Extreme Value Mixture Modelling: `evmix` Package and Simulation Study

2013 Joint NZSA+ORSNZ Conference

Carl Scarrott & Yang Hu

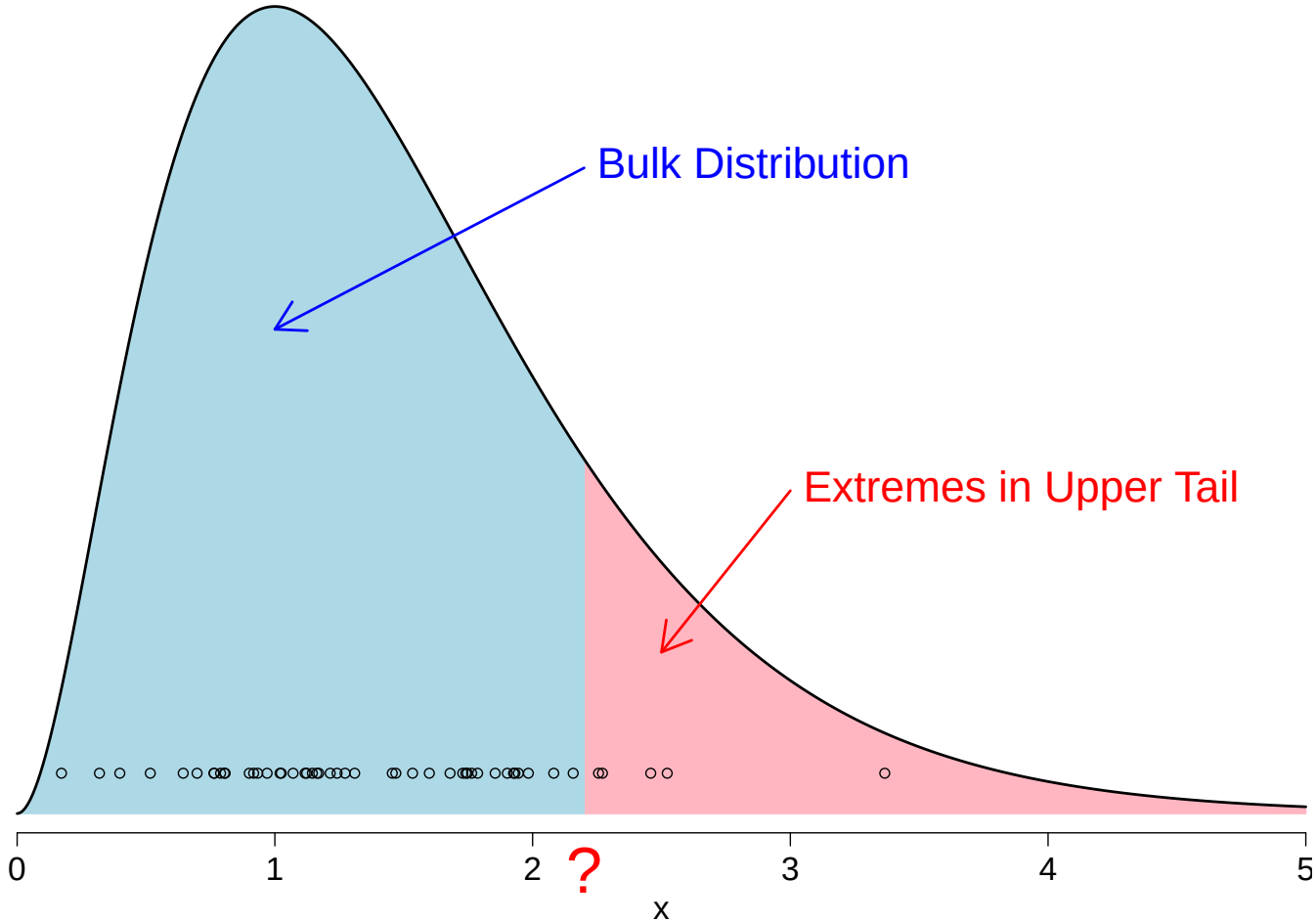
University of Canterbury

25-27 Nov 2013

Talk Outline

- Quick Intro to Extremes
- Outline Extreme Value Mixture Models
- New `evmix` package on CRAN
- Simulation Study
- Some Closing Advice

Quick Intro to Extremes



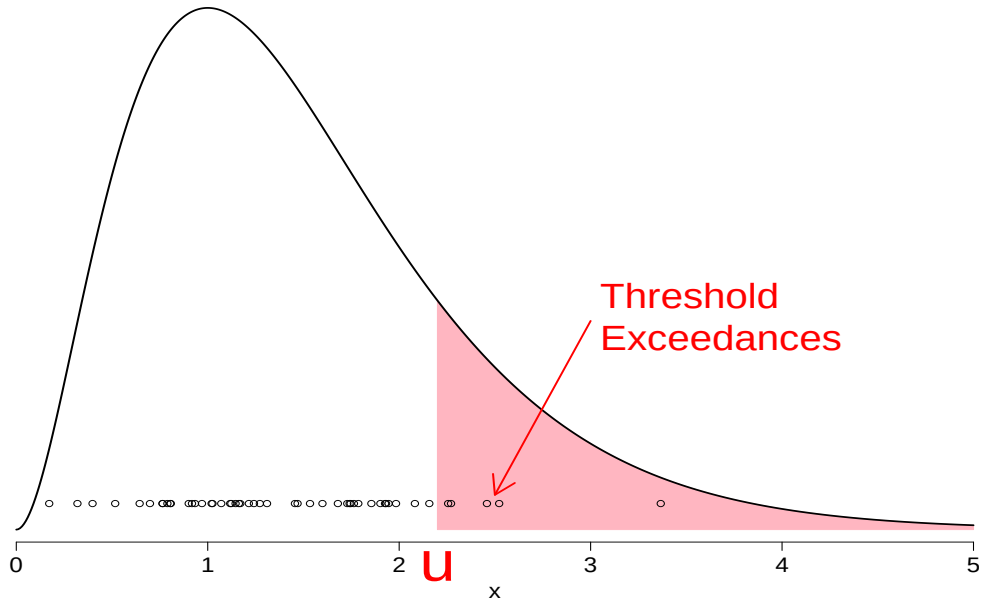
- Typically, upper (or lower tail) of distribution
- Intrinsically about extrapolation
- Limited information from data, supplement by asymptotically justified models
- Bayesian inference can also be beneficial

