# Package: vscc (via r-universe)

September 12, 2024

**Type** Package

**Title** Variable Selection for Clustering and Classification

**Version** 0.7

**Date** 2023-10-17

**Author** Jeffrey L. Andrews [aut], Mackenzie R. Neal [aut], Paul D. McNicholas [aut, cre] (<https://orcid.org/0000-0002-2482-523X>)

**Maintainer** Paul D. McNicholas <mcnicholas@math.mcmaster.ca>

**Description** Performs variable selection/feature reduction under a clustering or classification framework. In particular, it can be used in an automated fashion using mixture model-based methods ('teigen' and 'mclust' are currently supported). Can account for mixtures of non-Gaussian distributions via Manly transform (via 'ManlyMix'). See Andrews and McNicholas (2014) <doi:10.1007/s00357-013-9139-2> and Neal and McNicholas (2023) <doi:10.48550/arXiv.2305.16464>.

**License** GPL (>= 2)

**Imports** teigen, mclust, MixGHD

**Depends** ManlyMix

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2023-10-17 22:20:02 UTC

# Contents

---

**vscc-package**                    *Variable Selection for Clustering and Classification*

---

### Description

Performs variable selection under a clustering or classification framework. Automated implementation using model-based clustering is based on `teigen` and `mclust`.

### Details

|           |             |
|-----------|-------------|
| Package:  | vscc        |
| Type:     | Package     |
| Version:  | 0.7         |
| Date:     | 2023-10-17  |
| License:  | GPL>="2"    |

### Author(s)

Jeffrey L. Andrews, Mackenzie Neal, Paul D. McNicholas

Maintainer: Paul D. McNicholas <mcnicholas@math.mcmaster.ca>

### References

See `citation("vscc")`.

### See Also

[vscc](#)

---

  **plot.vscc**                              *Plotting for VSCC Objects*

---

### Description

Dedicated plot function for objects of class vscc.

### Usage

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

## Arguments

| | |
|---|---|
| x | An object of class vscc. |
| ... | Further arguments to be passed on |

## Details

Provides a scatterplot matrix of the selected variables with colours corresponding to each group.

## Value

No return value.

## Author(s)

Jeffrey L. Andrews

## See Also

[vscc](#)

## Examples

```
require("mclust")
data(banknote)
X<-banknote[,-1]
bankrun <- vscc(X)
plot(bankrun)
```

---

| print.vscc | *Printing for VSCC* |
|---|---|

---

## Description

Dedicated print function for objects of class vscc.

## Usage

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

## Arguments

| | |
|---|---|
| x | An object of class vscc |
| ... | Further arguments to be passed on |

## Details

Same as summary.

## Value

No return value.

## Author(s)

Jeffrey L. Andrews

## See Also

[summary.vscc](), [vscc]()

## Examples

```
require("mclust")
data(banknote)
X<-banknote[,-1]
vscc(X)
```

---

summary.vscc          *Summary for VSCC Objects*

---

## Description

Dedicated summary function for objects of class vscc

## Usage

```
## S3 method for class 'vscc'
summary(object, ...)
```

## Arguments

| | |
|---|---|
| object | An object of class vscc |
| ... | Additional arguments to be passed |

## Value

No return value.

## Author(s)

Jeffrey L. Andrews

## See Also

[vscc]()

## Examples

```
require("mclust")
data(banknote)
summary(vscc(banknote[,-1]))
```

---

vscc                     *Variable Selection for Clustering and Classification*

---

## Description

Performs variable selection under a clustering or classification framework. Automated implementation using model-based clustering is based on teigen version 2.0 and mclust version 4.0; issues *may* arise when using different versions.

## Usage

```
vscc(x, G=1:9, automate = "mclust", initial = NULL, initunc=NULL, train = NULL,
    forcereduction = FALSE)
```

## Arguments

| | |
|---|---|
| x | Data frame or matrix to perform variable selection on |
| G | Vector for the number of groups to consider during initialization and/or post-selection analysis. Default is 1-9. |
| automate | Character string ("teigen", "mclust" (default), or NULL only) indicating which mixture model family to implement as initialization and/or post-selection analysis. If NULL, the function assumes manual operation of the algorithm (meaning an initial clustering vector must be given, and no post-selection analysis is performed). |
| initial | Optional vector giving the initial clustering. |
| initunc | Optional scalar indicating the total uncertainty of the initial clustering solution. Only used when initial is non-null. |
| train | Optional vector of training data (for classification framework). |
| forcereduction | Logical indicating if the full data set should be considered (FALSE) when selecting the 'best' variable subset via total model uncertainty. Not used if automate=NULL. |

