Continuous distributions

Mikis Stasinopoulos[1] Bob Rigby[1]

[1]STORM, London Metropolitan University

GAMLSS: Short course Cordoba, Argentina, Oct 2012

1. Introduction: continuous distributions in gamlss
   - Types of distribution in GAMLSS
   - Types of continuous distributions
   - Summary of methods of generating distributions
   - Comparison of properties
2. Theoretical comparison of distributions
3. Fitting distributions
   - Theoretical considerations for fitting distributions
   - How to fit distributions in R (by maximum likelihood estimation)
4. Examples
   - Strength of glass fibres data
   - Example of fitting regression models: CD4 count data
     - fitting orthogonal, fractional and piecewise polynomials
     - fitting cubic and penalized splines
     - fitting P-splines
     - Validation Global Deviance
     - Fitting different distributions
5. End

Different types of distribution

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

Types of continuous distributions

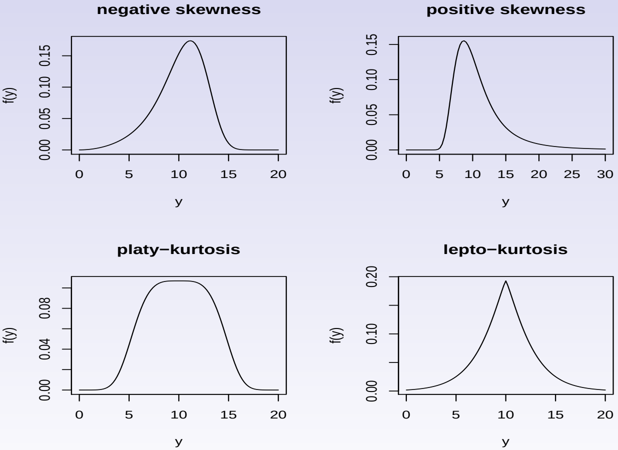


Figure: Showing different types of continuous distributions

Two parameter continuous distributions in GAMLSS

Two parameter distributions

BE	Beta
GA	Gamma
GU	Gumbel
LO	Logistic
LNO	Log Normal
NO	Normal
IG	Inverse Gaussian
RG	Reverse Gumbel
WEI	Weibull (also WEI2, WEI3)

Three parameter continuous distributions in GAMLSS

Three parameter distributions

BCCG	Box-Cox Normal
exGAUS	Exponential-Gaussian
GG	Generalized Gamma
GIG	Generalized Inverse Gaussian
PE	Power Exponential
TF	t family
SN1	skew normal type 1
SN2	skew normal type 1

Two and three parameters comparison: demos

1. demo.PE.NO()
2. demo.TF.NO()

Four parameter continuous distributions in GAMLSS

Four parameter distributions

BCT	Box-Cox *t*
BCPE	Box-Cox Power Exponential
EGB1-EGB2	Exponential Generalized Beta
GT	Generalized *t*
JSU	Johnson Su
SHASH	Sinh Arc-Sinh
SEP1-SEP4	Skew Exponential Power
ST1-ST5	Skew *t*

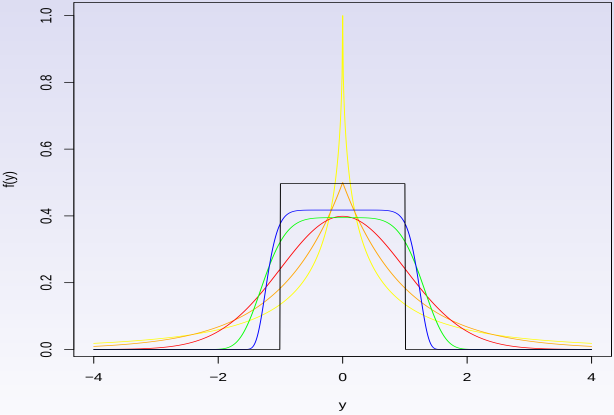
Continuous GAMLSS family distributions defined on $(-\infty, +\infty)$

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)

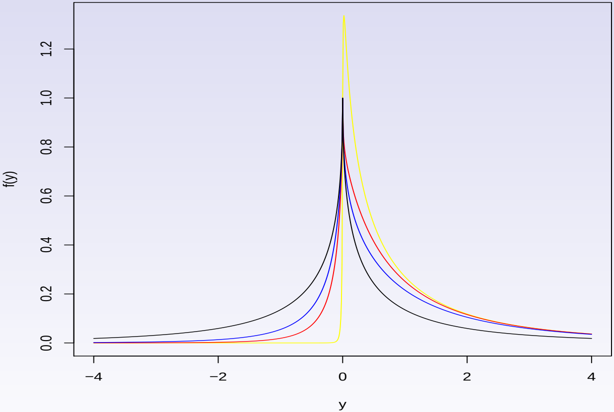
Continuous GAMLSS family distributions defined on $(-\infty, +\infty)$

