Flexible Regression and Smoothing

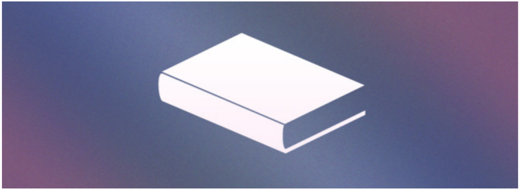
The gamlss.family Distributions

Bob Rigby Mikis Stasinopoulos

Dipartimento di Scienze Cliniche a de Comunita, Universita degli studi di Milano, Italy, January 2018

Information

Books & Vignettes



The GAMLSS Books

One book on GAMLSS is published and two others are in preparation

- `Flexible Regression and Smoothing: Using GAMLSS in R' was published in April 2017. See here for an old draft version.
- `Distributions for Location Scale and Shape: Using GAMLSS in R', is in preparation. See here for the November 2017 draft version of the book.
- `Generalised Additive Models for Location Scale and Shape: A Distributional Regression Approach' is in preparation. A draft version will be appear in the summer 2018.

gamlss

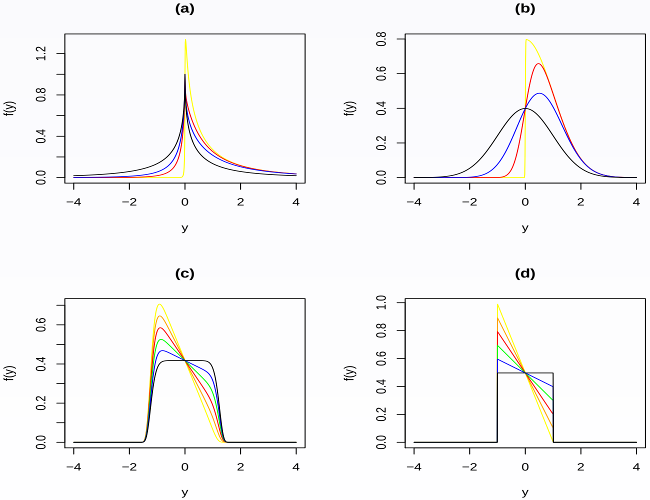
Different types of distribution in GAMLSS

1. *continuous distributions*: $f_Y(y|\boldsymbol{\theta})$, are usually defined on $(-\infty, +\infty)$, $(0, +\infty)$ or $(0, 1)$.
2. *discrete distributions*: $P(Y = y|\boldsymbol{\theta})$ are defined on $y = 0, 1, 2, \ldots, n$, where n is a known finite value or n is infinite, i.e. usually discrete (count) values.
3. *mixed distributions*: (finite mixture distributions) are mixtures of continuous and discrete distributions, i.e. continuous distributions where the range of Y has been expanded to include some discrete values with non-zero probabilities.

Different types of distribution: demos

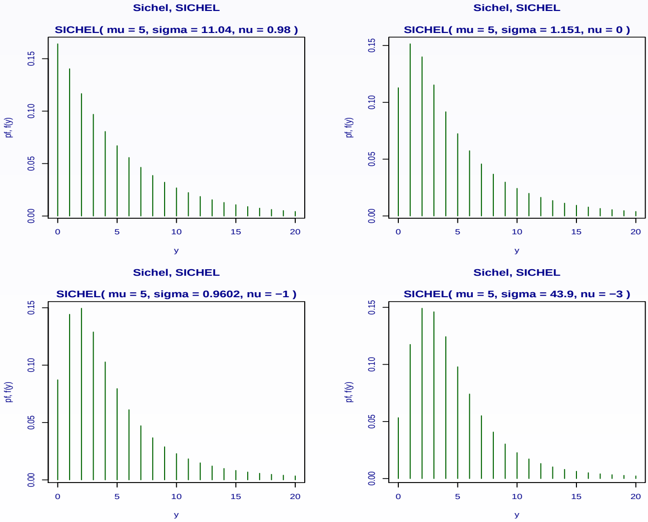
1. demo.GA()
2. demo.PO()
3. demo.BI()
4. demo.BE()
5. demo.BEINF()

