Bayesian Multivariate Extreme Value Thresholding for Environmental Hazards

D. Lupton K. Abayomi M. Lacer

School of Industrial and Systems Engineering
Georgia Institute of Technology

Institute for Operations Research and Management Sciences, 2010

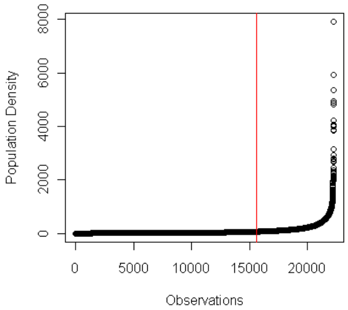
Outline

High Risk Hotspot

Between 1994-1998: Volcano eruption in Rabaul, Cyclone Justin in the Milne Bay (SE from map selection), and El Niño-induced drought

One Variable

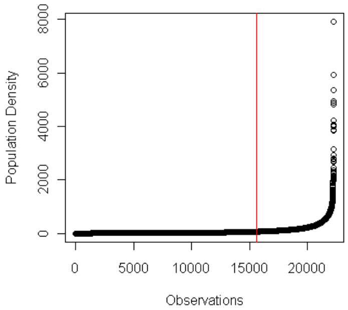
- In the univariate setting thresholding is straightforward...



- ..the separation of data into regular-valued and extreme-valued portions.

One Variable

- In the univariate setting thresholding is straightforward...



- ..the separation of data into regular-valued and extreme-valued portions.

Multivariate Data

Taking multivariate **q**, say, we want to return the set $\mathcal{T}$ such that

$$\mathcal{T} = \{t | F(\mathbf{T} > \mathbf{t}) > c\} \tag{1}$$

Censor the data:

$$\mathcal{T} \supset \mathcal{T}_* = \{\mathbf{t} \mid t_i > c, \forall i\} \tag{2}$$

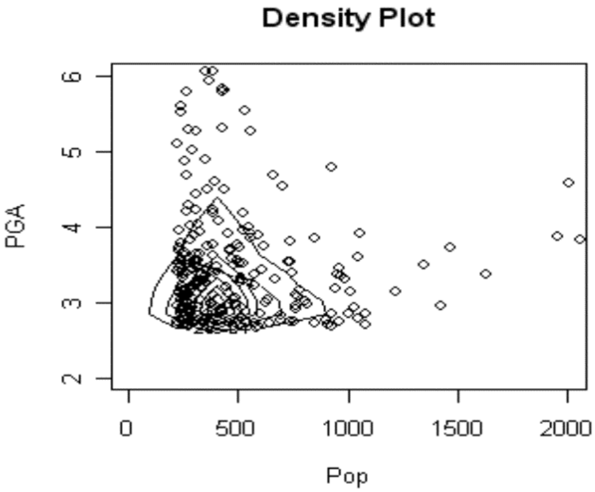
And the output is:F for $i = 1, 2$ is $F(\mathbf{T} \leq \mathbf{t}_*) = F_1 + F_2 - F_1 F_2$ and $F_1 = Pr(\mathbf{T} \leq \mathbf{t}_*)$; $F_2 = F_1 = Pr(\mathbf{T} \leq \mathbf{t} \mid \mathbf{T} > \mathbf{t}_*)$

Multivariate Data

In the Multivariate setting this is to fit some contour that partitions multivariate data into

- Regular valued
- Extreme valued

Pop vs. PGA



Global Natural Disaster Risk Hotspots

Worldwide data has been gridded to $1\frac{1}{2}^{\circ}$ boxes for 8 predictor variables.

- GDP
- Population
- Peak Ground Acceleration (PGA)
- Floods
- Cyclones
- Drought
- Volcanoes
- Landslides

Global Natural Disaster Risk Hotspots

Worldwide data has been gridded to $1\frac{1}{2}^{\circ}$ boxes for 8 predictor variables.

