SCOUTer

Table of Contents

barwithucl 1
custombar 2
distplot 2
distplotsimple 3
dscplot 3
ht2info 4
obscontribpanel 4
pcamb_classic 5
pcame 6
scoreplot 6
scoreplotsimple 7
scout 7
scoutgrid 9
scoutsimple 10
scoutsteps 10
speinfo 11
xshift 12

SCOUTer is a set of functions that open the door towards a new way of simulating outliers. Using Principal Component Analysis (PCA) as a base model, SCOUTer enables the user to control exactly how the simulated outliers will be.

This is possible by controlling the pair of properties that define outliers: the Squared Prediction Error (*SPE*) and the Hotelling's T^2. By setting the desired target values to these statistics, the user can obtain **all** types of outliers.

May you want to simulate observations which do not follow the pattern described by your model at all? Then simulate observations with high *SPE* target value. Or, if you want extreme observations which still follow the model pattern, then set a high T^2 target in the simulation. Or do both!

Control your simulations with SCOUTer in an easy and interactive way!

In this document, there is a **full documentation** of the functions implemented in the SCOUTer Matlab version

barwithucl

Description

Single bar plot with Upper Control Limis. Customized title and labels. Y-Axis limits are fixed according to the range of the values in x.

Inputs

- `x`: vector with the values of the statistic.
- `iobs`: index of the observation whose value will be displayed.
- `ucl`: Upper Control Limit of the statistic.

Name - Value pair Input Arguments:

- `'plotname'`: string with the title of the plot. Default set to `" "`.
- `'xlabelname'`: string with the x-axis label. Default set to `" "`.
- `'ylabelname'`: string with the y-axis label. Default set to `" "`.

Outputs

`axobj`: parent axes of the bar plot

custombar

Description

Bar plot with with customized title and labels. Y-Axis limits are fixed according to the range of the values in the `X` input argument.

Inputs

- `X`: matrix with the values of the statistic.
- `iobs`: index of the observation (row) whose values will be displayed.

Name - Value pair Input Arguments:

- `'plotname'`: string with the title of the plot. Default set to `" "`.
- `'xlabelname'`: string with the x-axis label. Default set to `" "`.
- `'ylabelname'`: string with the y-axis label. Default set to `" "`.

Outputs

`axobj`: parent axes of the bar plot.

distplot

Description

Returns the distance plot according to the input arguments.

Inputs

- `X`: vector with the values of the statistic.
- `pcamodel`: struct with the information of the PCA model, at least with fields m (mean vector), s (standard deviation vector), P (loading matrix), prepro (string with preprocessing), lambda (score variances vector), ncomp (number of PCs).

Name - Value pair Input Arguments:

- `'clicktoggle'`: string to control interactive plot options with values `"on"` or `"off"` . Default value set to `"on"` .
- `'obstag'`: column vector of integers indicating the group of each observation. Default value set to `zeros(size(X,1),1)` .

- `'steps_spe'`: column vector of integers indicating the *SPE* step of each observation. Default value set to `zeros(size(X,1),1)`.
- `'steps_spe'`: column vector of integers indicating the T^2 step of each observation. Default value set to `zeros(size(X,1),1)`.

Outputs

(none)

distplotsimple

Description

Returns the distance plot. Does not have interactive options.

Inputs

- `T2`: double input with the Hotelling's T^2 statistic vector.
- `SPE`: double input with the Squared Prediction Error (*SPE*) vector.
- `limt2`: double input with the Upper Control Limit of the T^2.
- `limspe`: double input with the Upper Control Limit of the *SPE*.
- `alpha` (optional): double input with the Type I error assumed for the UCLs. Default set to 0.05.
- `obstag` (optional): double vector indicating the tag (0 for reference and 1 for new) of the observations. Default set to `zeros(size(T2))`.

Outputs

(none)

dscplot

Description

Returns the distance plot (left) and score plot (right) according to the input arguments.