Extreme Value Threshold Model

- Asymptotically motivated model for excesses above threshold: generalised Pareto distribution (GPD)

$$P(X > x|X > u) = 1 - \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi}$$

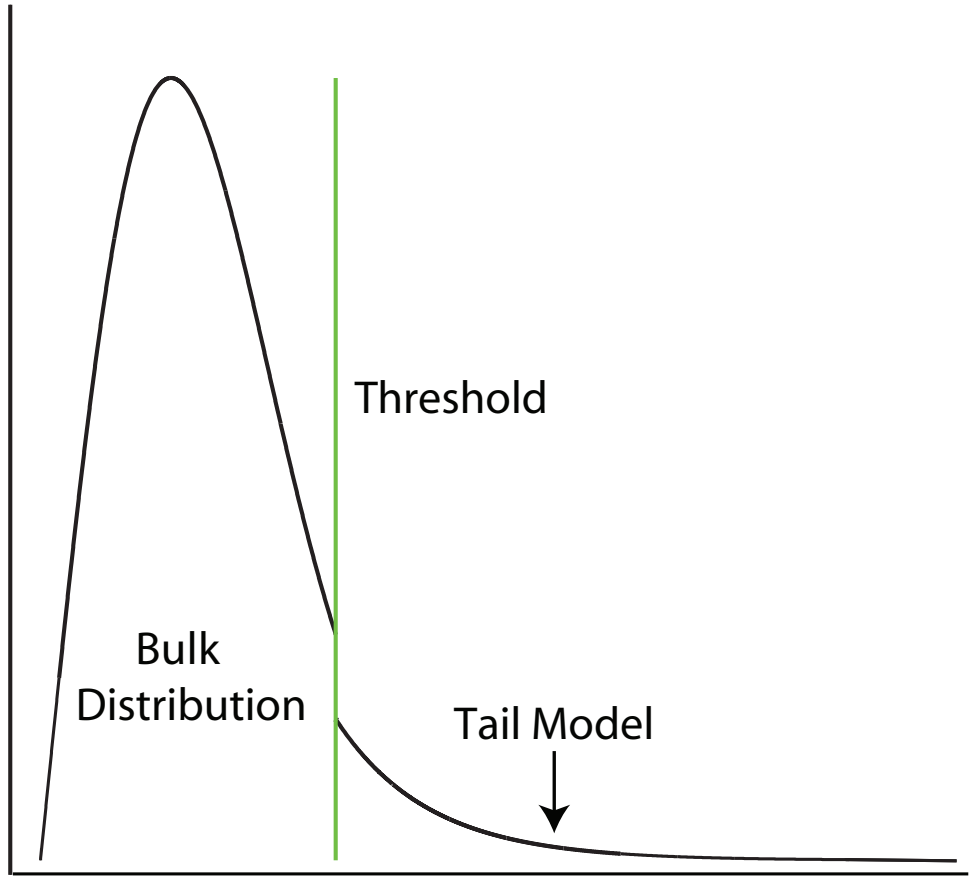
- Scale $\sigma_u > 0$ and shape ξ parameters
- Shape determines tail behaviour:
  - $\xi = 0$ - exponential tail
  - $\xi > 0$ - heavier tail
  - $\xi < 0$ - short tail (upper end-point: $u - \frac{\sigma_u}{\xi}$)



- Implicit parameter: tail fraction above threshold $\phi_u = P(X > u)$:

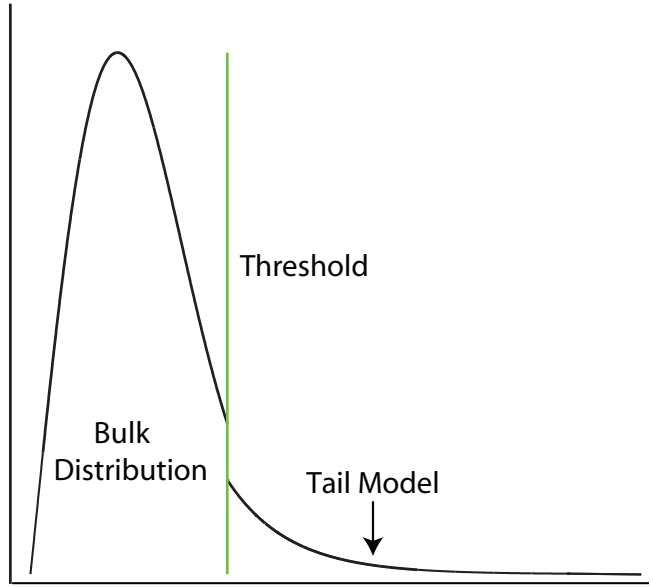
$$P(X > x) = \phi_u P(X > x|X > u)$$

Why Use Extreme Value Mixture Models?



