Теория экстремальных значений

Количественная аналитика

Часть 1. Одномерный случай

Распределение максимумов потерь

$\vec{x} = (x_1; \dots; x_n)$ — убытки, $\vec{x} \sim iid. F(x)$

$M_n = \max(x_1; \dots; x_n)$ — максима

$F_{M_n}(x) = P(M_n \leq x) = F^n(x)$ — распределение максимы

Пусть нормализованные максимумы сходятся к некоторому распределению $H(x)$, это означает, что

$\exists d_n, c_n > 0$: $\lim\limits_n P\left(\frac{M_n - d_n}{c_n} \leq x\right) = \lim\limits_n F^n(c_n x + d_n) = H(x)$,

тогда $F \in MDA(H)$

Generalized Extreme Value distribution (GEV)

Теорема Фишера-Типпетта-Гнеденко

Если $F \in MDA(H)$ и H не сосредоточено в одной точке, то $H \sim GEV(\mu(c_n, d_n); \sigma(c_n, d_n); \xi)$

$$GEV(0;1;\xi):\ H_\xi(x) = \begin{cases} e^{-(1+\xi x)^{-\frac{1}{\xi}}}, \xi \neq 0 \\ e^{-e^{-x}}, \xi = 0 \end{cases}, \text{ где } 1 + \xi x > 0$$

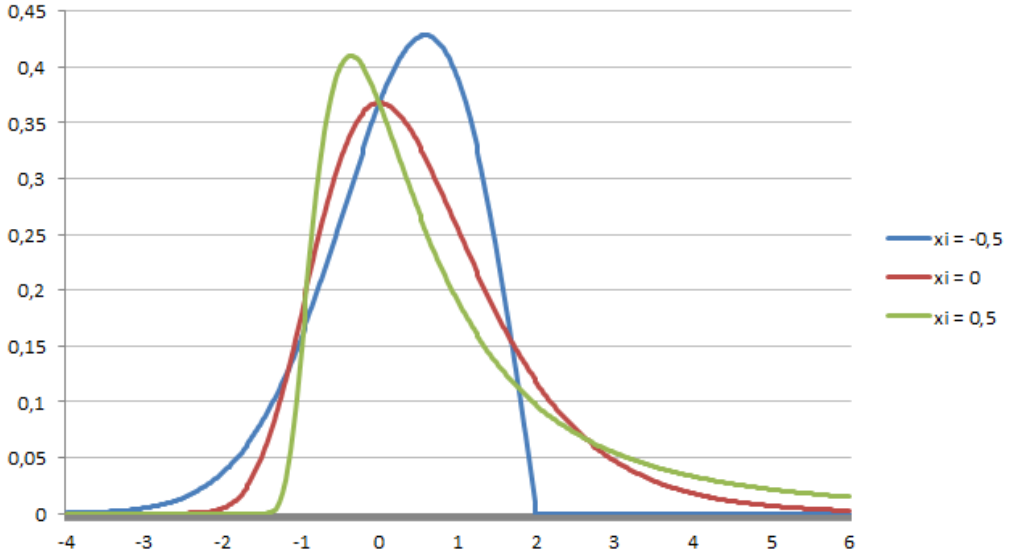
Частные случаи GEV:

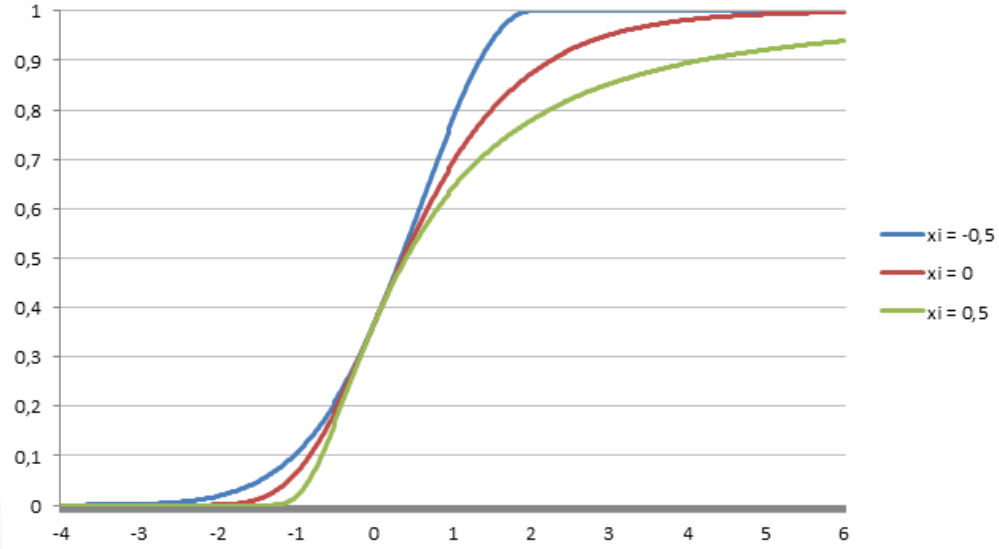
- $\xi > 0$ — распределение Фрешѐ
- $\xi = 0$ — распределение Гумбеля
- $\xi < 0$ — распределение Вейбулла

