Package 'mixSSG'

August 22, 2022

Type Package

Title Clustering Using Mixtures of Sub Gaussian Stable Distributions

Date 2022-08-19

Author Mahdi Teimouri [aut, cre, cph, ctb]
(<https://orcid.org/0000-0002-5371-9364>)

Maintainer Mahdi Teimouri <teimouri@aut.ac.ir>

Description Developed for model-based clustering using the finite mixtures of skewed sub-Gaussian stable distributions developed by Teimouri (2022) <arXiv:2205.14067>.

Encoding UTF-8

License GPL (>= 2)

Depends R(>= 3.4.3)

Imports ars, MASS, rootSolve

Repository CRAN

Version 1.1.1

NeedsCompilation no

Date/Publication 2022-08-22 14:40:07 UTC

R topics documented:

AIS *AIS data*

Description

The set of AIS data involves recorded body factors of 202 athletes including 100 women 102 men, see Cook (2009). Among factors, two variables body mass index (BMI) and body fat percentage (Bfat) are chosen for cluster analysis.

Usage

```
data(AIS)
```

Format

A text file with 3 columns.

References

R. D. Cook and S. Weisberg, (2009).*An Introduction to Regression Graphics*, John Wiley & Sons, New York.

Examples

```
data(AIS)
```

bankruptcy *bankruptcy data*

Description

The bankruptcy dataset involves ratio of the retained earnings (RE) to the total assets, and the ratio of earnings before interests and the taxes (EBIT) to the total assets of 66 American firms, see Altman (1969).

Usage

```
data(bankruptcy)
```

Format

A text file with 3 columns.

References

E. I. Altman, 1969. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, 23(4), 589-609.

Examples

```
data(bankruptcy)
```

`dssg` *Approximating the density function of skewed sub-Gaussian α-stable distribution.*

Description

Suppose d-dimensional random vector $\boldsymbol{Y}$ follows a skewed sub-Gaussian α-stable distribution with density function $f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\Theta})$ for $\boldsymbol{\Theta} = (\alpha, \boldsymbol{\mu}, \Sigma, \boldsymbol{\lambda})$ where α, $\boldsymbol{\mu}$, Σ, and $\boldsymbol{\lambda}$ are tail thickness, location, dispersion matrix, and skewness parameters, respectively. Herein, we give a good approximation for $f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\Theta})$. First , for $\mathcal{N} = 50$, define

$$L = \frac{\Gamma(\frac{\mathcal{N}\alpha}{2} + 1 + \frac{\alpha}{2})\Gamma\big(\frac{d+\mathcal{N}\alpha}{2} + \frac{\alpha}{2}\big)}{\Gamma(\frac{\mathcal{N}\alpha}{2} + 1)\Gamma\big(\frac{d+\mathcal{N}\alpha}{2}\big)(\mathcal{N}+1)}.$$

If $d(\boldsymbol{y}) \leq 2L^{\frac{2}{\alpha}}$, then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\Theta}) \simeq \frac{C_0\sqrt{2\pi\delta}}{N}\sum_{i=1}^{N}\exp\Big\{-\frac{d(\boldsymbol{y})}{2p_i}\Big\}\Phi\big(m|0,\sqrt{\delta p_i}\big)p_i^{-\frac{d}{2}},$$

where, $p_1, p_2, \cdots, p_N$ (for $N = 3000$) are independent realizations following positive α-stable distribution that are generated using command `rpstable(3000, alpha)`. Otherwise, if $d(\boldsymbol{y}) > 2L^{\frac{2}{\alpha}}$, we have

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\Theta}) \simeq \frac{C_0\sqrt{d(\boldsymbol{y})\delta}}{\sqrt{\pi}}\sum_{j=1}^{\mathcal{N}}\frac{(-1)^{j-1}\Gamma(\frac{j\alpha}{2}+1)\sin\big(\frac{j\pi\alpha}{2}\big)}{\Gamma(j+1)\big[\frac{d(\boldsymbol{y})}{2}\big]^{\frac{d+1+j\alpha}{2}}}\Gamma\Big(\frac{d+j\alpha}{2}\Big)T_{d+j\alpha}\bigg(m\sqrt{\frac{d+j\alpha}{d(\boldsymbol{y})\delta}}\bigg),$$

where $T_\nu(x)$ and $\Phi(x|a,b)$ denote the distribution function of the Student's t (distribution)with ν degrees of freedom) and normal (mean a and standard deviation b) distributions at point x, respectively, and

$C_0 = 2(2\pi)^{-\frac{d+1}{2}}|\Sigma|^{-\frac{1}{2}}$, $d(\boldsymbol{y}) = (\boldsymbol{y}-\boldsymbol{\mu})'\Omega^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$, $m = \boldsymbol{\lambda}'\Omega^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$, $\Omega = \Sigma + \boldsymbol{\lambda}\boldsymbol{\lambda}'$, $\delta = 1 - \boldsymbol{\lambda}'\Omega^{-1}\boldsymbol{\lambda}$.