- Treat threshold as parameter to be estimated
- Provide automated and objective "threshold" estimation
- Or avoid threshold choice together
- Allow for threshold uncertainty to be taken into account
- **Key issue: sensitivity of tail fit to that of bulk**

Some Terminology



- Tail model typically generalised Pareto distribution (GPD)
- Bulk model has many forms, "loosely" categorised:
  - **parametric**: normal, Weibull, gamma, log-normal, beta
  - **semi-parametric**: mixtures of gamma, normal, log-normal
  - **nonparametric**: mixture of uniforms, kernel density estimation, smoothing splines
- Most of these (and few related extreme value mixture models) implemented in `evmix` package

Tail Fraction Specification

- How should tail fraction $P(X > u)$ be specified?
  1. proportion of bulk model above threshold $\phi_u = 1 - H(u)$, where $H(.)$ is cdf of bulk model
  2. extra parameter $\phi_u = P(X > u)$
- First approach most common, but no theoretical justification
- Second approach consistent with classical GPD modelling (note that it requires bulk model to be renormalised to $1 - \phi_u$)
- Which is better? Sensitivity of tail estimates to bulk model specification?

`evmix` Package in `R`

- Two key suite of overlapping tools:
  - extreme value threshold estimation and uncertainty quantification, including mixture models; and
  - univariate kernel density estimation
- Named after `evd` package as similar syntax for basic GPD and threshold diagnostic plots
- **Kernel density estimation functionality in `R` extended to boundary corrected kernel density estimators**, where support is bounded (above, below or both!)

- Current version 0.2.0
- Designed using readable, native `R` code, so totally open (some speed penalties to achieve this)
- Available on `CRAN`
- Any feedback and bug reports welcome!

`evmix` Syntax

- Functions follow usual naming conventions, e.g. for gamma bulk with GPD for tail:
  - `dgammagpd` - density function
  - `pgammagpd` - cumulative distribution function
  - `qammagpd` - quantile function
  - `rgammagpd` - random number generation

  - `fgammagpd` - maximum likelihood estimation
  - `lgammagpd` - (log-)likelihood function
  - `nlgammagpd` - negative log-likelihood function
- Fitting function provides sensible initial values for parameters for numerical optimisation routines
- `evmix.diag` function provides usual four model fit diagnostics for all mixture models:
  - return level plot;
  - QQ and PP plots
  - density plots
- `tcplot` and `mrlplot` provide threshold stability plots and mean residual life plots respectively

Example Usage Code 1

- Example of fitting two variants of normal bulk with GPD tail:

```
library(evmix)
set.seed(0)
x = rnorm(1000)

# Fit normal bulk model with GPD for upper tail
fit = fnormgpd(x)

# plot fit over sample density histogram
xx = seq(-5, 5, 0.01)
hist(x, breaks = 100, freq = FALSE)
with(fit, lines(xx, dnormgpd(xx, nmean, nsd, u, sigmau, xi), col="blue"))
abline(v = fit$u, col="blue")

# Add constraint of continuous density at threshold
fitcon = fnormgpdcon(x)

with(fitcon, lines(xx, dnormgpdcon(xx, nmean, nsd, u, xi), col="red"))
abline(v = fitcon$u, col="red")
```

Example Usage Code 2

- Nonparametric KDE's uses cross-validation likelihood so much slower, sit back and take a sip of coffee!:

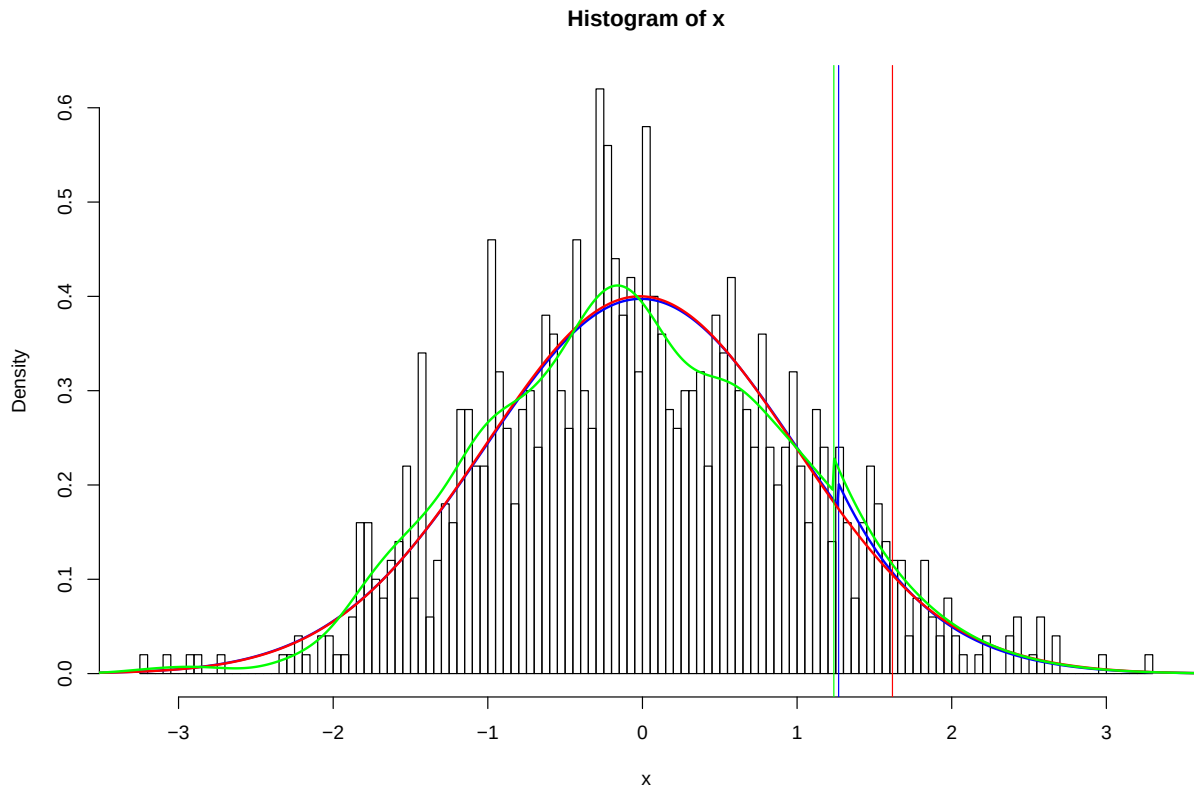
```
# Nonparametric bulk fit
fitkde = fkdengpd(x)

with(fitkde, lines(xx, dkdengpd(xx, x, lambda, u, sigmau, xi), col="green"))
abline(v = fitkde$u, col="green")
```