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	4	both	lepto
t Family	TF	3	(symmetric)	lepto

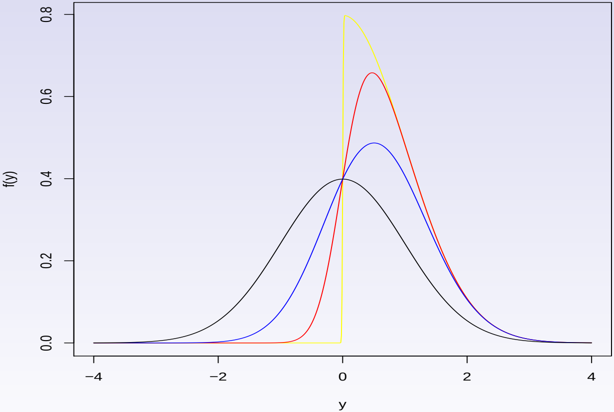
Symmetric SEP1 distribution ($\nu = 0$)



Skew SEP1 distributions ($\tau = 0.5$)

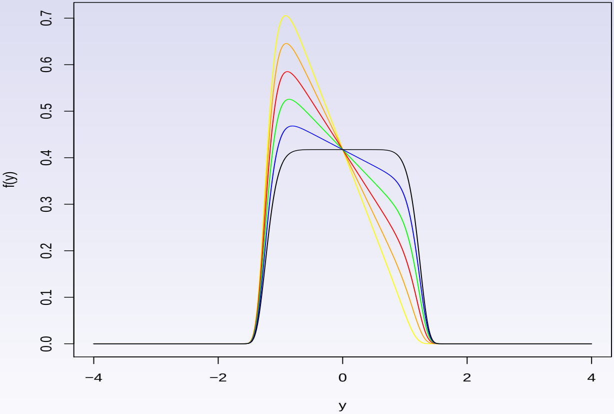


Skew SEP1 distributions ($\tau = 2$)

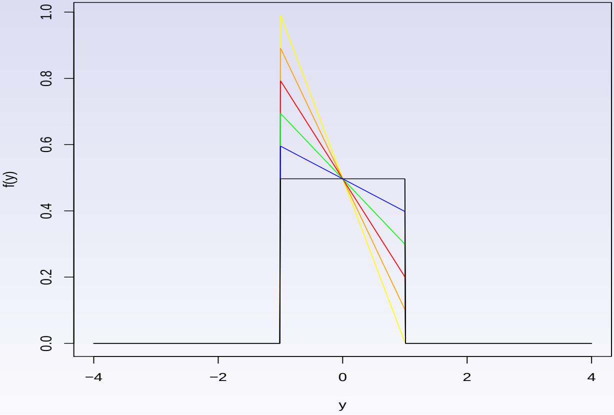


gamlss

Skew SEP1 distributions ($\tau = 10$)



Skew SEP1 distributions ($\tau = 1000$)



gamlss

Four parameters: demos

1. demo.SHASH()
2. demo.SEP1()

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)	-

Positive response: demos

1. demo.NO.LOGNO()
2. demo.BCPE()

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

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

Four parameters from 0 to 1: demos

1. demo.GB1()

gamlss

Summary of methods of generating distributions

1. univariate transformation from a single random variable
2. transformation from two or more random variables
3. truncation distributions
4. a (continuous or finite) mixture of distributions
5. Azzalini type methods
6. splicing distributions
7. stopped sums
8. systems of distributions

gamlss

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

Comparison of properties

Univariate transformation: Explicit pdf, cdf, inverse cdf, centiles

Azzalini type methods: Non-explicit cdf, inverse cdf, centiles, but flexible

Splicing methods: Explicit pdf, cdf, inverse cdf, centiles Lack of continuity of second derivative of pdf.

gamlss

Theoretical comparison of distributions

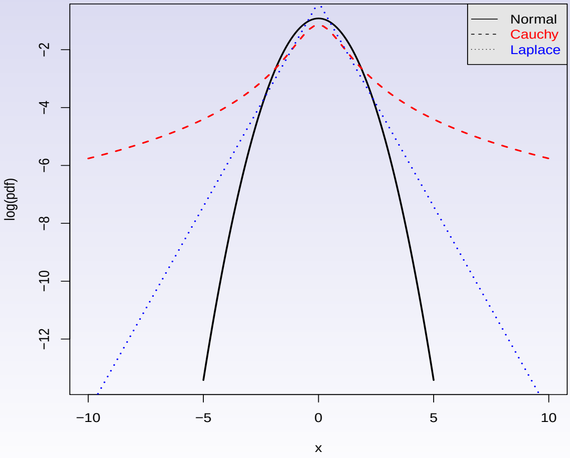
- OK we have a lot of distributions in GAMLSS.
  Do we need any more?
- Do we need all of them or some of them are redundant?

