Practical Issues in Applications of Multivariate Extreme Values

Jonathan Tawn

with

Caroline Keef and Mark Latham

Lancaster, UK

Two Applications

- **Sea-surge data**

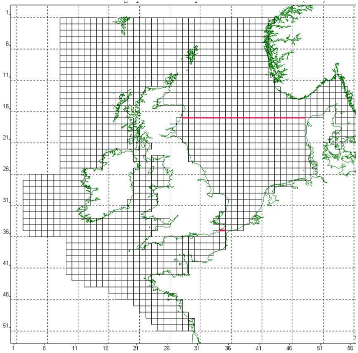
Modelling of surge process over space for joint flood risk assessment for coastal sites and for offshore sites needed for insurance industry

Two Applications

- **Sea-surge data**
Modelling of surge process over space for joint flood risk assessment for coastal sites and for offshore sites needed for insurance industry
- **River flow data**
Modelling of river flow for network for joint flood risk assessment for planning purposes and insurance

Surge Data

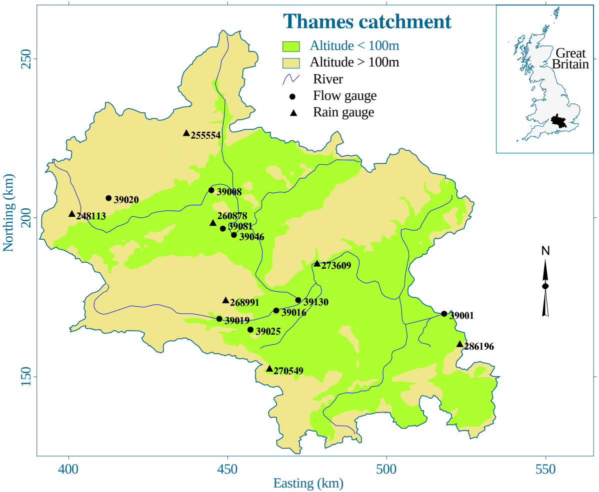
Hindcast output from the CSX model, a 2d numerical surge model for the European Continental Shelf forced by DNMI pressure data for the period 1955-2000



Data are: hourly maxima over 5-day blocks for 46 years at 259 sites

River Flow Data

Daily river flows for a network of sites in River Thames catchment in UK



Marginal Standardisation and Notation

X: univariate variable of most interest
$\mathbf{Y}$: d-dimensional variable

Transform marginals to Gumbel distributions

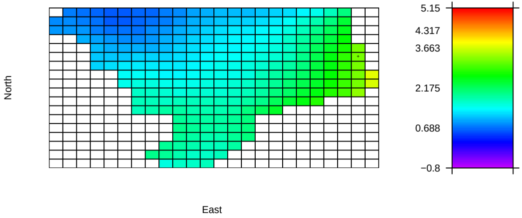
$$\Pr(X > x) = \Pr(Y_i > x) \sim \exp(-x) \text{ as } x \to \infty \text{ for } i = 1, \ldots, d$$

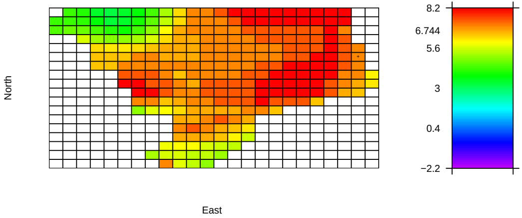
Lack of Memory Property

$$\Pr(X > t + x) \sim \exp(-t) \Pr(X > x) \text{ for large } x$$

Allows focus on dependence structure

A large surge event on the Danish coast in original and transformed margins





What is the Aim of Analysis?

- **Sea-surge data**

Simulation of surge events large at a given location
Estimation of spatial risk measure

$$E(\#\{\mathbf{Y} > x\} \mid X > x)$$

Dimension reduction for physical understanding

What is the Aim of Analysis?