Распределение Вейбулла имеет конечную правую точку $x_F = \sup\{x \in R: F(x) < 1\}$

Фреше и Гумбель не имеют конечных правых точек, но Фреше убывает значительно медленнее

Функции распределения и плотности GEV





Максима временного ряда

Пусть $(x_1, \dots, x_n) \sim F$ — стационарный временной ряд,

$(\bar{x}_1, \dots, \bar{x}_n) \sim WN(F)$ — соответствующий ему белый шум

Пусть $\bar{M}_n = \max(\bar{x}_1, \dots, \bar{x}_n)$, тогда

$\exists \theta \in (0; 1]: \lim\limits_n P\left(\frac{\bar{M}_n - d_n}{c_n} \le x\right) = H(x) \Leftrightarrow \lim\limits_n P\left(\frac{M_n - d_n}{c_n} \le x\right) = H^\theta(x)$, т.е.

если нормализованные максимы независимых величин сходятся к $H(x)$, то нормализованные максимы временного ряда сходятся к $H^\theta(x)$, причём

$H(x) \sim GEV(\mu; \sigma; \xi) \Longrightarrow H^\theta(x) \sim GEV(\mu(\theta); \sigma(\theta); \xi)$

θ — экстремальный индекс процесса $(x_1, \dots, x_n)$

Содержательная интерпретация экстремального индекса

Пусть $u = c_n x + d_n$, тогда при большом *n* имеем:

$P(M_n \leq u) \approx P^{\theta}(\overline{M}_n \leq u) = F^{n\theta}(u)$, таким образом

распределение максимумов временного ряда длиной *n* может быть аппроксимировано распределением максимумов соответствующего ему белого шума длиной $n\theta < n$

При этом θ интерпретируется как количество относительно независимых кластеров во временном ряде

$\theta = 1 \Rightarrow$ экстремальные значения не кластеризуются,

$\theta < 1 \Rightarrow$ экстремумы имеют тенденцию кластеризоваться

- $\vec{x} \sim WN, ARMA(m; n) \Rightarrow \theta = 1$
- $\vec{x} \sim ARCH(q), GARCH(p; q) \Rightarrow \theta < 1$

Оценка параметров GEV

$\vec{x} = (x_1, \dots, x_T),\ T = m \cdot n$

$M_{n,j} = \max(x_{n(j-1)+1}, \dots, x_{nj})$

$M_n = \left(M_{n,1}, \dots, M_{n,m}\right) \sim GEV(\mu, \sigma, \xi)$

Пусть $h(x; \mu, \sigma, \xi)$ — плотность $GEV(\mu, \sigma, \xi)$, тогда

$$l\left(M_{n,1}, \dots, M_{n,m}; \mu, \sigma, \xi\right) = \sum_{i=1}^{m} \ln h\left(M_{n,i}; \mu, \sigma, \xi\right) = -m \cdot \ln \sigma -$$

$$- \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \ln\left(1 + \xi \cdot \frac{M_{n,i} - \mu}{\sigma}\right) - \sum_{i=1}^{m} \ln\left(1 + \xi \cdot \frac{M_{n,i} - \mu}{\sigma}\right)^{-\frac{1}{\xi}} \rightarrow$$

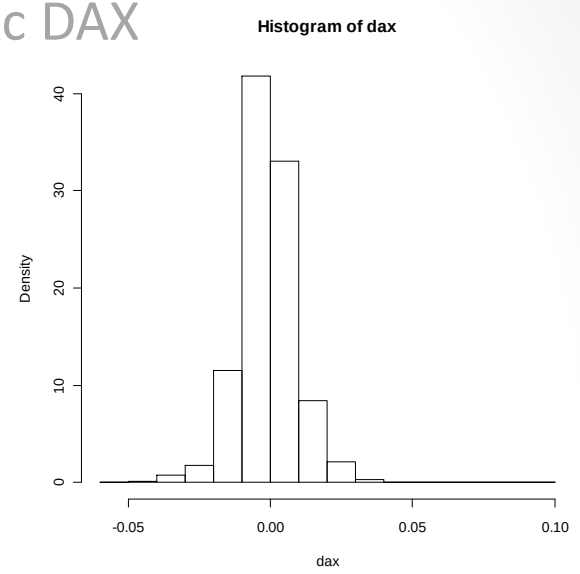
$$\rightarrow \max_{\mu, \sigma > 0, \xi, 1 + \frac{\xi\left(M_{n,i} - \mu\right)}{\sigma} > 0}$$

