Multivariate GPD & mixtures

Holger Rootzén
GMMC & Stochastic Centre
Chalmers & Gothenburg University

Joint work with

Erik Brodin
Anne-Laure Fougères
John Nolan
Nader Tajvidi

www.math.chalmers.se/~rootzen/

Don’t look at the stars with a microscope --- and don’t use statistical methods tailored to means and typical behavior to study extreme occurrences: Use Extreme Value Statistics! (if not -- you will not see the important things)

This talk is about two recent “instruments” for looking at extreme values

Outline of talk → what to take home

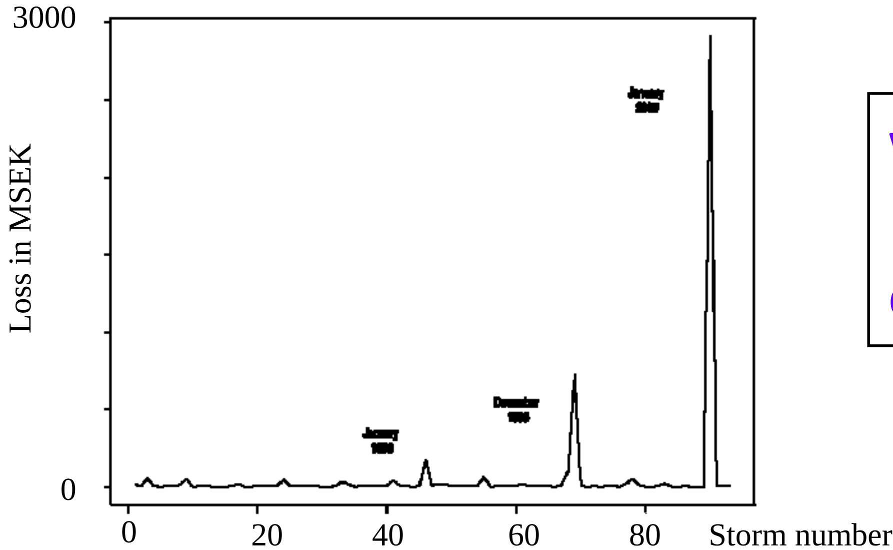
- Background: univariate GPD, multivariate EVD

- The multivariate Generalized Pareto distribution
  → limit of conditional distribution given at least one component is large: definition, independence, density, lower-dimensional marginal distributions

- Mixture models
  → analogues of Gaussian time series and spatial models

- Wind storm insurance
  → prediction intervals
  → bivariate models may be more realistic
  → structured thinking for "not yet seen catastrophes"

1-dim Peaks over Thresholds model – GP distribution



windstorm losses for Länsförsäkringar 1982 – 2005: excesses of 1.5 MSEK

excess times Poisson process, excess losses GP (Generalized Pareto):

$$H(x) = 1 - \left(1 + \frac{\gamma}{\sigma}x\right)_+^{-1/\gamma}$$

all mutually independent

conditional distribution of excesses

$$P\left(\frac{X-u}{\sigma_u} \leq x \,|\, X > u\right) \to H(x), \text{ as } u \to \infty$$

X = windstorm loss

dependence

Multivariate Extreme Value model

$M = (M_1 \ldots M_d)$ vector of componentwise maxima

Example: $d=2$ and M_1 = largest building loss
M_2 = largest forest loss

The multivariate extreme value distributions are the natural models for M. They are described in terms of marginal distributions and dependence, typically specified in terms of a "spectral measure" which gives the "angular distribution". Much studied, but still only a beginning. In example the observed maxima might be yearly, but the aim prediction for 10 or 100 or more years.

Basis for Generalized Pareto and Extreme Value distributions

- stability: maxima of vectors which are EV distributed are also EV; going to higher levels preserves the GP distribution of excesses

- asymptotics: maxima of many independent vectors are often (approximately) EV distributed; asymptotically excesses of high levels are are GP when maxima are EV

- ”transition”: easy to go back and forth between GP and EV

.

The multivariate Generalized Pareto distribution should:

- be the natural distribution for excesses over high thresholds by multivariate random variables -- i.e. it should have the stability and asymptotics properties from previous slide

- should describe what happens to the other variables when one or more of the variables exceed their threshold(s)

The multivariate GP distribution: conditional distribution of a vector given that at least one component is ”large” (cf also Segers 2004)

