Package 'PEMM'

February 19, 2015

Type Package

Title A Penalized EM algorithm incorporating missing-data mechanism

Version 1.0

Date 2013-11-12

Author Lin Chen <lchen@health.bsd.uchicago.edu> and Pei Wang <pwang@fhcrc.org>

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

Description This package provides functions to perform multivariate Gaussian parameter estimation based on data with abundance-dependent missingness. It implements a penalized Expectation-Maximization (EM) algorithm. The package is tailored for but not limited to proteomics data applications, in which a large proportion of the data are often missing-not-at-random with lower values (or absolute values) more likely to be missing.

License GPL

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2014-01-25 00:37:55

R topics documented:

PEMM *A penalized EM algorithm incorporating missing-data mechanism for multivariate parameter estimation*

Description

For many modern high-throughput technologies, missing values arise at high rates and the missingness probabilities may depend on the values to be measured. In mass spectrometry based proteomics experiments, the smaller the abundance value of a protein is, the harder the protein can be detected. That is, the probability of values being missing depends on the values to be measured.

Motivated by data characteristics in mass spectrometry based proteomics studies, we consider the problem of estimating mean and covariance of multivariate data with ignorable and non-ignorable missingness. The current R package will provide functions to perform a penalized Expectation-Maximization (EM) algorithm in which abundance-dependent missing-data mechanisms if present will be incorporated. The package is tailored for but not limited to proteomics data, in which sample sizes are typically small, and a large proportion of the data are missing-not-at-random. The package can be used to jointly estimate the mean abundance and covariance structure of multiple functionally-related proteins.

Details

Package:	PEMM
Type:	Package
Version:	1.0
Date:	2013-11-12
License:	GPL
LazyLoad:	yes

The package contains a PEMM function, which utilizes a penalized EM algorithm to estimate the mean and covariance of multivariate Gaussian data with ignorable or abundence-dependent missing-data mechanisms. The PEMM function incorporate the abundance-dependent missing-data mechanism in the penalized likelihood, and obtain the maximum penalized likelihood estimates for multvariate mean and covariance via the PEMM algorithm.

Author(s)

Lin S. Chen and Pei Wang

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

References

Lin S. Chen, Ross Prentice and Pei Wang. (2014) A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. Biometrics, in revision.

See Also

PEMM_fun

PEMM_fun *A penalized EM algorithm incorporating missing-data mechanism for multivariate parameter estimation*

Description

The PEMM function utilizes a penalized EM algorithm to estimate the mean and covariance of multivariate Gaussian data with ignorable or abundence-dependent missing-data mechanism. For example, in proteomics data, the smaller the abundance value of a protein is, the more likely the protein cannot be detected in the experiment. The PEMM function incorporates the known or estimated abundance-dependent missing-data mechanism in the penalized likelihood, and obtains the maximum penalized likelihood estimates for multvariate mean and covariance via the PEMM algorithm.

Usage

```
PEMM_fun(X, phi, lambda = NULL, K = NULL, pos = NULL, tol = 0.001,
        maxIter = 100)
```

Arguments

X	a n-by-p matrix, the value of the multivariate incomplete data. Each rown is a sample and each column is one feature (e.g. protein).
phi	the parameter in the missing-data mechanism. It is a non-negative number. We model the abundence-dependent missing-data mechanism as P(missing Xi)=constant*exp(-phi*Xi) if Xi is strictly positive; or P(missing Xi)=constant*exp(-phi*Xi^2) if Xi can be positive and negative. When phi is set to 0, the data will be treated as missing-at-random (MAR) and a penalized EM algorithm will be performed without the consideration of missing-data mechanism. When phi is specified as being greater than zero, the PEMM algorithm will be utilized.
lambda	the tuning parameter in the Inverse-Wishart penalty function. Appropriate value of lambda helps to bound the eigen-values of the covariance matrix away from zero and thus assures the positive-definiteness of the estimated covariance .
K	the second tuning parameter in the Inverse-Wishart penalty function. Appropriate choice of K helps to stabalize the estimation of the covariance matrix. Here we suggest K being a small positive integer (=5), which seems work well when dimensionality and sample size are both moderate (<100).
pos	a logical value. If pos=TRUE, all X are one-sided and the abundance-dependent missing-data mechanism is given by P(missing Xi)=constant*exp(-phi*Xi). If pos=FALSE, X can be two-sided. The abundance-dependent missing-data mechanism is given by P(missing Xi)=constant*exp(-phi*Xi^2).

