Package ‘RMallow’

October 12, 2022

Type Package
Title Fit Multi-Modal Mallows’ Models to Ranking Data
Version 1.1
Date 2020-01-23
Depends combinat
Description An EM algorithm to fit Mallows’ Models to full or partial rankings, with or without ties.
License GPL (>= 2)
Repository CRAN
Date/Publication 2020-02-07 01:10:16 UTC
NeedsCompilation no
Author Erik Gregory [aut, cre]
Maintainer Erik Gregory <egregory2011@gmail.com>

R topics documented:

RMallow-package ........................................ 2
AllKendall .................................................. 3
AllSeqDists .................................................. 3
BestFit ....................................................... 4
ConstructSeqs .............................................. 5
C_lam ......................................................... 5
datas ....................................................... 6
DistanceDistribution ...................................... 6
elect .......................................................... 7
EStep ......................................................... 7
FormatOut ................................................... 8
hello ......................................................... 9
KendallInfo ................................................ 9
Lambda ....................................................... 10
Likelihood .................................................. 11
Mallows ..................................................... 11
Fit Multi-modal Mallows' models to ranking data.

Description
Fits the Mallows’ model to ranking data. Data can be partially or fully-ranked.

Details

<table>
<thead>
<tr>
<th>Package:</th>
<th>RMallow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Package</td>
</tr>
<tr>
<td>Version:</td>
<td>1.0</td>
</tr>
<tr>
<td>Date:</td>
<td>2012-02-18</td>
</tr>
<tr>
<td>License:</td>
<td>GPL (&gt;= 2)</td>
</tr>
</tbody>
</table>

Author(s)

Erik Gregory Maintainer: <egregory2007@yahoo.com>

References


AllKendall

All Kendall’s distances between two sets of rankings.

Description
Calculates all of the Kendall’s distances between two different sets of rankings.

Usage
AllKendall(r, seqs, data.info = NULL)

Arguments
- r: One set of sequences.
- seqs: Another set of sequences.
- data.info: Optional argument, a 0/1/NA matrix specifying all of the relevant information to calculate Kendall’s difference for "r". Used for efficiency in "Solve".

Value
Matrix where output[i, j] represents the distance from sequence "i" in "r" to sequence "j" in "seqs".

Author(s)
Erik Gregory

Examples
data1 <- do.call("rbind", list(1:5, 5:1, c(3, 2, 1, 4, 5)))
data2 <- do.call("rbind", list(1:5, 5:1))
# AllKendall(data1, data2)

AllSeqDists

Calculate all distances between a set of sequences and a fixed sequence.

Description
Used to calculate the sequence Kendall distance distribution in N! space.

Usage
AllSeqDists(seqs)
Arguments

seqs Matrix or data frame of sequences.

Value

Vector of the distances from the sequences to 1:N.

Author(s)

Erik Gregory

---

BestFit Fit Mallows model N times and select most likely model. The EM algorithm to fit Multi-Modal Mallows' models is prone to getting stuck in local maxima, so we run it several times and select the best one.

Description

Fit Mallows model N times and select most likely model. The EM algorithm to fit Multi-Modal Mallows' models is prone to getting stuck in local maxima, so we run it several times and select the best one.

Usage

BestFit(datas, N, iter, G)

Arguments

N number of times to run the model
iter maximum number of iterations for each run
G Number of cluster centers
data set to fit

Value

best fitting model.
ConstructSeqs

ConstructSeqs

Constructs sequences from Kendall Information matrices.

Description

Sequences in a fully-ordered sequence space have a unique Kendall Information vector associated with them. This function creates the sequence from the Kendall information vector.

Usage

ConstructSeqs(prefs, n.abils)

Arguments

- `prefs` Ordering preference between columns in the data. 1 corresponds to an increase, 0 to a decrease.
- `n.abils` Number of columns in the original data set.

Value

List of fully-ordered sequences, one for each row of prefs.

