

Package ‘AdaptGauss’

June 30, 2016

Type Package

Title Gaussian Mixture Models (GMM)

Version 1.2.4

Date 2016-06-30

Author Michael Thrun, Onno Hansen-
Goos, Rabea Griese, Catharina Lippmann, Jorn Lotsch, Alfred Ultsch

Maintainer Michael Thrun <mthrun@mathematik.uni-marburg.de>

Description Multimodal distributions can be modelled as a mixture of components. The model is derived using the Pareto Density Estimation (PDE) for an estimation of the pdf. PDE has been designed in particular to identify groups/classes in a dataset. Precise limits for the classes can be calculated using the theorem of Bayes. Verification of the model is possible by QQ plot, Chi-squared test and Kolmogorov-Smirnov test.

Imports shiny, caTools, mclust, methods, ggplot2, grid

Depends R (>= 2.10)

License GPL-3

LazyLoad yes

URL <https://www.uni-marburg.de/fb12/datenbionik/software-en>

NeedsCompilation no

Repository CRAN

Date/Publication 2016-06-30 18:16:06

R topics documented:

AdaptGauss-package	2
AdaptGauss	3
Bayes4Mixtures	4
BayesDecisionBoundaries	6
CDFMixtures	7
Chi2testMixtures	8
ClassifyByDecisionBoundaries	9
EMGauss	10

InformationCriteria4GMM	11
Intersect2Mixtures	12
KStestMixtures	13
LikelihoodRatio4Mixtures	14
LogLikelihood4Mixtures	15
OptimalNoBins	16
ParetoDensityEstimation	17
ParetoRadius	18
Pdf4Mixtures	19
PlotMixtures	20
PlotMixturesAndBoundaries	22
QQplotGMM	23
RandomLogGMM	24

Index	26
--------------	-----------

AdaptGauss-package	<i>AdaptGauss-package</i>
--------------------	---------------------------

Description

Multimodal distributions can be modelled as a mixture of components. The model is derived using the Pareto Density Estimation (PDE) for an estimation of the pdf [Ultsch 2005]. PDE has been designed in particular to identify groups/classes in a dataset. The expectation maximization algorithm estimates a Gaussian mixture model of density states [Bishop 2006] and the limits between the different states are defined by Bayes decision boundaries [Duda 2001]. The model can be verified with Chi-squared test, Kolmogorov-Smirnov test and QQ plot.

The correct number of modes may be found with AIC or BIC.

Author(s)

Michael Thrun, Onno Hansen-Goos, Rabea Griese, Catharina Lippmann, Jorn Lotsch, Alfred Ultsch
 Maintainer: Michael Thrun <mthrun@mathematik.uni-marburg.de>

References

- Ultsch, A., Thrun, M.C., Hansen-Goos, O., Loetsch, J.: Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox(AdaptGauss), International Journal of Molecular Sciences, doi:10.3390/ijms161025897, 2015.
- Duda, R.O., P.E. Hart, and D.G. Stork, Pattern classification. 2nd. Edition. New York, 2001, p 512 ff
- Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006, p 435 ff
- Ultsch, A.: Pareto density estimation: A density estimation for knowledge discover, in Baier, D.; Werrnecke, K. D., (Eds), Innovations in classification, data science, and information systems, Proc Gfkl 2003, pp 91-100, Springer, Berlin, 2005.
- Thrun M.C., Ultsch, A.: Models of Income Distributions for Knowledge Discovery, European Conference on Data Analysis, DOI 10.13140/RG.2.1.4463.0244, Colchester 2015.

AdaptGauss

*Adapt Gaussian Mixture Model (GMM)***Description**

Adapt interactively a Gaussians Mixture Model GMM to the empirical PDF of the data (generated by ParetoDensityEstimation) such that $N(\text{Means}, \text{SDs}) * \text{Weights}$ is a model for Data

Usage

```
AdaptGauss(Data, Means, SDs, Weights, ParetoRadius, LB, HB, ListOfAdaptGauss, fast)
```

Arguments

Data	Data for empirical PDF. Has to be an Array of values. NaNs and NULLs will be deleted
Means	Optional: Means of gaussians of GMM.
SDs	Optional: StandardDeviations of gaussians of GMM. (Has to be the same length as Means)
Weights	Optional: Weights of gaussians of GMM. (Has to be the same length as Means)
ParetoRadius	Optional: Pareto Radius of Pareto Desity Estimation (PDE).
LB	Optional: Low boundary of estimation. All values below LB will be deleted. Default: min(Data)
HB	Optional: High boundary of estimation. All values above HB will be deleted. Default: max(Data)
ListOfAdaptGauss	Optional: If editing of an existing Model is the goal, enables to give the Output of AdaptGaus as the Input of AdaptGauss() instead of setting Means, SDs and Weights separately
fast	Default=TRUE; FALSE: Using mclust's EM see function densityMclust of that package, TRUE: Naive but faster EM implementation, which may be numerical unstable, because log(gauss) is not used

Details

Data: maximum length is 10000. If larger, Data will be randomly reduced to 10000 Elements. MeansIn/DeviationsIn/WeightsIN: If empty, either one or three Gaussian's are generated by kmeans algorithm. Pareto Radius: If empty: will be generated by ParetoDensityEstimation RMS: Root Mean Square error is normalized by RMS of Gaussian's with Mean=mean(data) and SD=sd(data), see [Ultsch et.al., 2015] for further details.

