A Conditional Approach to Modelling Multivariate Extremes

Janet E. Heffernan

Lancaster University & J. Heffernan Consulting

INTERFACE 2008 RISK : Reality

May 21 – 24, 2008

Durham, NC

Looking at Tails: why we need asymptotics

- Statistical maxim: extrapolate at your peril!
- Risk assessment: need to understand tails!

How can we reconcile these opposing requirements?

Looking at Tails: why we need asymptotics

- Statistical maxim: extrapolate at your peril!
- Risk assessment: need to understand tails!

How can we reconcile these opposing requirements?

Asymptotic models for tail behaviour:

- componentwise maxima
- point process: simultaneous excess of high marginal thresholds

Inadequacies of previous methods

Componentwise maxima / Multivariate Regular Variation

- rich class of models for Asymptotically Dependent variables
- models Asymptotically Independent variables as exactly independent

Inadequacies of previous methods

Componentwise maxima / Multivariate Regular Variation

- rich class of models for Asymptotically Dependent variables
- models Asymptotically Independent variables as exactly independent

Ledford and Tawn / Hidden Regular Variation

- richer class of models for Asymptotically Independent variables
- only allows investigation of joint tail

Conditional extreme value models

Wish list:

- rich class of models for Asymptotically Dependent/Independent and Negatively dependent variables
- multivariate (not just bivariate)
- examination of tail region where "action" is !

Conditional extreme value models

Wish list:

- rich class of models for Asymptotically Dependent/Independent and Negatively dependent variables
- multivariate (not just bivariate)
- examination of tail region where "action" is !

Approach using Conditional Asymptotics:

1. condition on one variable being extreme
2. examine behaviour of remaining variables conditional on having an extreme component

method developed with Jonathan Tawn, appeared in JRSS B, 2004.

Characterising the tail of a multivariate distribution

- $\boldsymbol{X} = (X_1, \ldots, X_d)$; $\boldsymbol{X}_{-i}$ is $\boldsymbol{X}$ excluding X_i.
- n i.i.d. observations of $\boldsymbol{X}$ from unknown distribution F

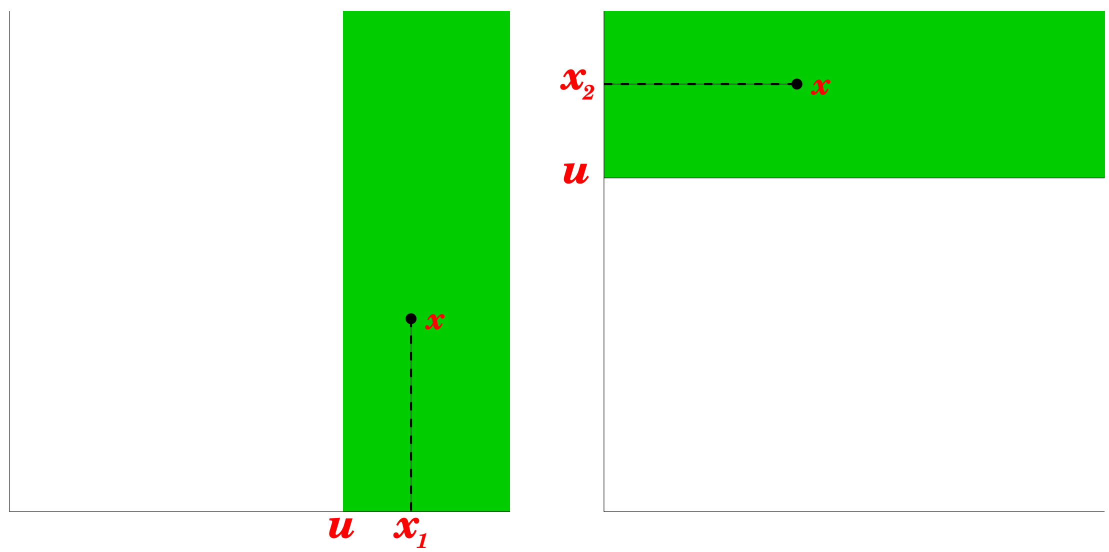
Three stages of understanding required:

$\Pr(X_i > u)$ assessed empirically;

$X_i \mid X_i > u$ needs a marginal model;

$\boldsymbol{X}_{-i} \mid X_i = x_i$ for $x_i > u$ additionally needs a dependence model.

