Package 'BayesCVI'

February 26, 2024

Type Package

Title Bayesian Cluster Validity Index

Version 1.0.0

Imports e1071, mclust, ggplot2, UniversalCVI

Description Algorithms for computing and generating plots with and without error bars for Bayesian cluster validity index (BCVI) (N. Wiroonsri, O. Preedasawakul (2024) <arXiv:2402.02162>) based on several underlying cluster validity indexes (CVIs) including Calinski-Harabasz, Chou-Su-Lai, Davies-Bouldin, Dunn, Pakhira-Bandyopadhyay-Maulik, Point biserial correlation, the score function, Starczewski, and Wiroonsri indices for hard clustering, and Correlation Cluster Validity, the generalized C, HF, KWON, KWON2, Modified Pakhira-Bandyopadhyay-Maulik, Pakhira-Bandyopadhyay-Maulik, Tang, Wiroonsri-Preedasawakul, Wu-Li, and Xie-Beni indices for soft clustering. The package is compatible with K-means, fuzzy C means, EM clustering, and hierarchical clustering (single, average, and complete linkage). Though BCVI is compatible with any underlying existing CVIs, we recommend users to use either WI or WP as the underlying CVI.

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

Depends R (>= 2.10)

NeedsCompilation no

Author Nathakhun Wiroonsri [aut] (<https://orcid.org/0000-0003-2167-9641>), Onthada Preedasawakul [cre, aut] (<https://orcid.org/0000-0002-4186-3158>)

Maintainer Onthada Preedasawakul <o.preedasawakul@gmail.com>

Repository CRAN

Date/Publication 2024-02-26 16:20:12 UTC

R topics documented:

`B1_data` *B1 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2024) generated from 1 Gaussian and 1 Uniform distributions labeled as 1-2.

Usage

```
B1_data
```

Format

A data frame with 5500 data points and 3 variables

- `x` Numeric values generated from Gaussian and Uniform distributions
- `y` Numeric values generated from Gaussian and Uniform distributions
- `label` Categorical labels 1,2

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A Bayesian cluster validity index, arXiv:2402.02162, 2024

See Also

B2_data, B3_data, B_WP.IDX, B_Wvalid, B_XB.IDX

`B2_data` *B2 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2024) generated from 5 different Gaussian distributions labeled as 1-5.

Usage

```
B2_data
```

Format

A data frame with 850 data points and 3 variables

- `x` Numeric values generated from Gaussian distributions
- `y` Numeric values generated from Gaussian distributions
- `label` Categorical labels 1,2,3,4,5

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A Bayesian cluster validity index, arXiv:2402.02162, 2024

See Also

B1_data, B3_data, B_WP.IDX, B_Wvalid, B_XB.IDX

`B3_data` *B3 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2024) generated from 5 different Gaussian distributions labeled as 1-5.

Usage

```
B3_data
```

Format

A data frame with 2300 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A Bayesian cluster validity index, arXiv:2402.02162, 2024

See Also

B2_data, B4_data, B_WP.IDX, B_Wvalid, B_XB.IDX

B4_data *B4 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2024) generated from 6 different Gaussian distributions labeled as 1-6.

Usage

```
B4_data
```

Format

A data frame with 740 data points and 3 variables

`x` Numeric values generated from Gaussian distributions

`y` Numeric values generated from Gaussian distributions

`label` Categorical labels 1,2,3,4,5,6

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A Bayesian cluster validity index, arXiv:2402.02162, 2024

See Also

B3_data, B5_data, B_WP.IDX, B_Wvalid, B_XB.IDX

B5_data *B5 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2024) generated from 7 different Gaussian and 2 Uniform distributions labeled as 1-9.

Usage

```
B5_data
```

Format

A data frame with 1820 data points and 3 variables

`x` Numeric values generated from Gaussian and Uniform distributions

`y` Numeric values generated from Gaussian and Uniform distributions

`label` Categorical labels 1,2,3,4,5,6,7,8,9

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A Bayesian cluster validity index, arXiv:2402.02162, 2024

See Also

B4_data, B6_data, B_WP.IDX, B_Wvalid, B_XB.IDX

`B6_data` *B6 Artificial Dataset*

Description

A `2`-dimensional dataset from Wiroonsri and Preedasawakul (2024) generated from `3` different Gaussian and `2` Uniform distributions labeled as `1-5`.

Usage

```
B6_data
```

Format

A data frame with 1000 data points and 3 variables

`x` Numeric values generated from Gaussian and Uniform distributions

`y` Numeric values generated from Gaussian and Uniform distributions

`label` Categorical labels 1,2,3,4,5

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A Bayesian cluster validity index, arXiv:2402.02162, 2024

See Also

B5_data, B7_data, B_WP.IDX, B_Wvalid, B_XB.IDX

`B7_data` *B7 Artificial Dataset*

Description

A 2-dimensional dataset from Wiroonsri and Preedasawakul (2024) generated from 3 different Gaussian and 2 Uniform distributions labeled as 1-5.

