Package ‘GenOrd’

April 1, 2015

Type Package

Title Simulation of Ordinal and Discrete Variables with Given Correlation Matrix and Marginal Distributions

Version 1.3.0

Date 2015-03-31

Author Alessandro Barbiero, Pier Alda Ferrari

Maintainer Alessandro Barbiero <alessandro.barbiero@unimi.it>

Description A gaussian copula based procedure is implemented for generating samples from ordinal/discrete random variables with pre-specified correlation matrix and marginal distributions.

License GPL

LazyLoad yes

Depends mvtnorm, Matrix, MASS

NeedsCompilation no

Repository CRAN

Date/Publication 2015-04-01 12:57:50

R topics documented:

GenOrd-package ...................................................... 2
contord ............................................................ 3
corrcheck .......................................................... 4
ordcont ............................................................. 5
ordsample ........................................................... 7

Index 10
GenOrd-package

Simulation of Ordinal and Discrete Variables with Given Correlation Matrix and Marginal Distributions

Description

The package implements a procedure for generating samples from ordinal/discrete random variables with pre-specified correlation matrix and marginal distributions. The marginal distribution are linked together through a gaussian copula. The procedure is developed in two steps: the first step (function `ordcont`) sets up the gaussian copula in order to achieve the desired correlation matrix on the target ordinal/discrete variables; the second step (`ordsample`) generates samples from the target variables. The procedure can handle both Pearson’s and Spearman’s correlations, and any finite support for the discrete variables. The intermediate function `contord` computes the correlations of discrete/ordinal variables derived from correlated normal variables through discretization. Function `corrcheck` returns the lower and upper bounds of the correlation coefficient of each pair of ordinal/discrete variables given their marginal distributions, i.e., returns the range of feasible bivariate correlations.

This version has fixed some drawbacks of `ordcont`, which now stops and returns an error if the resulting correlation matrix of the gaussian copula is not a valid correlation matrix. Validity checks on the marginal distributions and correlation matrix provided by the user in the four functions have been added.

Details

<table>
<thead>
<tr>
<th>Package:</th>
<th>GenOrd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Package</td>
</tr>
<tr>
<td>Version:</td>
<td>1.2.0</td>
</tr>
<tr>
<td>Date:</td>
<td>2015-03-31</td>
</tr>
<tr>
<td>License:</td>
<td>GPL</td>
</tr>
<tr>
<td>LazyLoad:</td>
<td>yes</td>
</tr>
</tbody>
</table>

Author(s)

Alessandro Barbiero, Pier Alda Ferrari Maintainer: Alessandro Barbiero <alessandro.barbiero@unimi.it>

References


See Also

`contord, ordcont, corrcheck, ordsample`
Correlations of discretized variables

Description

The function computes the correlation matrix of the \( k \) variables, with given marginal distributions, derived discretizing a \( k \)-variate standard normal variable with given correlation matrix.

Usage

\[
\text{contord}(\text{marginal}, \text{Sigma}, \text{support} = \text{list()}, \text{Spearman} = \text{FALSE})
\]

Arguments

- \text{marginal}: a list of \( k \) elements, where \( k \) is the number of variables. The \( i \)-th element of \text{marginal} is the vector of the cumulative probabilities defining the marginal distribution of the \( i \)-th component of the multivariate variable. If the \( i \)-th component has \( k_i \) categories, the \( i \)-th element of \text{marginal} will contain \( k_i - 1 \) probabilities (the \( k_i \)-th is obviously 1 and shall not be included).
- \text{Sigma}: the correlation matrix of the standard multivariate normal variable
- \text{support}: a list of \( k \) elements, where \( k \) is the number of variables. The \( i \)-th element of \text{support} contains the ordered values of the support of the \( i \)-th variable. By default, the support of the \( i \)-th variable is \( 1, 2, ..., k_i \).
- \text{Spearman}: if \text{TRUE}, the function finds Spearman’s correlations (and it is not necessary to provide \text{support}), if \text{FALSE} (default) Pearson’s correlations.