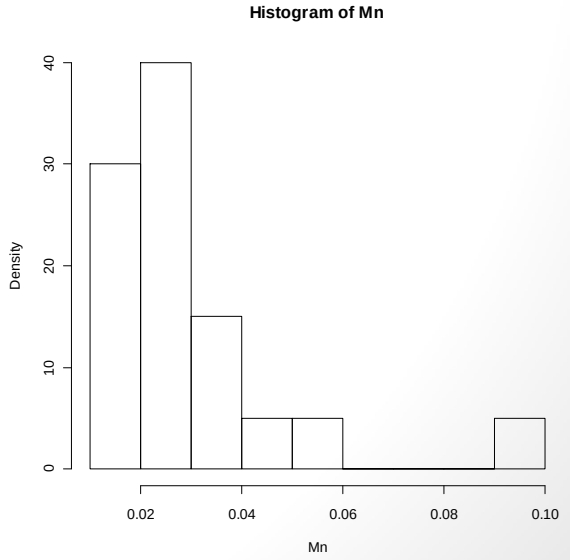
Оценка параметров GEV в R

Практический пример 1. Биржевой индекс DAX

```
# загрузка данных
library(datasets)
dax <- EuStockMarkets[,1]
n <- 90; m <- 20; T <- m*n
dax <- dax[2:(T+1)]/dax[1:T]-1
dax <- -dax

# расчёт максим
Mn <- rep(0,times=m)
for (i in 1:m)
Mn[i] <- max(dax[((i-1)*n+1):(i*n)])
```