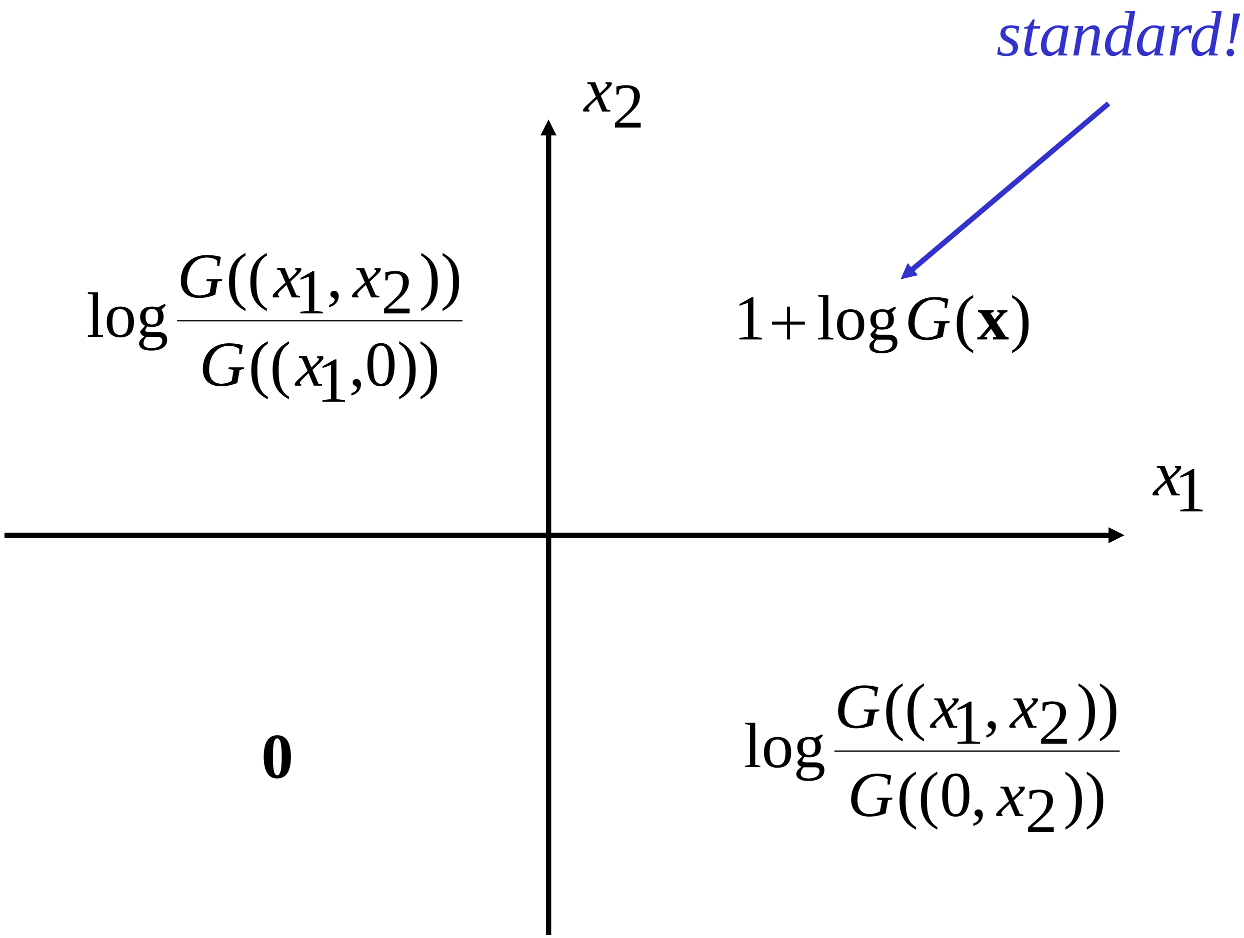
$$H(\mathbf{x}) = \log \frac{G(\mathbf{x})}{G(\mathbf{x} \wedge \mathbf{0})}, \qquad G \quad \text{multivariate EV}, \quad G(0) = e^{-1}$$

from a multivariate EV distribution you get a multivariate GP distribution; parametric families of EV distributions give corresponding families of GP distributions; and vice versa

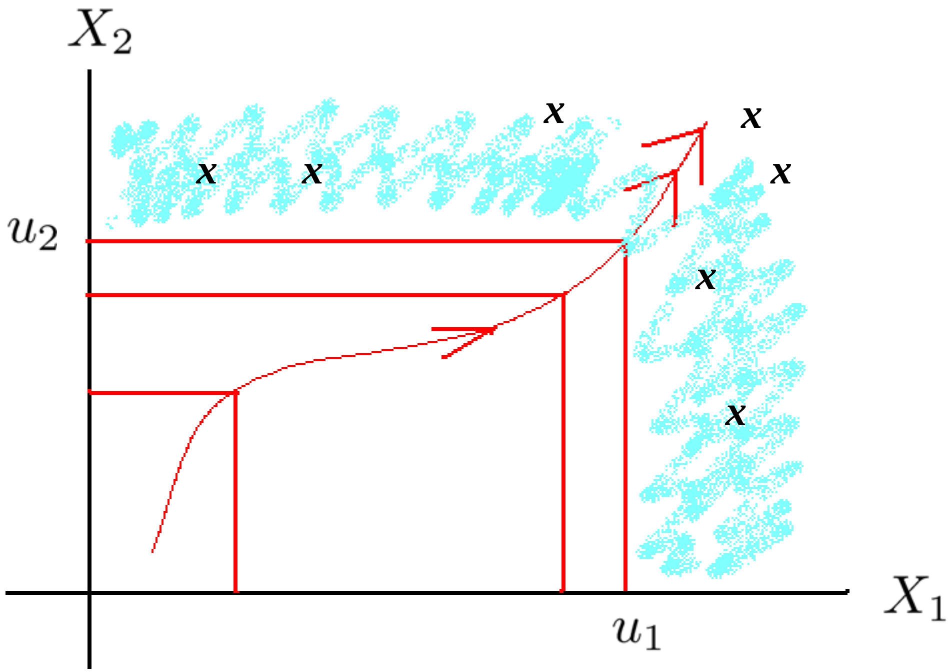
- this family is the only one which is stable under (a suitably coordinated) change of excess levels

- exceedances asymptotically have a multivariate GP distribution if and only if maxima are asymptotically multivariate EV

$$\log \frac{G(\mathbf{x})}{G(\mathbf{x} \wedge \mathbf{0})}$$



- **stability**
- **asymptotics**



GP: approximate (conditional) distribution for $(X_1 - u_1, X_2 - u_2)$ in shaded region

