Package: dcorVS (via r-universe)

November 3, 2024

Type Package

Title Variable Selection Algorithms Using the Distance Correlation

Version 1.0

Date 2023-10-17

Author Michail Tsagris [aut, cre]

Maintainer Michail Tsagris <mtsagris@uoc.gr>

Depends R (>= 4.0)

Imports dcov, Rfast, stats

Description The 'FBED' and 'mmpc' variable selection algorithms have been implemented using the distance correlation. The references include: Tsamardinos I., Aliferis C. F. and Statnikov A. (2003). ``Time and sample efficient discovery of Markovblankets and direct causal relations". In Proceedings of the ninth ACM SIGKDD international Conference. <doi:10.1145/956750.956838>. Borboudakis G. and Tsamardinos I. (2019). ``Forward-backward selection with early dropping". Journal of Machine Learning Research, 20(8): 1--39. <doi:10.48550/arXiv.1705.10770>. Huo X. and Szekely G.J. (2016). ``Fast computing for distance covariance". Technometrics, 58(4): 435--447. <doi:10.1080/00401706.2015.1054435>.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2023-10-18 14:10:02 UTC

Contents

dcorVS-package *Variable Selection Algorithms Using the Distance Correlation.*

Description

The 'FBED' and 'mmpc' variable selection algorithms have been implemented using the distance correlation.

Details

Package:	dcorVS
Type:	Package
Version:	1.0
Date:	2023-10-17
License:	GPL-2

Maintainers

Michail Tsagris <mtsagris@uoc.gr>.

Author(s)

Michail Tsagris <mtsagris@uoc.gr>.

References

Szekely G.J., Rizzo M.L. and Bakirov N.K. (2007). Measuring and Testing Independence by Correlation of Distances. Annals of Statistics, 35(6): 2769–2794.

Szekely G.J. and Rizzo M. L. (2014). Partial distance correlation with methods for dissimilarities. Annals of Statistics, 42(6): 2382–2412.

Huo X. and Szekely G.J. (2016). Fast computing for distance covariance. Technometrics, 58(4): 435–447.

Tsamardinos I., Aliferis C. F. and Statnikov A. (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. In Proceedings of the ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (pp. 673–678). ACM.

Brown L. E., Tsamardinos I. and Aliferis C. F. (2004). A novel algorithm for scalable and accurate Bayesian network learning. Medinfo, 711–715.

Borboudakis G. and Tsamardinos I. (2019). Forward-backward selection with early dropping. Journal of Machine Learning Research, 20(8): 1–39.

`Backward selection algorithms using the distance correlation`
Backward selection algorithms using the distance correlation

Description

Backward selection algorithms using the distance correlation.

Usage

```
dcor.bsmmpc(y, x, max_k = 3, alpha = 0.05, B = 999)
dcor.bs(y, x, alpha = 0.05)
```

Arguments

`y`	A numerical vector with the response variable.
`x`	A numerical matrix with the predictor variables.
`max_k`	The maximum conditioning set to use in the conditional indepedence test (see Details). Integer, default value is 3.
`alpha`	The significance level for assessing the p-values. Default value is 0.05.
`B`	The number of permutations to execute to compute the p-value of the distance correlation.

Details

The max_k option in the mmpc algorithm: the maximum size of the conditioning set to use in the conditioning independence test. Larger values provide more accurate results, at the cost of higher computational times. When the sample size is small (e.g., < 50 observations) the max_k parameter should be 3 for example, otherwise the conditional independence test may not be able to provide reliable results.

The dcor.bs() performs the classical backward selection.

Value

A list including:

`runtime`	The duration of the algorithm.
`res`	A matrix with all variables and their corresponding (logarithm) of the p-values of the updated associations. For the mmpc algorithm, the final p-value is the maximum p-value among the two p-values in the end.

Author(s)

Michail Tsagris.

R implementation and documentation: Michail Tsagris <mtsagris@uoc.gr>.

References

Szekely G.J., Rizzo M.L. and Bakirov N.K. (2007). Measuring and Testing Independence by Correlation of Distances. Annals of Statistics, 35(6): 2769–2794.

Szekely G.J. and Rizzo M. L. (2014). Partial distance correlation with methods for dissimilarities. Annals of Statistics, 42(6): 2382–2412.

