Package 'krm'

July 2, 2014

Version 13.11-03

Date 2013-11-03

Title Kernel-based Regression Models

Imports

Author Youyi Fong <youyifong@gmail.com>, Saheli Datta, Krisztian Sebestyen

Maintainer Youyi Fong <youyifong@gmail.com>

Depends R (>= 3.0.0)

Suggests RUnit, xtable, MASS

Description Testing methods for kernel-based regression models.

License GPL-2

LazyLoad yes

LazyData yes

Date/Publication 2013-11-05 08:57:30

NeedsCompilation yes

Repository CRAN

R topics documented:

`aa.prop.list` *Amino Acid Properties*

Description

Amino Acid Properties

Format

A data frame with 20 observations on the following 13 variables.

`Symbol` a chacracter vector with 20 values: A through Y

`AA_Name` a chacracter vector with 20 values: Alanine through Tyrosine

`AA_Symbol` a chacracter vector with 20 values: Ala through Tyr

`Surface_Area_Chothia` a numeric vector

`Residue_Volume_Zamayatin` a numeric vector

`Bulkiness_Jones` a numeric vector

`Polarity_Jones` a numeric vector

`Refractivity_Jones` a numeric vector

`Hydrophobicity_Engleman` a numeric vector

`Hydrophobicity_Prabha` a numeric vector

`Hydrophilicity_Hopp` a numeric vector

`Hydrophilicity_Levitt` a numeric vector

`RelMutability_Jones` a numeric vector

calcPairwiseIdentity *Functions Related to Sequence Alignment*

Description

Functions related to sequence alignment

Usage

```
calcPairwiseIdentity(alignment, dissimilarity, removeGap)
alignment2count (alignment, level=20, weight=rep(1,nrow(alignment)))
alignment2trancount (alignment, weight=rep(1,nrow(alignment)))
removeGap (seq)
```

Arguments

alignment	matrix of arabic representation of sequences (1 based)
dissimilarity	Boolean.
removeGap	Boolean
level	integer. Size of alphabet
weight	numeric vector. Weights given to each sequence
seq	string. A string of amino acids

Value

alignment2count return T by 20 matrix, where T is the number of column in the alignment. alignment2trancount return a T by 4 matrix, each row is the count of MM, MD, DM, DD for each position.

chi.norm *A Transformation of Chi-squared Random Variable*

Description

A transformation of Chi-squared random variable to make it normal like.

Usage

```
chi.norm(q, v)
```

Arguments

q	numeric. A random variable following chi-squared distribution
v	numeric. A random variable following normal distribution

cloud9 *9-Component Mixture Dirichlet Prior for Protein Sequences*

Description

9-Component Mixture Dirichlet Prior for Protein Sequences

Format

List of 2. The alpha element is a 9 by 20 matrix, where each row represents one Dirichlet distribution of 20 dimensions. The mix.coef element contains the mixing probability, a vector 0f 9 numbers that add up to 1.

dmdirichlet *Functions related to mixture Dirichlet distribution*

Description

Functions related to mixture Dirichlet distribution

Usage

```
dmdirichlet(x, mAlpha, mixtureCoef)
ddirichlet (x, alpha)
rdirichlet (n, alpha)
rmdirichlet (mAlpha, mixtureCoef)
modifyDirichlet (prior, y)
logIntegrateMixDirichlet(y, prior, tau=1)
logIntegrateDirichlet (y, alpha)
```

Arguments

x	A vector containing a single deviate or matrix containing one random deviate per row.
mAlpha	matrix. Each row is a parameter of Dirichlet
alpha	numeric vector. Parameter for a Dirichlet distribution
mixtureCoef	numeric vector
n	integer
prior	list of two components: alpha and mix.coef
y	numeric vector of counts
tau	numeric

Details

ddirichlet andn rdirichlet are identically copied from MCMCpack

---

`getSeqKernel` *Protein Sequence Kernels*

Description

Get mutual information and other kernels for protein sequences

Usage

```
getSeqKernel (sequences, kern.type=c("mm", "prop", "mi"), tau, call.C=TRUE)
```

Arguments

`sequences`	String or list. If string, the name of a fasta file containing aligned sequences. If list, a list of strings, each string is a protein sequence. If list, call.C will be set to FALSE internally because C/C++ function needs sequence file name as input
`kern.type`	string. Type of kernel. mm: match-mismatch, prop: physicochemical properties, mi: mutual information.
`tau`	Numeric. It is the same as rho^-2.
`call.C`	Boolean. If TRUE, do a .C call. If FALSE, the implementation is in R. The .C call is 50 times faster.