- GDP
- Population
- Peak Ground Acceleration (PGA)
- Floods
- Cyclones
- Drought
- Volcanoes
- Landslides

2003 Global Natural Disaster Risk Hotspots Data

Incidence Maps, gridded to 1.5° lat-lon, 8 variables

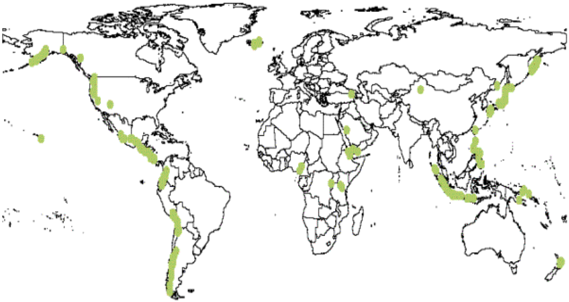
- Floods
- Volcano
- Drought
- Earthquake
- GNP: 1990 Gross National Product in US dollars
- Population: Gridded population count (estimate) 1995

Floods

.9 ptile of Flood counts

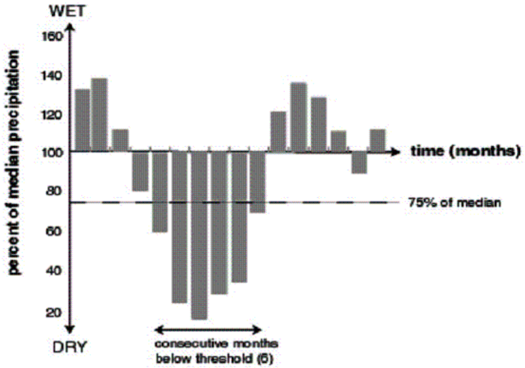
Volcanos

'.9' ptile of Volcano incidence



Droughts

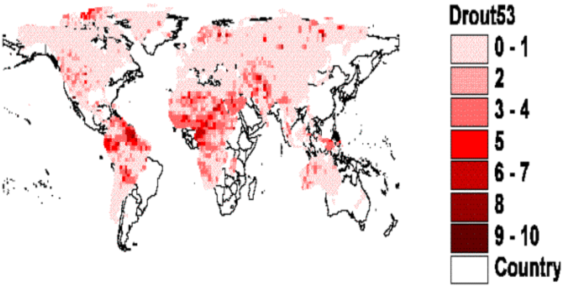
Droughts: Classifying a drought.



Example of a drought event defined by monthly precipitation being below a threshold of 75% of the long-term median value for at least 3 consecutive months. In this case, the duration of the event was 6 months.

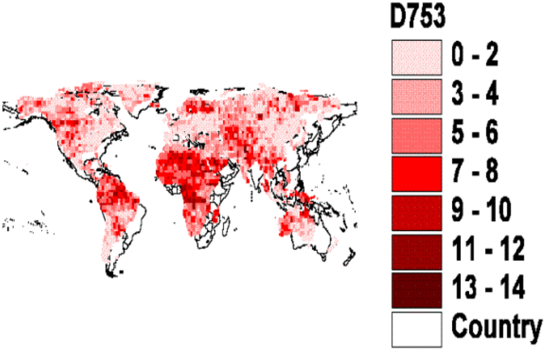
Droughts

50 pct Weighted Anomaly Standardized Precipitation (WASP)



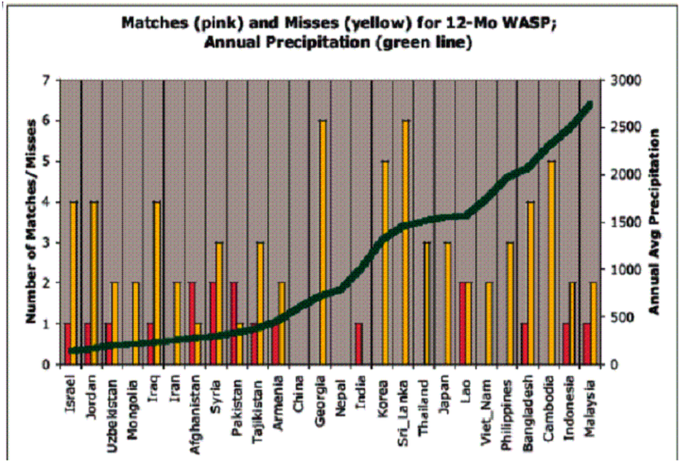
Droughts

75 pct Weighted Anomaly Standardized Precipitation (WASP)