We look at all three characteristics for each conditioning variable i:



Marginal model:

Semi-parametric:

- asymptotic justification for the generalised Pareto distribution (GPD) as a conditional model above a threshold

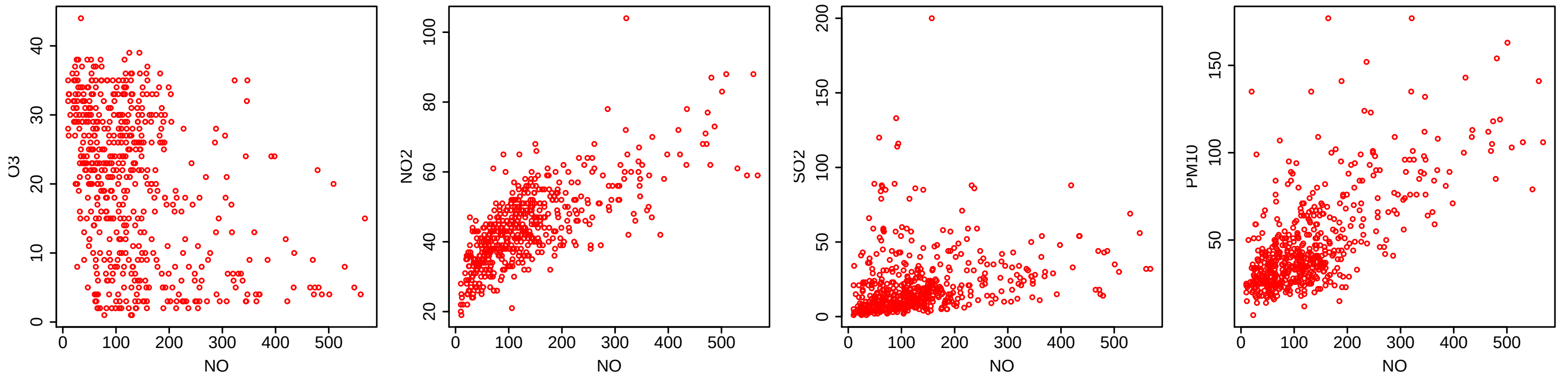
$$\Pr(X_i > u + x \mid X_i > u) \sim \text{GPD}(x)$$

  where $x > 0$ and u is a high threshold for variable X_i.

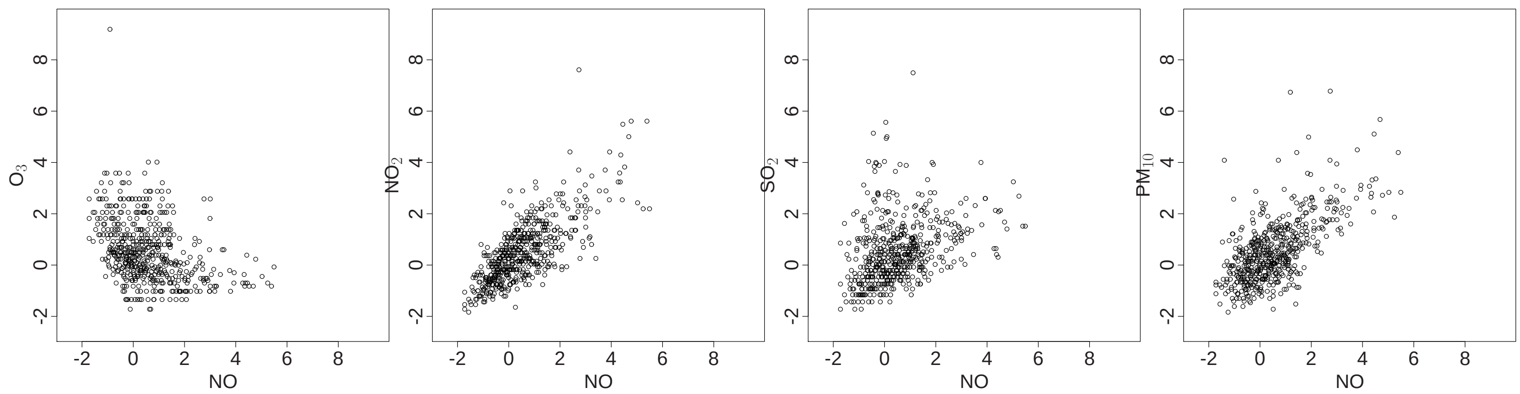
- Below the threshold, we use the empirical distribution $\tilde{F}_{X_i}(x)$.

