Package: AdaptGauss (via r-universe)

November 5, 2024

Type Package

Title Gaussian Mixture Models (GMM)

Version 1.5.8

Date 2022-10-15

Maintainer Michael Thrun <m.thrun@gmx.net>

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. The package is based on the publication of Ultsch, A., Thrun, M.C., Hansen-Goos, O., Lotsch, J. (2015) <DOI:10.3390/ijms161025897>.

Imports Rcpp, shiny, pracma, methods, DataVisualizations

Suggests mclust, grid, foreach, dqrng, parallelDist, knitr (>= 1.12), rmarkdown (>= 0.9), reshape2, ggplot2, plotly

LinkingTo Rcpp

Depends R (>= 2.10)

License GPL-3

LazyLoad yes

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

Encoding UTF-8

NeedsCompilation yes

VignetteBuilder knitr

BugReports https://github.com/Mthrun/AdaptGauss/issuesConfig/pak/sysreqs: make libicu-dev zlib1g-dev

Repository https://mthrun.r-universe.dev

RemoteUrl https://github.com/mthrun/adaptgauss

RemoteRef HEAD

RemoteSha d0a5d20d90eb51faf41af16c01a507d6e41d9416

Contents

AdaptGauss-package 2
AdaptGauss 4
Bayes4Mixtures 6
BayesClassification 7
BayesDecisionBoundaries 8
BayesFor2GMM 9
CDFMixtures 10
Chi2testMixtures 11
ClassifyByDecisionBoundaries 12
EMGauss 13
GMMplot_ggplot2 14
InformationCriteria4GMM 15
Intersect2Mixtures 17
KStestMixtures 18
LikelihoodRatio4Mixtures 19
LKWFahrzeitSeehafen2010 20
LogLikelihood4Mixtures 21
Pdf4Mixtures 22
PlotMixtures 23
PlotMixturesAndBoundaries 24
QQplotGMM 25
RandomLogGMM 27
Symlognpdf 28

Index **29**

`AdaptGauss-package` *Gaussian Mixture Models (GMM)*

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. The package is based on the publication of Ultsch, A., Thrun, M.C., Hansen-Goos, O., Lotsch, J. (2015) <DOI:10.3390/ijms161025897>.

Details

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.

Index: This package was not yet installed at build time.

Author(s)

Michael Thrun, Onno Hansen-Goos, Rabea Griese, Catharina Lippmann, Florian Lerch, Jorn Lotsch, Alfred Ultsch Maintainer: Michael Thrun <m.thrun@gmx.net>

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 discovery, 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.

Examples

```
## Statistically  significant GMM

data=c(rnorm(3000,2,1),rnorm(3000,7,3),rnorm(3000,-2,0.5))

gmm=AdaptGauss::AdaptGauss(data,

Means = c(-2, 2, 7),

SDs = c(0.5, 1, 4),

Weights = c(0.3333, 0.3333, 0.3333))

AdaptGauss::Chi2testMixtures(data,

gmm$Means,gmm$SDs,gmm$Weights,PlotIt=T)

AdaptGauss::QQplotGMM(data,gmm$Means,gmm$SDs,gmm$Weights)


## Statistically non significant GMM

data('LKWFahrzeitSeehafen2010')

gmm=AdaptGauss::AdaptGauss(LKWFahrzeitSeehafen2010,
```

```
Means = c(52.74, 385.38, 619.46, 162.08),

SDs = c(38.22, 93.21, 57.72, 48.36),

Weights = c(0.2434, 0.5589, 0.1484, 0.0749))

AdaptGauss::Chi2testMixtures(LKWFahrzeitSeehafen2010,

gmm$Means,gmm$SDs,gmm$Weights,PlotIt=T)

AdaptGauss::QQplotGMM(LKWFahrzeitSeehafen2010,gmm$Means,gmm$SDs,gmm$Weights)
```

AdaptGauss *Adapt Gaussian Mixture Model (GMM)*

Description

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

Usage

```
AdaptGauss(Data, Means = NaN, SDs = NaN, Weights = NaN,

                    ParetoRadius = NaN, LB = NaN, HB = NaN,

                    ListOfAdaptGauss, fast = T)
```

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: StandardDevations 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 DataVisualizations::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 enerated 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.,Ultsch, 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`

`BayesClassification` *BayesClassification*

Description

Bayes Klassifikation den Daten zuordnen

Usage

```
BayesClassification(Data, Means, SDs, Weights, IsLogDistribution = Means
  * 0, ClassLabels = c(1:length(Means)))
```

Arguments

`Data`	vector of Data
`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
`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) classiffication 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

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`

`BayesFor2GMM` *Posterioris of Bayes Theorem for a two group GMM*

Description

Calculates the posterioris of Bayes theorem, splits the GMM in two groups beforehand.