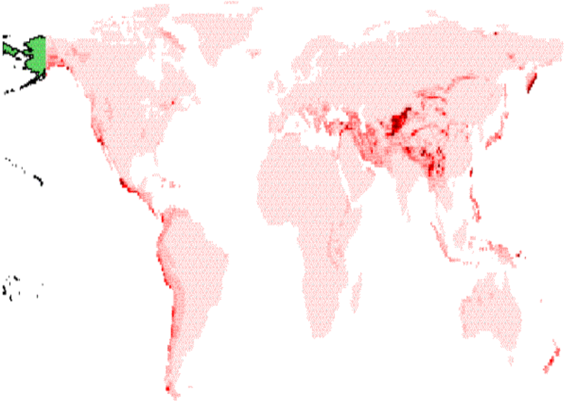
Droughts

Drought declaration vs. Drought classification



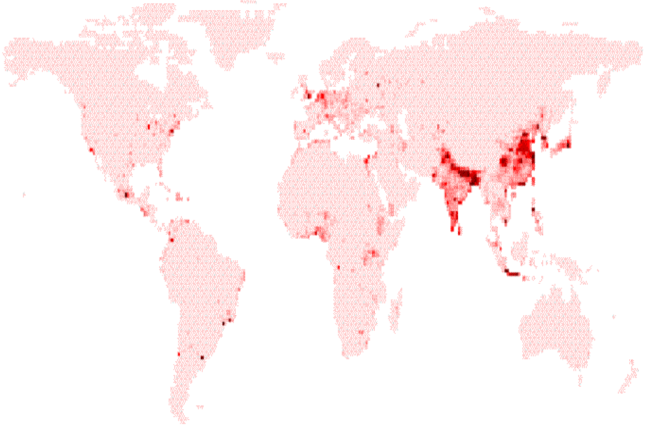
Quakes

Peak Ground Acceleration



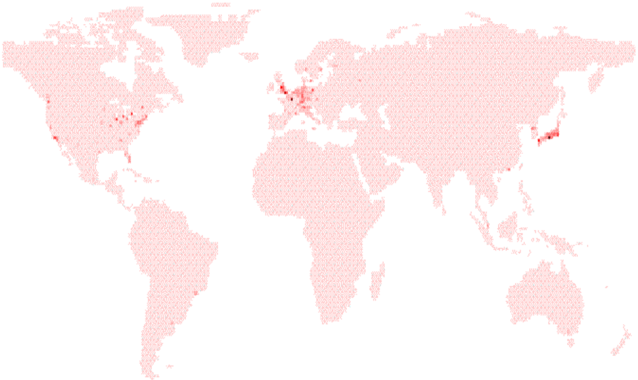
Population

Population Density



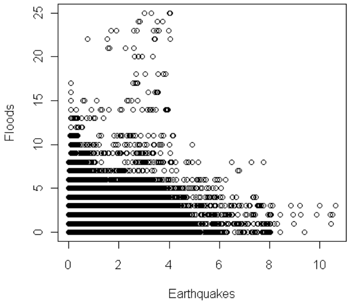
Income

GNP



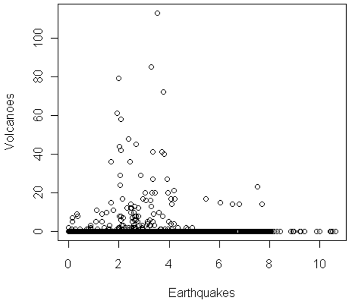
Select Bivariate Plots

- PGA vs. Floods
- PGA vs. Volcanoes



Select Bivariate Plots

- PGA vs. Volcanoes



Multivariate Extreme Value Thresholding

We proceed as follows:

- Select a thresholding level
- Fit an extreme-valued parametric model to the data's tail
- Measure distance between the parametric model and an empirical distribution function

Multivariate Extreme Value Thresholding

We proceed as follows:

- Select a thresholding level
- Fit an extreme-valued parametric model to the data's tail
- Measure distance between the parametric model and an empirical distribution function

Multivariate Extreme Value Thresholding

We proceed as follows:

- Select a thresholding level
- Fit an extreme-valued parametric model to the data's tail
- Measure distance between the parametric model and an empirical distribution function

Multivariate Extreme Value Thresholding

We proceed as follows:

- Select a thresholding level
- Fit an extreme-valued parametric model to the data’s tail
- Measure distance between the parametric model and an empirical distribution function

Parametric Model

Asymmetric Logistic Distribution (Tawn 1990):

$$F_\Theta(x_1, \ldots, x_d) = \exp\left[- \sum_{b \in B} \left[\sum_{j \in b} \left(\frac{\theta_{j,b}}{y_j} \right)^{1/\alpha_b} \right]^{\alpha_b} \right]$$