Usage

```
B7_data
```

Format

A data frame with 800 data points and 3 variables

`x` Numeric values generated from Gaussian and Uniform distributions

`y` Numeric values generated from Gaussian and Uniform distributions

`label` Categorical labels 1,2,3,4,5

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, A Bayesian cluster validity index, arXiv:2402.02162, 2024

See Also

B6_data, B1_data, B_WP.IDX, B_Wvalid, B_XB.IDX

`BayesCVIs` *Bayesian cluster validity index*

Description

Compute Bayesian cluster validity index (BCVI) from two to `kmax` groups using an underlying cluster validity index (CVI) and Dirichlet prior parameters of the user's choice. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
BayesCVIs(CVI, n, kmax, opt.pt, alpha = "default", mult.alpha = 1/2)
```

Arguments

`CVI`	the CVI values for `k` from `2` to `kmax` to be used as the underlying index for computing BCVI.
`n`	a number of data point.
`kmax`	a maximum number of clusters to be considered.
`opt.pt`	a character string indicating whether the maximum or the minimum of `CVI` specifies the optimal number of groups (`"min"` or `"max"`).
`alpha`	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter `mult.alpha` to be its multiplier. The default is `"default"`.)
`mult.alpha`	the power s from n^s to be multiplied to the Dirichlet prior parameters `alpha` (selecting `mult.alpha` in `[0,1)` is recommended). The default is $\frac{1}{2}$.

Details

BCVI is defined as follows. Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{CVI(j)} - \mathbf{CVI(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{CVI(j)} - \mathbf{CVI(i)})}$$

for a CVI such that the smallest value indicates the optimal number of clusters and

$$r_k(\mathbf{x}) = \frac{\mathbf{CVI(k)} - \min_{\mathbf{j}} \mathbf{CVI(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{CVI(i)} - \min_{\mathbf{j}} \mathbf{CVI(j)})}$$

for a CVI such that the largest value indicates the optimal number of clusters. Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

`BCVI`	the dataframe where the first and the second columns are the number of groups `k` and BCVI(k), respectively, for `k` from `2` to `kmax`.
`VAR`	the data frame where the first and the second columns are the number of groups `k` and the variance of p_k, respectively, for `k` from `2` to `kmax`.
`CVI`	the data frame where the first and the second columns are the number of groups `k` and the original CVI(k), respectively, for `k` from `2` to `kmax`.
`opt.pt`	a character string indicating whether the maximum or the minimum of `CVI` specifies the optimal number of groups (`"min"` or `"max"`) that user select.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

`B2_data`, `B_TANG.IDX`, `B_WP.IDX`, `B_Wvalid`, `B_DB.IDX`

Examples

```
# install a package for computing an underlying CVI
# install.packages("UniversalCVI")

library(UniversalCVI)
library(BayesCVI)

data = R1_data[,-3]
```

```
# Compute WP index by WP.IDX using default gamma
FCM.WP = WP.IDX(scale(data), cmax = 10, cmin = 2, corr = 'pearson', method = 'FCM', fzm = 2,
                iter = 100, nstart = 20, NCstart = TRUE)


# WP.IDX values
result = FCM.WP$WP$WPI


aalpha = c(20,20,20,5,5,5,0.5,0.5,0.5)
B.WP = BayesCVIs(CVI = result,
          n = nrow(data),
          kmax = 10,
          opt.pt = "max",
          alpha = aalpha,
          mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.WP)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_CCV.IDX *BCVI-Correlation Cluster Validity (CCV) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using the pearson correlation cluster validity (CCVP) and/or the spearman's (rho) correlation cluster validity (CCVS) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_CCV.IDX(x, kmax, indexlist = "all", method = "FCM", fzm = 2,
        iter = 100, nstart = 20, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
indexlist	a character string indicating which The generalized C index be computed ("all","CCVP","CCVS"). More than one indexes can be selected.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".

fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
iter	a maximum number of iterations for method = "FCM". The default is 100.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-CCV is defined as follows. Let

$$r_k(\mathbf{x}) = \frac{\mathbf{CVI(k)} - \min_{\mathbf{j}} \mathbf{CVI(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{CVI(i)} - \min_{\mathbf{j}} \mathbf{CVI(j)})}$$

where CVI is either CCVP or CCVS index.

