Package ‘dirmult’

March 21, 2022

Version 0.1.3-5
Date 2022-03-08
Title Estimation in Dirichlet-Multinomial Distribution
Author Torben Tvedebrink <tvede@math.aau.dk>
Maintainer Torben Tvedebrink <tvede@math.aau.dk>
Description Estimate parameters in Dirichlet-Multinomial and compute log-likelihoods.
Depends R (>= 2.5.0)
License GPL (>= 2)
NeedsCompilation no
Repository CRAN
Date/Publication 2022-03-21 10:30:02 UTC

R topics documented:

adapGridProf ......................................................... 2
dirmult .............................................................. 3
dirmult.summary .................................................... 4
equalTheta ............................................................. 5
estProfLogLik ....................................................... 6
gridProf .............................................................. 7
nullTest .............................................................. 8
rdirichlet ............................................................ 8
simPop ............................................................ 9
us ............................................................ 10
weirMoM ............................................................ 10

Index 12
adapGridProf

Profile log-likelihood of Dirichlet-multinomial model

Description

Computes the profile log-likelihood of \( \ell(\pi, \theta; x) \) for an interval determined by a given difference in log-likelihood value from the maximum log-likelihood value.

Usage

adapGridProf(data, delta, stepsize=50)

Arguments

data A matrix or table with counts. Rows represent subpopulations and columns the different categories of the data. Zero rows or columns are automatically removed.
delta The difference between max of log-likelihood and the profile log-likelihood. May be used to construct approximate confidence intervals, e.g. with delta = qchisq(0.95,df=1)*2.
stepsize The stepsize used when stepping left/right of the MLE. The stepsize used by the algorithm is given by the MLE of theta divided by stepsize. Default value is 50.

Value

Gives a data frame with theta values and associated profile log-likelihood values.

See Also

estProfLogLik

Examples

data(us)
fit <- dirmult(us[[1]],epsilon=10^(-12),trace=FALSE)
adapGridProf(us[[1]],delta=0.5)
## Not run: adapGridProf(us[[1]],delta=qchisq(0.95,df=1)*2)
Parameter estimation in Dirichlet-multinomial distribution

Description

Consider allele frequencies from different subpopulations. The allele counts, $X$, (or equivalently allele frequencies) are expected to vary between subpopulation. This variability are sometimes referred to as identity-by-decent, but may be modelled as overdispersion due to intra-class correlation $\theta$. The allele counts within each subpopulation is assumed to follow a multinomial distribution conditioned on the allele probabilities, $\pi_1, \ldots, \pi_{k-1}$. When $\pi$ follows a Dirichlet distribution the marginal distribution of $X$ is Dirichlet-multinomial with parameters $\pi$ and $\theta$ with density:

$$P(X = x) = \binom{n}{x} \frac{\prod_{j=1}^{k} \prod_{r=1}^{x_j} \{\pi_j(1 - \theta) + (r - 1)\theta\}}{\prod_{r=1}^{n} (1 - \theta + (r - 1)\theta)}.$$

Using an alternative parameterization the density may be written as:

$$P(X = x) = \binom{n}{x} \frac{\prod_{j=1}^{k} \Gamma(x_j + \gamma_j)}{\Gamma(n + \gamma_+)} \prod_{j=1}^{k} \frac{\Gamma(x_j + \gamma_j)}{\Gamma(\gamma_j)},$$

where $\gamma_+ = (1 - \theta)/\theta$ and $\gamma_j = \pi_j\theta$.

This formulation second parameterization is used in the iterations since it converges much faster than the original parameterization. The function `dirmult` estimates the parameters $\gamma$ in the Dirichlet-multinomial distribution and transform these into $\pi_1, \ldots, \pi_{k-1}$ and $\theta$.

Usage

dirmult(data, init, initscalar, epsilon=10^(-4), trace=TRUE, mode)

Arguments

data A matrix or table with counts. Rows represent subpopulations and columns the different categories of the data. Zero rows or columns are automatically removed.

init Initial values for the $\gamma$-vector. Default is empty implying the column-proportions are used as initial values.

initscalar Initial value for $(1 - \theta)/\theta$. Default value is (1-MoM)/MoM where MoM a the method of moment estimate.

epsilon Convergence tolerance. On termination the difference between to succeeding log-likelihoods must be smaller than epsilon.

trace Logical. If TRUE the parameter estimates and log-likelihood value is printed to the screen after each iteration, otherwise no out-put is produces while iterating.

mode Takes values "obs" (default) or "exp" determining whether the observed or expected FIM should be used in the Fisher Scoring. All other arguments produces an error message, but the observed FIM is used in the iterations.
Value

Returns a list containing:

- **loglik**: The final log-likelihood value.
- **ite**: Number of iterations used.
- **gamma**: A vector of $\gamma$ estimates.
- **pi**: A vector of $\pi$ estimates.
- **theta**: Estimated $\theta$-value.