Value

List with	
Means	Means of Gaussian's.
SDs	Standard SDs of Gaussian's.
Weights	Weights of Gaussian's.
ParetoRadius	Pareto Radius: Either ParetoRadiusIn, the pareto radius enenerated by PretoDensityEstimation(if no Pareto Radius in Input).
RMS	Root Mean Square of Deviation between Gaussian Mixture Model GMM to the empirical PDF. Normalized by RMS of one Gaussian with mean=meanrobust(data) and sdev=stdrobust(data). Further Details in [Ultsch et al 2015]
BayesBoundaries	vector[1:L-1], Bayes decision boundaries

Author(s)

Onno Hansen-Goos, Michael Thrun

References

Ultsch, A., Thrun, M.C., Hansen-Goos, O., Loetsch, J.: Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox(AdaptGauss), International Journal of Molecular Sciences, doi:10.3390/ijms161025897, 2015.

Thrun M.C., Ultsch, A.: Models of Income Distributions for Knowledge Discovery, European Conference on Data Analysis, DOI 10.13140/RG.2.1.4463.0244, Colchester 2015.

Examples

```
data1=c(rnorm(1000))
## Not run: Vals1=AdaptGauss(data1)

data2=c(rnorm(1000),rnorm(2000)+2,rnorm(1000)*2-1)
## Not run: Vals2=AdaptGauss(data2,c(-1,0,2),c(2,1,1),c(0.25,0.25,0.5),0.3,-6,6)
```

Bayes4Mixtures

Posterioris of Bayes Theorem

Description

Calculates the posterioris of Bayes theorem

Usage

```
Bayes4Mixtures(Data, Means, SDs, Weights, IsLogDistribution,
  PlotIt, CorrectBorders, Color, xlab, lwd)
```

Arguments

Data	vector (1:N) of data points
Means	vector[1:L] of Means of Gaussians (of GMM), L == Number of Gaussians
SDs	vector of standard deviations, estimated Gaussian Kernels, has to be the same length as Means
Weights	vector of relative number of points in Gaussians (prior probabilities), has to be the same length as Means
IsLogDistribution	Optional, ==1 if distribution(i) is a LogNormal, default vector of zeros of length L
PlotIt	Optional, Default: FALSE; TRUE do a Plot
CorrectBorders	Optional, ==TRUE data at right borders of GMM distribution will be assigned to last gaussian, left border vice versa. (default ==FALSE) normal Bayes Theorem
Color	Optional, character vector of colors, default rainbow()
xlab	Optional, label of x-axis, default 'Data', see intern R documentation
lwd	Width of Line, see intern R documentation

Details

See conference presentation for further explanation.

Value

List with

Posteriors	(1:N,1:L) of Posteriors corresponding to Data
NormalizationFactor	(1:N) denominator of Bayes theorem corresponding to Data

Author(s)

Catharina Lippmann, Onno Hansen-Goos, Michael Thrun

References

Thrun M.C., Utsch, A.: Models of Income Distributions for Knowledge Discovery, European Conference on Data Analysis, DOI 10.13140/RG.2.1.4463.0244, Colchester 2015.

See Also

[BayesDecisionBoundaries](#), [AdaptGauss](#)

 BayesDecisionBoundaries

Decision Boundaries calculated through Bayes Theorem

Description

Function finds the intersections of Gaussians or LogNormals

Usage

BayesDecisionBoundaries(Means,SDs,Weights,IsLogDistribution,MinData,MaxData,Ycoor)

Arguments

Means	vector[1:L] of Means of Gaussians (of GMM)
SDs	vector of standard deviations, estimated Gaussian Kernels, has to be the same length as Means
Weights	vector of relative number of points in Gaussians (prior probabilities), has to be the same length as Means
IsLogDistribution	Optional, ==1 if distribution(i) is a LogNormal, default vector of zeros of length 1:L
MinData	Optional, Beginning of range, where the Boundaries are searched for, default min(M)
MaxData	Optional, End of range, where the Boundaries are searched for, default max(M)
Ycoor	Optional, Bool, if TRUE instead of vector of DecisionBoundaries list of DecisionBoundaries and DBY is returned

Value

DecisionBoundaries	vector[1:L-1], Bayes decision boundaries
DBY	if (Ycoor==TRUE), y values at the cross points of the Gaussians is also returned, that the return is a list of DecisionBoundaries and DBY

Author(s)

Michael Thrun, Rabea Griese

References

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. 2nd. Edition. New York, p. 512ff

See Also

[AdaptGauss](#),[Intersect2Mixtures](#),[Bayes4Mixtures](#)

Chi2testMixtures *Pearson's chi-squared goodness of fit test*

Description

Chi2testMixtures is goodness of fit test which establishes whether an observed distribution (data) differs from a Gauss Mixture Model (GMM). Returns a P value of a special case of a chi-square test and visualizes data versus a given GMM.