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

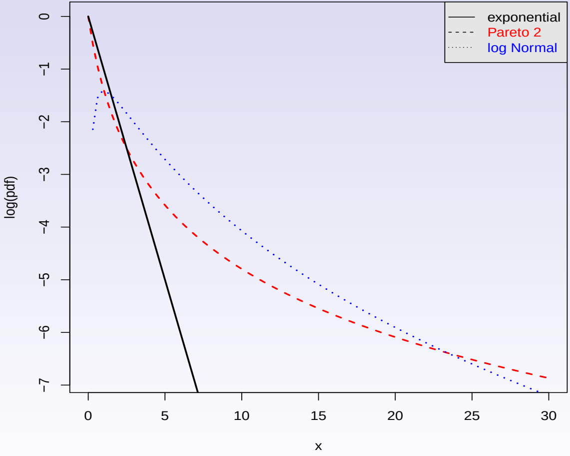
gamlss

Comparing tails: real line



gamlss

Comparing tails: positive real line



gamlss

Types of tails

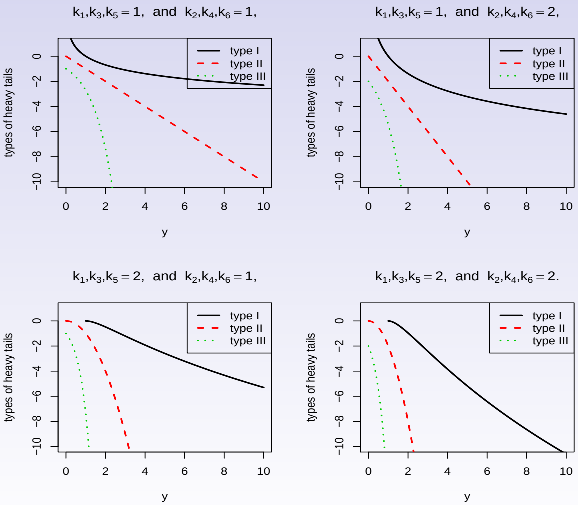
There are three main forms for $\log f_Y(y)$ for a tail of Y

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

decreasing k's results in a heavier tail,

gamlss

Types of tails



Types of tails: Bob's classification (real line)

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

gamlss

Types of tails: Bob's 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	$\text{IG}(\mu, \sigma)$

gamlss

Types of tails

- Given we are dealing with a long tail situation can we use the tail classification in order to chose an appropriate distribution for data?

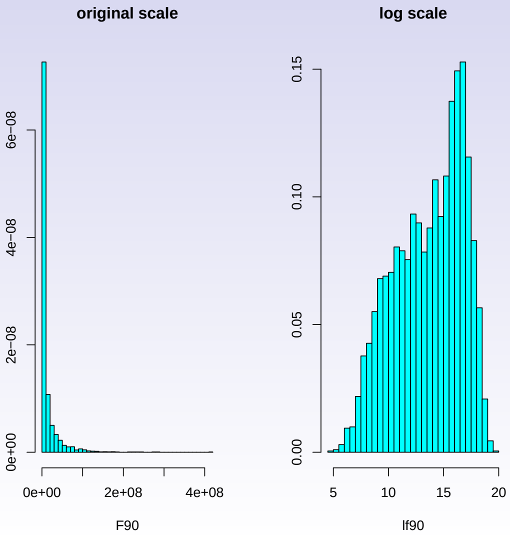
- For regression problem: ?? (research under progress)
- For single response variable (no explanatory variables): YES

Types of tails: single variable

- complementary cumulative distribution function (CCDF) ($\log[S_Y(y)]$ against $\log y$).
- complementary log log plot of the survival function.
- Hill 'type' plot