Value

- the correlation matrix of the discretized variables

Author(s)

Alessandro Barbiero, Pier Alda Ferrari

See Also

\text{ordcont}, \text{ordsample}, \text{corrcheck}

Examples

\[
\begin{align*}
\text{# consider 4 discrete variables} \\
k <- 4 \\
\text{# with these marginal distributions} \\
\text{marginal} <- \text{list}(0.4, c(0.3, 0.6), c(0.25, 0.5, 0.75), c(0.1, 0.2, 0.8, 0.9)) \\
\text{# generated discretizing a multivariate standard normal variable} \\
\text{# with correlation matrix} \\
\text{Sigma} <- \text{matrix}(0.5, 4, 4)
\end{align*}
\]
diag(Sigma) <- 1
# the resulting correlation matrix for the discrete variables is
contord(marginal, Sigma)
# note all the correlations are smaller than the original 0.6
# change Sigma, adding a negative correlation
Sigma[1,2] <- -0.15
Sigma[2,1] <- Sigma[1,2]
Sigma
# checking whether Sigma is still positive definite
eigen(Sigma)$values # all >0, OK
contord(marginal, Sigma)

---

**corrcheck**  
*Checking correlations for feasibility*

**Description**

The function returns the lower and upper bounds of the correlation coefficients of each pair of ordinal/discrete variables given their marginal distributions, i.e., returns the range of feasible bivariate correlations.

**Usage**

```
corrcheck(marginal, support = list(), Spearman = FALSE)
```

**Arguments**

- `marginal`: a list of $k$ elements, where $k$ is the number of variables. The $i$-th element of `marginal` is the vector of the cumulative probabilities defining the marginal distribution of the $i$-th component of the multivariate variable. If the $i$-th component has $k_i$ categories, the $i$-th element of `marginal` will contain $k_i - 1$ probabilities (the $k_i$-th is obviously 1 and shall not be included).

- `support`: a list of $k$ elements, where $k$ is the number of variables. The $i$-th element of `support` contains the ordered values of the support of the $i$-th variable. By default, the support of the $i$-th variable is $1, 2, ..., k_i$

- `Spearman`: `TRUE` if we consider Spearman's correlation, `FALSE` (default) if we consider Pearson's correlation

**Value**

The functions returns a list of two matrices: the former contains the lower bounds, the latter the upper bounds of the feasible pairwise correlations (on the extra-diagonal elements)

**Author(s)**

Alessandro Barbiero, Pier Alda Ferrari
ordcont

See Also

contord, ordcont, ordsample

Examples

# four variables
k <- 4
# with 2, 3, 4, and 5 categories (Likert scales, by default)
kj <- c(2,3,4,5)
# and these marginal distributions (set of cumulative probabilities)
marginal <- list(c(0.4, c(0.6,0.9), c(0.4,0.2,0.4), c(0.6,0.7,0.8,0.9))
corrcheck(marginal) # lower and upper bounds for Pearson's rho
corrcheck(marginal, Spearman=TRUE) # lower and upper bounds for Spearman's rho
# change the supports
support <- list(c(0,1), c(1,2,4), c(1,2,3,4), c(0,1,2,5,10))
corrcheck(marginal, support=support) # updated bounds

ordcont

Computing the "intermediate" correlation matrix for the multivariate standard normal in order to achieve the "target" correlation matrix for the ordinal/discrete variables

Description

The function computes the correlation matrix of the $k$-dimensional standard normal r.v. yielding the desired correlation matrix Sigma for the $k$-dimensional r.v. with desired marginal distributions marginal

Usage

ordcont(marginal, Sigma, support=list(), Spearman=FALSE, epsilon=1e-06, maxit=100)

Arguments

marginal a list of $k$ elements, where $k$ is the number of variables. The $i$-th element of marginal is the vector of the cumulative probabilities defining the marginal distribution of the $i$-th component of the multivariate variable. If the $i$-th component has $k_i$ categories, the $i$-th element of marginal will contain $k_i - 1$ probabilities (the $k_i$-th is obviously 1 and shall not be included).