Usage

Chi2testMixtures(Data,Means,SDs,Weights,IsLogDistribution,PlotIt,UpperLimit,VarName)

Arguments

Data	vector of data points (1:n)
Means	vector of Means of Gaussians (1:c)
SDs	vector of standard deviations, estimated Gaussian Kernels (1:c)
Weights	vector of relative number of points in Gaussians (prior probabilities) (1:c)
IsLogDistribution	Optional, if IsLogDistribution(i)==1, then mixture is lognormal, default vector of zeros of length 1:L
PlotIt	Optional, Default: FALSE, do a Plot of the compared cdfs and the KS-test distribution (Diff)
UpperLimit	Optional. test only for Data <= UpperLimit, Default = max(Data) i.e all Data.
VarName	If PlotIt=TRUE, the name of the inspected variable, default 'Data'

Details

The null hypothesis is that the estimated data distribution differs significantly from the GMM. Let O_i be the observed features and E_i be the expected number E , than the test statistic is defined with the minimum chi-square estimate $T = \sum((O_i - E_i)^2 / E_i) * 1/m$, where m the number of data points. The expected number E_i may be derived for each bin. If there is a significant difference between the O_i and the E_i , the Pvalue is large and the null hypothesis cannot be rejected.

Further details, see [Thrun & Ullsch, 2015].

Value

List with	
Pvalue	Pvalue of a suiting chi-square , Pvalue ==0 if Pvalue <0.001
BinCenters	bin centers
ObsNrInBin	No. of data in bin
ExpectedNrInBin	No. of data that should be in bin according to GMM

Chi2Value the TestStatistic T i.e.: $\sum((\text{ObsNrInBin}(\text{Ind}) - \text{ExpectedNrInBin}(\text{Ind}))^2 / \text{ExpectedNrInBin}(\text{Ind}))$
with $\text{Ind} = \text{find}(\text{ExpectedNrInBin} \geq 10)$ The value of Chi2Value is compared to a chi-squared distribution.

Note

The statistic assumption is that the the test statistic follows a chi square distribution. The number of degrees of freedom is equal to the number of datapoints $n - 1 - 3 * c$

Author(s)

Rabea Griese, Michael Thrun

References

Hartung, J., Elpelt, B., and Kloesener, K.H.: Statistik, 8. Aufl. Verlag Oldenburg (1991).
Thrun, M. C., Ultsch, A.: Models of Income Distributions for Knowledge Discovery, European Conference on Data Analysis, DOI 10.13140/RG.2.1.4463.0244, pp. 28-29, Colchester 2015.

ClassifyByDecisionBoundaries

Classify Data according to decision Boundaries

Description

The Decision Boundaries calculated through Bayes Theorem.

Usage

ClassifyByDecisionBoundaries(Data, DecisionBoundaries, ClassLabels)

Arguments

Data vector of Data
DecisionBoundaries decision boundaries, [BayesDecisionBoundaries](#)
ClassLabels Optional numbered class labels that are assigned to the classes. default (1:L), L number of different components of gaussian mixture model

Value

Cls(1:n,1:d) classification of Data, such that 1= first component of gaussian mixture model, 2= second component of gaussian mixture model and so on. For Every datapoint a number is returned.

Author(s)

Michael Thrun

References

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. 2nd. Edition. New York, p. 512ff

See Also

[BayesDecisionBoundaries](#), [Bayes4Mixtures](#)

 EMGauss

EM Algorithm for GMM

Description

Expectation-Maximization algorithm to calculate optimal Gaussian Mixture Model for given data in one Dimension.

