Package ‘factree’

December 10, 2025

Type Package

Title Factor-Augmented Clustering Tree

Date 2025-11-10

Version 0.1.0

Description Implements the Factor-Augmented Clustering Tree (FACT) algorithm for clustering time series data. The method constructs a classification tree where splits are determined by covariates, and the splitting criterion is based on a group factor model representation of the time series within each node. Both threshold-based and permutation-based tests are supported for splitting decisions, with an option for parallel computation. For methodological details, see Hu, Li, Luo, and Wang (2025, in preparation), Factor-Augmented Clustering Tree for Time Series.

License MIT + file LICENSE

Encoding UTF-8

Depends R (>= 3.5.0)

RoxygenNote 7.3.3

Imports irlba, foreach, doParallel, parallel, doRNG, mvtnorm

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

NeedsCompilation no

Author Jiaqi Hu [cre, aut],
Ting Li [ctb] (Assisted with methodology),
Zidan Luo [ctb] (Assisted with methodology),
Xueqin Wang [ctb] (Assisted with methodology)

Maintainer Jiaqi Hu <hujiaqi@mail.ustc.edu.cn>

Repository CRAN

Date/Publication 2025-12-10 21:30:12 UTC

Contents

COR *Correlation-Based Clustering Tree*

Description

Builds a binary tree for clustering time series data based on covariates. The splitting criterion minimizes the average absolute Pearson correlation between time series across child nodes.

Usage

```
COR(X, Y, control = list())
```

Arguments

`X` A numeric matrix of covariates with dimension $N \times p$, where N is the number of time series and p is the number of features. Each row corresponds to the covariates for one time series.

`Y` A numeric matrix of time series data with dimension $T \times N$, where T is the length of each series. Each column represents one time series.

`control` A list of control parameters for tree construction:

- `minsplit` Minimum number of observations required to attempt a split. Default: `90`.
- `minbucket` Minimum number of observations in any terminal node. Default: `30`.
- `alpha` Significance level for the permutation test. Default: `0.01`.
- `R` Number of permutations for the hypothesis test. Default: `199`.
- `parallel` Logical; if `TRUE`, enables parallel computation for permutation tests. Default: `FALSE`.
- `n_cores` Number of cores for parallel processing. If `NULL` (default), uses `detectCores() - 1`.

Details

The algorithm recursively partitions the data by finding splits that minimize the average absolute correlation between time series in different child nodes. Statistical significance of each split is assessed via a permutation test.

At each node, the optimal split is found by exhaustively searching over all covariates and candidate split points. The permutation test shuffles the time series labels to generate a null distribution for the test statistic.

Value

An object of class "FACT" containing:

frame A data frame describing the tree structure, with one row per node containing split variable, split value, test statistic, and p-value. A smaller test statistic suggests more heterogeneity between child nodes.

membership An integer vector of length N indicating the terminal node assignment for each observation.

control The control parameters used.

terms Metadata including covariate names and data dimensions.

See Also

FACT for factor model-based clustering, gendata for generating synthetic data, print.FACT and plot.FACT for visualization.

Examples

```
# Generate synthetic data
data <- gendata(seed = 42, T = 100, N = c(50, 50, 50, 50))

# Build correlation-based tree
result <- COR(data$X, data$Y, control = list(R = 99, alpha = 0.05))

# Examine results
print(result)
plot(result)
table(result$membership, data$group)
```

FACT *Factor-Augmented Clustering Tree*

Description

Builds a binary tree for clustering time series data based on covariates, using a group factor model framework. The splitting criterion evaluates whether child nodes exhibit distinct factor structures.

Usage

```
FACT(
  X,
  Y,
  r_a = 8,
  r_b = 4,
  method = c("threshold", "permutation"),
  control = list()
)
```

Arguments

X
: A numeric matrix of covariates with dimension $N \times p$, where N is the number of time series and p is the number of features. Each row corresponds to the covariates for one time series.

Y
: A numeric matrix of time series data with dimension $T \times N$, where T is the length of each series. Each column represents one time series.

r_a
: A positive integer specifying the number of singular vectors to extract from each child node for constructing the projection matrices, default is 8.

r_b
: A positive integer specifying the number of leading singular values to sum for the split statistic. Must satisfy `r_b` <= `r_a`, default is 2.

method
: Character string specifying the splitting decision rule:

  `"threshold"` Uses a data-adaptive threshold based on signal-to-noise ratio estimation. Faster but may be less accurate. Suitable for large datasets.

  `"permutation"` Uses a permutation test for hypothesis testing. More rigorous but computationally intensive.

control
: A list of control parameters for tree construction:

  `minsplit` Minimum number of observations required to attempt a split. Default: `90`.

  `minbucket` Minimum number of observations in any terminal node. Default: `30`.

  `alpha` Significance level for the permutation test (used only when `method = "permutation"`). Default: `0.01`.

  `R` Number of permutations for the hypothesis test (used only when `method = "permutation"`). Default: `199`.

  `sep` Controls the density of candidate split points. If `"auto"` (default), subsamples candidates when $n > 800$. If numeric, evaluates every `sep` candidate point.

  `parallel` Logical; if `TRUE`, enables parallel computation. Default: `FALSE`.

  `n_cores` Number of cores for parallel processing. If `NULL` (default), uses `detectCores() - 1`.

