Association-Based Spectral Clustering for Mixed Data

Intermediate Workshop of the PRIN 2022 project, 5-6 February 2026, Bologna

Alfonso Iodice D'Enza, Cristina Tortora and Francesco Palumbo

Ministero dell'Università e della Ricerca
PRIN Progetti di Ricerca di Rilevante Interesse Nazionale
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

outline

independence-based (IB) vs association-based (AB) dissimilarities

association-based for mixed data

taking into account continuous/categorical interactions

AB spectral clustering

example and future work

(adveRtising)

independence-based (IB) vs association-based (AB) dissimilarities

learning from dissimilarities

some unsupervised learning methods take as input a dissimilarity matrix

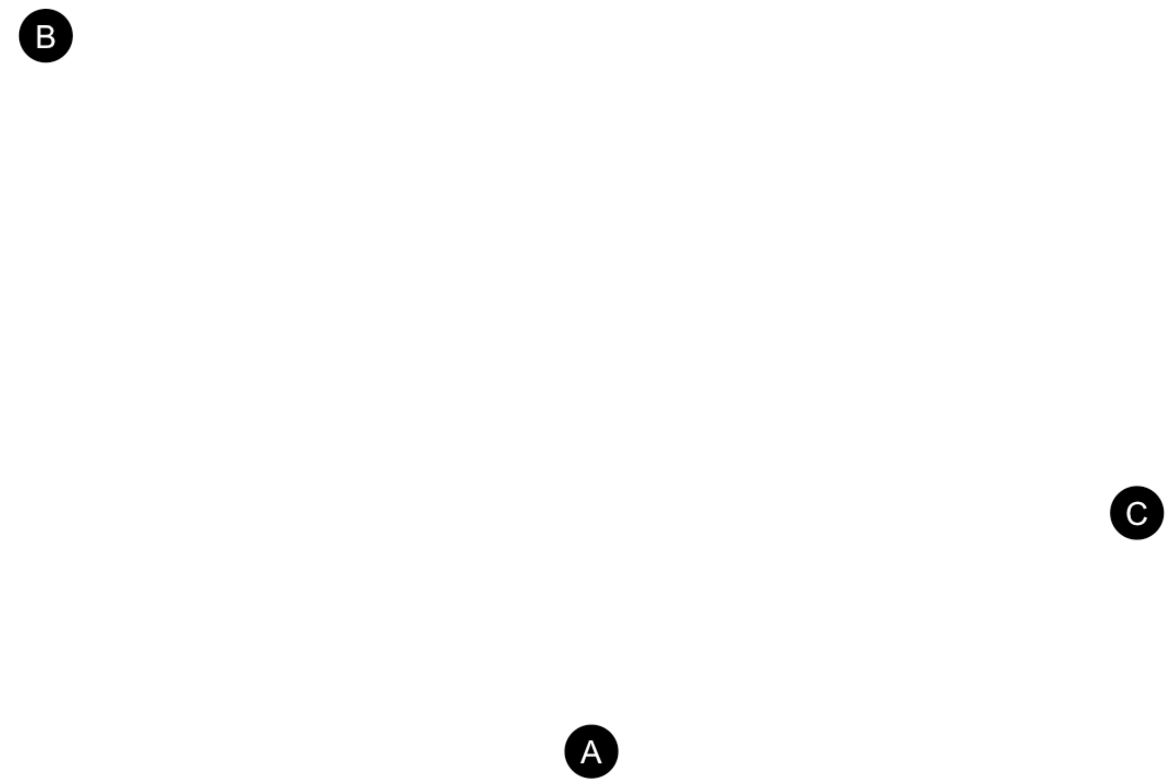
dimension reduction: multidimensional scaling (MDS)[1]

clustering methods: hierarchical (HC) and partitioning around medoids (PAM)[2]

the dissimilarity measure of choice is key, obviously

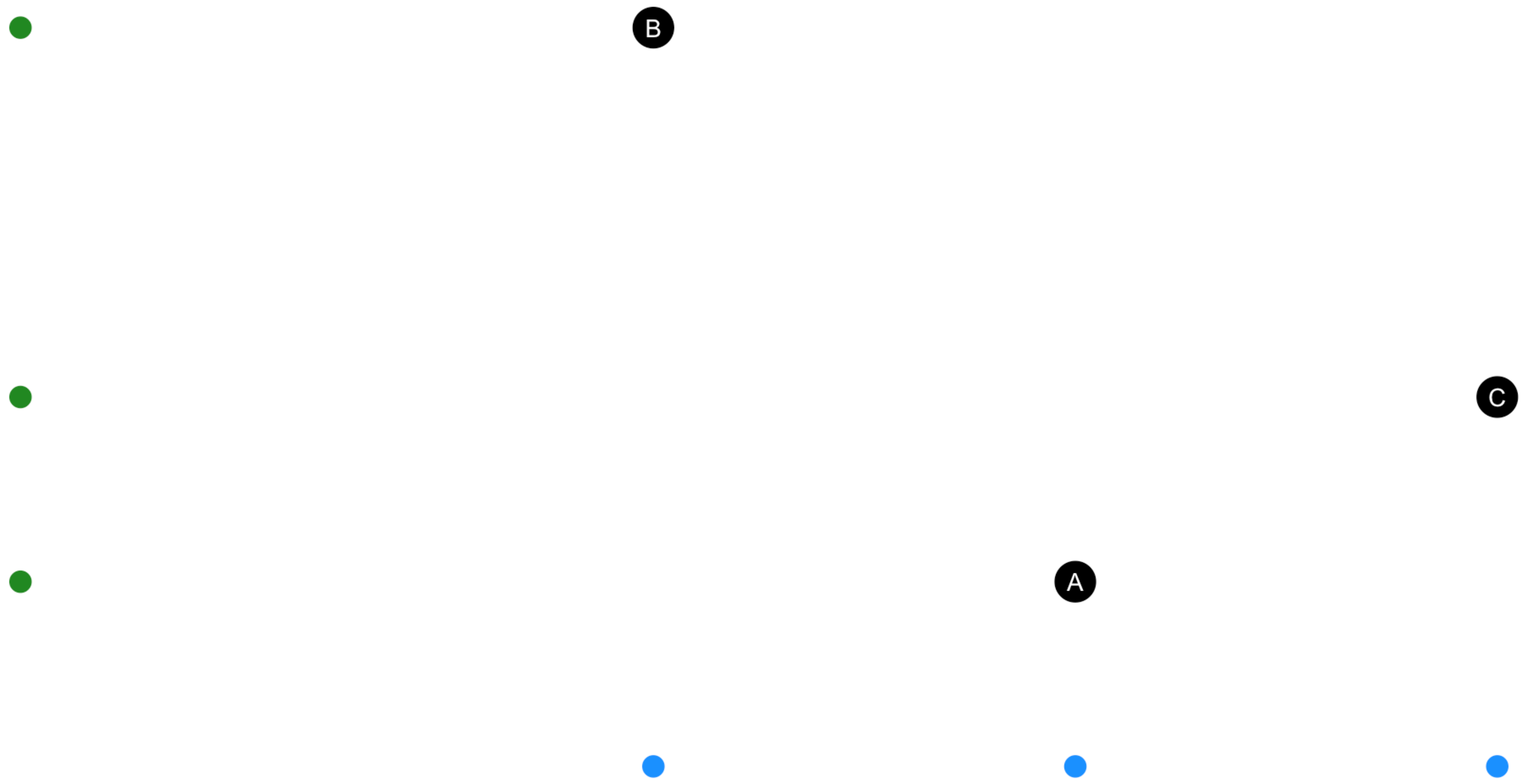
intuition

2 continuous variables: add up by-variable (absolute value or squared) differences



