

# Package: hetcorFS (via r-universe)

May 23, 2026

**Type** Package

**Title** Unsupervised Feature Selection using the Heterogeneous Correlation Matrix

**Version** 1.0.1

**Description** Unsupervised multivariate filter feature selection using the UFS-rHCM or UFS-cHCM algorithms based on the heterogeneous correlation matrix (HCM). The HCM consists of Pearson's correlations between numerical features, polyserial correlations between numerical and ordinal features, and polychoric correlations between ordinal features. Tortora C., Madhvani S., Punzo A. (2025). ``Designing unsupervised mixed-type feature selection techniques using the heterogeneous correlation matrix." International Statistical Review <doi:10.1111/insr.70016>. This work was supported by the National Science foundation NSF Grant N 2209974 (Tortora) and by the Italian Ministry of University and Research (MUR) under the PRIN 2022 grant number 2022XRHT8R (CUP: E53D23005950006), as part of 'The SMILE Project: Statistical Modelling and Inference to Live the Environment', funded by the European Union – Next Generation EU (Punzo).

**License** GPL-2

**Imports** polycor, dplyr, cluster, graphics,psych

**Depends** R (>= 3.5.0)

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.3.1

**NeedsCompilation** no

**Author** Cristina Tortora [aut, cre, fnd], Antonio Punzo [aut], Shaam Madhvani [aut]

**Maintainer** Cristina Tortora <grikris1@gmail.com>

**Repository** https://cran.r-universe.dev

**Date/Publication** 2025-11-24 13:30:09 UTC

**RemoteUrl** <https://github.com/cran/hetcorFS>

**RemoteRef** HEAD

**RemoteSha** be3f598431f6e86edf8e05f70c76e27ab22df2c5

## Contents

ESI . . . . .	2
FS_barplot . . . . .	3
HCPM . . . . .	4
JaccardRate . . . . .	5
RedRate . . . . .	6
UFS . . . . .	6
<b>Index</b>	<b>9</b>

---

ESI *Employee Satisfaction Index (ESI) Data Set*

---

### Description

The Employee Satisfaction Index (ESI) data set, from Kaggle (Harris, 2023), is a fictional data set that measures employee satisfaction

### Usage

`data(ESI)`

### Format

A data frame with 500 rows and 10 features.

**emp\_id** label.

**age** continuous from 23 to 45.

**Dept** categorical.

**location** binary.

**education** binary.

**recruitment\_type** categorical.

**job\_level** ordinal from 1 to 5.

**rating** ordinal from 1 to 5.

**onsite** binary.

**awards** number of awards 0-9.

**certifications** binary.

**salary** continuous from 24.1 to 86.8.

**satisfied** binary.

**Source**

Harris, M. (2023). Employee Satisfaction Index Dataset. Evanston, Illinois: Kaggle. Version 1

---

 FS\_barplot

*Feature importance bar plot*


---

**Description**

Displays retained features for different values of alpha in a bar plot.

**Usage**

```
FS_barplot(
  data = NULL,
  grid.alpha = seq(0.01, 0.99, by = 0.01),
  missing = FALSE,
  pv_adj = "none",
  smooth.tol = 10^-12,
  method = "c"
)
```

**Arguments**

data	A data frame. Values of type 'numeric' or 'integer' are treated as numerical.
grid.alpha	A vector of alpha values to be plotted, default = seq(0.01,0.99,by=0.01).
missing	Pairwise complete by default, set to TRUE for complete deletion.
pv_adj	Correction method for p-value, "none" by default. For options see p.adjust.
smooth.tol	Minimum acceptable eigenvalue for the smoothing, default 10^-12.
method	Algorithm used. c (cell-wise) by default, r (row-wise) as the alternative.

**Value**

Displays a bar plot depicting which features are selected at each value of alpha (multiplied by 100) and a list with elements:

survivors	Vector depicting how many alphas a variable is selected for
data_names	Vector depicting the corresponding names of the features

**References**

Tortora C., Madhvani S., Punzo A. (2025). Designing unsupervised mixed-type feature selection techniques using the heterogeneous correlation matrix. *International Statistical Review*. <https://doi.org/10.1111/insr.70016>

## Examples

```
data(ESI)
data=ESI[,-c(1,3,4,6,9)]##removing categorical features
FS_barplot(data, pv_adj='BH') #using BH adjustment for the p-values
```

---

HCPM

*Heterogeneous correlation and p-value matrices*

---

## Description

Extends the traditional correlation matrix (between numerical data) to also include binary and ordinal categorical data and computes the p-values for the tests of uncorrelation.

## Usage

```
HCPM(data = NULL)
```

## Arguments

`data` A data frame. Values of type 'numeric' or 'integer' are treated as numerical.