- $j \in \{1, \ldots, d\}$, and y_j is the transformed data
- $B = \text{PowerSet}\{1, \ldots, d\} \setminus \emptyset$. Hence, $|B| = 2^d - 1$
- Say, $b = \{2, 4, 7\}$, then the inner sum is over $j = 2, 4, 7$
- $\alpha_b \in (0, 1] \; \forall \; b \in B \setminus B_1$ are dependence parameters
- $\theta_{j,b}$ are asymmetry parameters, with the constraint: $\sum_{b \in B_{(j)}} \theta_{j,b} = 1$ for $j = 1, \ldots, d$ to force univariate margins to be of the correct form. Here, $B_{(j)} = \{b \in B \; : \; j \in b\}$.
- $|B_{(j)}| = 2^{d-1}$.

Parametric Model

Asymmetric Logistic Distribution (Tawn 1990):

$$F_{\Theta}(x_1, \ldots, x_d) = \exp\left[-\sum_{b \in B} \left[\sum_{j \in b} \left(\frac{\theta_{j,b}}{y_j} \right)^{1/\alpha_b} \right]^{\alpha_b} \right]$$

- $j \in \{1, \ldots, d\}$, and y_j is the transformed data
- $B = \text{PowerSet}\{1, \ldots, d\} \setminus \emptyset$. Hence, $|B| = 2^d - 1$
- Say, $b = \{2, 4, 7\}$, then the inner sum is over $j = 2, 4, 7$
- $\alpha_b \in (0, 1] \; \forall \; b \in B \setminus B_1$ are dependence parameters
- $\theta_{j,b}$ are asymmetry parameters, with the constraint: $\sum_{b \in B_{(j)}} \theta_{j,b} = 1$ for $j = 1, \ldots, d$ to force univariate margins to be of the correct form. Here, $B_{(j)} = \{b \in B \; : \; j \in b\}$.
- $|B_{(j)}| = 2^{d-1}$.

Parametric Model

Asymmetric Logistic Distribution (Tawn 1990):

$$F_{\Theta}(x_1, \ldots, x_d) = \exp\left[-\sum_{b \in B} \left[\sum_{j \in b} \left(\frac{\theta_{j,b}}{y_j} \right)^{1/\alpha_b} \right]^{\alpha_b} \right]$$

- $j \in \{1, \ldots, d\}$, and y_j is the transformed data
- $B = \text{PowerSet}\{1, \ldots, d\} \setminus \emptyset$. Hence, $|B| = 2^d - 1$
- Say, $b = \{2, 4, 7\}$, then the inner sum is over $j = 2, 4, 7$
- $\alpha_b \in (0, 1] \; \forall \; b \in B \setminus B_1$ are dependence parameters
- $\theta_{j,b}$ are asymmetry parameters, with the constraint: $\sum_{b \in B_{(j)}} \theta_{j,b} = 1$ for $j = 1, \ldots, d$ to force univariate margins to be of the correct form. Here, $B_{(j)} = \{b \in B \; : \; j \in b\}$.
- $|B_{(j)}| = 2^{d-1}$.

Parametric Model

Asymmetric Logistic Distribution (Tawn 1990):

$$F_\Theta(x_1, \ldots, x_d) = \exp\left[-\sum_{b\in B}\left[\sum_{j\in b}\left(\frac{\theta_{j,b}}{y_j}\right)^{1/\alpha_b}\right]^{\alpha_b}\right]$$

- $j \in \{1, \ldots, d\}$, and y_j is the transformed data
- $B = \text{PowerSet}\{1, \ldots, d\} \setminus \emptyset$. Hence, $|B| = 2^d - 1$
- Say, $b = \{2, 4, 7\}$, then the inner sum is over $j = 2, 4, 7$
- $\alpha_b \in (0, 1]\ \forall\ b \in B \setminus B_1$ are dependence parameters
- $\theta_{j,b}$ are asymmetry parameters, with the constraint: $\sum_{b\in B_{(j)}} \theta_{j,b} = 1$ for $j = 1, \ldots, d$ to force univariate margins to be of the correct form. Here, $B_{(j)} = \{b \in B\ :\ j \in b\}$.
- $|B_{(j)}| = 2^{d-1}$.

