PCA for extremes

Sam Morris[1], Brian J Reich[1], Emeric Thibauld[2], and Dan Cooley[2]

March 17, 2016

Abstract

words...

Key words: Max-stable process.

[1]North Carolina State University
[2]Colorado State University

1 Introduction

2 Model

Let Y_{it} be the observation at location $\mathbf{s}_i$ for $i \in \{1, ..., n_s\}$ and time $t \in \{1, ..., n_t\}$. To focus attention on the extreme values, we consider data above a threshold T. The marginal distribution of Y_{it} is then determined by the probability of exceeding the threshold and the distribution of the excursions. Denote the exceedance probability as Prob$[Y_{it} > T] = p_{it}$. Extreme value theory says that for sufficiently large T the excursion distribution can be approximated using a generalized Pareto distribution (GPD). Therefore we model $Y_{it}|Y_{it} > T \sim$ GDP(σ_{it}, ξ), where the GDP scale and shape parameters are denoted $\sigma_{it} > 0$ and ξ, respectively.

spectral, max-linear...finally we settle on... Spatial extremal dependence is captured using a max-stable copula (define). Let Z_{it} be a max-stable process with Fréchet marginal distributions (define GEV etc...). Our objective is to identify a low-rank model for spatial dependence in Z_{it}. Decompose Z_{it} as $Z_{it} = \theta_{it}\varepsilon_{it}$ where θ_{it} is a spatial process and $\varepsilon_{it} \stackrel{iid}{\sim}$ GEV$(1, \alpha, \alpha)$ is a nugget. The spatial component is written as a combination of L basis functions B_{il}

$$\theta_{it} = \left(\sum_{l=1}^{L} B_{il}^{1/\alpha} A_{lt} \right)^{\alpha}. \tag{1}$$

If $B_{il} > 0$, $\sum_{l=1}^{L} B_{il} = 1$, and the A_{lt} have positive stable (PS) distribution $A_{lt} \sim$ PS(α) (define), then Z_{it} is max-stable and has Fréchet marginal distributions.

The Z_{it} are conditionally independent given the spatial random effects, with conditional distribution $Z_{it}|\theta_{it} \sim$. As a result, the likelihood is $Y_{it}|\theta_{it} \stackrel{indep}{\sim} g(y; \theta_{it}, p_{it}, \sigma_{it}, \xi)$ where

$$g(y; \theta, p, \sigma, \xi) = \tag{2}$$

Therefore, the likelihood factors across observations which is computationally convenient. Marginalizing over the random effect θ_{it} induces extremal spatial dependence in the Z_{it}, and thus the Y_{it}. Spatial dependence can be summarized by the extremal coefficient (EC) $\vartheta_{ij} \in [1, 2]$, where

$$\text{Prob}(Z_{it} < c, Z_{jt} < c) = \text{Prob}(Z_{it} < c)^{\vartheta_{ij}}. \tag{3}$$

For the PS random effects model the EC has the form

$$\vartheta_{ij} = \sum_{l=1}^{L} \left(B_{il}^{1/\alpha} + B_{jl}^{1/\alpha} \right)^{\alpha}. \tag{4}$$

In particular, $\vartheta_{ii} = 2^{\alpha}$ for all i. Since $\sum_{l=1}^{L} B_{il} = 1$ for all i, we have $\sum_{l=1}^{L}(\sum_{i=1}^{n_s} B_{il}/n_s) = 1$. Therefore, the relative contribution of term l can be measured by

$$v_l = \sum_{i=1}^{n_s} B_{il}/n_s, \tag{5}$$

with $\sum_{l=1}^{L} v_l = 1$. The order of the terms is arbitrary, and so we assume without loss of generality that $v_1 \geq ... \geq v_L$.

3 Estimating the extremal coefficient function

In this section we develop an algorithm to estimate the spatial dependence parameter α and the $n_s \times L$ matrix $\mathbf{B} = \{B_{il}\}$. Given these parameters, we plug them into our model and proceed with Bayesian analysis as described in Section 4. Our algorithm has the following steps:

(1) Obtain an initial estimate of the extremal coefficient for each pair of locations, $\hat{\vartheta}_{ij}$.