Example of continuous distribution: SEP1

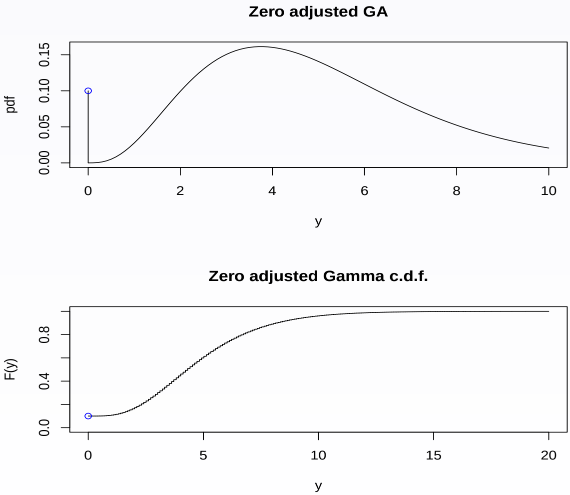


gamlss

Example of discrete distribution: Sichel



Example of mixed distribution distributions: ZAGA



Skewness and kurtosis

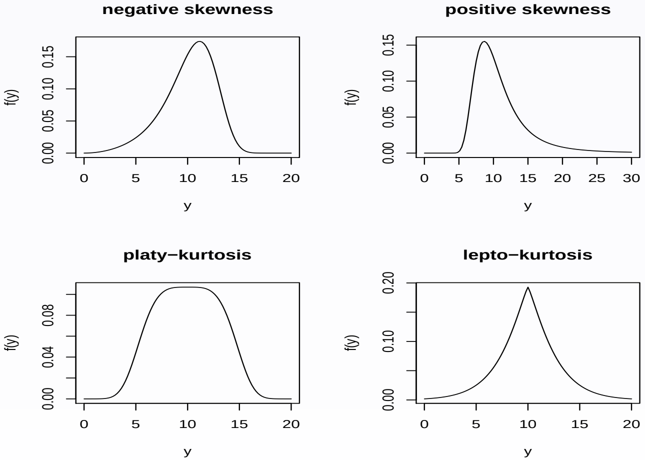


Figure: Showing different types of continuous distributions

Types of continuous distributions

$(-\infty, \infty)$: real line $\Re$

$(0, \infty)$: positive real line $\Re^+$

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

gamlss

Location and scale family

For these distributions, if

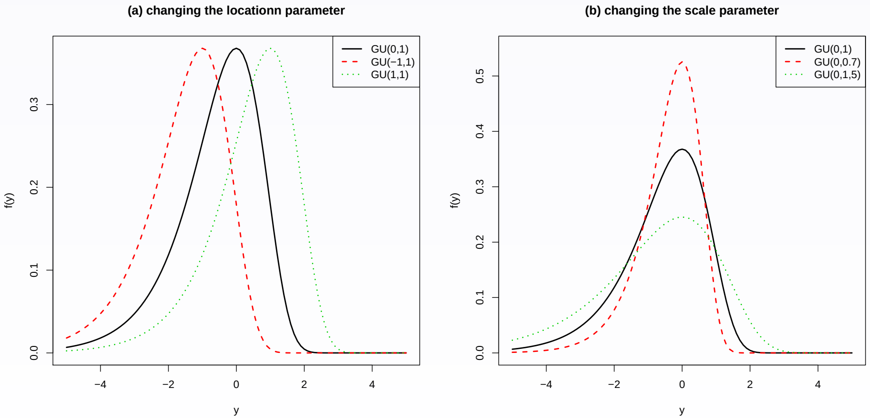
$$Y \sim \mathcal{D}(\mu, \sigma, \nu, \tau)$$

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

Location-scale family



gamlss

Two parameter on $\Re$

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

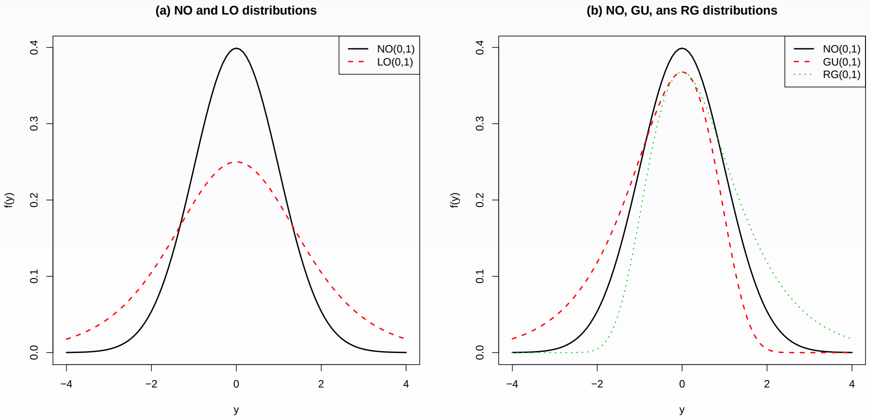
Gumbel `GU`(μ, σ) left skew

Logistic `LO`(μ, σ) lepto

Normal `NO`(μ, σ) and `NO2`(μ, σ)

Reverse Gumbel `RG`(μ, σ) right skew

Two parameter on ℜ



gamlss

Three parameter on $\Re$

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

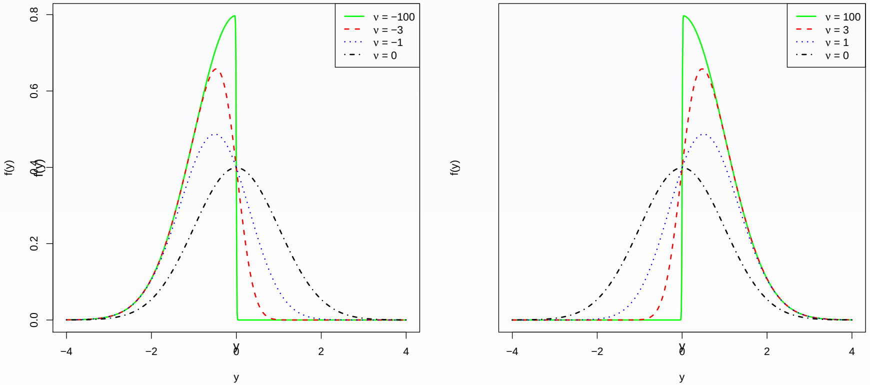
normal family: `NOF`(μ, σ, ν) for modelling mean and variance relationships following the power law;