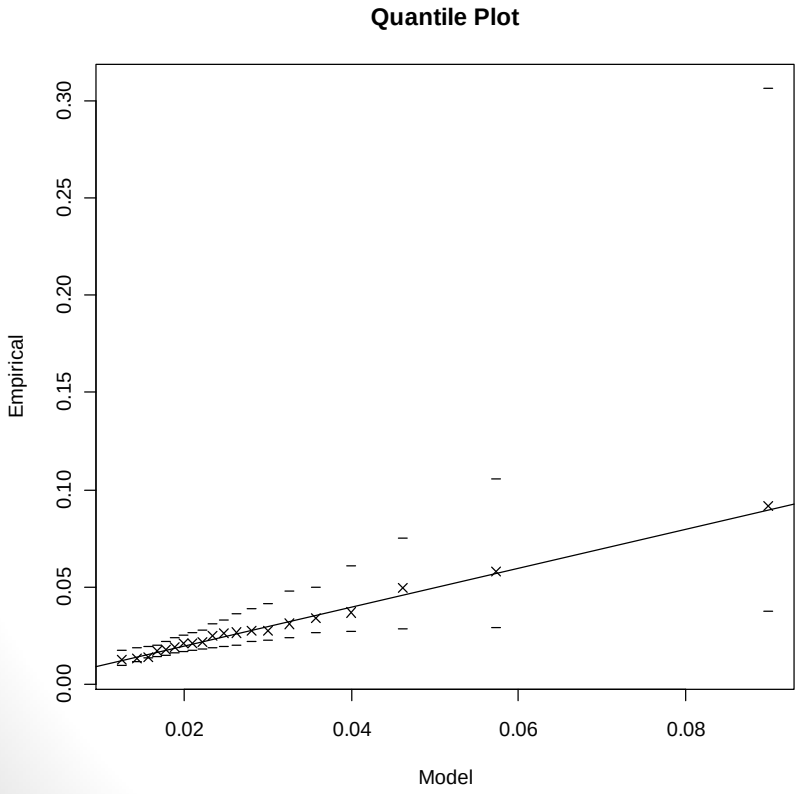


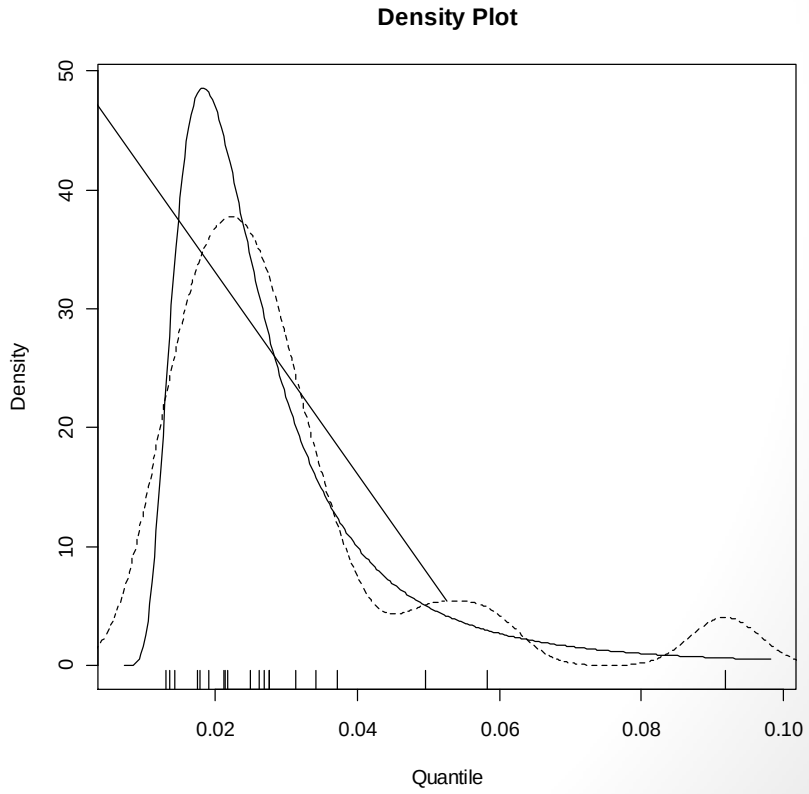


Оценка параметров GEV в R

распределение максим на основе GEV

```
library(evd)
Mn.fit <- fgev(Mn)
plot(Mn.fit)
```





Пороговый уровень и средний период наступления события

$$r_{n,k} = q_{1-\frac{1}{k}}(H) = H^{-1}_{\xi,\mu,\sigma}\left(1-\frac{1}{k}\right) = \mu + \frac{\sigma}{\xi}\left(\left(-\ln\left(1-\frac{1}{k}\right)\right)^{-\xi} - 1\right)$$

уровень, который будет превзойдён в среднем 1 раз за *k* блоков по *n* наблюдений

$k_{n,u} = \frac{1}{1-H(u)}$ — средний период наступления события $M_n > u$

$r_{n,k_{n,u}} = u$

расчёт этих показателей в R

```
mu <- Mn.fit$estimate[1]; sigma <- Mn.fit$estimate[2]
xi <- Mn.fit$estimate[3]; k <- 4; u <- 0.09
r.nk <- mu+sigma/xi*((-log(1-1/k))^(-xi)-1)
k.nr <- 1/(1-pgev(u,loc=mu,scale=sigma,shape=xi))
```

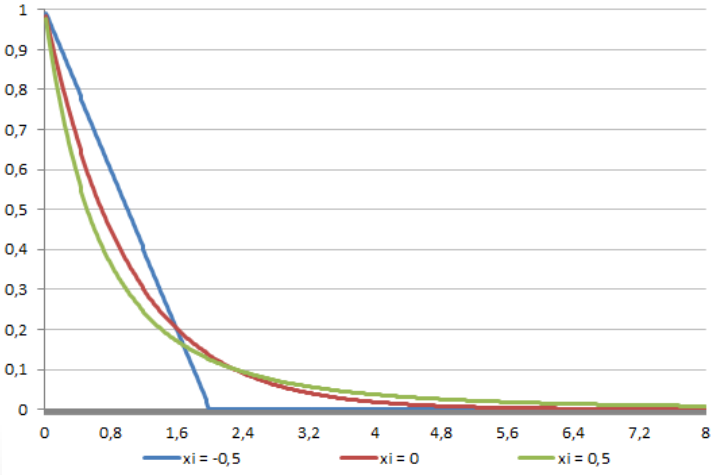
$r_{n,k}$	0.034
$k_{n,u}$	40.14

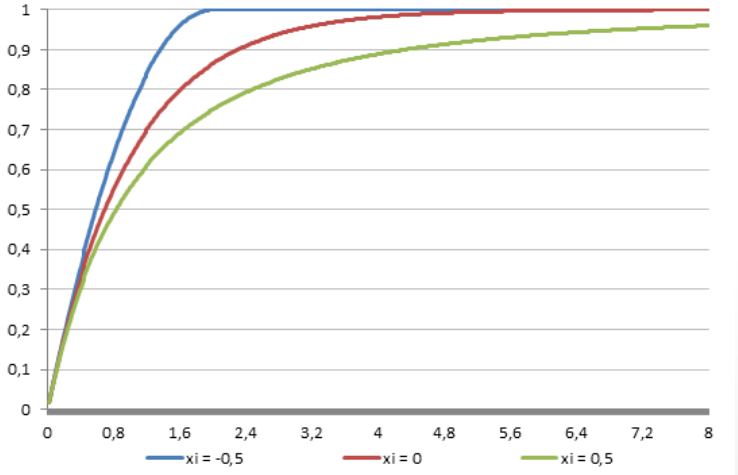
Generalized Pareto distribution (GPD)