## Value

A list with with elements:

`cor_mat` An  $p$  by  $p$  heterogeneous correlation matrix

`p_value` An  $p$  by  $p$  heterogeneous p-values matrix

## References

Tortora C., Madhvani S., Punzo A. (2025). Designing unsupervised mixed-type feature selection techniques using the heterogeneous correlation matrix. *International Statistical Review*. <https://doi.org/10.1111/insr.70016>

## Examples

```
data(ESI)
data=ESI[,-c(1,3,4,6,9)]##removing categorical features
HCPM(data)
```

---

JaccardRate	<i>Jaccard Rate</i>
-------------	---------------------

---

**Description**

Computes the Jaccard index using Gower's dissimilarity.

**Usage**

```
JaccardRate(  
  data,  
  data_red,  
  k=6  
)
```

**Arguments**

data	A data frame. Values of type 'numeric' or 'integer' are treated as numerical.
data_red	A data frame. A subset of data with the selected features.
k	number of neighbors

**Value**

Jaccard Index    numeric

**References**

Zhao, Z., L. Wang, and H. Liu (2010). Efficient spectral feature selection with minimum redundancy. In Proceedings of the AAAI conference on artificial intelligence, Volume 24, pp. 673–678.

**Examples**

```
data(ESI)  
data=ESI[,-c(1,3,4,6,9)] ##removing categorical features  
out=UFS(data,alpha=0.01,method='c',pv_adj='BH')  
JR=JaccardRate(data,out$selected.features)  
JR #visualize the index
```

---

RedRate	<i>Redundancy Rate</i>
---------	------------------------

---

**Description**

Computes the Redundancy Rate using heterogeneous correlation matrix.

**Usage**

```
RedRate(
  data_red
)
```

**Arguments**

data\_red      A data frame. A subset of data with the selected features.

**Value**

Redundancy Rate  
numeric

**References**

Zhao, Z., L. Wang, and H. Liu (2010). Efficient spectral feature selection with minimum redundancy. In Proceedings of the AAAI conference on artificial intelligence, Volume 24, pp. 673–678.

**Examples**

```
data(ESI)
data=ESI[,-c(1,3,4,6,9)] ##removing categorical features
out=UFS(data,alpha=0.01,method='c',pv_adj='BH')
RR=RedRate(out$selected.features)
RR #visualize the index
```

---

UFS	<i>Unsupervised Feature Selection</i>
-----	---------------------------------------

---

**Description**

Performs unsupervised feature selection for mixed type data. Both algorithms are based on the heterogeneous correlation matrix.

**Usage**

```

UFS(
  data = NULL,
  alpha = 0.05,
  missing = FALSE,
  pv_adj = "none",
  smooth.tol = 10^-12,
  method = "c"
)

```

**Arguments**

<code>data</code>	A data frame. Values of type 'numeric' or 'integer' are treated as numerical, factors as ordinal categorical.
<code>alpha</code>	Significance level to be used for testing, default = 0.05.
<code>missing</code>	Pairwise complete by default, set to TRUE for complete deletion.
<code>pv_adj</code>	Correction method for p-value, "none" by default. For options see <code>p.adjust</code> .
<code>smooth.tol</code>	Minimum acceptable eigenvalue for the smoothing, default = $10^{-12}$ .
<code>method</code>	Algorithm used. c (cell-wise) by default, r (row-wise) as the alternative.

**Value**

An list of elements:

<code>rearranged.data.set</code>	Original data frame with with numerical features first
<code>selected.features</code>	A data frame of the selected features
<code>feature.indices</code>	The indices of the selected features from the original data frame
<code>original.corr.matrix</code>	The $p$ by $p$ extended correlation matrix of all the inputted features
<code>corr.matrix</code>	The $d$ by $d$ extended correlation matrix of the selected features
<code>original.p.value.matrix</code>	The $p$ by $p$ p-values matrix of all the inputted features
<code>p.value.matrix</code>	The $d$ by $d$ p-values matrix of the selected features

**References**

Tortora C., Madhvani S., Punzo A. (2025). Designing unsupervised mixed-type feature selection techniques using the heterogeneous correlation matrix. *International Statistical Review*. <https://doi.org/10.1111/insr.70016>

**Examples**

```
data(ESI)#Loading the data
data = ESI[,-c(1,3,4,6,9)]##removing categorical features
res = UFS(data)

### visualize selected features
colnames(res$selected.features)
```

# Index

\* **datasets**

ESI, [2](#)

ESI, [2](#)

FS\_barplot, [3](#)

HCPM, [4](#)

JaccardRate, [5](#)

RedRate, [6](#)

UFS, [6](#)