tol	the tolerance to assess the convergence of the iterative algorithm. We set tol=0.001 as default, i.e., when the change in log-likelihood of the current iteration versus the last iteration is less than 0.1%, the algorithm stops.
maxIter	the maximum number of iterations allowed. We set maxIter=100 as default, i.e., if after 100 iterations the algorithm still do not converge, we force the algorithm to stop and give the users a warning.

Details

The algorithm is motivated by data characteristics in proteomics data with substantial abundance-dependent missing values. The algorithm calculates the maximum penalized likelihood estimates of mean and covariance of multivariate incomplete data, with abundence-dependent missing data mechanism incorporated.

Value

The algorithm will return a list of estimated mean and covariance, and the imputed missing-values from the last iteration.

mu	the estimated mean of the multivariate incomplete data
Sigma	the estimated covariance matrix of the multivariate incomplete data
Xhat	the "imputed" missing-values from the last iteration.

Note

Motivated by characteristics and the abundance-dependent missing-data mechanism in proteomics data, the current function takes the known or estimated abundance-dependent missing-data parameter as input and estimates the mean and covariance of multivariate incomplete data.

When the missing-data mechansim is MAR or MCAR, one could specify phi being zero so that a penalized EM algorithm will be performed without modelling of the missing-data mechanism.

When the missing data are non-ignorable and the missing-data mechanism is unknown, one would first examine if the data follow abundance-dependent pattern and if so, one can estimate the abundance-dependent parameters and perform the analysis based on the PEMM algorithm. Otherwise, if the missing data are not abundence-dependent, one would need alternative approaches.

Author(s)

Lin S. Chen and Pei Wang

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

References

Lin S. Chen, Ross Prentice and Pei Wang. (2014) A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. Biometrics, in revision.

Examples

```
set.seed(111)
library(PEMM)
data(sim_dat)
X.mar=sim_dat
X.mar[sample(1:length(X.mar),round(length(X.mar)*0.2))]<-NA

## If data are MAR or MCAR, by only specifying phi=0,
## a penalized EM algorithm will be performed at default.
PEM.result = PEMM_fun(X.mar, phi=0)

## By specifying phi=0, lambda=0, K=0, an EM algorithm will be performed.
## Although when n is small, EM may not converge.
EM.result = PEMM_fun(X.mar, phi=0, lambda=0, K=0)

## Generate data with non-ignorable missingness -- observations with
## lower absolute values are more likely to be missing
phi=1
prob <- 0.5*exp(-phi*(sim_dat)^2)
X.mnar=sim_dat
X.mnar[which(rbinom(length(X.mnar),1,prob)==1)] <- NA
mean(is.na(X.mnar))  ## proportion of data being missing

## Geting the estimated results
PEMM.result.mnar = PEMM_fun(X.mnar, phi=1)
PEM.result.mnar = PEMM_fun(X.mnar, phi=0)  ## ignoring MNAR mechanism
EM.result.mnar = PEMM_fun(X.mnar, phi=0, lambda=0, K=0)  ## ignoring MNAR

## Compare the mean estimates for data with MNAR from different methods
## complete data
colMeans(sim_dat)

## EM results ignoring MNAR mechanism
EM.result.mnar$mu

## PEMM estimates
PEMM.result.mnar$mu

cor(colMeans(sim_dat),PEMM.result.mnar$mu)
cor(colMeans(sim_dat),EM.result.mnar$mu)
```

---

sim_dat *A simulated multivariate data*

Description

A simulated multivariate data with 30 samples on 15 features.

Details

The data set contains a 30 by 15 matrix, with mean of each feature being -4/15, -3/15 ... to 10/15. Covariance among different features being 0.1 and the variance for each is 1. This is the complete data without any missingness.

Index