Flexible Regression and Smoothing

Continuous Distributions

Mikis Stasinopoulos, Rob Rigby, Gilllian Heller,
and Fernanda De Bastiani

BCAM – Basque Center for Applied Mathematics, Bilbao, Basque Country - Spain, October 2019

gamlss

gamlss

Types of continuous distributions

$(-\infty, \infty)$: real line $\Re$ [Chapter 4 of Rigby *et al.* (2019)]

$(0, \infty)$: positive real line $\Re^+$ [Chapter 5]

$(0, 1)$ real line on interval $\Re_{(0,1)}$ (not containing zero or one) [Chapter 6]

gamlss

Location and scale family

For these distributions, if

$$Y \sim \mathcal{D}(\mu, \sigma, \nu, \tau)$$

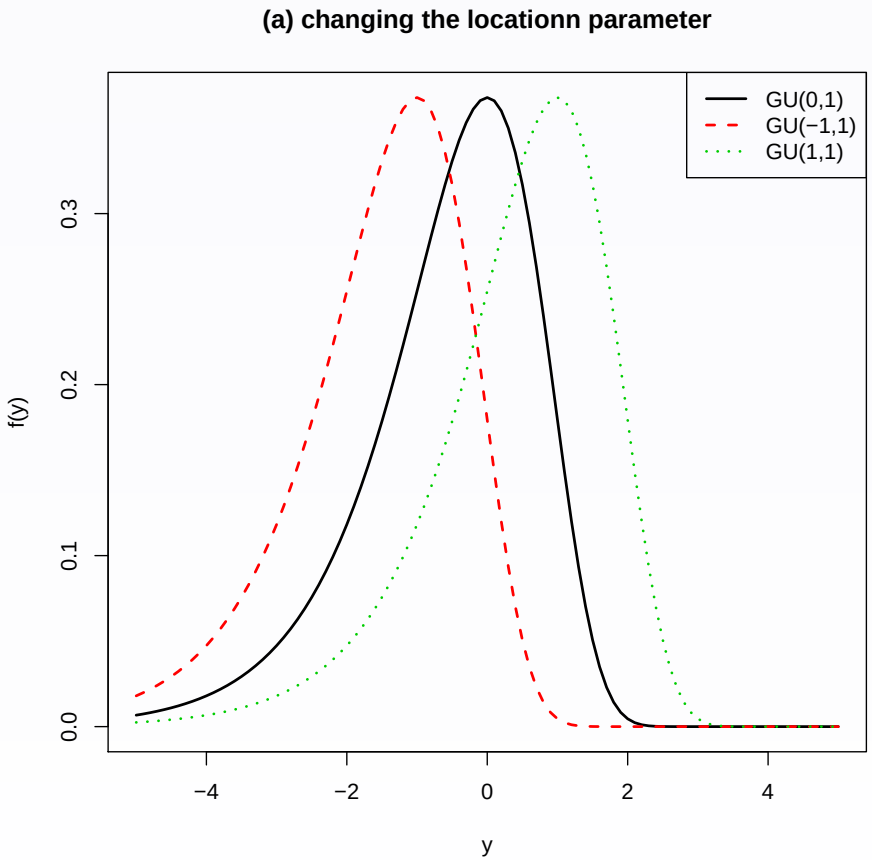
then

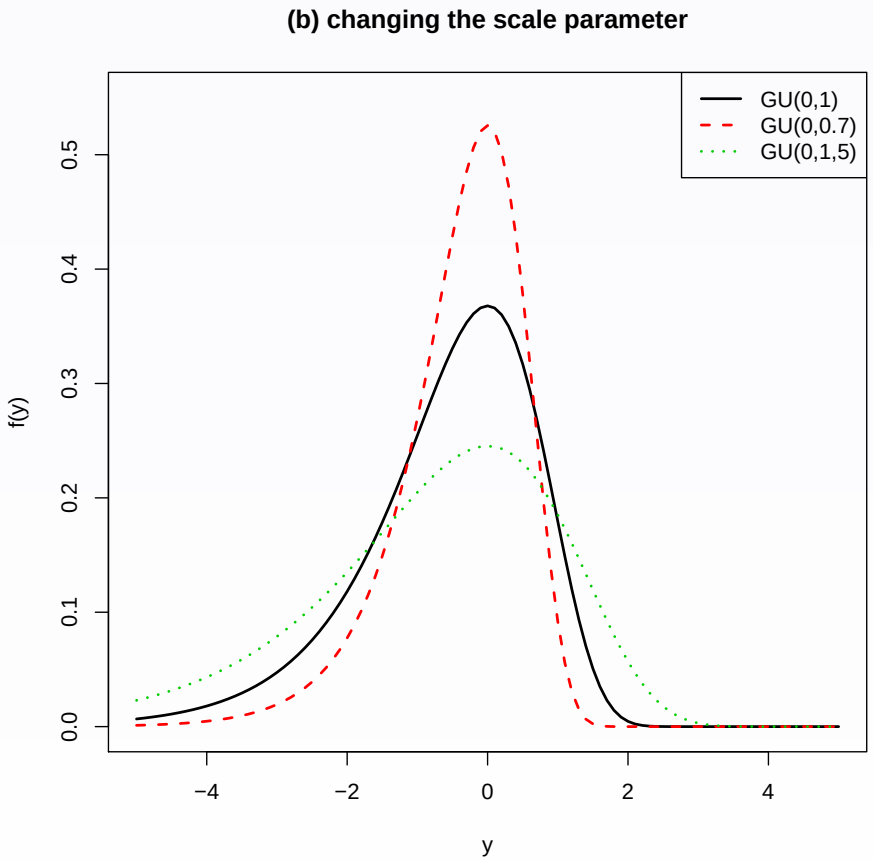
$$\varepsilon = (Y - \mu)/\sigma \sim \mathcal{D}(0, 1, \nu, \tau),$$