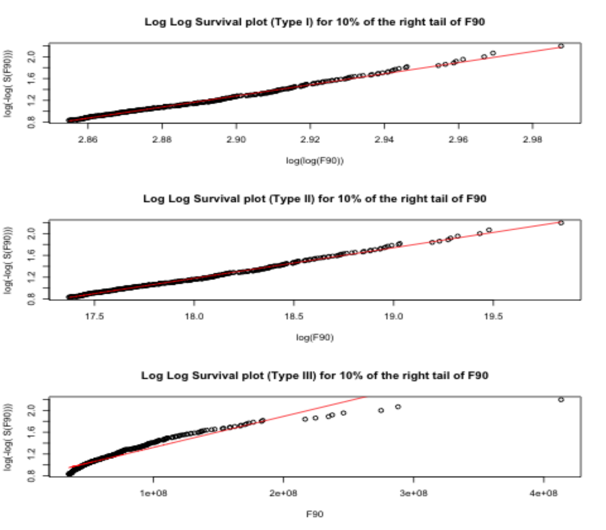
gamlss

The 90's Film data



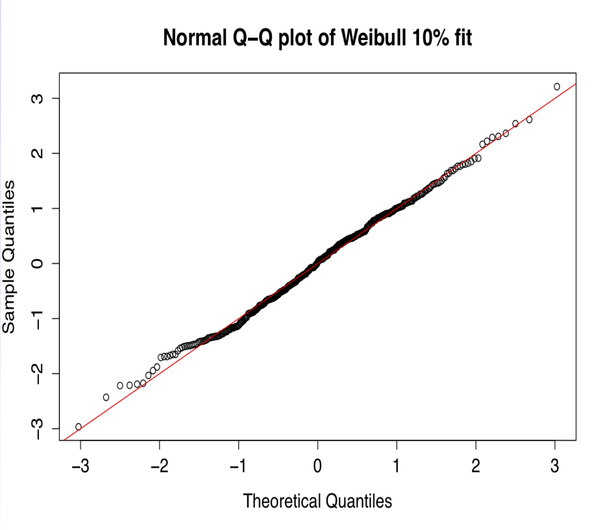
gamlss

Comparing tails: complementary log log survival.



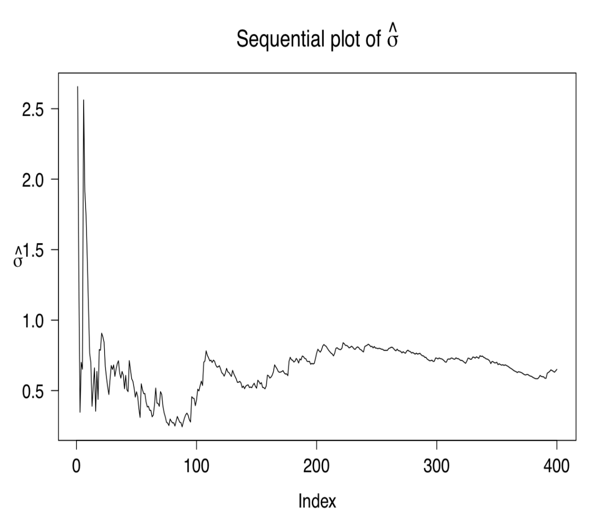
gamlss

Comparing tails: QQ plot from a Weibull



gamlss

Comparing tails: Hill 'type' plot for Weibull



Comparing skewness and kurtosis

Since moment based measures may not exist or may be unreliable, centile based measures of skewness and kurtosis are used for comparison

A general measure of skewness;

$$s_p = \frac{(y_p + y_{1-p})/2 - y_{0.5}}{(y_{1-p} - y_p)/2}$$

where $y_p = F_Y^{-1}(p)$. For example $p = 0.25$, giving Galton's measure of skewness

$$s_{0.25} = \frac{(Q_1 + Q_3)/2 - m}{(Q_3 - Q_1)/2}$$

a measure of central skewness

gamlss

Comparing skewness and kurtosis

A general measure of kurtosis

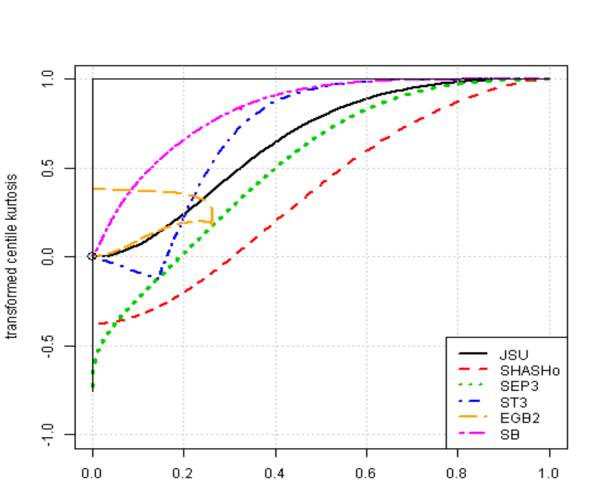
$$k_p = \frac{(y_{1-p} - y_p)}{(Q_3 - Q_1)}$$