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original CCVP(k) or CCVS(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

M. Popescu, J. C. Bezdek, T. C. Havens and J. M. Keller (2013). "A Cluster Validity Framework Based on Induced Partition Dissimilarity." https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6246717&isnumber=6340245

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_TANG.IDX, B_XB.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(20,20,20,5,5,5,0.5,0.5,0.5)

B.CCV = B_CCV.IDX(x = scale(data), kmax=10, indexlist = "CCVP", method = "FCM", fzm = 2, iter = 100,
              nstart = 20, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI-CCVP

pplot = plot_BCVI(B.CCV$CCVP)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_CH.IDX *BCVI-Calinski–Harabasz (CH) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Calinski–Harabasz (CH) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_CH.IDX(x, kmax, method = "kmeans", nstart = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-CH is defined as follows. Let

$$r_k(\mathbf{x}) = \frac{\mathbf{CH(k)} - \min_{\mathbf{j}} \mathbf{CH(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{CH(i)} - \min_{\mathbf{j}} \mathbf{CH(j)})}$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original CH(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

T. Calinski, J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, 3, 1-27 (1974).

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B2_data, B_TANG.IDX, B_XB.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B2_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.CH = B_CH.IDX(x = scale(data), kmax=10, method = "kmeans",
                nstart = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.CH)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_CSL.IDX *BCVI-Chou-Su-Lai (CSL) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Chou-Su-Lai (CSL) as the underlying cluster validity index (CVI) and Dirichlet prior parameters of the user's choice. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_CSL.IDX(x, kmax, method = "kmeans", nstart = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-CSL is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{CSL(j)} - \mathbf{CSL(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{CSL(j)} - \mathbf{CSL(i)})}.$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

`BCVI`	the dataframe where the first and the second columns are the number of groups `k` and BCVI(k), respectively, for `k` from `2` to `kmax`.
`VAR`	the data frame where the first and the second columns are the number of groups `k` and the variance of p_k, respectively, for `k` from `2` to `kmax`.
`CVI`	the data frame where the first and the second columns are the number of groups `k` and the original CSL(k), respectively, for `k` from `2` to `kmax`.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