Usage

```
EMGauss(Data, Means, SDs,Weights, MaxNumberOfIterations,fast)
```

Arguments

Data	vector of data points
Means	vector(1:L), Means of Gaussians, L == Number of Gaussians
SDs	estimated Gaussian Kernels = standard deviations
Weights	optional, relative number of points in Gaussians (prior probabilities): sum(Weights) ==1, default weight is 1/L
MaxNumberOfIterations	Optional, Number of Iterations; default=10
fast	Default: FALSE: Using mclust's EM see function densityMclust of that package, TRUE: Naive but faster EM implementation, which may be numerical unstable, because log(gauss) is not used

Details

No adding or removing of Gaussian kernels. Number of Gaussian hast to be set by the length of the vector of Means, SDs and Weights. This EM is only for univariate data. For multivariate data see package mclust

Value

List with	
Means	means of GMM generated by EM algorithm
SDs	standard deviations of GMM generated by EM algorithm
Weights	prior probabilities of Gaussians

Author(s)

Onno Hansen-Goos, Michael Thrun

References

Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006, p 435 ff

See Also

[AdaptGauss](#)

InformationCriteria4GMM

Information Criteria For GMM

Description

Calculates the AIC and BIC criteria

Usage

InformationCriteria4GMM(Data, Means, SDs, Weights, IsLogDistribution)

Arguments

Data	vector (1:N) of data points
Means	vector[1:L] of Means of Gaussians (of GMM), L == Number of Gaussians
SDs	vector of standard deviations, estimated Gaussian Kernels, has to be the same length as Means
Weights	vector of relative number of points in Gaussians (prior probabilities), has to be the same length as Means
IsLogDistribution	Optional, ==1 if distribution(i) is a LogNormal, default vector of zeros of length L, LogNormal Modes are at this point only experimental

Details

$AIC = 2*k - 2*\text{LogLikelihood}$, $k = \text{nr. of model parameter} = 3*\text{Nr. of Gaussians}$ One Gaussian: $K=2$ (Weight is then not an parameter!) SMALL SAMPLE CORRECTION: for $n = \text{nr of Data}$ and $n < 40 * k$, AIC is adjusted to $AIC = AIC + (2*k*(k+1))/(n-k-1)$

$BIC = k * \log(n) - 2*\text{LogLikelihood}$

Only for a Gaussian Mixture Model (GMM) verified, for the Log Gaussian, Gaussian, Log Gaussian (LGL) Model only experimental

Value

List with	
K	Number of gaussian mixtures
AIC	Akaike Informations criterium
BIC	Bayes Information criterium
LogLikelihood	LogLikelihood of GMM, see LogLikelihood4Mixtures
PDFmixture	probability density function of GMM, see Pdf4Mixtures
LogPDFdata	log(PDFmixture)

Author(s)

Michael Thrun

References

- Aubert, A. H., Thrun, M. C., Breuer, L., & Ultsch, A.: Knowledge discovery from data structure: hydrology versus biology controlled in-stream nitrate concentration, Scientific reports, Vol. (in revision), pp., 2016.
- Aho, K., Derryberry, D., & Peterson, T.: Model selection for ecologists: the worldviews of AIC and BIC. Ecology, 95(3), pp. 631-636, 2014.

Intersect2Mixtures *Intersect of two Gaussians*

Description

Finds the intersect of two gaussians or log gaussians

Usage

```
Intersect2Mixtures(Mean1,SD1,Weight1,Mean2,SD2,Weight2,IsLogDistribution,MinData,MaxData)
```

Arguments

Mean1	mean of 1.gaussian
SD1	standard deviations of 1.gaussian
Weight1	weight of 1. gaussian
Mean2	mean of 2.gaussian
SD2	standard deviations of 2.gaussian
Weight2	weight of 2. gaussian
IsLogDistribution	Optional, ==1 if distribution(i) is a LogNormal, default vector of zeros of length 2
MinData	Optional, Beginning of range, where the intersect is searched for, default min(Mean1,Mean2)
MaxData	Optional, End of range, where the intersect is searched for, default max(Mean1,Mean2)

Value

CutX x value, where gaussian 1=gaussian2
 CutY y value, where gaussian 1=gaussian2

Author(s)

Michael Thrun, Rabea Griese

See Also

[BayesDecisionBoundaries](#)

KStestMixtures	<i>Kolmogorov-Smirnov test</i>
----------------	--------------------------------

Description

Returns a P value and visualizes for Kolmogorov-Smirnov test of Data versus a given Gauss Mixture Model

Usage

KStestMixtures(Data, Means, SDs, Weights, IsLogDistribution, PlotIt, UpperLimit, Silent)

Arguments

Data	vector of data points
Means	vector of Means of Gaussians
SDs	vector of standard deviations, estimated Gaussian Kernels
Weights	vector of relative number of points in Gaussians (prior probabilities)
IsLogDistribution	Optional, if IsLogDistribution(i)==1, then mixture is lognormal, default vector of zeros of length 1:L
PlotIt	Optional, Default: FALSE, do a Plot of the compared cdfs and the KS-test distribution (Diff)
UpperLimit	Optional. test only for Data <= UpperLimit, Default = max(Data) i.e all Data.
Silent	Optional, default=TRUE, If FALSE, shows progress of computation by points (On windows systems a progress bar)

Details

...