Inputs

- `X`: vector with the values of the statistic.
- `pcamodel`: struct with the information of the PCA model, at least with fields m (mean vector), s (standard deviation vector), P (loading matrix), prepro (string with preprocessing), lambda (score variances vector), ncomp (number of PCs).

Name - Value pair Input Arguments:

- `'clicktoggle'`: string to control interactive plot options with values `"on"` or `"off"`. Default set to `"on"`.
- `'pcx'`: integer with the PC on the horizontal axis. Default set to 1.

- `'pcy'`: integer with the PC on the vertical axis. Default set to 2.
- `'obstag'`: column vector of integers indicating the group of each observation. Default value set to `zeros(size(X,1),1)` .
- `'steps_spe'`: column vector of integers indicating the *SPE* step of each observation. Default value set to `zeros(size(X,1),1)` .
- `'steps_spe'`: column vector of integers indicating the T^2 step of each observation. Default value set to `zeros(size(X,1),1)` .

Outputs

(none)

ht2info

Description

Information about Hotelling's T^2_A (T^2) for an observation, i.e.: information about the Mahalanobis distance of the observation on the PCA model.

Inputs

- `T2`: double vector with values of the T^2 statistic.
- `T2mat`: double matrix with the contributions of each variable (columns) for each observation (rows) to the T^2.
- `limt2`: double with the value of the T^2 Upper Control Limit (with confidence level ($1-\alpha$) $*$ 100 %).
- `iobs`: integer with the index of the observation of interest.

Outputs

- `barobs` : axis of the bar plot with the T^2 value.
- `barcont` : axis of the bar plot with the contributions to the T^2.

obscontribpanel

Description

Information about Squared Prediction Error and Hotelling's T^2_A (T^2) for an observation, i.e.: information about the Mahalanobis distance of the observation on the PCA model.

Inputs

- `pcaout`: struct containing the following fields:

`SPE` : double vector with values of the *SPE* statistic.

`E` : double matrix with the contributions of each variable (columns) for each observation (rows) to the *SPE*.

`T2` : double vector with values of the T^2 statistic.

`T2cont` : double matrix with the contributions of each variable (columns) for each observation (rows) to the T^2.

- `limspe`: double with the value of the SPE Upper Control Limit (with confidence level ($1-\alpha) * 100$ %).
- `limt2`: double with the value of the T^2 Upper Control Limit (with confidence level ($1-\alpha) * 100$ %).
- `iobs`: integer with the index of the observation of interest.

Outputs

(none) figure with 2 x 2 subplots layout

pcamb_classic

Description

Performs PCA Model Building using the data in X using the SVD approach.

Inputs

- `X`: double matrix of dimensions *NxK* with observations used for the PCA-MB.
- `ncomp`: integer indicating the number of Principal Components of the model.
- `alpha` (optional): value of the Type I risk assumed for the Upper Control Limits (UCL) calculation. Default value set to $\alpha = 0.05$.
- `prepro` (optional): string indicating preprocessing applied to X, its possible values are `'cent'` , `'autosc'` or `'none'` . Default value is set to `'none'`.

Outputs

`pcamodel` : struct returning the parameters of the PCA model fit with data in X.

- `m`: mean vector ($1 \times K$).
- `s`: mean vector ($1 \times K$).
- `P`: loading matrix ($K \times ncomp$).
- `Pfull`: loading matrix ($K \times K$).
- `lambda`: vector with variances of the scores ($1 \times Ncomp$).
- `limspe`: Upper Control Limit (for α value) for the SPE.
- `limt2`: Upper Control Limit (for α value) for the T^2.
- `prepro`: string indicating preprocessing applied to X.
- `ncomp`: integer indicating the number of PCs of the model.