standarised

$$s\,k_{0.01} = \frac{(y_{0.99} - y_{0.01})}{3.49\,(Q_3 - Q_1)}\,, \tag{1}$$

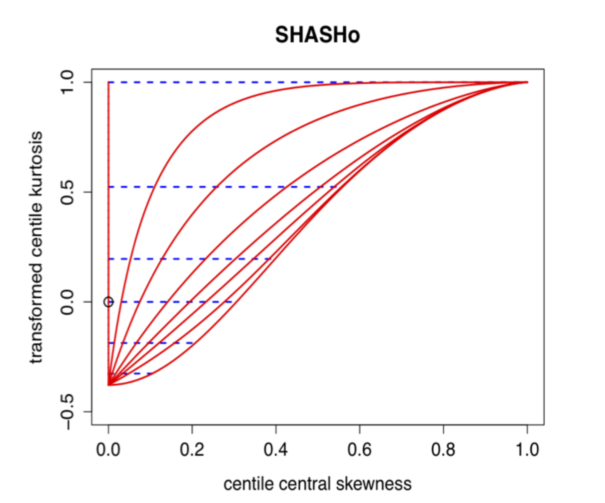
gamlss

Comparing skewness and kurtosis: a plot of the boundary of central skewness against transformed kurtosis



gamlss

Comparing skewness and kurtosis: The SHASHo distribution



gamlss

Comparing skewness and kurtosis

- There is some progress in classifying the continuous `gamlss.family` distributions
- the results apply to all continuous distributions (new or old)

Theoretical considerations for fitting distributions

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} f_Y(y_i|\boldsymbol{\theta}) \\
L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[F_Y\left(y_i + \Delta_i|\boldsymbol{\theta}\right) - F_Y\left(y_i - \Delta_i|\boldsymbol{\theta}\right)\right] \\
L(\boldsymbol{\theta}) &\approx \prod_{i=1}^{n} f_Y(y_i|\boldsymbol{\theta})\Delta_i = \left[\prod_{i=1}^{n} \Delta_i\right]\left[\prod_{i=1}^{n} f_Y(y_i|\boldsymbol{\theta})\right] \\
\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log f_Y(y_i|\boldsymbol{\theta}) + \sum_{i=1}^{n} \log \Delta_i
\end{aligned}
$$

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 smaller GAIC

gamlss

Strength of glass fibres data

Data summary:

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

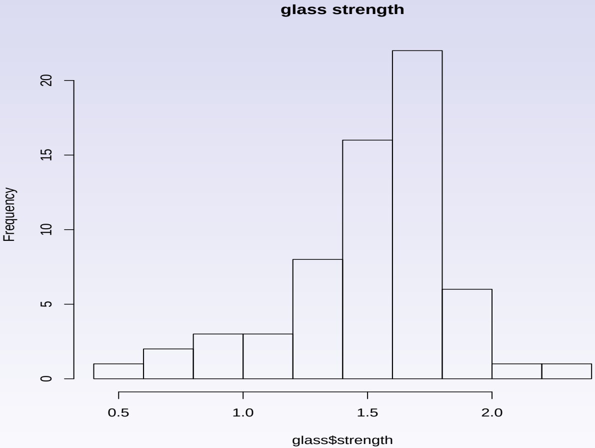
sourse: Smith and Naylor (1987)

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

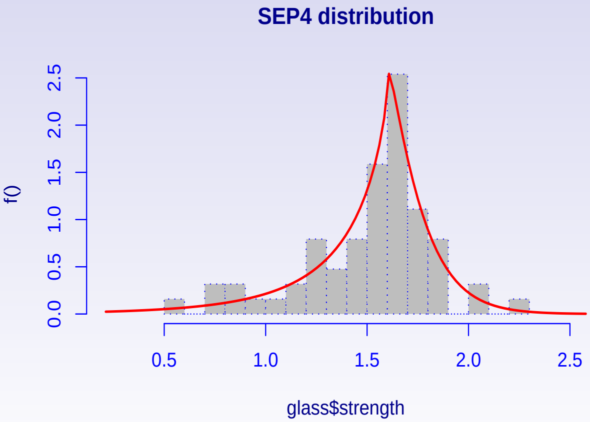
```
> m1<-fitDist(strength, data=glass, k=2, extra="TFtr")
          # AIC
> m2<-fitDist(strength, data=glass, k=log(length(strength)
           extra="TFtr") # SBC
> m1$fit[1:8]
    SEP4     SEP3   SHASHo     EGB2      JSU    BCPEo
27.68790 27.98191 28.01800 28.38691 30.41380 31.17693 31.4
> m2$fit[1:8]
    SEP4     SEP3   SHASHo     EGB2       GU     WEI3
36.26044 36.55444 36.59054 36.95945 38.19839 38.69995 38.9
```