-- exceedance times approximately Poisson process

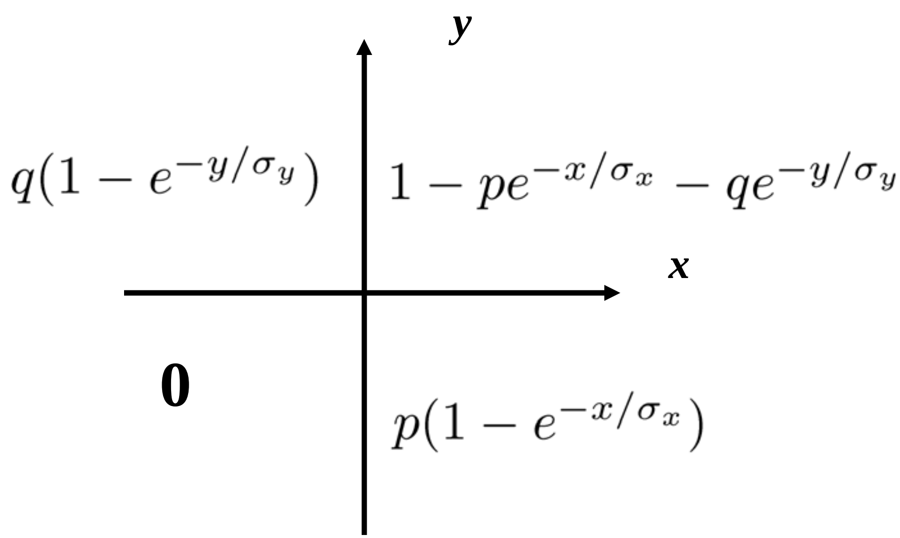
An independent margins example

$(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. $M_n = (\max_{1 \le i \le n} X_i, \max_{1 \le i \le n} Y_i)$

X, Y independent, standard exponential

$$\left(\frac{\max_{1\le i\le n} X_i - \log pn}{\sigma_x}, \frac{\max_{1\le i\le n} Y_i - \log qn}{\sigma_y}\right) \stackrel{d}{\to} \exp(-pe^{-x/\sigma_x} - ge^{-y/\sigma_y}) =: G(x,y)$$

$$P\left(\frac{X - \log pt}{\sigma_x}, \frac{Y - \log qt}{\sigma_y} \le (x,y) \middle| X > \log pt \text{ or } Y > \log qt\right) \stackrel{t\to\infty}{\to} \log \frac{G(x,y)}{G(\min(0,x), \min(0,x))}$$



$$\mathbf{X}_\infty = \begin{cases} (\exp(\sigma_x), -\infty), & \text{prob } p \\ (-\infty, \exp(\sigma_y)), & \text{prob } q \end{cases}$$

An absolutely continuous example: a bivariate symmetric logistic distribution with margins normalized to Frechet, $\alpha = 2$

$$G(x,y) = \exp\{-(\frac{1}{(x+\sqrt{2})^2} + \frac{1}{(y+\sqrt{2})^2})^{1/2}\}, \qquad x, y > \sqrt{2},$$

$$G(0,0) = e^{-1}$$

y

$H(x,y)$

$$1-(\frac{1}{(x+\sqrt{2})^2} + \frac{1}{(y+\sqrt{2})^2})^{1/2}\}$$

0

Change
x to y

x

0

$$(\frac{1}{2} + \frac{1}{(y+\sqrt{2})^2})^{1/2} - (\frac{1}{(x+\sqrt{2})^2} + \frac{1}{(y+\sqrt{2})^2})^{1/2}$$

0

If $(X_1, \dots X_d)$ has a multivariate GP distribution then the marginal distribution of a component X_i is not a univariate GP distribution

However, if in the marginal distribution of X_i one conditions on $X_i > 0$ the result is a univariate GP distribution

The reason is that in the multivariate GP distribution the conditioning is on one of the d components being large, while in the univariate GP distribution the condition is that the variable itself is large

Similar results hold for higher-dimensional marginal distributions

Mixture models for maxima

→ mixture models for multivariate GPD

S pos. stable if $\quad E(e^{-tS}) = \exp(-t^{\alpha}),\quad \text{where} \quad 0 < \alpha < 1$

Gumbel distribution $\quad G(x) = \exp\left(-\exp(-\frac{x-\mu}{\sigma})\right)$

if $\quad P(X \le x \mid S) = \exp(-S\exp(-\frac{x-\mu}{\sigma}))$

then $\quad P(X \le x) = E(\exp(-S\exp(-\frac{x-\mu}{\sigma})) = \exp(-\exp(-\frac{x-\mu}{\sigma/\alpha}))$

Gumbel!

Watson & Smith (1985), Hougaard (1986), Crowder (1989), Tawn (1990), Coles & Tawn (1991), Stephenson (2003)

“same” holds for the general EV distribution

→ large, flexible and interpretable class of "logistic" models with Gumbel margins and with maxima over all kinds of sets Gumbel distributed

$$X_t = G_t + \sigma \log S_t, \qquad t \in T$$

i.i.d. Gumbel (μ_t, σ) **pos** α**-stable process**

- components of variance --- S_t sum of effects
- time series --- S_t ARMA
- spatial --- S_t spatial ARMA process
- continuous parameter --- S_t continous parameter MA
- hierarchical models (McFadden)