Parametric Model

Asymmetric Logistic Distribution (Tawn 1990):

$$F_{\Theta}(x_1, \ldots, x_d) = \exp\left[-\sum_{b \in B} \left[\sum_{j \in b} \left(\frac{\theta_{j,b}}{y_j} \right)^{1/\alpha_b} \right]^{\alpha_b} \right]$$

- $j \in \{1, \ldots, d\}$, and y_j is the transformed data
- $B = \text{PowerSet}\{1, \ldots, d\} \setminus \emptyset$. Hence, $|B| = 2^d - 1$
- Say, $b = \{2, 4, 7\}$, then the inner sum is over $j = 2, 4, 7$
- $\alpha_b \in (0, 1] \; \forall \; b \in B \setminus B_1$ are dependence parameters
- $\theta_{j,b}$ are asymmetry parameters, with the constraint: $\sum_{b \in B_{(j)}} \theta_{j,b} = 1$ for $j = 1, \ldots, d$ to force univariate margins to be of the correct form. Here, $B_{(j)} = \{b \in B \;:\; j \in b\}$.
- $|B_{(j)}| = 2^{d-1}$.

Parametric Model

Asymmetric Logistic Distribution (Tawn 1990):

$$F_\Theta(x_1, \ldots, x_d) = \exp\left[-\sum_{b \in B} \left[\sum_{j \in b} \left(\frac{\theta_{j,b}}{y_j} \right)^{1/\alpha_b} \right]^{\alpha_b} \right]$$

- $j \in \{1, \ldots, d\}$, and y_j is the transformed data
- $B = \text{PowerSet}\{1, \ldots, d\} \setminus \emptyset$. Hence, $|B| = 2^d - 1$
- Say, $b = \{2, 4, 7\}$, then the inner sum is over $j = 2, 4, 7$
- $\alpha_b \in (0, 1] \ \forall \ b \in B \setminus B_1$ are dependence parameters
- $\theta_{j,b}$ are asymmetry parameters, with the constraint: $\sum_{b \in B_{(j)}} \theta_{j,b} = 1$ for $j = 1, \ldots, d$ to force univariate margins to be of the correct form. Here, $B_{(j)} = \{b \in B \ : \ j \in b\}$.
- $|B_{(j)}| = 2^{d-1}$.

Parametric Model

Asymmetric Logistic Distribution (Tawn 1990):

$$F_{\Theta}(x_1, \ldots, x_d) = \exp\left[-\sum_{b \in B}\left[\sum_{j \in b}\left(\frac{\theta_{j,b}}{y_j}\right)^{1/\alpha_b}\right]^{\alpha_b}\right]$$

- $j \in \{1, \ldots, d\}$, and y_j is the transformed data
- $B = \text{PowerSet}\{1, \ldots, d\} \setminus \emptyset$. Hence, $|B| = 2^d - 1$
- Say, $b = \{2, 4, 7\}$, then the inner sum is over $j = 2, 4, 7$
- $\alpha_b \in (0, 1]\ \forall\ b \in B \setminus B_1$ are dependence parameters
- $\theta_{j,b}$ are asymmetry parameters, with the constraint: $\sum_{b \in B_{(j)}} \theta_{j,b} = 1$ for $j = 1, \ldots, d$ to force univariate margins to be of the correct form. Here, $B_{(j)} = \{b \in B\ :\ j \in b\}$.
- $|B_{(j)}| = 2^{d-1}$.

Conditional Representation

To derive the pdf, we make use of the positive stable (PS) distribution and its Laplace transform (Stephenson 2009):

- $\int_0^\infty h_1(s)\exp(-st)ds = \exp(-t^\alpha)$
- Take $S_b \sim \mathrm{PS}(\alpha_b)\ \forall\ b \in \mathrm{B} \setminus \mathrm{B}_1$, and $\mathbf{S} = \{S_b \mid b \in \mathrm{B} \setminus \mathrm{B}_1\}$.
- Then we have for $j = 1, \ldots, d$:

  $$\Pr(X_j < x_j \mid \mathbf{S} = \mathbf{s}) = \exp\left[-\sum_{b \in B_{(j)}} s_b \left(\frac{\theta_{j,b}}{y_j}\right)^{1/\alpha_b}\right]$$

  while $X_1, \ldots, X_d$ are conditionally independent given $\mathbf{S} = \mathbf{s}$
- Thus, each marginal asymmetric logistic pdf can be given by:

  $$f_j(x_j|s) = \sigma_j^{-1} y_j^{-xi_j} \left[\sum_{b \in B_{(j)}} (z_{j,b}/\alpha_b)\right] \exp\left(-\sum_{b \in B_{(j)}} z_j, b\right)$$

  where $z_{j,b} = s_b(\theta_{j,b}/y_j)^{1/\alpha_b}$

Conditional Representation

To derive the pdf, we make use of the positive stable (PS) distribution and its Laplace transform (Stephenson 2009):