$$G(x;\xi,\beta) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - e^{-\frac{x}{\beta}}, & \xi = 0 \end{cases}, \text{ где } \begin{cases} \begin{cases} 0 \leq x \leq -\frac{\beta}{\xi}, & \xi < 0 \\ x \geq 0, & \xi \geq 0 \end{cases} \\ \beta > 0 \end{cases}$$

Частные случаи GPD:

- $\xi > 0$ — распределение Парето
- $\xi = 0$ — экспоненциальное распределение
- $\xi = 0$ — короткохвостое распределение Парето





Превышение порогового значения

Пусть $x_i \sim F$, тогда распределение превышений порога *u* равно

$$F_u(x) = P(x_i - u \le x | x > u) = \frac{F(x+u)-F(u)}{1-F(u)},\ 0 \le x \le x_F - u$$

$e(u) = E(x_i - u | x_i > u)$ — среднее превышение порога

Если $F \equiv G_{\xi,\beta}$, то $F_u(x) \equiv G_{\xi,\beta(u)}(x),\ \beta(u) = \beta + \xi u,$

$e(u) = \frac{\beta(u)}{1-\xi} = \frac{\beta+\xi u}{1-\xi}$, т.е. распределение превышений остаётся GDP с тем же параметром формы ξ, а среднее превышение является линейной функцией относительно порога *u*

Теорема Пикандса-Балкема-де Хаана

$$\exists \beta(u):\ \lim_{u \to x_F} \sup \left| F_u(x) - G_{\xi,\beta(u)}(x) \right| = 0,\ 0 \le x < x_F - u \Leftrightarrow$$

$$F \in MDA(H_\xi),\ \xi \in R,$$

т.о. если распределение максимумов сходится к H_ξ, то превышения для высокого порога *u* описываются GPD

Моделирование хвостов

Пусть $F_u(x) = G_{\xi,\beta}(x),\ 0 \le x < x_F - u, \beta > 0, \xi \in R$, тогда для $x \ge u\ \tilde{F}_u(x) = P(x_i > u) \cdot P(x_i > x | x_i > u) = \bar{F}(u) \cdot P(x_i - u > x - u | x_i > u) = \bar{F}(u) \cdot \bar{F}_u(x-u) = \bar{F}(u) \cdot \left(1 + \frac{\xi(x-u)}{\beta}\right)^{-\frac{1}{\xi}}$ — распределение хвоста доходностей

Используя эту формулу, можно находить квантили убытков:

$$VaR_\alpha = q_\alpha(F) = u + \frac{\beta}{\xi}\left(\left(\frac{1-\alpha}{\bar{F}(u)}\right)^{-\xi} - 1\right),\ \alpha \ge F(u)$$

$ES_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 q_x(F)dx = \frac{VaR_\alpha}{1-\xi} + \frac{\beta - \xi u}{1-\xi}$, $\xi < 1$, также верно, что $ES_\alpha = VaR_\alpha + e(VaR_\alpha)$

Smith (1987): $\hat{\bar{F}}(x) = \frac{N_u}{n}\left(1 + \frac{\hat{\xi}(x-u)}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}},\ x \ge u$