distribution funktion easy --- density hard

A spatial moving average process

(u,v) spatial coordinates, G_t i.i.d ~ $Gumbel(\mu,\sigma)$

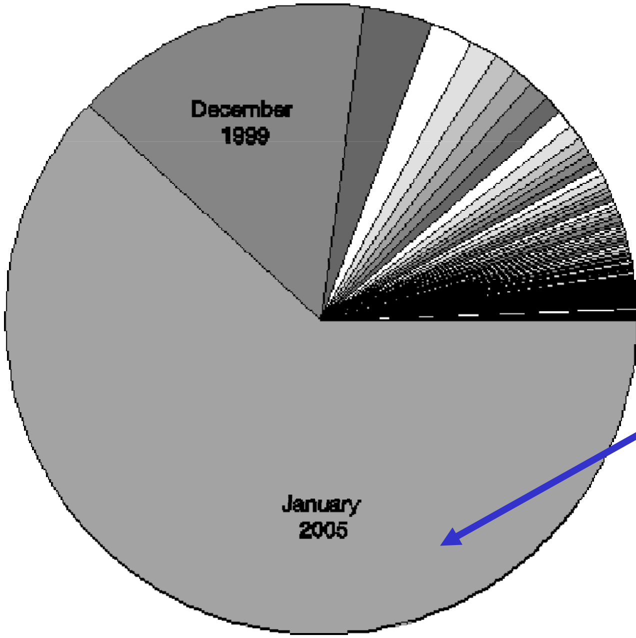
$$X_{(u,v)} = G_{(u,v)} + \sigma \log \int \exp(-\beta|(u,v)-(x,y)|^{\gamma}) S(d(x,y))$$

$$P\big(X_{(u_i,v_i)} \le x_{t_i}, i = 1,\dots n\big)$$
$$= \exp(-\int (\textstyle\sum_{i=1}^{n} \exp(-\beta|(u_{t_i},v_{t_i})-(x,y)|^{\gamma}) e^{-\frac{x_{t_i}-\mu}{\sigma}})^{\alpha} dxdy)$$

The d.f. of the corresponding multivariate GPD then, in the third quadrant, is (up to appropriate normalization)

$$1 - \int (\textstyle\sum_{i=1}^{n} \exp(-\beta|(u_{t_i},v_{t_i})-(x,y)|^{\gamma}) e^{-\frac{x_{t_i}-\mu}{\sigma}})^{\alpha} dxdy$$

Perhaps tractable for $n = 5$ to 10



Windstorm losses for Länsförsäkringar 1982-2005

Gudrun January 2005
326 MEuro loss
72 % due to forest losses
4 times larger than second largest

The real problem!

The insurance problems

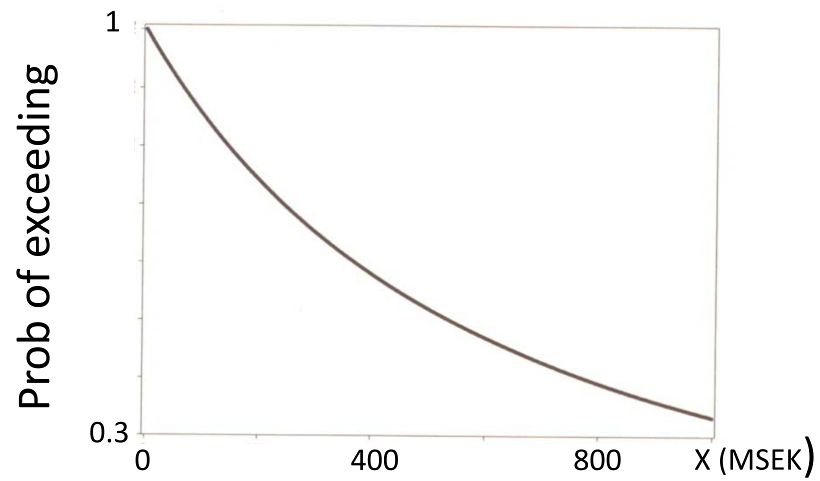
How much reinsurance should Länsförsäkringar buy?

How should Länsförsäkringar adjust if its forest insurance portfolio grows?

What statistics can provide: estimates of high quantiles of distribution of maximum loss (→ Extreme Value Statistics – we used PoT).

History: result of 1994 analysis of 1982-1993 LFAB data

Risk (MSEK)	next year	next 5 years	next 15 years
10%	66	215	473
1%	366	1149	2497



$$Y_i \ \mathrm{GP}(y;\sigma_t,\gamma)$$

$$\sigma_t = \exp(\alpha + \beta t)$$

$$\hat{\alpha} = 15.1$$

$$\hat{\beta} = .013 \pm .013$$

no evidence of trend in extremes

conditional probability that a loss in excees of the reinsurance level 850 MSEK exceeds x

