Package 'mix'

October 13, 2022

Version 1.0-11

Date 2022-05-31

Title Estimation/Multiple Imputation for Mixed Categorical and Continuous Data

Author Original by Joseph L. Schafer <jls@stat.psu.edu>.

Maintainer Brian Ripley <ripley@stats.ox.ac.uk>

Depends stats

Description Estimation/multiple imputation programs for mixed categorical and continuous data.

License Unlimited

LazyData yes

NeedsCompilation yes

Repository CRAN

Date/Publication 2022-05-31 10:54:22 UTC

R topics documented:

da.mix 2
dabipf.mix 3
ecm.mix 5
em.mix 6
getparam.mix 8
imp.mix 9
loglik.mix 10
mi.inference 11
prelim.mix 12
rngseed 13
stlouis 13

Index **15**

da.mix *Data Augmentation for Unrestricted General Location Model*

Description

Markov Chain Monte Carlo method for generating posterior draws of the parameters of the unrestricted general location model, given a matrix of incomplete mixed data. At each step, missing data are randomly imputed under the current parameter, and a new parameter value is drawn from its posterior distribution given the completed data. After a suitable number of steps are taken, the resulting value of the parameter may be regarded as a random draw from its observed-data posterior distribution. May be used together with `imp.mix` to create multiple imputations of the missing data.

Usage

```
da.mix(s, start, steps=1, prior=0.5, showits=FALSE)
```

Arguments

`s`	summary list of an incomplete data matrix created by the function `prelim.mix`.
`start`	starting value of the parameter. This is a parameter list such as one created by the function `em.mix`.
`steps`	number of data augmentation steps to be taken.
`prior`	Optional vector or array of hyperparameter(s) for a Dirichlet prior distribution. The default is the Jeffreys prior (all hyperparameters = .5). If structural zeros appear in the table, prior counts for these cells should be set to `NA`.
`showits`	if `TRUE`, reports the iterations so the user can monitor the progress of the algorithm.

Details

The prior distribution used by this function is a combination of a Dirichlet prior for the cell probabilities, an improper uniform prior for the within-cell means, and the improper Jeffreys prior for the covariance matrix. The posterior distribution is not guaranteed to exist, especially in sparse-data situations. If this seems to be a problem, then better results may be obtained by imposing restrictions on the parameters; see `ecm.mix` and `dabipf.mix`.

Value

A new parameter list. The parameter can be put into a more understandable format by the function `getparam.mix`.

Note

The random number generator seed must be set at least once by the function `rngseed` before this function can be used.

References

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, Chapter 9.

See Also

prelim.mix, getparam.mix, em.mix, and rngseed.

Examples

```
data(stlouis)
s <- prelim.mix(stlouis,3)  # preliminary manipulations
thetahat <- em.mix(s) # find ML estimate
rngseed(1234567)   # set random number generator seed
newtheta <- da.mix(s, thetahat, steps=100, showits=TRUE)  # take 100 steps
ximp1 <- imp.mix(s, newtheta) # impute under newtheta
```

---

dabipf.mix *Data Augmentation/Bayesian IPF Algorithm for Restricted General Location Models*

Description

Markov Chain Monte Carlo method for generating posterior draws of the parameters of the unrestricted general location model, given a matrix of incomplete mixed data. After a suitable number of steps are taken, the resulting value of the parameter may be regarded as a random draw from its observed-data posterior distribution. May be used together with imp.mix to create multiple imputations of the missing data.

Usage

```
dabipf.mix(s, margins, design, start, steps=1, prior=0.5,
           showits=FALSE)
```

Arguments

s	summary list of an incomplete data matrix created by the function prelim.mix.
margins	vector describing the sufficient configurations or margins in the desired loglinear model. The variables are ordered in the original order of the columns of x, so that 1 refers to x[,1], 2 refers to x[,2], and so on. A margin is described by the factors not summed over, and margins are separated by zeros. Thus c(1,2,0,2,3,0,1,3) would indicate the (1,2), (2,3), and (1,3) margins in a three-way table, i.e., the model of no three-way association.
design	design matrix specifying the relationship of the continuous variables to the categorical ones. The dimension is c(D,r) where D is the number of cells in the contingency table, and r is the number of effects which must be less than or equal to D. The order of the rows corresponds to the storage order of the cell probabilities in the contingency table; see getparam.mix for details.

start	starting value of the parameter. This is a parameter list such as one created by this function or by ecm.mix.
steps	number of steps of data augmentation-Bayesian IPF to be taken.
prior	Optional vector or array of hyperparameter(s) for a Dirichlet prior distribution. The default is the Jeffreys prior (all hyperparameters = .5). If structural zeros appear in the table, prior counts for these cells should be set to NA.
showits	if TRUE, reports the iterations so the user can monitor the progress of the algorithm.