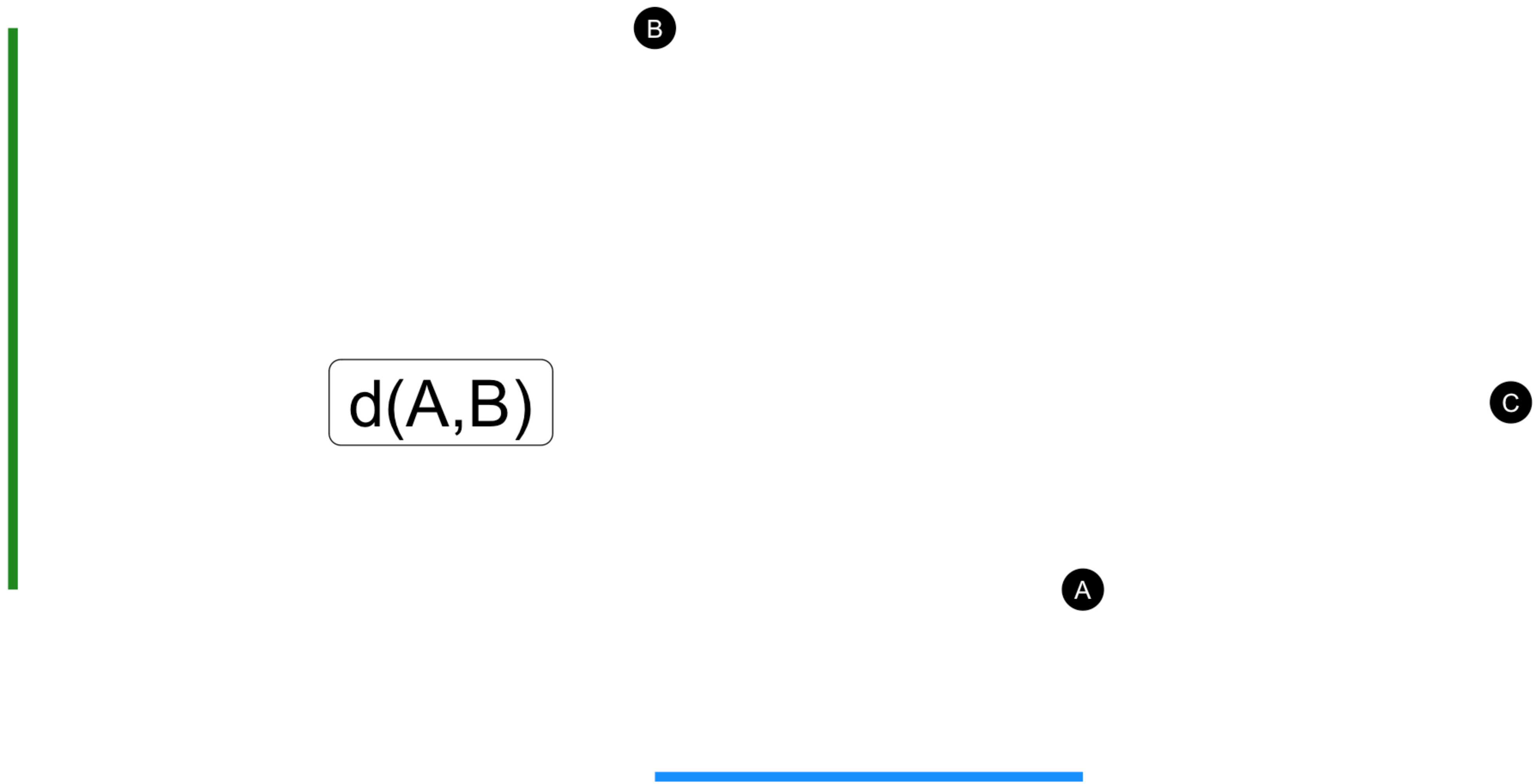
intuition

2 continuous variables: add up by-variable (absolute value or squared) differences



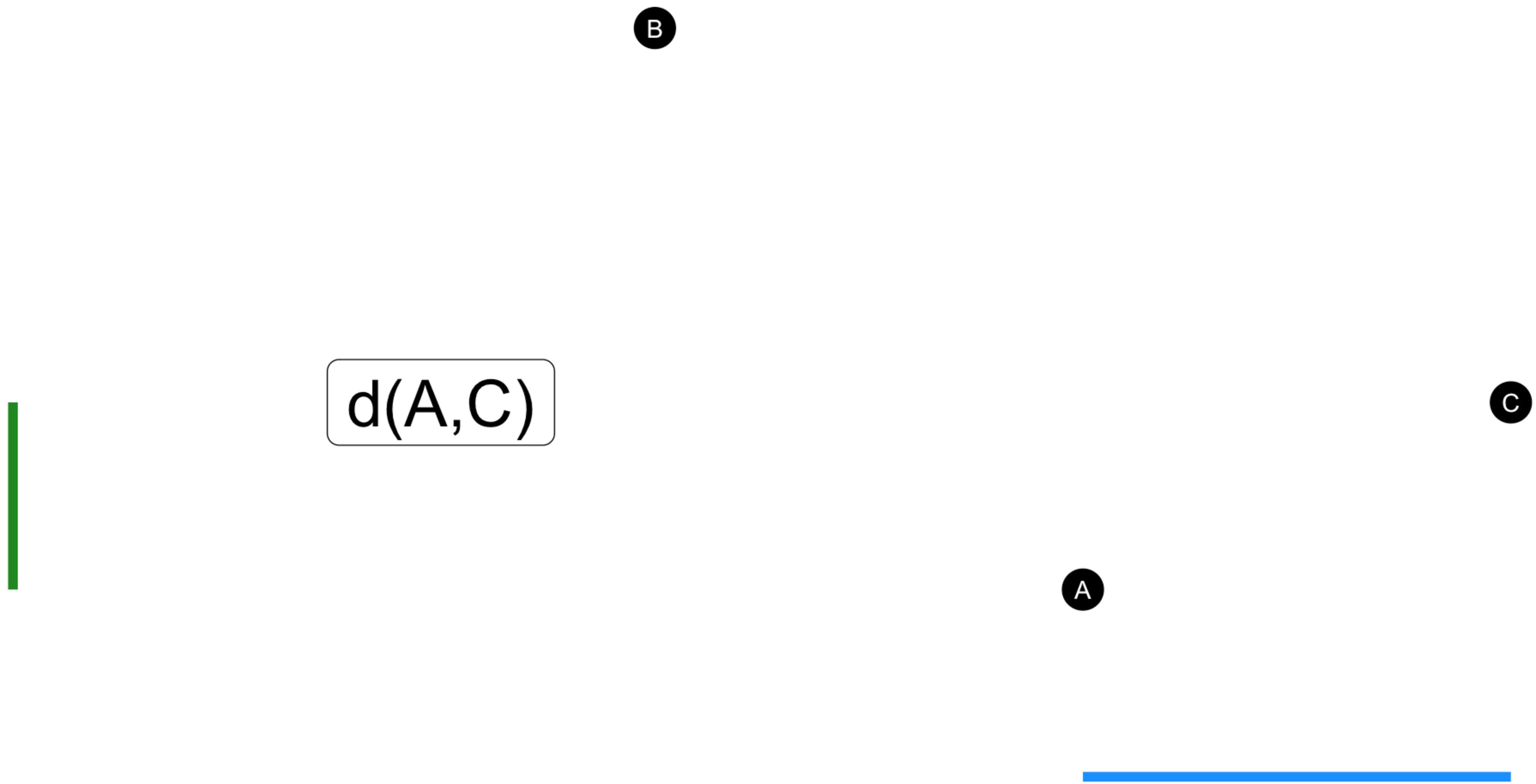
intuition

2 continuous variables: add up by-variable (absolute value or squared) differences