See Also

dirmult.summary

Examples

data(us)
fit <- dirmult(us[[1]], epsilon=10^-4, trace=FALSE)
dirmult.summary(us[[1]], fit)

dirmult.summary

Summary table of parameter estimates from dirmult

Description

Produces a summary table based on the estimated parameters from dirmult. The table contains MLE estimates and standard errors together with method of moment (MoM) estimates and standard errors based on MoM estimates from 'Weir and Hill (2002)'.

Usage

dirmult.summary(data, fit, expectedFIM=FALSE)

Arguments

- **data**: A matrix or table with counts. Rows represent subpopulations and columns the different categories of the data. Zero rows or columns are automatically removed.
- **fit**: Output from dirmult used on the same data table as above.
- **expectedFIM**: Logical. Determines whether the observed or expected Fisher Information Matrix should be used. For speed use observed (i.e. FALSE) - for accuracy (and theoretical support) use expected (i.e. TRUE).

Value

Summary table with estimates and standard errors for $\pi$ and $\theta$. 
equalTheta

See Also

dirmult

Examples

data(us)
fit <- dirmult(us[[1]],epsilon=10^(-4),trace=FALSE)
dirmult.summary(us[[1]],fit)

equalTheta

Test whether theta is equal for several tables

Description

Estimates parameters $\pi$ for each table under the constraint that $\theta$ is equal for all tables.

Usage

equalTheta(data, theta, epsilon=10^(-4), trace=TRUE, initPi, maxit=1000)

Arguments

data A list of matrix or table with counts. Rows in the tables represent subpopulations and columns the different categories of the data. Zero columns are automatically removed.
theta Initial value of the common theta parameter.
epsilon Tolerance of the convergence, see dirmult.
trace Logical. TRUE: print estimates while iterating.
initPi Initial values for each pi vector (one of each table).
maxit Maximum number of iterations.

Value

Returns a list similar to the output of dirmult.

See Also

dirmult
Examples

```r
## Not run: data(us)
fit <- lapply(us[1:2], dirmult, epsilon=10^(-12), trace=FALSE)
thetas <- unlist(lapply(fit, function(x) x$theta))
logliks <- unlist(lapply(fit, function(x) x$loglik))
fit1 <- equalTheta(us[c(1:2)], theta=mean(thetas), epsilon=10^(-12))
lr <- -2*(fit1$loglik-sum(logliks))
1-pchisq(lr, df=1)
fit1$theta[[1]]

## End(Not run)
```

---

**estProfLogLik**  Profile log-likelihood of Dirichlet-multinomial model

**Description**

Computes the profile log-likelihood of $\ell(\pi, \theta; x)$ for a given value of $\theta$, i.e. $\hat{\ell}(\theta) = \max_{\pi} \ell(\pi, \theta; x)$.

**Usage**

```
estProfLogLik(data, theta, epsilon=10^(-4), trace=TRUE, initPi, maxit=1000)
```

**Arguments**

- `data` A matrix or table with counts. Rows represent subpopulations and columns the different categories of the data. Zero rows or columns are automatically removed.
- `theta` The theta-value of which the profile log-likelihood is to be computed.
- `epsilon` Tolerance used in the iterations. Succeeding log-likelihood values need to be within epsilon for convergence.
- `trace` Logical. Whether parameter estimates and log-likelihood values should be printed to the screen while iterating.
- `initPi` Initial pi vector.
- `maxit` Maximum number of iterations. Default is 1000 and will often not be envoked, but if theta is to extreme compared to the MLE of theta the log-likelihood may misbehave near theta.

**Value**

Gives a list of components (similar to output from `dirmult` where loglik and lambda (the Lagrange multiplier) are the most interesting.

**See Also**

- `dirmult`
Examples

data(us)
fit <- dirmult(us[[1]], epsilon=10^(-12), trace=FALSE)
estProfLogLik(us[[1]], fit$theta*1.2, epsilon=10^(-12), trace=FALSE)

gridProf

Profile log-likelihood of Dirichlet-multinomial model

Description

Computes the profile log-likelihood of $\ell(\pi, \theta; x)$ for a given sequence of $\theta$ by calling estProfLogLik.

Usage

gridProf(data, theta, from, to, len)

Arguments

- **data**: A matrix or table with counts. Rows represent subpopulations and columns the different categories of the data. Zero rows or columns are automatically removed.
- **theta**: A theta-value used as offset for the interval: $[\theta+\text{from}; \theta+\text{to}]$.
- **from**: Left endpoint in the interval: $[\theta+\text{from}; \theta+\text{to}]$.
- **to**: Right endpoint in the interval: $[\theta+\text{from}; \theta+\text{to}]$.
- **len**: Number of points in the $[\text{from}; \text{to}]$ interval. Similar to the $\text{len}$ argument in seq.