Details

The prior distribution used by this function is a combination of a constrained Dirichlet prior for the cell probabilities, an improper uniform prior for the regression coefficients, and the improper Jeffreys prior for the covariance matrix. The posterior distribution is not guaranteed to exist, especially in sparse-data situations. If this seems to be a problem, then better results may be obtained by imposing restrictions further restrictions on the parameters.

Value

a new parameter list. The parameter can be put into a more understandable format by the function getparam.mix.

Note

The random number generator seed must be set at least once by the function rngseed before this function can be used.

The starting value should satisfy the restrictions of the model and should lie in the interior of the parameter space. A suitable starting value can be obtained by running ecm.mix, possibly with the prior hyperparameters set to some value greater than 1, to ensure that the mode lies in the interior.

References

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, Chapter 9.

See Also

prelim.mix, getparam.mix, ecm.mix, rngseed, imp.mix.

Examples

```
data(stlouis)
s <- prelim.mix(stlouis,3)      # do preliminary manipulations
margins <- c(1,2,3)       # saturated contingency table model
design <- diag(rep(1,12))  # identity matrix  D=no of cells
thetahat <- ecm.mix(s,margins,design) # find ML estimate
rngseed(1234567)       # random generator seed
newtheta <- dabipf.mix(s,margins,design,thetahat,steps=200)
ximp <- imp.mix(s,newtheta,stlouis)   # impute under newtheta
```

ecm.mix *ECM Algorithm for Restricted General Location Model*

Description

Computes maximum-likelihood estimates for the parameters of the general location model from an incomplete mixed dataset.

Usage

```
ecm.mix(s, margins, design, start, prior=1, maxits=1000,
        showits=TRUE, eps=0.0001)
```

Arguments

s	summary list of an incomplete data matrix x produced by the function prelim.mix.
margins	vector describing the sufficient configurations or margins in the desired loglinear model. The variables are ordered in the original order of the columns of x, so that 1 refers to x[,1], 2 refers to x[,2], and so on. A margin is described by the factors not summed over, and margins are separated by zeros. Thus c(1,2,0,2,3,0,1,3) would indicate the (1,2), (2,3), and (1,3) margins in a three-way table, i.e., the model of no three-way association.
design	design matrix specifying the relationship of the continuous variables to the categorical ones. The dimension is c(D,r) where D is the number of cells in the contingency table, and r is the number of effects which must be less than or equal to D. The order of the rows corresponds to the storage order of the cell probabilities in the contingency table; see getparam.mix for details.
start	optional starting value of the parameter. This is a list such as one created by this function or by dabipf.mix. If structural zeros appear in the table, start$pi should contain zeros in those positions and ones elsewhere. If no starting value is supplied, ecm.mix chooses its own appropriate starting value.
prior	Optional vector or array of hyperparameter(s) for a Dirichlet prior distribution. By default, uses a uniform prior on the cell probabilities. ECM finds the posterior mode, which under a uniform prior is the same as a maximum-likelihood estimate. If structural zeros appear in the table, hyperparameters for those cells should be set to NA..
maxits	maximum number of iterations performed. The algorithm will stop if the parameter still has not converged after this many iterations.
showits	if TRUE, reports the iterations of ECM so the user can monitor the progress of the algorithm.
eps	optional convergence criterion. The algorithm stops when the maximum relative difference in every parameter from one iteration to the next is less than or equal to this value.

Value

a list representing the maximum-likelihood estimates (or posterior mode) of the normal parameters. This list contains cell probabilities, cell means, and covariances. The parameter can be transformed back to the original scale and put into a more understandable format by the function getparam.mix.