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

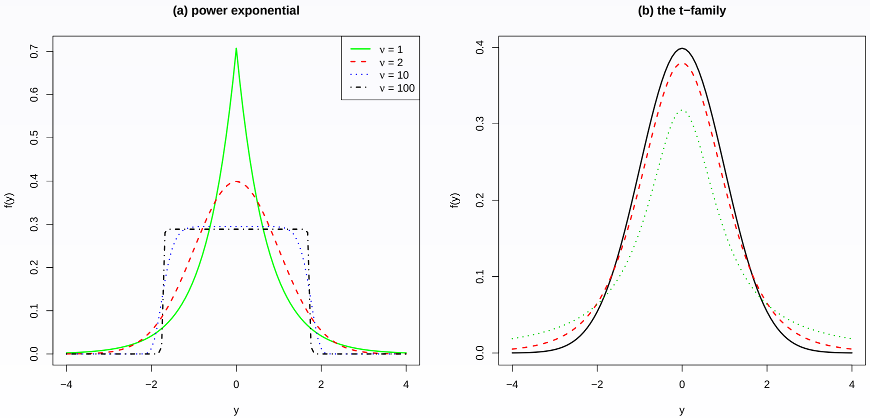
t family: `TF`(μ, σ, ν) and `TF2`(μ, σ, ν) for modelling lepro kurtotic data;

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

Three parameter on $\Re$: skew normal type 1



Three parameter on $\Re$: 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;

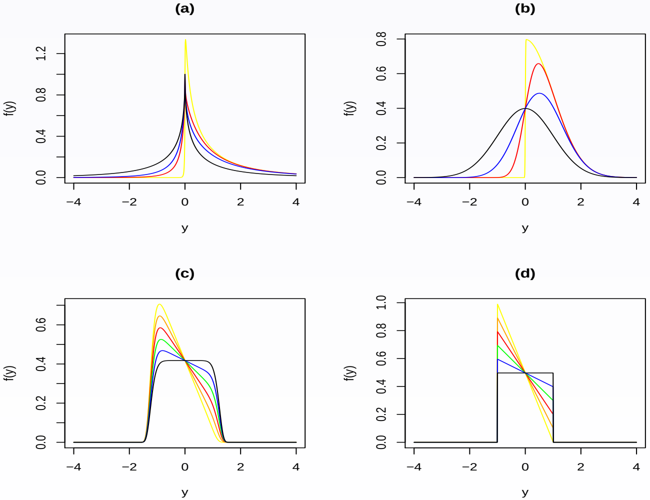
normal-exponential-*t*: $\texttt{NET}(\mu, \sigma, \nu, \tau)$, robustly location and scale;

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;

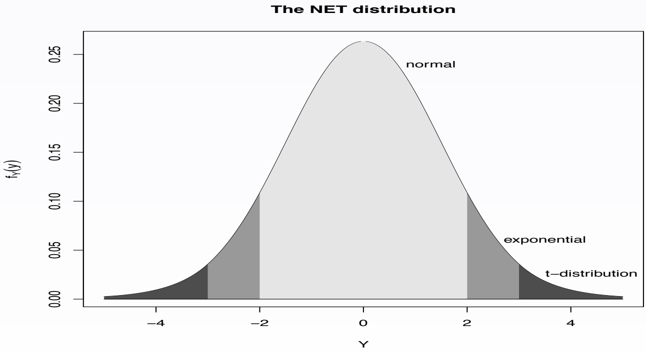
sinh-arcsinh: $\texttt{SHASH}(\mu, \sigma, \nu, \tau)$, $\texttt{SHASHo}(\mu, \sigma, \nu, \tau)$ and $\texttt{SHASHo2}(\mu, \sigma, \nu, \tau)$ skewness and lepto-platy;

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.

Example of four parameter on $\Re$: SEP1



The NET distribution



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

Continuous distribution on $\Re^+$: scale family

If a random variable is distributed as

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

and

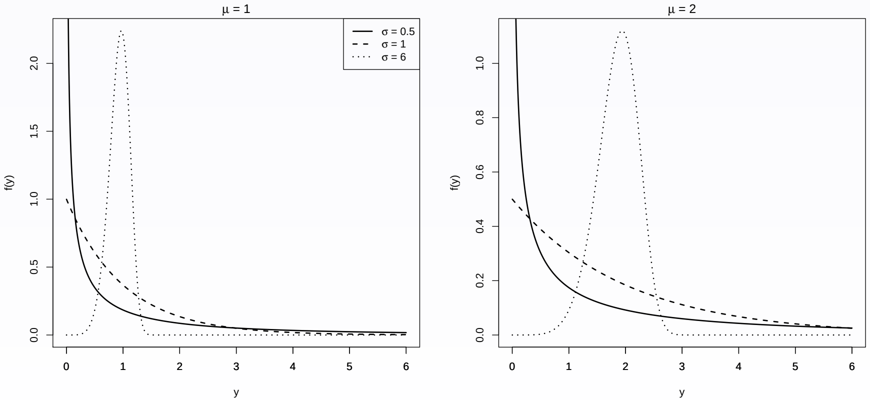
$$\varepsilon = (Y/\mu) \sim \mathcal{D}(1, \sigma, \nu, \tau)$$