Transforming data to known margins

Leeds U.K. air quality data:



We transform variable X to have standard Gumbel marginal distributions.



Conditional asymptotics for multivariate extremes

For d dimensional $\boldsymbol{Y}$, we consider for each $i = 1, \ldots, d$

$$\mathrm{P}\left(\boldsymbol{Y}_{-i} < \boldsymbol{z}_{-i} \,|\, Y_i = y_i\right)$$

as $y_i \to \infty$.

(Here, $\boldsymbol{Y}$ has Gumbel margins.)

Asymptotic assumptions

We assume that there exist vector normalising functions

- $\boldsymbol{a}_{\,|\,i}(y_i) : \mathbb{R} \to \mathbb{R}^{(d-1)}$
- $\boldsymbol{b}_{\,|\,i}(y_i) : \mathbb{R} \to \mathbb{R}^{(d-1)}$

such that for all fixed $\boldsymbol{z}_{-i}$ and any sequence $y_i \to \infty$

$$\lim_{y_i \to \infty} \Pr\left\{ \frac{\boldsymbol{Y}_{-i} - \boldsymbol{a}_{\,|\,i}(y_i)}{\boldsymbol{b}_{\,|\,i}(y_i)} \leq \boldsymbol{z}_{-i} \,\middle|\, Y_i = y_i \right\} = G_{\,|\,i}(\boldsymbol{z}_{-i})$$

where $G_{\,|\,i}$ has non-degenerate margins.

Operations involving vectors are componentwise.

Form of normalising functions

Analogous to parameters of GEV in univariate EVT:

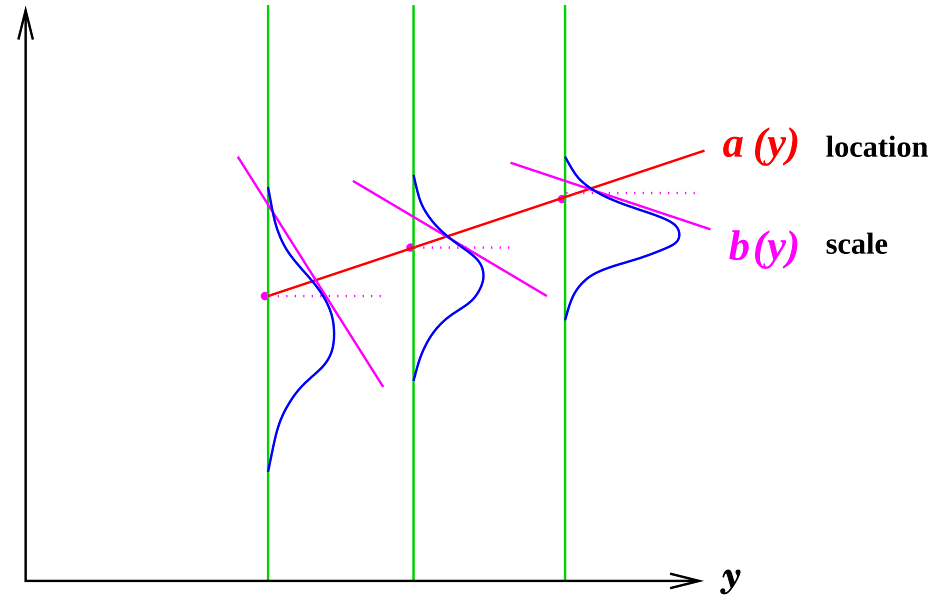
$a_{j\,|\,i}(y_i)$ satisfies:

$$\lim_{y_i \to \infty} \Pr\{Y_j > a_{j\,|\,i}(y_i) \,|\, Y_i = y_i\} > 0$$

$b_{j\,|\,i}(y_i)$ is then given by:

$$b_{j\,|\,i}(y_i) = h_{j\,|\,i}\{a_{j\,|\,i}(y_i) \,|\, Y_i = y_i\}^{-1}$$

where $h_{j\,|\,i}$ is the conditional hazard function for $Y_j \,|\, Y_i = y_i$.



What do the functions look like?