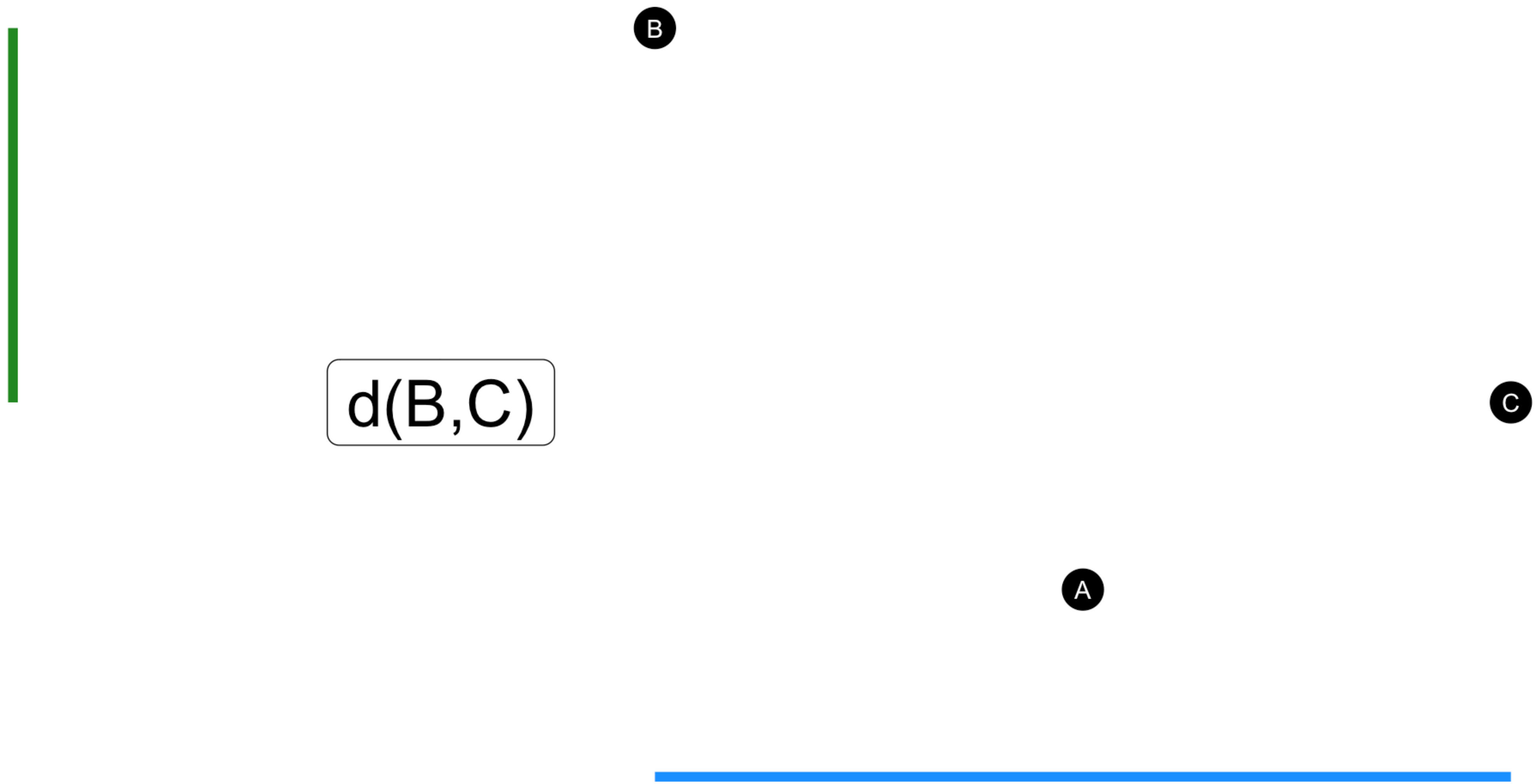
intuition

2 continuous variables: add up by-variable (absolute value or squared) differences



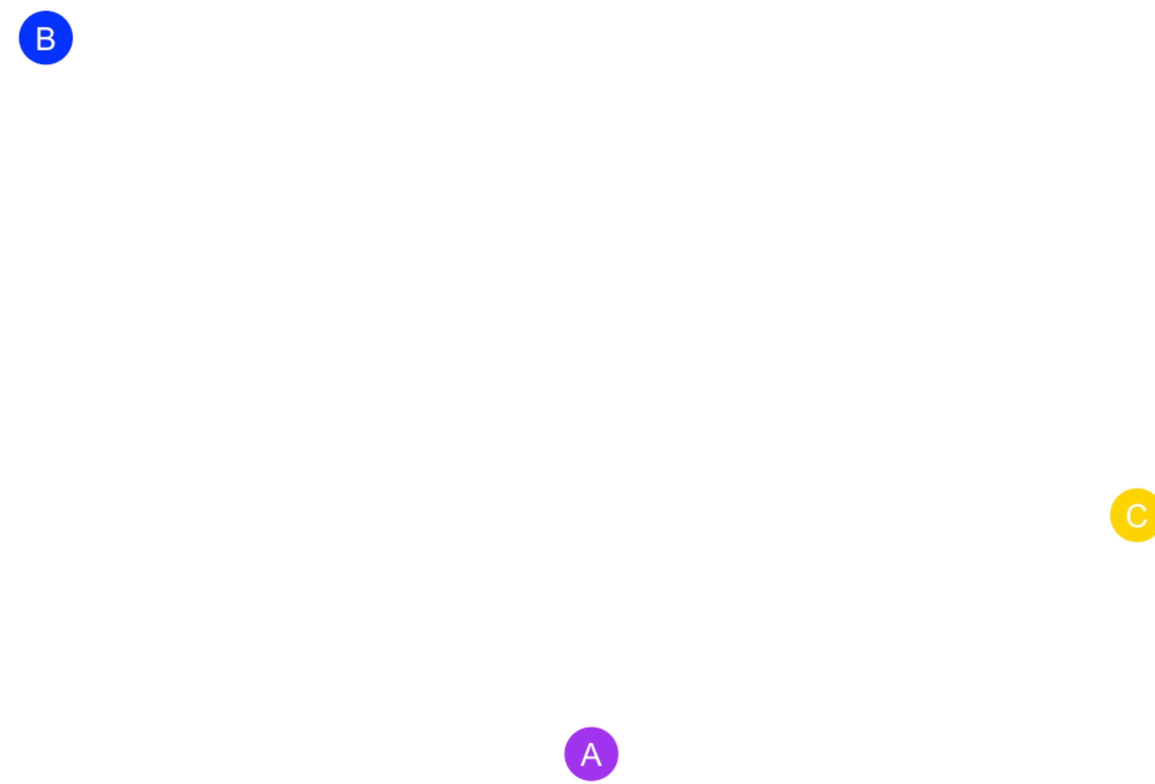
intuition

2 continuous variables: add up by-variable (absolute value or squared) differences