then Y has a scale family and the random variable

$$Y = \mu\varepsilon$$

is a scaled version of the random variable ε.
μ does not effect the shape.

Continuous distribution on $\Re^+$: Weibull pdf

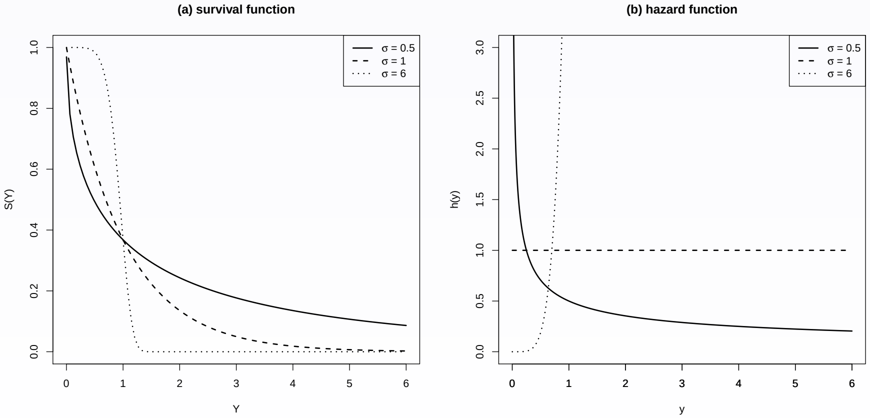


Continuous distribution on $\Re^+$: one and two parameters

- exponencial $\texttt{EXP}(\mu, \sigma)$
- gamma: $\texttt{GA}(\mu, \sigma)$ member of the exponential family;
- inverse gamma: $\texttt{IGAMMA}(\mu, \sigma)$;
- inverse Gaussian: $\texttt{IG}(\mu, \sigma)$ a member of the exponential family;
- log-normal: $\texttt{LOGNO}(\mu, \sigma)$ and $\texttt{LOGNO2}(\mu, \sigma)$.
- Pareto: $\texttt{PARETO}(\mu, \sigma)$, $\texttt{PARETO2o}(\mu, \sigma)$ and $\texttt{GP}(\mu, \sigma)$ for heavy tail;
- Weibull: $\texttt{WEI}(\mu, \sigma)$, $\texttt{WEI2}(\mu, \sigma)$ and $\texttt{WEI3}(\mu, \sigma)$ used in survival analysis.

gamlss

Continuous distribution on $\Re^+$: Weibull survival and hazard functions



gamlss

Continuous distribution on $\Re^+$: three parameters

Box-Cox Cole and Green $\texttt{BCCG}(\mu, \sigma, \nu)$ known also as the LMS method in centile estimation.

generalised gamma $\texttt{GG}(\mu, \sigma, \nu)$

generalised inverse Gaussian $\texttt{GIG}(\mu, \sigma, \nu)$

log-normal family $\texttt{LNO}(\mu, \sigma, \nu)$ based on the standard Box-Cox transformation

$$Z = \begin{cases} \frac{(Y^{\nu}-1)}{\nu} & \text{if} \quad \nu \neq 0 \\ \log(Y) & \text{if} \quad \nu = 0 \end{cases} \tag{1}$$

gamlss

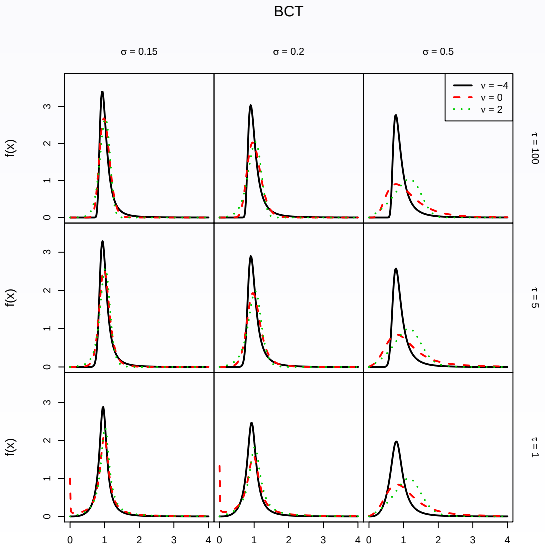
Continuous distribution on $\Re^+$: four parameters

Box-Cox power exponential `BCPE`(μ, σ, ν, τ) and `BCPEo`(μ, σ, ν, τ) skewness and platy-lepto;

Box-Cox *t* `BCT`(μ, σ, ν, τ) and `BCTo`(μ, σ, ν, τ) skewness and leptokurtosis;

generalised beta type 2 `GB2`(μ, σ, ν, τ) skewness and platy-lepto.

