Package ‘de.bias.cca’

September 24, 2021

Type Package

Title BIAS CORRECTION OF SCCA CANONICAL DIRECTIONS

Version 0.1.0

Maintainer Nilanjana Laha <nlaha@hsph.harvard.edu>

Description Given sparse estimator of the leading canonical correlation directions (SCCA), this package performs a de-biasing procedure. The bias-correction targets to correct the first order bias of the SCCA estimators under concern. The de-biasing procedure is as discussed in Laha et al. (2021).

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports MASS,
glmnet,
pls,
CCA,
CVXR,
quadprog,
olpsR,
stats,
mvtnorm,
parallel,
dplyr,
Matrix,
methods

RoxygenNote 7.1.0

R topics documented:

`cca.mai` *Leading canonical correlates based on Mai and Zhang (2019).*

Description

The code for the general set up is provided in Qing Mai's website. We include the special rank one case here for completeness because we use this SCCA estimator in an example.

Usage

```
cca.mai(xmat, ymat)
```

Arguments

xmat	The X matrix, a matrix with n rows and p columns.
ymat	The Y matrix, a matrix with n rows and q columns.

Details

The l_1 penalty parameters corresponding to α and β are taken to be $\sqrt{(\log p)/n}$ and $\sqrt{(\log q)/n}$, respectively.

Value

A list. Contains the followings:

sa Estimator of α, leading left canonical direction.

sb Estimator of β, leading right canonical direction.

srho Estimator of leading canonical correlation.

References

Mai, Q. and Zhang, X. (2019). *An iterative penalized least squares approach to sparse canonical correlation analysis*, Biometrics, 75, 734-744.

`give_CCA` *De-biased CCA estimators*

Description

Suppose X and Y are random vectors with variance Σ_x and Σ_y, and cross-covariance matrix Σ_{xy}. The leading canonical correlation coefficient is defined as

$$\rho = \max_{\alpha,\beta} \alpha' \Sigma_{xy} \beta$$

where the maximization is over $\alpha : \alpha'\Sigma_x\alpha = 1$ and $\beta : \beta'\Sigma_y\beta = 1$. The solutions to the above optimization problems are known as the canonical directions. If the estimators of α and β are sparse, then a first order bias is incurred. Laha et al. (2021) introduces a bias correction method, which provides elementwise $\sqrt{n}$-consistent estimators of $\sqrt{\rho}\alpha$ and $\sqrt{\rho}\beta$, provided the preliminary estimators of α and β satisfy some l_1 and l_2 consistency properties. We also provide a $\sqrt{n}$-consistent estimator of ρ^2. See Laha et al. (2021) for more details. The following function provides de-biased estimators of $\sqrt{\rho}\alpha$, $\sqrt{\rho}\beta$, ρ^2, and estimates of their variances.

Usage

```
give_CCA(alpha, beta, X, Y, C, elements, nlC)
```

Arguments

alpha	A vector of length p, estimator of α, the leading canonical vector corresponding to X.
beta	A vector of length q, estimator of β, the leading canonical vector corresponding to Y.
X	A matrix with n rows and p columns, the first dataset.
Y	A matrix with n rows and q columns, the second dataset.
C	An optional constant. Default is 2.
elements	Optional. A vector of integers taking values in the set 1,2,...,p+q. Corresponds to variance estimates of the debiased estimators of α_i and β_j. See details. The default value is the vector 1,2,...,p+q.
nlC	An optional constant. Must be positive. The default is $\sqrt{\log(p+q)/n}$.

Details

`elements` The variance estimates corresponding to α_i (or β_j) are calculated only if i (or p+i) are in `elements`. Otherwise, NA will be returned.

`nlC` The node-wise Lasso parameter λ_j^{nl} in Laha et al. (2021). Should be a constant multiple of $\sqrt{\log(p+q)/n}$.

`C` The constant C in Lemma 1 of Laha et al. (2021). The value of 2 has been used in this paper.

`nlC` Penalty parameter for a l_1 penalty term, required in the nodewise Lasso step, during inversion of the hessian matrix.

Value

A list. Contains the followings:

matx A data frame with de-biased and initial estimators of α in first and second columns, respectively.

maty A data frame with de-biased and initial estimators of β in first and second columns, respectively.

var_calc A (p + q)-length vector with the variance of the parameters specified by the `elements` argument. Non-specified positions return "NA".

rhosq_db The debiased estimate of the squared first canonical correlation.

rhosq_var The variance estimate of the debiased squared first canonical correlation.

References

Laha, N., Huey, N., Coull, B., and Mukherjee, R. (2021). *On Statistical Inference with High Dimensional Sparse CCA*, submitted.

Examples

```
library(mvtnorm)
library(dplyr)
library(CVXR)

#Simulate standard normal data matrix: first generate alpha and beta
p <- 50; q <- 50; al <- c(rep(1, 10), rep(0, 40));
be <- c(rep(0,25), rnorm(25,1))

#Normalize alpha and beta
al <- al/sqrt(sum(al^2)); be <- be/sqrt(sum(be^2))

 #Set n and rho
n <- 300; rho <- 0.5

#Creating the covariance matrix
Sigma_mat <- function(p,q,al,be, rho)
{
 Sx <- diag(rep(1,p), p, p)
 Sy <- diag(rep(1,q), q, q)
 Sxy <- tcrossprod(crossprod(rho*Sx, outer(al, be)), Sy)
 Syx <- t(Sxy)
 rbind(cbind(Sx, Sxy), cbind(Syx, Sy))
}
truesigma <- Sigma_mat(p,q,al,be, rho)

#Finally simulating the data
Z <- mvtnorm::rmvnorm(n, sigma = truesigma)
x <- Z[,1:p]
y <- Z[,(p+1):(p+q)]
elements <- 1:p
nlC <- log(p+q)/n

 # Preliminary estimators: Mai(2019)'s SCCA estimators
temp <- cca.mai(x,y)
ha <- temp[[1]]
hb <- temp[[2]]

#Call give_CCA
give_CCA(ha, hb, x, y)
```

Index