Package ‘bandsfdp’

May 12, 2023

Title Compute Upper Prediction Bounds on the FDP in Competition-Based Setups

Version 1.1.0

Suggests fdpbandsdata

Description Implements functions that calculate upper prediction bounds on the false discovery proportion (FDP) in the list of discoveries returned by competition-based setups, implementing Ebadi et al. (2022) <arXiv:2302.11837>. Such setups include target-decoy competition (TDC) in computational mass spectrometry and the knockoff construction in linear regression (note this package typically uses the terminology of TDC). Included is the standardized (TDC-SB) and uniform (TDC-UB) bound on TDC's FDP, and the simultaneous standardized and uniform bands. Requires pre-computed Monte Carlo statistics available at <https://github.com/uni-Arya/fdpbandsdata>. This data can be downloaded by running the command 'devtools::install_github(``uni-Arya/fdpbandsdata")' in R and restarting R after installation. The size of this data is roughly 81Mb.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.1.2

URL https://github.com/uni-Arya/bandsfdp

BugReports https://github.com/uni-Arya/bandsfdp/issues

NeedsCompilation no

Author Arya Ebadi [aut, cre],
Dong Luo [aut],
Jack Freestone [aut],
William Stafford Noble [aut],
Uri Keich [aut] (<https://orcid.org/0000-0002-3209-5011>)

Maintainer Arya Ebadi <aeba3842@uni.sydney.edu.au>

Repository CRAN

Date/Publication 2023-05-12 07:20:03 UTC

R topics documented:

gen_bound *Generalized band*

Description

This function computes an upper prediction bound on the FDP among target wins in any set R of hypotheses of TDC. See details for more information.

Usage

```
gen_bound(
  labels,
  indices,
  gamma,
  type,
  d_max = NULL,
  max_fdp = 0.5,
  c = 0.5,
  lambda = 0.5
)

genband(
  labels,
  indices,
  gamma,
  type,
  d_max = NULL,
  max_fdp = 0.5,
  c = 0.5,
  lambda = 0.5
)
```

Arguments

`labels`	A vector of (ordered) labels. See details below.
`indices`	A vector specifying the indices of hypotheses for which an upper prediction bound on the FDP is computed.
`gamma`	The confidence parameter of the band. Typical values include `gamma = 0.05` or `gamma = 0.01`.

`type` A character string specifying which band to use. Must be one of `"stband"` or `"uniband"`.

`d_max` An optional positive integer specifying the maximum number of decoy wins considered in calculating the bands.

`max_fdp` A number specifying the maximum FDP considered by the user in calculating the bands. Used to compute `d_max` if `d_max` is set to `NULL`.

`c` Determines the ranks of the target score that are considered winning. Defaults to `c = 0.5` for (single-decoy) TDC.

`lambda` Determines the ranks of the target score that are considered losing. Defaults to `lambda = 0.5` for (single-decoy) TDC.

Details

In (single-decoy) TDC, each hypothesis is associated to a winning score and a label (1 for a target win, -1 for a decoy win). This function assumes that the hypotheses are ordered in decreasing order of winning scores (with ties broken at random). The argument `labels`, therefore, must be ordered according to this rule.

This function also supports the extension of TDC that uses multiple decoys. In that setup, the target score is competed with multiple decoy scores and the rank of the target score after competition is used to determine whether the hypothesis is a target win (label = 1), decoy win (-1) or uncounted (0). The top `c` proportion of ranks are considered winning, the bottom `1-lambda` losing, and all the rest uncounted.

The threshold of TDC is given by the formula (assuming hypotheses are ordered):

$$\max\{k : \frac{D_k + 1}{T_k \vee 1} \cdot \frac{c}{1-\lambda} \leq \alpha\}$$

where T_k is the number of target wins among the top k hypotheses, and D_k is the number of decoy wins similarly.

The argument `gamma` sets a confidence level of `1-gamma`. Both the uniform and standardized bands require pre-computed Monte Carlo statistics, so only certain values of `gamma` are available to use. Commonly used confidence levels, like 0.95 and 0.99, are available. We refer the reader to the README of this package for more details.

