평활 스플라인 회귀모형을 이용한 극단값의 공간적 분석

한이정

University of Seoul

December 19, 2019

연구 배경

- 집중호우, 돌발홍수, 해일 등의 극단 사건들 (extreme events) 은 발생 빈도가 낮지만 우리에게 미치는 영향이 큼
- 극단 사건의 분석 및 예측은 인명피해 예방과 재산피해 최소화를 위해 중요한 과제임
- 통계학 분야에서는 극단값을 모형화 하기 위해 일반화 극단값 분포 (Generalized Extreme Value; GEV)에 대한 연구가 이루어져 왔음 [Coles et al., 2001]

- 강수량과 같은 기후자료들은 지리적 특성들과 밀접한 관련이 있음
- 공간적인 종속성을 반영하면서 관측되지 않는 다른 위치에서의 분포 추정이 가능한 공간모형을 고려하는 것이 필요함

연구 목적 및 방법

공간 회귀모형을 이용하여 2차원 공간상의 극단값 분포 패턴을 분석하고자 함

- 기존에 알려진 2차원 평활스플라인 회귀모형을 극단값 분포 모형에 적용하였음
- 함수의 곡률에 패널티를 부여하고 정규화 모수 (평활 모수)를 통해 모형의 복잡도를 조절함
- 이를 통해 극단값의 공간적 분포를 유연하게 추정할 수 있음

공간 GEV 모형

일반화 극단값 분포

일반화 극단값 분포(generalized extreme value distribution, GEV)는 극단값 자료를 분석하는 데 사용하는 분포

$$G(y;\mu,\sigma,\kappa)=\begin{cases}\exp\left(-\left[1+\kappa(\frac{y-\mu}{\sigma})\right]_{+}^{-1/\kappa}\right) & \text{if } \kappa\neq 0\\ \exp\left(-\exp(-\frac{(y-\mu)}{\sigma})\right) & \text{if } \kappa=0\end{cases}$$

여기에서 G는 $\{y:1+\kappa(y-\mu)/\sigma>0\}$에서 정의되며, $\mu\in\mathbb{R}$는 위치 모수 (location parameter), $\sigma>0$는 규모 모수(scale parameter), $\kappa\in\mathbb{R}$는 형상 모수 (shape parameter)

공간 모형

- 먼저, $Y_{s,i}$를 s번째 지역의 i번째 단위 시간에서 아래와 같은 분포를 따르는 확률변수라고 가정

$$Y_{s,i} \sim \mathrm{GEV}(\mu_s, \sigma_s, \kappa_s), \quad s = 1, ..., S, \ i = 1, ..., T$$

- 여기에서 $\mu_s, \sigma_s, \kappa_s$는 s번째 지역에서의 GEV 분포의 모수

- 극단값의 공간적 분포를 설명하는 수학적 표면 (surface)을 추정하기 위해 Thin plates spline (TPS) [Duchon, 1977] 회귀모형을 이용함

$$
\begin{aligned}
\mu_{\mathsf{s}} &= \beta_{\mu,0} + \mathsf{h}_{\mu}(\mathsf{x}_{\mathsf{s}}) \\
\log(\sigma_{\mathsf{s}}) &= \beta_{\sigma,0} + \mathsf{h}_{\sigma}(\mathsf{x}_{\mathsf{s}}) \\
\kappa_{\mathsf{s}} &= \beta_{\kappa,0} + \mathsf{h}_{\kappa}(\mathsf{x}_{\mathsf{s}})
\end{aligned}
$$

- 이때, $\beta_{\eta,0}$는 공간에 대해 전역적인 절편항, η는 GEV 모수 (μ,σ,κ) 중 하나를 의미
- $\mathsf{x}_{\mathsf{s}}^{\top} \in \mathbb{R}^{\mathsf{p}}$는 공간에 대한 설명변수로, 지리학적 변수를 사용할 수 있음
- 예) $\mathsf{x}_{\mathsf{s}}^{\top} = (\text{경도}_{\mathsf{s}}, \text{위도}_{\mathsf{s}}) = (\mathsf{x}_{1\mathsf{s}}, \mathsf{x}_{2\mathsf{s}})$