- Code available on package website:

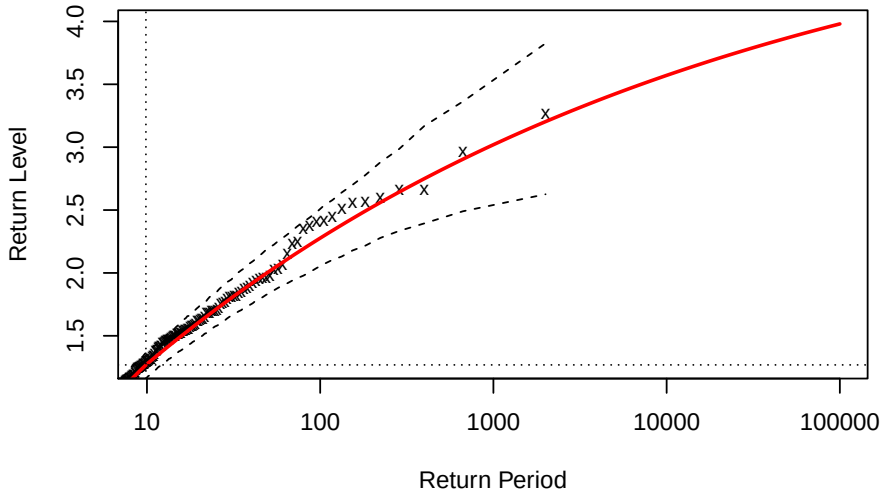
  `http://www.math.canterbury.ac.nz/~c.scarrott/evmix`