$\alpha \ge 1 - \frac{N_u}{n} \to \widehat{VaR}_\alpha, \widehat{ES}_\alpha$

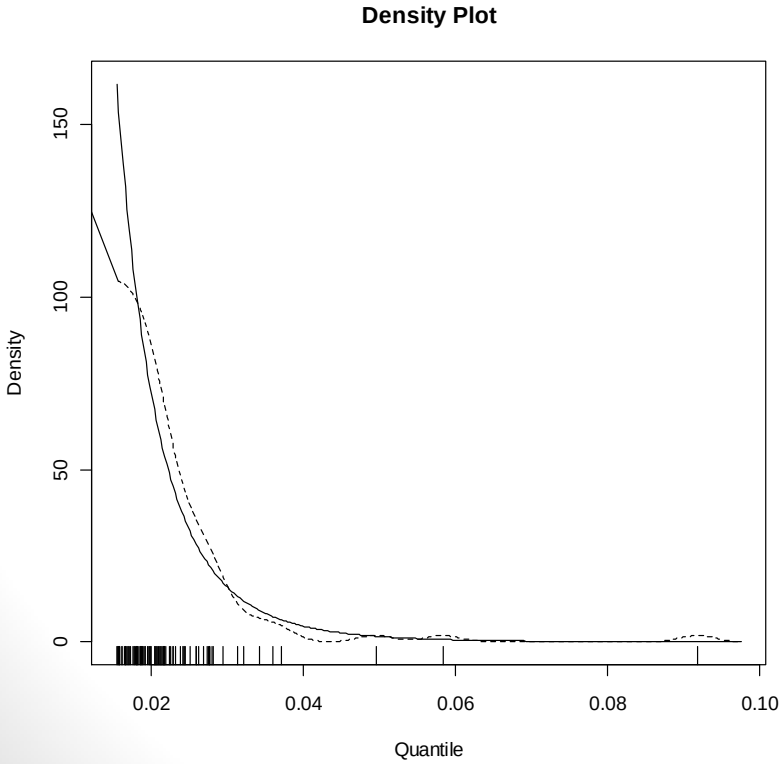
GPD в R

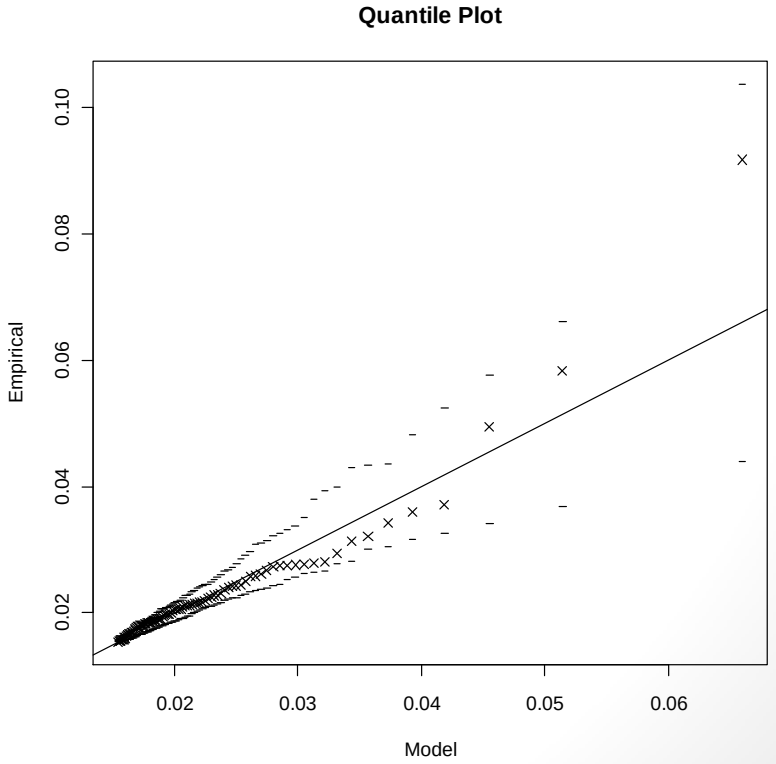
пороговое значение - 95% квантиль

```
u <- sort(dax)[0.95*T]
gpd.fit <- fpot(dax,threshold=u,model="gpd",method="SANN")
```

оценки параметров

```
beta <- gpd.fit$estimate[1]; xi <- gpd.fit$estimate[2]
```





GPD в R

```
# расчёт мер риска
Fu <- gpd.fit$pat
alpha <- 1-1/260      # соответствует k = 4
VaR <- u+beta/xi*(((1-alpha)/Fu)^(-xi)-1)
ES <- (VaR+beta-xi*u)/(1-xi)
```

VaR	0.036
ES	0.048

Часть 2. Многомерный случай

Многомерная максима

Пусть $\vec{x}_1, \dots, \vec{x}_n \sim F,\ \vec{x}_i \in R^d,\ \vec{x}_i \sim F_i$

$\vec{x}_i = \left(x_{i,1}, \dots, x_{i,d}\right)^T$ — убытки различных видов

$M_{n,j} = \max\left(x_{1,j}, \dots, x_{n,j}\right)$

$M_n = \left(M_{n,1}, \dots, M_{n,d}\right)^T$ — покомпонентная блочная максима

Нас интересует сходимость нормализованной максимы:

$\frac{M_n - d_n}{c_n} = \left(\frac{M_{n,1} - d_{n,1}}{c_{n,1}}, \dots, \frac{M_{n,d} - d_{n,d}}{c_{n,d}}\right)^T \xrightarrow{n} H,\ c_n > 0$