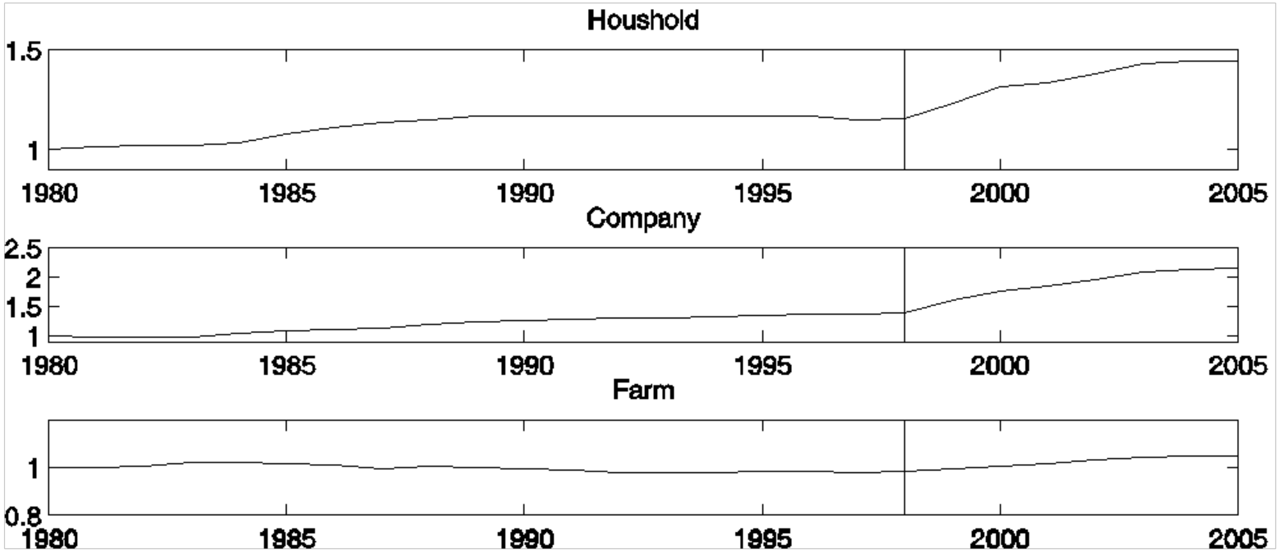
Gudrun: 2912 MSEK, after 12 years

Windstorms of 1902 and 1969 probably comparable to Gudrun

The data

- all individual claims for windstorm damage to buildings and forest paid out by Länsförsäkringar during 1982-2005
- inflation adjusted into 2005 prices using the factor price index for building
- appr 80 storm events where selected based on exceedances of three-day moving sums, different selection for univariate and bivariate analysis
- simplistic correction for portfolio changes

relative change in number of policies

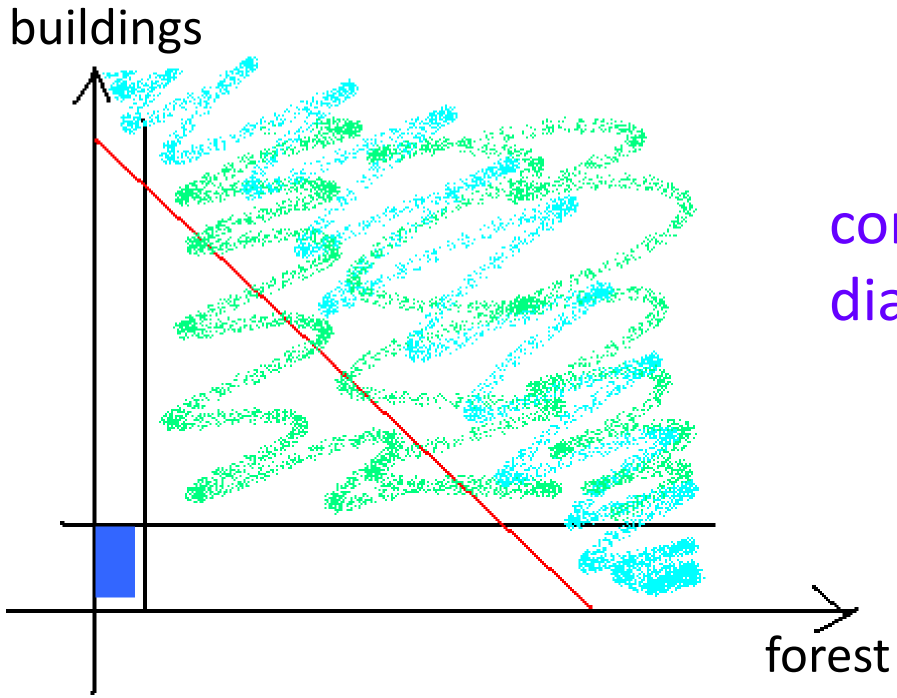


One-dimensional analysis: total loss, standard PoT, ML estimation

Two-dimensional analysis: (loss from buildings , loss from forest) bivariate GP model with symmetric logistic distribution, simultaneous ML estimation of all parameters, numerical computation of quantiles

Covariates may be incorporated in parameters, in the "usual way"

Modelling, estimation and computation in different areas!



computation in area over diagonal line

estimation using data in open rectangle

assumed GP model above and to the right of blue square

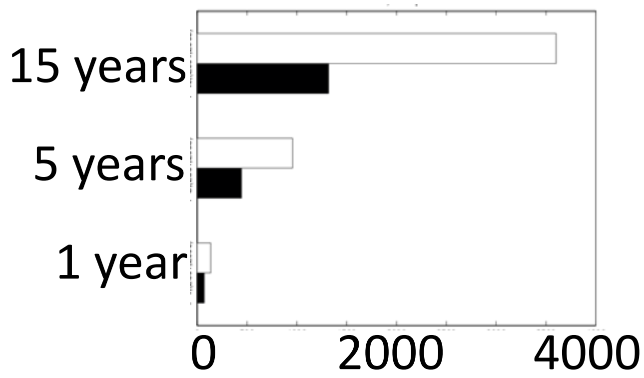
Prediction intervals

A level *p* prediction interval includes the predicted quantity (say, the maximum loss during the next 15 years) with probability *1-p.*

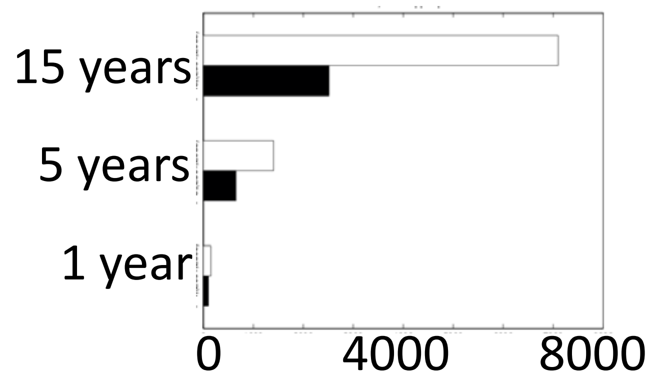
A "naïve" prediction interval ends at the estimated *p*-th quantile from the top. However, this usually doesn't achieve the level *p,* because of estimation uncertainty.