Strength of glass fibres data: Results

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

gamlss

Strength of glass fibres data: SBC



gamlss

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
Global Deviance:     19.8111
               AIC:     27.8111
```

Example of fitting regression models: CD4 count data

Data summary:

R data file: CD4 in package **gamlss.data** of dimensions 609×2

source: Wade and Ades (1994)

variables:

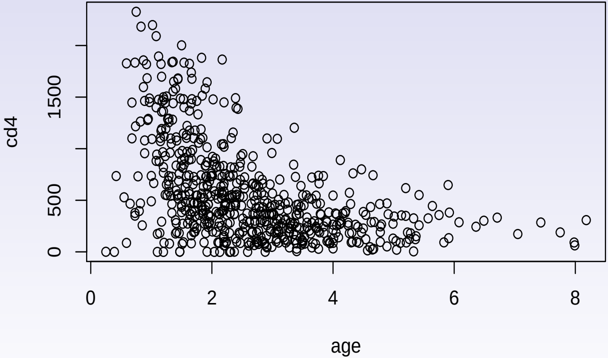
cd4 : CD4 counts from uninfected children born to HIV-1 mothers

age : age in years of the child.

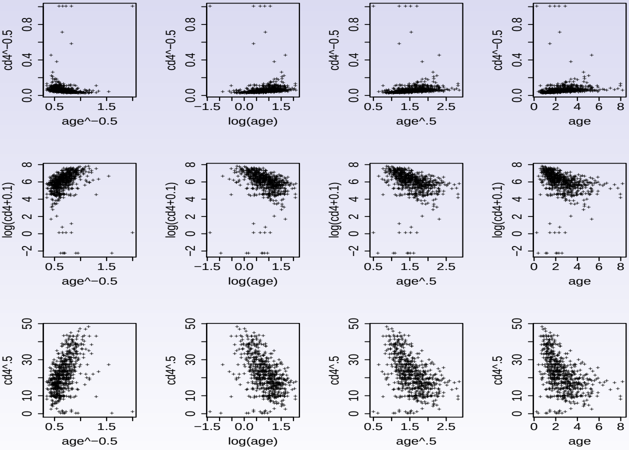
purpose: to demonstrate regression fitting of different functional forms for the explanatory variable age.

conclusion: models for both μ and σ are needed with a 4-parameter distribution for the response variable

CD4 count data: plotting the data



CD4 count data: transforming the data



gamlss

CD4 count data: fitting polynomials

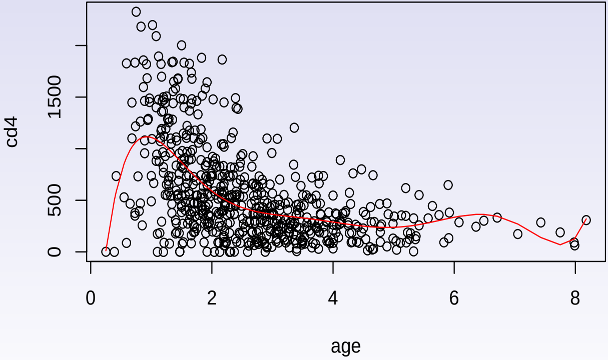
```
m1 <- gamlss(cd4 ~ age, sigma.fo = ~1, data = CD4, control
m2 <- gamlss(cd4 ~ poly(age, 2), sigma.fo = ~1, data = CD4)
m3 <- gamlss(cd4 ~ poly(age, 3), sigma.fo = ~1, data = CD4)
m4 <- gamlss(cd4 ~ poly(age, 4), sigma.fo = ~1, data = CD4)
m5 <- gamlss(cd4 ~ poly(age, 5), sigma.fo = ~1, data = CD4)
m6 <- gamlss(cd4 ~ poly(age, 6), sigma.fo = ~1, data = CD4)
m7 <- gamlss(cd4 ~ poly(age, 7), sigma.fo = ~1, data = CD4)
m8 <- gamlss(cd4 ~ poly(age, 8), sigma.fo = ~1, data = CD4)
```

gamlss

CD4 count data: fitting polynomials

	df	AIC
m7	9	8963.263
m8	10	8963.874
m5	7	8977.383
m4	6	8988.105
m3	5	8993.351
m2	4	8995.636
m1	3	9044.145

CD4 count data: fitting polynomials of degree 7



gamlss

CD4 count data: fractional polynomials