(2) Spatially smooth these initial estimates $\hat{\vartheta}_{ij}$ using kernel smoothing to obtain $\tilde{\vartheta}_{ij}$.

(3) Estimate the spatial dependence parameters by minimizing the difference between model-based coefficients, ϑ_{ij}, and smoothed coefficients, $\tilde{\vartheta}_{ij}$.

To estimate the spatial dependence we first remove variation in the marginal distribution through a sample rank transformation so that the marginal distribution at each location is approximately uniform. Let $U_{it} = r_t(\mathbf{Y}_i)/(n_t + 1)$, where $r_t(\mathbf{Y}_i)$ is the sample rank of the tth entry of $\mathbf{Y}_i$. Then for some extreme probability $q \in (0, 1)$, solving (3) suggest the estimate

$$\hat{\vartheta}_{ij}(q) = \frac{\log[Q_{ij}(q)]}{\log(q)}, \tag{6}$$

where $Q_{ij}(q) = \sum_{t=1}^{n_t} I[U_{it} < q, U_{jt} < q]/n_t$ is the sample proportion of the time points at which both sites are less then q. Since all large q give valid estimates, we average over a grid of q with $q_1 < \cdots < q_{n_q}$

$$\hat{\vartheta}_{ij} = \frac{1}{n_q} \sum_{j=1}^{n_q} \hat{\vartheta}_{ij}(q_j). \tag{7}$$

If the data are max-stable (e.g. block maxima) then we can set $q_1 = 0$ and average over all observations.

Assuming the true B_{il} are smooth over space, the initial estimates $\hat{\vartheta}_{ij}$ can be improved by smoothing. Let

$$\tilde{\vartheta}_{ij} = \frac{\sum_{u=1}^{n_s} \sum_{v=1}^{n_s} w_{iu} w_{jv} \hat{\vartheta}_{uv}}{\sum_{u=1}^{n_s} \sum_{v=1}^{n_s} w_{iu} w_{jv}}, \tag{8}$$

where $w_{iu} = \exp(-\phi||\mathbf{s}_i - \mathbf{s}'_u||^2)$ is the Gaussian kernel function with bandwidth ϕ. The elements $\hat{\vartheta}_{ii}$ do not contributed any information as $\hat{\vartheta}_{ii} = 1$ for all i by construction. To eliminate the influence of these estimates we set $w_{ii} = 0$. However, this approach does give imputed values $\tilde{\vartheta}_{ii}$, which provides information about small-scale spatial variability.

The dependence parameters are estimated by comparing estimates $\tilde{\vartheta}_{ij}$ with the model-based values ϑ_{ij}.

For all i, $\vartheta_{ii} = 2^{\alpha}$, and therefore we set α to $\hat{\alpha} = \log_2(\sum_{i=1}^{n_s} \tilde{\vartheta}_{ii}/n_s)$. Given $\alpha = \hat{\alpha}$, it remains to estimate $\mathbf{B}$. The estimate, $\hat{\mathbf{B}}$, is taken as the minimizer of

$$m(\mathbf{B}) = \sum_{i<j} \left(\tilde{\vartheta}_{ij} - \vartheta_{ij}\right)^2 = \sum_{i<j} \left\{\tilde{\vartheta}_{ji} - \sum_{l=1}^{L} [B_{il}^{1/\hat{\alpha}} + B_{jl}^{1/\hat{\alpha}}]^{\hat{\alpha}}\right\}^2 \tag{9}$$

under the restrictions that $B_{il} \geq 0$ for all i and l and $\sum_{l=1}^{L} B_{il} = 1$ for all i. The order of the B_{il} is not defined. Therefore, we sort the terms so that $v_1 > ... > v_L$.