Sigma the target correlation matrix of the ordinal/discrete variables

support a list of $k$ elements, where $k$ is the number of variables. The $i$-th element of support contains the ordered values of the support of the $i$-th variable. By default, the support of the $i$-th variable is $1, 2, ..., k_i$

Spearman if TRUE, the function finds Spearman’s correlations (and it is not necessary to provide support), if FALSE (default) Pearson’s correlations

epsilon the maximum tolerated error among target and actual correlations

maxit the maximum number of iterations allowed for the algorithm
Value

- A list of five elements:
  - `SigmaC`: The correlation matrix of the multivariate standard normal variable.
  - `SigmaO`: The actual correlation matrix of the discretized variables (it should approximately coincide with the target correlation matrix `sigma`).
  - `Sigma`: The target correlation matrix of the ordinal/discrete variables.
  - `niter`: A matrix containing the number of iterations performed by the algorithm, one for each pair of variables.
  - `maxerr`: The actual maximum error (the absolute maximum deviation between actual and target correlations of the ordinal/discrete variables).

Note

For some choices of `marginal` and `Sigma`, there may not exist a feasible \( k \)-variate probability mass function or the algorithm may not provide a feasible correlation matrix `SigmaC`. In this case, the procedure stops and exits with an error. The value of the maximum tolerated absolute error `epsilon` on the elements of the correlation matrix for the target r.v. can be set by the user: a value between 1e-6 and 1e-2 seems to be an acceptable compromise assuring both the precision of the results and the convergence of the algorithm; moreover, a maximum number of iterations can be chosen (maxit), in order to avoid possible endless loops.

Author(s)

Alessandro Barbiero, Pier Alda Ferrari

See Also

- `contord`, `ordsample`, `corrcheck`

Examples

```r
# Consider a 4-dimensional ordinal variable
k <- 4
# with different number of categories
kj <- c(2,3,4,5)
# and uniform marginal distributions
marginal <- list(P.ULc(1/OSLROS)L c(1/OULROULSOULTOU))
corrcheck(marginal)
# and the following correlation matrix
sigma <- matrix(c(1,0.5,0.4,0.3,0.5,1,0.5,0.4,0.4,0.5,1,0.5,0.3,0.4,0.5,1),4,4,byrow=TRUE)
Sigma
# the correlation matrix of the standard 4-dimensional standard normal
# ensuring Sigma is
res <- ordcont(marginal, Sigma)
res[[1]]
# change some marginal distributions
marginal <- list(P.SL c(1/OULROULSOU)L c(P.1LP.RLP.TLP.6))
corrcheck(marginal)
```
# and notice how the correlation matrix of the multivariate normal changes...
res <- ordcont(marginal, Sigma)
res[[1]]

# change Sigma, adding a negative correlation
Sigma[1,2] <- -0.2
Sigma[2,1] <- Sigma[1,2]
Sigma

# checking whether Sigma is still positive definite
eigen(Sigma)$values  # all >0, OK
res <- ordcont(marginal, Sigma)
res[[1]]

---

ordsample

### Drawing a sample of ordinal/discrete data

**Description**

The function draws a sample from a multivariate ordinal/discrete variable with correlation matrix Sigma and pre-specified marginals marginal

**Usage**

```r
ordsample(n, marginal, Sigma, support=list(), Spearman=FALSE, cormat="ordinal")
```

**Arguments**

- `n` the sample size
- `marginal` a list of \(k\) elements, where \(k\) is the number of variables. The \(i\)-th element of marginal is the vector of the cumulative probabilities defining the marginal distribution of the \(i\)-th component of the multivariate variable. If the \(i\)-th component has \(k_i\) categories, the \(i\)-th element of marginal will contain \(k_i - 1\) probabilities (the \(k_i\)-th is obviously 1 and shall not be included).
- `Sigma` the target correlation matrix of the ordinal/discrete variables
- `support` a list of \(k\) elements, where \(k\) is the number of variables. The \(i\)-th element of support contains the ordered values of the support of the \(i\)-th variable. By default, the support of the \(i\)-th variable is 1, 2, ..., \(k_i\)
- `Spearman` if TRUE, the function finds Spearman’s correlations (and it is not necessary to provide support), if FALSE (default) Pearson’s correlations
- `cormat` ordinal if the Sigma in input is the target correlation matrix of ordinal/discrete variables; continuous if the Sigma in input is the intermediate correlation matrix of the multivariate standard normal