- **Sea-surge data**

Simulation of surge events large at a given location

Estimation of spatial risk measure

$$E(\#\{\mathbf{Y} > x\} \mid X > x)$$

Dimension reduction for physical understanding

- **River flow data**

Estimation of $\Pr(\mathbf{Y} > x \mid X > x)$

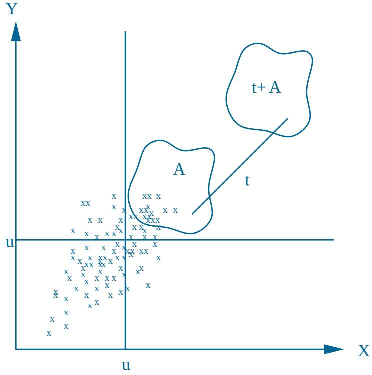
Schematic of Threshold Approach

Under assumption of asymptotic dependence

$$\lim_{x \to \infty} \Pr(Y > x \mid X > x) > 0$$

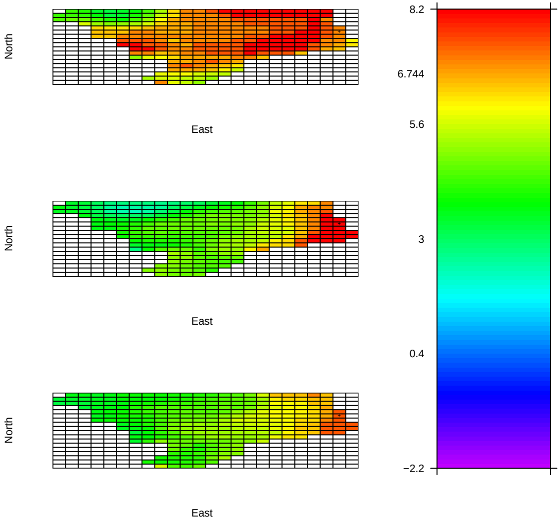
the following homogeneity property holds for all sets A extreme in at least one variable

$$\Pr((X, \mathbf{Y}) \in t + A) \approx \exp(-t) \Pr((X, \mathbf{Y}) \in A)$$



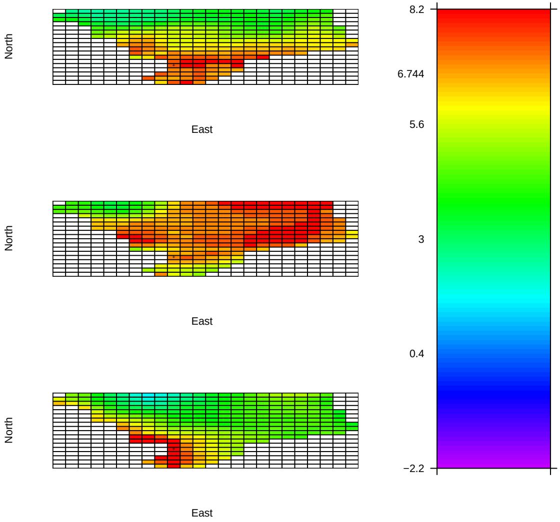
Is Surge Process Asymptotically Dependent?

X: Danish Site



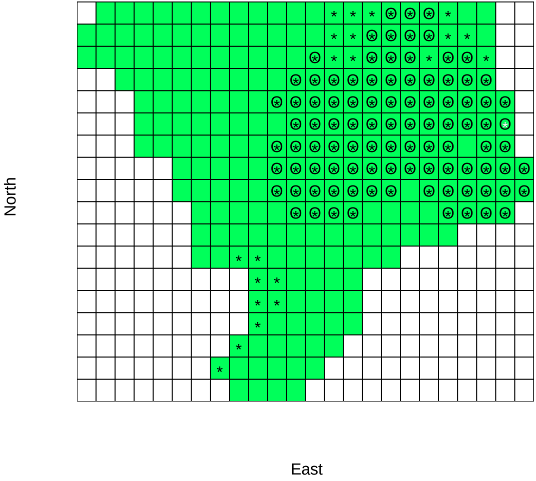
Is Surge Process Asymptotically Dependent?

X: UK Site



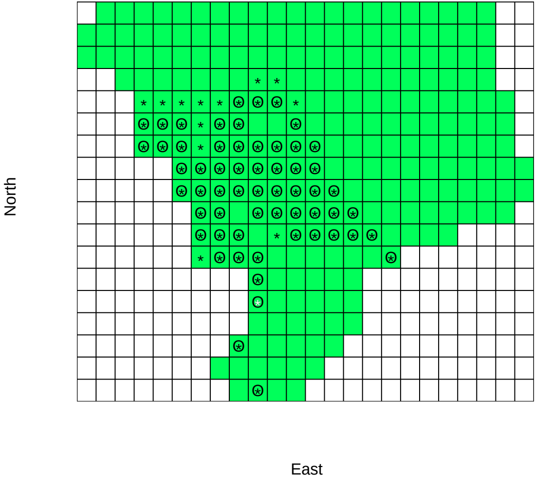
Sites Significant on Testing for Asymptotic Dependence