- 1 dim: used bootstrap approach due to Hall, Peng & Tajvidi
- >1 dim: no method available

Results of univariate analysis

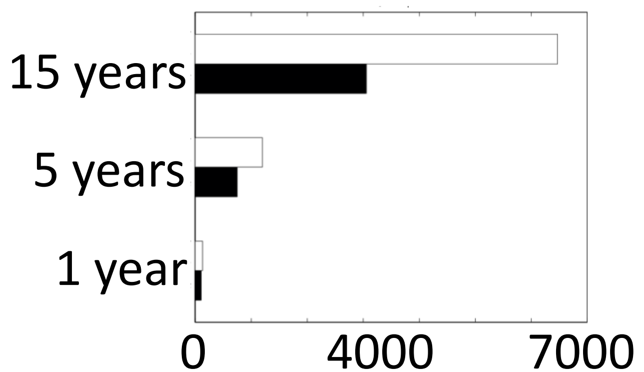


"Naïve" 10% prediction intervals. Black1982-2004 data, white 1982-2005

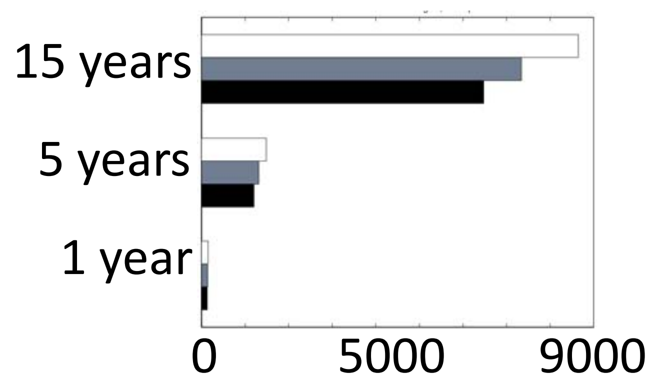


Bootstrapped 10% prediction intervals. Black 1982-2004 data, white 1982-2005

Results of bivariate analysis



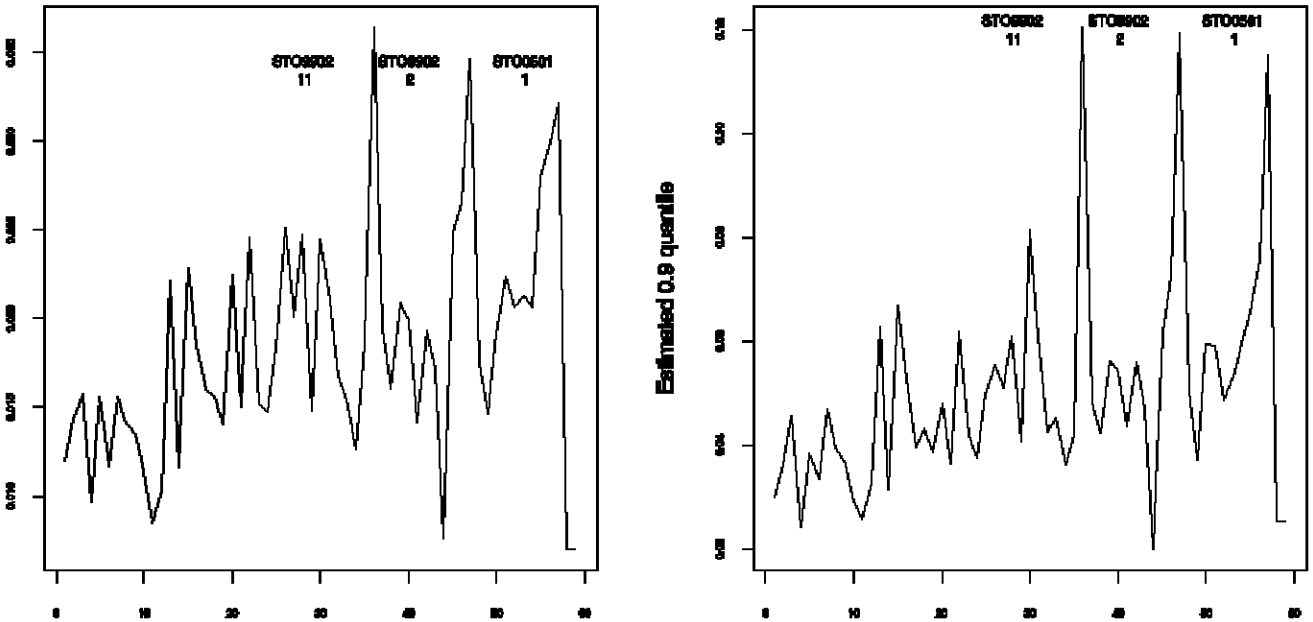
"Naïve" 10% prediction intervals. Black1982-2004 data, white 1982-2005



Black: no portfolio change, grey: 20% higher forest exposure, white 50% higher

Are windstorm losses getting worse?

- LR-tests of linear trend in shape parameter gave *p=.90*
- LR-test of exp linear trend in scale parameter gave *p=.10*
- 5 records in 73 observations: as expected from i.i.d theory



.7 and .9 quantiles of individual claims for storm events with more than 100 claims → significant trends in individual claims

Conclusions

- both univariate and bivariate models fitted the data and gave credible prediction intervals – quantiles substantially different, changes in probabilities of exceeding much less dramatic

- bivariate analysis may give the most correct evaluation of the real uncertainties

- predicted losses were rather insensitive to changes in portfolio size

- **organizations should develop systematic ways of thinking about "not yet seen" types of disasters**

... and conclusions for statistics

- existing models can handle the problems, but still:
- much more thinking about prediction intervals is desirable
- much more thinking about multivariate peaks over thresholds modelling is needed
- should statistics involve itself in thinking about “not yet seen” disasters?