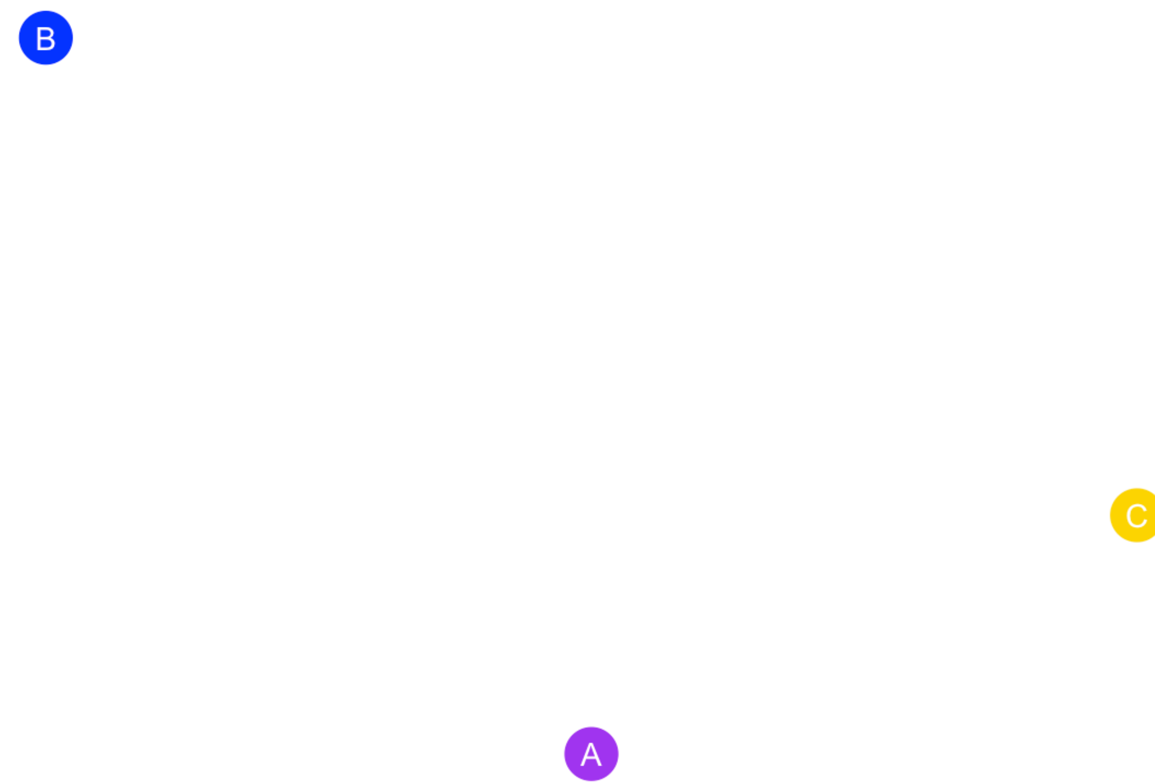
intuition

2 continuous and 1 categorical variables



intuition

one might consider purple and blue *closer* than e.g. purple and yellow



independence-based

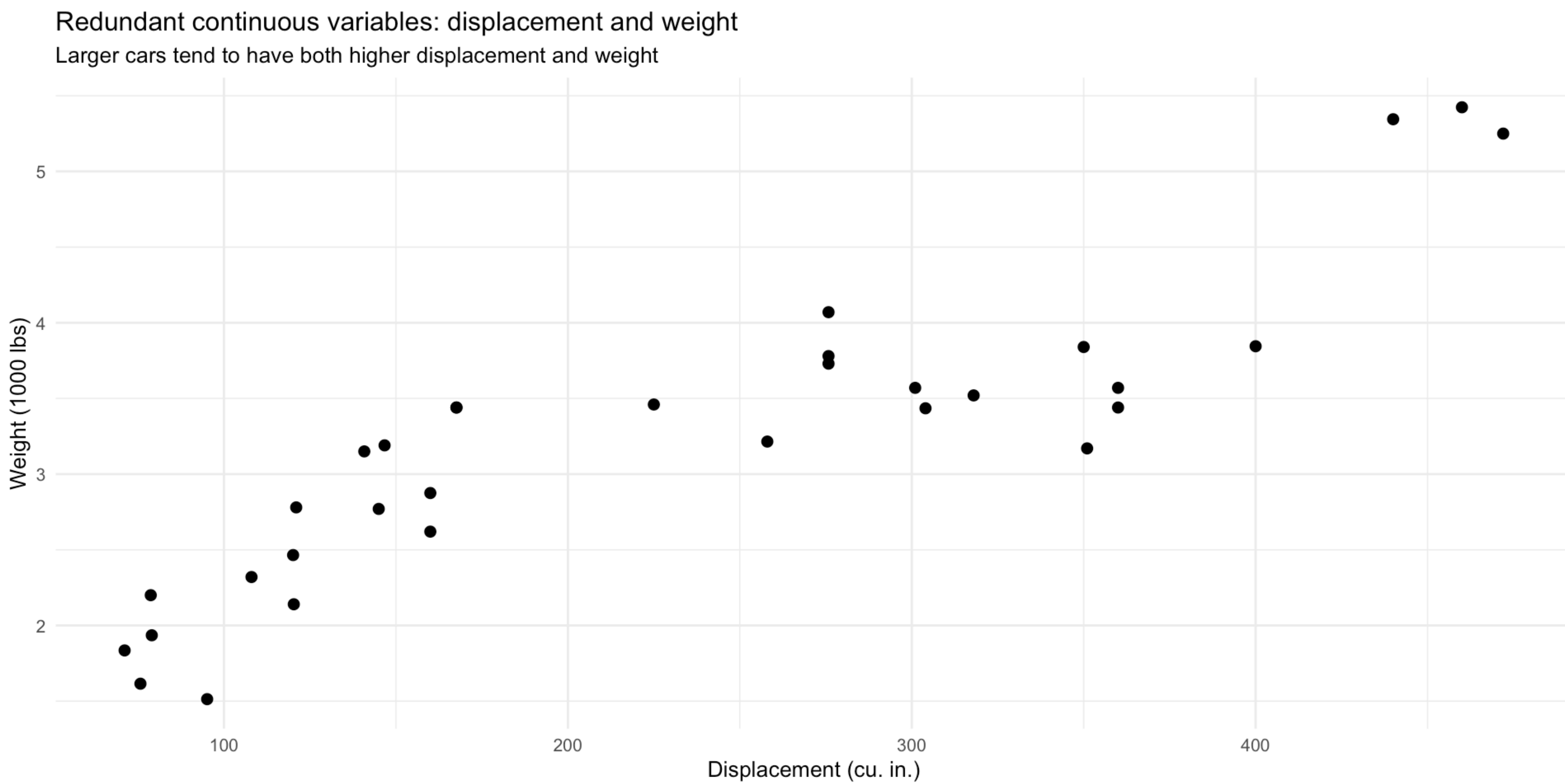
Most commonly used dissimilarity (or, distance) measures are based on by-variable differences that are then added together

- in the continuous case: Euclidean or Manhattan distances
- in the categorical case: Hamming (matching) distance (among MANY others)
- in the mixed data case: Gower dissimilarity index

no inter-variable relations are considered $\rightarrow$ independence-based

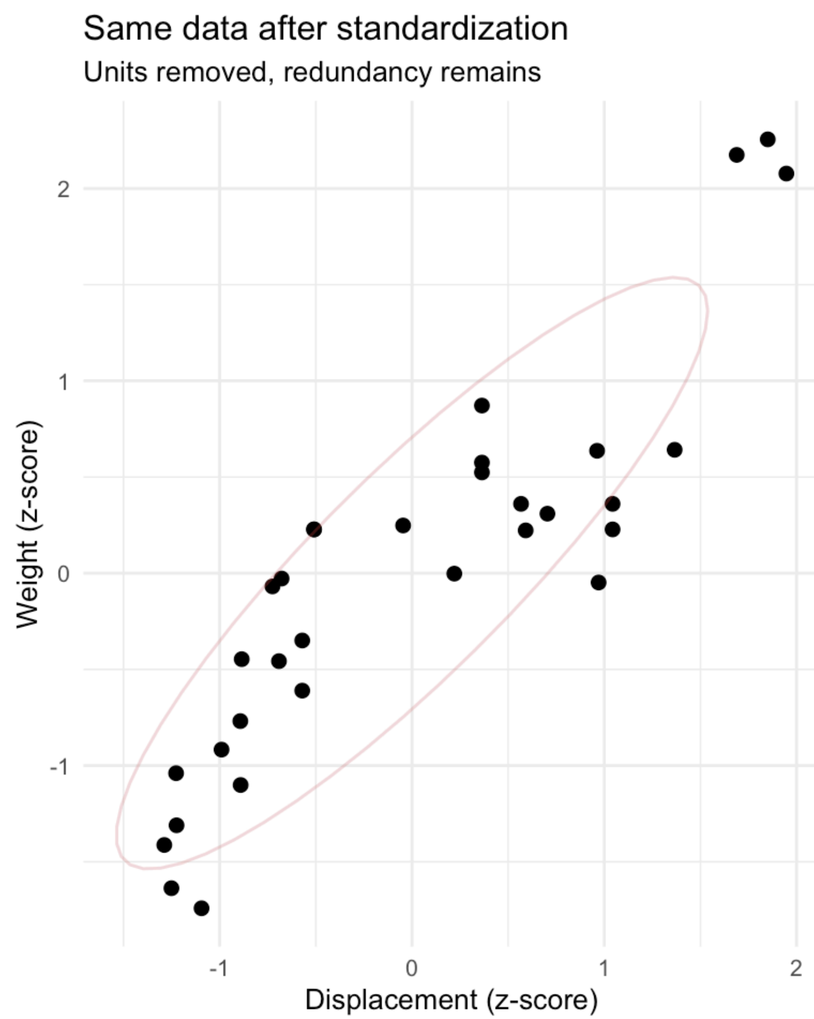
independence-based

- When variables are correlated or associated, shared information is effectively counted multiple times
- inflated dissimilarities may cause potential distortions in downstream unsupervised learning tasks.