Author(s)

Erik Gregory

Examples

ConstructSeqs(matrix(c(1, 1, 1, 0, 0, 0), nrow = 1), 4)
# Should output (4, 1, 2, 3)

C_lam

Calculate the normalizing coefficient for Mallow's model in a sequence space.

Description

Calculate the normalizing coefficient, as a function of the lambda parameter, and the size of the sequence space.

Usage

C_lam(lambda, dists = NULL, dists.table = NULL)
**DistanceDistribution**

**Arguments**

- **lambda**  
  Spread parameter for Mallows’ model.

- **dists**  
  Vector of all distances from each sequence to 1:N

- **dists.table**  
  Table version of “dists” above.

**Value**

Normalizing coefficient of Mallows’ model in N! space with lambda = lambda.

**Author(s)**

Erik Gregory

---

**datas**

*Sample data set.*

**Description**

Simple synthetic data set containing 3 modal sequences in 15! space, with some noise added.

**Format**

The format is: num [1:1700, 1:15] 1 15 1 15 1 15 12 10 4 1 15 ...

**Examples**

```
data(datas)
head(datas)
```

---

**DistanceDistribution**  
*Calculate the Kendall distance distribution in N! space.*

**Description**

This function counts the number of fully-ordered vectors at each distance in N! space.

**Usage**

```
DistanceDistribution(N = 3)
```

**Arguments**

- **N**  
  Integer value, greater than or equal to 3.
Value

Table-like structure, where the names represent the distance from the modal sequence of each sequence in N! space, and the values represent the number of sequences at that distance in the sequence space.

Author(s)

Erik Gregory

Description

This data is a pre-processed version of the 1980 American Psychological Association Presidential candidate ranking data. It has uninformative rankings removed, and values pre-simplified into partial rankings.

Format

The format is: int [1:1378, 1:3] 1 1 1 2 1 1 2 2 ... - attr(*, "dimnames")=List of 2 ..$ : chr [1:1378] "1" "2" "3" "6" ... ..$ : chr [1:3] "Carter" "Reagan" "Anderson"

Source


Examples

data(elect)
head(elect)

EStep

The Expectation step of the EM algorithm.

Description

Assigns each ranking the probability that it belongs to each cluster, given current parameters.

Usage

EStep(R, r, p, lambda, G, N, C, all.dists = NULL)
Arguments

R  Current cluster modal sequences.

r  The data of partial or full rankings.

p  The proportion of the data currently assigned to each cluster.

lambda  The lambda parameters from Mallow’s model for each cluster.

G  Number of clusters, length(R).

N  Number of rows in the data.

C  Vector of normalizing coefficients for the clusters.

all.dists  For efficiency, provide all of the Kendall distances between each sequence and each cluster mode.

Value

Matrix where output[i, j] represents the current probability that subject "i" belongs to cluster "j".

Author(s)

Erik Gregory

References


---

FormatOut  Formats the data in the "Solve" function for output.

Description

Data formatting function.

Usage

FormatOut(R, p, lambda, z, datas, likelihood)

Arguments

R  The modal sequences.

p  Proportion of data in each cluster.

lambda  Mallows’ spread parameters for each cluster.

z  Probability of cluster membership for each individual.

datas  Matrix of partial sequences.

likelihood  Vector of the log-likelihood of the model at each iteration.
**Value**

- **R**  The modal sequences  
- **p**  Proportion in each cluster  
- **lambda**  Spread parameters for each cluster  
- **datas**  Rankings merged with their cluster membership, distance from each cluster center, and probability of each cluster membership  
- **min.like**  Likelihood at each iteration

**Author(s)**

Erik Gregory

**hello**  

*Hello, World!*

**Description**

Prints 'Hello, world!'.

**Usage**

```r
hello()
```

**Examples**

```r
hello()
```

**KendallInfo**

*All information used to calculate Kendall’s distance.*

**Description**

Performs each column-wise comparison on a matrix of sequences. A 0 value denotes that there is an increase between the two columns, 1 a decrease, and NA indicates that the column values are identical in the row.