X: Danish Site



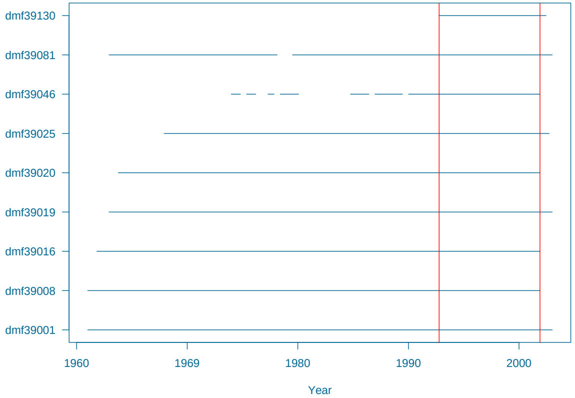
Sites Significant on Testing for Asymptotic Dependence

X: UK Site



Problems for River Flow Application

Plot of data availability for Thames catchment sites



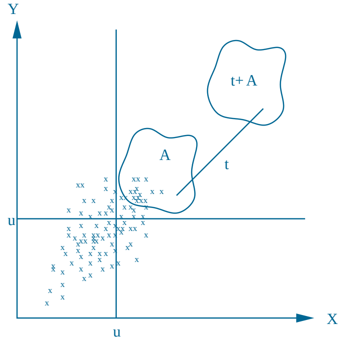
Regression Interpretation of Threshold Method

For $X > u$

$$\mathbf{Y} = X + \mathbf{Z}$$

where $\mathbf{Z}$ is independent of X

$$\hat{\Pr}((X, \mathbf{Y}) \in t + A) = \exp(-v) \int_v^\infty \frac{1}{m} \sum_{i=1}^m 1_{\{(x, x+\mathbf{z}_i) \in t+A\}} \exp(-x) dx$$



Extension of Regression/Conditional Method

Heffernan and Tawn (2004,JRSS B)

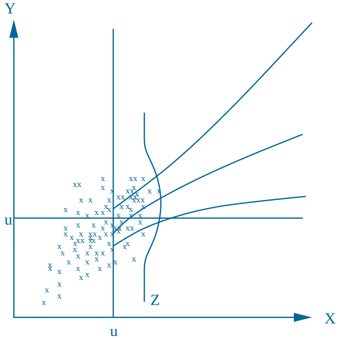
For $X > u$

$$\mathbf{Y} = \mathbf{a}X + X^{\mathbf{b}}\mathbf{Z}$$

where $\mathbf{Z}$ **is independent of** X

d**-dimensional parameters** $\mathbf{0} \leq \mathbf{a} \leq \mathbf{1}$ **and** $\mathbf{b}$