i.e. $Y = \mu + \sigma\varepsilon$, so Y is a scaled and shifted version of the random variable ε.
All $(-\infty, \infty)$ distributions in GAMLSS except $\texttt{exGAUS}(\mu, \sigma, \tau)$ are *location-scale* families

gamlss

Location-scale family





gamlss

Two parameter on $\Re$

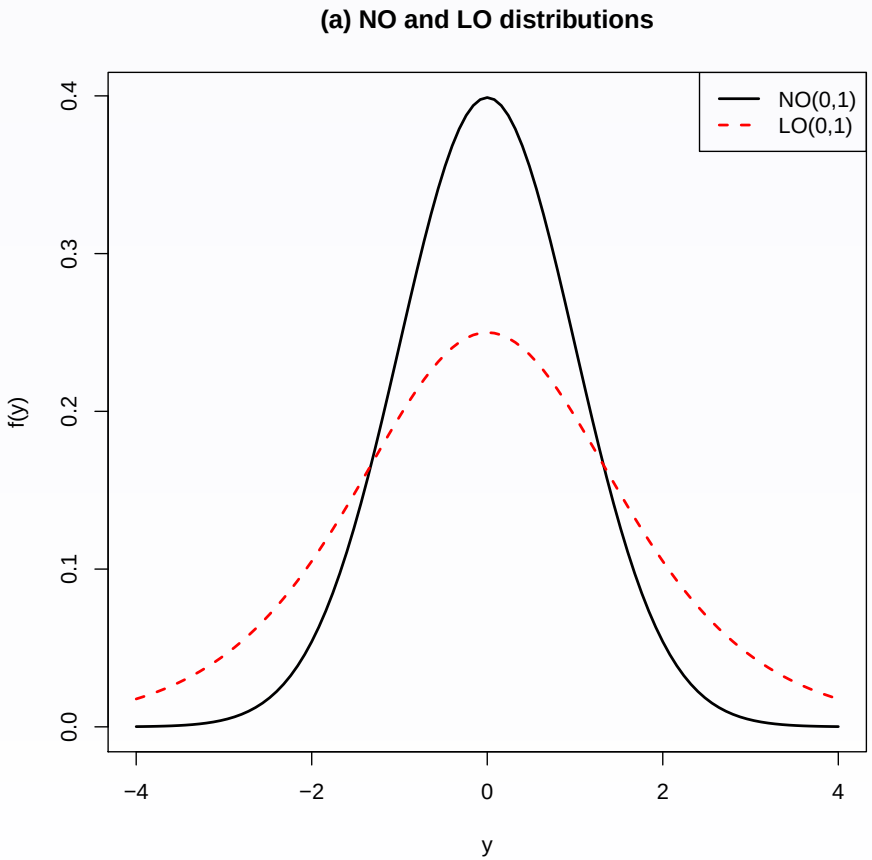
Two parameter distributions on $(-\infty, \infty)$

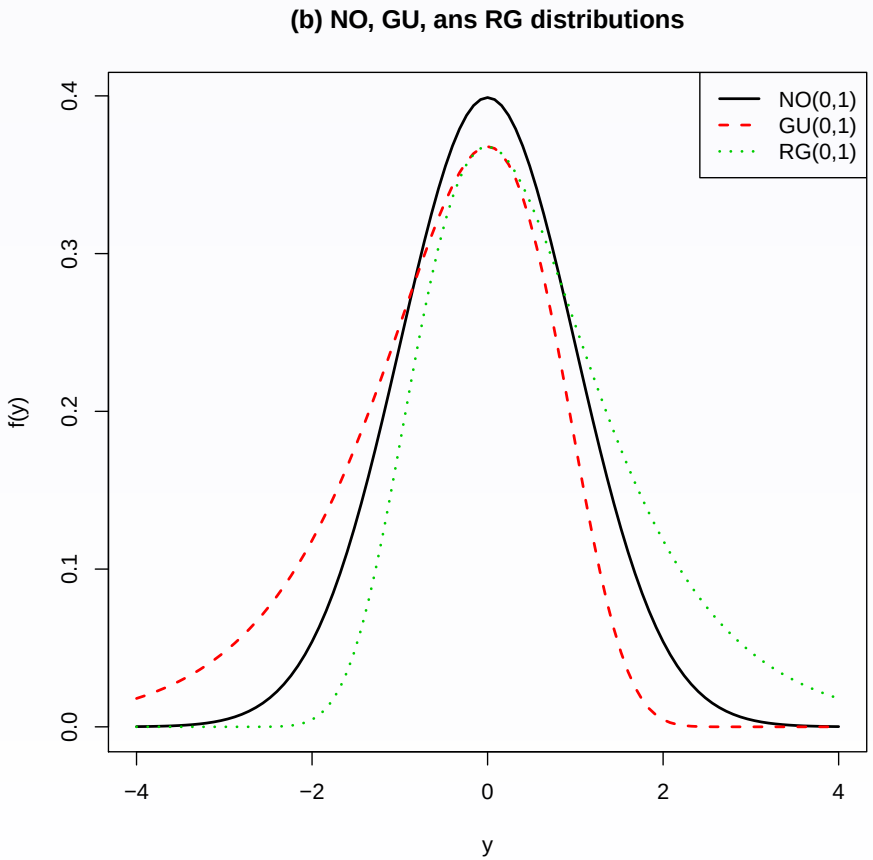
- Gumbel `GU`(μ, σ) left skew
- Logistic `LO`(μ, σ) lepto
- Normal `NO`(μ, σ) and `NO2`(μ, σ)