C. H. Chou, M. C. Su, E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Anal Applic*, 7, 205-220 (2004).

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

`B2_data`, `B_TANG.IDX`, `B_WP.IDX`, `B_Wvalid`, `B_DB.IDX`

Examples

```
library(BayesCVI)

# The data included in this package.
data = B2_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.CSL = B_CSL.IDX(x = scale(data), kmax=10, method = "kmeans",
                nstart = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI
```

```
pplot = plot_BCVI(B.CSL)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_DB.IDX *BCVI-Davies–Bouldin (DB) and DB* (DBs) indexes*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using DB and/or DBs as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_DB.IDX(x, kmax, method = "kmeans", indexlist = "all", p = 2, q = 2,
        nstart = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
indexlist	a character string indicating which cluster validity indexes to be computed ("all", "DB", "DBs"). More than one indexes can be selected.
p	the power of the Minkowski distance between centroids of clusters. The default is 2.
q	the power of dispersion measure of a cluster. The default is 2.
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-DB is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{CVI(j)} - \mathbf{CVI(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{CVI(j)} - \mathbf{CVI(i)})}.$$

where CVI indicates DB or DBs index.
Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

`BCVI`	the dataframe where the first and the second columns are the number of groups `k` and BCVI(k), respectively, for `k` from `2` to `kmax`.
`VAR`	the data frame where the first and the second columns are the number of groups `k` and the variance of p_k, respectively, for `k` from `2` to `kmax`.
`CVI`	the data frame where the first and the second columns are the number of groups `k` and the original DB(k) or DBs(k), respectively, for `k` from `2` to `kmax`.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

D. L. Davies, D. W. Bouldin, "A cluster separation measure," *IEEE Trans Pattern Anal Machine Intell*, 1, 224-227 (1979).

M. Kim, R. S. Ramakrishna, "New indices for cluster validity assessment," *Pattern Recognition Letters*, 26, 2353-2363 (2005).

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B2_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DI.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B2_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.DB = B_DB.IDX(x = scale(data), kmax=10, method = "kmeans", indexlist = "all",
              p = 2, q = 2, nstart = 100, alpha = "default", mult.alpha = 1/2)

# plot the BCVI-DB

pplot = plot_BCVI(B.DB$DB)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot

# plot the BCVI-DBs

pplot = plot_BCVI(B.DB$DBs)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_DI.IDX *BCVI-Dunn index (DI)*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Dunn index (DI) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_DI.IDX(x, kmax, method = "kmeans", nstart = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.

method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-DI is defined as follows. Let

$$r_k(\mathbf{x}) = \frac{\mathbf{DI(k)} - \min_{\mathbf{j}} \mathbf{DI(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{DI(i)} - \min_{\mathbf{j}} \mathbf{DI(j)})}$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original DI(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J Cybern*, 3(3), 32-57 (1973).

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B2_data, B_TANG.IDX, B_XB.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B2_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.DI = B_DI.IDX(x = scale(data), kmax=10, method = "kmeans",
                nstart = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.DI)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_GC.IDX *BCVI-The generalized C (GC) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using all or part of GC1 GC2 GC3 and GC4 as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_GC.IDX(x, kmax, indexlist = "all", method = "FCM", fzm = 2, iter = 100,
        nstart = 20, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
indexlist	a character string indicating which The generalized C index be computed ("all","GC1","GC2","GC3","GC4" More than one indexes can be selected.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
iter	a maximum number of iterations for method = "FCM". The default is 100.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-GC is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{CVI(j)} - \mathbf{CVI(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{CVI(j)} - \mathbf{CVI(i)})}.$$

where CVI is one of the GC1 GC2 GC3 or GC4 index.

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original GC1(k) GC2(k) GC3(k) GC4(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

J. C. Bezdek, M. Moshtaghi, T. Runkler, and C. Leckie, "The generalized c index for internal fuzzy cluster validity," IEEE Transactions on Fuzzy Systems, vol. 24, no. 6, pp. 1500–1512, 2016. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7429723&isnumber=7797168

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_TANG.IDX, B_XB.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.GC = B_GC.IDX(x = scale(data), kmax = 10, indexlist = "GC1",
              method = "FCM", fzm = 2, iter = 100,
                nstart = 20, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI-GC1

pplot = plot_BCVI(B.GC$GC1)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_HF.IDX *BCVI-HF index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using HF as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_HF.IDX(x, kmax, method = "FCM", fzm = 2, nstart = 20,
        iter = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-HF is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{HF(j)} - \mathbf{HF(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{HF(j)} - \mathbf{HF(i)})}.$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

`BCVI`	the dataframe where the first and the second columns are the number of groups `k` and BCVI(k), respectively, for `k` from `2` to `kmax`.
`VAR`	the data frame where the first and the second columns are the number of groups `k` and the variance of p_k, respectively, for `k` from `2` to `kmax`.
`CVI`	the data frame where the first and the second columns are the number of groups `k` and the original HF(k), respectively, for `k` from `2` to `kmax`.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

F. Haouas, Z. Ben Dhiaf, A. Hammouda and B. Solaiman, "A new efficient fuzzy cluster validity index: Application to images clustering," 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 2017, pp. 1-6. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8015651&isnumber=8015374

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
```

```
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.HF = B_HF.IDX(x = scale(data), kmax =10, method = "FCM", fzm = 2,
              nstart = 20, iter = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.HF)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_KPBM.IDX *BCVI-Modified Kernel form of Pakhira-Bandyopadhyay-Maulik (KPBM) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Modified Kernel form of Pakhira-Bandyopadhyay-Maulik (KPBM) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_KPBM.IDX(x, kmax, method = "FCM", fzm = 2, nstart = 20,
          iter = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-KPBM is defined as follows. Let

$$r_k(\mathbf{x}) = \frac{\mathbf{KPBM(k)} - \min_{\mathbf{j}} \mathbf{KPBM(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{KPBM(i)} - \min_{\mathbf{j}} \mathbf{KPBM(j)})}$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original KPBM(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

C. Alok. (2010). "An investigation of clustering algorithms and soft computing approaches for pattern recognition," Department of Computer Science, Assam University. http://hdl.handle.net/10603/93443

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.KPBM = B_KPBM.IDX(x = scale(data), kmax =10, method = "FCM", fzm = 2, nstart = 20,
                iter = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.KPBM)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_KWON.IDX *BCVI-KWON index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using KWON as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_KWON.IDX(x, kmax, method = "FCM", fzm = 2, nstart = 20,
          iter = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.

`alpha` Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter `mult.alpha` to be its multiplier. The default is `"default"`.)

`mult.alpha` the power s from n^s to be multiplied to the Dirichlet prior parameters `alpha` (selecting `mult.alpha` in `[0,1)` is recommended). The default is $\frac{1}{2}$.

Details

BCVI-KWON is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{KWON(j)} - \mathbf{KWON(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{KWON(j)} - \mathbf{KWON(i)})}.$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

`BCVI` the dataframe where the first and the second columns are the number of groups `k` and BCVI(k), respectively, for `k` from `2` to `kmax`.

`VAR` the data frame where the first and the second columns are the number of groups `k` and the variance of p_k, respectively, for `k` from `2` to `kmax`.

`CVI` the data frame where the first and the second columns are the number of groups `k` and the original KWON(k), respectively, for `k` from `2` to `kmax`.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

S. H. Kwon, “Cluster validity index for fuzzy clustering,” Electronics letters, vol. 34, no. 22, pp. 2176–2177, 1998. doi:10.1049/el:19981523

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.KWON = B_KWON.IDX(x = scale(data), kmax =10, method = "FCM", fzm = 2, nstart = 20,
                    iter = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.KWON)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_KWON2.IDX *BCVI-KWON2 index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using KWON2 as the underling cluster validity index (CVI) with the user’s selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_KWON2.IDX(x, kmax, method = "FCM", fzm = 2, nstart = 20,