4 Implementation details

The model has three tuning parameters: the quantile threshold q_1, the kernel bandwidth ϕ, and the number of terms L. If the data are already max-stable (e.g. block maxima), then we can use $q_l = 0$; however, in the case of data that are not max-stable, then q should be high enough to give GPD marginals. We set ϕ to be twice the distance between the two closest sites. To pick the optimal number of knots, L, we run the model with multiple options and select the L that gives the best predictive performance as measured by 5-fold cross validation and the quantile and Brier score (Gneiting and Raftery, 2007).

Given the estimates of α and $\mathbf{B}$, the hierarchical model is

$$\begin{aligned} Y_{it}|\theta_{ij} &\overset{indep}{\sim} g(y;\theta_{it}, p_{it}, \sigma_{it}, \xi) \\ \theta_{it} &= \left(\sum_{l=1}^{L} \hat{B}_{il}^{1/\hat{\alpha}} A_{lt}\right)^{\hat{\alpha}} \quad \text{where} \quad A_{lt} \overset{iid}{\sim} PS(\hat{\alpha}) \\ \text{logit}(p_{it}) &= \mathbf{X}_{it}^T \boldsymbol{\beta}_1 \quad \text{and} \quad \log(\sigma_{it}) = \mathbf{X}_{it}^T \boldsymbol{\beta}_2 \end{aligned} \tag{10}$$

where g is given in (2) and $\mathbf{X}_{it} = (X_{it1}, ..., X_{itp})^T$ is a vector of spatiotemporal covariates. To complete the Bayesian model, we select independent normal priors with mean zero and variance $\sigma_{\beta_1}^2, \sigma_{\beta_1}^2$ for the components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and normal prior with mean 0 and variance 0.25 for ξ. Finally, we use independent IG$(0.1, 0.1)$ priors on $\sigma^2_{\beta_1}, \sigma^2_{\beta_1}$.

We estimate parameters $\Theta = \{A_{lt}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \xi, \sigma^2_{\beta_1}, \sigma^2_{\beta_2}\}$ using Markov chain Monte Carlo methods. We use a Metropolis-Hastings algorithm to update the model parameters with random walk candidate distributions for all parameters except $\sigma^2_{\beta_1}, \sigma^2_{\beta_2}$. To update $\sigma^2_{\beta_1}$ and $\sigma^2_{\beta_2}$, we use Gibbs sampling.

5 Data analysis

The dataset used for our application is composed of yearly acreage burned due to wildfires for each county in Georgia from 1965 – 2014 (`http://weather.gfc.stat.ga.us/FireData/`). Figure 1 shows the time series of $\log$(acres burned) for 25 randomly selected counties. Based on this plot, and some other exploratory analysis, we see no evidence of non-linear trends and proceed with linear time trends for the GEV location and scale parameters.

We estimate the extremal coefficient function $\hat{\theta}_{ij}$ by setting $q_1 = 0.90$ and using $n_q = 100$. With more data, it would possible to increase q_1, but we set $q_1 = 0.90$ to increase the stability when estimating $\hat{\vartheta}_{ij}$.

Because these data are not max-stable, we select a site-specific threshold T_i to use in the analysis with the following algorithm. Without some adjustment to the data, it is challenging to borrow information across sites to inform the threshold selection. We first compute

$$\tilde{\mathbf{Y}}_i = \frac{\mathbf{Y}_i - \text{med}(\mathbf{Y}_i)}{\text{IQR}(\mathbf{Y}_i)} \tag{11}$$

where med$(\cdot)$ is the median, and IQR$(\cdot)$ is the inter-quartile range. Then we combine all sites together and plot a mean residual plot for $\tilde{\mathbf{Y}}_i, i = 1, \ldots, n_s$. The mean residual plot is given in Figure 3, with a vertical line indicating the quantile we use for the county-specific values $\mathbf{T}$. Based upon the mean residual plot,