## Value

| | |
|---|---|
| selected | A list containing the subsets of variables selected for each relation. Each set is numbered according to the number in the exponential of the relationship. For instance, vscc_object$selected[[3]] corresponds to the variable subset selected by the cubic relationship. |
| family | The family used as initialization and/or post selection. (Same as user input automate, and can be NULL). |

| | |
|---|---|
| wss | The within-group variance associated with each variable from the full data set. |

The remaining values are provided as long as `automate` is not `NULL`:

| | |
|---|---|
| topselected | The best variable subset according to the total model uncertainty. |
| initialrun | Results from the initialization; an object of class `teigen` or `mclust`. |
| bestmodel | Results from the best model on the selected variable subset; an object of class `teigen` or `mclust`. |
| chosenrelation | Numeric indication of the relationship chosen according to total model uncertainty. The number corresponds to exponent in the relationship: for instance, a value of '4' suggests the quartic relationship. If the value `"Full dataset"` is given, then the unreduced data provides the best model uncertainty; can be avoided by specifying `forcereduction=TRUE` in the function call. |
| uncertainty | Total model uncertainty associated with the best relationship. |
| allmodelfit | List containing the results (`teigen` or `mclust` objects) from the post-selection analysis on each variable subset. Number corresponds to the exponent in the relationship. For instance, `vscc_object$allmodelfit[[1]]` gives the results from the analysis on the variables selected by the linear relationship. |

**Author(s)**

Jeffrey L. Andrews, Paul D. McNicholas

**References**

See `citation("vscc")` for the variable selection references. See also `citation("teigen")` and `citation("mclust")` if using those families of models via the `automate` call.

**See Also**

[teigen](#), [Mclust](#)

**Examples**

```
require("mclust")
data(banknote)
head(banknote)
bankrun <- vscc(banknote[,-1])
head(bankrun$topselected) #Show preview of selected variables
table(banknote[,1], bankrun$initialrun$classification) #Clustering results on full data set
table(banknote[,1], bankrun$bestmodel$classification) #Clustering results on reduced data set
```

---

| | |
|---|---|
| vsccmanly | *Variable Selection for Skewed Clustering and Classification* |

---

## Description

Performs variable selection under a clustering framework. Accounts for mixtures of non-Gaussian distributions via the ManlyTransform (via 'ManlyMix').

## Usage

```
vsccmanly(x, G=2:9, numstart=100, selection="backward",forcereduction=FALSE,
                  initstart="k-means", seedval=2354)
```

## Arguments

| | |
|---|---|
| x | Data frame or matrix to perform variable selection on |
| G | Vector for the number of groups to consider during initialization and/or post-selection analysis. Default is 2-9. |
| numstart | Number of random starts. |
| selection | Forward or backward transformation parameter selection. User may also choose to fit a full Manly mixture (options are 'forward', 'backward', or 'none'). |
| forcereduction | Logical indicating if the full data set should be considered (FALSE) when selecting the 'best' variable subset via total model uncertainty. |
| initstart | Method for initial starting values (options are 'k-means' or 'hierarchical'). |
| seedval | Value of seed, used for k-means initialization. |

## Value

| | |
|---|---|
| selected | A list containing the subsets of variables selected for each relation. Each set is numbered according to the number in the exponential of the relationship. For instance, vscc_object$selected[[3]] corresponds to the variable subset selected by the cubic relationship. |
| wss | The within-group variance associated with each variable from the full data set. |
| topselected | The best variable subset according to the total model uncertainty. |
| initialrun | Results from the initial model, prior to variable selection; an object of class ManlyMix. |
| bestmodel | Results from the best model on the selected variable subset; an object of class ManlyMix. |
| variables | Variables used to fit the final model. |
| chosenrelation | Numeric indication of the relationship chosen according to total model uncertainty. The number corresponds to exponent in the relationship: for instance, a value of '4' suggests the quartic relationship. If the value "Full dataset" is given, then the unreduced data provides the best model uncertainty; can be avoided by specifying forcereduction=TRUE in the function call. |

| uncertainty | Total model uncertainty associated with the best relationship. |
|---|---|
| allmodelfit | List containing the results (ManlyMix objects) from the post-selection analysis on each variable subset. Number corresponds to the exponent in the relationship. For instance, vscc_object$allmodelfit[[1]] gives the results from the analysis on the variables selected by the linear relationship. |

## Author(s)

Jeffrey L. Andrews, Mackenzie R. Neal, Paul D. McNicholas

## References

See citation("vscc") for the variable selection references.

## See Also

[vscc](#)

## Examples

```
## Not run:
data(ais)
X=ais[,3:13]
aisfor=vsccmanly(as.data.frame(scale(X)),G=2:9,selection = "forward", forcereduction = TRUE,
                        initstart = "k-means",seedval=2354)
aisfor$variables #Show selected variables
table(ais[,1], aisfor$bestmodel$id) #Clustering results on reduced data set

## End(Not run)
```

# Index