H. Rootzén and N. Tajvidi (1997). Extreme value statistics and wind storm losses: a case study. *Scand. Actuarial J.,* 70-94, reprinted in “Extremes and integrated risk management”, Risk Books 2000.

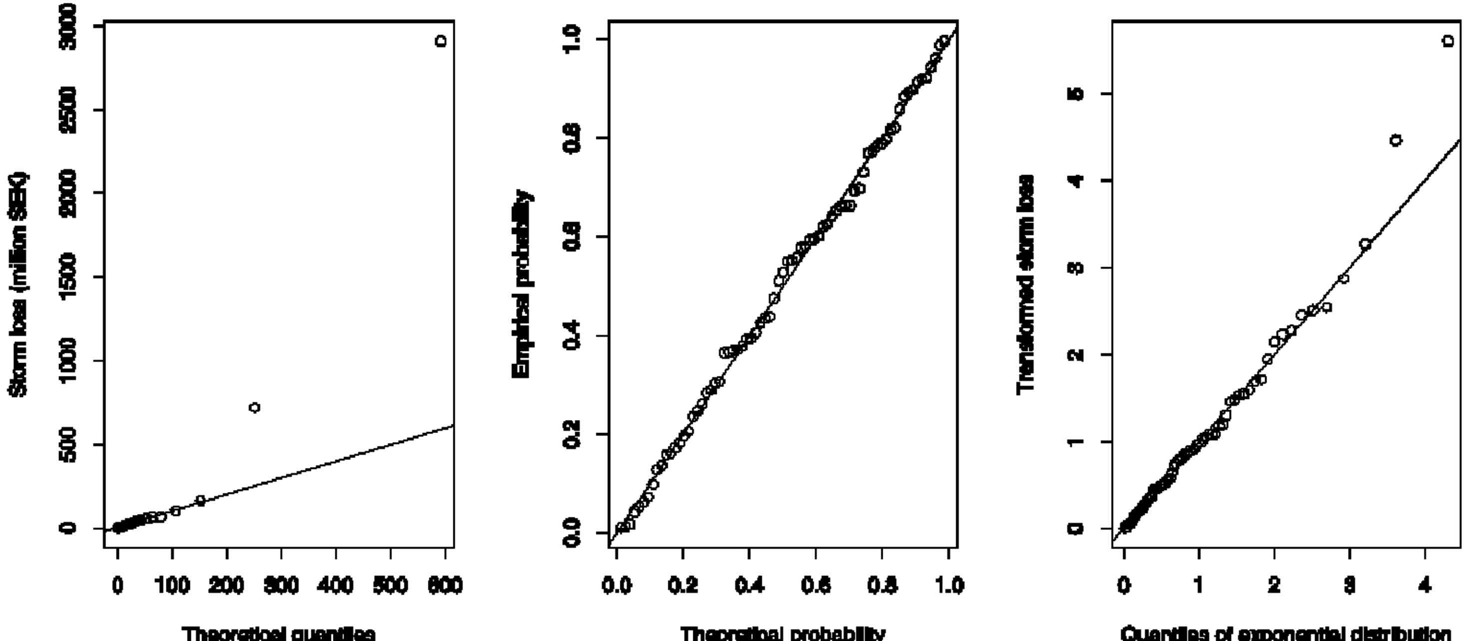
H. Rootzén, and N.Tajvidi (2000). Can losses caused by wind storms be predicted from meteorological observations? *Scand. Actuarial J.*, 162-175.

H. Rootzén, and N.Tajvidi (2006). The multivariate Generalized Pareto Distribution. *Bernoulli* **12**, 917-930

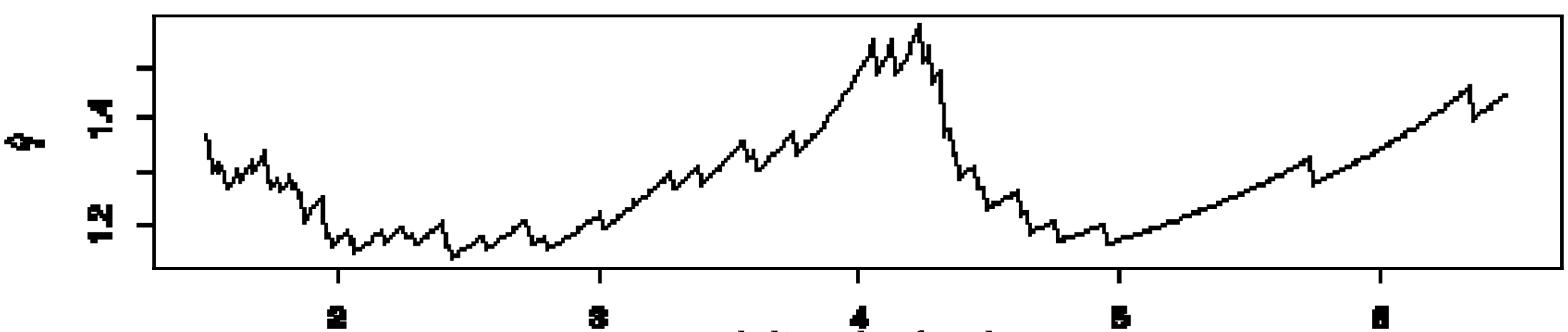
E. Brodin and H. Rootzén (2009). Modelling and predicting extreme wind storm losses. *Insurance: Mathematics and Economics*,

Does the univariate model fit?