Reverse Gumbel `RG`(μ, σ) right skew

gamlss

Two parameter on ℜ





gamlss

Three parameter on $\Re$

exponential Gaussian : exGAUS(μ, σ, ν) for modelling right skew data,

normal family: NOF(μ, σ, ν) for modelling mean and variance relationships following the power law;

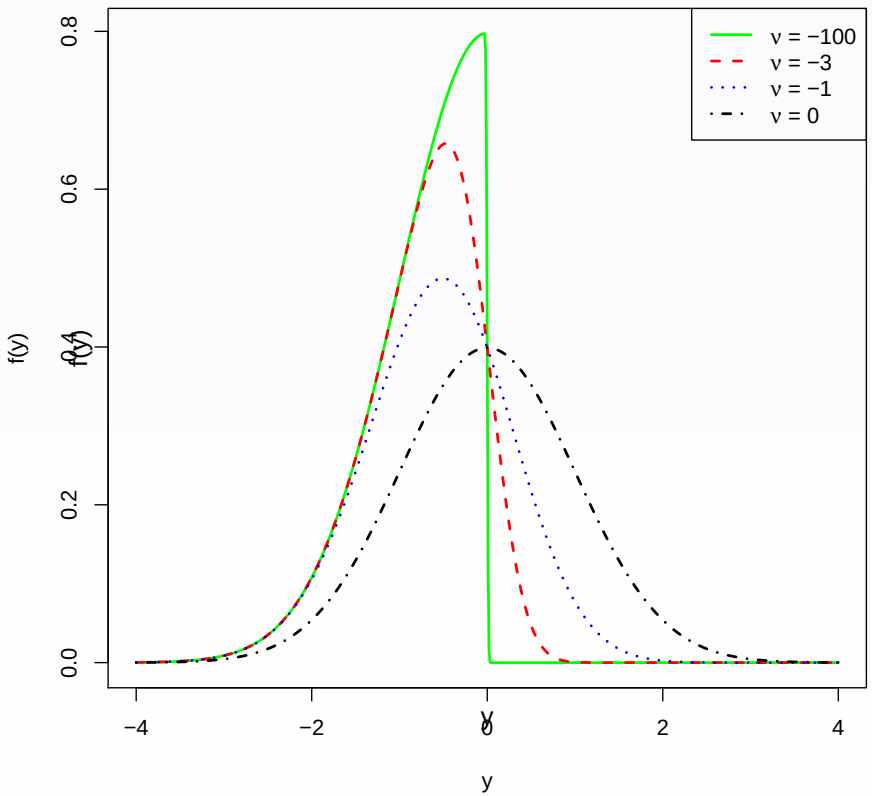
power exponential: PE(μ, σ, ν) and PE2(μ, σ, ν) for modelling lepto and platy kurtotic data;

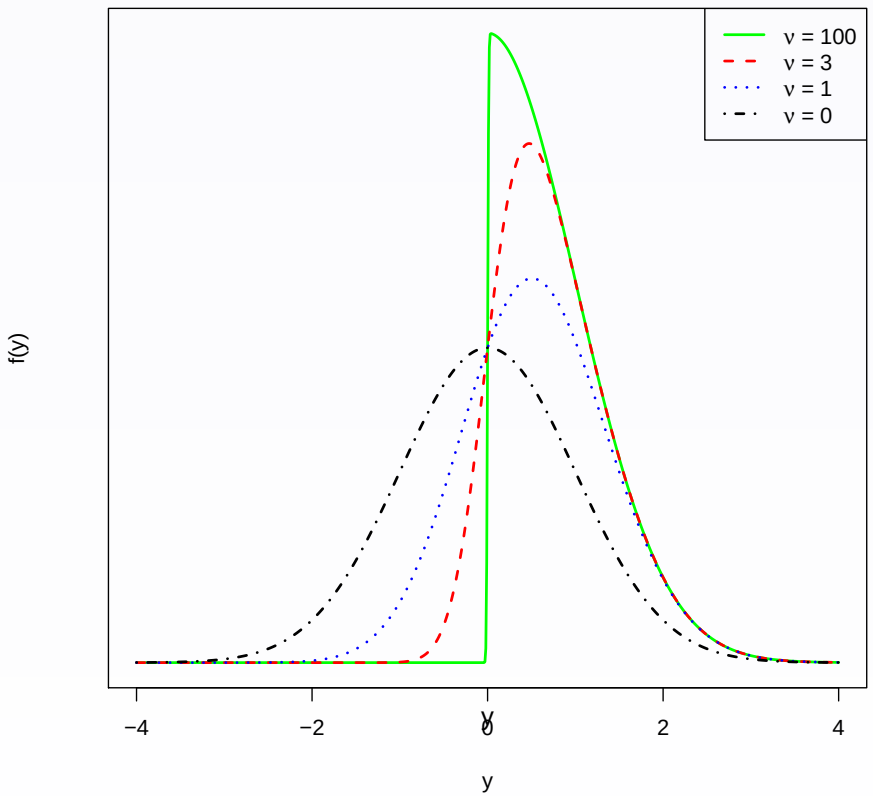
t family: TF(μ, σ, ν) and TF2(μ, σ, ν) for modelling lepro kurtotic data;

skew normal: SN1(μ, σ, ν) and SK2(μ, σ, ν) for modellinng skewness in data.

