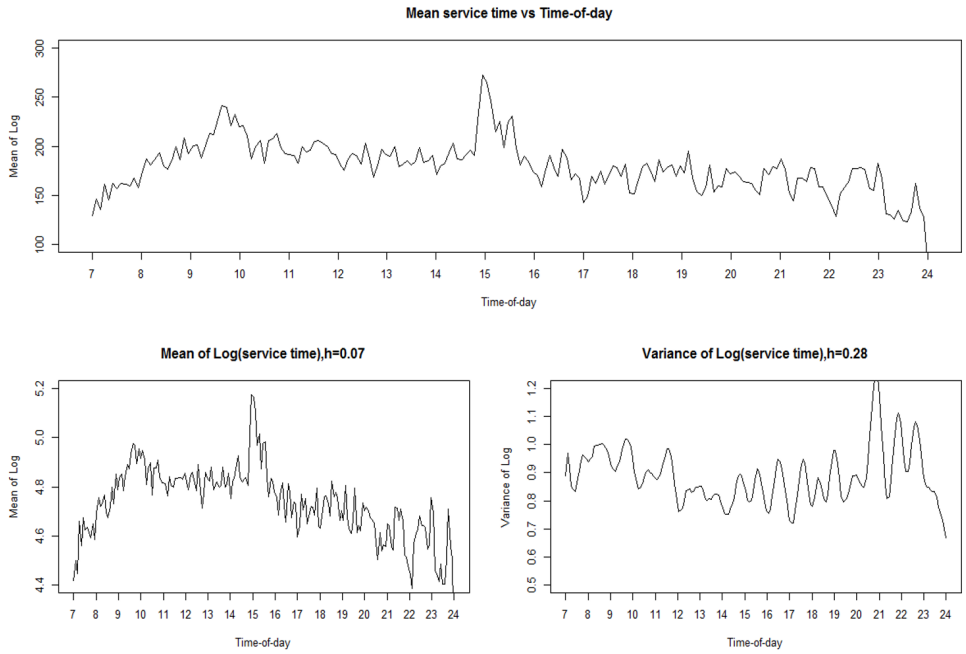

Figure 4.1. Mean service time and parameters of lognormal distribution (local linear regression).

두어 국소선형(local linear) 모형을 적용하였으며 띠너비 선택을 위해서 5차 교차타당성을 이용하였다. 그 결과를 Figure 3.1과 Figure 3.2에 나타냈다. 검은색 실선이 참 모수의 값이며 붉은색 점선이 추정된 값이다. 전반적으로 참 모수와 매우 가깝게 적절하게 추정이 됨을 알 수 있으나 양측 가장자리가 어긋나는 것을 관찰할 수 있는데, 이는 비모수추정에서 발생하는 경계 효과(boundary effect) 때문인 것으로 사료된다. 따라서, 우리가 제안한 시간에 따라 변하는 LN-GPD 모형의 추정방법이 참값을 잘 추정하고 있음을 살펴볼 수 있다. 제안한 추정법의 성능을 크게 좌우하는 띠너비의 추정에 있어서도 각 그림의 오른쪽 패널에서 교차타당성 함수가 볼록함수의 형태로 $\hat{h}_1 = 0.082$, $\hat{h}_2 = 0.06$을 찾아내어 수치적으로 안정된 값을 추정함을 확인할 수 있다.

4. 실증분석

본 장에서는 본 논문이 제안한 시간에 따라 변하는 모형에 대한 실증 분석 결과를 보고한다. 우리가 사용한 자료는 1999년 11월부터 12월까지 이스라엘 은행의 콜센터에서 수집된 자료로 Shen과 Brown (2006) 논문에서 사용한 자료이며, 총 표본의 수는 46,762이다. 분석하고자 하는 변수는 시간(time-of-day)에 따른 콜센터 서비스 시간(service time)으로, 이는 은행 고객이 콜센터 상담원과 통화한 평균 서비스 시간을 시간대별로 보고자 한다.