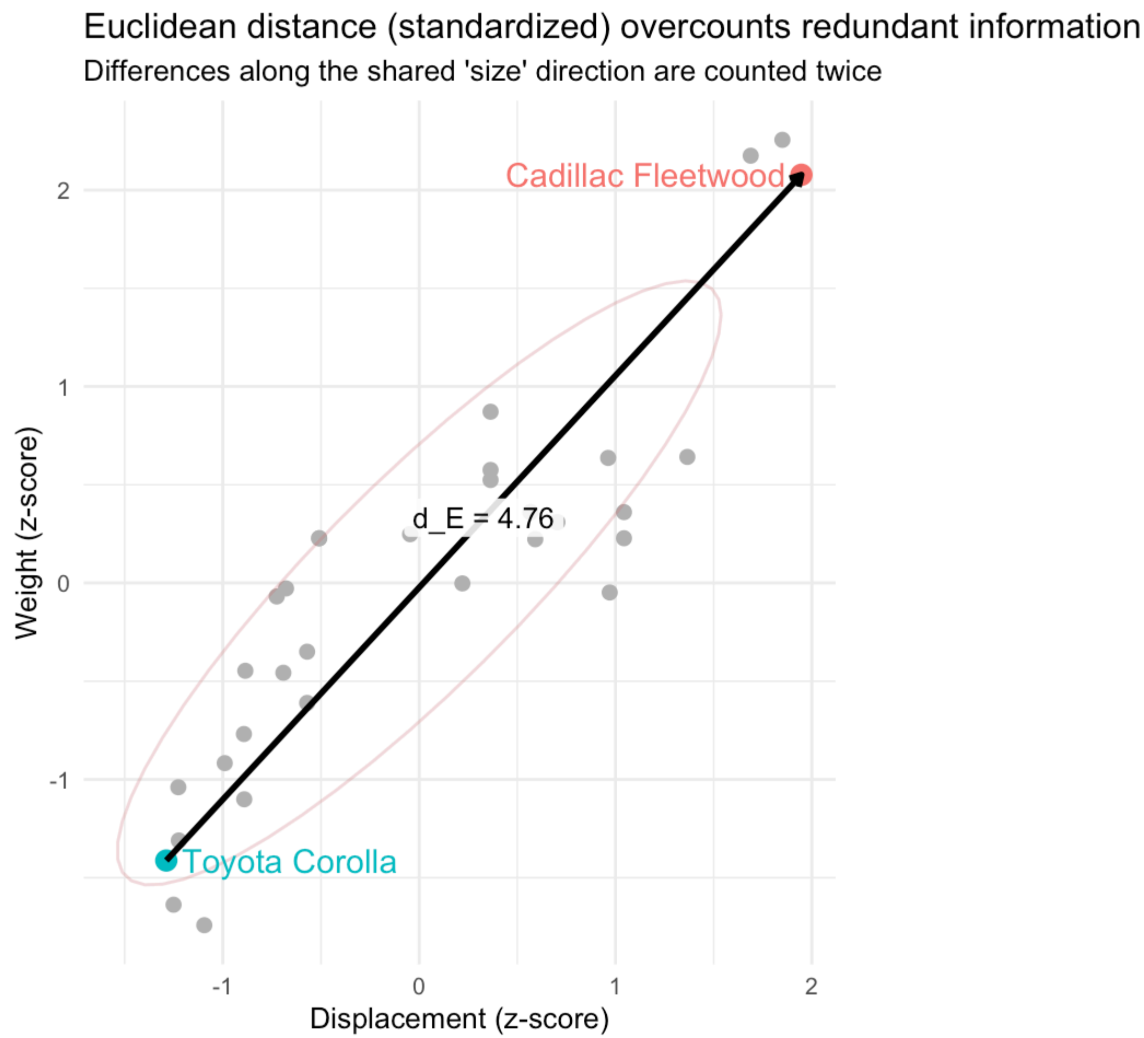
independence-based

- When variables are correlated or associated, shared information is effectively counted multiple times
- inflated dissimilarities may cause potential distortions in downstream unsupervised learning tasks.



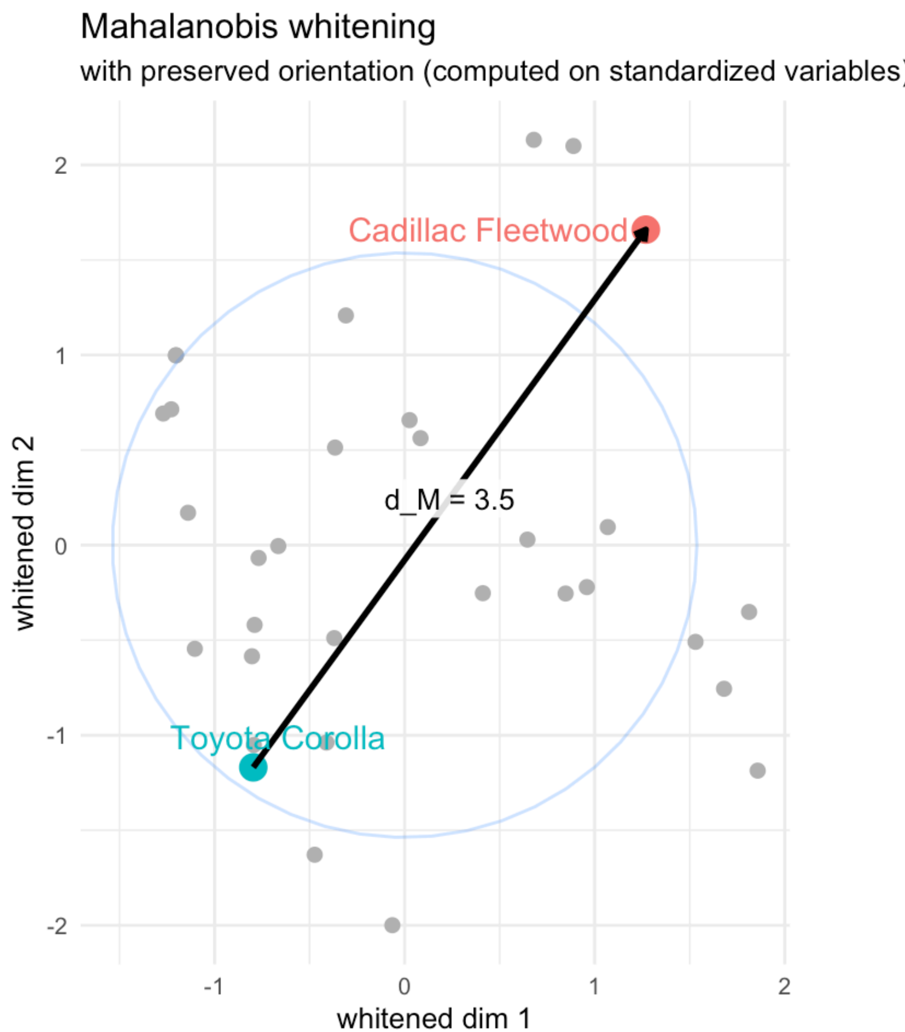
independence-based

The Euclidean distance ⟶ shared information is over-counted



association-based

The Mahalanobis distance $\longrightarrow$ shared information is not over-counted



this is an association-based distance for continuous data

association-based pairwise distance

- differences in line with the inter-variables association/correlation are down-weighted

Association-based for continuous: Mahalanobis distance

Let $\mathbf{X}_{con}$ be $n \times Q_d$ a data matrix of n observations described by Q_d continuous variables, and let $\mathbf{S}$ the sample covariance matrix, the Mahalanobis distance matrix is

$$\mathbf{D}_{mah} = \left[\text{diag}(\mathbf{G})\,\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\ \text{diag}(\mathbf{G})^\mathsf{T} - 2\mathbf{G}\right]^{\odot 1/2}$$

where

- $[\cdot]^{\odot 1/2}$ denotes the element-wise square root
- $\mathbf{G} = (\mathbf{C}\mathbf{X}_{con})\mathbf{S}^{-1}(\mathbf{C}\mathbf{X}_{con})^\mathsf{T}$ is the Mahalanobis Gram matrix
- $\mathbf{C} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ is the centering operator

association-based pairwise distance

- differences in line with the inter-variables association/correlation are down-weighted

Association-based for categorical: total variation distance (TVD)[1]

To distance matrix $\mathbf{D}_{tvd}$ is defined using the so-called delta framework[2] a general way to define categorical data distances.

