Package 'mvmise'

May 27, 2017

Title A General Framework for Multivariate Mixed-Effects Selection Models with Potential Missing Data

Version 1.0

Date 2017-06-01

Author Jiebiao Wang and Lin S. Chen

Maintainer Jiebiao Wang <randel.wang@gmail.com>

Description It offers a general framework for multivariate mixed-effects models with potential missing data in the outcome. The missingness can depend on the outcome (missing not at random), the covariate of interest (missing at random), or both. The multiple outcome variables can have correlated outcome-specific random effects with a factor-analytic structure, or correlated outcome-specific error terms subject to a graphical lasso penalty. Although it is designed for multivariate clustered data with missing values, it can also work for univariate clustered data, and fully observed data.

License GPL

Depends lme4

URL https://github.com/randel/mvmise

BugReports https://github.com/randel/mvmise/issues

RoxygenNote 5.0.1

R topics documented:

`mvmise_b` *Multivariate mixed-effects selection model with correlated outcome-specific random intercepts*

Description

This function fits a multivariate mixed-effects selection model with potential missing values in the outcome and correlated outcome-specific random intercepts.

Usage

```
mvmise_b(Y, X, Z = NULL, id, maxIter = 100, tol = 0.001, verbose = FALSE,
    specific_eff = FALSE, miss_mechanism = "y", sigma_diff = FALSE)
```

Arguments

Y	an outcome matrix, each row is an observation, each column is an outcome variable, with potential missing values (NAs).
X	a covariates matrix, each row is an observation, each column is a covariate. Now covariates are assumed to be common for outcomes.
Z	a design matrix for random effects, each row is an observation, each column is a random effect. If it is NULL (the default), a matrix with each column as an indicator for each outcome is generated.
id	a vector for cluster/grouping index, matching with the rows of Y, X, Z (if specified).
maxIter	maximum number of iterations for the EM algorithm.
tol	tolerance level for the relative change in the observed-data log-likelihood function.
verbose	logical. If TRUE, the iteration history of each step of the EM algorithm will be printed. The default is FALSE.
specific_eff	logical. If TRUE, outcome-specific fixed-effects are estimated for the last covariate in X. The default is FALSE.
miss_mechanism	one of "y" (the default), "x", "yx", and "none", indicating the missingness of outcome k in cluster i depends on the mean of the outcome, the mean of the covariate of interest, both, or none. The missing probability is modelled as exp(phi0 + phi1*mean(y) + phi2*mean(x)). If there is no missing values in Y, it should be set as "none".
sigma_diff	logical. If TRUE, the sample error variance of the first sample is different from that for the rest of samples within each cluster. This is the case for the reference sample in the iTRAQ proteomics data. The default is FALSE.

Details

The multivariate mixed-effects selection model consists of two components, the outcome model and the missing-data model. Here the outcome model is a multivariate mixed-effects model, with correlations among multivariate outcomes modelled via outcome-specific random intercepts with a factor-analytic structure

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\boldsymbol{\tau}b_i + \mathbf{e}_i,$$

where i denotes a cluster, $\boldsymbol{\tau}$ is a $K \times 1$ vector for the outcome-specific variance components corresponding to the random effect b_i (a standard normal random variable), and K is the number of outcomes. The factor-analytic structure is used to facilitate the computation. It assumes that the random effects are derived from a latent variable b_i with a loading vector $\boldsymbol{\tau}$. In this way, only K rather parameters are needed in the estimation for the covariance matrix of random effects. The last fixed effect in $\boldsymbol{\alpha}$ can be outcome-specific, if specific_eff is specified as TRUE.