QQ-plot PP-plot Exp QQ-plot



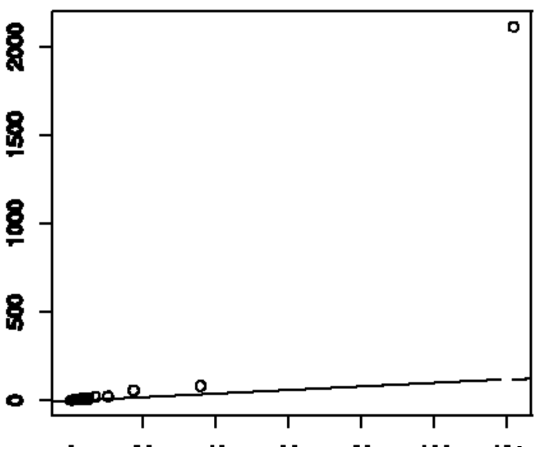
Theoretical quantiles Theoretical probability Quantiles of exponential distr



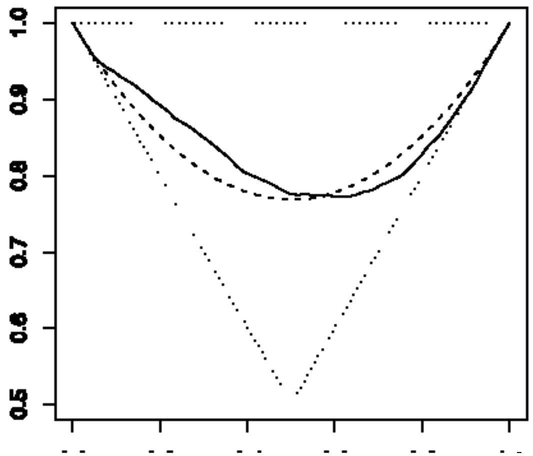
Parameter stability plot for shape parameter

Does the bivariate model fit?

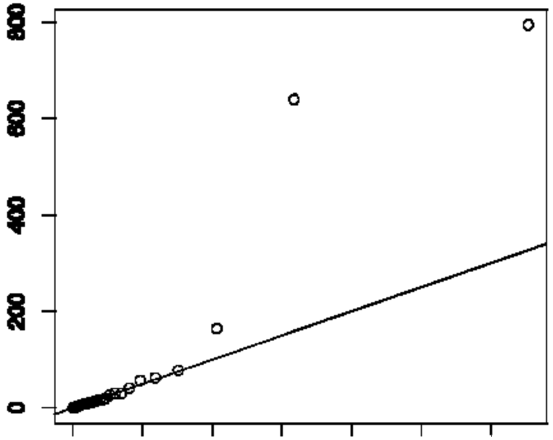
QQ-plot forest | Pickand's dependence funct | QQ-plot buildings



Theoretical quantiles



t



Theoretical quantiles

Bivariate GPD-model with symmetric logistic dependence function, all parameters estimated simultaneously

Ex2: 2 dim, totally dependent margins,

$$\mathbf{X}_i = (X_i, Y_i), \quad M_n = (\max_{1\le i\le n}(X_i), \max_{1\le i\le n}(Y_i))$$

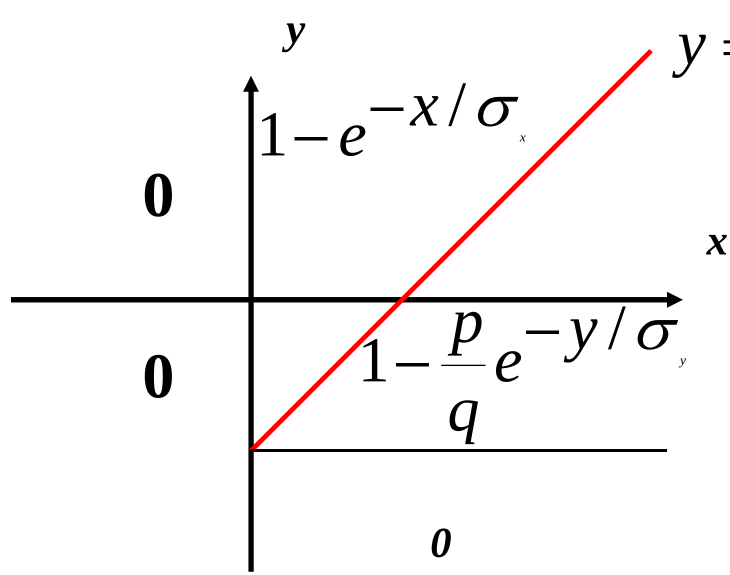
X=Y exponential,

$$\sigma = (\sigma_x^{-1}, \sigma_y^{-1}), \ \mathbf{u}_t = (\log pt, \log qt), \ \mathrm{p} < \mathrm{q}$$

$$\frac{(\mathbf{M}_n - \mathbf{u}_n)}{\sigma} \Rightarrow \exp(-\exp(\min(x/\sigma_x + \log p, y/\sigma_y + \log q)) = G(x, y)$$

$$\mathbf{X_u} = \frac{(X, Y) - \mathbf{u}_t}{\sigma}, \qquad \mathrm{P}(\mathbf{X}_u \le \mathbf{x} \mid (\mathbf{X}_u \le \mathbf{0})^c) \to p \log \frac{G(\mathbf{x})}{G(\mathbf{x} \wedge \mathbf{0})}$$

$$y = \sigma_y(x/\sigma_x + \log\frac{p}{q}),$$



$$\boxed{\begin{array}{l} \mathbf{X}_\infty = (Z, \sigma_y(Z/\sigma_x + \log\frac{p}{q}), \\ \mathrm{Z} \sim \exp(\sigma_x) \end{array}}$$