Let $\mathbf{X}_{cat}$ be $n \times Q_c$ a data matrix of n observations described by Q_c categorical variables.

$$\mathbf{D} = \mathbf{Z}\Delta\mathbf{Z}^{\mathsf{T}} = \begin{bmatrix}\mathbf{Z}_1 & \dots & \mathbf{Z}_{Q_c}\end{bmatrix} \begin{bmatrix}\mathbf{\Delta}_1 & & \\ & \ddots & \\ & & \mathbf{\Delta}_{Q_c}\end{bmatrix} \begin{bmatrix}\mathbf{Z}_1^{\mathsf{T}} \\ \vdots \\ \mathbf{Z}_{Q_c}^{\mathsf{T}}\end{bmatrix}$$

- where $\mathbf{Z} = [\mathbf{Z}_1, \dots, \mathbf{Z}_{Q_c}]$ is the super-indicator matrix, with $Q^* = \sum_{j=1}^{Q_c} q_j$

- Δ_j is the category dissimilarity matrix for variable j, i.e., the jth diagonal block of the block-diagonal matrix Δ.
- setting Δ_j determines the categorical distance measure of choice (independent- or association-based)

association-based pairwise distance

- differences in line with the inter-variables association/correlation are down-weighted

Association-based for categorical: total variation distance (TVD)[1] (2)

Consider the empirical joint probability distributions stored in the off-diagonal blocks of $\mathbf{P}$:

$$\mathbf{P} = \frac{1}{n}\begin{bmatrix} \mathbf{Z}_1^\top\mathbf{Z}_1 & \mathbf{Z}_1^\top\mathbf{Z}_2 & \cdots & \mathbf{Z}_1^\top\mathbf{Z}_{Q_c} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{Z}_{Q_c}^\top\mathbf{Z}_1 & \mathbf{Z}_{Q_c}^\top\mathbf{Z}_2 & \cdots & \mathbf{Z}_{Q_c}^\top\mathbf{Z}_{Q_c} \end{bmatrix}.$$

We refer to the conditional probability distributions for each variable j given each variable i ($i, j = 1, \ldots, Q_c, i \neq j$), stored in the block matrix

$$\mathbf{R} = \mathbf{P}_z^{-1}(\mathbf{P} - \mathbf{P}_z).$$

where $\mathbf{P}_z = \mathbf{P} \odot \mathbf{I}_{Q^*}$, and $\mathbf{I}_{Q^*}$ is the $Q^* \times Q^*$ identity matrix.

association-based pairwise distance

- differences in line with the inter-variables association/correlation are down-weighted

Association-based for categorical: total variation distance (TVD)[1] (3)

Let $\mathbf{r}_a^{ji}$ and $\mathbf{r}_b^{ji}$ be the rows of $\mathbf{R}_{ji}$, the (j,i)th off-diagonal block of $\mathbf{R}$ The category dissimilarity between a and b for variable j based on the total variation distance (TVD) is defined as

$$\delta_{tvd}^j(a,b) = \sum_{i\neq j}^{Q_c} w_{ji}\Phi^{ji}(\mathbf{r}_a^{ji}, \mathbf{r}_b^{ji}) = \sum_{i\neq j}^{Q_c} w_{ji}\left[\frac{1}{2}\sum_{\ell=1}^{q_i}|\mathbf{r}_{a\ell}^{ji} - \mathbf{r}_{b\ell}^{ji}|\right],$$

where $w_{ji} = 1/(Q_c - 1)$ for equal weighting (can be user-defined).

TVD-based dissimilarity matrix is, therefore,

$$\mathbf{D}_{tvd} = \mathbf{Z}\Delta^{(tvd)}\mathbf{Z}^{\mathsf{T}}.$$

association-based for mixed?

association-based for mixed

A straightforward AB-distance for mixed data is given by the convex combination of Mahalanobis and TVD distances:

$$\mathbf{D}_{mix} = \frac{Q_d}{Q}\mathbf{D}_{mah} + \left(1 - \frac{Q_d}{Q}\right)\mathbf{D}_{tvd}.$$

- this distance only accounts for correlations or associations among variables of the same type
- no continuous–categorical interactions are considered.

how to measure interactions?

how to measure interactions

define $\mathbf{\Delta}^{int}$, that accounts for the interactions and augment $\mathbf{\Delta}^{tvd}$

- the dissimilarity measure becomes

$$\mathbf{D}_{mix}^{(int)} = \mathbf{D}_{mah} + \mathbf{D}_{cat}^{(int)}.$$

where

$$\mathbf{D}_{cat}^{(int)} = \mathbf{Z}\tilde{\mathbf{\Delta}}\mathbf{Z}^{\top}$$

and

$$\tilde{\mathbf{\Delta}} = (1-\alpha)\mathbf{\Delta}^{tvd} + \alpha\mathbf{\Delta}^{int}$$

where $\alpha = \frac{1}{Q_c}$.

how to measure interactions

What is Δ^{int}?

- the general entry for the j^{th} diagonal block is $\delta^{j}_{int}(a,b)$ accounts for the interaction by measuring how the continuous variables help in discriminating between the observations choosing category a and those choosing category b for the j^{th} categorical variable

- consider the computation of $\delta^{ij}_{int}(ab)$ as a two-class (a/b) classification problem, with the continuous variables as predictors
  - use a distance-based classifier: nearest-neighbors

Δ_j^{int}

- consider $\mathbf{D}_{mah}$ and sort it to identify the neighbors for each observation.
- set a proportion of neighbors to consider, say $\hat{\pi}_{nn} = 0.1$
- for each pair of categories $(a, b), a, b = 1, \ldots, q_j, a \neq b$ of the j^{th} categorical variable:
- classify the observations using the prior corrected[1] decision rule

$$\text{if } i \text{ is such that } \quad \frac{\hat{\pi}_{nn}(a)}{\hat{\pi}(a)} \geq \frac{\hat{\pi}_{nn}(b)}{\hat{\pi}(b)} \quad \text{then assign } i \text{ to class } a \text{ else to class } b$$

- compute the balanced accuracy[2] (average of class-wise sensitivities)

$$\delta_{int}^{j}(a,b) = \frac{1}{2}\left(\frac{\texttt{true}\ a}{\texttt{true}\ a + \texttt{false}\ a} + \frac{\texttt{true}\ b}{\texttt{true}\ b + \texttt{false}\ b}\right)$$