          iter = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-KWON2 is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{KWON2(j)} - \mathbf{KWON2(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{KWON2(j)} - \mathbf{KWON2(i)})}.$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original KWON2(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

S. H. Kwon, J. Kim, and S. H. Son, “Improved cluster validity index for fuzzy clustering,” Electronics Letters, vol. 57, no. 21, pp. 792–794, 2021.

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.KWON2 = B_KWON2.IDX(x = scale(data), kmax =10, method = "FCM", fzm = 2,
                  nstart = 20, iter = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.KWON2)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_PB.IDX *BCVI-Point biserial correlation (PB)*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Point biserial correlation (PB) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_PB.IDX(x, kmax, method = "kmeans", corr = "pearson", nstart = 100,
        alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
corr	a character string indicating which correlation coefficient is to be computed ("pearson", "kendall" or "spearman"). The default is "pearson".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-PB is defined as follows. Let

$$r_k(\mathbf{x}) = \frac{\mathbf{PB(k)} - \min_{\mathbf{j}} \mathbf{PB(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{PB(i)} - \min_{\mathbf{j}} \mathbf{PB(j)})}$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original PB(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

G. W. Miligan, "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrika*, 45, 325-342 (1980).

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B2_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B2_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.PB = B_PB.IDX(x = scale(data), kmax=10, method = "kmeans", corr = "pearson", nstart = 100,
```

```
                alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.PB)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_PBM.IDX *BCVI-Pakhira-Bandyopadhyay-Maulik (PBM) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Pakhira-Bandyopadhyay-Maulik (PBM) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_PBM.IDX(x, kmax, method = "FCM", fzm = 2, nstart = 20,
        iter = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-PBM is defined as follows. Let

$$r_k(\mathbf{x}) = \frac{\mathbf{PBM(k)} - \min_{\mathbf{j}} \mathbf{PBM(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{PBM(i)} - \min_{\mathbf{j}} \mathbf{PBM(j)})}$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original PBM(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," Pattern recognition, vol. 37, no. 3, pp. 487–501, 2004.

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.PBM = B_PBM.IDX(x = scale(data), kmax =10, method = "FCM", fzm = 2, nstart = 20,
                iter = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.PBM)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

`B_SF.IDX` *BCVI-The score function*

Description

Compute Bayesian cluster validity index (BCVI) from two to `kmax` groups using the score function (SF) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_SF.IDX(x, kmax, method = "kmeans", nstart = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

`x`	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
`kmax`	a maximum number of clusters to be considered.
`method`	a character string indicating which clustering method to be used (`"kmeans"`, `"hclust_complete"`, `"hclust_average"`, `"hclust_single"`). The default is `"kmeans"`.
`nstart`	a maximum number of initial random sets for kmeans for `method = "kmeans"`. The default is `100`.
`alpha`	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter `mult.alpha` to be its multiplier. The default is `"default"`.
`mult.alpha`	the power s from n^s to be multiplied to the Dirichlet prior parameters `alpha` (selecting `mult.alpha` in `[0,1)` is recommended). The default is $\frac{1}{2}$.

Details

BCVI-SF is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{SF(j)} - \mathbf{SF(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{SF(j)} - \mathbf{SF(i)})}.$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original SF(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

S. Saitta, B. Raphael, I. Smith, "A bounded index for cluster validity," *In Perner, P.: Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science*, 4571, Springer (2007).

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B2_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B2_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.SF = B_SF.IDX(x = scale(data), kmax=10, method = "kmeans",
                nstart = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.SF)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_STRPBM.IDX *BCVI-Starczewski and Pakhira-Bandyopadhyay-Maulik for crisp clustering indexes*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Starczewski (STR) and/or Pakhira-Bandyopadhyay-Maulik (PBM) as the underlying cluster validity index (CVI) and Dirichlet prior parameters of the user's choice. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_STRPBM.IDX(x, kmax, method = "kmeans", indexlist = "all",
          nstart = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".

`indexlist`	a character string indicating which cluster validity indexes to be computed (`"all"`, `"STR"`, `"PBM"`). More than one indexes can be selected.
`nstart`	a maximum number of initial random sets for kmeans for `method = "kmeans"`. The default is `100`.
`alpha`	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter `mult.alpha` to be its multiplier. The default is `"default"`.)
`mult.alpha`	the power s from n^s to be multiplied to the Dirichlet prior parameters `alpha` (selecting `mult.alpha` in `[0,1)` is recommended). The default is $\frac{1}{2}$.

Details

BCVI-STRPBM is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\mathbf{CVI(k)} - \min_{\mathbf{j}} \mathbf{CVI(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{CVI(i)} - \min_{\mathbf{j}} \mathbf{CVI(j)})}$$

where CVI is either STR or PBM index.
Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

`BCVI`	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
`VAR`	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
`CVI`	the data frame where the first and the second columns are the number of groups k and the original STR(k) or PBM(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

M. K. Pakhira, S. Bandyopadhyay and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recogn* 37(3):487–501 (2004).

A. Starczewski, "A new validity index for crisp clusters," *Pattern Anal Applic* 20, 687–700 (2017).

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B2_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B2_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.STRPBM = B_STRPBM.IDX(x = scale(data), kmax=10, method = "kmeans",
            indexlist = "all", nstart = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI-STR

pplot = plot_BCVI(B.STRPBM$STR)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot

# plot the BCVI-PBM

pplot = plot_BCVI(B.STRPBM$PBM)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_TANG.IDX *BCVI-Tang index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Tang as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_TANG.IDX(x, kmax, method = "FCM", fzm = 2, nstart = 20,
          iter = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-TANG is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{TANG(j)} - \mathbf{TANG(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{TANG(j)} - \mathbf{TANG(i)})}.$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original TANG(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

Y. Tang, F. Sun, and Z. Sun, "Improved validation index for fuzzy clustering," in Proceedings of the 2005, American Control Conference, 2005., pp. 1120–1125 vol. 2, 2005. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1470111&isnumber=31519

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_DI.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.TANG = B_TANG.IDX(x = scale(data), kmax =10, method = "FCM", fzm = 2,
                  nstart = 20, iter = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.TANG)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_WL.IDX *BCVI-Wu and Li (WL) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Wu and Li (WL) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_WL.IDX(x, kmax, method = "FCM", fzm = 2, nstart = 20,
        iter = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-WL is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{WL(j)} - \mathbf{WL(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{WL(j)} - \mathbf{WL(i)})}.$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original WL(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

C. H. Wu, C. S. Ouyang, L. W. Chen, and L. W. Lu, "A new fuzzy clustering validity index with a median factor for centroid-based clustering," IEEE Transactions on Fuzzy Systems, vol. 23, no. 3, pp. 701–718, 2015.https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6811211&isnumber=7115244

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_TANG.IDX, B_WP.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)
```

```
B.WL = B_WL.IDX(x = scale(data), kmax =10, method = "FCM", fzm = 2,
              nstart = 20, iter = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.WL)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

`B_WP.IDX` *BCVI-Wiroonsri and Preedasawakul (WP) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to `kmax` groups using Wiroonsri and Preedasawakul (WP) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_WP.IDX(x, kmax, corr = "pearson", method = "FCM", fzm = 2,
        gamma = (fzm^2 * 7)/4, sampling = 1, iter = 100, nstart = 20,
        NCstart = TRUE, alpha = "default", mult.alpha = 1/2)
```