Value

List with

Pvalue Pvalue of a suiting Kolmogorov-Smirnov test, Pvalue ==0 if Pvalue <0.001
 DataKernels such that plot(DataKernels,DataCDF) gives the cdf(Data)
 DataCDF such that plot(DataKernels,DataCDF) gives the cdf(Data)
 CDFGaussMixture
 No. of data that should be in bin according to GMM

Author(s)

Michael Thrun, Alfred Ultsch

References

Smirnov, N., Table for Estimating the Goodness of Fit of Empirical Distributions. 1948, (2), 279-281.

LikelihoodRatio4Mixtures

Likelihood Ratio for Gaussian Mixtures

Description

Computes the likelihood ratio for two Gaussian Mixture Models.

Usage

LikelihoodRatio4Mixtures(Data,NullMixture,OneMixture,PlotIt,LowerLimit,UpperLimit)

Arguments

Data Data points.
 NullMixture A Matrix: cbind(Means0,SDs0,Weights0) or cbind(Means0,SDs0,Weights0,IsLog0).
 The null model; usually with less Gaussians than the OneMixture
 OneMixture A Matrix: cbind(Means1,SDs1,Weights1) or cbind(Means1,SDs1,Weights1,IsLog1).
 The alternative model usually with more Gaussians than the OneMixture.
 PlotIt Optional: zero or one. o a Plot of the compared cdf's and the KS-test distribution
 (Diff)
 LowerLimit Optional: test only for Data >= LowerLimit, Default = min(Data) i.e all Data.
 UpperLimit Optional: test only for Data <= UpperLimit, Default = max(Data) i.e all Data.

Value

List with	
Pvalue	the error that we make, if we accept OneMixture as the better Model over the NullMixture
NullLogLikelihood	log likelihood of GMM Null
OneLogLikelihood	log likelihood of GMM One

Author(s)

Alfred Ultsch, Michael Thrun, Catharina Lippmann

LogLikelihood4Mixtures

LogLikelihood for Gaussian Mixture Models

Description

Computes the LogLikelihood for Gaussian Mixture Models.

Usage

LogLikelihood4Mixtures(Data, Means, SDs, Weights, IsLogDistribution)

Arguments

Data	Data for empirical PDF. Has to be an Array of values. NaNs and NULLs will be deleted
Means	Optional: Means of gaussians of GMM.
SDs	Optional: StandardDeviations of gaussians of GMM. (Has to be the same length as Means)
Weights	Optional: Weights of gaussians of GMM. (Has to be the same length as Means)
IsLogDistribution	Optional, ==1 if distribution(i) is a LogNormal, default vector of zeros of length 1:L

Value

List with	
LogLikelihood	LogLikelihood = = sum(log(PDFmixture))
LogPDF	=log(PDFmixture)
PDFmixture	die Probability density function for each point

Author(s)

Alfred Ultsch, Catharina Lippmann

References

Pattern Recognition and Machine Learning, C.M. Bishop, 2006, isbn: ISBN-13: 978-0387-31073-2, p. 433 (9.14)

OptimalNoBins

Optimal Number Of Bins

Description

Calculation of the optimal number of bins for a histogram.

Usage

OptimalNoBins(Data)

Arguments

Data Data

Details

The bin width ist defined with $bw=3.49*\text{stdrobust}(1/(n)^{1/3})$

Value

optNrOfBins The best possible number of bins. Not less than 10 though

Note

OptimalNoBins() is a kernel density estimation for fixed intervals.

Author(s)

Alfred Ultsch, Michael Thrun

References

David W. Scott Jerome P. Keating: A Primer on Density Estimation for the Great Home Run Race of 98, STATS 25, 1999, pp 16-22.

See Also

[ParetoRadius](#)

Examples

```
Data = c(rnorm(1000),rnorm(2000)+2,rnorm(1000)*2-1)
optNrOfBins = OptimalNoBins(Data)
minData = min(Data,na.rm = TRUE)
maxData = max(Data,na.rm = TRUE)
i = maxData-minData
optBreaks = seq(minData, maxData, i/optNrOfBins) # bins in fixed intervals
hist(Data, breaks=optBreaks)
```

ParetoDensityEstimation

Pareto Density Estimation

Description

This function estimates the Pareto Density for the distribution of one variable.

Usage

```
ParetoDensityEstimation(Data,paretoRadius,kernels,MinAnzKernels)
```

Arguments

Data	numeric vector of data.
paretoRadius	Optional, numeric value, see ParetoRadius
kernels	Optional, numeric vector. data values where pareto density is measured at. If 0 (by default) kernels will be computed.
MinAnzKernels	Optional, minimal number of kernels, default MinAnzKernels==100

Details

Pareto Density Estimation (PDE) is a method for the estimation of probability density functions using hyperspheres. The Pareto-radius of the hyperspheres is derived from the optimization of information for minimal set size. It is shown, that Pareto Density is the best estimate for clusters of Gaussian structure. The method is shown to be robust when cluster overlap and when the variances differ across clusters.