**Value**

a \(n \times k\) matrix of discrete/ordinal data drawn from the \(k\)-variate discrete/ordinal r.v. with the desired marginal distributions and correlation matrix
Author(s)
Alessandro Barbiero, Pier Alda Ferrari

See Also
contord, ordcont, corrcheck

Examples

# Example 1
# draw a sample from a bivariate ordinal variable
# with 4 of categories and asymmetrical marginal distributions
# and correlation coefficient 0.6 (to be checked)
k <- 2
marginal <- list(c(0.1, 0.3, 0.6), c(0.4, 0.7, 0.9))
corrcheck(marginal) # check ok
Sigma <- matrix(c(1, 0.6, 0.6, 1), 2, 2)
# sample size 1000
n <- 1000
# generate a sample of size n
m <- ordsample(n, marginal, Sigma)
head(m)
# sample correlation matrix
cor(m) # compare it with Sigma
cumsum(table(m[,1]))/n
cumsum(table(m[,2]))/n # compare it with the two marginal distributions

# Example 1bis
# draw a sample from a bivariate ordinal variable
# with 4 of categories and asymmetrical marginal distributions
# and Spearman correlation coefficient 0.6 (to be checked)
k <- 2
marginal <- list(c(0.1, 0.3, 0.6), c(0.4, 0.7, 0.9))
corrcheck(marginal, Spearman=TRUE) # check ok
Sigma <- matrix(c(1, 0.6, 0.6, 1), 2, 2)
# sample size 1000
n <- 1000
# generate a sample of size n
m <- ordsample(n, marginal, Sigma, Spearman=TRUE)
head(m)
# sample correlation matrix
cor(rank(m[,1]), rank(m[,2])) # compare it with Sigma
cumsum(table(m[,1]))/n
cumsum(table(m[,2]))/n # compare it with the two marginal distributions

# Example 2
# draw a sample from a 4-dimensional ordinal variable
# with different number of categories and uniform marginal distributions
# and different correlation coefficients
k <- 4
marginal <- list(0.5, c(1/3,2/3), c(1/4,2/4,3/4), c(1/5,2/5,3/5,4/5))
corrcheck(marginal)
# select a feasible correlation matrix
Sigma <- matrix(c(1,0.5,0.4,0.3,0.5,1,0.5,0.4,0.4,0.4,0.5,1,0.5,0.3,0.4,0.5,1),4,4,byrow=TRUE)
Sigma
# sample size 100
n <- 100
# generate a sample of size n
set.seed(1)
m <- ordsample(n, marginal, Sigma)
# sample correlation matrix
cor(m) # compare it with Sigma
cumsum(table(m[4]))/n # compare it with the fourth marginal
head(m)
# or equivalently...
set.seed(1)
res <- ordcont(marginal, Sigma)
res[[1]] # the intermediate correlation matrix of the multivariate normal
m <- ordsample(n, marginal, res[[1]], cormat="continuous")
head(m)
Index

*Topic datagen
   contord, 3
   ordcont, 5
   ordsample, 7

*Topic distribution
   contord, 3
   corrcheck, 4
   ordcont, 5
   ordsample, 7

*Topic htest
   contord, 3
   corrcheck, 4
   ordcont, 5
   ordsample, 7

*Topic models
   contord, 3
   corrcheck, 4
   ordcont, 5
   ordsample, 7

*Topic multivariate
   contord, 3
   corrcheck, 4
   ordcont, 5
   ordsample, 7

*Topic package
   GenOrd-package, 2
   contord, 2, 3, 5, 6, 8
   corrcheck, 2, 3, 4, 6, 8
   GenOrd-package, 2
   ordcont, 2, 3, 5, 8
   ordsample, 2, 3, 5, 6, 7