Note

If zero cell counts occur in the complete-data table, the maximum likelihood estimate may not be unique, and the algorithm may converge to different stationary values depending on the starting value. Also, if zero cell counts occur in the complete-data table, the ML estimate may lie on the boundary of the parameter space.

References

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, Chapter 9.

See Also

prelim.mix, em.mix, getparam.mix, loglik.mix.

Examples

```
data(stlouis)
s <- prelim.mix(stlouis,3)         # preliminary manipulations
margins <- c(1,2,3)          # saturated loglinear model
design <- diag(rep(1,12))     # identity matrix, D=no of cells
thetahat <- ecm.mix(s,margins,design)  # should be same as em.mix(s)
loglik.mix(s,thetahat)             # loglikelihood at thetahat
```

---

em.mix *EM Algorithm for Unrestricted General Location Model*

Description

Computes maximum-likelihood estimates for the parameters of the unrestricted general location model from an incomplete mixed dataset.

Usage

```
em.mix(s, start, prior=1, maxits=1000, showits=TRUE, eps=0.0001)
```

Arguments

s	summary list of an incomplete data matrix produced by the function prelim.mix.
start	optional starting value of the parameter. This is a parameter list in packed storage, such as one returned by this function or by da.mix. If structural zeros appear in the contingency table, start$pi should contain zeros in those positions and ones elsewhere. If no starting value is supplied, em.mix chooses its own appropriate starting value.

`prior` Optional vector or array of hyperparameters for a Dirichlet prior distribution. By default, uses a uniform prior on the cell probabilities (all hyperparameters set to one). EM algorithm finds the posterior mode, which under a uniform prior is the same as a maximum-likelihood estimate. If structural zeros appear in the table, the corresponding hyperparameters should be set to NA.

`maxits` maximum number of iterations performed. The algorithm will stop if the parameter still has not converged after this many iterations.

`showits` if `TRUE`, reports the iterations of EM so the user can monitor the progress of the algorithm.

`eps` optional convergence criterion. The algorithm stops when the maximum relative difference in every parameter from one iteration to the next is less than or equal to this value.

Value

a list representing the maximum-likelihood estimates (or posterior mode) of the normal parameters. This list contains cell probabilities, cell means, and covariances. The parameter can be transformed back to the original scale and put into a more understandable format by the function `getparam.mix`.

Note

If zero cell counts occur in the complete-data table, the maximum likelihood estimate may not be unique, and the algorithm may converge to different stationary values depending on the starting value. Also, if zero cell counts occur in the complete-data table, the ML estimate may lie on the boundary of the parameter space.

References

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, Chapter 9.

See Also

`prelim.mix`, `getparam.mix`, and `ecm.mix`.

Examples

```
data(stlouis)
s <- prelim.mix(stlouis,3)    # do preliminary manipulations
thetahat <- em.mix(s)   # compute ML estimate
getparam.mix(s,thetahat, corr=TRUE) # look at estimated parameters
```

`getparam.mix` *Present Parameters of General Location Model in an Understandable Format*

Description

Present parameters of general location model in an understandable format.

Usage

```
getparam.mix(s, theta, corr=FALSE)
```

Arguments

`s`	summary list of an incomplete normal data matrix created by the function `prelim.mix`.
`theta`	list of parameters such as one produced by the function `em.mix`, `da.mix`, `ecm.mix`, or `dabipf.mix`.
`corr`	if `FALSE`, returns a list containing an array of cell probabilities, a matrix of cell means, and a variance-covariance matrix. If `TRUE`, returns a list containing an array of cell probabilities, a matrix of cell means, a vector of standard deviations, and a correlation matrix.

Value

if `corr=FALSE`, a list containing the components `pi`, `mu` and `sigma`; if `corr=TRUE`, a list containing the components `pi`, `mu`, `sdv`, and `r`.