gamlss

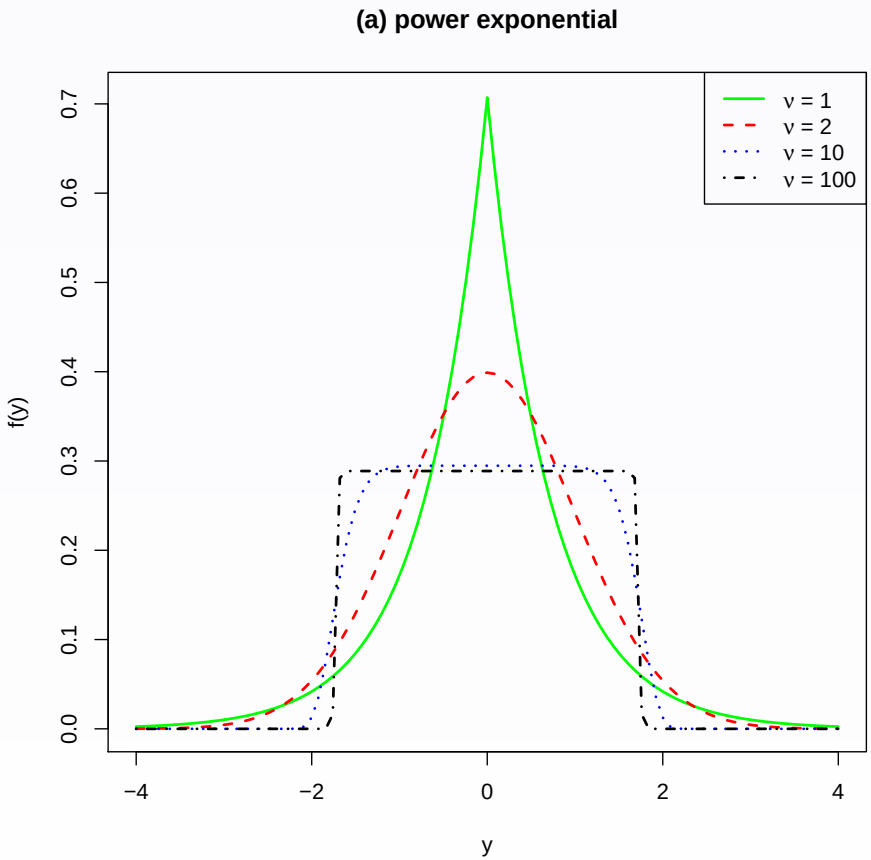
Three parameter on ℜ: skew normal type 1

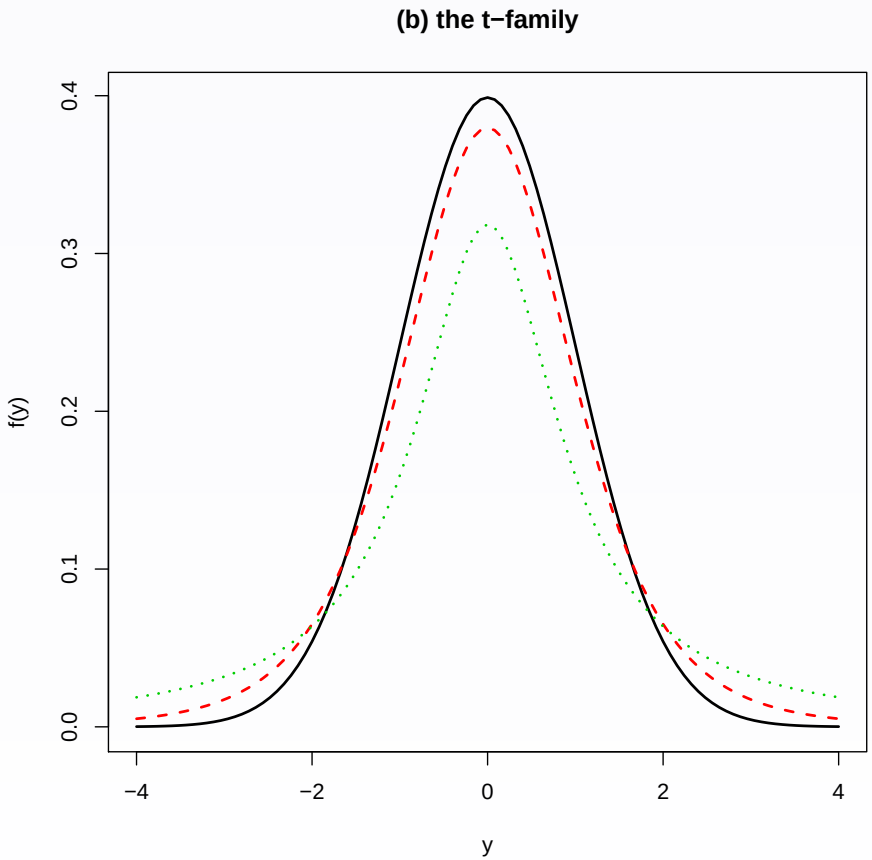




gamlss

Three parameter on ℜ: PE and TF distributions





gamlss

Four parameter on $\Re$

exponential generalised beta type 2: $\texttt{EGB2}(\mu, \sigma, \nu, \tau)$ skewness and leptokutrosis;

generalised t: $\texttt{GT}(\mu, \sigma, \nu, \tau)$ kurtosis;