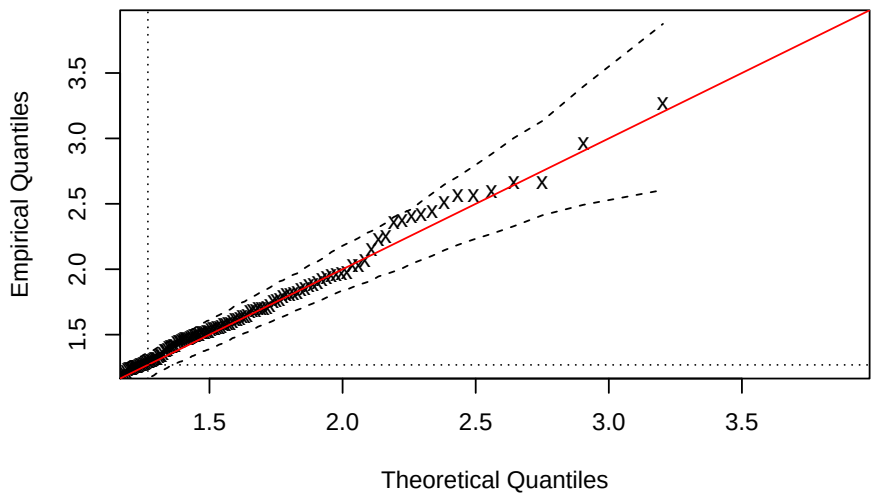
Example Usage Results

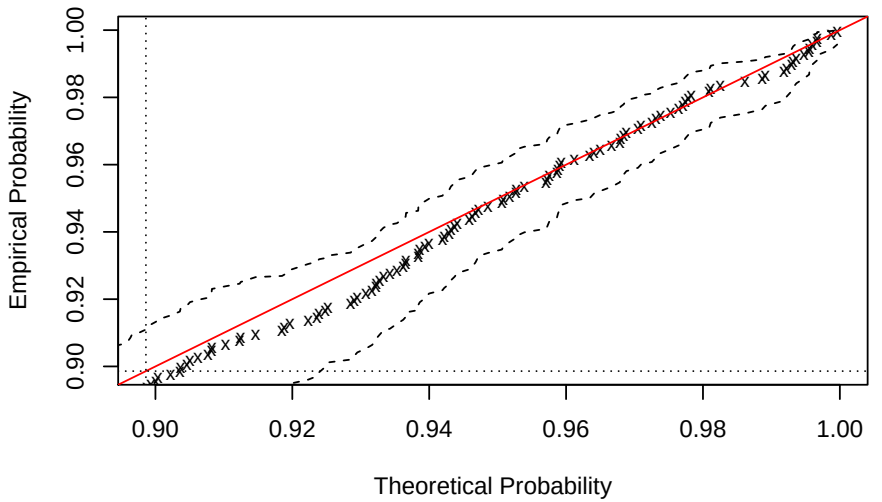


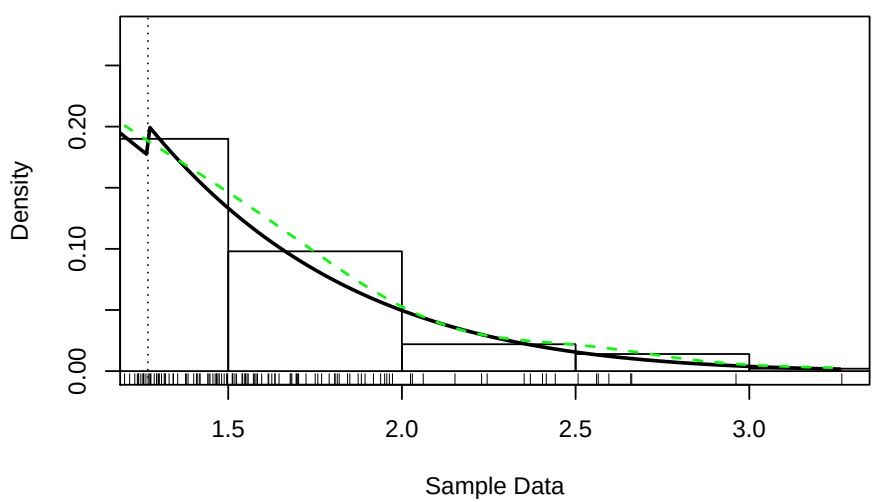
Model Fit Diagnostics

```
# Usual model diagnostics default to focus on upper tail
evmix.diag(fit)
```