We examined a wide range of multivariate distributions and found

- $\boldsymbol{a}_{\mid i}(y)$ and $\boldsymbol{b}_{\mid i}(y)$ fall in the parametric family:

$$\begin{aligned} \boldsymbol{a}_{\mid i}(y) &= \boldsymbol{a}_{\mid i} y + I_{\{\boldsymbol{a}_{\mid i}=\mathbf{0}, \boldsymbol{b}_{\mid i}<\mathbf{0}\}}(\boldsymbol{c}_{\mid i} - \boldsymbol{d}_{\mid i} \log y) \\ \boldsymbol{b}_{\mid i}(y) &= y^{\boldsymbol{b}_{\mid i}} \end{aligned}$$

- There is no general form for $G_{\mid i}$ or for its marginal distributions $G_{j \mid i}$.

- $\boldsymbol{a}_{\mid i}, \boldsymbol{b}_{\mid i}, \boldsymbol{c}_{\mid i}$ and $\boldsymbol{d}_{\mid i}$ are vector constants and I is an indicator function
- $0 \leq a_{j \mid i} \leq 1$, $-\infty < b_{j \mid i} < 1$, $-\infty < c_{j \mid i} < \infty$, $0 \leq d_{j \mid i} \leq 1$ for all $j \neq i$.

Model assumptions

Asymptotic structure assumed to hold exactly above high threshold

$$\Pr\left\{ \frac{\boldsymbol{Y}_{-i} - \boldsymbol{a}_{\,|\,i}(y_i)}{\boldsymbol{b}_{\,|\,i}(y_i)} \leq \boldsymbol{z}_{-i} \,\middle|\, Y_i = y_i \right\} = G_{\,|\,i}(\boldsymbol{z}_{-i})$$

for $y_i > u_Y$.

Then

$$\boldsymbol{Z}_{-i} = \frac{\boldsymbol{Y}_{-i} - \boldsymbol{a}_{\,|\,i}(y_i)}{\boldsymbol{b}_{\,|\,i}(y_i)}, \text{ for } y_i > u_Y,$$

are assumed to follow $G_{\,|\,i}$ and be independent of Y_i.

Model assumptions

Asymptotic structure assumed to hold exactly above high threshold

$$\Pr\left\{\frac{\boldsymbol{Y}_{-i}-\boldsymbol{a}_{\,|\,i}(y_i)}{\boldsymbol{b}_{\,|\,i}(y_i)} \le \boldsymbol{z}_{-i} \,\middle|\, Y_i = y_i\right\} = G_{\,|\,i}(\boldsymbol{z}_{-i})$$

for $y_i > u_Y$.

Then

$$\boldsymbol{Z}_{-i} = \frac{\boldsymbol{Y}_{-i}-\boldsymbol{a}_{\,|\,i}(y_i)}{\boldsymbol{b}_{\,|\,i}(y_i)}, \text{ for } y_i > u_Y,$$

are assumed to follow $G_{\,|\,i}$ and be independent of Y_i.

Model is semiparametric:

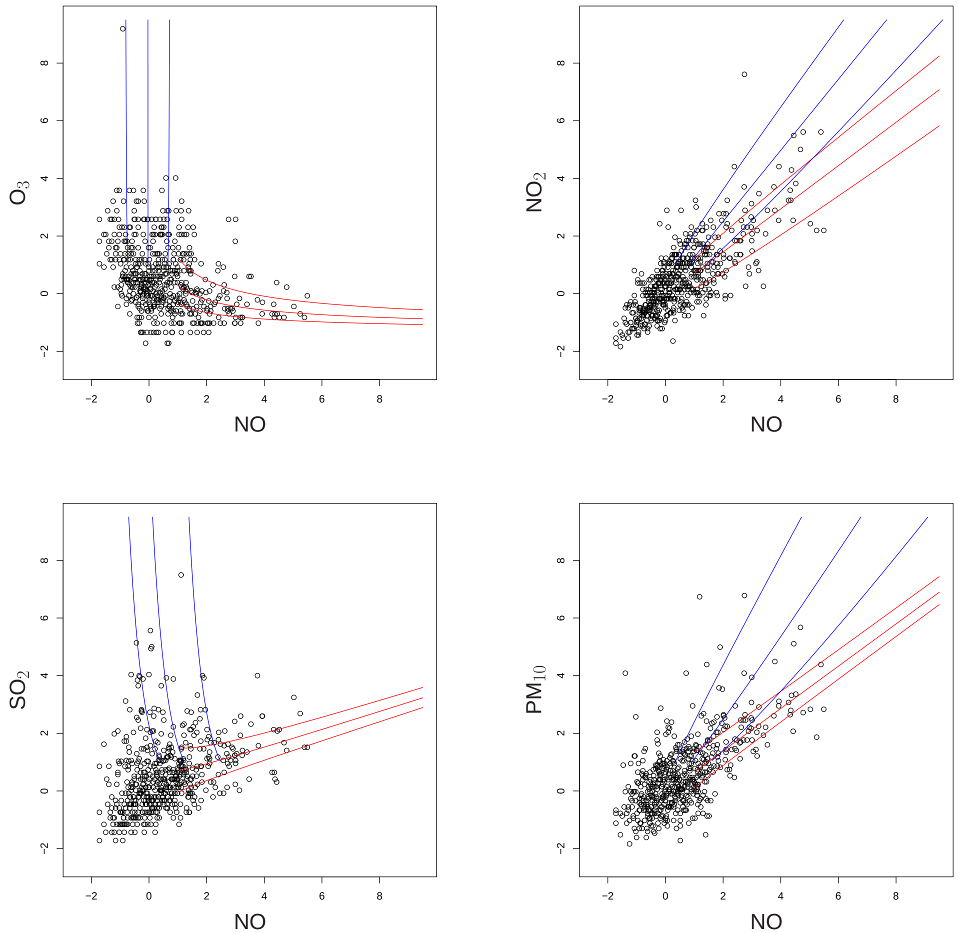
- parametric forms for $\boldsymbol{a}_{\,|\,i}(y)$ and $\boldsymbol{b}_{\,|\,i}(y)$;
- a nonparametric model for $G_{\,|\,i}$.