- $\mathsf{h} : \mathbb{R}^2 \mapsto \mathbb{R}^1$ 는 TPS에서 정의되는 기저함수 (basis function)로 구성된 공간 매핑 (spatial mapping) 함수
- $\mathsf{x}_\ell,\ \ell = 1, 2$에 대해 $\mathsf{B}_\ell(\cdot) = \{\mathsf{B}_{\ell 1}(\cdot), \mathsf{B}_{\ell 2}(\cdot), ..., \mathsf{B}_{\ell \mathsf{M}_\ell}(\cdot)\}^\top$
- $\mathsf{B}(\mathsf{x}_{1\mathsf{s}}, \mathsf{x}_{2\mathsf{s}}) = \mathsf{B}_1(\mathsf{x}_{1\mathsf{s}}) \circ \mathsf{B}_2(\mathsf{x}_{2\mathsf{s}}) : \mathbb{R}^2 \mapsto \mathbb{R}^{\mathsf{M}_1\mathsf{M}_2}$
- 공간 매핑 함수 h는 다음과 같이 표현할 수 있음

$$
\begin{aligned}
\mathsf{h}_\mu(\mathsf{x}_\mathsf{s}) &= \mathsf{B}(\mathsf{x}_{1\mathsf{s}}, \mathsf{x}_{2\mathsf{s}})^\top \boldsymbol{\beta}_\mu = \mathsf{z}_\mathsf{s}^\top \boldsymbol{\beta}_\mu \\
\mathsf{h}_\sigma(\mathsf{x}_\mathsf{s}) &= \mathsf{B}(\mathsf{x}_{1\mathsf{s}}, \mathsf{x}_{2\mathsf{s}})^\top \boldsymbol{\beta}_\sigma = \mathsf{z}_\mathsf{s}^\top \boldsymbol{\beta}_\sigma \\
\mathsf{h}_\kappa(\mathsf{x}_\mathsf{s}) &= \mathsf{B}(\mathsf{x}_{1\mathsf{s}}, \mathsf{x}_{2\mathsf{s}})^\top \boldsymbol{\beta}_\kappa = \mathsf{z}_\mathsf{s}^\top \boldsymbol{\beta}_\kappa
\end{aligned}
$$

- 여기에서 $\boldsymbol{\beta}_\eta \in \mathbb{R}^{\mathsf{M}_1\mathsf{M}_2}$

- TPS 회귀모형은 추정된 표면이 관측지역에서의 모수 추정치를 가장 가깝게 통과하는 동시에 추정된 표면의 곡률 (curvature) 을 최소화하는 방법
- 추정된 표면의 곡률은 모형의 복잡도와 같으며, $\mathbb{R}^2$ 에서 정의되는 $h_\eta(\cdot)$ 의 패널티 함수는 다음과 같음

$$J[h_\eta(\cdot)] = \iint_{\mathbb{R}^2} \left[\left(\frac{\partial^2 h_\eta(x)}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 h_\eta(x)}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 h_\eta(x)}{\partial x_2^2} \right)^2 \right] dx_1 dx_2$$

모형 추정

- 모든 지역 $s = 1, ..., S$에 대해 같은 기간 $i = 1, ..., T$의 관측치 $y_{s,i}$가 존재한다고 가정할 때, 음의 로그가능도 함수는

$$L(\boldsymbol{\beta}_0, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\kappa) = -\sum_{s=1}^{S}\sum_{i=1}^{T} \log g(y_{s,i}; \boldsymbol{\beta}_0, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\kappa)$$

- 여기에서 $\boldsymbol{\beta}_0 = (\beta_{\mu,0}, \beta_{\sigma,0}, \beta_{\kappa,0})$

- 따라서 공간 GEV 모형의 모수를 다음과 같은 벌점화된 음의 로그가능도함수를 최소화하여 추정함

$$\mathsf{L}_\lambda(\boldsymbol{\beta}_0, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\kappa) = \mathsf{L}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\kappa) + \lambda_\mu \mathsf{J}[\mathsf{h}_\mu(\cdot)] + \lambda_\sigma \mathsf{J}[\mathsf{h}_\sigma(\cdot)] + \lambda_\kappa \mathsf{J}[\mathsf{h}_\kappa(\cdot)] \quad (1)$$

- $\lambda_\eta \geq 0$는 정규화 모수 (regularization parameter) 또는 평활 모수 (smoothing parameter)
- 식 (1)의 벌점화 MLE는 뉴튼-랩슨 방법을 통해 추정할 수 있음