Пусть $\frac{M_n - d_n}{c_n}$ сходится к некоторой векторной случайной величине с совместной функцией распределения *H*, тогда

$\lim_n P\left(\frac{M_n - d_n}{c_n} \le \vec{x}\right) = \lim_n F^n(c_n\vec{x} + d_n) = H(\vec{x})$, т.е. $F \in MDA(H)$

Экстремальная копула

Если у *H* есть невырожденные частные функции распределения, то они должны быть Фреше, Гумбеля или Вейбулла. По теореме Шкляра существует копула

$$C(F_1(x_1), \dots, F_d(x_d)) = H(\vec{x})$$

Теорема о копуле эктремальных значений

Пусть $F \in MDA(H)$ и $H_i \sim GEV$, тогда $C(\vec{u}^t) = C^t(\vec{u}),\ \forall t > 0$

Теорема о представлении Пикандса

Копула *C* — экстремальная тогда и только тогда, когда её можно представить в виде

$$C(\vec{u}) = e^{B\left(\frac{\ln u_1}{\sum_{k=1}^{d} \ln u_k}, \dots, \frac{\ln u_d}{\sum_{k=1}^{d} \ln u_k}\right) \sum_{i=1}^{d} \ln u_i}$$, где

$$B(\vec{w}) = \int_{S_d} \max(x_1 w_1, \dots, x_d w_d)\, dH(\vec{x}),$$

$$S_d = \left\{\vec{x} : x_i \geq 0,\ i = 1, \dots, d,\ \sum_{i=1}^{d} x_i = 1\right\}$$

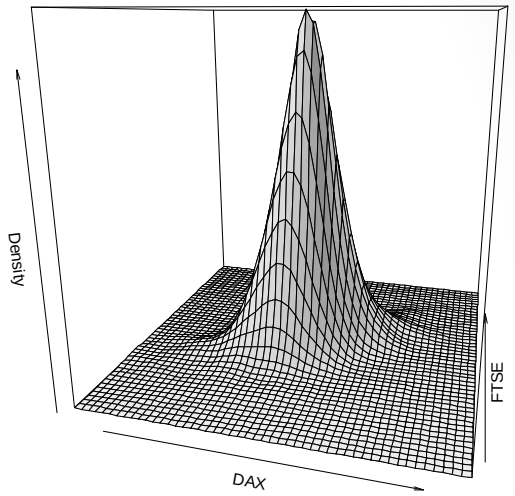
MGEV в R

Практический пример 2. Биржевые индексы DAX и FTSE

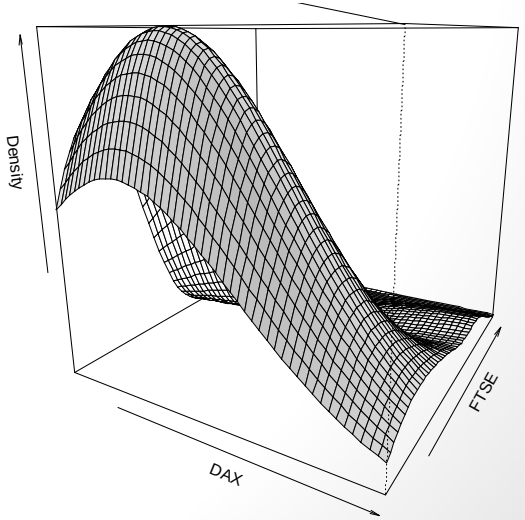
```
# загрузка данных
ftse <- EuStockMarkets[,4]
ftse <- ftse[2:(T+1)]/ftse[1:T]-1
ftse <- - ftse
ESM <- cbind(dax,ftse)

# расчёт максим
Mn <- rep(0,times=m*2)
dim(Mn) <- c(m,2)
for (i in 1:2) {
  for (j in 1:m)
  Mn[j,i] <- max(ESM[((j-1)*n+1):(j*n),i])
}
```

Эмпирическое распределение доходностей DAX и FTSE



Эмпирическое распределение максим



MGEV в R

частные распределения на основе GED

```
fit1 <- fgev(Mn[,1])
fit2 <- fgev(Mn[,2])
```

экстремальные копулы

```
library(copula)
gumb.cop <- gumbelCopula(2)
gal.cop <- galambosCopula(2)
```

значения частных функций распределения

```
cdf1 <- pgev(Mn[,1],loc=fit1$estimate[1],
    scale=fit1$estimate[2],shape=fit1$estimate[3])
cdf2 <- pgev(Mn[,2],loc=fit2$estimate[1],
    scale=fit2$estimate[2],shape=fit2$estimate[3])
cdf <- cbind(cdf1,cdf2)
```