- `alpha`: value of the Type I risk assumed for the UCL.
- `n`: number of observations used in the PCA-MB.
- `S`: covariance matrix of observations used in the PCA-MB.
- `limits_t`: Control Limits for the scores with a confidence level $(1-\alpha)\times 100$ %

pcame

Description

Performs PCA Model Exploitation to the data in **X** using the PCA model information stored in the pcamodel struct.

Inputs

- `X`: double matrix of dimensions $N \times K$ with observations to be projected onto the PCA model stored in the `pcamodel` input argument.
- `pcamodel`: struct with the information of the PCA model, at least with fields m (mean vector), s (standard deviation vector), P (loading matrix), prepro (string with preprocessing), lambda (score variances vector), ncomp (number of PCs).

Outputs

`pcaout`: struct containing the information from the observations in **X** projected onto the PCA model of pcamodel. With fields:

- `T` : Scores matrix of ($N \times Ncomp$)
- `E` : Error matrix of ($N \times K$).
- `Xhat` : Prediction of **X** with the PCA model ($N \times K$).
- `SPE` : Squared Prediction Error vector ($N \times 1$).
- `T2` : Hotelling's T^2 vector ($N \times 1$).
- `T2cont` : Contributions to the T^2 ($N \times Ncomp$).

scoreplot

Description

Returns the score plot according to the input arguments.

Inputs

- `X`: matrix that will be projected onto the PCA model in `pcamodel` and whose scores will be displayed on the plot.
- `pcamodel`: struct with the information of the PCA model, at least with fields m (mean vector), s (standard deviation vector), P (loading matrix), prepro (string with preprocessing), lambda (score variances vector), ncomp (number of PCs).

Name - Value pair Input Arguments:

- `'clicktoggle'`: string to control interactive plot options with values `"on"` or `"off"` . Default set to `"on"` .
- `'pcx'`: integer with the x-axis PC. Default set to 1.
- `'pcy'`: integer with the y-axis PC. Default set to 2.
- `'obstag'`: column vector of integers indicating the group of each observation. Default value set to `zeros(size(X,1),1)` .
- `'steps_spe'`: column vector of integers indicating the *SPE* step of each observation. Default value set to `zeros(size(X,1),1)` .
- `'steps_spe'`: column vector of integers indicating the T^2 step of each observation. Default value set to `zeros(size(X,1),1)` .

Outputs

(none)

scoreplotsimple

Description

Returns the score plot.

Inputs

- `T`: matrix of dimensions $N \times Ncomp$ with the scores, **T**.
- `pcx`: integer indicating the PC in the horizontal axis. Default value set to 1.
- `pcy`: integer indicating the PC in the vertical axis.Default value set to 2.
- `obstag`: vector indicating the tag (0 for reference and 1 for new) of the observations. Default set to `zeros(size(T2))`.
- `alpha`: input with the Type I error assumed for the UCLs. Default set to 0.05. Must be within the (0; 1) interval.
- `varT`: vector with the variances of the scores of the PCA model. Default set to `var(T)`.

Outputs

(none) It returns the score plot.

scout

Description

Performs the SCOUTing on the observations of **X** according to the provided input parameters.

Inputs

- `X`: matrix with observations to be shifted as row-vectors.
- `pcamodel`: struct with the information of the PCA model.
- `mode`: string with procedure to generate steps. Accepted values are `'simple'`, `'steps'` and `'grid'`. Default value is set to `'simple'`.

Name-Value pair Input Arguments:

- `'T2y'`: Hotelling's T^2 target value for each observation in **X**. If no value is provided, the T^2 value of the observation is set as target, i.e.: the T^2 remains constant.
- `'SPEy'`: *SPE* target value for each observation in **X**. If no value is provided, the *SPE* value of the observation is set as target, i.e.: the *SPE* remains constant.
- `'nsteps'`: integer with number of steps for the *SPE* and the T^2. Default value set to 1.
- `'nstepsspe'`: integer with number of steps for the *SPE*. Default value set to 1.
- `'nstepst2'`: integer with number of steps for the T^2. Default value set to 1.
- `'gt2'`: number with T^2 speed parameter (γ_{T^2}). Default value set to 1.
- `'gspe'`: number with *SPE* speed parameter (γ_{SPE}). Default value set to 1.