모형 선택

- 모형 선택 기준으로 AIC (Akaike Information Criterion)를 사용함
- $\mathsf{df}(\hat{\eta}) = \mathsf{Tr}(\mathsf{H}_{\lambda_\eta})$, where $\mathsf{H}_{\lambda_\eta} = \mathsf{Z}(\mathsf{Z}^\top \mathsf{Z} + \lambda_\eta \boldsymbol{\Omega}_{\mathsf{Z}})^{-1}\mathsf{Z}^\top$
- $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_\mu, \hat{\boldsymbol{\beta}}_\sigma, \hat{\boldsymbol{\beta}}_\kappa)^\top$가 고정된 $(\lambda_\mu, \lambda_\sigma, \lambda_\kappa)$에 대한 (1)의 해일 때,

$$\mathsf{AIC} = 2\mathsf{L}(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_\mu, \hat{\boldsymbol{\beta}}_\sigma, \hat{\boldsymbol{\beta}}_\kappa) + 2(\mathsf{Tr}(\mathsf{H}_{\lambda_\mu}) + \mathsf{Tr}(\mathsf{H}_{\lambda_\sigma}) + \mathsf{Tr}(\mathsf{H}_{\lambda_\kappa})) \quad (2)$$

모의 실험

목적

- 정규화 모수에 따른 추정 모형의 예측 성능을 평가
- 최적의 정규화 모수로 추정된 모형의 예측 성능을 기반으로 AIC의 결과를 평가
  - 학습데이터셋을 통해 주어진 지역에서의 예측성능을 평가
  - 테스트데이터셋을 통해 새로운 지역에 대한 예측성능을 평가

평가 기준

- 두 확률 밀도 함수 f와 g가 주어졌을 때, 헬링거 거리의 제곱은 $H^2(f, g) = \frac{1}{2}\int(\sqrt{f(x)} - \sqrt{g(x)})^2 dx$와 같이 정의됨
- 헬링거 거리를 기반으로 한 예측 성능 척도를 다음과 같이 정의함

$$HD = \sum_{s=1}^{S} H^2(g(\cdot;\theta_s), g(\cdot;\hat{\theta}_s))$$

- 여기에서 g는 GEV 분포의 확률함수, $\theta_s = (\mu_s, \sigma_s, \kappa_s)$는 s번째 지역의 참모수, $\hat{\theta}_s = (\hat{\mu}_s, \hat{\sigma}_s, \hat{\kappa}_s)$는 s지역의 추정된 모수

실험 설계

- GEV 모수들에 대해 세 가지 공간상 패턴을 가정
  - Plane

    $f_1(x_1, x_2) = -3x_1 + 3x_2$
  - Unimodal

    $f_2(x_1, x_2) = \varphi(x; \mu, \Sigma),\ \mu = (0, 0),\ \Sigma = \mathrm{diag}(30, 30)$
  - Bimodal

    $f_3(x_1, x_2) = \pi_1 \varphi(x; \mu_1, \Sigma_1) + \pi_2 \varphi(x; \mu_2, \Sigma_2)$

$$\mu_1 = (5, 0),\ \ \mu_2 = (-5, 0)$$

$$\Sigma_1 = \mathrm{diag}(10, 10),\ \ \Sigma_2 = \mathrm{diag}(20, 20)$$

$$\pi_1 = 0.4,\ \ \pi_2 = 0.6$$

(여기에서, $\varphi(x; \mu, \Sigma)$는 $(d \times 1)$ 평균 벡터 μ와 $(d \times d)$ 공분산 행렬 Σ을 갖는 다변량 정규분포 $x \sim \mathcal{N}(\mu, \Sigma)$ 의 확률 함수)

	Plane	Unimodal	Bimodal
Location (range)	$100 + f_1(x_1, x_2)$ (40, 160)	$90 + 4000\ f_2(x_1, x_2)$ (90.00, 153.66)	$90 + 3000\ f_3(x_{1s}, x_{2s})$ (90.04, 110.38)
Scale (range)	$40 + 0.5\ f_1(x_1, x_2)$ (10, 70)	$30 + 4000\ f_2(x_1, x_2)$ (30.00, 93.66)	$30 + 3000\ f_3(x_{1s}, x_{2s})$ (30.04, 50.38)
Shape (range)	$0.1 + 0.005\ f_1(x_1, x_2)$ (-0.2, 0.4)	$50\ f_2(x_1, x_2)$ (0.00003, 0.79)	$50\ f_3(x_{1s}, x_{2s})$ (0.0007 , 0.34)