```
m1f <- gamlss(cd4 ~ fp(age, 1), sigma.fo = ~1, data = CD4)
m2f <- gamlss(cd4 ~ fp(age, 2), sigma.fo = ~1, data = CD4)
m3f <- gamlss(cd4 ~ fp(age, 3), sigma.fo = ~1, data = CD4)
```

CD4 count data: fractional polynomials

```
     df      AIC
m3f   8 8966.375
m2f   6 8978.469
m1f   4 9015.321
```

CD4 count data: fractional polynomials

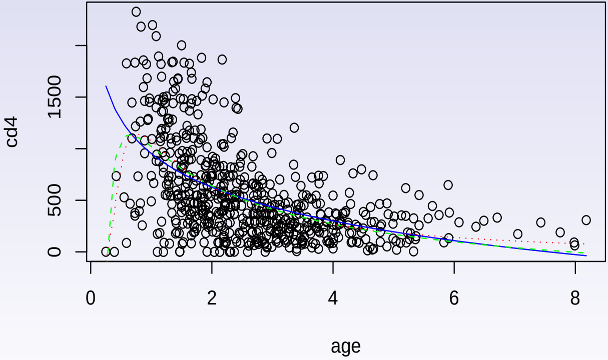
```
 m3f$mu.coefSmo

[[1]]
[[1]]$coef
     one
-599.2970 1116.7924 1776.1937  698.6097
[[1]]$power
[1] -2 -2 -2
[[1]]$varcoeff
[1]  2016.238  4316.743 22835.146  7521.417
```

terms : age^{-2}, $age^{-2}\log(age)$ and $age^{-2}\,[\log(age)]^2$.

gamlss

CD4 count data: fractional polynomials fits



gamlss

CD4 count data: piecewise polynomials

```
m2b <- gamlss(cd4 ~ bs(age), data = CD4)
m3b <- gamlss(cd4 ~ bs(age, df = 3), data = CD4)
m4b <- gamlss(cd4 ~ bs(age, df = 4), data = CD4)
m5b <- gamlss(cd4 ~ bs(age, df = 5), data = CD4)
m6b <- gamlss(cd4 ~ bs(age, df = 6), data = CD4)
m7b <- gamlss(cd4 ~ bs(age, df = 7), data = CD4)
m8b <- gamlss(cd4 ~ bs(age, df = 8), data = CD4)
```

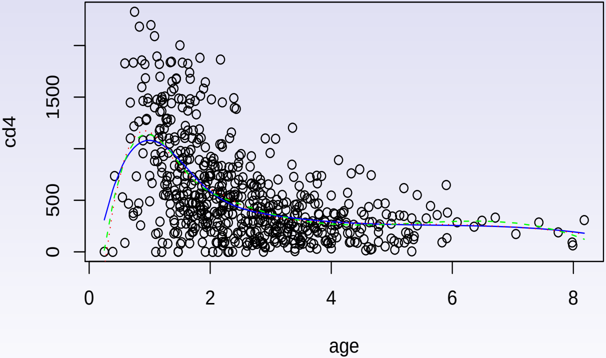
CD4 count data: piecewise polynomials GAIC(2)

	df	AIC
m7b	9	8959.519
m6b	8	8960.353
m8b	10	8961.073
m5b	7	8964.022
m4b	6	8977.475
m2b	5	8993.351
m3b	5	8993.351

CD4 count data: piecewise polynomials GAIC(log(n))

	df	AIC
m5b	7	8994.904
m6b	8	8995.648
m7b	9	8999.225
m4b	6	9003.946
m8b	10	9005.191
m2b	5	9015.410
m3b	5	9015.410

CD4 count data:piecewise polynomials fits



gamlss

CD4 count data: cubic smoothing splines