Usage

```
BayesFor2GMM(Data, Means, SDs, Weights, IsLogDistribution = Means * 0,
  Ind1 = c(1:floor(length(Means)/2)), Ind2 = c((floor(length(Means)/2)
  + 1):length(Means)), PlotIt = 0, CorrectBorders = 0)
```

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
`Ind1`	indices from (1:C) such that [M(Ind1),S(Ind1) ,W(Ind1)]is one mixture, [M(Ind2),S(Ind2) ,W(Ind2)] the second mixture default Ind1= 1:C/2, Ind2= C/2+1:C
`Ind2`	indices from (1:C) such that [M(Ind1),S(Ind1) ,W(Ind1)]is one mixture, [M(Ind2),S(Ind2) ,W(Ind2)] the second mixture default Ind1= 1:C/2, Ind2= C/2+1:C
`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

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)

Alfred Ultsch, Michael Thrun

References

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.

See Also

BayesDecisionBoundaries,AdaptGauss

`CDFMixtures` *cumulative distribution of mixture model*

Description

returns the cdf (cumulative distribution function) of a mixture model of gaussian or log gaussians

Usage

```
CDFMixtures(Kernels,Means,SDs,Weights,IsLogDistribution)
```

Arguments

`Kernels`	at these locations N(Means,Sdevs)*Weights is used for cdf calcuation, NOTE: Kernels are usually (but not necessarily) sorted and unique
`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
`IsLogDistribution`	Optional, if IsLogDistribution(i)==1, then mixture is lognormal default == 0*(1:L)

Value

List with

`CDFGaussMixture`	(1:N,1), cdf of Sum of SingleGaussians at Kernels
`CDFSingleGaussian`	(1:N,1:L) ,cdf of mixtures at Kernels

Author(s)

Rabea Griese

See Also

`Chi2testMixtures`

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,NoRepetitionst)
```

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'
`NoRepetitions`	Optional, scalar, default =1000, Number of Repetitions for monte carlo sampling

Details

The null hypothesis is that the estimated data distribution does not differ 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 Ei may be derived for each bin. If there is a significant difference between the O_i and the E_i, the Pvalue is small and the null hypothesis can be rejected.

Further details, see [Thrun & Ultsch, 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((ObsNrInBin(Ind)-ExpectedNrInBin(Ind))^2/ExpectedNrInBin(Ind)) with Ind = find(ExpectedNrInBin>=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) classiffication 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, K, Means, SDs,Weights, MaxNumberofIterations,fast)
```

Arguments

`Data`	vector of data points
`K`	estimated amount of Gaussian Kernels
`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, Florian Lerch

References

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

See Also

`AdaptGauss`

`GMMplot_ggplot2` *Plots the Gaussian Mixture Model (GMM) withing ggplot2*

Description

PlotMixtures and PlotMixturesAndBoundaries for ggplot2

Usage

```
GMMplot_ggplot2(Data, Means, SDs, Weights,

BayesBoundaries, SingleGausses = TRUE, Hist = FALSE)
```

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

`BayesBoundaries`
: Optional, x values for baye boundaries, if missing 'BayesDecisionBoundaries' is called

`SingleGausses`
: Optional, SingleGausses=T than components of the mixture in blue will be shown.

`Hist`
: Optional, geom_histogram overlayed

Value

ggplot2 object

Note

MT standardized code for CRAN and added dec boundaries and doku

Author(s)

Joern Loetsch, Michael Thrun (ctb)

See Also

`PlotMixturesAndBoundaries`, `PlotMixtures`, `BayesDecisionBoundaries`

Examples

```
data=c(rnorm(1000),rnorm(2000)+2,rnorm(1000)*2-1)

GMMplot_ggplot2(data,c(-1,0,2),c(2,1,1),c(0.25,0.25,0.5),SingleGausses=TRUE)
```