Outputs

- `outscout` : struct with fields containing:

`X` : matrix with the shifted observations from **X**. Structure:

$$\begin{matrix} obs1 & stepT2_1 & stepSPE_1 \\ obs2 & ... & ... \\ ... & ... & ... \\ obsN & stepT2_1 & stepSPE_1 \\ obs1 & stepT2_2 & stepSPE_1 \\ ... & ... & ... \\ obsN & stepT2_M & stepSPE_1 \\ obs1 & stepT2_1 & stepSPE_2 \\ ... & ... & ... \\ obsN & stepT2_M & stepSPE_M \end{matrix}$$

`T2` : column vector with the T^2 values of the shifted observations.

`SPE` : column vector with the *SPE* values of the shifted observations.

`tag` : column vector indicating if the observation belongs to the reference data set (0) or to the new generated data (1).

`step_spe` : column vector indicating the step between SPE_x and SPE_y.

`step_t2` : column vector indicating the step between T^2_x and T^2_y.

- `SPE_0` : vector with the initial *SPE* values.

- `T2_0` : vector with the initial Hotelling's T^2 values.

scoutgrid

Description

Performs grid-wise SCOUTing on the observations of **X** according to the provided input parameters.

Inputs

- `X`: matrix with observations to be shifted as row-vectors.
- `pcamodel`: struct with the information of the PCA model.
- `T2target`: Hotelling's T^2 target value for each observation in **X**. If no value is provided, the T^2 value of the observation is set as target, i.e.: the T^2 remains constant.
- `SPEtarget`: *SPE* target value for each observation in **X**. If no value is provided, the *SPE* value of the observation is set as target, i.e.: the *SPE* remains constant.
- `nstepsspe`: integer with number of steps for the *SPE*. Default value set to 1.
- `nstepst2`: integer with number of steps for the T^2. Default value set to 1.
- `gt2`: number with T^2 speed parameter (γ_{T^2}). Default value set to 1.
- `gspe`: number with *SPE* speed parameter (γ_{SPE}). Default value set to 1.

Outputs

- `outscout` : struct with fields containing:

`X` : matrix with the shifted observations from **X**. Structure:

$$\begin{array}{ccc} obs1 & stepT2_1 & stepSPE_1 \\ obs2 & ... & ... \\ ... & ... & ... \\ obsN & stepT2_1 & stepSPE_1 \\ obs1 & stepT2_2 & stepSPE_1 \\ ... & ... & ... \\ obsN & stepT2_M & stepSPE_1 \\ obs1 & stepT2_1 & stepSPE_2 \\ ... & ... & ... \\ obsN & stepT2_M & stepSPE_M \end{array}$$

`T2` : column vector with the T^2 values of the shifted observations.

`SPE` : column vector with the *SPE* values of the shifted observations.

`tag` : column vector indicating if the observation belongs to the reference data set (0) or to the new generated data (1).

`step_spe` : column vector indicating the step between SPE_x and SPE_y.

`step_t2` : column vector indicating the step between T_x^2 and T_y^2.

- `SPE_0` : vector with the initial *SPE* values.
- `T2_0` : vector with the initial Hotelling's T^2 values.

scoutsimple

Description

Performs one-step SCOUTing on the observations of **X** according to the provided input parameters.

Inputs

- `X`: matrix with observations to be shifted as row-vectors.
- `pcamodel`: struct with the information of the PCA model.
- `T2target`: Hotelling's T^2 target value for each observation in **X**. If no value is provided, the T^2 value of the observation is set as target, i.e.: the T^2 remains constant.
- `SPEtarget`: *SPE* target value for each observation in **X**. If no value is provided, the *SPE* value of the observation is set as target, i.e.: the *SPE* remains constant.