**Usage**

```r
KendallInfo(r, inds = NULL)
```

**Arguments**

- **r**  Matrix of sequences.  
- **inds**  Possibly efficiency increase when doing repeated calculations, currently not used.
Value
Matrix of 0s, 1s, and NAs representing pairwise comparisons of vector values.

Author(s)
Erik Gregory

References
http://en.wikipedia.org/wiki/Kendall_tau_distance

Description
Objective function to find the root of in calculating the lambda parameters for each cluster.

Usage
Lambda(lambda, rhs, dists, dists.table = NULL)

Arguments
lambda lambda value to calculate the function output at.
rhs Right-hand side of the equation in the referenced paper.
dists Not used.
dists.table Table of distances between each sequence and the modal sequence in N! space.

Value
Output of the objective function to determine the root of. Goal is zero.

Author(s)
Erik Gregory

References
Likelihood

**Description**

Calculates the log-likelihood of the data with the current parameters and Kendall’s distance.

**Usage**

```r
Likelihood(z, p, C.lam, lambda, all.dists.data)
```

**Arguments**

- `z`: Probability of each cluster membership.
- `p`: Proportion in each cluster.
- `C.lam`: Vector of normalizing coefficients for Mallows’ model.
- `lambda`: Current spread parameters
- `all.dists.data`: All distances from the data to the modal sequences.

**Value**

Current log-likelihood of the data with the current parameters.

**Author(s)**

Erik Gregory

**References**


Mallows

**Description**

Fits a Multi-Modal Mallows’ model to ranking data.

**Usage**

```r
Mallows(datas, G, iter = 10, hyp = NULL, plot.like = FALSE)
```

**Description**

Fits the Multi-Modal Mallows’ model to partial or full ranking data, using Kendall’s metric and an EM algorithm. This is essentially metric sequence clustering.
Arguments

- `datas`: Matrix of partial or fully-ranked data.
- `G`: Number of modes, 2 or greater.
- `iter`: Maximum number of iterations.
- `hyp`: Hypothesis sequence vector, to initialize one of the cluster centers at.
- `plot.like`: Should the likelihood be printed at each iteration?

Value

See output of `FormatOut`.

Author(s)

Erik Gregory

References


---

**NextTable**

Calculates the table of Kendall distances in \((N+1)!\) space, given those in \(N!\) space.

---

**Description**

This is identical to counting the number of fully-ordered vectors at each bubble sort distance in \((N+1)!\) space.

**Usage**

```r
NextTable(last.table, N.last)
```

**Arguments**

- `last.table`: Table of distances in \(N!\) space.
- `N.last`: \(N\)

**Value**

Table of distances in \((N+1)!\) space.

**Author(s)**

Erik Gregory
Rgen

**Description**
Initialize sequence modes for the clustering process.

**Usage**
Rgen(G, hyp = NULL, abils)

**Arguments**
- **G**: number of cluster centers, including the hypothesis if provided
- **hyp**: a single sequence of length abils to initialize one of the cluster centers
- **abils**: number of items being ranked

**Value**
A list of G cluster centers, each of length abils

**Author(s)**
Erik Gregory

**Examples**
Rgen(3, 1:5, 5)

---

SeqDistribution

**Description**
Calculates distances in N! space. This is VERY Inefficient for N >= 8. See DistanceDistribution for an astronomical improvement (possibly on the order of 10^10).

**Usage**
SeqDistribution(N)

**Arguments**
- **N**: Length of the ranking. Preferably less than 9.
Value
Vector of Kendall distances from 1:N to each sequence in N! space.

Author(s)
Erik Gregory

SimplifySequences
Change the form of ordered sequences.

Description
Simplifies sequences so that each tie group is only of distance 1 to the next tie group. For example, we would simplify (1, 1, 2, 4, 4, 5) to (1, 1, 2, 3, 3, 4).

Usage
SimplifySequences(loss.time)

Arguments
loss.time Matrix of sequences to be simplified.

Value
Simplified sequences, as described in Description.

Author(s)
Erik Gregory

three.mode
Fitted version of the toy datas data set, with three modal sequences.