먼저 기본이 되는 모형은 Shen과 Brown (2006)에서 분석한 로그정규분포를 이용한 콜센터 서비스 평균 시간에 대한 추정이다. Figure 4.1은 $p = 1$인 국소선형모형을 이용한 시간에 따른 모수의 추정값을 나타내고 있으며, 5차 교차 타당성으로 띠너비를 선택하였다. 로그정규분포의 평균이 $\exp\left(\mu(x) + \sigma^2(x)/2\right)$으로 주어짐을 고려한다면 시간에 따라 변하는 로그정규분포 모형을 사용했을 때

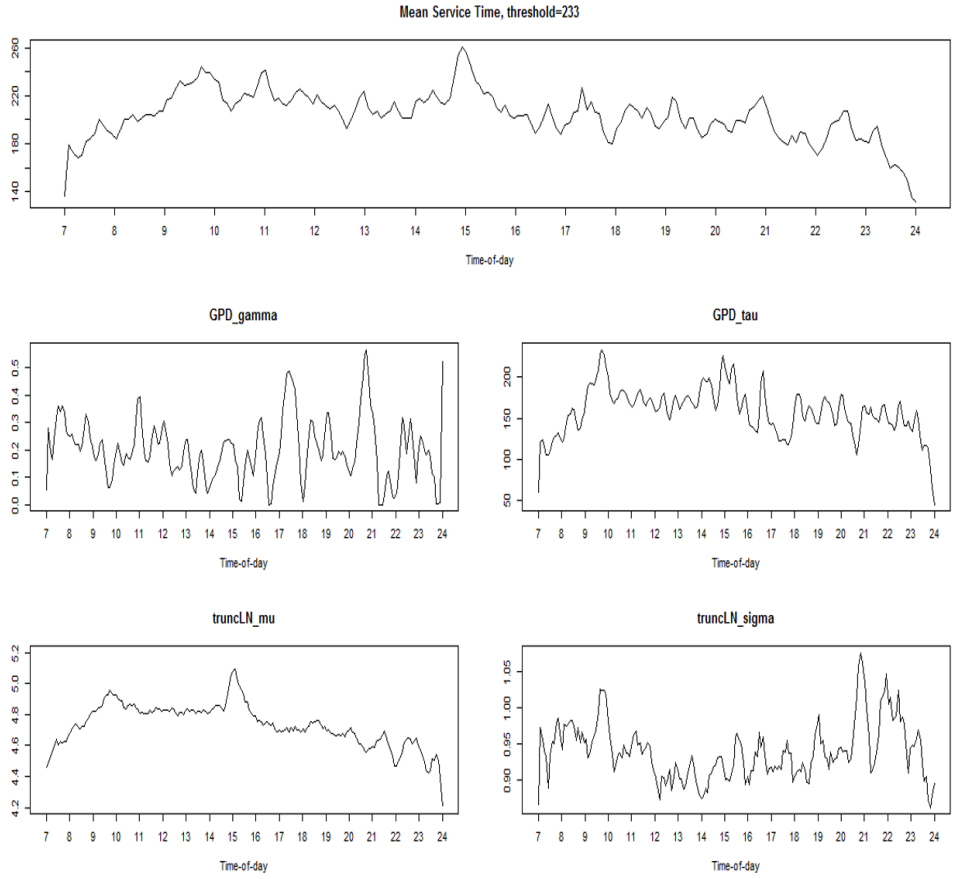

Figure 4.2. Mean service time and parameters of LN-GPD (local polynomial MLE, $\theta = 233$). LN-GPD = lognormal-generalized Pareto distribution; MLE = mean squared error.

평균이 $\mu(x)$의 모수값에 의해서 결정됨을 알 수 있다. 예를 들어, 평균 서비스 시간은 오전 9시30분과 오후 3시에 가장 크며 저녁 이후에는 대체로 평균 서비스 시간이 줄어듦을 알 수 있다.