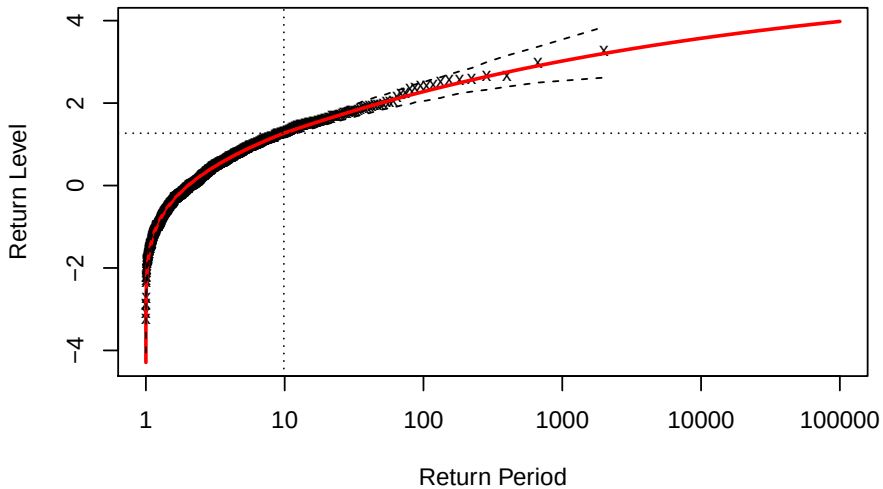


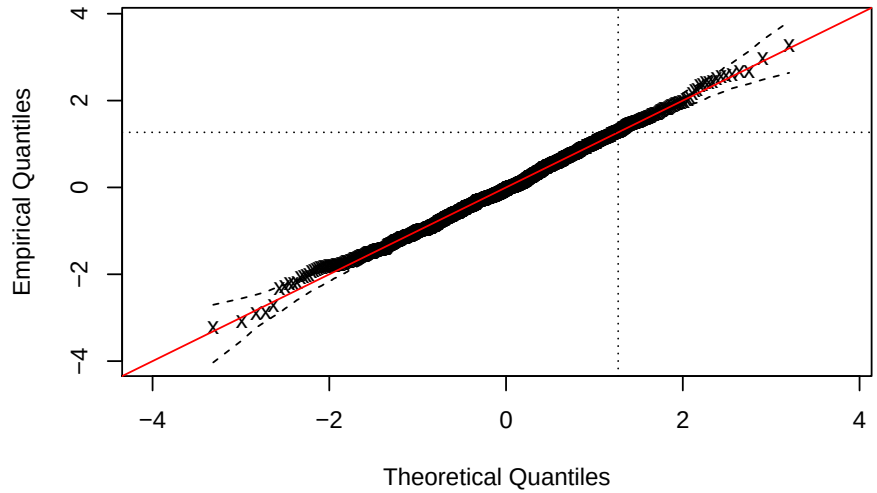


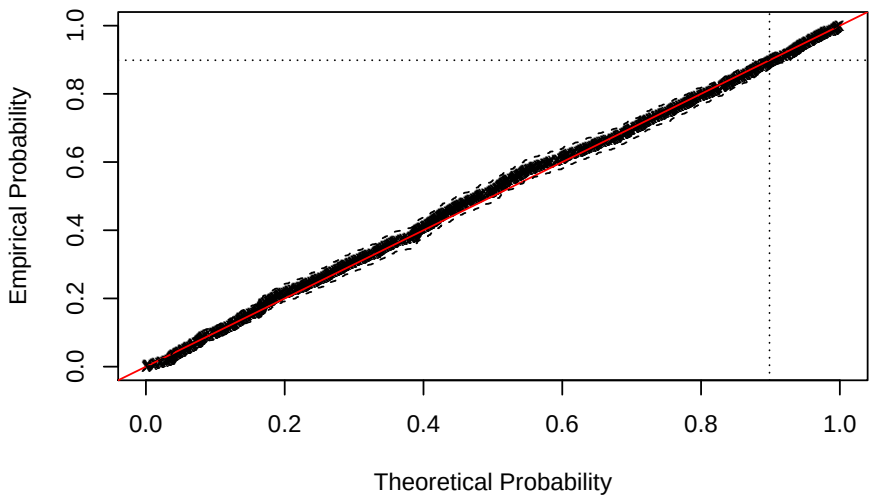


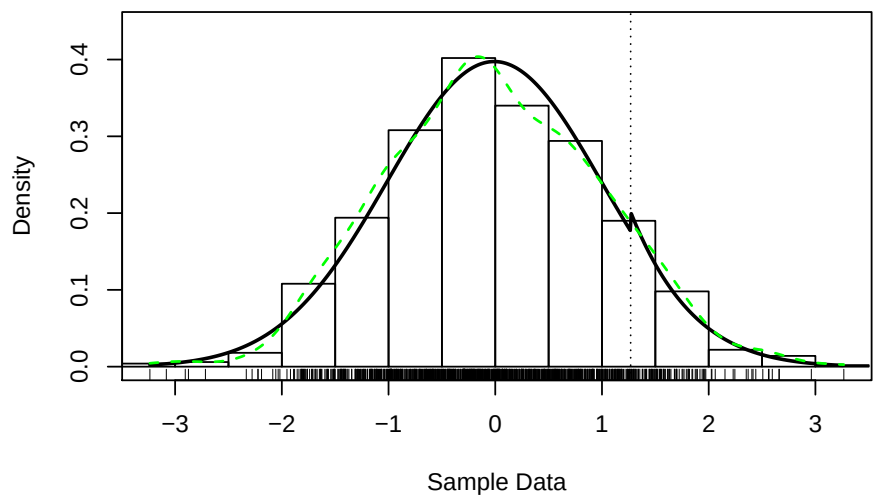
Model Fit Diagnostics

```
evmix.diag(fit, upperfocus=FALSE)
```