Table: 세 가지 패턴에 대한 GEV 모수 생성 모형

실험 절차

1. 모의분포의 27개 $(3 \times 3 \times 3)$ 시나리오에 따라 자료를 생성
   예) 시나리오 1 (Plane, Plane, Plane)

$$
\begin{aligned}
\mathsf{y}_{\mathsf{s,i}} &\sim \mathsf{GEV}(\mu_{\mathsf{s}}, \sigma_{\mathsf{s}}, \kappa_{\mathsf{s}}), \\
\mu_{\mathsf{s}} &= 100 + \mathsf{f}_1(\mathsf{x}_{1\mathsf{s}}, \mathsf{x}_{2\mathsf{s}}), \\
\sigma_{\mathsf{s}} &= 40 + 0.5\ \mathsf{f}_1(\mathsf{x}_{1\mathsf{s}}, \mathsf{x}_{2\mathsf{s}}), \\
\kappa_{\mathsf{s}} &= 0.1 + 0.005\ \mathsf{f}_1(\mathsf{x}_{1\mathsf{s}}, \mathsf{x}_{2\mathsf{s}}).
\end{aligned}
$$

2. 식(1)의 목적함수 $\mathsf{L}_\lambda(\boldsymbol{\beta}_0, \boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \boldsymbol{\beta}_\kappa)$를 최소화하는 모수를 뉴튼-랩슨 알고리즘을 이용하여 추정
3. 식(2)의 AIC가 가장 작은 값을 가지는 정규화 모수의 집합 $(\hat{\lambda}_\mu, \hat{\lambda}_\sigma, \hat{\lambda}_\kappa)$을 찾음
4. 1단계부터 3단계를 100번 반복

실험 설정

- $\mathsf{S} = 30$, $\mathsf{T} = 100$
- 좌표 $(\mathsf{x}_1, \mathsf{x}_2)$는 $\mathbb{R}^2 = [-10, 10]^2$의 공간 영역에서 랜덤하게 추출하였으며, 모든 실험에서 동일한 좌표값을 적용
- 테스트 지역은 $\mathsf{S} = 10$에 대해 새로운 좌표 $(\mathsf{x}_1^*, \mathsf{x}_2^*) \in \mathbb{R}^2 = [-10, 10]^2$를 추출하여 생성하여 예측성능을 평가

실험결과: 시나리오 27 (Bimodal, Bimodal, Bimodal)

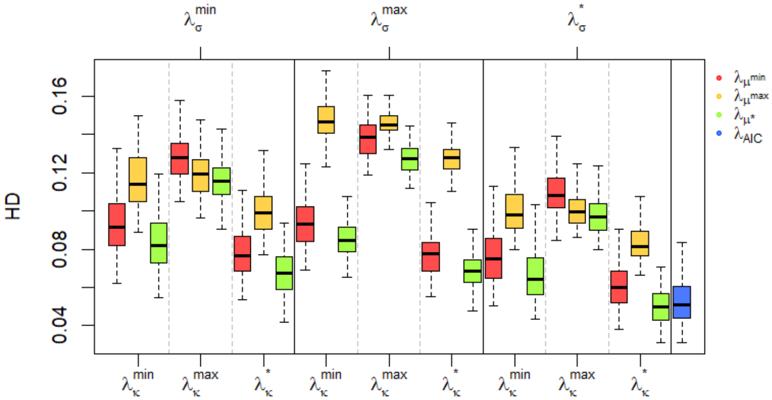

Figure: 시나리오 27의 HD 상자그림 (train sites)