Description
The data has 3 modal sequences, and we can compare this to the two.mode data set.
Format

The format is: List of 5 $ R : List of 3 $ : int [1:15] 1 2 3 4 5 6 7 8 9 10 ... $ : int [1:15] 13 5 7 9 2 4 6 8 10 ... $ : int [1:15] 15 14 13 12 11 10 9 8 7 6 ... $ p : num [1:3] 0.447 0.118 0.435 $ lambda : num [1:3] 2.01 1000 2.04 $ datas : 'data.frame': 1700 obs. of 23 variables: X1 : num [1:1700] 1 15 15 15 12 10 4 11 15 14 ... $ X2 : num [1:1700] 2 14 2 14 14 13 12 2 14 13 ... $ X3 : num [1:1700] 3 13 3 13 2 4 6 3 13 ... $ X4 : num [1:1700] 4 12 4 12 8 7 14 12 ... $ X5 : num [1:1700] 5 11 5 11 9 14 5 11 ... $ X6 : num [1:1700] 6 10 6 10 10 8 10 6 10 8 ... $ X7 : num [1:1700] 7 9 9 9 15 1 13 7 9 ... $ X8 : num [1:1700] 8 8 8 8 8 10 9 8 8 ... $ X9 : num [1:1700] 9 7 9 7 6 5 14 9 7 ... $ X10 : num [1:1700] 10 6 10 6 6 11 8 10 6 ... $ X11 : num [1:1700] 11 5 11 5 3 15 11 5 ... $ X12 : num [1:1700] 12 4 12 4 4 14 12 11 12 4 ... $ X13 : num [1:1700] 13 3 13 3 3 7 2 7 13 3 ... $ X14 : num [1:1700] 14 2 14 2 2 5 3 15 14 2 ... $ X15 : num [1:1700] 15 1 15 1 1 4 6 3 15 1 ... $ clust : int [1:1700] 1 3 1 3 3 1 3 1 3 3 ... $ pvals.1: num [1:1700] 1.00 1.03e-91 1.00 2.04e-93 1.03e-91 ... $ pvals.2: num [1:1700] 0 0 0 0 0 0 0 0 0 0 ... $ pvals.3: num [1:1700] 1.02e-92 1.00 1.34e-93 1.00 1.00 ... $ seq : Factor w/ 3 levels "1 2 3 4 5 6 7 8 9 10 11 12 13 14 15",..: 1 3 1 3 3 1 3 1 3 3 ... $ dists.1: num [1:1700] 0 105 0 105 105 61 58 46 0 105 ... $ dists.2: num [1:1700] 105 0 105 0 0 44 47 59 105 0 ... $ min.like: num [1:100] -178063 -139298 -58290 -54074 -53902 ...
Examples

```r
data(two.mode)
head(two.mode[[4]])
```

**two.seq**  
Bi-modal Mallow’s model fit to the APA data set.

**Description**

The two-modes seem to divide well between Democrats and Republicans...

**Format**

The format is: List of 5 $ R : List of 2 ..$ : int [1:3] 1 3 2 ..$ : int [1:3] 3 1 2 $ p : num [1:2] 0.541 0.459 $ lambda : num [1:2] 2.19 2.32 $ datas : ‘data.frame’: 1378 obs. of 9 variables: ..$ Carter : int [1:1378] 1 1 1 1 2 2 1 1 2 2 ... ..$ Reagan : int [1:1378] 1 2 2 1 1 2 3 1 1 ... ..$ Anderson: int [1:1378] 1 2 2 3 3 3 2 3 3 ... ..$ clust : int [1:1378] 1 1 1 1 2 2 1 1 2 2 ... ..$ pvals.1 : num [1:1378] 0.541 0.992 0.992 0.932 0.131 ... ..$ pvals.2 : num [1:1378] 0.45893 0.00809 0.00809 0.06802 0.86945 ... ..$ seq : Factor w/ 2 levels "1 3 2","3 1 2": 1 1 1 1 2 2 1 1 2 2 ... ..$ dists.1 : num [1:1378] 0 0 0 1 2 2 1 0 2 2 ... ..$ dists.2 : num [1:1378] 0 2 2 1 2 3 1 1 ... $ min.like: num [1:100] -6421 -3386 -2916 -2811 -2799 ...