Air quality data — fitted model

$$Y_j \mid Y_i = y_i \quad : \quad Y_j = \hat{a}_{j\,|\,i}(y_i) + \hat{b}_{j\,|\,i}(y_i) Z_{j\,|\,i}$$

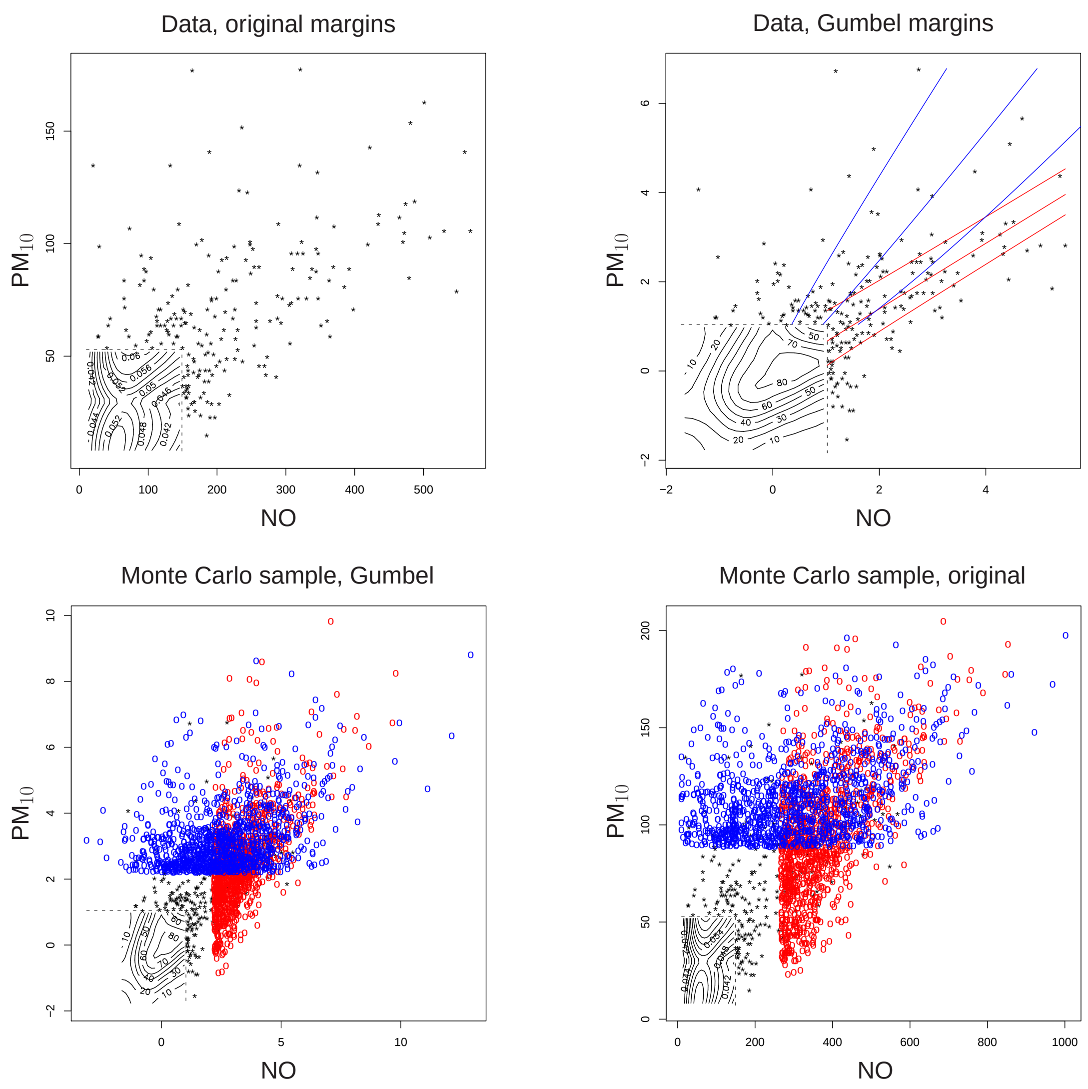
for $y_i > u_Y$.