Arguments

`x`	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
`kmax`	a maximum number of clusters to be considered.
`corr`	a character string indicating which correlation coefficient is to be computed (`"pearson"`, `"kendall"` or `"spearman"`). The default is `"pearson"`.
`method`	a character string indicating which clustering method to be used (`"FCM"` or `"EM"`). The default is `"FCM"`.
`fzm`	a number greater than 1 giving the degree of fuzzification for `method = "FCM"`. The default is `2`.
`gamma`	adjusted fuzziness parameter for `indexlist = ("WP", "WPC", "WPCI1", "WPCI2")`. The default is computed from $7fzm^2/4$.
`sampling`	a number greater than 0 and less than or equal to 1 indicating the undersampling proportion of data to be used. This argument is intended for handling a large dataset. The default is `1`.
`iter`	a maximum number of iterations for `method = "FCM"`. The default is `100`.
`nstart`	a maximum number of initial random sets for FCM for `method = "FCM"`. The default is `20`.

NCstart	logical for indexlist = ("WP", "WPC", "WPCI1","WPCI2"), if TRUE, the WP correlation at c=1 is defined as an adjusted sd of the distances between all data points and their mean. Otherwise, the WP correlation at c=1 is defined as 0.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-WP is defined as follows. Let

$$r_k(\mathbf{x}) = \frac{\mathbf{WP(k)} - \min_{\mathbf{j}} \mathbf{WP(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{WP(i)} - \min_{\mathbf{j}} \mathbf{WP(j)})}$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original WP(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, "A correlation-based fuzzy cluster validity index with secondary options detector," arXiv:2308.14785, 2023

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B7_data, B_TANG.IDX, B_XB.IDX, B_Wvalid, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(20,20,20,5,5,5,0.5,0.5,0.5)