하지만 Kim 등 (2016)을 비롯한 Figure 1.1에서 살펴보았듯이 로그정규분포와 일반화파레토분포는 매우 구분하기 어려우며 본 자료는 LN-GPD 분포로 적합하였을 때, 더 좋은 결과를 보임을 밝혔다. 따라서 콜센터 서비스 시간이 LN-GPD를 따를 때 국소다항최대우도법을 이용하여 추정한 방법을 적용해보았다. 우리가 제안한 방법의 경우 수치적인 안정성을 위하여 임계값을 주어진 값으로 가정하였기에 제안한 방법을 적용하기 위해서는 적절한 임계값을 먼저 추정해야 한다. 임계값은 Kim 등 (2016)에서 제안한 콜모고로프-스미르노프 검정법을 통해 233으로 추정한 값을 먼저 사용하였고, 그 결과는 Figure 4.2에 나타내었다. 시간에 따른 평균 서비스 시간에 대한 패턴은 로그정규분포를 사용하였을 때와 비슷한 형태를 띤다. 오전 9시 30분 및 오후 3시 근처에서 가장 높은 평균 서비스 시간을 보이고 있으며, 놀랍게도 GPD 모수인 $\tau(x)$ 및 로그정규분포의 $\mu(x)$와 매우 흡사함을 알 수 있다. 또한 GPD 분포의 꼬리 지수 값인 $\gamma(x)$의 값이 모두 0.5보다 작아 분산이 유한인 모형을 따르고 있으며 로그정규분포의 $\sigma(x)$는 그 값의 변화가 대동소이하다. 이는 곧 위 두 시간대에서의 평균서비스 시간이 큰 것은 소수의

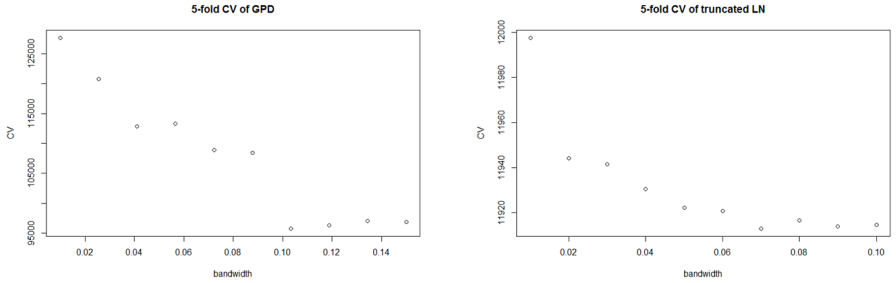

Figure 4.3. 5-fold cross-validation of GPD and truncated lognormal ditribution, $\theta = 233$. CV = coefficient of variation; LN = lognormal distribution;GPD = generalized Pareto distribution .

아주 큰 극단값에 의한 것 보다는 전반적으로 긴 평균서비스 시간 때문임을 알 수 있다. 따라서 이 시간대의 평균서비스 시간을 줄이기 위해서는 많은 수의 사람을 고용하는 것이 좋은 해결책임을 알 수 있다. 본 방법론을 적용할 때 필요한 띠너비 선택은 5차 교차타당성을 h가 0.103에서, 절삭된 로그정규분포에서는 0.07일 때 가장 작은 coefficient of variation (CV)값을 갖는 것을 Figure 4.3에서 확인할 수 있다.

