Package 'mixEMM'

June 8, 2017

Title A Mixed-Effects Model for Analyzing Cluster-Level Non-Ignorable Missing Data

Version 1.0

Date 2017-06-06

Author Lin S. Chen, Pei Wang, and Jiebiao Wang

Maintainer Lin S. Chen <lchen@health.bsd.uchicago.edu>

Description Contains functions for estimating a mixed-effects model for clustered data (or batch-processed data) with cluster-level (or batch-level) missing values in the outcome, i.e., the outcomes of some clusters are either all observed or missing altogether. The model is developed for analyzing incomplete data from labeling-based quantitative proteomics experiments but is not limited to this type of data.
We used an expectation conditional maximization (ECM) algorithm for model estimation. The cluster-level missingness may depend on the average value of the outcome in the cluster (missing not at random).

License GPL

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2017-06-08 15:21:36 UTC

R topics documented:

mixEMM *A mixed-effects model for analyzing cluster-level non-ignorable missing data*

Description

This function fits a mixed-effects model for clustered data with cluster-level missing values in the outcome.

Usage

```
mixEMM(Ym, Xm, Zm, gamma, maxIter = 100, tol = 0.001)
```

Arguments

`Ym`	is an N by p outcome data from N clusters/batches/experiments; p is the number of samples within each cluster. The first sample within each cluster is assumed to be a reference sample with different error variance. Missing values are coded as NAs.
`Xm`	is a covariate array of dimension N by k by p, where k is the number of covariates.
`Zm`	is a design array for random-effects, with a dimension of N by h by p, where h is the number of variables with random effects.
`gamma`	is the parameter for the missing-data mechanism. The missingness of the outcome in cluster i depends on the mean of the outcome. The missing probability is modelled as exp(-gamma0 - gamma*mean(y)). The parameter gamma can be estimated by borrowing information across outcomes and finding the common missing-data patterns in the high-dimensional data. For example, by estimating the relationship the observed average value of $\bar{\mathbf{y}}_i$ and the missing rate, or the parameter can be selected by the log-likelihood profile (see the Reference). If gamma = 0, the missingness is ignorable. The parameter gamma0 does not affect the estimation of the EM algorithm, and is mostly determined by the missing rate. So it is set as 0 in the estimation here.
`maxIter`	the maximum number of iterations in the estimation of the EM algorithm.
`tol`	the tolerance level for the absolute change in the observed-data log-likelihood function.

Details

The model consists of two parts, the outcome model and the missing-data model. The outcome model is a mixed-effects model,

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

where $\mathbf{y}_i$ is the outcome for the i-th cluster, $\mathbf{X}_i$ is the covariate matrix, $\boldsymbol{\alpha}$ is the fixed-effects, $\mathbf{Z}_i$ is the design matrix for the random-effects $\mathbf{b}_i$, and $\mathbf{e}_i$ is the error term.

The non-ignorable batch-level (or cluster-level) abundance-dependent missing-data model (BADMM) can be written as

$$\Pr\left(M_i = 1|\mathbf{y}_i\right) = \exp\left(-\gamma_0 - \gamma\bar{\mathbf{y}}_i\right),$$

where M_i is the missing indicator for the i-th cluster, and $\bar{\mathbf{y}}_i$ is the average of $\mathbf{y}_i$. If $M_i = 1$, the outcome of the i-th cluster $\mathbf{y}_i$ would be missing altogether. The estimation of the mixEMM model is implemented via an ECM algorithm. If $\gamma \neq 0$, i.e., the missingness depends on the outcome, the missing-data mechanism is missing not at random (MNAR), otherwise it is missing completely at random (MCAR) for the current model. The parameter γ can be estimated by borrowing information across outcomes and finding the common missing-data patterns in the high-dimensional data. For example, by estimating the relationship the observed average value of $\bar{\mathbf{y}}_i$ and the missing rate, or the parameter can be selected by the log-likelihood profile (see the Reference).

Value

A list containing

`alpha.hat`	the estimated fixed-effects.
`alpha.se`	the standard errors for the estimated fixed-effects.
`sigma0.hat, sigma2.hat`	the estimated sample error variances. It returns the variances for the first (reference) sample and the other samples within each cluster/batch.
`D`	the estimated covariance matrix for the random-effects.
`RE`	the estimated random-effects.
`loglikelihood`	the observed-data log-likelihood values.

References

Chen, L. S., Wang, J., Wang, X., & Wang, P. (2017). A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments. The Annals of Applied Statistics, 11(1), 114-138. doi: 10.1214/16AOAS994

Examples

```
data(sim_dat)

Z = sim_dat$X[, 1, , drop = FALSE]
fit0 = mixEMM(Ym = sim_dat$Ym, Xm = sim_dat$X, Zm = Z, gamma = 0.14)
```

`sim_dat` *An example data set*

Description

This simulated data list is for demonstration.

Value

A list containing

Ym — A N by p outcome data from N clusters/batches/experiments; and p is the number of samples within each cluster. The first sample within each cluster is a reference sample with a different error variance than other samples. Missing values are coded as NAs. Note the model allows unbalanced data.

X — A covariate array of dimension of N by k by p, where k is the number of covariates.

Examples

```
data(sim_dat)
```

Index