Johnson's SU: $\texttt{JSU}(\mu, \sigma, \nu, \tau)$ and $\texttt{JSUo}(\mu, \sigma, \nu, \tau)$ skewness and leptokurtosis;

normal-exponential-t: $\texttt{NET}(\mu, \sigma, \nu, \tau)$, robustly location and scale;

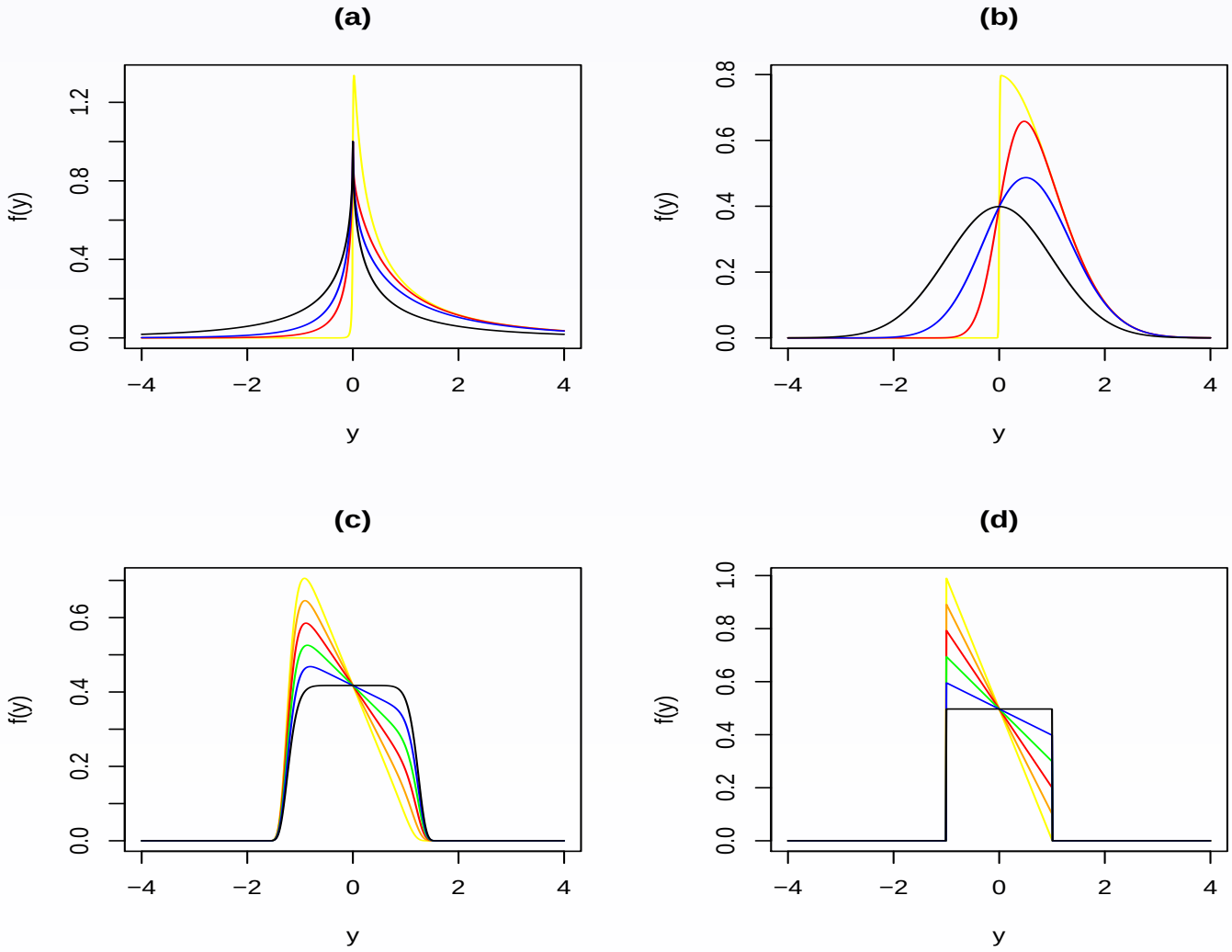
skew exponential power: $\texttt{SEP1}(\mu, \sigma, \nu, \tau)$, $\texttt{SEP2}(\mu, \sigma, \nu, \tau)$, $\texttt{SEP3}(\mu, \sigma, \nu, \tau)$ and $\texttt{SEP4}(\mu, \sigma, \nu, \tau)$ skewness and lepto-platy;

sinh-arcsinh: $\texttt{SHASH}(\mu, \sigma, \nu, \tau)$, $\texttt{SHASHo}(\mu, \sigma, \nu, \tau)$ and $\texttt{SHASHo2}(\mu, \sigma, \nu, \tau)$ skewness and lepto-platy;

skew t: $\texttt{ST1}(\mu, \sigma, \nu, \tau)$, $\texttt{ST2}(\mu, \sigma, \nu, \tau)$, $\texttt{ST3}(\mu, \sigma, \nu, \tau)$, $\texttt{ST4}(\mu, \sigma, \nu, \tau)$, $\texttt{ST5}(\mu, \sigma, \nu, \tau)$ and $\texttt{SST}(\mu, \sigma, \nu, \tau)$ skewness and leptokurtosis.