임계값을 최대우도추정량으로 추정할 경우 Kim 등 (2016)은 435를 제안하였고, 이 때 시간에 따른 평균 서비스 시간에 대한 추정은 Figure 4.4에서 찾아볼 수 있다. 앞서 살펴본 두 경우와는 조금 다른 형태의 시간에 따른 평균 서비스 시간을 보여주고 있다. 오전 9시 30분과 오후 3시의 뚜렷한 봉우리를 가지는 쌍봉형의 그림이 아니라 크기가 비슷한 여러 봉우리를 가지는 패턴을 보여준다. 오전 7시 30분, 오전 11시, 오후 3시, 오후 5시 30분, 오후 8시 30분 등을 찾아볼 수 있다. 그 이유를 찾기 위해 LN-GPD 분포의 모수를 좀 더 자세히 살펴보면 우선 GPD의 모수 $\tau(x)$ 및 로그정규분포의 모수 $\sigma(x)$는 앞선 두 경우와 비슷한 양상을 띠고 있으나, GPD 분포의 꼬리지수인 $\gamma(x)$의 그래프의 경우 0.5를 넘는 값들이 오전 7시 30분, 오후 5시 30분, 오후 8시 30분 등에 나타남을 알 수 있다. 따라서 우리는 임계값이 435로 높을 경우의 분석을 통해서 오전 7시 30분, 오후 5시 30분 및 오후 8시 30분의 경우 큰 값의 평균 서비스 시간은 소수의 긴 통화 시간 때문임을 알 수 있다. 즉, 위 특별한 시간대의 경우, 아침 일찍 혹은 퇴근 직후나 저녁 이후에는 꽤 복잡한 문제를 문의해오는 사용자가 많음을 알 수 있다. 이 시간대의 평균 서비스 시간을 줄이기 위해서는 숙련된 상담원을 배치함으로써 복잡한 문제를 해결할 수 있다. 띠너비 값의 경우 GPD에서는 0.12, 절삭된 로그정규분포에서는 0.08임을 5차 교차 타당성을 통해 선택하였고 Figure 4.5에 나타내었다.

5. 결론

본 논문은 시간에 따라 변하는 모수를 가지는 LN-GPD 모형을 국소다항최대우도법을 이용하여 추정하는 방법에 대해서 연구하였다. LN-GPD 모형은 몸통(body) 부분은 로그정규분포를, 꼬리 부분은 일반화파레토 분포를 사용하여 두터운 꼬리를 갖는 자료를 자료의 손실 없이 분석할 수 있는 매우 유용한 분포이다. 하지만 임계값 역시 모수로 최대우도법을 사용하기에는 많은 어려움이 있다. 따라서 본 논문은 Kim 등 (2016)에서 제안한 2단계 추정법을 통해 임계값을 먼저 추정하고, 주어진 임계값에 대해서 나머지 모수들을 추정하는 방법을 통해 모수를 효율적으로 추정하는 방법에 대해서 제안하였다. 모의실험을 통해 본 논문에서 제안한 방법이 추정을 적절히 하고 있음을 살펴 보았다. 이스라엘 은행 콜센터 자료 분석을 통해 기존의 로그정규분포를 사용할 경우에는 오전 9시 30분 및 오후 3시에 평균 서비스 시간이 큰 것 이외에도 오전 7시 30분, 오후 5시 30분, 오후 8시 30분의 경우 소수의 긴 통화들이 평균 서비스 시간을 크게 만듦을 밝혀 숙련된 상담원을 배치한다면 서비스의 질 및 비용을 줄일 수 있을 것이라