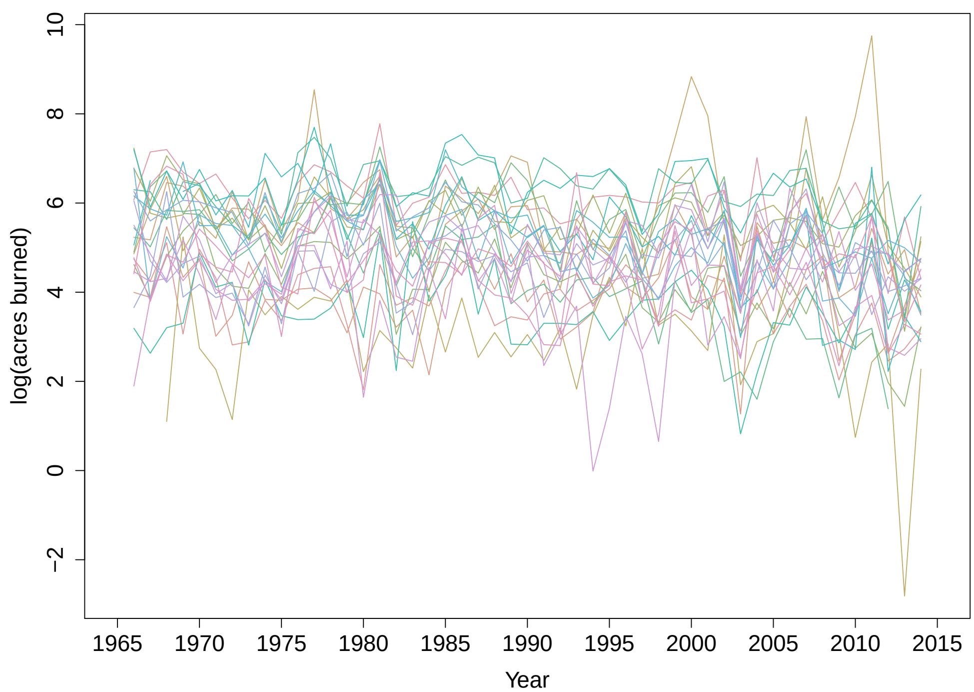

Figure 1: Time series of $\log$(acres burned) for 25 randomly selected counties.

we select $q(0.95)$ for the spatially smoothed threshold. To calculate T_i for each county, we find $\hat{q}(0.95)$ by taking the $95th$ quantile for county i and the five closest counties.

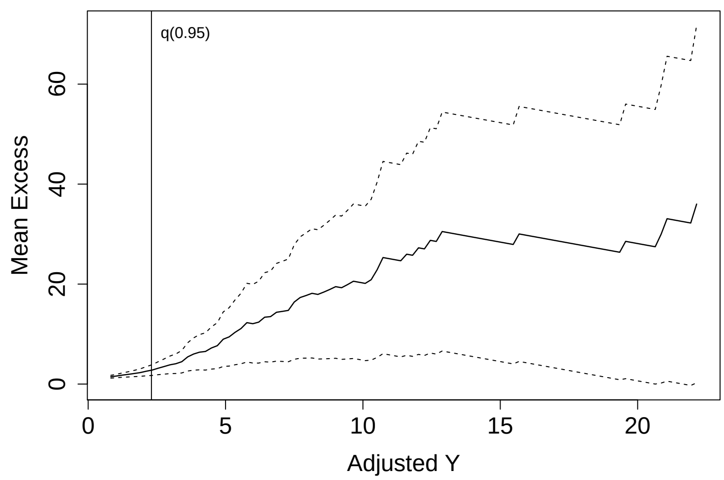

Figure 2: Mean residual plot with line indicating the $T = q(0.95)$ for the analysis.

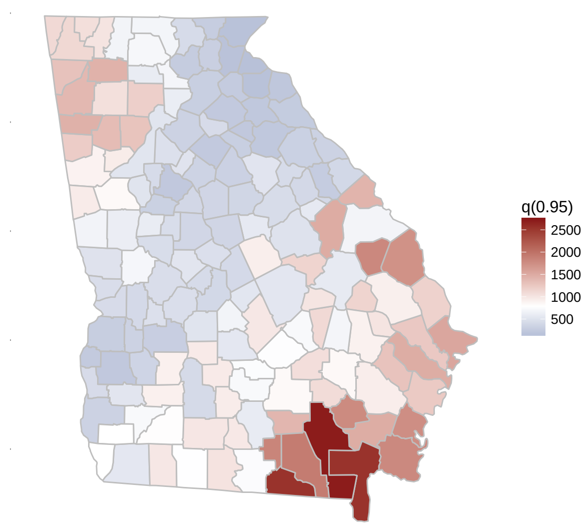

Figure 3: Spatially smoothed threshold values for each county.