Continuous distribution on $\Re^+$: four parameters, BCT



Continuous GAMLSS family distributions defined on $(0, +\infty)$

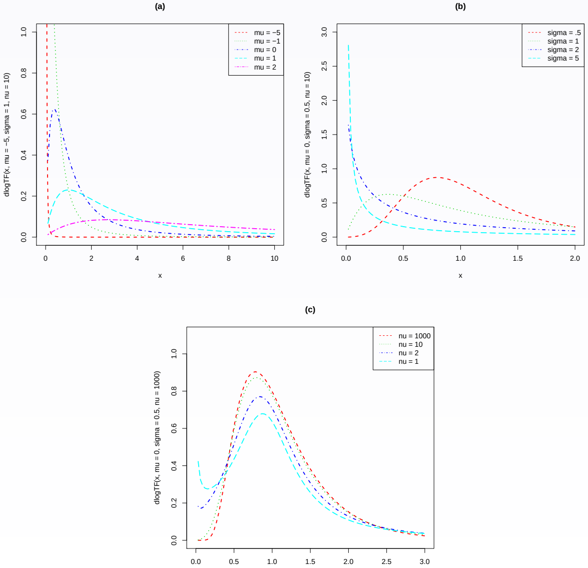
Distributions	family	no par.	skewness	kurtosis
BCCG	BCCG	3	both	
BCPE	BCPE	4	both	both
BCT	BCT	4	both	lepto
Exponential	EXP	1	(positive)	-
Gamma	GA	2	(positive)	-
Gen. Beta type 2	GB2	4	both	both
Gen. Gamma	GG-GG2	3	positive	-
Gen. Inv. Gaussian	GIG	3	positive	-
Inv. Gaussian	IG	2	(positive)	-
Log Normal	LOGNO	2	(positive)	-
Log Normal family	LNO	2,(1)	positive	
Reverse Gen. Extreme	RGE	3	positive	-
Weibull	WEI-WEI3	2	(positive)	-

Transformation from $(-\infty, \infty)$ to $(0, +\infty)$

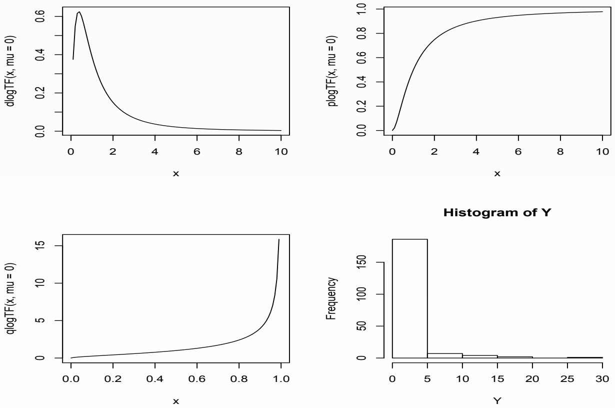
- Any distribution for Z on $(-\infty, \infty)$ can be transformed to a corresponding distribution for $Y = \exp(Z)$ on $(0, +\infty)$
- For example: from t distribution to log t distribution
- `gen.Family("TF", type="log")`

gamlss

Positive response: log T distribution



Positive response: log T distribution

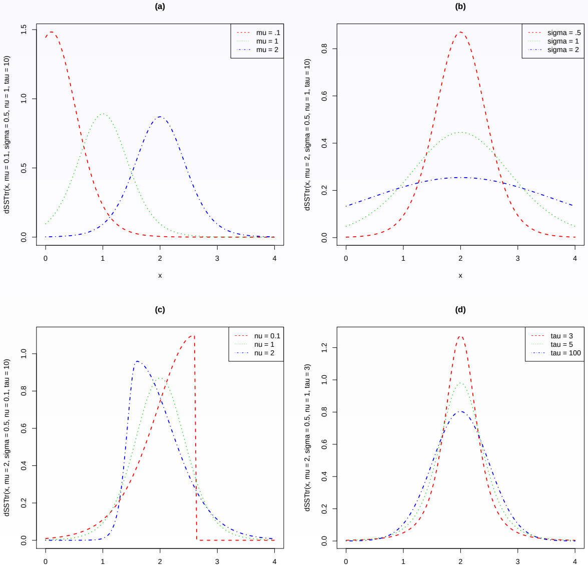


Truncating from $(-\infty, \infty)$ to $(0, +\infty)$

- Any distribution for Z on $(-\infty, \infty)$ can be truncated to a corresponding distribution for Y on $(0, +\infty)$
- For example: from `SST`(μ, σ, ν, τ) distribution to `flipped-SST`(μ, σ, ν, τ) distribution
- `gen.trun(0,"SST",type="left")`