MGEV в R

подгонка копулы

```
gumb.fit <- fitCopula(cdf,copula=gumb.cop)
gal.fit <- fitCopula(cdf,copula=gal.cop)
```

gumb.fit@loglik	5.798
gal.fit@loglik	5.846

модельные значения максим

```
N <- 10^5
cdf.sim <- rcopula(n=N,copula=gal.fit@copula)
sim1 <- qgev(cdf.sim[,1],loc=fit1$estimate[1],
    scale=fit1$estimate[2],shape=fit1$estimate[3])
sim2 <- qgev(cdf.sim[,2],loc=fit2$estimate[1],
    scale=fit2$estimate[2],shape=fit2$estimate[3])
```

MGEV в R

модельные убытки

```
w <- c(0.5,0.5)
loss <- sort(w[1]*sim1+w[2]*sim2)
```

расчёт мер риска

```
k <- 4
alpha <- 1-1/k
VaR <- loss[alpha*N]
ES <- mean(loss[(alpha*N+1):N])
```

VaR	0.029
ES	0.043

Превышение многомерного порога

Пусть $\vec{x}_1, \dots, \vec{x}_n \sim F(\vec{x}) = C(F_1(x_1), \dots, F_d(x_d)) \in MDA(MGEV)$

Согласно теории для одномерного случая частные распределения величин, превышающих многомерный порог $\vec{u} = (u_1, \dots, u_d)$ имеет вид:

$$\tilde{F}_i(x_i) = 1 - \bar{F}_i(u_i)\left(1 + \frac{\xi_i(x_i - \mu_i)}{\beta_i}\right)^{\frac{1}{\xi_i}}, \ \vec{x} \geq \vec{u}$$

Для многомерного случая используется приближение

$$F(\vec{x}) \approx C\left(\tilde{F}_1(x_1), \dots, \tilde{F}_d(x_d)\right), \ \vec{x} \geq \vec{u}$$

Поскольку исходное распределение $F(\vec{x})$ неизвестно, копулу *C(.)* также нужно аппроксимировать

Для этого применяется предельная копула:

$$F(\vec{x}) \approx C_0\left(\tilde{F}_1(x_1), \dots, \tilde{F}_d(x_d)\right), \ \vec{x} \geq \vec{u}$$

Превышение многомерного порога в R

```
# выборка значений, превышающих многомерный порог
u <- c(sort(dax)[0.9*T],sort(ftse)[0.9*T])
t.ESM <- ESM[(ESM[,1]>u[1])&(ESM[,2]>u[2]),]

# частные распределения на основе GED
fit1 <- fpot(t.ESM[,1],threshold=u[1],
    model="gpd",method="SANN")
fit2 <- fpot(t.ESM[,2],threshold=u[2],
    model="gpd",method="SANN")

# значения частных функций распределения
cdf1 <- pgpd(t.ESM[,1],loc=u[1],scale=fit1$par[1],
    shape=fit1$par[2])
cdf2 <- pgpd(t.ESM[,2],loc=u[2],scale=fit2$par[1],
    shape=fit2$par[2])
cdf <- cbind(cdf1,cdf2)
```

Превышение многомерного порога в R

подгонка копулы

```
gumb.fit <- fitCopula(cdf,copula=gumb.cop)
gal.fit <- fitCopula(cdf,copula=gal.cop)
```

gumb.fit@loglik	12.27
gal.fit@loglik	12.69

модельные значения убытков

```
cdf.sim <- rcopula(n=N,copula=gal.fit@copula)
sim1 <- qgpd(cdf.sim[,1],loc=u[1],scale=fit1$par[1],
    shape=fit1$par[2])
sim2 <- qgpd(cdf.sim[,2],loc=u[2],scale=fit2$par[1],
    shape=fit2$par[2])
```

Превышение многомерного порога в R

убытки по портфелю

```
loss <- sort(w[1]*sim1+w[2]*sim2)
```

расчёт мер риска

```
Fu <- nrow(t.ESM)/T
alpha <- 1-1/(260*Fu)
VaR <- loss[alpha*N]
ES <- mean(loss[(alpha*N+1):N])
```

VaR	0.029
ES	0.037

Домашнее задание

В файле «loss_train.csv» находятся данные о возникновении просроченной задолженности по кредитам предприятиям пяти укрупнённых отраслей за периоды $t \in \{1; \dots; 1000\}$

Вашей задачей является оценка трёх таких уровней совокупной просроченной задолженности, которые не будут превышены в 1, 5 и 10 процентах случаев в периоды $t \in \{1001; \dots; 2000\}$

Ответ состоит из двух файлов:

- таблица из трёх чисел
- описание решения