Package 'PoissonMultinomial'

January 13, 2022

Type Package

Title The Poisson-Multinomial Distribution

Version 1.0

Date 2022-01-12

Maintainer Yili Hong <yilihong@vt.edu>

Description Implementation of the exact, normal approximation, and simulation-based methods for computing the probability mass function (pmf) and cumulative distribution function (cdf) of the Poisson-Multinomial distribution, together with a random number generator for the distribution. The exact method is based on multi-dimensional fast Fourier transformation (FFT) of the characteristic function of the Poisson-Multinomial distribution. The normal approximation method uses a multivariate normal distribution to approximate the pmf of the distribution based on central limit theorem. The simulation method is based on the law of large numbers. Details about the methods are available in Lin, Wang, and Hong (2022) <arXiv:2201.04237>.

License GPL (>= 2)

Encoding UTF-8

Imports mvtnorm, Rcpp

LinkingTo Rcpp, RcppArmadillo

SystemRequirements fftw3(>=3.3)

RoxygenNote 7.1.1

NeedsCompilation yes

Author Yili Hong [aut, cre],
Zhengzhi Lin [aut, ctb],
Yueyao Wang [aut, ctb],
Florian Junge [aut, ctb]

Repository CRAN

Date/Publication 2022-01-13 20:22:42 UTC

R topics documented:

dpmd *Probability Mass Function of Poisson-Multinomial Distribution*

Description

Computes the pmf of Poisson-Multinomial distribution (PMD), specified by the success probability matrix, using various methods. This function is capable of computing all probability mass points as well as of pmf at certain point(s).

Usage

```
dpmd(pmat, xmat = NULL, method = "DFT-CF", B = 1000)
```

Arguments

`pmat`	An $n \times m$ success probability matrix. Here, n is the number of independent trials, and m is the number of categories. Each row of pmat describes the success probability for the corresponding trial and it should add up to 1.
`xmat`	A matrix with m columns that specifies where the pmf is to be computed. Each row of the matrix should has the form $x = (x_1, \ldots, x_m)$ which is used for computing $P(X_1 = x_1, \ldots, X_m = x_m)$, the values of x should sum up to n. It can be a vector of length m. If `xmat` is `NULL`, the pmf at all probability mass points will be computed.
`method`	Character string stands for the method selected by users to compute the cdf. The method can only be one of the following three: `"DFT-CF"`, `"NA"`, `"SIM"`.
`B`	Number of repeats used in the simulation method. It is ignored for methods other than the `"SIM"` method.

Details

Consider n independent trials and each trial leads to a success outcome for exactly one of the m categories. Each category has varying success probabilities from different trials. The Poisson multinomial distribution (PMD) gives the probability of any particular combination of numbers of successes for the m categories. The success probabilities form an $n \times m$ matrix, which is called the success probability matrix and denoted by `pmat`. For the methods we applied in `dpmd`, `"DFT-CF"` is an exact method that computes all probability mass points of the distribution, using multi-dimensional FFT algorithm. When the dimension of `pmat` increases, the computation burden of `"DFT-CF"` may challenge the capability of a computer because the method automatically computes all probability mass points regardless of the input of `xmat`.

`"SIM"` is a simulation method that generates random samples from the distribution, and uses relative frequency to estimate the pmf. Note that the accuracy and running time will be affected by user choice of `B`. Usually `B`=1e5 or 1e6 will be accurate enough. Increasing `B` to larger than 1e8 will heavily increase the computational burden of the computer.

`"NA"` is an approximation method that uses a multivariate normal distribution to approximate the pmf at the points specified in `xmat`. This method requires an input of `xmat`.

Notice if `xmat` is not specified then it will be set as `NULL`. In this case, `dpmd` will compute the entire pmf if the chosen method is `"DFT-CF"` or `"SIM"`. If `xmat` is provided, only the pmf at the points specified by `xmat` will be outputted.

Value

For a given `xmat`, `dpmd` returns the pmf at points specified by `xmat`.

If `xmat` is `NULL`, all probability mass points for the distribution specified by the success probability matrix `pmat` will be computed, and the results are stored and outputted in a multi-dimensional array, denoted by `res`. Note the dimension of `pmat` is $n \times m$, thus `res` will be an $(n+1)^{(m-1)}$ array. Then the value of the pmf $P(X_1 = x_1, \ldots, X_m = x_m)$ can be extracted as $res[x_1 + 1, \ldots, x_{m-1} + 1]$.