Fitted models (Gumbel scale):



Air quality data — extrapolation

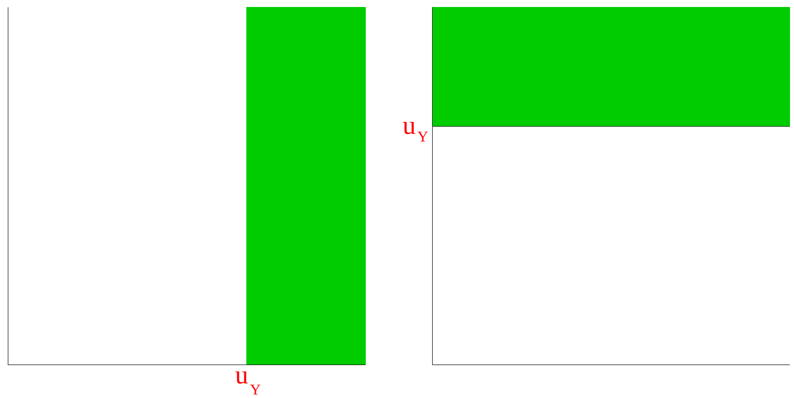
PM_{10} and NO:



Combining conditional models

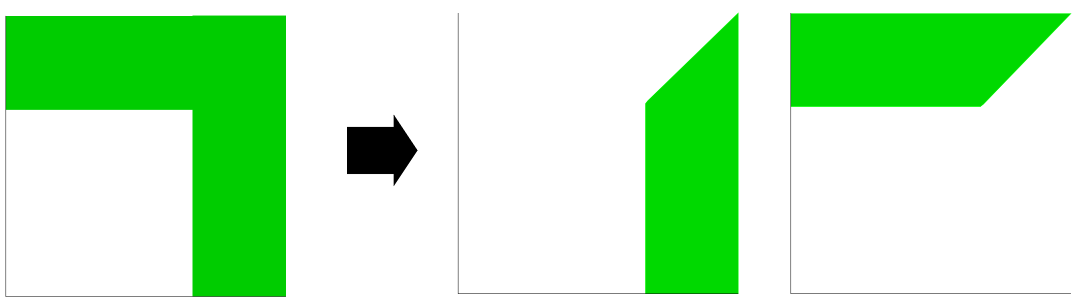
What about self-consistency on $\{\boldsymbol{Y} : Y_i > u_Y, Y_j > u_Y\}$?

$$\boldsymbol{Y} \mid Y_i > u_Y \qquad\qquad \boldsymbol{Y} \mid Y_j > u_Y$$



We combine estimates from both models by partitioning:

$$\Pr(\boldsymbol{Y} \in C) \quad = \quad \Pr(\boldsymbol{Y} \in C_1) + \Pr(\boldsymbol{Y} \in C_2)$$