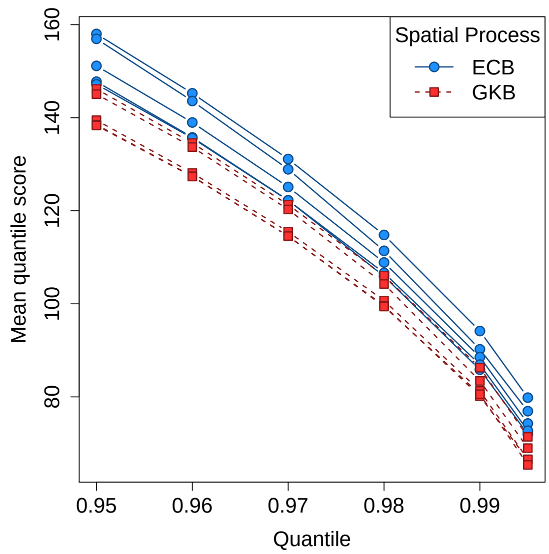

Figure 4: Average quantile score for selected quantiles

5.1 Results

We use 5-fold cross-validation to assess the predictive performance of a model. For each method, we randomly select 80% of the observations across counties and years to be used as a training set to fit the model. The remaining 20% of sites and years are withheld for testing model predictions. To assess the predictions for the test set, we use quantile scores and Brier scores citation. The quantile score is given by give formula. The Brier score is given by give formula. For both of these methods, we use a negative orientation, so a lower score indicates a better fit.

Table 1: Average quantile scores for selected quantiles and Brier scores ($\times 1000$) for selected thresholds

	L = 2		L = 5		L = 10		L = 15		L = 20	
	Basis	Kernel	Basis	Kernel	Basis	Kernel	Basis	Kernel	Basis	Kernel
QS(0.95)	157.23	146.18	156.96	145.07	151.15	139.47	147.76	138.49	147.19	138.35
QS(0.99)	93.47	86.27	90.23	83.43	88.56	81.29	85.84	80.14	86.93	80.57
BS(0.95)	91.61	84.21	91.45	84.04	88.43	82.29	87.57	81.83	87.17	81.70
BS(0.99)	44.63	41.53	44.55	41.70	43.18	40.72	42.81	40.71	42.43	40.64

Based upon the cross-validation results, we reran the full data analysis using $L = 15$ basis functions.

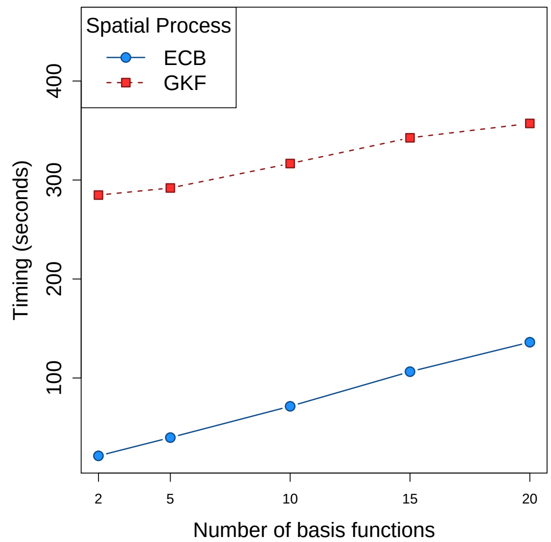

Figure 5: Timing comparison of basis functions to kernel functions for the spatial process (100 iterations)

5.2 Model checking and sensitivity analysis

6 Conclusions

Acknowledgements

References

Gneiting, T. and Raftery, A. E. (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**, 359–378.