Usage

```
dssg(Y, alpha, Mu, Sigma, Lambda)
```

Arguments

`Y`	a vector (or an $n \times d$ matrix) at which the density function is approximated.
`alpha`	tail thickness parameter.
`Mu`	a vector giving the location parameter.
`Sigma`	a positive definite symmetric matrix specifying the dispersion matrix.
`Lambda`	a vector giving the skewness parameter.

Value

simulated realizations of size n from positive α-stable distribution.

Author(s)

Mahdi Teimouri

Examples

```
n <- 4
alpha <- 1.4
Mu <- rep(0, 2)
Sigma <- diag(2)
Lambda <- rep(2, 2)
Y <- rssg(n, alpha, Mu, Sigma, Lambda)
dssg(Y, alpha, Mu, Sigma, Lambda)
```

`fitmssg` *Computing the maximum likelihood estimator for the mixtures of skewed sub-Gaussian α-stable distributions using the EM algorithm.*

Description

Each d-dimensional skewed sub-Gaussian α-stable (SSG) random vector $\mathbf{Y}$, admits the representation given by (Teimouri (2022)):

$$\mathbf{Y} \stackrel{d}{=} \boldsymbol{\mu} + \sqrt{P}\boldsymbol{\lambda}|Z_0| + \sqrt{P}\Sigma^{\frac{1}{2}}\mathbf{Z}_1,$$

where $\boldsymbol{\mu}$ (location vector in R^d, $\boldsymbol{\lambda}$ (skewness vector in R^d), Σ (positive definite symmetric dispersion matrix), and $0 < \alpha \leq 2$ (tail thickness) are model parameters. Furthermore, P is a positive α-stable random variable, $Z_0 \sim N(0,1)$, and $\mathbf{Z_1} \sim \mathbf{N_d}\big(\mathbf{0}, \mathbf{\Sigma}\big)$. We note that Z, Z_0, and $\boldsymbol{Z}_1$ are mutually independent.

Usage

```
fitmssg(Y, K, eps = 0.15, initial = "FALSE", method = "moment", starts = starts)
```

Arguments

`Y`	an $n \times d$ matrix of observations.
`K`	number of component.
`eps`	threshold value for stopping EM algorithm. It is 0.15 by default. The algorithm can be implemented faster if `eps` is larger.
`initial`	logical statement. If `initial = TRUE`, then a list of the initial values must be given. Otherwise it is determined by `method`.

`method` either `em` or `moment`. If `method = "moment"`, then the initial values are determined through the method of moment applied to each of K clusters that are obtained through the k-means method of Hartigan and Wong (1979). Otherwise, the initial values for each cluster are determined through the EM algorithm (Teimouri et al., 2018) developed for sub-Gaussian α-stable distributions applied to each of K clusters.

`starts` a list of initial values if `initial="TRUE"`. The list contains a vector of length K of mixing (weight) parameters, a vector of length K of tail thickness parameters, K vectors of length of d of location parameters, K dispersion matrices, K vectors of length of d of skewness parameters, respectively.

Value

a list of estimated parameters corresponding to K clusters, predicted labels for clusters, the log-likelihood value across iterations, the Bayesian information criterion (BIC), and the Akaike information criterion (AIC).

Author(s)

Mahdi Teimouri

References

M. Teimouri, 2022. Finite mixture of skewed sub-Gaussian stable distributions. https://arxiv.org/abs/2205.14067.

M. Teimouri, S. Rezakhah, and A. Mohammadpour, 2018. Parameter estimation using the em algorithm for symmetric stable random variables and sub-Gaussian random vectors, *Journal of Statistical Theory and Applications*, 17(3), 439-41.

J. A. Hartigan, M. A. Wong, 1979. Algorithm as 136: A k-means clustering algorithm, *Journal of the Royal Statistical Society. Series c (Applied Statistics)*, 28, 100-108.