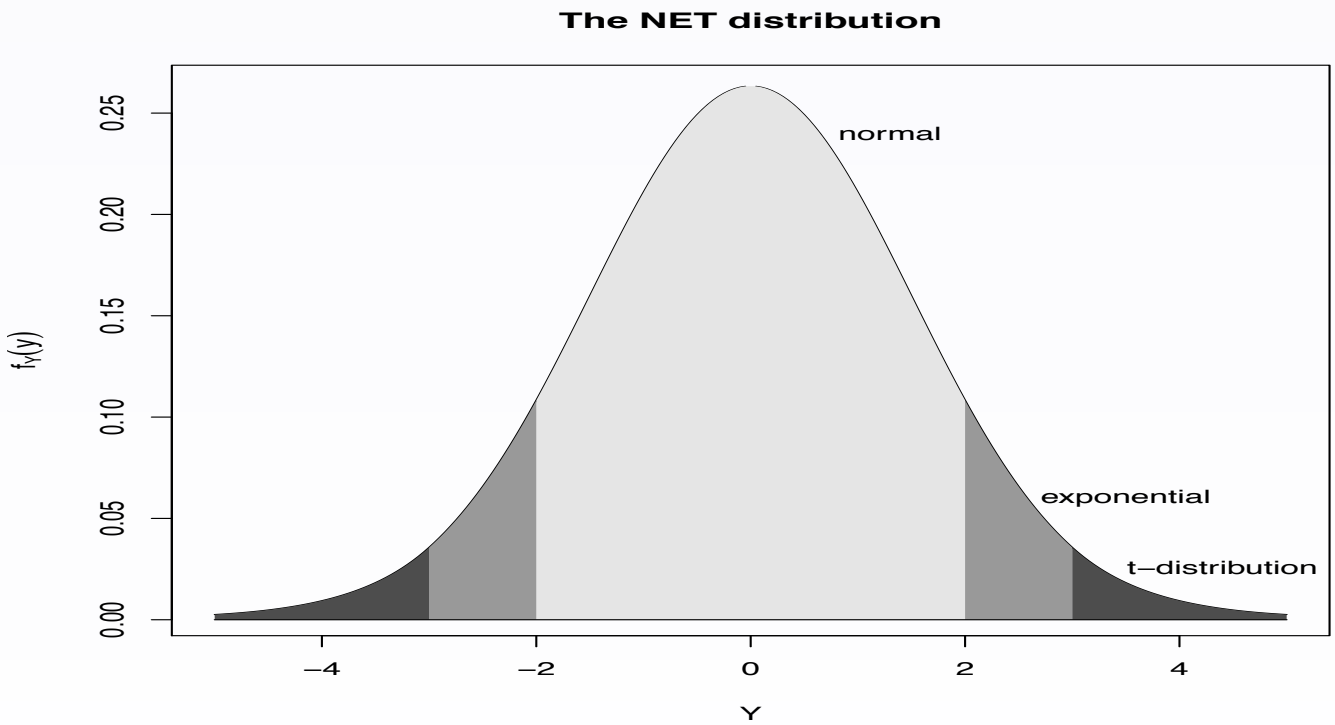
gamlss

Example of four parameter on ℜ: SEP1



gamlss

The NET distribution



gamlss

Summary of GAMLSS family distributions defined on $(-\infty, +\infty)$

Distributions	family	no par.	skewness	kurtosis
Exp. Gaussian	exGAUS	3	positive	-
Exp.G. beta 2	EGB2	4	both	lepto
Gen. t	GT	4	(symmetric)	lepto
Gumbel	GU	2	(negative)	-
Johnson's SU	JSU, JSUo	4	both	lepto
Logistic	LO	2	(symmetric)	(lepto)
Normal-Expon.-t	NET	2,(2)	(symmetric)	lepto
Normal	NO-NO2	2	(symmetric)	
Normal Family	NOF	3	(symmetric)	(meso)

gamlss

Summary of GAMLSS family distributions defined on $(-\infty, +\infty)$

Power Expon.	PE-PE2	3	(symmetric)	both
Reverse Gumbel	RG	2	positive	
Sinh Arcsinh	SHASH,	4	both	both
Skew Exp. Power	SEP1-SEP4	4	both	both
Skew t	ST1-ST5, SST	4	both	lepto
t Family	TF	3	(symmetric)	lepto

gamlss

Choosing an appropriate distribution

global deviance

$$\texttt{GDEV} = -2\hat{\ell}(\hat{\boldsymbol{\theta}}) = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) \tag{1}$$

generalized Akaike information criterion

$$\texttt{GAIC}(k) = \texttt{GDEV} + k \cdot df \tag{2}$$

special cases

$$\begin{aligned} \text{AIC} &= \texttt{GDEV} + 2\, df \\ \text{SBC} &= \texttt{GDEV} + \log n \cdot df \ . \end{aligned}$$

gamlss

Choosing an appropriate distribution

prediction global deviance

$$\text{VDEV} = \text{TDEV} = -2\,\tilde{\ell}(\tilde{\boldsymbol{\theta}}) = -2\sum_{i=1}^{n_P} \log f(\tilde{y}_i|\tilde{\boldsymbol{\theta}})$$