The missing-data model can be written as

$$\Pr\left(r_{ik} = 1|\mathbf{y}_{ik}\right) = \exp\left(\phi_0 + \phi_1\mathbf{1}_{n_i}^{'}\mathbf{y}_{ik} + \phi_2\mathbf{1}_{n_i}^{'}\mathbf{x}_i\right),$$

where r_{ik} is the missing indicator for the kth outcome in the ith cluster. If missing, the kth outcome in the ith cluster $\mathbf{y}_{ik}$ is missing altogether. The estimation is implemented via an EM algorithm. Parameters in the missing-data models can be specified via the argument miss_mechanism. If miss_mechanism = "y" or "yx", i.e., the missingness depends on the outcome, the missing-data mechanism is missing not at random (MNAR), otherwise it is missing at random (MAR).

It works for fully observed data if miss_mechanism = "none". It also works for univariate outcome with potential missing values, if the outcome Y is a matrix with one column.

Value

A list containing

`beta`	the estimated fixed effects.
`se`	the standard errors for the estimated fixed effects.
`sigma2`	the estimated sample error variance(s). If sigma_diff is TRUE, it returns a vector of two elements, the variances for the first sample and the rest of samples within each cluster.
`tau`	the estimated variance components for the outcome-specific factor-analytic random effects.
`phi`	the estimated parameters for the missing-data mechanism. The missing probability is modelled as exp(phi0 + phi1*mean(y) + phi2*mean(x)). A zero value implies that parameter is ignored via the specification of miss_mechanism.
`loglikelihood`	the observed-data log-likelihood values.
`iter`	the number of iterations for the EM algorithm.

References

Jiebiao Wang, Pei Wang, Donald Hedeker, and Lin S. Chen. A multivariate mixed-effects selection model framework for labelling-based proteomics data with non-ignorable missingness. (In preparation).

Examples

```
data(sim_dat)

fit0 = mvmise_b(Y = sim_dat$Y, X = sim_dat$X, id = sim_dat$id)
```

`mvmise_e` *Multivariate mixed-effects selection model with correlated outcome-specific error terms*

Description

This function fits a multivariate mixed-effects selection model with potential missing values in the outcome and correlated outcome-specific error terms. It can shrink the error precision matrix with a graphical lasso penalty for high-dimensional outcomes.

Usage

```
mvmise_e(Y, X, Zidx = 1, id, maxIter = 100, tol = 0.001, lambda = 0.05, admm = TRUE,
   verbose = FALSE, specific_eff = FALSE, miss_mechanism = "y", sigma_diff = FALSE)
```

Arguments

Y	an outcome matrix, each row is an observation, each column is an outcome variable, with potential missing values (NAs).
X	a covariates matrix, each row is an observation, each column is a covariate. Now covariates are assumed to be common for outcomes.
Zidx	column indexes of matrix X used as the design matrix of random effects. The default is 1, i.e., a random intercept is included if the first column of X is a vector of 1s.
id	a vector for cluster/grouping index, matching with the rows of Y and X.
maxIter	maximum number of iterations for the EM algorithm.
tol	tolerance level for the relative change in the observed-data log-likelihood function.
lambda	tuning parameter for the graphical lasso penalty of the error precision matrix. It can be selected by AIC (an output).
admm	logical. If TRUE (the default), the alternating direction method of multipliers (ADMM) is used to estimate the error precision matrix with a graphical lasso penalty. This works for multivariate outcomes. For an univariate outcome, it should be set as FALSE.
verbose	logical. If TRUE, the iteration history of each step of the EM algorithm will be printed. The default is FALSE.
specific_eff	logical. If TRUE, outcome-specific fixed-effects are estimated for the last covariate in X. The default is FALSE.
miss_mechanism	one of "y" (the default), "x", "yx", and "none", indicating the missingness of outcome k in cluster i depends on the mean of the outcome, the mean of the covariate of interest, both, or none. The missing probability is modelled as exp(phi0 + phi1*mean(y) + phi2*mean(x)). If there is no missing values in Y, it should be set as "none".
sigma_diff	logical. If TRUE, the sample error variance of the first sample is different from that for the rest of samples within each cluster. This is the case for the reference sample in the iTRAQ proteomics data. The default is FALSE.