Examples

```
data(bankruptcy)
fitmssg(bankruptcy[, 2:3], K = 2, eps = 0.15, initial = "FALSE", method = "moment", starts = starts)
```

---

rpstable *Simulating positive α-stable random variable.*

Description

The cumulative distribution function of positive α-stable distribution is given by

$$F_P(x) = \frac{1}{\pi}\int_0^{\pi} \exp\Big\{-x^{-\frac{\alpha}{2-\alpha}}a(\theta)\Big\}d\theta,$$

where $0 < \alpha \leq 2$ is tail thickness or index of stability and

$$a(\theta) = \frac{\sin\Big(\big(1-\frac{\alpha}{2}\big)\theta\Big)\Big[\sin\big(\frac{\alpha\theta}{2}\big)\Big]^{\frac{\alpha}{2-\alpha}}}{[\sin(\theta)]^{\frac{2}{2-\alpha}}}.$$

Kanter (1975) used the above integral transform to simulate positive α-stable random variable as

$$P \stackrel{d}{=} \Big(\frac{a(\theta)}{W}\Big)^{\frac{2-\alpha}{\alpha}},$$

in which $\theta \sim U(0,\pi)$ and W independently follows an exponential distribution with mean unity.

Usage

```
rpstable(n, alpha)
```

Arguments

`n`	the number of samples required.
`alpha`	tail thickness parameter.

Value

simulated realizations of size n from positive α-stable distribution.

Author(s)

Mahdi Teimouri

References

M. Kanter, 1975. Stable densities under change of scale and total variation inequalities, *Annals of Probability*, 3(4), 697-707.

Examples

```
rpstable(10, alpha = 1.2)
```

`rssg` *Simulating skewed sub-Gaussian α-stable random vector.*

Description

Each skewed sub-Gaussian α-stable (SSG) random vector $\mathbf{Y}$, admits the representation

$$\mathbf{Y} \stackrel{d}{=} \boldsymbol{\mu} + \sqrt{P}\boldsymbol{\lambda}|Z_0| + \sqrt{P}\Sigma^{\frac{1}{2}}\mathbf{Z}_1,$$

where $\boldsymbol{\mu} \in R^d$ is location vector, $\boldsymbol{\lambda} \in R^d$ is skewness vector, Σ is a positive definite symmetric dispersion matrix, and $0 < \alpha \leq 2$ is tail thickness. Further, P is a positive α-stable random variable, $Z_0 \sim N(0,1)$, and $\mathbf{Z}_1 \sim N_d\big(\mathbf{0}, \Sigma\big)$. We note that Z, Z_0, and $\mathbf{Z}_1$ are mutually independent.

Usage

```
rssg(n, alpha, Mu, Sigma, Lambda)
```

Arguments

n	the number of samples required.
alpha	tail thickness parameter.
Mu	a vector giving the location parameter.
Sigma	a positive definite symmetric matrix specifying the dispersion matrix.
Lambda	a vector giving the skewness parameter.

Value

simulated realizations of size n from the skewed sub-Gaussian α-stable distribution.

Author(s)

Mahdi Teimouri

Examples

```
n <- 4
alpha <- 1.4
Mu <- rep(0, 2)
Sigma <- diag(2)
Lambda <- rep(2, 2)
rssg(n, alpha, Mu, Sigma, Lambda)
```

---

stoch *Estimating the tail index of the skewed sub-Gaussian α-stable distribution using the stochastic EM algorithm given that other parameters are known.*

Description

Suppose $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_n$ are realizations following d-dimensional skewed sub-Gaussian α-stable distribution. Herein, we estimate the tail thickness parameter $0 < \alpha \leq 2$ when $\boldsymbol{\mu}$ (location vector in R^d, $\boldsymbol{\lambda}$ (skewness vector in R^d), and Σ (positive definite symmetric dispersion matrix are assumed to be known.

Usage

```
stoch(Y, alpha0, Mu0, Sigma0, Lambda0)
```

Arguments

Y	a vector (or an $n \times d$ matrix) at which the density function is approximated.
alpha0	initial value for tail thickness parameter.
Mu0	a vector giving the initial value for location parameter.
Sigma0	a positive definite symmetric matrix specifying the initial value for dispersion matrix.
Lambda0	a vector giving the initial value for skewness parameter.

Details

Here, we assume that parameters $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$, and Σ are known and only the tail thickness parameter needs to be estimated.

Value

Estimated tail thickness parameter α of skewed sub-Gaussian α-stable distribution.

Author(s)

Mahdi Teimouri

Examples

```
n <- 100
alpha <- 1.4
Mu <- rep(0, 2)
Sigma <- diag(2)
Lambda <- rep(2, 2)
Y <- rssg(n, alpha, Mu, Sigma, Lambda)
stoch(Y, alpha, Mu, Sigma, Lambda)
```

Index