`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*LogLikelihood, k = nr. of model parameter = 3*Nr. of Gaussians One Gaussian: K=2 (Weight is then not an parameter!) SMALL SAMPLE CORRECTION: for n= 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*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. guassian
`Mean2`	mean of 2.gaussian
`SD2`	standard deviations of 2.gaussian
`Weight2`	weight of 2. guassian
`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` *Kolmogorov-Smirnov test*

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,NoRepetitions,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.
`NoRepetitions`	Optional, default =1000, scalar, Number of Repetitions for monte carlo sampling
`Silent`	Optional, default=TRUE, If FALSE, shows progress of computation by points (On windows systems a progress bar)

Details

The null hypothesis is that the estimated data distribution does not differ significantly from the GMM. If there is a significant difference, then the Pvalue is small and the null hypothesis is rejected.

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

Examples

```
  data2=c(rnorm(1000),rnorm(2000)+2,rnorm(1000)*2-1)
  ## Not run: Vals=AdaptGauss(data2,c(-1,0,2),c(2,1,1),c(0.25,0.25,0.5),0.3,-6,6)
  NullMixture=cbind(Vals$Means,Vals$SDs,Vals$Weights)

## End(Not run)
 ## Not run: Vals2=AdaptGauss(data2,c(-1,0,2,3),c(2,1,1,1),c(0.25,0.25,0.25,0.25),0.3,-6,6)
  OneMixture=cbind(Vals2$Means,Vals2$SDs,Vals2$Weights)

## End(Not run)
  ## Not run:
  res=LikelihoodRatio4Mixtures(Data,NullMixture,OneMixture,T)

## End(Not run)
```

`LKWFahrzeitSeehafen2010` *Truck driving time seaport 2010*

Description

Truck driving time to seaports measured in 2010.

Usage

```
data("LKWFahrzeitSeehafen2010")
```

Format

The format is: num [1:11441] 84.7 13.2 11.5 41.4 52.9 ...

References

Behnisch, M., Ultsch, A.: Knowledge Discovery in Spatial Planning Data - A Concept for Cluster Understanding, in: Helbich, M., Arsanjani, J. J., Leitner, M. (eds.): Computational Approaches for Urban Environments, in: Gatrell, J.D., Jensen, R.R.: Geotechnologies and the Environment Series, Vol, 13, Springer, Berlin, pp. 49-75, 2015.

Examples

```
data(LKWFahrzeitSeehafen2010)
## maybe str(LKWFahrzeitSeehafen2010) ; plot(LKWFahrzeitSeehafen2010) ...
```

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

`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

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
`Plotter`	Optional, plotting package, either native or plotly
`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
`SingleGausses`	Optional, If TRUE, single gaussians are shown, default FALSE
`axes`	Optional,Default:TRUE with axis, see argument `axis` of `plot`
`xlab`	Optional, see `plot`
`ylab`	Optional, see `plot`
`xlim`	Optional, see `plot`
`ylim`	Optional, see `plot`
`ParetoRad`	Optional: Precalculated Pareto Radius to use
`...`	other plot arguments like xlim = c(1,10)

Details

Example shows that overlapping variances of gaussians will result in inappropriate decision boundaries.

Author(s)

Michael Thrun, Quirin Stier

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 = rep(FALSE, length(Means)), Plotter="native",

SingleColor = "blue", MixtureColor = "red", DataColor = "black",

BoundaryColor = "magenta", xlab, ylab,

 SingleGausses =TRUE, ...)
```

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
`Plotter`	Optional, plotting package, either native or plotly

`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`
`SingleGausses`	Optional, SingleGausses=T than components of the mixture in blue will be shown.
`...`	Optional, see `plot` for plot properties and for `SingleGausses` `PlotMixtures`

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

References

Michael, 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 genrated points for log oder gaussian mixture model

Author(s)

Alfred Ultsch,Michael Thrun, Rabea Griese

See Also

`QQplotGMM`,`Chi2testMixtures`

`Symlognpdf` *computes a special case of log normal distribution density*

Description

Symlognpdf is an internal function for **AdaptLGL**.

Usage