```
fn <- function(p) AIC(gamlss(cd4 ~ cs(age, df = p[1]),
        data = CD4, trace = FALSE), k = 2)
opAIC <- optim(par = c(3), fn, method = "L-BFGS-B",
        lower = c(1), upper = c(15))
fn <- function(p) AIC(gamlss(cd4 ~ cs(age, df = p[1]),
        data = CD4, trace = FALSE), k = log(length(CD4$age)))
opSBC <- optim(par = c(3), fn, method = "L-BFGS-B",
        lower = c(1), upper = c(15))
opAIC$par
[1] 10.85157
opSBC$par
[1] 1.854689
```

effective df should be between 1.8 and 10.8

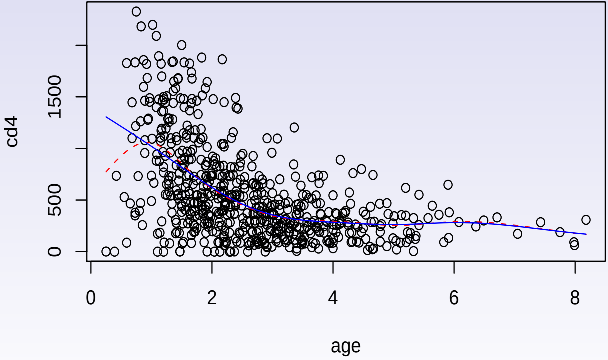
CD4 count data: cubic P-splines

```
mpb1 <- gamlss(cd4 ~ pb(age), data = CD4)
mpb1$mu.df}
[1] 8.606468
```

CD4 count data: Modelling sigma using cubic P-splines

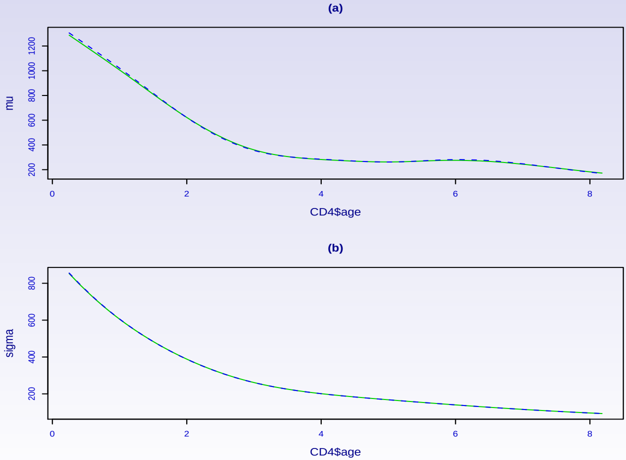
```
mpb2 <- gamlss(cd4 ~ pb(age), sigma.fo = ~pb(age),
     data = CD4, gd.tol = 10)
 mpb2$mu.df
[1] 6.200118    compare to 5.72 for cs()
mpb2$sigma.df   compare to 3.81 for cs()
[1] 3.803651
```

CD4 count data: P-spline fits



gamlss

CD4 count data: Comparing P-splines with Cubic splines

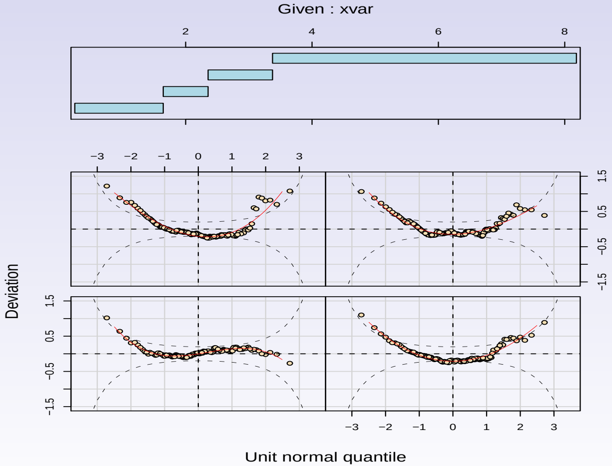


gamlss

CD4 count data: Validation Global Deviance

```
set.seed(1234)
rSample6040 <- sample(2, length(CD4$cd4),replace=T,
               prob=c(0.6, 0.4))
fn <- function(p) VGD(cd4~cs(age,df=p[1]),
                  sigma.fo=~cs(age,df=p[2]),
                  data=CD4, rand=rSample6040)
op<-optim(par=c(3,1), fn, method="L-BFGS-B", lower=c(1,1),
                 upper=c(10,10))
op$par
[1] 4.779947 1.376534
```

CD4 count data: Worm plot



gamlss

CD4 count data: Fitting different distributions

distributions	df	AIC
SEP2	8.86	8685.69
SEP1	8.77	8688.08
SEP3	11.24	8693.68
ST2	9.31	8693.91
ST1	9.31	8693.91
ST3	11.33	8694.21
JSU	13.28	8701.05
SHASH	12.75	8705.26
ST5	8.99	8724.25
ST4	12.54	8749.01
TF	11.48	8785.20
PE	11.17	8790.16

END

for more information see

www.gamlss.org

gamlss