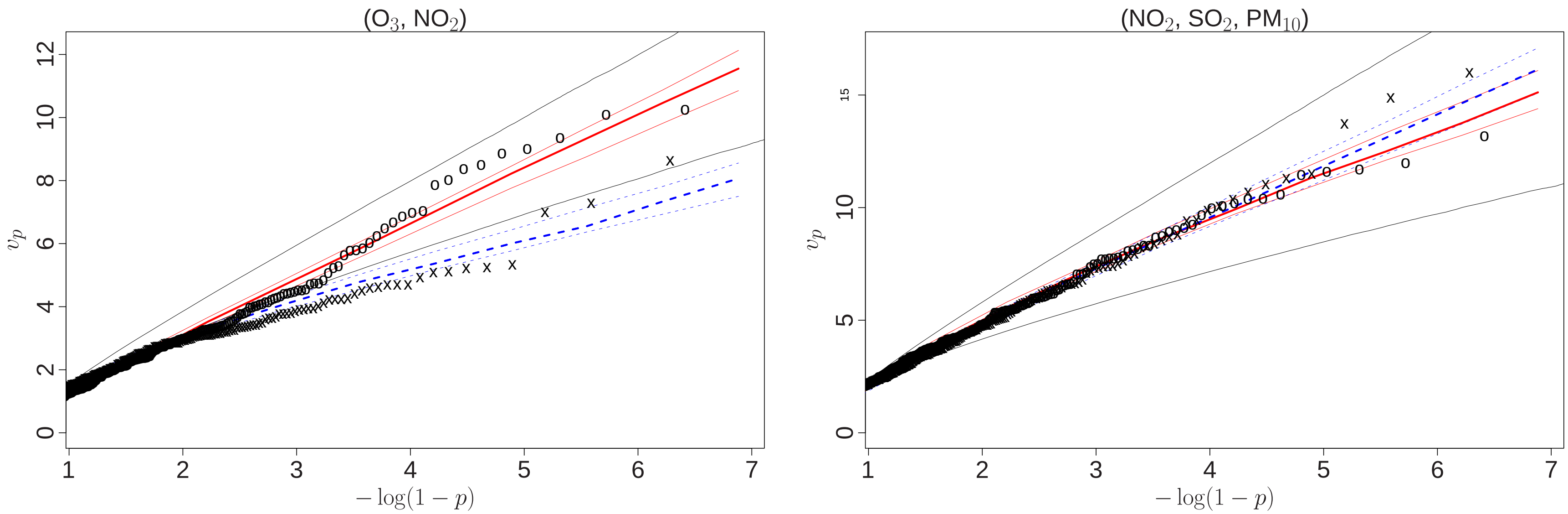


Air quality — estimating functionals

Example: estimate return levels v_p satisfying

$$\Pr\left(\sum_{i=1}^{m} Y_i > v_p\right) = p$$

for a given p.



——— summer   - - - - - winter

——— perfect dependence and independence

Conditional Approach – Conclusions

Advantages:

- accommodates any functional of multivariate extreme events
- handles asymptotic dependence and asymptotic independence including negative dependence
- is truly multivariate – not just bivariate
- has straightforward regression interpretation

Conditional Approach – Conclusions

Advantages:

- accommodates any functional of multivariate extreme events
- handles asymptotic dependence and asymptotic independence including negative dependence
- is truly multivariate – not just bivariate
- has straightforward regression interpretation

Difficulties:

- Consistency of different conditional models genuine concern
- Semi-parametric nature makes inference non-standard

Recent Applications and Developments

Applications of conditional modelling approach

- Mendes and Pericchi (2008) Extremal risk of flooding in Puerto Rico: flood risk of 5 rivers
- Paulo et al (2006) Multivariate intakes of food chemicals: functionals of 3 dimensional variables

Recent Applications and Developments

Applications of conditional modelling approach

- Mendes and Pericchi (2008) Extremal risk of flooding in Puerto Rico: flood risk of 5 rivers
- Paulo et al (2006) Multivariate intakes of food chemicals: functionals of 3 dimensional variables

Theoretical developments

- Heffernan and Resnick (2007) Limit laws for random vectors with an extreme component Ann. Appl. Probab.
- Eastoe, Heffernan and Tawn (2008) Using the Conditional Model to estimate Multivariate Extreme Value model Dependence Functions.
- Resnick and Das (2008) Consistency of models from different conditioning variables.

References

Heffernan, J.E. and Tawn, J.A. (2004). *A Conditional Approach for Multivariate Extreme Values (with discussion).* JRSS B **66** Part 3, pp. 497-546.

Heffernan, J.E. and Resnick, S.I. (2007) Limit laws for random vectors with an extreme component *Ann. Appl. Probab.*, **17**, No. 2, 537–571.