B.WP = B_WP.IDX(x = scale(data), kmax =10, corr = "pearson", method = "FCM",
                fzm = 2, sampling = 1, iter = 100, nstart = 20, NCstart = TRUE,
                alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.WP)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_Wvalid *BCVI-Wiroonsri (WI) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Wiroonsri (WI) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_Wvalid(x, kmax, method = "kmeans", corr = "pearson", nstart = 100,
      sampling = 1, NCstart = TRUE, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("kmeans", "hclust_complete", "hclust_average", "hclust_single"). The default is "kmeans".
corr	a character string indicating which correlation coefficient is to be computed ("pearson", "kendall" or "spearman"). The default is "pearson".
nstart	a maximum number of initial random sets for kmeans for method = "kmeans". The default is 100.
sampling	a number greater than 0 and less than or equal to 1 indicating the undersampling proportion of data to be used. This argument is intended for handling a large dataset. The default is 1.
NCstart	logical for indexlist includes the "NC", "NCI", "NCI1", and "NCI2"), if TRUE, the NC correlation at k=1 is defined as the ratio introduced in the reference. Otherwise, it is assigned as 0.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-WI is defined as follows. Let

$$r_k(\mathbf{x}) = \frac{\mathbf{WI(k)} - \min_{\mathbf{j}} \mathbf{WI(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{WI(i)} - \min_{\mathbf{j}} \mathbf{WI(j)})}$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

BCVI	the dataframe where the first and the second columns are the number of groups k and BCVI(k), respectively, for k from 2 to kmax.
VAR	the data frame where the first and the second columns are the number of groups k and the variance of p_k, respectively, for k from 2 to kmax.
CVI	the data frame where the first and the second columns are the number of groups k and the original WI(k), respectively, for k from 2 to kmax.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, "Clustering performance analysis using a new correlation based cluster validity index," Pattern Recognition, 145, 109910, 2024. doi:10.1016/j.patcog.2023.109910

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

B2_data, B_TANG.IDX, B_WP.IDX, B_STRPBM.IDX, B_DB.IDX

Examples

```
library(BayesCVI)

# The data included in this package.
data = B2_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)

B.WI = B_Wvalid(x = scale(data), kmax = 10, method = "kmeans", corr = "pearson",
              nstart = 100, sampling = 1, NCstart = TRUE, alpha = aalpha,
              mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.WI)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

B_XB.IDX *BCVI-Xie and Beni (XB) index*

Description

Compute Bayesian cluster validity index (BCVI) from two to kmax groups using Xie and Beni (XB) as the underling cluster validity index (CVI) with the user's selected Dirichlet prior parameters. The full detail of BCVI can be found in the paper Wiroonsri and Preedasawakul (2024).

Usage

```
B_XB.IDX(x, kmax, method = "FCM", fzm = 2, nstart = 20,
        iter = 100, alpha = "default", mult.alpha = 1/2)
```

Arguments

x	a numeric data frame or matrix where each column is a variable to be used for cluster analysis and each row is a data point.
kmax	a maximum number of clusters to be considered.
method	a character string indicating which clustering method to be used ("FCM" or "EM"). The default is "FCM".
fzm	a number greater than 1 giving the degree of fuzzification for method = "FCM". The default is 2.
nstart	a maximum number of initial random sets for FCM for method = "FCM". The default is 20.
iter	a maximum number of iterations for method = "FCM". The default is 100.
alpha	Dirichlet prior parameters $\alpha_2, ..., \alpha_k$ where α_k is the parameter corresponding to "the probability of having k groups" (selecting each α_k between 0 to 30 is recommended and using the other parameter mult.alpha to be its multiplier. The default is "default".)
mult.alpha	the power s from n^s to be multiplied to the Dirichlet prior parameters alpha (selecting mult.alpha in [0,1) is recommended). The default is $\frac{1}{2}$.

Details

BCVI-XB is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{XB(j)} - \mathbf{XB(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{XB(j)} - \mathbf{XB(i)})}.$$

Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

`BCVI`	the dataframe where the first and the second columns are the number of groups `k` and BCVI(k), respectively, for `k` from `2` to `kmax`.
`VAR`	the data frame where the first and the second columns are the number of groups `k` and the variance of p_k, respectively, for `k` from `2` to `kmax`.
`CVI`	the data frame where the first and the second columns are the number of groups `k` and the original XB(k), respectively, for `k` from `2` to `kmax`.

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

X. Xie and G. Beni, “A validity measure for fuzzy clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 8, pp. 841–847, 1991.

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

`B7_data`, `B_TANG.IDX`, `B_WP.IDX`, `B_Wvalid`, `B_DB.IDX`

Examples

```
library(BayesCVI)

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)
```

```
B.XB = B_XB.IDX(x = scale(data), kmax =10, method = "FCM",
              fzm = 2, nstart = 20, iter = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.XB)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

---

plot_BCVI *Plots for visualizing BCVI*

Description

Plot Bayesian cluster validity index (BCVI) with and without standard deviation error bars and the underlying index.