Value

List with

kernels	numeric vector. data values at with Pareto Density is measured.
paretoDensity	numeric vector containing the determined density by ParetoRadius.
paretoRadius	numeric value.

Note

This is the best density estimation to judge Gaussian Mixtures of the data see [Ultsch 2003]

Author(s)

Michael Thrun

References

Ultsch, A.: Pareto density estimation: A density estimation for knowledge discover, in Baier, D.; Werrnecke, K. D., (Eds), Innovations in classification, data science, and information systems, Proc Gfkl 2003, pp 91-100, Springer, Berlin, 2005.

See Also

[ParetoRadius](#)

Examples

```
data = c(rnorm(1000),rnorm(2000)+2,rnorm(1000)*2-1)
pdeVal <- ParetoDensityEstimation(data)
plot(pdeVal$kernelns,pdeVal$paretoDensity,type='l',xaxs='i',
yaxs='i',xlab='Data',ylab='PDE')
```

ParetoRadius

ParetoRadius for distributions

Description

Calculation of the ParetoRadius i.e. the 18 percentiles of all mutual Euclidian distances in data.

Usage

```
ParetoRadius(Data,maximumNrSamples,plotDistancePercentiles)
```

Arguments

Data numeric data vector

maximumNrSamples

Optional, numeric. Maximum number for which the distance calculation can be done. 1000 by default.

plotDistancePercentiles

Optional, logical. If TRUE, a plot of the percentiles of distances is produced. FALSE by default.

Details

The Pareto-radius of the hyperspheres is derived from the optimization of information for minimal set size.

Value

numeric value, the Pareto radius.

Note

ParetoRadius() is a kernel density estimation for variable intervals. It works only on Data without missing values (NA) or NaN. In other cases, please use [ParetoDensityEstimation](#) directly.

Author(s)

Michael Thrun

References

Utsch, A.: Pareto density estimation: A density estimation for knowledge discover, in Baier, D.; Wernicke, K. D., (Eds), Innovations in classification, data science, and information systems, Proc Gfkl 2003, pp 91-100, Springer, Berlin, 2005.

See Also

[ParetoDensityEstimation](#), [OptimalNoBins](#)

Pdf4Mixtures

Calculates pdf for GMM

Description

Calculate Gaussianthe probability density function for a Mixture Model

Usage

```
Pdf4Mixtures(Data, Means, SDs, Weights,IsLogDistribution,PlotIt)
```

Arguments

Data	vector (1:N) of data points
Means	vector[1:L] of Means of Gaussians (of GMM),L == Number of Gaussians
SDs	vector of standard deviations, estimated Gaussian Kernels, has to be the same length as Means
Weights	vector of relative number of points in Gaussians (prior probabilities), has to be the same length as Means

IsLogDistribution Optional, ==1 if distribution(i) is a LogNormal, default vector of zeros of length 1:L

PlotIt Optional: =TRUE plot of pdf

Value

List with

PDF4modes matrix, where the columns are the gaussians

PDF matrix, where the columns are the gaussians weighted by Weights

PDFmixture linear superpositions of PDF - prior probabilities of Gaussians

Author(s)

Michael Thrun

See Also

[PlotMixtures](#)

Examples

```
data=c(rnorm(1000),rnorm(2000)+2,rnorm(1000)*2-1)
Pdf4Mixtures(data,c(-1,0,2),c(2,1,1),c(0.25,0.25,0.5), PlotIt=TRUE)
```

PlotMixtures

Shows GMM

Description

Plots Gaussian Mixture Model without Bayes decision boundaries, such that:

Black is the PDE of Data

Red is color of the GMM

Blue is the color of components of the mixture

Usage

```
PlotMixtures(Data, Means, SDs, Weights,IsLogDistribution, SingleColor,
MixtureColor,DataColor, SingleGausses,axes, xlab, ylab,xlim, ylim, ...)
```

Arguments

Data	vector (1:N) of data points
Means	vector[1:L] of Means of Gaussians (of GMM), L == Number of Gaussians
SDs	vector of standard deviations, estimated Gaussian Kernels, has to be the same length as Means
Weights	vector of relative number of points in Gaussians (prior probabilities), has to be the same length as Means
IsLogDistribution	Optional, ==1 if distribution(i) is a LogNormal, default vector of zeros of length 1:L
SingleColor	Optional, Color for line plot of all the single gaussians, default magenta
MixtureColor	Optional, Color of line lot for the mixture default red
DataColor	Optional, Color of line plot for the data, default black
axes	Optional, Default: TRUE with axis, see argument axis of plot
xlim	Optional, see plot
ylim	Optional, see plot
xlab	Optional, see plot
ylab	Optional, see plot
SingleGausses	Optional, If TRUE, single gaussians are shown, default FALSE
...	other plot arguments like xlim = c(1,10)