gamlss

Truncating from $(-\infty, \infty)$ to $(0, +\infty)$



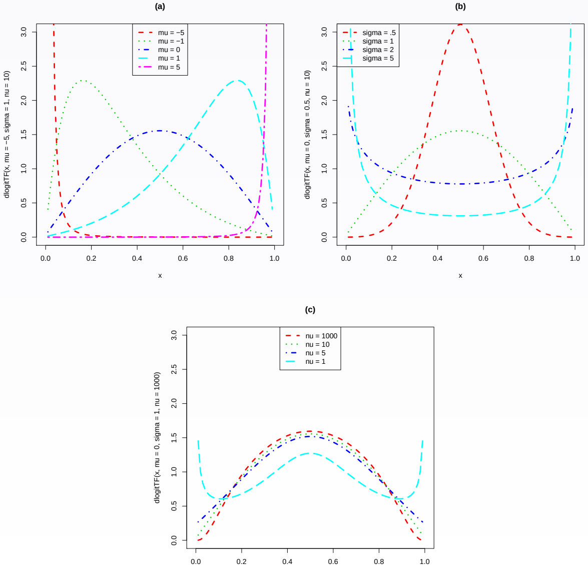
Continuous GAMLSS family distributions defined on $(0, 1)$

Distributions	family	no par.	skewness	kurtosis
Beta	BE	2	(both)	-
Beta original	BEo	2	(both)	-
Logit normal	LOGITNO	2	(both)	-
Generalized beta type 1	GB1	4	(both)	(both)

Transformation from $(-\infty, \infty)$ to $(0, 1)$

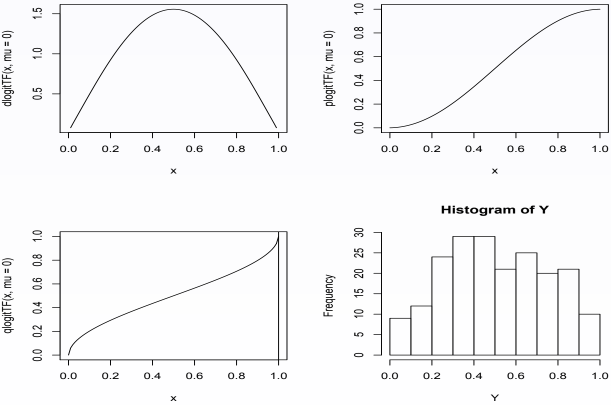
- Any distribution for Z on $(-\infty, \infty)$ can be transformed to a corresponding distribution for $Y = \frac{1}{1+e^{-Z}}$ on $(0, 1)$
- For example: from t distribution to *logit* t distribution
- `gen.Family("TF", "logit")`

logit T distribution



gamlss

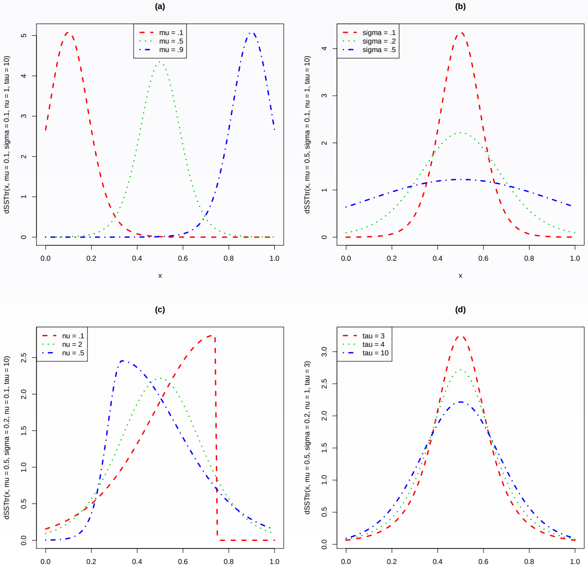
logit T distribution



Truncating from $(-\infty, \infty)$ to $(0, 1)$

- Any distribution for Z on $(-\infty, \infty)$ can be truncated to a corresponding distribution for Y on $(0, 1)$
- For example: from SST(μ, σ, ν, τ) distribution to trun-SST(μ, σ, ν, τ) distribution
- `gen.trun(c(0,1),"SST",type="both")`

Truncated SST on $(0, 1)$



Summary of methods of generating distributions

1. univariate transformation from a single random variable (LOGNO, BCCG, BCPE, BCT, EGB2, SHASHo, inverse log and logic transformations ...)
2. transformation from two or more random variables (TF, ST2, exGAUS)
3. truncation distributions
4. a (continuous or finite) mixture of distributions (TF, GT, GB2, EGB2)
5. Azzalini type methods (SN1, SEP1, SEP2, ST1, ST2)
6. splicing distributions (SN2, SEP3, SEP4, ST3, ST4, NET)
7. stopped sums
8. systems of distributions