well separated or not

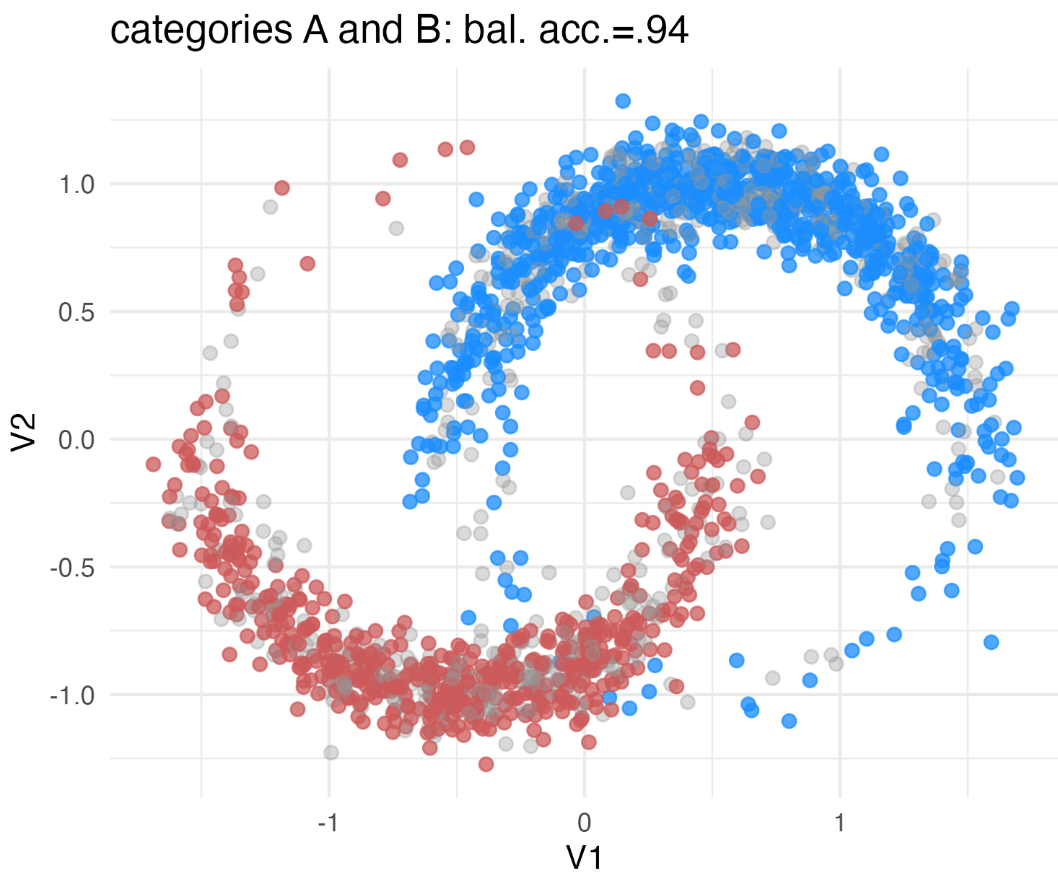
Building Δ_j^{int}

for the general categorical variable j with q_j you compute the quantities for $\frac{q_j(q_j-1)}{2}$ pairs

$$\Delta_{int} = \begin{pmatrix} 0 & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot \\ \cdot & \cdot & \cdot & 0 \end{pmatrix}$$

Building Δ_j^{int}

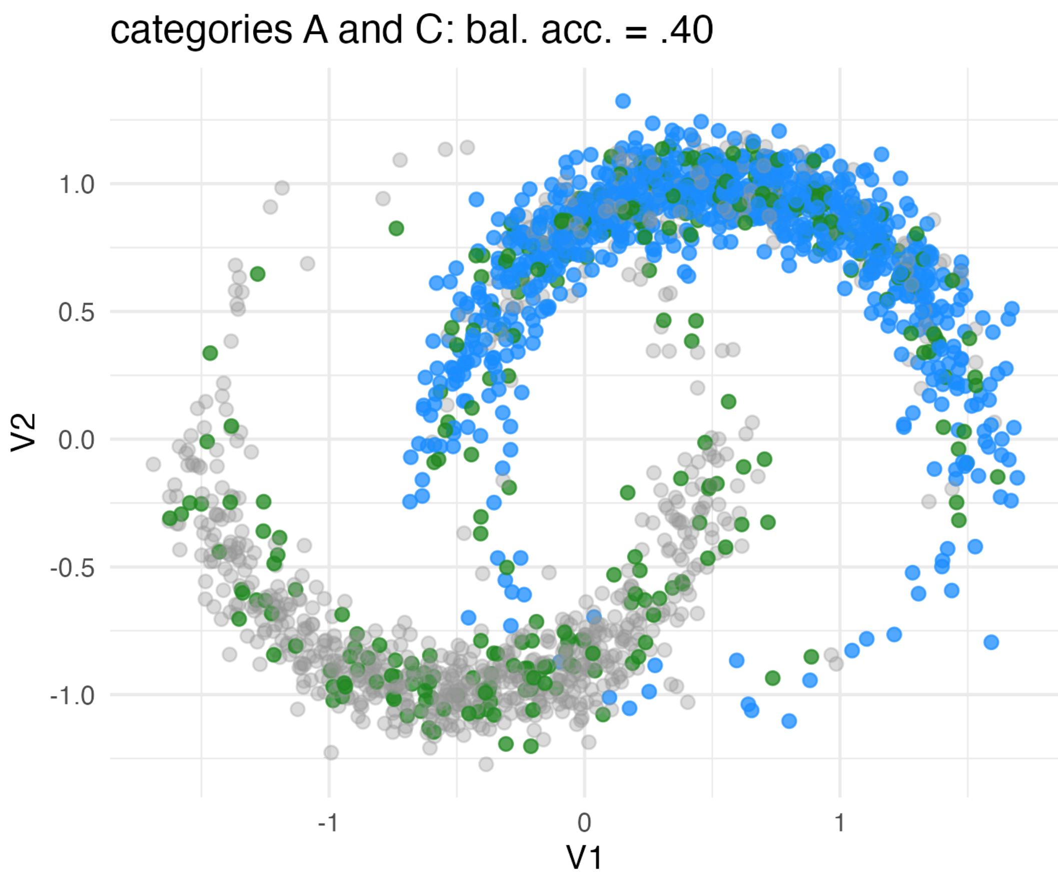
for the general categorical variable j with q_j you compute the quantities for $\frac{q_j(q_j-1)}{2}$ pairs



$$\Delta_{int} = \begin{pmatrix} 0 & 0.94 & \cdot & \cdot \\ 0.94 & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot \\ \cdot & \cdot & \cdot & 0 \end{pmatrix}$$

Building Δ_j^{int}

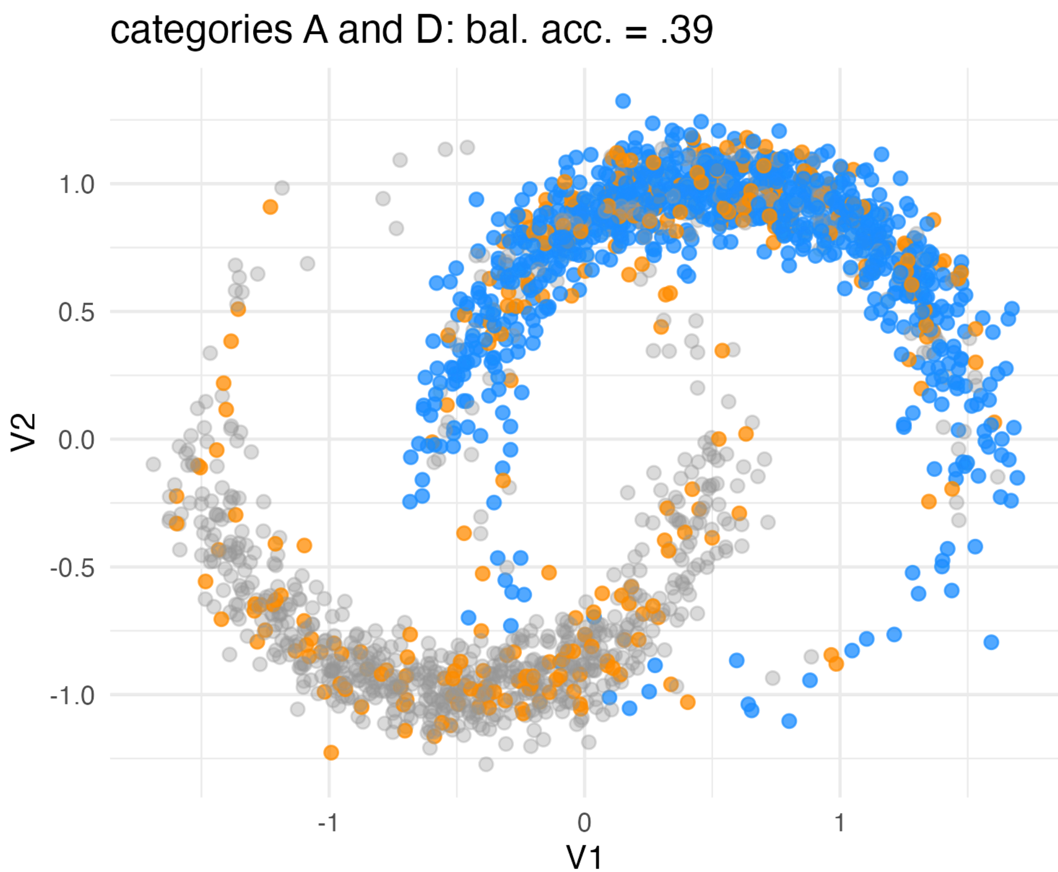
for the general categorical variable j with q_j you compute the quantities for $\frac{q_j(q_j-1)}{2}$ pairs



$$\Delta_{int} = \begin{pmatrix} 0 & 0.94 & \color{red}{0.4} & \cdot \\ 0.94 & 0 & \cdot & \cdot \\ \color{red}{0.4} & \cdot & 0 & \cdot \\ \cdot & \cdot & \cdot & 0 \end{pmatrix}$$