Details

call.C option is to allow comparison of R and C implementation. The two should give the same results and C implementation is 50 times faster.

when kern.type is mi and call.C is TRUE and when running on linux, this function will print messages like "read ...". This message is generated from U::openRead

Examples

```
fileName=paste(system.file(package="krm")[1],'/misc/SETpfamseed_aligned_for_testing.fasta',
     sep="")
K=getSeqKernel (fileName, kern.type="mi", tau=1, call.C=TRUE)
K
```

hmmMargLlik *Functions related to profile HMM*

Description

Functions related to profile HMM

Usage

```
hmmMargLlik(dat, aaPrior, tau)
readPriorFromFile(priorFileName)
```

Arguments

dat	a matrix representation of a multiple sequence alignment, each row is a sequence, each column is a position
aaPrior	a list of two elements, "alpha" "mix.coef", representing mixture Dirichlet prior
tau	numeric
priorFileName	string

krm *Kernel-based Regression Model*

Description

Tests for kernel-based regression model

krm.most *Kernel-based Regression Model Maximum of adjusted Score Test*

Description

Computes maximum of adjusted score test. Obtain p value through parametric bootstrap

Usage

```
krm.most (formula, data, regression.type=c("logistic","linear"),
    kern.type=c("rbf","mi","mm","prop"),
    n.rho=10, range.rho=0.99, n.mc=2000, seq.file.name=NULL, formula.kern=NULL,
    inference.method=c("parametric.bootstrap", "perturbation", "LGL2008"),
    verbose=FALSE)
```

Arguments

formula	a formula object describing the null model
data	data frame
regression.type	a string
formula.kern	formula. The formula for the covariates used to form the kernel
seq.file.name	string. Name of a file containing sequences in fasta format
kern.type	string. Type of kernel. mm: match-mismatch, prop: physicochemical properties, mi: mutual information, rbf: radial basis function
n.rho	integer. Number of rhos to maximize over
range.rho	numeric. A number between 0 and 1. It controls the range of rhos to use to compute kernel
n.mc	integer. Number of bootstrap samples
inference.method	string
verbose	boolean

Examples

```
## Not run:
# the examples are not run during package build because it takes a little too long to run

data=sim.liu.2008 (n=100, a=.1, seed=1)
test = krm.most(y~x, data, formula.kern=~z.1+z.2+z.3+z.4+z.5, kern.type="rbf")


dat.file.name=paste(system.file(package="krm")[1],'/misc/y1.txt', sep="")
seq.file.name=paste(system.file(package="krm")[1],'/misc/sim1.fasta', sep="")
dat=read.table(dat.file.name); names(dat)="y"
test = krm.most (y~1, dat, seq.file.name=seq.file.name, kern.type="mi")



## End(Not run)
```

---

krm.score.test *Adjusted Score Test*

Description

Performs adjusted score test for logistic models with kernel random effect.

Usage

```
krm.score.test(formula, data, K, regression.type=c("logistic","linear"), verbose=FALSE)
```

Arguments

formula	a formula object. Model under null.
data	a data frame
K	a n by n kernel/correlation matrix
regression.type	a string
verbose	Boolean

Examples

```
dat=sim.liu.2008(n=100, a=0, seed=1)
z=as.matrix(subset(dat, select=c(z.1,z.2,z.3,z.4,z.5)))
rho=1
K=krm:::getK(z,kernel="rbf",para=rho^-2)
krm.score.test (y~x, dat, K, regression.type="logistic")
```

---

readFastaFile *Read a Fasta Sequence File*

Description

Read a Fasta Sequence File

Usage

```
readFastaFile(fileName, sep = " ")
writeFastaFile (seqList, fileName)
aa2arabic (seq1)
string2arabic (seqList)
fastaFile2arabicFile (fastaFile, arabicFile, removeGapMajor=FALSE)
selexFile2arabicFile (selexFile, arabicFile, removeGapMajor=FALSE)
stringList2arabicFile (seqList, arabicFile, removeGapMajor=FALSE)
arabic2arabicFile (alignment, arabicFile)
readSelexFile (fileName)
readSelexAsMatrix (fileName)
arabic2fastaFile (alignment, fileName)
readArabicFile (fileName)
readBlockFile (fileName)
```

Arguments

fileName	string
fastaFile	string
arabicFile	string
selexFile	string
sep	string
seq1	string. A string of amino acids
seqList	list of string.
removeGapMajor	Boolean
alignment	matrix of arabic representation of sequences (1 based)

Value

string2arabic returns a matrix of arabic numbers representing aa. readSelexFile return a list of strings. readArabicFile return a matrix of n by p alignment.

Examples

```
library(RUnit)
fileName=paste(system.file(package="krm")[1],'/misc/SETpfamseed_aligned_for_testing.fasta', sep="")
seqs = readFastaFile (fileName, sep=" ")
checkEquals(length(seqs),11)
```

---

sim.liu.2008 *Simulate sDataset*

Description

Per Liu et al (2008) and Liu et al (2007)

Usage

```
sim.liu.2008(n, a, seed = NULL)
sim.liu.2007(n, a, seed = NULL)
```

Arguments

n	sample size
a	numeric. If a is 0, then the data is used to study size, otherwise power
seed	optional random number generator seed

Index