- $\int_0^\infty h_1(s)\exp(-st)ds = \exp(-t^\alpha)$
- Take $S_b \sim \text{PS}(\alpha_b)\ \forall\ b \in \text{B} \setminus \text{B}_1$, and $\mathbf{S} = \{S_b \mid b \in \text{B} \setminus \text{B}_1\}$.
- Then we have for $j = 1, \ldots, d$:

$$\Pr(X_j < x_j \mid \mathbf{S} = \mathbf{s}) = \exp\left[-\sum_{b \in B_{(j)}} s_b \left(\frac{\theta_{j,b}}{y_j}\right)^{1/\alpha_b}\right]$$

while $X_1, \ldots, X_d$ are conditionally independent given $\mathbf{S} = \mathbf{s}$

- Thus, each marginal asymmetric logistic pdf can be given by:

$$f_j(x_j|s) = \sigma_j^{-1} y_j^{-xi_j}\left[\sum_{b \in B_{(j)}} (z_{j,b}/\alpha_b)\right]\exp\left(-\sum_{b \in B_{(j)}} z_j, b\right)$$

where $z_{j,b} = s_b(\theta_{j,b}/y_j)^{1/\alpha_b}$

Conditional Representation

To derive the pdf, we make use of the positive stable (PS) distribution and its Laplace transform (Stephenson 2009):

- $\int_0^\infty h_1(s)\exp(-st)ds = \exp(-t^\alpha)$
- Take $S_b \sim \text{PS}(\alpha_b)\ \forall\ b \in \text{B} \setminus \text{B}_1$, and $\mathbf{S} = \{S_b \mid b \in \text{B} \setminus \text{B}_1\}$.
- Then we have for $j = 1, \ldots, d$:

$$\Pr(X_j < x_j \mid \mathbf{S} = \mathbf{s}) = \exp\left[-\sum_{b \in B_{(j)}} s_b \left(\frac{\theta_{j,b}}{y_j}\right)^{1/\alpha_b}\right]$$

while $X_1, \ldots, X_d$ are conditionally independent given $\mathbf{S} = \mathbf{s}$

- Thus, each marginal asymmetric logistic pdf can be given by:

$$f_j(x_j|s) = \sigma_j^{-1} y_j^{-x_{i_j}} \left[\sum_{b \in B_{(j)}} (z_{j,b}/\alpha_b)\right] \exp\left(-\sum_{b \in B_{(j)}} z_j, b\right)$$

where $z_{j,b} = s_b(\theta_{j,b}/y_j)^{1/\alpha_b}$

Conditional Representation

To derive the pdf, we make use of the positive stable (PS) distribution and its Laplace transform (Stephenson 2009):

- $\int_0^\infty h_1(s)\exp(-st)ds = \exp(-t^\alpha)$
- Take $S_b \sim \text{PS}(\alpha_b)\ \forall\ b \in \text{B} \setminus \text{B}_1$, and $\mathbf{S} = \{S_b \mid b \in \text{B} \setminus \text{B}_1\}$.
- Then we have for $j = 1, \ldots, d$:

  $$\Pr(X_j < x_j \mid \mathbf{S} = \mathbf{s}) = \exp\left[-\sum_{b \in B_{(j)}} s_b \left(\frac{\theta_{j,b}}{y_j}\right)^{1/\alpha_b}\right]$$

  while $X_1, \ldots, X_d$ are conditionally independent given $\mathbf{S} = \mathbf{s}$
- Thus, each marginal asymmetric logistic pdf can be given by:

  $$f_j(x_j|s) = \sigma_j^{-1} y_j^{-xi_j}\left[\sum_{b \in B_{(j)}} (z_{j,b}/\alpha_b)\right] \exp\left(-\sum_{b \in B_{(j)}} z_j, b\right)$$

  where $z_{j,b} = s_b(\theta_{j,b}/y_j)^{1/\alpha_b}$

Parameter Estimation

- We begin by estimating the marginal parameters ($\mu_j, \sigma_j,$ and ξ_j) from univariate data and keep them fixed throughout.
- Simplifying assumptions: we consider high-dimensional (5 and more) asymmetry parameters to be trivial; also, we assume a non-informative prior.
- To obtain estimates for α and θ, we use Metropolis-Hastings within Gibbs to calculate conditional posterior means.