Author(s)

Michael Thrun

See Also

[PlotMixturesAndBoundaries](#)

Examples

```
data=c(rnorm(1000),rnorm(2000)+2,rnorm(1000)*2-1)
PlotMixtures(data,c(-1,0,2),c(2,1,1),c(0.25,0.25,0.5),SingleColor='blue',SingleGausses=TRUE)
```

 PlotMixturesAndBoundaries

Shows GMM with Boundaries

Description

Plots Gaussian Mixture Model with Bayes decision boundaries, such that:

Black is the PDE of Data

Red is color of the GMM

Magenta are the Bayes boundaries

Usage

```
PlotMixturesAndBoundaries(Data, Means, SDs, Weights, IsLogDistribution,
  SingleColor, MixtureColor, DataColor, BoundaryColor, xlabel, ylabel, ...)
```

Arguments

Data	vector (1:N) of data points
Means	vector[1:L] of Means of Gaussians (of GMM), L == Number of Gaussians
SDs	vector of standard deviations, estimated Gaussian Kernels, has to be the same length as Means
Weights	vector of relative number of points in Gaussians (prior probabilities), has to be the same length as Means
IsLogDistribution	Optional, ==1 if distribution(i) is a LogNormal, default vector of zeros of length 1:L
SingleColor	Optional, Color for line plot of all the single gaussians, default magenta
MixtureColor	Optional, Color of line plot for the mixture, default red
DataColor	Optional, Color of line plot for the data, default black
BoundaryColor	Optional, Color of bayesian boundaries
xlab	Optional, x label, see plot
ylab	Optional, y label, ee plot
...	Optional, see plot for plot properties and for SingleGausses PlotMixtures

Details

you may also set SingleGausses=T than components of the mixture in blue will be shown.

Author(s)

Michael Thrun

See Also

[BayesDecisionBoundaries](#), [PlotMixtures](#)

 QQplotGMM

Quantile Quantile Plot of Data

Description

Quantile Quantile plot of data against gaussian distribution mixture model with optional best-fit-line

Usage

```
QQplotGMM(Data, Means, SDs, Weights, IsLogDistribution, Line,
PlotSymbol, xug, xog, LineWidth, PointWidth, ylab, main, ...)
```

Arguments

Data	vector (1:N) of data points
Means	vector[1:L] of Means of Gaussians (of GMM), L == Number of Gaussians
SDs	vector of standard deviations, estimated Gaussian Kernels, has to be the same length as Means
Weights	vector of relative number of points in Gaussians (prior probabilities), has to be the same length as Means
IsLogDistribution	Optional, ==1 if distribution(i) is a LogNormal, default Zeros of Length L
Line	Optional, Default: TRUE=Regression Line is drawn
xug	Optional, lower limit of the interval [xug, xog], in which a line will be interpolated
xog	Optional, upper limit of the interval [xug, xog], in which a line will be interpolated
PlotSymbol	Optional, plot symbol. Default is 20.
LineWidth	Optional, width of regression line, if Line==TRUE
PointWidth	Optional, width of points
ylab	Optional, see plot
main	Optional, see plot
...	Note: xlab cannot be changed, other parameters see qqplot

Details

Only verified for a Gaussian Mixture Model, usage of IsLogDistribution for LogNormal Modes is experimental!

Value

List with

x The x coordinates of the points that were plotted
 y The original data vector, i.e., the corresponding y coordinates

Author(s)

Michael Thrun

ReferencesMichael, J. R. (1983). The stabilized probability plot. *Biometrika*, 70(1), 11-17.**See Also**[qqplot](#)**Examples**

```
data=c(rnorm(1000),rnorm(2000)+2,rnorm(1000)*2-1)
QQplotGMM(data,c(-1,0,2),c(2,1,1),c(0.25,0.25,0.5))
```

 RandomLogGMM

Random Number Generator for Log or Gaussian Mixture Model

Description

Function finds the intersections of Gaussians or LogNormals

Usage

```
RandomLogGMM(Means,SDs,Weights,IsLogDistribution,TotalNoPoints)
```

Arguments

Means vector[1:L] of Means of Gaussians (of GMM)
 SDs vector of standard deviations, estimated Gaussian Kernels, has to be the same length as Means
 Weights vector of relative number of points in Gaussians (prior probabilities), has to be the same length as Means
 IsLogDistribution Optional, ==1 if distribution(i) is a LogNormal, default vector of Zeros of Length L
 TotalNoPoints Optional, number of point for log or GMM generated

Value

Returns vector of [1:TotalNoPoints] of generated points for log oder gaussian mixture model

Author(s)

Alfred Ultsch, Michael Thrun, Rabea Griese

See Also

[QQplotGMM](#), [Chi2testMixtures](#)