Value

Gives a data frame with theta values and associated profile log-likelihood values.

See Also

estProfLogLik

Examples

data(us)
fit <- dirmult(us[[1]], epsilon=10^(-12), trace=FALSE)
## Not run: grid <- gridProf(us[[1]], fit$theta, from=-0.001, to=0.001, len=10)
plot(loglik ~ theta, data=grid, type="l")
## End(Not run)
### nullTest

*Simulation based test for null-hypothesis, $H_0: \theta=0$*

**Description**

Simulates data sets under the null-hypothesis, $H_0 : \theta = 0$. This corresponds to an ordinary multinomial model without any overdispersion. Based on the returned data frame simulated $p$-values may be computed.

**Usage**

```r
nullTest(data, m=1000, prec=6)
```

**Arguments**

- **data**: A matrix or table with counts. Rows represent subpopulations and columns the different categories of the data. Zero rows or columns are automatically removed.
- **m**: Number of simulated data tables.
- **prec**: The tolerance of the iterations. Corresponds to epsilon=1e-prec in `dirmult`.

**Value**

Returns a data frame with theta estimates and log-likelihood values.

**See Also**

- `dirmult`

**Examples**

```r
data(us)
## Not run: nullTest(us[[1]], m=50)
```

---

### rdirichlet

*Simulate from Dirichlet distribution*

**Description**

Simulates from a Dirichlet distribution

**Usage**

```r
rdirichlet(n=1, alpha)
```
simPop

Arguments

n
The number of samples.

alpha
The shape parameters, need to be positive.

Value

Return an n x length(alpha) matrix where each row is drawn from a Dirichlet.

See Also
dirmult

Examples

rdirichlet(n=100, alpha=rep(1,10))

Description

Simulate data from Dirichlet-multinomial distribution

Usage

simPop(J=10, K=20, n, pi, theta)

Arguments

J
The number of subpopulations sampled.

K
Number of different alleles. If argument pi is given, the length of pi is used as K.

n
The number of alleles sampled in each subpopulation. If scalar repeated for all subpopulations, otherwise a vector of length J is needed with subpopulation specific total sampled alleles.

pi
Vector of allele probabilities. If missing a random vector of length K is generated.

theta
The theta-value used for simulations.

Value

Return an J x K matrix with allelic counts.

See Also
dirmult
Examples

```r
simPop(n=100, theta=0.03)
```

us

*Allele counts for six US subpopulations.*

Description

9 STR loci were typed in sample populations of African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians.

Format

A list of tables with allele counts.

Source


References


weirMoM

*Method of moment estimator of theta*

Description

Estimates \( \theta \) using a method of moment (MoM) estimate by 'Weir and Hill (2002).'

Usage

```r
weirMoM(data, se=FALSE)
```

Arguments

- `data` A matrix or table with counts. Rows represent subpopulations and columns the different categories of the data. Zero rows or columns are automatically removed.
- `se` Logical. Determines if a standard error of theta should be computed or not. The variance is based on an expression by Li cited in 'Weir and Hill (2002)'.

Value

MoM-estimate (and standard error) of theta.
weirMoM

References


See Also

dirmult.summary

Examples

data(us)
weirMoM(us[[1]], se=TRUE)
Index

* Dirichlet-multinomial
  adapGridProf, 2
  dirmult, 3
  dirmult.summary, 4
  equalTheta, 5
  estProfLogLik, 6
  gridProf, 7
  nullTest, 8
  rdirichlet, 8
  simPop, 9
  weirMoM, 10

* Genetics
  adapGridProf, 2
  dirmult, 3
  dirmult.summary, 4
  equalTheta, 5
  estProfLogLik, 6
  gridProf, 7
  nullTest, 8
  rdirichlet, 8
  simPop, 9
  weirMoM, 10

* Overdispersion
  adapGridProf, 2
  dirmult, 3
  dirmult.summary, 4
  equalTheta, 5
  estProfLogLik, 6
  gridProf, 7
  nullTest, 8
  rdirichlet, 8
  simPop, 9
  weirMoM, 10

* datasets
  us, 10

adapGridProf, 2

dirmult, 3, 4–6, 8, 9
dirmult.summary, 4, 4, 11

equalTheta, 5
estProfLogLik, 2, 6, 7
gridProf, 7
nullTest, 8
rdirichlet, 8
seq, 7
simPop, 9
us, 10
weirMoM, 10