$\tilde{\mathbf{y}}$ indicates the response variable values at the validation (or test) sample,

$$\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\nu}}, \tilde{\boldsymbol{\mu}})$$

are the fitted parameters for the validation (or test) sample (which in **gamlss** can be obtained using the functions `predict()`,

gamlss

How to fit distributions in R

optim() or mle() R functions requiring initial parameter values

gamlsssML() fits a distribution (using MLE) on the response[1]

gamlss(): fits a distribution on the response using RS or CG or mixed algorithms,

histDist() fits a distribution using `gamlsssML()`, and plots a histogram of the response with the fitted distribution[1]

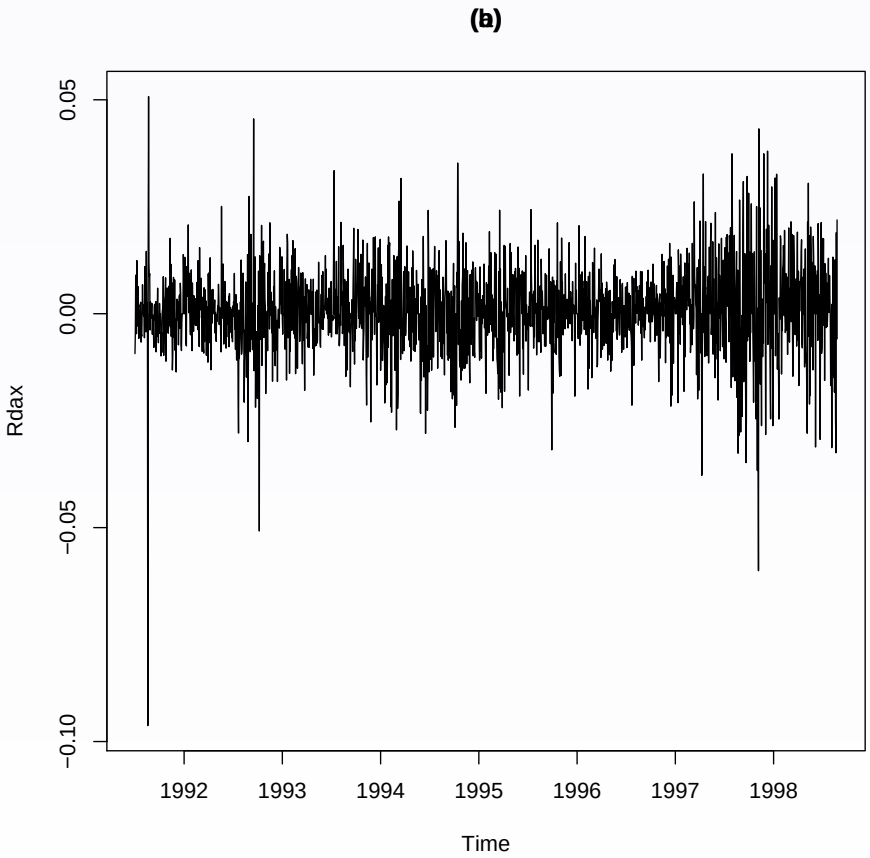
fitDist() fits a set of distributions to the response and chooses the one with the smallest GAIC[1]

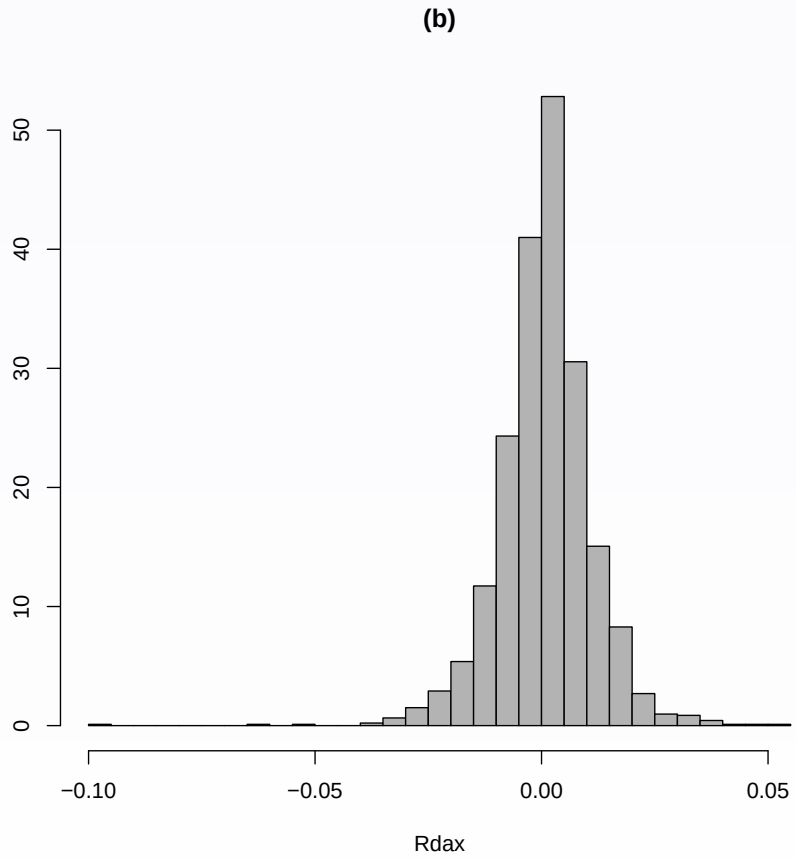
chooseDist() fits a set of distributions on a fitted model and chooses the one with the smallest GAIC