Comparison of properties of distributions

1. Explicit pdf cdf and inverse cdf
2. Explicit centiles and centile based measures (e.g median)
3. Explicit moment based measures (e.g. mean)
4. Explicit mode(s)
5. Continuity of the pdf and its derivatives with respect to y
6. Continuity of the pdf with respect to μ, σ, ν and τ
7. Flexibility in modelling skewness and kurtosis
8. Range of Y

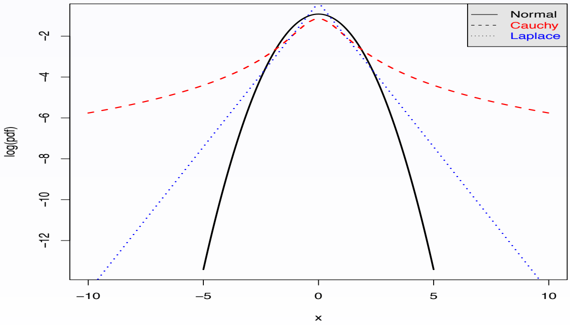
Theoretical comparison of distributions

- OK we have a lot of distributions in GAMLSS.
  Do we need any more?
- Do we need all of them or some of them are redundant?

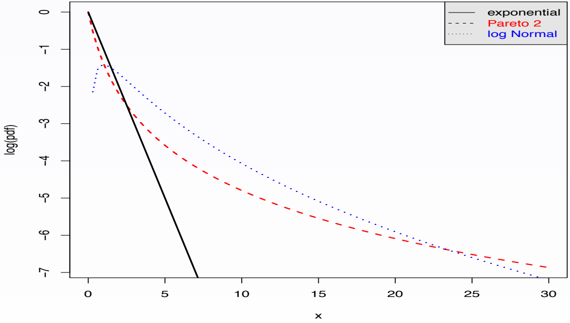
Is there any way to compare them theoretically?

- We can compare the tail behaviour in of the distributions
- Since most of then are location-scale we can compare their flexibility in terms of skewness and kurtosis

Comparing tails: real line



Comparing tails: positive real line



Types of tails

There are three main forms for $\log f_Y(y)$ for a tail of Y

$$
\begin{aligned}
\text{Type I:} &\quad -k_2 \left(\log|y|\right)^{k_1}, \\
\text{Type II:} &\quad -k_4 \, |y|^{k_3}, \\
\text{Type III:} &\quad -k_6 \, \mathrm{e}^{k_5|y|},
\end{aligned}
$$

decreasing k's results in a heavier tail,

Types of tails: classification (real line)

Value of k_1–k_6	Distribution name	Distribution
$k_1 = 1$	Cauchy	$\text{CA}(\mu, \sigma)$
	Generalized t	$\text{GT}(\mu, \sigma, \nu, \tau)$
	Skew t type 3	$\text{ST3}(\mu, \sigma, \nu, \tau)$
	Skew t type 4	$\text{ST4}(\mu, \sigma, \nu, \tau)$
	Stable	$\text{SB}(\mu, \sigma, \nu, \tau)$
	t	$\text{TF}(\mu, \sigma, \nu)$
$k_1 = 2$	Johnson's SU	$\text{JSU}(\mu, \sigma, \nu, \tau)$
	Johnson's SU original	$\text{JSUo}(\mu, \sigma, \nu, \tau)$
$0 < k_3 < \infty$	Power exponential	$\text{PE}(\mu, \sigma, \nu)$
	Power exponential type 2	$\text{PE2}(\mu, \sigma, \nu)$
	Sinh-arcsinh original	$\text{SHASHo}(\mu, \sigma, \nu, \tau)$
	Sinh-arcsinh	$\text{SHASH}(\mu, \sigma, \nu, \tau)$
	Skew exponential power type 3	$\text{SEP3}(\mu, \sigma, \nu, \tau)$
	Skew exponential power type 4	$\text{SEP4}(\mu, \sigma, \nu, \tau)$
$k_3 = 1$	Exponential generalized beta type 2	$\text{EGB2}(\mu, \sigma, \nu, \tau)$
	Gumbel	$\text{GU}(\mu, \sigma)$
	Laplace	$\text{LA}(\mu, \sigma)$
	Logistic	$\text{LG}(\mu, \sigma)$
	Reverse Gumbel	$\text{RG}(\mu, \sigma)$
$k_3 = 2$	Normal	$\text{NO}(\mu, \sigma)$
$0 < k_5 < \infty$	Gumbel	$\text{GU}(\mu, \sigma)$
	Reverse Gumbel	$\text{RG}(\mu, \sigma)$

Types of tails: classification (positive real line)