Index

- *Topic **AIC**
 - InformationCriteria4GMM, 11
- *Topic **AdaptGauss**
 - AdaptGauss-package, 2
- *Topic **Akaike informations criterium**
 - InformationCriteria4GMM, 11
- *Topic **BIC**
 - InformationCriteria4GMM, 11
- *Topic **Bayes information criterium**
 - InformationCriteria4GMM, 11
- *Topic **BayesDecisionBoundaries**
 - BayesDecisionBoundaries, 6
- *Topic **Bayes**
 - Bayes4Mixtures, 4
 - BayesDecisionBoundaries, 6
 - PlotMixturesAndBoundaries, 22
- *Topic **Boundaries**
 - Bayes4Mixtures, 4
 - BayesDecisionBoundaries, 6
 - PlotMixturesAndBoundaries, 22
- *Topic **ClassifyByDecisionBoundaries**
 - ClassifyByDecisionBoundaries, 9
- *Topic **EM algorithm**
 - EMGauss, 10
- *Topic **EM**
 - AdaptGauss-package, 2
 - EMGauss, 10
- *Topic **Expectation-Maximization algorithm**
 - EMGauss, 10
- *Topic **Expectation-Maximization**
 - EMGauss, 10
- *Topic **Expectation**
 - EMGauss, 10
- *Topic **GMM**
 - AdaptGauss, 3
 - AdaptGauss-package, 2
 - Pdf4Mixtures, 19
 - PlotMixtures, 20
 - RandomLogGMM, 24
- *Topic **Maximization**
 - EMGauss, 10
- *Topic **Minimum chi-square estimation**
 - Chi2testMixtures, 8
- *Topic **MultiModal**
 - AdaptGauss, 3
- *Topic **Multimodal**
 - AdaptGauss-package, 2
- *Topic **Pearson's chi-squared test**
 - Chi2testMixtures, 8
- *Topic **best-fit-line**
 - QQplotGMM, 23
- *Topic **bins**
 - OptimalNoBins, 16
- *Topic **chi-square estimation**
 - Chi2testMixtures, 8
- *Topic **chi-square goodness-of-fit**
 - Chi2testMixtures, 8
- *Topic **chi-square test for independence**
 - Chi2testMixtures, 8
- *Topic **chi-squared test**
 - Chi2testMixtures, 8
- *Topic **chi-square**
 - Chi2testMixtures, 8
- *Topic **expectation maximization**
 - AdaptGauss-package, 2
- *Topic **gaussian mixture model**
 - AdaptGauss, 3
 - AdaptGauss-package, 2
 - Pdf4Mixtures, 19
 - PlotMixtures, 20
- *Topic **histogram**
 - OptimalNoBins, 16
- *Topic **kernel density estimation**
 - OptimalNoBins, 16
 - ParetoRadius, 18

- *Topic **log GMM**
 - RandomLogGMM, 24
 - *Topic **mixture of components**
 - AdaptGauss-package, 2
 - *Topic **mixture**
 - AdaptGauss, 3
 - AdaptGauss-package, 2
 - *Topic **pareto density estimation**
 - AdaptGauss-package, 2
 - *Topic **pdf**
 - AdaptGauss-package, 2
 - Pdf4Mixtures, 19
 - *Topic **plot**
 - QQplotGMM, 23
 - *Topic **posterioris**
 - Bayes4Mixtures, 4
 - *Topic **posterior**
 - Bayes4Mixtures, 4
 - *Topic **probability density function**
 - Pdf4Mixtures, 19
 - *Topic **qq-plot**
 - QQplotGMM, 23
 - *Topic **qqplot**
 - QQplotGMM, 23
 - *Topic **quantile/quantile-plot**
 - QQplotGMM, 23
- AdaptGauss, 3, 5, 6, 11
- AdaptGauss-package, 2
- Bayes4Mixtures, 4, 6, 10
- BayesDecisionBoundaries, 5, 6, 9, 10, 13, 23
- CDFMixtures, 7
- Chi2testMixtures, 7, 8, 25
- ClassifyByDecisionBoundaries, 9
- EMGauss, 10
- InformationCriteria4GMM, 11
- Intersect2Mixtures, 6, 12
- KStestMixtures, 13
- LikelihoodRatio4Mixtures, 14
- LogLikelihood4Mixtures, 12, 15
- MultiModal (AdaptGauss-package), 2
- MultiModal-package
 (AdaptGauss-package), 2
- OptimalNoBins, 16, 19
- ParetoDensityEstimation, 17, 19
- ParetoRadius, 16–18, 18
- Pdf4Mixtures, 12, 19
- plot, 21–23
- PlotMixtures, 20, 20, 22, 23
- PlotMixturesAndBoundaries, 21, 22
- qqplot, 23, 24
- QQplotGMM, 23, 25
- RandomLogGMM, 24