For example, for the `pmat` matrix in the example section, the array element `res[1,2,1]=0.90` gives the value of the pmf $P(X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 2) = 0.90$.

References

Lin, Z., Wang, Y., and Hong, Y. (2022). The Poisson Multinomial Distribution and Its Applications in Voting Theory, Ecological Inference, and Machine Learning, arXiv:2201.04237.

Examples

```
pp <- matrix(c(.1, .1, .1, .7, .1, .3, .3, .3, .5, .2, .1, .2), nrow = 3, byrow = TRUE)
x <- c(0,0,1,2)
x1 <- matrix(c(0,0,1,2,2,1,0,0),nrow=2,byrow=TRUE)

dpmd(pmat = pp)
dpmd(pmat = pp, xmat = x1)
dpmd(pmat = pp, xmat = x)

dpmd(pmat = pp, xmat = x, method = "NA" )
dpmd(pmat = pp, xmat = x1, method = "NA" )

dpmd(pmat = pp, method = "SIM", B = 1e3)
dpmd(pmat = pp, xmat = x, method = "SIM", B = 1e3)
dpmd(pmat = pp, xmat = x1, method = "SIM", B = 1e3)
```

---

ppmd — *Cumulative Distribution Function of Poisson-Multinomial Distribution*

Description

Computes the cdf of Poisson-Multinomial distribution that is specified by the success probability matrix, using various methods.

Usage

```
ppmd(pmat, xmat, method = "DFT-CF", B = 1000)
```

Arguments

`pmat`	An $n \times m$ success probability matrix. Here, n is the number of independent trials, and m is the number of categories. Each row of pmat describes the success probability for the corresponding trial and it should add up to 1.
`xmat`	A matrix with m columns. Each row has the form $x = (x_1, \ldots, x_m)$ for computing the cdf at x, $P(X_1 \leq x_1, \ldots, X_m \leq x_m)$. It can also be a vector with length m.
`method`	Character string stands for the method selected by users to compute the cdf. The method can only be one of the following three: `"DFT-CF"`, `"NA"`, `"SIM"`.
`B`	Number of repeats used in the simulation method. It is ignored for methods other than the `"SIM"` method.

Details

See Details in `dpmd` for the definition of the PMD, the introduction of notation, and the description of the three methods (`"DFT-CF"`, `"NA"`, and `"SIM"`). `ppmd` computes the cdf by adding all probability mass points within hyper-dimensional space bounded by `x` as in the cdf.

Value

The value of cdf $P(X_1 \leq x_1, \ldots, X_m \leq x_m)$ at $x = (x_1, \ldots, x_m)$.

Examples

```
pp <- matrix(c(.1, .1, .1, .7, .1, .3, .3, .3, .5, .2, .1, .2), nrow = 3, byrow = TRUE)
x <- c(3,2,1,3)
x1 <- matrix(c(0,0,1,2,2,1,0,0),nrow=2,byrow=TRUE)

ppmd(pmat = pp, xmat = x)
ppmd(pmat = pp, xmat = x1)

ppmd(pmat = pp, xmat = x, method = "NA")
ppmd(pmat = pp, xmat = x1, method = "NA")

ppmd(pmat = pp, xmat = x, method = "SIM", B = 1e3)
ppmd(pmat = pp, xmat = x1, method = "SIM", B = 1e3)
```

---

rpmd — *Poisson-Multinomial Distribution Random Number Generator*

Description

Generates random samples from the PMD specified by the success probability matrix.

Usage

```
rpmd(pmat, s = 1)
```

Arguments

`pmat` An $\mathrm{n} \times \mathrm{m}$ success probability matrix, where n is the number of independent trials and m is the number of categories. Each row of pmat contains the success probabilities for the corresponding trial, and each row adds up to 1.

`s` The number of samples to be generated.

Value

An $s \times m$ matrix of samples, each row stands for one sample from the PMD with success probability matrix `pmat`.

Examples

```
pp <- matrix(c(.1, .1, .1, .7, .1, .3, .3, .3, .5, .2, .1, .2), nrow = 3, byrow = TRUE)

rpmd(pmat = pp, s = 5)
```

Index