Parameter Estimation

- We begin by estimating the marginal parameters ($\mu_j, \sigma_j,$ and ξ_j) from univariate data and keep them fixed throughout.
- Simplifying assumptions: we consider high-dimensional (5 and more) asymmetry parameters to be trivial; also, we assume a non-informative prior.
- To obtain estimates for α and θ, we use Metropolis-Hastings within Gibbs to calculate conditional posterior means.

Parameter Estimation

- We begin by estimating the marginal parameters $(\mu_j, \sigma_j,$ and $\xi_j)$ from univariate data and keep them fixed throughout.
- Simplifying assumptions: we consider high-dimensional (5 and more) asymmetry parameters to be trivial; also, we assume a non-informative prior.
- To obtain estimates for α and θ, we use Metropolis-Hastings within Gibbs to calculate conditional posterior means.

Thresholding

To select the best threshold, we minimize distances between our parametric fit $F_{\hat{\theta}}$ and the empirical distribution function $\hat{F}_n$ – which is given by:

$$\hat{F}_n(t_1, \ldots, t_d) = \frac{1}{nk} \sum_{j=1}^{d} \sum_{i=1}^{n} \mathbf{1}\{x_{ij} < t_j\}$$

Thresholding

To select the best threshold, we minimize distances between our parametric fit $F_{\hat{\theta}}$ and the empirical distribution function $\hat{F}_n$ – which is given by:

$$\hat{F}_n(t_1, \ldots, t_d) = \frac{1}{nk} \sum_{j=1}^{d} \sum_{i=1}^{n} \mathbf{1}\{x_{ij} < t_j\}$$

Thresholding

To select the best threshold, we minimize distances between our parametric fit $F_{\hat{\theta}}$ and the empirical distribution function $\hat{F}_n$ – which is given by:

$$\hat{F}_n(t_1, \ldots, t_d) = \frac{1}{nk} \sum_{j=1}^{d} \sum_{i=1}^{n} \mathbf{1}\{x_{ij} < t_j\}$$

Pickands Type

Pickands suggesting minimizing KS distance

$$d_k = sup_{\mathbf{q}} |\hat{F}_n(\mathbf{t}) - \hat{F}_\theta(\mathbf{t})|$$

with $k = 1, 2, ...[n/4]$

Joe Type

Joe suggests computing measure of association and setting cutoff to maximize tail dependence

$$max_k \; \tau_{1-k/n} = max \; \tau(\mathbf{t}|\mathbf{t} > \mathbf{C}_k)$$
$$= max_k \; 4E[C_\theta(\mathbf{t}|\mathbf{t} > \mathbf{C}_k)] - 1$$

[Joe 1992]

Generalization of Joe Type

Maximum likelihood over minimum distance:

$$max_\theta \ min_k \ d_\theta(\mathbf{q}, \mathbf{C}_{k,\theta})$$
$$= max_\theta \ min_k \ E[ln(\frac{dG_\theta(\mathbf{q})}{dG_\theta(\mathbf{C}_k)})]$$

Kendall's Tau on tails

$\tau_{1-k/n}$	$\tau_{.9}$	$\tau_{.95}$	$\tau_{.99}$
Pop-Pga	.072	.186	.472
GNP-Flood	.113	.270	.326
GNP-Drought	.208	.290	.168

70-percentile

75-percentile

80-percentile

85-percentile

90-percentile

95-percentile

99-percentile

Summary

- We fit a flexible model to high-dimensional data.
- This framework allows for the identification of multivariate extremes via either
  - $\mathcal{L}^1$ or Pickands distance
  - Kullback-Liebler or Expected Entropic Distance.

Summary

- The method (on data ending in 2003) identified several, *post hoc*, locations $\rightarrow$ Haiti.
- Compare thresholded ‘hotspots’ with disaster record from 2003-2010.