The components are:

`pi`	array of cell probabilities whose dimensions correspond to the columns of the categorical part of x. The dimension is `c(max(x[,1]),max(x[,2]),...,max(x[,p]))` where p is the number of categorical variables.
`mu`	Matrix of cell means. The dimension is `c(q,D)` where q is the number of continuous variables in x, and D is `length(pi)`. The order of the rows, corresponding to the elements of `pi`, is the same order we would get by vectorizing `pi`, as in `as.vector(pi)`; it is the usual lexicographic order used by S and Fortran, with the subscript corresponding to `x[,1]` varying the fastest, and the subscript corresponding to `x[,p]` varying the slowest.
`sigma`	matrix of variances and covariances corresponding to the continuous variables in `x`.
`sdv`	vector of standard deviations corresponding to the continuous variables in `x`.
`r`	matrix of correlations corresponding to the continuous variables in `x`.

Note

In a restricted general location model, the matrix of means is required to satisfy `t(mu)=A%*%beta` for a given design matrix `A`. To obtain `beta`, perform a multivariate regression of `t(mu)` on `A` — for example, `beta <- lsfit(A, t(mu), intercept=FALSE)$coef`.

References

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, Chapter 9.

See Also

prelim.mix, em.mix, ecm.mix, da.mix, dabipf.mix.

Examples

```
data(stlouis)
s <- prelim.mix(stlouis,3)    # do preliminary manipulations
thetahat <- em.mix(s)   # compute ML estimate
getparam.mix(s, thetahat, corr=TRUE)$r   # look at estimated correlations
```

---

imp.mix *Impute Missing Data Under General Location Model*

Description

This function, when used with da.mix or dabipf.mix, can be used to create proper multiple imputations of missing data under the general location model with or without restrictions.

Usage

```
imp.mix(s, theta, x)
```

Arguments

s	summary list of an incomplete data matrix x created by the function prelim.mix.
theta	value of the parameter under which the missing data are to be randomly imputed. This is a parameter list such as one created by da.mix or dabipf.mix.
x	the original data matrix used to create the summary list s. If this argument is not supplied, then the data matrix returned by this function may disagree slightly with the observed values in x due to rounding errors.

Details

This function is essentially the I-step of data augmentation.

Value

a matrix of the same form as x, but with all missing values filled in with simulated values drawn from their predictive distribution given the observed data and the specified parameter.

Note

The random number generator seed must be set at least once by the function rngseed before this function can be used.

References

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, Chapter 9.

See Also

prelim.mix, da.mix, dabipf.mix, rngseed

Examples

```
data(stlouis)
s <- prelim.mix(stlouis,3)    # do preliminary manipulations
thetahat <- em.mix(s)   # ML estimate for unrestricted model
rngseed(1234567)     # set random number generator seed
newtheta <- da.mix(s,thetahat,steps=100) # data augmentation
ximp <- imp.mix(s, newtheta, stlouis)  # impute under newtheta
```

---

loglik.mix *Loglikelihood for Incomplete Data under the General Location Model*

Description

Calculates the observed-data loglikelihood under the general location model at a user-specified parameter value.

Usage

```
loglik.mix(s, theta)
```

Arguments

s	summary list of an incomplete data matrix x created by the function prelim.mix.
theta	parameter list, such as one produced by ecm.mix or da.mix.

Value

the value of the loglikelihood function at theta.

References

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, Chapter 9.

See Also

prelim.mix, em.mix, ecm.mix.

Examples

```
data(stlouis)
s <- prelim.mix(stlouis,3)       # preliminary manipulations
thetahat <- em.mix(s)      # MLE under unrestricted general location model
loglik.mix(s, thetahat)  # loglikelihood at thetahat
```

`mi.inference` *Multiple Imputation Inference*

Description

Combines estimates and standard errors from m complete-data analyses performed on m imputed datasets to produce a single inference. Uses the technique described by Rubin (1987) for multiple imputation inference for a scalar estimand.

Usage

```
mi.inference(est, std.err, confidence=0.95)
```

Arguments

`est`	a list of m (at least 2) vectors representing estimates (e.g., vectors of estimated regression coefficients) from complete-data analyses performed on m imputed datasets.
`std.err`	a list of m vectors containing standard errors from the complete-data analyses corresponding to the estimates in `est`.
`confidence`	desired coverage of interval estimates.