[1] with no explanatory variables

gamlss

DAX returns data





gamlss

DAX returns data: `fitDist()`

```
f1 <- fitDist(Rdax, type="realline")
f1$fits
     GT      SEP2       PE2       PE       JSU      JSUo      SEP1      SEP4
-11967.6 -11965.5 -11962.4 -11962.4 -11961.4 -11961.4 -11961.3 -11961.2
     SEP3       TF2        TF       ST4       ST1      EGB2       ST5       ST2
-11960.8 -11960.6 -11960.6 -11960.1 -11959.8 -11959.7 -11959.4 -11959.2
     ST3       SST     SHASH    SHASHo2     SHASHo        LO
-11958.8 -11958.8 -11957.5    -11956.2    -11956.2    -11932.1
     NET       SN1       SN2        NO    exGAUS        GU        RG
-11919.1 -11759.8 -11739.1 -11733.2 -11733.1 -11262.1 -10249.8
f1$failed
list()
```

gamlss

DAX returns data: chooseDist()

```
t1 <- chooseDist(m1, type = "realline")
minimum GAIC(k= 2 ) family:  GT
minimum GAIC(k= 3.84 ) family:  GT
minimum GAIC(k= 7.53 ) family:  PE2
t1
            2         3.84         7.53
NO -11733.21 -11729.53 -11722.15
GU -11262.12 -11258.44 -11251.06
RG -10249.88 -10246.20 -10238.82
....
ST5 -11959.48 -11952.12 -11937.36
SST -11958.87 -11951.51 -11936.75
 GT -11967.68 -11960.32 -11945.56
```

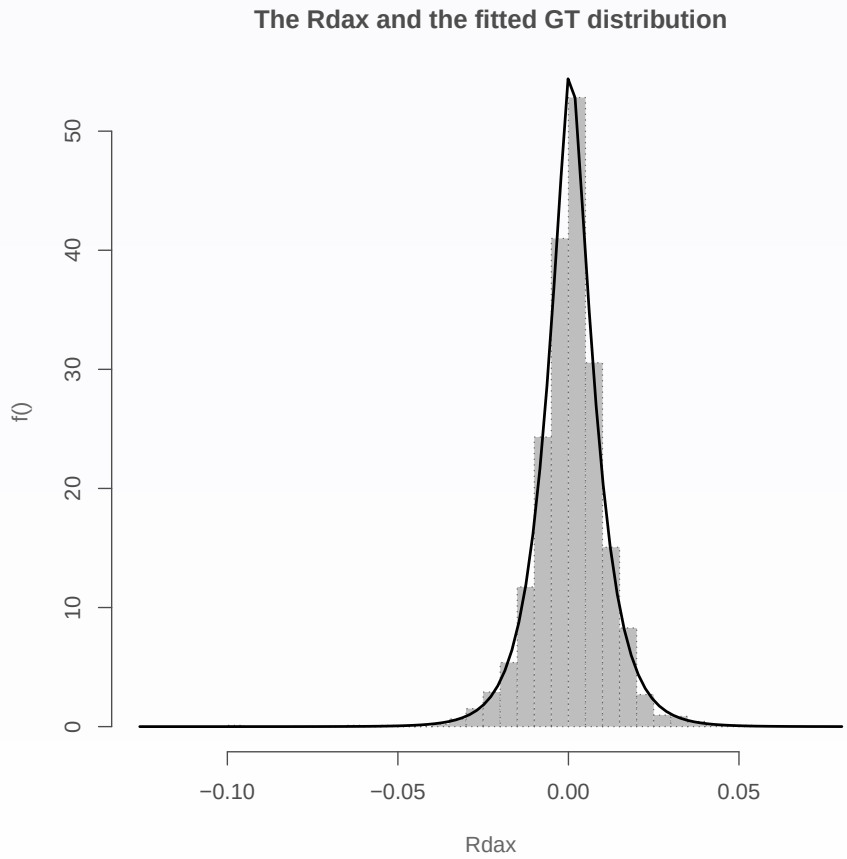
gamlss

DAX returns data: chooseDist()

```
getOrder(t1,1)[1:6]
      GT       SEP2        PE2         PE        JSU       JSUo
-11967.68 -11965.53 -11962.46 -11962.46 -11961.48 -11961.48
mf <- update(m1, family="GT")
summary(f1)
fitted(f1,"mu")[1]
[1] 0.0006838042
fitted(f1,"sigma")[1]
[1] 0.009610674
fitted(f1,"nu")[1]
[1] 5.96016
fitted(f1, "tau")[1]
[1] 1.417194
fh<-histDist(Rdax, family=GT, nbins=30, line.col="black")
```

gamlss

DAX returns data fitted model



gamlss

END

for more information see

www.gamlss.org

gamlss