Details

The multivariate mixed-effects selection model consists of two components, the outcome model and the missing-data model. Here the outcome model is a multivariate mixed-effects model, with correlations among multivariate outcomes modelled via outcome-specific error terms. For the ith cluster, the outcome $\mathbf{Y}_i$ is a matrix of n_i samples (rows) and K outcomes (columns). Let $\mathbf{y}_i = \text{vec}(\mathbf{Y}_i)$. The outcome vector $\mathbf{y}_i$ can be modelled as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i,$$

where the random effects ($\mathbf{b}_i$) follow a normal distribution $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$; and the error term $\mathbf{e}_i = \text{vec}(\mathbf{E}_i) \sim N(\mathbf{0}, \boldsymbol{\Sigma}\otimes\mathbf{S}_i)$. The matrix $\mathbf{S}_i$ is an $n_i \times n_i$ diagonal matrix with diagonal elements corresponding to the error variances of the n_i samples within the ith cluster. The variances for the first and other samples can be different if sigma_diff = TRUE. The matrix $\boldsymbol{\Sigma}$ captures the error (or unexplained) covariances among K outcomes. To facilitate the computation for high-dimensional outcomes, the off-diagonal elements of the inverse of $\boldsymbol{\Sigma}$ can be shrinked by a graphical lasso penalty. If admm = TRUE (the default), the alternating direction method of multipliers (ADMM) is used to estimate $\boldsymbol{\Sigma}$. The last fixed effect in $\boldsymbol{\alpha}$ can be outcome-specific, if specific_eff is specified as TRUE.

The missing-data model can be written as

$$\Pr\left(r_{ik}=1|\mathbf{y}_{ik}\right)=\exp\left(\phi_0+\phi_1\mathbf{1}_{n_i}^{'}\mathbf{y}_{ik}+\phi_2\mathbf{1}_{n_i}^{'}\mathbf{x}_i\right),$$

where r_{ik} is the missing indicator for the kth outcome in the ith cluster. If missing, the kth outcome in the ith cluster $\mathbf{y}_{ik}$ is missing altogether. The estimation is implemented via an EM algorithm. Parameters in the missing-data models can be specified via the argument miss_mechanism. If miss_mechanism = "y" or "yx", i.e., the missingness depends on the outcome, the missing-data mechanism is missing not at random (MNAR), otherwise it is missing at random (MAR).

It works for fully observed data if miss_mechanism = "none". It also works for univariate outcome with potential missing values, if the outcome Y is a matrix with one column.

Value

A list containing

`beta`	the estimated fixed effects.
`se`	the standard errors for the estimated fixed effects.
`Sigma`	the estimated error covariance matrix for the outcomes.
`sigma2`	the estimated sample error variance(s). If sigma_diff is TRUE, it returns a vector of two elements, the variances for the first sample and the rest of samples within each cluster.
`D`	the estimated covariance matrix for the random effects.
`phi`	the estimated parameters for the missing-data mechanism. The missing probability is modelled as exp(phi0 + phi1*mean(y) + phi2*mean(x)). A zero value implies that parameter is ignored via the specification of miss_mechanism.
`loglikelihood`	the observed-data log-likelihood values.
`iter`	the number of iterations for the EM algorithm.
`AIC`	The Akaike information criterion (AIC) calculated for selecting the tuning parameter lambda.

References

Jiebiao Wang, Pei Wang, Donald Hedeker, and Lin S. Chen. A multivariate mixed-effects selection model framework for labelling-based proteomics data with non-ignorable missingness. (In preparation).

Examples

```
data(sim_dat)

fit0 = mvmise_e(Y = sim_dat$Y, X = sim_dat$X, id = sim_dat$id)
```

sim_dat *Example data*

Description

This simulated data list is for demonstration.

Value

A list containing

Y	an outcome matrix, each row is an observation, each column is an outcome variable, with potential missing values (NAs).
X	a covariates matrix, each row is an observation, each column is a covariate. Now covariates are assumed to be common for outcomes.
id	a vector for cluster/grouping index, matching with the rows of Y and X.

Examples

```
data(sim_dat)
```

Index