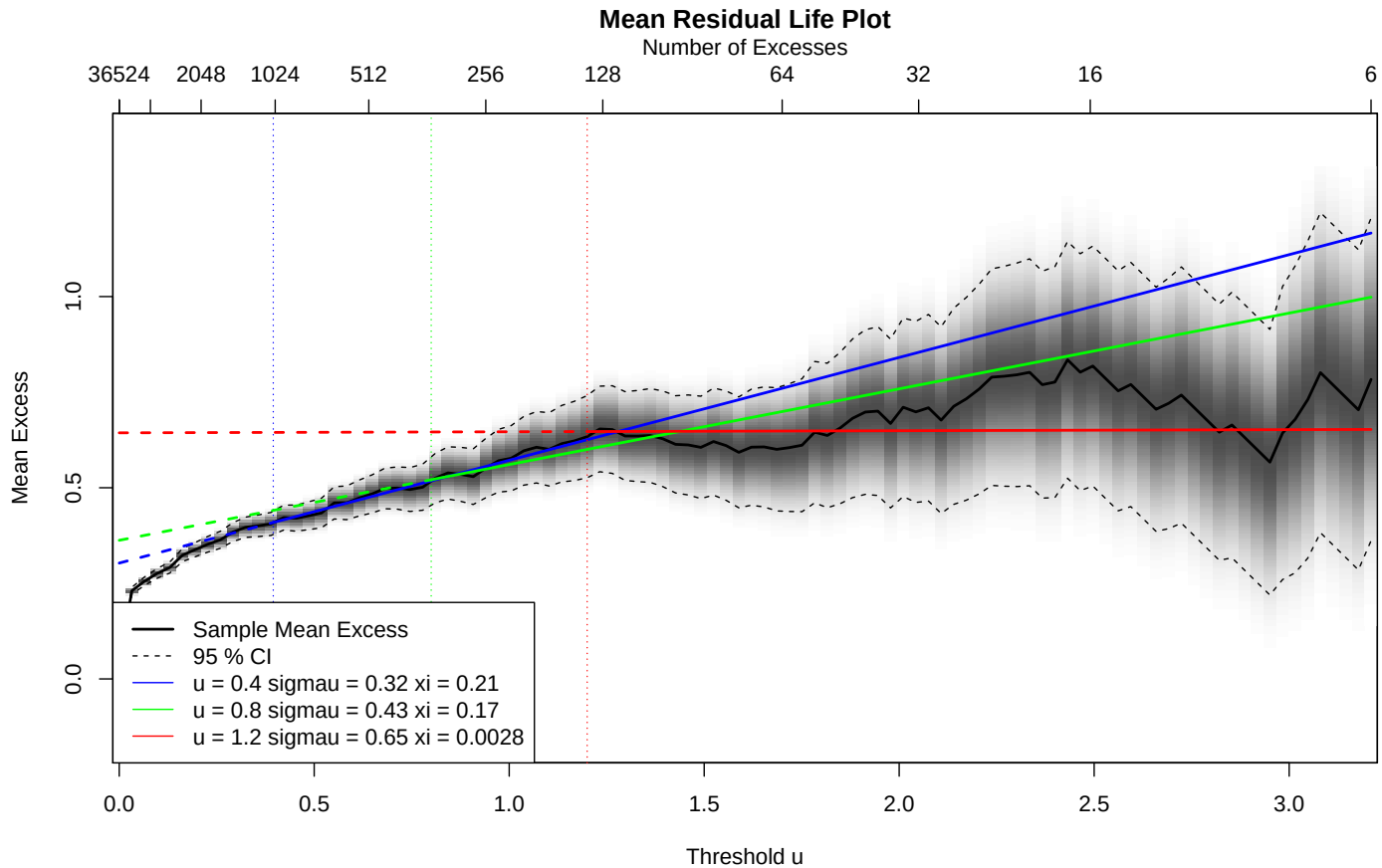




Threshold Choice Plots

- Mean residual life plot is commonly used diagnostic
- Upto sample variation it is linear above a suitably high threshold, for which the GPD is a good approximation

```
# Usual MRL plot with some extra features
data(FtCoPrec,package="extRemes")
mrlplot(FtCoPrec[,5], try.thresh=c(0.395, 0.8, 1.2))
```



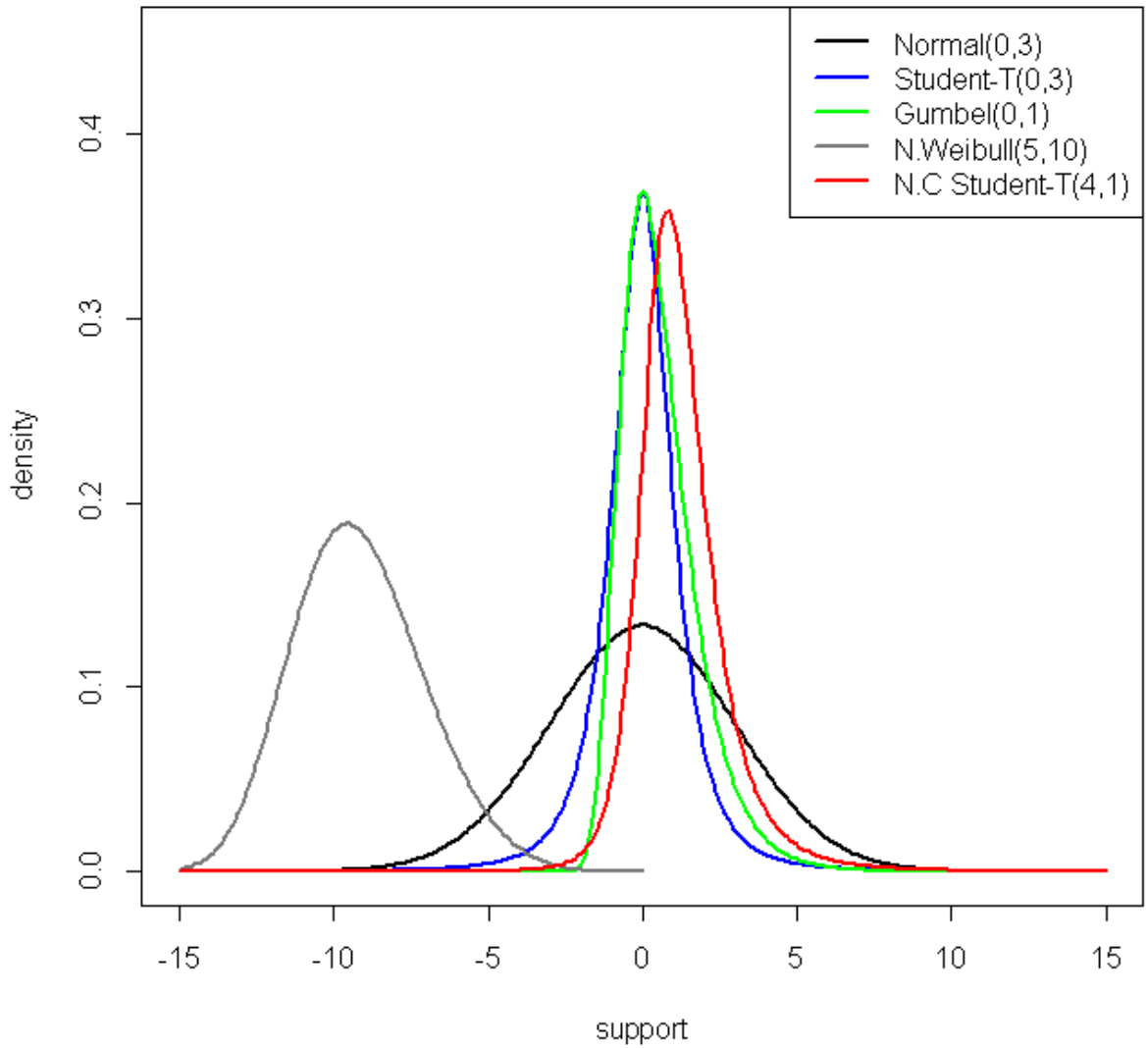
Simulation Study

- Compare all relevant extreme value mixture models depending on range of support:
  - entire reals (one tailed and two-tailed models)
  - positive/non-negative support

- Aim at answering following questions:
  1. In which situations is it best to use the bulk model versus parameter for tail fraction?
  2. In which situation is it best to use parametric, semi or nonparametric mixture models?
  3. In which situation is it best to have constraint of continuous density at the threshold?