The argument `d_max` controls the rate at which the returned bounds increase: a larger `d_max` results in a more conservative bound. If, however, $D_k + 1$ exceeds `d_max` for some index k, each target win thereafter is considered a false discovery when computing the bound. Thus it is important that `d_max`, chosen a priori, is large enough. Given it is sufficiently large, the precise value of `d_max` does not have a significant effect on the resulting bounds (see `https://arxiv.org/abs/2302.11837` for more details).

We recommend setting `d_max = NULL` so that it is computed automatically using `max_fdp`. This argument ensures that $D_k + 1$ never exceeds `d_max` when the (non-interpolated) FDP bound on the top k hypotheses is less than `max_fdp`.

Value

An upper prediction bound on the FDP among target wins in the set of hypotheses whose `indices` are given as input.

References

Ebadi et al. (2022), Bounding the FDP in competition-based control of the FDR https://arxiv.org/abs/2302.11837.

Examples

```
if (requireNamespace("fdpbandsdata", quietly = TRUE)) {
  set.seed(123)
  labels <- c(
    rep(1, 250),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.9, 0.1)),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.5, 0.5)),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.1, 0.9))
  )
  indices <- c(1:100, 300:400, 600:650)
  gamma <- 0.05
  gen_bound(labels, indices, gamma, "stband")
}
```

sim_bound *Simultaneous Band*

Description

This function computes upper prediction bounds on the target wins among the top k hypotheses of TDC, for each $k = 1, \dots, n$ where n is the total number of hypotheses.

Usage

```
sim_bound(
  labels,
  gamma,
  type,
  d_max = NULL,
  max_fdp = 0.5,
  c = 0.5,
  lambda = 0.5
)

simband(
  labels,
  gamma,
  type,
  d_max = NULL,
  max_fdp = 0.5,
  c = 0.5,
  lambda = 0.5
)
```

Arguments

`labels` A vector of (ordered) labels. See details below.

`gamma` The confidence parameter of the band. Typical values include `gamma = 0.05` or `gamma = 0.01`.

`type` A character string specifying which band to use. Must be one of `"stband"` or `"uniband"`.

`d_max` An optional positive integer specifying the maximum number of decoy wins considered in calculating the bands.

`max_fdp` A number specifying the maximum FDP considered by the user in calculating the bands. Used to compute `d_max` if `d_max` is set to `NULL`.

`c` Determines the ranks of the target score that are considered winning. Defaults to `c = 0.5` for (single-decoy) TDC.

`lambda` Determines the ranks of the target score that are considered losing. Defaults to `lambda = 0.5` for (single-decoy) TDC.

Details

In (single-decoy) TDC, each hypothesis is associated to a winning score and a label (1 for a target win, -1 for a decoy win). This function assumes that the hypotheses are ordered in decreasing order of winning scores (with ties broken at random). The argument `labels`, therefore, must be ordered according to this rule.

This function also supports the extension of TDC that uses multiple decoys. In that setup, the target score is competed with multiple decoy scores and the rank of the target score after competition is used to determine whether the hypothesis is a target win (label = 1), decoy win (-1) or uncounted (0). The top `c` proportion of ranks are considered winning, the bottom `1-lambda` losing, and all the rest uncounted.

The threshold of TDC is given by the formula (assuming hypotheses are ordered):

$$\max\{k : \frac{D_k + 1}{T_k \vee 1} \cdot \frac{c}{1 - \lambda} \leq \alpha\}$$

where T_k is the number of target wins among the top k hypotheses, and D_k is the number of decoy wins similarly.

The argument `gamma` sets a confidence level of `1-gamma`. Both the uniform and standardized bands require pre-computed Monte Carlo statistics, so only certain values of `gamma` are available to use. Commonly used confidence levels, like 0.95 and 0.99, are available. We refer the reader to the README of this package for more details.

The argument `d_max` controls the rate at which the returned bounds increase: a larger `d_max` results in a more conservative bound. If, however, $D_k + 1$ exceeds `d_max` for some index k, each target win thereafter is considered a false discovery when computing the bound. Thus it is important that `d_max`, chosen a priori, is large enough. Given it is sufficiently large, the precise value of `d_max` does not have a significant effect on the resulting bounds (see `https://arxiv.org/abs/2302.11837` for more details).