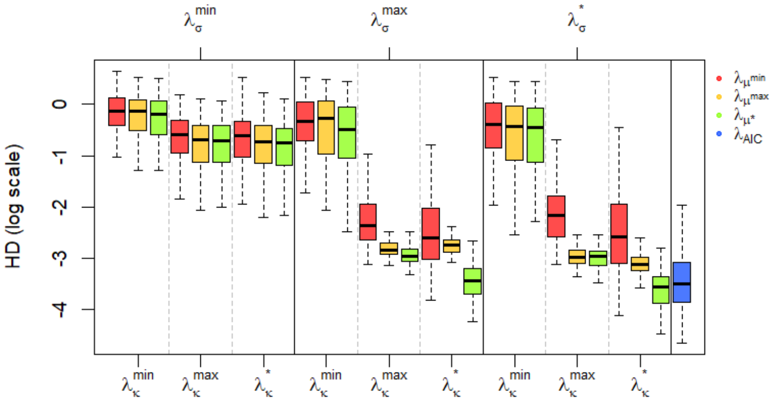

Figure: 시나리오 27의 HD 상자그림 (test sites)

실제 자료 분석

국내 강수 자료 분석

- 우리나라에서 발생하는 자연재해의 대부분이 강수에 의해 발생하는 홍수재해 [행정안전부 기상재해현황, 2018]
- 여름철 빈번하게 발생하는 수해재난을 방지하고 수자원을 효율적으로 이용 · 관리하기 위하여 연최대강수량의 공간분석은 홍수유출량 산정에 매우 중요한 문제임
- 기상청 (KMA)에서 제공하는 강수량 일별 자료 (단위: mm)를 이용하여 요약한 연최대강수량을 사용
- 1973년 1월 1일부터 2018년 12월 31일까지 46년동안 강수량 자료가 모두 존재하고 도서지역이 아닌 56개의 지점에 대한 자료만을 고려

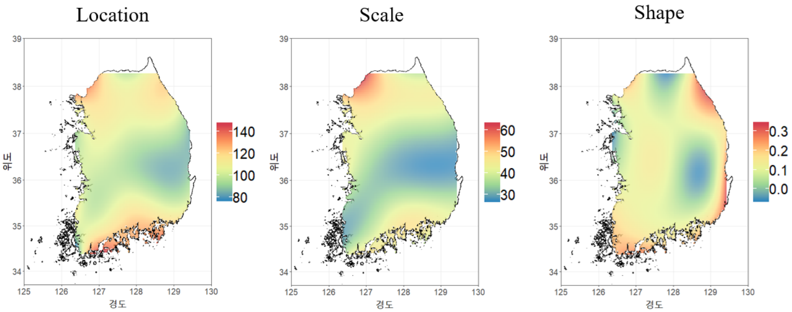

Figure: 연최대강수량 GEV 모수의 공간적 분포

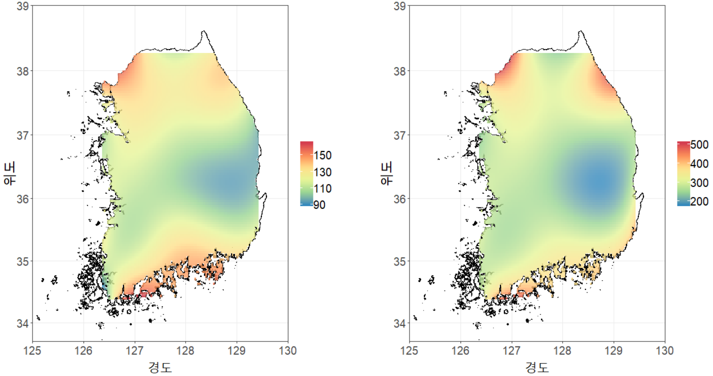

Figure: 연최대강수량의 재현수준 (2년, 50년)

- 2차원 공간상의 극단값 분포의 패턴을 탐지하기 위해 기존에 알려진 다차원 평활 스플라인 방법을 극단값 분포 모형에 적용
- 모의실험에서는 정규화 모수에 따른 추정 모형의 예측 성능을 평가
- 소개한 모형을 우리나라의 일강수량 자료에 적용하였으며, 우리나라 연최대강수량의 공간분포를 추정
- 분석 결과를 통해 우리나라의 수자원 설계시 지역에 따른 효율적인 관리에 도움이 될 것으로 기대함

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001).
An introduction to statistical modeling of extreme values, volume 208.
Springer.

Duchon, J. (1977).
Splines minimizing rotation-invariant semi-norms in sobolev spaces.
In *Constructive theory of functions of several variables*, pages 85–100.
Springer.