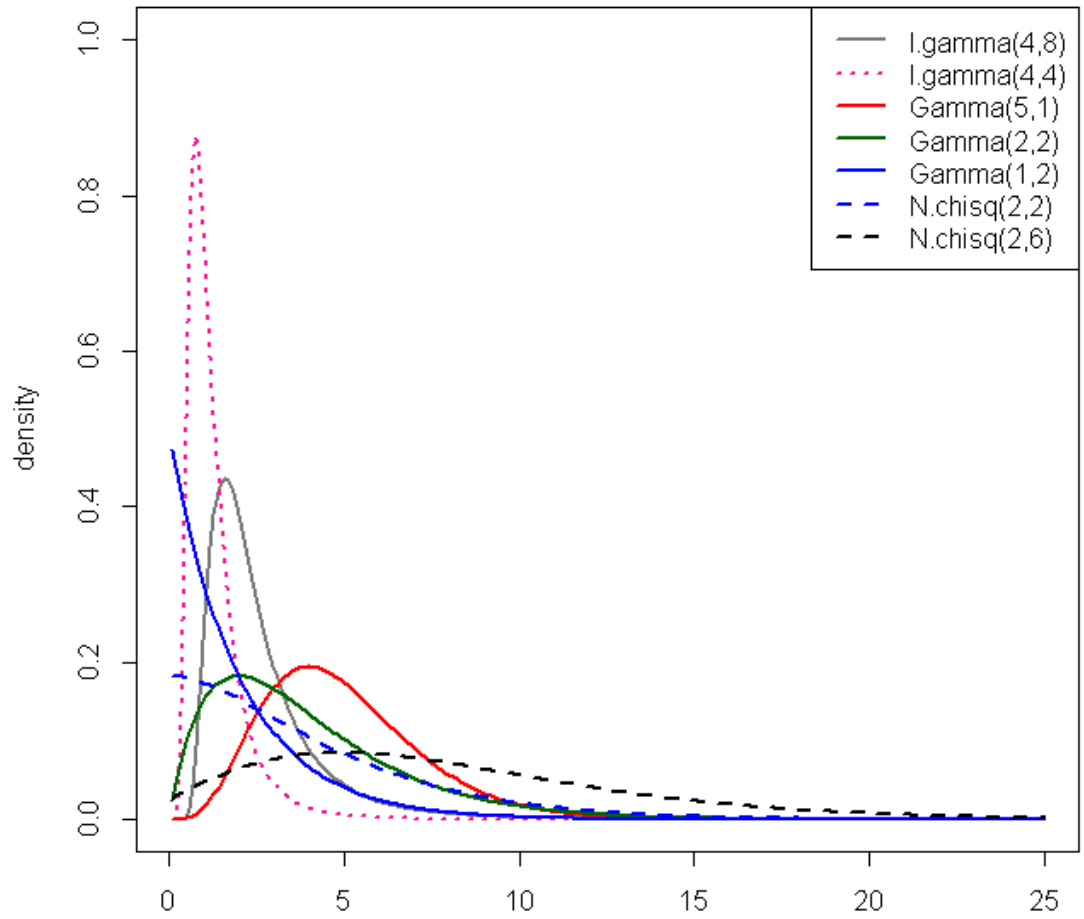
Simulation Setup

- 100 simulations of sample sizes 1,000 and 5,000
- Variety of different bulk and tail behaviour combinations



Simulation Setup

- Fit using maximum likelihood estimation (and for some models MCMC for posterior sampling in Bayesian inference also considered)
- Estimate high quantiles (90, 95, 99, 99.9%)
- Compare performance using RMSE



Advice for Practitioners and Future Mixture Models

1. If the bulk model is correct, then should use it to calculate tail fraction $P(X > u)$, as **borrowing information from bulk for tail inference**. A "small" advantage if density constrained to be continuous at threshold
2. If bulk model is mis-specified (i.e. unknown population), then **better to use extra parameterise tail fraction as ϕ_u which robustifies tail fit to that of bulk**. BUT(!), little to be gained by the continuity constraint at the threshold and can reduce robustness so should be avoided
3. If the bulk model is correctly specified, then the parametric mixture models are easy to understand, quick to fit and have lowest RMSE so are preferred
4. However, in more usual situation of unknown population distribution, the nonparametric mixture models perform consistently well for low and high quantiles

Advice for Practitioners and Future Mixture Models

1. Little difference between "sensibly chosen" mixture models for highest quantiles (e.g. 99.9%)
2. Substantial variation between models for lower quantile (90, 95, 99%)
3. Poorest performing mixture model, by far, was hybrid Pareto (Carreau and Bengio, 2009) which is due to it completely ignoring the tail fraction scaling of GPD
4. The dynamically weighted mixture model also had variable performance
5. Note: limitation on results so far - no penalty for complexity, profile likelihood approach for threshold estimation (implement in `evmix`) since been shown to be beneficial

References and Website

Review paper:

Scarrott and MacDonald (2012). A review of extreme value threshold estimation and uncertainty quantification. REVSTAT Statistical Journal 10(1), 33-60.
(all references in here)

Package: `evmix` available on CRAN (all feedback appreciated)

Website:
`http://www.math.canterbury.ac.nz/~c.scarrott/evmix`

Yang Hu's thesis with all simulation results on website

Thanks for your attention...