**Source**


Examples

```r
data(two.seq)
head(two.seq[[4]])
```

**UpdateLambda**

Update the Lambda parameters of clusters.

**Description**

Updates the Lambda parameters to maximize the likelihood of the data under Mallows’ model.

**Usage**

```r
UpdateLambda(r, R, z, G, dists.to.Rg, dists.table, top.bound = 1000)
```
**Arguments**

- $r$: Matrix of partial rankings.
- $R$: Current modal sequences.
- $z$: Current probabilities of memberships in each cluster.
- $G$: Number of modal sequences.
- $dists.to.Rg$: Matrix of the distances between the data and the current modal sequences.
- $dists.table$: Table of the distance distribution in $N!$ space, under Kendall's metric.
- $top.bound$: The maximum value for the lambda parameter.

**Value**

Vector of new lambda parameters for the clusters.

**Author(s)**

Erik Gregory

**References**


---

**UpdateP**

*Update Proportion in each cluster.*

**Description**

Updates the proportion of data assigned to each cluster.

**Usage**

`UpdateP(z)`

**Arguments**

- $z$: Probabilities that each sequence is in each cluster.

**Value**

Proportion of data in each cluster.

**Author(s)**

Erik Gregory
References

"Mixtures of distance-based models for ranking data". Thomas Brendan Murphy & Donal Martin.

Description

Maximizes the likelihood of the data by updating the cluster centers of the model.

Usage

UpdateR(r, z, infos = NULL)

Arguments

r Matrix of sequences being clustered.
z Probability of cluster membership for each sequence and each cluster.
infos The KendallInfo matrix for "r".

Value

New cluster centers for each cluster.

Author(s)

Erik Gregory

References

"Mixtures of distance-based models for ranking data". Thomas Brendan Murphy & Donal Martin.
Index

* BubbleSort
  FormatOut, 8
* DistanceDistribution(10)
  DistanceDistribution, 6
* Distance
  AllSeqDists, 3
  KendallInfo, 9
* Kendall
  AllKendall, 3
  AllSeqDists, 3
  DistanceDistribution, 6
  FormatOut, 8
  KendallInfo, 9
  NextTable, 12
* Mallow
  Lambda, 10
  Likelihood, 11
  Mallows, 11
* Sequences
  ConstructSeqs, 5
* #
  DistanceDistribution, 6
* bubblesort
  DistanceDistribution, 6
  NextTable, 12
  SeqDistribution, 13
* center
  UpdateR, 18
* cluster
  Mallows, 11
  UpdateR, 18
* datasets
  datas, 6
  elect, 7
  three.mode, 14
  two.mode, 15
  two.seq, 16
* dataset
  two.seq, 16
* distance
  AllKendall, 3
  SeqDistribution, 13
* expectation
  EStep, 7
* lambda
  Lambda, 10
  UpdateLambda, 16
* likelihood
  Likelihood, 11
* maximization
  EStep, 7
  UpdateLambda, 16
* normalize
  C_lam, 5
* proportion
  UpdateP, 17
* ranking
  RMallow-package, 2
* sequence
  SimplifySequences, 14
* simplify
  SimplifySequences, 14

AllKendall, 3
AllSeqDists, 3
BestFit, 4
C_lam, 5
ConstructSeqs, 5
datas, 6
DistanceDistribution, 6
elect, 7
EStep, 7
FormatOut, 8
hello, 9
KendallInfo, 9
Lambda, 10
Likelihood, 11
Mallows, 11
NextTable, 12
Rgen, 13
RMallow (RMallow-package), 2
RMallow-package, 2
SeqDistribution, 13
SimplifySequences, 14
three.mode, 14
two.mode, 15
two.seq, 16
UpdateLambda, 16
UpdateP, 17
UpdateR, 18