Value

a list with the following components, each of which is a vector of the same length as the components of `est` and `std.err`:

`est`	the average of the complete-data estimates.
`std.err`	standard errors incorporating both the between and the within-imputation uncertainty (the square root of the "total variance").
`df`	degrees of freedom associated with the t reference distribution used for interval estimates.
`signif`	P-values for the two-tailed hypothesis tests that the estimated quantities are equal to zero.
`lower`	lower limits of the (100*confidence)% interval estimates.
`upper`	upper limits of the (100*confidence)% interval estimates.
`r`	estimated relative increases in variance due to nonresponse.
`fminf`	estimated fractions of missing information.

Method

Uses the method described on pp. 76-77 of Rubin (1987) for combining the complete-data estimates from m imputed datasets for a scalar estimand. Significance levels and interval estimates are approximately valid for each one-dimensional estimand, not for all of them jointly.

References

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys.* Wiley.

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall.

`prelim.mix` *Preliminary Manipulations on Matrix of Incomplete Mixed Data*

Description

This function performs grouping and sorting operations on a mixed dataset with missing values. It creates a list that is needed for input to `em.mix`, `da.mix`, `imp.mix`, etc.

Usage

```
prelim.mix(x, p)
```

Arguments

`x`	data matrix containing missing values. The rows of x correspond to observational units, and the columns to variables. Missing values are denoted by `NA`. The categorical variables must be in the first `p` columns of `x`, and they must be coded with consecutive positive integers starting with 1. For example, a binary variable must be coded as 1,2 rather than 0,1.
`p`	number of categorical variables in x

Value

a list of twenty-nine (!) components that summarize various features of x after the data have been collapsed, centered, scaled, and sorted by missingness patterns. Components that might be of interest to the user include:

`nmis`	a vector of length `ncol(x)` containing the number of missing values for each variable in `x`.
`r`	matrix of response indicators showing the missing data patterns in `x`. Observed values are indicated by 1 and missing values by 0. The row names give the number of observations in each pattern, and the columns correspond to the columns of `x`.

References

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, Chapter 9.

See Also

em.mix, ecm.mix, da.mix, dabipf.mix, imp.mix, getparam.mix

Examples

```
data(stlouis)
s <- prelim.mix(stlouis, 3)  # do preliminary manipulations
s$nmis              # look at nmis
s$r                 # look at missing data patterns
```

---

rngseed *Initialize Random Number Generator Seed*

Description

Initialize random number generator seed for mix package.

Usage

```
rngseed(seed)
```

Arguments

seed	a positive number, preferably a large integer.

Value

NULL.

Note

The random number generator seed must be set at least once by this function before the simulation or imputation functions in this package (da.mix, imp.mix, etc.) can be used.

---

stlouis *St. Louis Risk Research Project*

Description

The St. Louis Risk Research Project was an observational study to assess the affects of parental psychological disorders on child development. In the preliminary study, 69 families with 2 children were studied.

Usage

```
data(stlouis)
```

Format

This is a numeric matrix with 69 rows and 7 columns:

[,1]	G	Parental risk group
[,2]	D1	Symptoms, child 1
[,3]	D2	Symptoms, child 2
[,4]	R1	Reading score, child 1
[,5]	V1	Verbal score, child 1
[,6]	R2	Reading score, child 2
[,7]	V2	Verbal score, child 2

The parental risk group was coded 1, 2 or 3, from low or high, and the child symptoms 1 = low or 2 = high. Missing values occur on all variables except G.

Source

Little, R. J. A. and Schluchter, M. D. (1985), Maximum-likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, **72**, 492–512.

Schafer, J. L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall. pp. 359–367.

Index

∗ **datasets**
stlouis, 13
∗ **models**
da.mix, 2
dabipf.mix, 3
ecm.mix, 5
em.mix, 6
getparam.mix, 8
imp.mix, 9
loglik.mix, 10
mi.inference, 11
prelim.mix, 12
rngseed, 13

da.mix, 2, *6*, *8–10*, *12, 13*
dabipf.mix, *2*, 3, *5*, *8–10*, *13*

ecm.mix, *2*, *4, 5*, 5, *7–10*, *13*
em.mix, *2, 3*, *6*, 6, *8–10*, *12, 13*

getparam.mix, *3–7*, 8, *13*

imp.mix, *2–4*, 9, *12, 13*

loglik.mix, *6*, 10

mi.inference, 11

prelim.mix, *2–10*, 12

rngseed, *3, 4*, *9, 10*, 13

stlouis, 13