We recommend setting `d_max = NULL` so that it is computed automatically using `max_fdp`. This argument ensures that $D_k + 1$ never exceeds `d_max` when the (non-interpolated) FDP bound on the top k hypotheses is less than `max_fdp`.

Value

A vector of upper prediction bounds on the FDP of target wins among the top k hypotheses for each $k = 1, \ldots, n$ where n is the total number of hypotheses.

References

Ebadi et al. (2022), Bounding the FDP in competition-based control of the FDR https://arxiv.org/abs/2302.11837.

Examples

```
if (requireNamespace("fdpbandsdata", quietly = TRUE)) {
  set.seed(123)
  labels <- c(
    rep(1, 250),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.9, 0.1)),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.5, 0.5)),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.1, 0.9))
  )
  gamma <- 0.05
  head(sim_bound(labels, gamma, "stband"))
}
```

tdc_sb *Standardized Band*

Description

This function computes an upper prediction bound, derived from the standardized band, on the FDP in TDC's list of discoveries.

Usage

```
tdc_sb(
  thresholds,
  labels,
  alpha,
  gamma,
  c = 0.5,
  lambda = 0.5,
  n = length(labels),
  interpolate = TRUE
)

stband(
  thresholds,
  labels,
  alpha,
  gamma,
  c = 0.5,
  lambda = 0.5,
  n = length(labels),
  interpolate = TRUE
)
```

Arguments

`thresholds` The rejection threshold of TDC. If given as a vector, an upper prediction bound is returned for each element.

`labels` A vector of (ordered) labels. See details below.

`alpha` The FDR threshold.

`gamma` The confidence parameter of the bound. Typical values include `gamma = 0.05` or `gamma = 0.01`.

`c` Determines the ranks of the target score that are considered winning. Defaults to `c = 0.5` for (single-decoy) TDC.

`lambda` Determines the ranks of the target score that are considered losing. Defaults to `lambda = 0.5` for (single-decoy) TDC.

`n` The number of hypotheses. Defaults to the length of `labels`.

`interpolate` A boolean indicating whether the bands should be interpolated. Offers a slight boost in performance at the cost of computing power. Defaults to `TRUE`.

Details

In (single-decoy) TDC, each hypothesis is associated to a winning score and a label (1 for a target win, -1 for a decoy win). This function assumes that the hypotheses are ordered in decreasing order of winning scores (with ties broken at random). The argument `labels`, therefore, must be ordered according to this rule.

This function also supports the extension of TDC that uses multiple decoys. In that setup, the target score is competed with multiple decoy scores and the rank of the target score after competition is used to determine whether the hypothesis is a target win (label = 1), decoy win (-1) or uncounted (0). The top `c` proportion of ranks are considered winning, the bottom `1-lambda` losing, and all the rest uncounted.

The threshold of TDC is given by the formula:

$$\max\{k : \frac{D_k + 1}{T_k \vee 1} \cdot \frac{c}{1 - \lambda} \leq \alpha\}$$

where T_k is the number of target wins among the top k hypotheses, and D_k is the number of decoy wins similarly.

The argument `gamma` sets a confidence level of `1-gamma`. Since the standardized band requires precomputed Monte Carlo quantiles, only certain values of `gamma` are available to use. Commonly used confidence levels, like 0.95 and 0.99, are available. We refer the reader to the README of this package for more details.

The argument alpha, used to compute the threshold of TDC, is also used in this function. It serves to compute an appropriate d_max for a non-trivial bound. In particular, if the user inputs a vector of thresholds, a bound is returned for each element of thresholds using the same d_max. For more details, see: https://arxiv.org/abs/2302.11837.

We recommend the use of interpolate = TRUE (default), as it generally results in a tighter bound. This comes at the cost of performance: the bound for each threshold is computed in O(n) time with interpolation and O(1) without.

Value

An upper prediction bound on the FDP in TDC's list of discoveries. If thresholds is a vector, returns an upper prediction bound for each element of thresholds.

References

Ebadi et al. (2022), Bounding the FDP in competition-based control of the FDR https://arxiv.org/abs/2302.11837.