Value of k_1–k_6	Distribution name	Distribution
$k_1 = 1$	Box-Cox Cole-Green	$\text{BCCG}(\mu, \sigma, \nu)$
	Box-Cox power exponential	$\text{BCPE}(\mu, \sigma, \nu, \tau)$
	Box-Cox t	$\text{BCT}(\mu, \sigma, \nu, \tau)$
	Generalized beta type 2	$\text{GB2}(\mu, \sigma, \nu, \tau)$
	Generalized gamma	$\text{GG}(\mu, \sigma, \nu)$
	Inverse gamma	$\text{IGA}(\mu, \sigma)$
	log t	$\text{LOGT}(\mu, \sigma, \nu)$
	Pareto Type 2	$\text{PA2o}(\mu, \sigma)$
$k_1 = 2$	Box-Cox Cole-Green	$\text{BCCG}(\mu, \sigma, \nu)$
	Lognormal	$\text{LOGNO}(\mu, \sigma)$
	Log Weibull	$\text{LOGWEI}(\mu, \sigma)$
$1 \leq k_1 < \infty$	Box-Cox power exponential	$\text{BCPE}(\mu, \sigma, \nu, \tau)$
$0 < k_3 < \infty$	Box-Cox Cole-Green	$\text{BCCG}(\mu, \sigma, \nu)$
	Box-Cox power exponential	$\text{BCPE}(\mu, \sigma, \nu, \tau)$
	Generalized gamma	$\text{GG}(\mu, \sigma, \nu)$
	Weibull	$\text{WEI}(\mu, \sigma)$
$k_3 = 1$	Exponential	$\text{EX}(\mu)$
	Gamma	$\text{GA}(\mu, \sigma)$
	Generalized inverse Gaussian	$\text{GIG}(\mu, \sigma, \nu)$
	Inverse Gaussian	$\text{IG}(\mu, \sigma)$

How to fit distributions in R

- **optim() or mle()** requires initial parameter values
- **gamlsssML()** mle a using variation of the mle() function
- **gamlss():** mle using RS or CG or mixed algorithms,
- **histDist()** mle using RS or CG or mixed algorithms, for a univariate sample only, and plots a histogram of the data with the fitted distribution
- **fitDist()** fits a set of distributions and chooses the one with the smallest GAIC

Strength of glass fibres data

Data summary:

R data file: glass in package `gamlss.data` of dimensions 63×1

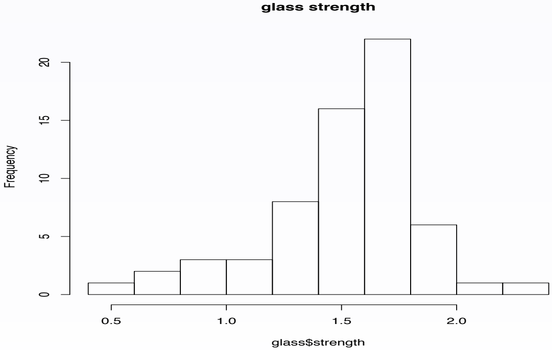
sourse: Smith and Naylor (1987)

strength: the strength of glass fibres (the unit of measurement are not given).

purpose: to demonstrate the fitting of a parametric distribution to the data.

conclusion: a SEP4 distribution fits adequately

Strength of glass fibres data: SBC



Strength of glass fibres data: creating truncated distribution

```
data(glass)
library(gamlss.tr)
gen.trun(par = 0, family = TF)
A truncated family of distributions from TF
 has been generated
 and saved under the names:
 dTFtr pTFtr qTFtr rTFtr TFtr
The type of truncation is left and the
truncation parameter is 0
```

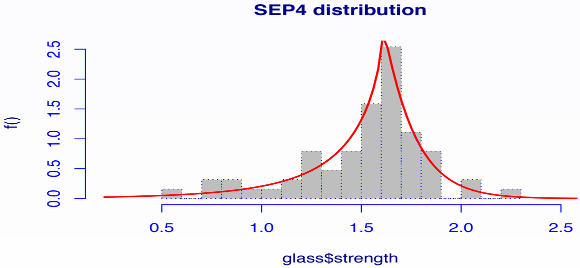
Strength of glass fibres data: fitting the models

```
> m1<-fitDist(strength, data=glass, k=2, extra="TFtr")
        # AIC
> m2<-fitDist(strength, data=glass, k=log(length(strength)),
         extra="TFtr") # SBC
> m1$fit[1:8]
    SEP4     SEP3   SHASHo     EGB2      JSU    BCPEo      ST2     BCTo
27.68790 27.98191 28.01800 28.38691 30.41380 31.17693 31.40101 31.47677
> m2$fit[1:8]
    SEP4     SEP3   SHASHo     EGB2       GU     WEI3      JSU       PE
36.26044 36.55444 36.59054 36.95945 38.19839 38.69995 38.98634 39.00792
```

Strength of glass fibres data: Results

```
histDist(glass$strength, SEP4, nbins = 13,
      main = "SEP4 distribution",
     method = mixed(20, 50))
```

Strength of glass fibres data: SBC



Strength of glass fibres data: summary

```
Call:  gamlss(formula = y~1, family = FA)
Mu Coefficients:
(Intercept)
      1.581
Sigma Coefficients:
(Intercept)
     -1.437
Nu Coefficients:
(Intercept)
    -0.1280
Tau Coefficients:
(Intercept)
     0.3183
 Degrees of Freedom for the fit: 4 Residual Deg. of Freedom   59
Global Deviance:     19.8111
            AIC:     27.8111
            SBC:     36.3836
```

END

for more information see

www.gamlss.org