Usage

```
plot_BCVI(B.result, mult.err.bar = 2)
```

Arguments

B.result	a result from one of the functions B_XB.IDX, B_Wvalid, B_WP.IDX, B_WL.IDX, B_TANG.IDX, B_STRPBM.IDX, B_SF.IDX, B_PBM.IDX, B_PB.IDX, B_KWON.IDX, B_KWON2.IDX, B_KPBM.IDX, B_HF.IDX, B_GC.IDX, B_DI.IDX, B_DB.IDX, B_CSL.IDX, B_CH.IDX, B_CCV.IDX and B_BayesCVIs.IDX
mult.err.bar	a multiplier of the stadard deviations to be used for plotting error bars

Details

BCVI is defined as follows.

Let

$$r_k(\mathbf{x}) = \frac{\max_{\mathbf{j}} \mathbf{CVI(j)} - \mathbf{CVI(k)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\max_{\mathbf{j}} \mathbf{CVI(j)} - \mathbf{CVI(i)})}$$

for a cluster validity index (CVI) such that the smallest value indicates the optimal number of clusters and

$$r_k(\mathbf{x}) = \frac{\mathbf{CVI(k)} - \min_{\mathbf{j}} \mathbf{CVI(j)}}{\sum_{\mathbf{i=2}}^{\mathbf{K}} (\mathbf{CVI(i)} - \min_{\mathbf{j}} \mathbf{CVI(j)})}$$

for a CVI such that the largest indicates the optimal number of clusters. Assume that

$$f(\mathbf{x}|\mathbf{p}) = C(\mathbf{p}) \prod_{k=2}^{K} p_k^{nr_k(x)}$$

represents the conditional probability density function of the dataset given $\mathbf{p}$, where $C(\mathbf{p})$ is the normalizing constant. Assume further that $\mathbf{p}$ follows a Dirichlet prior distribution with parameters $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_K)$. The posterior distribution of $\mathbf{p}$ still remains a Dirichlet distribution with parameters $(\alpha_2 + nr_2(\mathbf{x}), \ldots, \alpha_K + nr_K(\mathbf{x}))$.

The BCVI is then defined as

$$BCVI(k) = E[p_k|\mathbf{x}] = \frac{\alpha_k + nr_k(\mathbf{x})}{\alpha_0 + n}$$

where $\alpha_0 = \sum_{k=2}^{K} \alpha_k$.

The variance of p_k can be computed as

$$Var(p_k|\mathbf{x}) = \frac{(\alpha_k + nr_k(x))(\alpha_0 + n - \alpha_k - nr_k(x))}{(\alpha_0 + n)^2(\alpha_0 + n + 1)}.$$

Value

`plot_index`	a plot of the underlying index for the number of groups from 2 to $kmax$ according to `B.result`
`plot_BCVI`	a plot of BCVI for the number of groups from 2 to $kmax$ according to `B.result`
`error_bar_plot`	a plot of BCVI with error bars for the number of groups from 2 to $kmax$ according to `B.result`

Author(s)

Nathakhun Wiroonsri and Onthada Preedasawakul

References

N. Wiroonsri, O. Preedasawakul, "A Bayesian cluster validity index", arXiv:2402.02162, 2024.

See Also

`B_STRPBM.IDX`, `B_TANG.IDX`, `B_XB.IDX`, `B_Wvalid`, `B_WP.IDX`, `B_DB.IDX`

Examples

```
library(BayesCVI)
library(UniversalCVI)

##Soft clustering

# The data included in this package.
data = B7_data[,1:2]

# alpha
aalpha = c(5,5,5,20,20,20,0.5,0.5,0.5)
```

```
B.XB = B_XB.IDX(x = scale(data), kmax =10, method = "FCM", fzm = 2,
              nstart = 20, iter = 100, alpha = aalpha, mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.XB)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot


## Hard clustering

# The data included in this package.
data = B2_data[,1:2]

K.STR = STRPBM.IDX(scale(data), kmax = 10, kmin = 2, method = "kmeans",
  indexlist = "STR", nstart = 100)

# WP.IDX values
result = K.STR$STR$STR


aalpha = c(20,20,20,5,5,5,0.5,0.5,0.5)
B.STR = BayesCVIs(CVI = result,
          n = nrow(data),
          kmax = 10,
          opt.pt = "max",
          alpha = aalpha,
          mult.alpha = 1/2)

# plot the BCVI

pplot = plot_BCVI(B.STR)
pplot$plot_index
pplot$plot_BCVI
pplot$error_bar_plot
```

Index