Nonparametric model for $\mathbf{Z}$



Theoretical Examples

$$\mathbf{Y} = \mathbf{a}X + X^{\mathbf{b}}\mathbf{Z}$$

Asymptotic Dependence

$$\mathbf{a} = \mathbf{1} \text{ and } \mathbf{b} = \mathbf{0}$$

Asymptotic Independence with Y_j

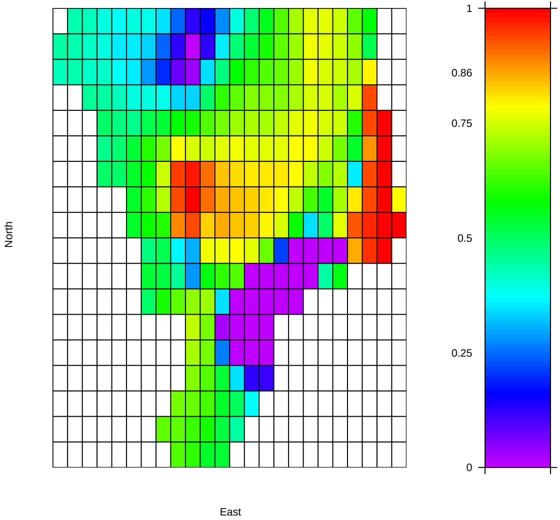
$$a_j < 1$$

Multivariate Normal Copula

$$a_j = \rho_j^2 \text{ and } b_j = \frac{1}{2} \text{ for } j = 1, \ldots, d$$

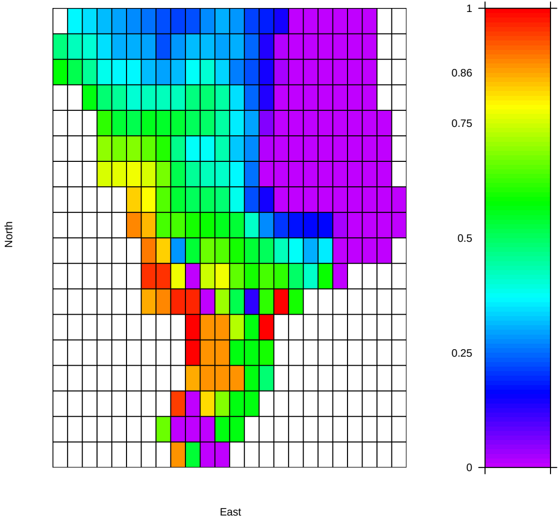
Estimates of a

X: Danish Site



Estimates of a

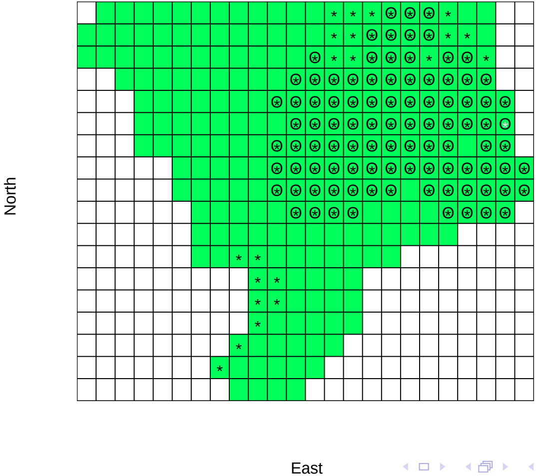
X: UK Site



Which Sites are Asymptotically Dependent?

Test $a_j = 1, b_j = 0$

X: Danish Site



Search for Parsimonious Model

Dimension of model parameters currently $259 \times 258 \times 2$

Dimension Reduction helpful/insightful

Search for Parsimonious Model

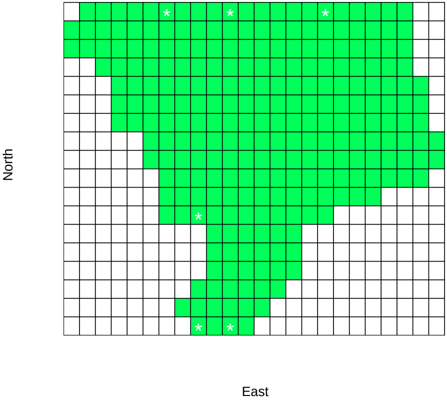
Dimension of model parameters currently $259 \times 258 \times 2$

- Dimension Reduction helpful/insightful
- How many sites do we need to condition on to get all sites asymptotically dependent on a conditioning site?

Search for Parsimonious Model

Dimension of model parameters currently $259 \times 258 \times 2$

- **Dimension Reduction helpful/insightful**
- **How many sites do we need to condition on to get all sites asymptotically dependent on a conditioning site?**



Parsimonious Spatial Model

Partition $(X, \mathbf{Y}) = (\mathbf{X}_C, \mathbf{Y}_C)$ where
$\mathbf{X}_C$ the six conditioning sites
$\mathbf{Y}_C$ the remaining sites

Then

$$[\mathbf{X}_C, \mathbf{Y}_C] = [\mathbf{X}_C][\mathbf{Y}_C \mid \mathbf{X}_C]$$

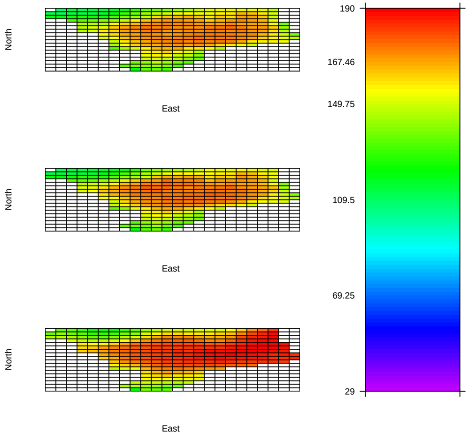
where $[\mathbf{X}_C]$ is low dimensional, and
$[\mathbf{Y}_C \mid \mathbf{X}_C]$ is simpler due to asymptotic dependence property

