

# Package ‘dirmult’

September 24, 2009

**Version** 0.1.2

**Date** 2009-09-24

**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 profile log-likelihoods.

**License** GPL (>= 2)

**Repository** CRAN

**Date/Publication** 2009-09-24 10:40:25

## R topics documented:

adapGridProf . . . . .	2
dirmult . . . . .	3
dirmult.summary . . . . .	4
equalTheta . . . . .	5
estProfLogLik . . . . .	6
gridProf . . . . .	7
nullTest . . . . .	8
simPop . . . . .	8
us . . . . .	9
weirMoM . . . . .	10
<b>Index</b>	<b>11</b>

---

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 <code>delta = qchisq(0.95,df=1)*2</code> .
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 <code>stepsize</code> . 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)
```

## 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-descent, 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,  $\pi_1, \dots, \pi_{k-1}$ . 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:

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

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  $\pi_1, \dots, \pi_{k-1}$  and  $\theta$ .

## Usage

```
dirmult(data, init, initscalar=30, epsilon=10^(-4), trace=TRUE, mode)
```

## Arguments

<code>data</code>	A matrix or table with counts. Rows represent subpopulations and columns the different categories of the data. Zero rows or columns are automatically removed.
<code>init</code>	Initial values for the $\gamma$ -vector. Default is empty implying the column-proportions are used as initial values.
<code>initscalar</code>	Initial value for $\theta$ . Default value is 30 which approximately corresponds to a $\theta$ value of 0.03 which is commonly assumed in genetics.
<code>epsilon</code>	Convergence tolerance. On termination the difference between to succeeding log-likelihoods must be smaller than epsilon.
<code>trace</code>	Logical. If TRUE the parameter estimates and log-likelihood value is printed to the screen after each iteration, otherwise no out-put is produced while iterating.
<code>mode</code>	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^(-12),trace=FALSE)
dirmult.summary(us[[1]],fit)
```

---

dirmult.summary	<i>Summary table of parameter estimates from dirmult</i>
-----------------	--

---

**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)
```

**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 <code>dirmult</code> used on the same data table as above.

**Value**

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

**See Also**

[dirmult](#)

**Examples**

```
data(us)
fit <- dirmult(us[[1]],epsilon=10^(-12),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 <a href="#">dirmult</a> .
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**

```
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]]
```

---

 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 invoked, but if theta is too 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+from; theta+to].
from	Left endpoint in the interval: [theta+from; theta+to].
to	Right endpoint in the interval: [theta+from; theta+to].
len	Number of points in the [from; to] interval. Similar to the <code>len</code> argument in <a href="#">seq</a> .

**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)
grid <- gridProf(us[[1]], fit$theta, from=-0.001, to=0.001, len=10)
plot(loglik ~ theta, data=grid, type="l")
```

---

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

```
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 <a href="#">dirmult</a> .

### Value

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

### See Also

[dirmult](#)

### Examples

```
data(us)
nullTest(us[[1]], m=50)
```

---

simPop

*Simulate data from Dirichlet-multinomial distribution*


---

### Description

Simulates data using user defined  $\theta$  value and allele probabilities in the reference population,  $\pi$ .

### Usage

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



**Arguments**

<code>J</code>	The number of subpopulations sampled.
<code>K</code>	Number of different alleles. If argument <code>pi</code> is given, the length of <code>pi</code> is used as <code>K</code> .
<code>n</code>	The number of alleles sampled in each subpopulation. If scalar repeated for all subpopulations, otherwise a vector of length <code>J</code> is needed with subpopulation specific total sampled alleles.
<code>pi</code>	Vector of allele probabilities. If missing a random vector of length <code>K</code> is generated.
<code>theta</code>	The theta-value used for simulations.

**Value**

Return an `J` x `K` matrix with allelic counts.

**See Also**

[dirmult](#)

**Examples**

```
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.

**Usage**

```
us
```

**Format**

A list of tables with allele counts.

**Source**

<http://www.fbi.gov/hq/lab/fsc/backissu/july1999/budowle.htm>

**References**

Budowle, B., Moretti, T. R., Baumstark, A. L., Defenbaugh, D. A., and Keys, K. M. Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians, *Journal of Forensic Sciences*. 1999.

---

`weirMoM`*Method of moment estimator of theta*

---

**Description**

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

**Usage**

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

**Arguments**

<code>data</code>	A matrix or table with counts. Rows represent subpopulations and columns the different categories of the data. Zero rows or columns are automatically removed.
<code>se</code>	Logical. Determines if a standard error of theta could 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.

**References**

Weir, B. S. and W. G. Hill (2002). 'Estimating F-statistics'. *Annu Rev Genet* 36: 721-750

**See Also**

[`dirmult.summary`](#)

**Examples**

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

# Index

## \*Topic **datasets**

us, [9](#)

## \*Topic **htest**

adapGridProf, [1](#)

dirmult, [2](#)

dirmult.summary, [4](#)

equalTheta, [4](#)

estProfLogLik, [5](#)

gridProf, [6](#)

nullTest, [7](#)

simPop, [8](#)

weirMoM, [9](#)

adapGridProf, [1](#)

dirmult, [2](#), [4–8](#)

dirmult.summary, [3](#), [4](#), [10](#)

equalTheta, [4](#)

estProfLogLik, [2](#), [5](#), [6](#), [7](#)

gridProf, [6](#)

nullTest, [7](#)

seq, [7](#)

simPop, [8](#)

us, [9](#)

weirMoM, [9](#)