```
Symlognpdf(Data, Mean, SD)
```

Arguments

`Data`	vector of data points used for sampling
`Mean`	Mean of log Gaussian
`SD`	Standard deviation of log Gaussian

Value

M>0 Log normal distribution density

M<0 Log normal distribution density mirrored at y axis

Note

not for external usage.

See Also

`AdaptLGL`

Index

∗ **AIC**
InformationCriteria4GMM, 15
∗ **AdaptGauss**
AdaptGauss-package, 2
∗ **Akaike informations criterium**
InformationCriteria4GMM, 15
∗ **BIC**
InformationCriteria4GMM, 15
∗ **Bayes information criterium**
InformationCriteria4GMM, 15
∗ **BayesDecisionBoundaries**
BayesDecisionBoundaries, 8
∗ **Bayes**
Bayes4Mixtures, 6
BayesDecisionBoundaries, 8
PlotMixturesAndBoundaries, 24
∗ **Boundaries**
Bayes4Mixtures, 6
BayesDecisionBoundaries, 8
PlotMixturesAndBoundaries, 24
∗ **ClassifyByDecisionBoundaries**
ClassifyByDecisionBoundaries, 12
∗ **EM algorithm**
EMGauss, 13
∗ **EM**
AdaptGauss-package, 2
EMGauss, 13
∗ **Expectation-Maximization algorithm**
EMGauss, 13
∗ **Expectation-Maximization**
EMGauss, 13
∗ **Expectation**
EMGauss, 13
∗ **GMM**
AdaptGauss, 4
AdaptGauss-package, 2
GMMplot_ggplot2, 14
Pdf4Mixtures, 22
PlotMixtures, 23
RandomLogGMM, 27
∗ **Maximization**
EMGauss, 13
∗ **Minimum chi-square estimation**
Chi2testMixtures, 11
∗ **MultiModal**
AdaptGauss, 4
∗ **Multimodal**
AdaptGauss-package, 2
∗ **Pearson's chi-squared test**
Chi2testMixtures, 11
∗ **best-fit-line**
QQplotGMM, 25
∗ **chi-square estimation**
Chi2testMixtures, 11
∗ **chi-square goodness-of-fit**
Chi2testMixtures, 11
∗ **chi-square test for independence**
Chi2testMixtures, 11
∗ **chi-squared test**
Chi2testMixtures, 11
∗ **chi-square**
Chi2testMixtures, 11
∗ **datasets**
LKWFahrzeitSeehafen2010, 20
∗ **expectation maximization**
AdaptGauss-package, 2
∗ **gaussian mixture model**
AdaptGauss, 4
AdaptGauss-package, 2
Pdf4Mixtures, 22
PlotMixtures, 23
∗ **ggplot2**
GMMplot_ggplot2, 14
∗ **log GMM**
RandomLogGMM, 27
∗ **mixture of components**
AdaptGauss-package, 2
∗ **mixture**

AdaptGauss, 4
AdaptGauss-package, 2
∗ **pareto density estimation**
AdaptGauss-package, 2
∗ **pdf**
AdaptGauss-package, 2
Pdf4Mixtures, 22
∗ **plot**
QQplotGMM, 25
∗ **posterioris**
Bayes4Mixtures, 6
∗ **posterior**
Bayes4Mixtures, 6
∗ **probability density function**
Pdf4Mixtures, 22
∗ **qq-plot**
QQplotGMM, 25
∗ **qqplot**
QQplotGMM, 25
∗ **quantile/quantile-plot**
QQplotGMM, 25

AdaptGauss, 4, *7, 8*, *14*
AdaptGauss-package, 2
AdaptLGL, *28*

Bayes4Mixtures, 6, *8*, *13*
BayesClassification, 7
BayesDecisionBoundaries, *7*, 8, *12, 13*, *15*, *17*, *25*
BayesFor2GMM, 9

CDFMixtures, 10
Chi2testMixtures, *10*, 11, *27*
ClassifyByDecisionBoundaries, 12

EMGauss, 13

GMMplot_ggplot2, 14

InformationCriteria4GMM, 15
Intersect2Mixtures, *8*, 17

KStestMixtures, 18

LikelihoodRatio4Mixtures, 19
LKWFahrzeitSeehafen2010, 20
LogLikelihood4Mixtures, *16*, 21

MultiModal *(*AdaptGauss-package*)*, 2
MultiModal-package *(*AdaptGauss-package*)*, 2

Pdf4Mixtures, *16*, 22
plot, *23*, *25, 26*
PlotMixtures, *15*, *22*, 23, *25*
PlotMixturesAndBoundaries, *15*, *24*, 24

qqplot, *26*
QQplotGMM, 25, *27*

RandomLogGMM, 27

Symlognpdf, 28