Extremes for $[\mathbf{Y}_C]$ only arise when $[\mathbf{X}_C]$ is extreme in at least only component

Spatial Risk Measure

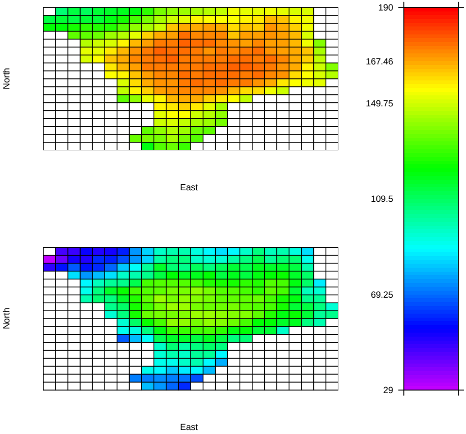
$E(\#\{\mathbf{Y} > x\} \mid X > x)$ **where x is the 97% quantile**

Comparison of empirical, global model, parsimonious model



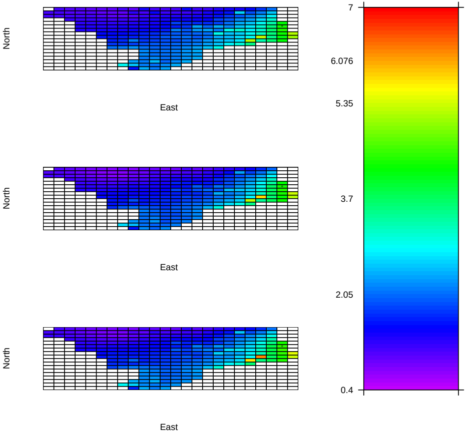
Extrapolation of Spatial Risk Measure

$E(\#\{\mathbf{Y} > x\} \mid X > x)$ **where x is the 97% and 99.9% quantiles for global model**



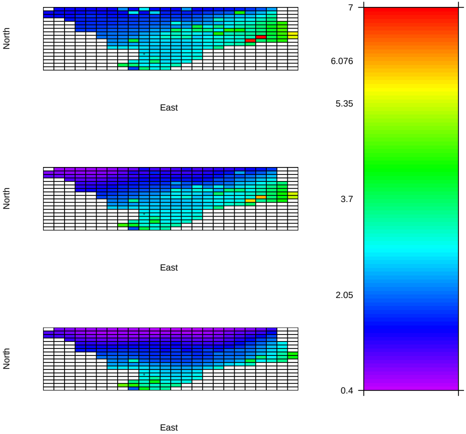
Simulated Fields on Original Scale

Exceeds 1000 year level on Danish coast site



Simulated Fields on Original Scale

Exceeds 1000 year level on UK coast site



Handling Missing Data for River Flows

Partition $\mathbf{Y} = (\mathbf{Y}_M, \mathbf{Y}_O)$ where $\mathbf{Y}_M$ missing; $\mathbf{Y}_O$ observed
Also $\mathbf{Z} = (\mathbf{Z}_M, \mathbf{Z}_O)$

Then need to model $[\mathbf{Z}_M \mid \mathbf{Z}_O]$

Approach is:

Handling Missing Data for River Flows

Partition $\mathbf{Y} = (\mathbf{Y}_M, \mathbf{Y}_O)$ where $\mathbf{Y}_M$ missing; $\mathbf{Y}_O$ observed
Also $\mathbf{Z} = (\mathbf{Z}_M, \mathbf{Z}_O)$

Then need to model $[\mathbf{Z}_M \mid \mathbf{Z}_O]$

Approach is:

- Transform margins

$$\mathbf{Z}^N = T(\mathbf{Z}) = \Phi^{-1}(\hat{F}(\mathbf{Z}))$$

Handling Missing Data for River Flows

Partition $\mathbf{Y} = (\mathbf{Y}_M, \mathbf{Y}_O)$ where $\mathbf{Y}_M$ missing; $\mathbf{Y}_O$ observed
Also $\mathbf{Z} = (\mathbf{Z}_M, \mathbf{Z}_O)$

Then need to model $[\mathbf{Z}_M \mid \mathbf{Z}_O]$

Approach is:

- Transform margins

$$\mathbf{Z}^N = T(\mathbf{Z}) = \Phi^{-1}(\hat{F}(\mathbf{Z}))$$

- Model dependence by MVN copula

$$\begin{pmatrix} \mathbf{Z}_M^N \\ \mathbf{Z}_O^N \end{pmatrix} \sim \mathbf{MVN}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