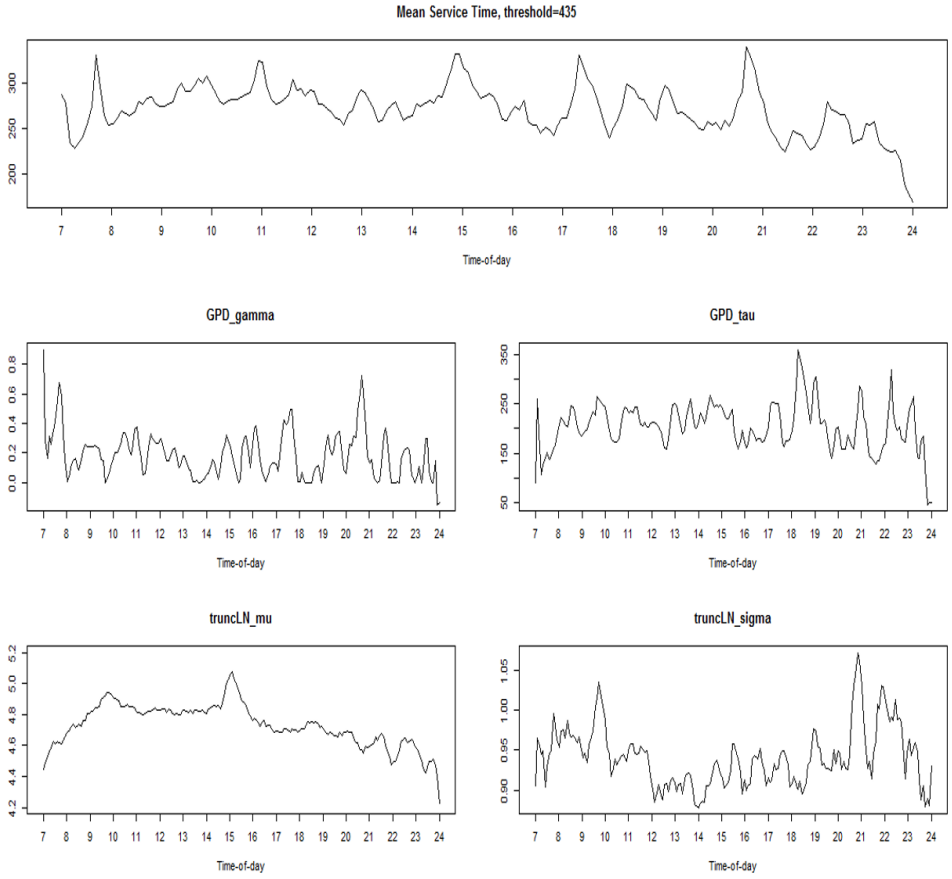

Figure 4.4. Mean service time and parameters of LN-GPD (local polynomial MLE $\theta = 435$). LN-GPD = lognormal-generalized Pareto distribution; MLE = mean squared error.

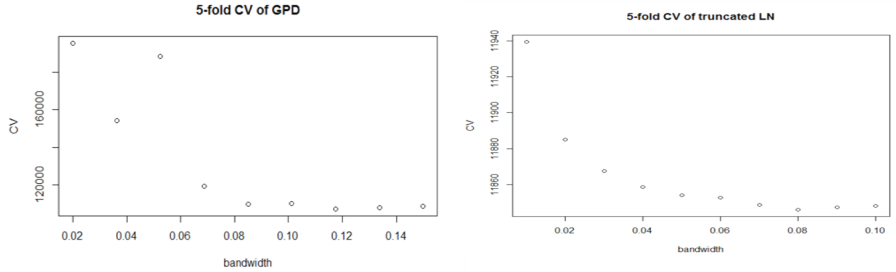

Figure 4.5. 5-fold cross-validation of GPD and truncated lognormal ditribution, $\theta = 435$. GPD = generalized Pareto distribution; CV = coefficient of variation.

판단한다.

본 논문에서는 시간에 따라 모수가 변하는 모형을 제안했지만, 일반적인 공변량에 대해서도 자연스럽게 적용할 수 있다. 빅데이터를 이용하여 다수의 공변량으로 정확한 분포의 추정을 통해 조건부 평균이 어

떻게 변화하는지 파악한다면 산업의 예산 관리, 직원 고용 등에 수 많은 응용분야에 적용할 수 있을 것이라 기대한다.