Details

The FACT algorithm clusters time series by recursively partitioning them based on their underlying factor structures. At each node, the method:

1. Searches for the optimal split across all covariates and candidate points.
2. Computes a test statistic based on the overlap of factor spaces between the two child nodes.
3. Decides whether to split using either a threshold rule or permutation test.

Value

An object of class `"FACT"` containing:

frame A data frame describing the tree structure, with one row per node. Includes split variable, split value, test statistic, and p-value (if applicable). A smaller test statistic indicates stronger evidence of heterogeneous factor structures between child nodes.

membership An integer vector of length N indicating the terminal node assignment for each observation.

control The control parameters used.

terms Metadata including covariate names, data dimensions, and the values of r_a and r_b.

method The splitting method used.

References

Hu, J., Li, T., Luo, Z., & Wang, X. Factor-Augmented Clustering Tree for Time Series.

See Also

COR for correlation-based clustering, gendata for generating synthetic data, print.FACT and plot.FACT for visualization.

Examples

```
data <- gendata(seed = 123, T = 200, N = c(50, 50, 50, 50))
tree1 <- FACT(data$X, data$Y, r_a = 8, r_b = 4, method = "threshold")
print(tree1)
```

gendata *Generate Synthetic Group Factor Model Data*

Description

Generates synthetic time series data with a multi-group factor structure, along with associated covariates. Useful for Monte Carlo simulation. the FACT and COR algorithms.

Usage

```
gendata(
  seed = 1,
  T = 100,
  N = c(100, 100, 100, 100),
  r0 = 2,
  r = c(2, 2, 2, 2),
  M = 4,
  sigma = 1,
  p = 10,
  mu = 3,
  type_F = "Independent",
  type_X = "Uniform",
  type_noise = "Gaussian"
)
```

Arguments

`seed` Integer. Random seed for reproducibility. Default: `1`.

`T` Integer. Number of time periods (rows in `Y`). Default: `100`.

`N` Integer vector of length `M`. Number of time series per group, such that `sum(N)` equals the total number of series. Default: `c(100, 100, 100, 100)`.

`r0` Integer. Number of global factors shared across all groups. Default: `2`.

`r` Integer vector of length `M`. Number of local (group-specific) factors for each group. Default: `c(2, 2, 2, 2)`.

`M` Integer. Number of groups. Default: `4`.

`sigma` Numeric. Standard deviation of the idiosyncratic noise. Default: `1`.

`p` Integer. Number of covariates (columns in `X`). Default: `10`.

`mu` Numeric. Controls separation between group covariate distributions when `type_X = "Gaussian"`. Larger values yield better-separated groups. Default: `3`.

`type_F` Character. Correlation structure for local factors:

- `"Independent"` Local factors are independent across groups (default). Each follows an AR(1) process.
- `"Correlated"` Local factors share a common correlation structure across groups.

`type_X` Character. Distribution for generating covariates:

- `"Uniform"` Groups differ by support on the real line (default).
- `"Gaussian"` Groups differ by mean shifts.

`type_noise` Character. Distribution for idiosyncratic errors:

- `"Gaussian"` Normal errors (default).
- `"t3"` Heavy-tailed errors from a t-distribution with 3 degrees of freedom, scaled to have the same variance.

Details

The data generating process follows a group factor model:

$$Y_m = G\Lambda_m' + F_m\Gamma_m' + E_m, \quad m = 1, \ldots, M$$

where:

- G: $T \times r_0$ matrix of global factors (shared across groups)
- Λ_m: $N_m \times r_0$ global factor loadings for group m
- F_m: $T \times r_m$ matrix of local factors for group m
- Γ_m: $N_m \times r_m$ local factor loadings for group m
- E_m: $T \times N_m$ idiosyncratic error matrix

Both global and local factors follow AR(1) processes with coefficient 0.5. Factor loadings are drawn from standard normal distributions.

Value

A list containing:

Y A $T \times N$ numeric matrix of time series, where $N = \sum N_m$.

X A $N \times p$ numeric matrix of covariates.

G The $T \times r_0$ matrix of true global factors.

r0 Number of global factors.

r Vector of local factor counts per group.

group Integer vector of length N indicating true group membership (values 1 through M).

Note

The default covariate generation (type_X = "Uniform" or "Gaussian") assumes M = 4 groups with a specific hierarchical structure: groups 1-2 vs 3-4 are separated by the first covariate, and within each pair, groups are separated by additional covariates.

See Also

FACT for building factor-augmented clustering trees, COR for correlation-based clustering.

Examples

```
data <- gendata(seed = 123, T = 200, N = c(100, 50, 50, 200), r0 = 1, r = c(2, 2, 2, 3), M = 4)
Y <- data$Y
X <- data$X
```

---

plot.FACT *Plot a FACT Tree Object*

Description

Generates a visual plot of the FACT tree structure.

Usage

```
## S3 method for class 'FACT'
plot(x, ...)
```

Arguments

x An object of class ‘FACT‘, typically the result of a call to ‘FACT()‘.

... Additional arguments (not used).

Value

No return value, called for side effects (plotting).

print.FACT *Print a FACT Tree Object*

Description

Provides a concise textual representation of the FACT tree structure.

Usage

```
## S3 method for class 'FACT'
print(x, ...)
```

Arguments

`x`	An object of class ‘FACT‘, typically the result of a call to ‘FACT()‘.
`...`	Additional arguments (not used).

Value

Invisibly returns the original ‘FACT‘ object.

Index