Examples

```
if (requireNamespace("fdpbandsdata", quietly = TRUE)) {
  set.seed(123)
  thresholds <- c(250, 500, 750, 1000)
  labels <- c(
    rep(1, 250),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.9, 0.1)),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.5, 0.5)),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.1, 0.9))
  )
  alpha <- 0.05
  gamma <- 0.05
  tdc_sb(thresholds, labels, alpha, gamma)
}
```

tdc_ub *Uniform Band*

Description

This function computes an upper prediction bound, derived from the uniform band, on the FDP in TDC's list of discoveries.

Usage

```
tdc_ub(
  thresholds,
  labels,
  alpha,
  gamma,
  c = 0.5,
  lambda = 0.5,
  n = length(labels),
  interpolate = TRUE
)

uniband(
  thresholds,
  labels,
  alpha,
  gamma,
  c = 0.5,
  lambda = 0.5,
  n = length(labels),
  interpolate = TRUE
)
```

Arguments

`thresholds` The rejection threshold of TDC. If given as a vector, an upper prediction bound is returned for each element.

`labels` A vector of (ordered) labels. See details below.

`alpha` The FDR threshold.

`gamma` The confidence parameter of the bound. Typical values include `gamma = 0.05` or `gamma = 0.01`.

`c` Determines the ranks of the target score that are considered winning. Defaults to `c = 0.5` for (single-decoy) TDC.

`lambda` Determines the ranks of the target score that are considered losing. Defaults to `lambda = 0.5` for (single-decoy) TDC.

`n` The number of hypotheses. Defaults to the length of `labels`.

`interpolate` A boolean indicating whether the bands should be interpolated. Offers a slight boost in performance at the cost of computing power. Defaults to `TRUE`.

Details

In (single-decoy) TDC, each hypothesis is associated to a winning score and a label (1 for a target win, -1 for a decoy win). This function assumes that the hypotheses are ordered in decreasing order of winning scores (with ties broken at random). The argument `labels`, therefore, must be ordered according to this rule.

This function also supports the extension of TDC that uses multiple decoys. In that setup, the target score is competed with multiple decoy scores and the rank of the target score after competition is used to determine whether the hypothesis is a target win (label = 1), decoy win (-1) or uncounted (0). The top `c` proportion of ranks are considered winning, the bottom `1-lambda` losing, and all the rest uncounted.

The threshold of TDC is given by the formula:

$$\max\{k : \frac{D_k + 1}{T_k \vee 1} \cdot \frac{c}{1-\lambda} \leq \alpha\}$$

where T_k is the number of target wins among the top k hypotheses, and D_k is the number of decoy wins similarly.

The argument `gamma` sets a confidence level of `1-gamma`. Since the uniform band requires pre-computed Monte Carlo statistics, only certain values of `gamma` are available to use. Commonly used confidence levels, like 0.95 and 0.99, are available. We refer the reader to the README of this package for more details.

The argument `alpha`, used to compute the threshold of TDC, is also used in this function. It serves to compute an appropriate d_max for a non-trivial bound. In particular, if the user inputs a vector of `thresholds`, a bound is returned for each element of `thresholds` using the same d_max. For more details, see: https://arxiv.org/abs/2302.11837.

We recommend the use of `interpolate = TRUE` (default), as it generally results in a tighter bound. This comes at the cost of performance: the bound for each threshold is computed in O(n) time with interpolation and O(1) without.

Value

An upper prediction bound on the FDP in TDC's list of discoveries. If `thresholds` is a vector, returns an upper prediction bound for each element of `thresholds`.

References

Ebadi et al. (2022), Bounding the FDP in competition-based control of the FDR https://arxiv.org/abs/2302.11837.

Examples

```
if (requireNamespace("fdpbandsdata", quietly = TRUE)) {
  set.seed(123)
  thresholds <- c(250, 500, 750, 1000)
  labels <- c(
    rep(1, 250),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.9, 0.1)),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.5, 0.5)),
    sample(c(1, -1), size = 250, replace = TRUE, prob = c(0.1, 0.9))
  )
  alpha <- 0.05
  gamma <- 0.05
  tdc_ub(thresholds, labels, alpha, gamma)
}
```

Index