또한 후속 연구를 위해 다음을 제안한다. Shen과 Brown (2006)에서 사용한 방법과 같이 가능도에 기반하지 않고 커널 가중합을 최소화 시키는 추정값을 사용한다면 계산에 있어서의 불안전성이나 속도를 개선할 수 있을 것이라 생각한다. 띠넓이 선택에서 사용한 교차 타당성 검증의 경우 시간에 의존하는 자료의 경우 이를 고려하지 않고 자료를 분할할 경우 성능이 매우 좋지 않음이 잘 알려져 있다. 본 논문에서 사용한 실증 자료의 경우 자료수가 46,762개로 충분히 크고 국소 다항 방법을 통해 시간에 대한 의존성이 줄어든 모형이여서 기존의 교차 타당성을 통한 띠넓이 선택이 큰 영향을 미치지 못하였다. 하지만, 블럭으로 자료를 뽑는 과정을 통한 교차 타당성 방법의 적용과 같은 방법을 적용한다면 보다 안정적인 띠넓이 선택이 가능할 것이라 기대한다.

References

Beirlant, J. and Goegebeur, Y. (2004). Local polynomial maximum likelihood estimation for Pareto-type distributions, *Journal of Multivariate Analysis*, **89**, 97–118.

Brown, L. D. and Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method, *The Annals of Statistics*, **35**, 2219–2232.

Cooray, K. and Ananda, M. M. A. (2005). Modeling actuarial data with a composite lognormal-Pareto model, *Scandinavian Actuarial Journal*, **2005**, 321–334.

Kim, B., Noh, J., and Baek, C. (2016). Threshold estimation for the composite lognormal-GPD models, *The Korean Journal of Applied Statistics*, **29**, 807–822.

Nadarajah, S. and Baker, S. S. A. (2014). New composite models for the Danish fire insurance data, *Scandinavian Actuarial Journal*, **2014**, 180–187.

Scollnik, D. P. M. (2007). On composite lognormal-Pareto models, *Scandinavian Actuarial Journal*, **2007**, 20–33.

Shen, H. and Brown, L. D. (2006). Non-parametric modelling of time-varying customer service times at a bank call centre, *Applied Stochastic Models in Business and Industry*, **22**, 297–311.

시간에 따라 변화하는 로그-정규분포와 파레토 합성 분포의 모형 추정

박소진[a] · 백창룡[a1]

[a]성균관대학교 통계학과

(2017년 11월 2일 접수, 2017년 12월 1일 수정, 2017년 12월 15일 채택)

요 약

임계값을 기준으로 그 보다 작은 값은 로그정규분포(lognormal distribution; LN)를, 큰 값은 일반화파레토분포(generalized Pareto distribution; GPD)를 따르는 합성 분포를 LN-GPD 합성분포라 한다. Scollnik (2007)은 LN-GPD 합성분포가 로그정규분포와 GPD를 합성 시킴으로써 자료의 손실 없이 꼬리가 두꺼운 분포에서 좋은 적합력을 가진다고 밝혔다. 본 논문에서는 시간에 따라 변하는 LN-GPD 평균모형을 다루었으며 방법론으로는 국소다항최대우도법을 기반으로 추정하는 방법에 대해서 연구하였다. 시간에 따라 변하는 분포를 추정함으로써 자료에 대한 훨씬 자세한 이해가 가능하며 이는 곧 상담원 배치나 자원배분과 같은 운영관리에 큰 도움을 줄 수 있다. 본 연구는 GPD 분포만을 고려한 Beirlant와 Goegebeur (2004)를 확장하여 절삭한 로그정규분포를 추가하여 자료의 손실 없이 자료의 특징을 살펴볼 수 있다는데도 의의가 있다. 모의실험을 통해 제안한 방법론의 적절함을 살펴 보았고 실증 자료 분석으로 이스라엘 은행의 콜센터 서비스 시간에 대해 분석하여 상담원 배치와 관련된 흥미로운 결과를 찾을 수 있었다.

주요용어: LN-GPD 합성분포, 국소다항최대우도법, 콜센터 분석

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: crbaek@skku.edu