Handling Missing Data for River Flows

Partition $\mathbf{Y} = (\mathbf{Y}_M, \mathbf{Y}_O)$ where $\mathbf{Y}_M$ missing; $\mathbf{Y}_O$ observed
Also $\mathbf{Z} = (\mathbf{Z}_M, \mathbf{Z}_O)$

Then need to model $[\mathbf{Z}_M \mid \mathbf{Z}_O]$

Approach is:

- Transform margins

$$\mathbf{Z}^N = T(\mathbf{Z}) = \Phi^{-1}(\hat{F}(\mathbf{Z}))$$

- Model dependence by MVN copula

$$\begin{pmatrix} \mathbf{Z}_M^N \\ \mathbf{Z}_O^N \end{pmatrix} \sim \mathbf{MVN}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

- Take a sample from this conditional distribution $[\hat{\mathbf{Z}}_M^N \mid \mathbf{Z}_O^N]$

Handling Missing Data for River Flows

Partition $\mathbf{Y} = (\mathbf{Y}_M, \mathbf{Y}_O)$ where $\mathbf{Y}_M$ missing; $\mathbf{Y}_O$ observed
Also $\mathbf{Z} = (\mathbf{Z}_M, \mathbf{Z}_O)$

Then need to model $[\mathbf{Z}_M \mid \mathbf{Z}_O]$

Approach is:

- Transform margins

$$\mathbf{Z}^N = T(\mathbf{Z}) = \Phi^{-1}(\hat{F}(\mathbf{Z}))$$

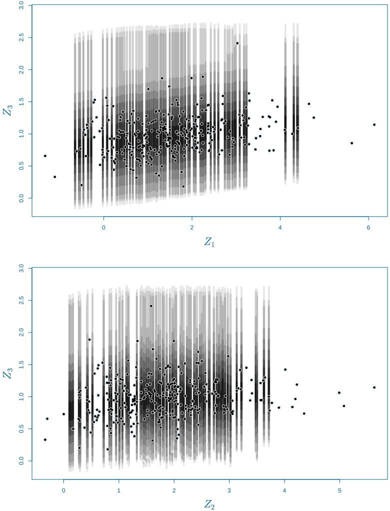
- Model dependence by MVN copula

$$\begin{pmatrix} \mathbf{Z}_M^N \\ \mathbf{Z}_O^N \end{pmatrix} \sim \mathbf{MVN}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

- Take a sample from this conditional distribution $[\hat{\mathbf{Z}}_M^N \mid \mathbf{Z}_O^N]$
- Back transform sample and downweight values in sample $\hat{\mathbf{Z}}_M = T^{-1}(\hat{\mathbf{Z}}_M^N)$

Example of Handling Missing Data

Joint distribution model for $\mathbf{Z} = (Z_1, Z_2, Z_3)$ with infilled sample to replace missing Z_3 values

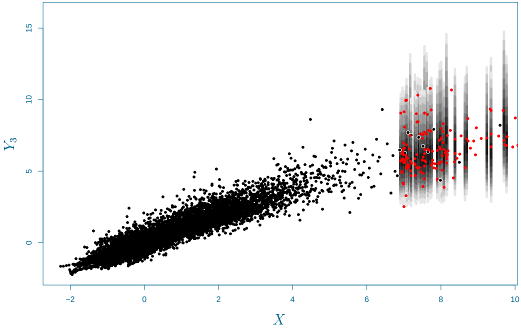


Extrapolation with Missing Data

Recall conditional model is for $X > u$

$$\mathbf{Y} = \mathbf{a}X + X^{\mathbf{b}}\mathbf{Z}$$

Extrapolation: simulate $X > v$ and independently simulate $\mathbf{Z}$ then join as above to give Y



Simulation Study to Assess Infill Method

Consider 3 different patterns of missingness with

$$X : \textbf{Full data};\ Y_1 : \textbf{50\%};\ Y_2 : \textbf{90\%};\ Y_3 : \textbf{80\%};$$

9 true distributions of $\mathbf{Z}$

Methods:
Use overlapping data only $\star$
Infill method $\circ$

Compare Estimators of:

$$P_i = \Pr(Y_i > x \mid X > x) \textbf{ for } i = 1, 2, 3$$

by RMSE efficiency relative to the Full Data case

Efficiency Results for Handling Missing Data

Results for P_1, P_2, P_3 respectively

The infill method does well!