Outputs

- `outscout` : struct with fields containing:

`X` : matrix with the shifted observations from **X**. Structure:

$$\begin{array}{cc} obs1 & 1 \quad 1 \\ obs2 & 1 \quad 1 \\ \ldots & 1 \quad 1 \\ obsN & 1 \quad 1 \end{array}$$

`T2` : column vector with the T^2 values of the shifted observations.

`SPE` : column vector with the *SPE* values of the shifted observations.

`tag` : column vector indicating if the observation belongs to the reference data set (0) or to the new generated data (1).

`step_spe` : column vector indicating the step between SPE_x and SPE_y.

`step_t2` : column vector indicating the step between T_x^2 and T_y^2.

- `SPE_0` : vector with the initial *SPE* values.
- `T2_0` : vector with the initial Hotelling's T^2 values

scoutsteps

Description

Performs the step-wise SCOUTing on the observations of **X** according to the provided input parameters.

Inputs

- `X`: matrix with observations to be shifted as row-vectors.
- `pcamodel`: struct with the information of the PCA model.
- `T2target`: Hotelling's T^2 target value for each observation in **X**. If no value is provided, the T^2 value of the observation is set as target, i.e.: the T^2 remains constant.
- `SPEtarget`: *SPE* target value for each observation in **X**. If no value is provided, the *SPE* value of the observation is set as target, i.e.: the *SPE* remains constant.
- `nsteps`: integer with number of steps for the *SPE* and the T^2. Default value set to 1.
- `'gt2'`: number with T^2 speed parameter (γ_{T^2}). Default value set to 1.
- `'gspe'`: number with *SPE* speed parameter (γ_{SPE}). Default value set to 1.

Outputs

- `outscout` : struct with fields containing:

`X` : matrix with the shifted observations from **X**. Structure:

$$\begin{array}{ccc} obs1 & step_1 & step_1 \\ obs2 & ... & ... \\ ... & ... & ... \\ obsN & step_1 & step_1 \\ obs1 & step_2 & step_2 \\ ... & ... & ... \\ obsN & step_M & step_M \end{array}$$

`T2` : column vector with the T^2 values of the shifted observations.

`SPE` : column vector with the *SPE* values of the shifted observations.

`tag` : column vector indicating if the observation belongs to the reference data set (0) or to the new generated data (1).

`step_spe` : column vector indicating the step between SPE_x and SPE_y.

`step_t2` : column vector indicating the step between T^2_x and T^2_y.

- `SPE_0` : vector with the initial *SPE* values.
- `T2_0` : vector with the initial Hotelling's T^2 values.

speinfo

Description

Information about Squared Prediction Error (*SPE*) for an observation, i.e.: information about the distance of the observation to the PCA model.

Inputs

- `SPE`: vector with values of the *SPE* statistic.
- `E`: matrix with the contributions of each variable (columns) for each observation (rows) to the *SPE*.
- `limspe`: value of the *SPE* Upper Control Limit (with confidence level $(1 - alpha) * 100$ %).
- `iobs`: integer with the index of the observation of interest.

Outputs

- `barobs`: axis of the bar plot with the *SPE* value.
- `barcont`: axis of the bar plot with the contributions to the *SPE*.

xshift

Description

Performs a shift to each row in `X`, increasing by factors a and b the distance of the observations according to the PCA model expressed in `P`.

Inputs

- `X`: data matrix with observations to be shifted.
- `P`: loading matrix of the PCA model according to which the observations in `X` will change their distance.
- `a`: column vector with the factor which tunes the increment of the Hotelling's T^2 for its corresponding row in `X` .
- `b`: column vector with the factor which tunes the increment of the *SPE* for its corresponding row in `X` .

Outputs

`Xnew` : data matrix with the same dimensions as `X` , with each observation shifted.

Published with MATLAB® R2020a