Huo X. and Szekely G.J. (2016). Fast computing for distance covariance. Technometrics, 58(4): 435–447.

Tsamardinos I., Aliferis C. F. and Statnikov A. (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. In Proceedings of the ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (pp. 673–678). ACM.

Brown L. E., Tsamardinos I. and Aliferis C. F. (2004). A novel algorithm for scalable and accurate Bayesian network learning. Medinfo, 711–715.

Borboudakis G. and Tsamardinos I. (2019). Forward-backward selection with early dropping. Journal of Machine Learning Research, 20(8): 1–39.

See Also

dcor.fbed

Examples

```
y <- rnorm(100)
x <- matrix( rnorm(100 * 10), ncol = 10 )
a <- dcor.bs(y, x)
```

`MMPC and the FBED variable selection algorithms using the distance correlation`

MMPC and the FBED variable selection algorithms using the distance correlation

---

Description

MMPC and the FBED variable selection algorithms using the distance correlation.

Usage

```
dcor.mmpc(y, x, max_k = 3, alpha = 0.05, B = 999, backward = TRUE)
dcor.fbed(y, x, alpha = 0.05, K = 0, backward = TRUE)
```

Arguments

y	A numerical vector with the response variable.
x	A numerical matrix with the predictor variables.
max_k	The maximum conditioning set to use in the conditional indepedence test (see Details). Integer, default value is 3.
alpha	The significance level for assessing the p-values. Default value is 0.05.
B	The number of permutations to execute to compute the p-value of the distance correlation.
K	How many times should the process of the Forward Early Dropping be repeated? The default value is 0.
backward	Should the backward selection take place? The default value is set to TRUE.

Details

The FBED algorithm is a variation of the usual forward selection. At every step, the most significant variable enters the selected variables set. In addition, only the significant variables stay and are further examined. The non signifcant ones are dropped. This goes until no variable can enter the set. The user has the option to re-do this step 1 or more times (the argument K). In the end, a backward selection is performed to remove falsely selected variables. Note that you may have specified, for example, K=10, but the maximum value FBED used can be 4 for example.

The max_k option in the mmpc algorithm: the maximum size of the conditioning set to use in the conditioning independence test. Larger values provide more accurate results, at the cost of higher computational times. When the sample size is small (e.g., < 50 observations) the max_k parameter should be 3 for example, otherwise the conditional independence test may not be able to provide reliable results.

Both the MMPC (Tsamardinos, Aliferis and Statnikov, 2003) and FBED algortihms (Borboudakis and Tsamardinos, 2019) are performed though by utilizing the distance correlation (Szekely et al., 2007, Szekely and Rizzo 2014, Huo and Szekely, 2016).

Value

A list including:

runtime	The duration of the algorithm.
res	A matrix with the selected variables and their corresponding (logarithm) of the p-values of the updated associations. For the mmpc algorithm, the final p-value is the maximum p-value among the two p-values in the end.

Author(s)

Michail Tsagris.

R implementation and documentation: Michail Tsagris <mtsagris@uoc.gr>.

References

Szekely G.J., Rizzo M.L. and Bakirov N.K. (2007). Measuring and Testing Independence by Correlation of Distances. Annals of Statistics, 35(6): 2769–2794.

Szekely G.J. and Rizzo M. L. (2014). Partial distance correlation with methods for dissimilarities. Annals of Statistics, 42(6): 2382–2412.

Huo X. and Szekely G.J. (2016). Fast computing for distance covariance. Technometrics, 58(4): 435–447.

Tsamardinos I., Aliferis C. F. and Statnikov A. (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. In Proceedings of the ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (pp. 673–678). ACM.

Brown L. E., Tsamardinos I. and Aliferis C. F. (2004). A novel algorithm for scalable and accurate Bayesian network learning. Medinfo, 711–715.

Borboudakis G. and Tsamardinos I. (2019). Forward-backward selection with early dropping. Journal of Machine Learning Research, 20(8): 1–39.

See Also

dcor.bs

Examples

```
y <- rnorm(100)
x <- matrix( rnorm(100 * 50), ncol = 50 )
a <- dcor.fbed(y, x, backward = FALSE)
```

Index