Building Δ_j^{int}

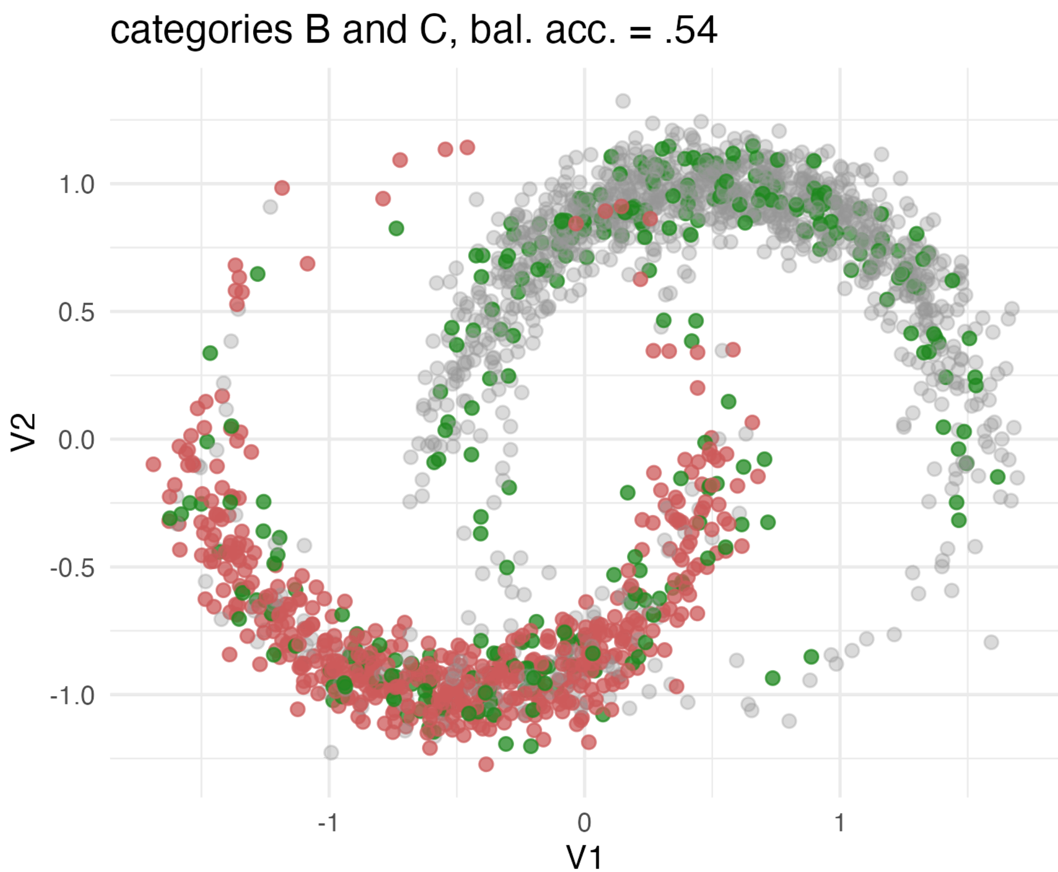
for the general categorical variable j with q_j you compute the quantities for $\frac{q_j(q_j-1)}{2}$ pairs



$$\Delta_{int} = \begin{pmatrix} 0 & 0.94 & 0.4 & 0.39 \\ 0.94 & 0 & \cdot & \cdot \\ 0.4 & \cdot & 0 & \cdot \\ 0.39 & \cdot & \cdot & 0 \end{pmatrix}$$

Building Δ_j^{int}

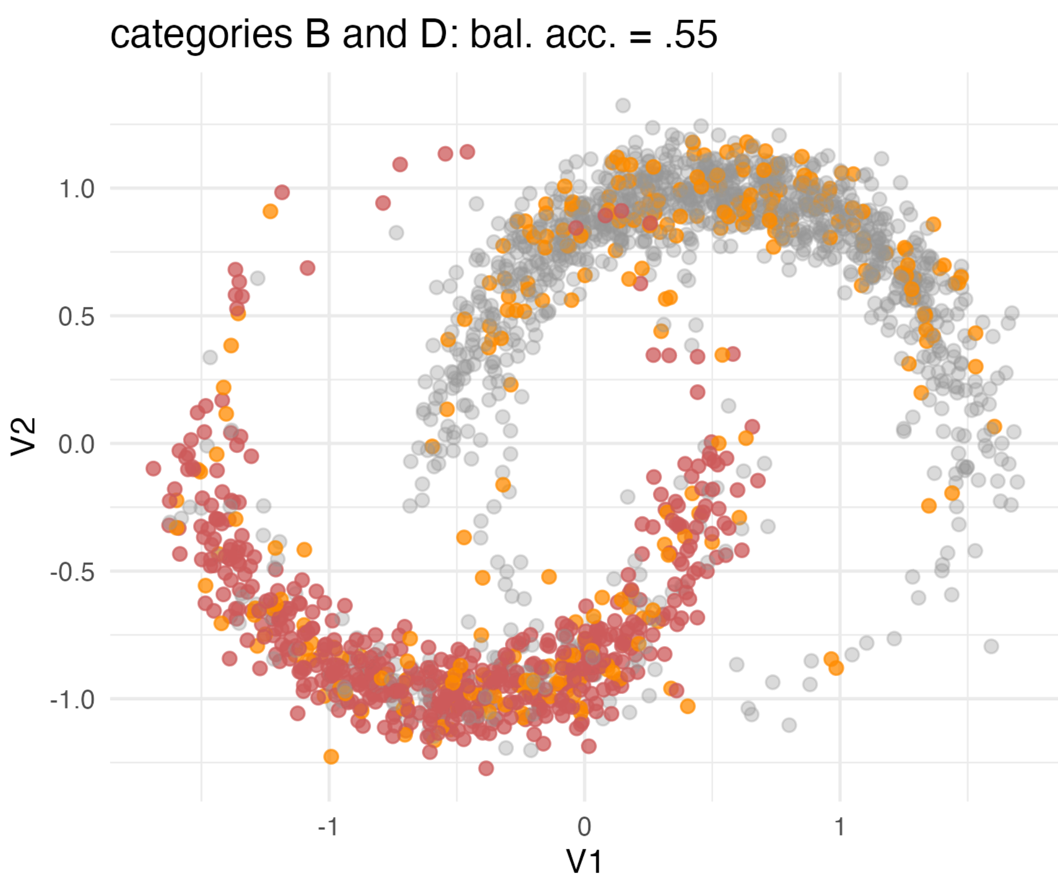
for the general categorical variable j with q_j you compute the quantities for $\frac{q_j(q_j-1)}{2}$ pairs



$$\Delta_{int} = \begin{pmatrix} 0 & 0.94 & 0.4 & 0.39 \\ 0.94 & 0 & 0.54 & \cdot \\ 0.4 & 0.54 & 0 & \cdot \\ 0.39 & \cdot & \cdot & 0 \end{pmatrix}$$

Building Δ_j^{int}

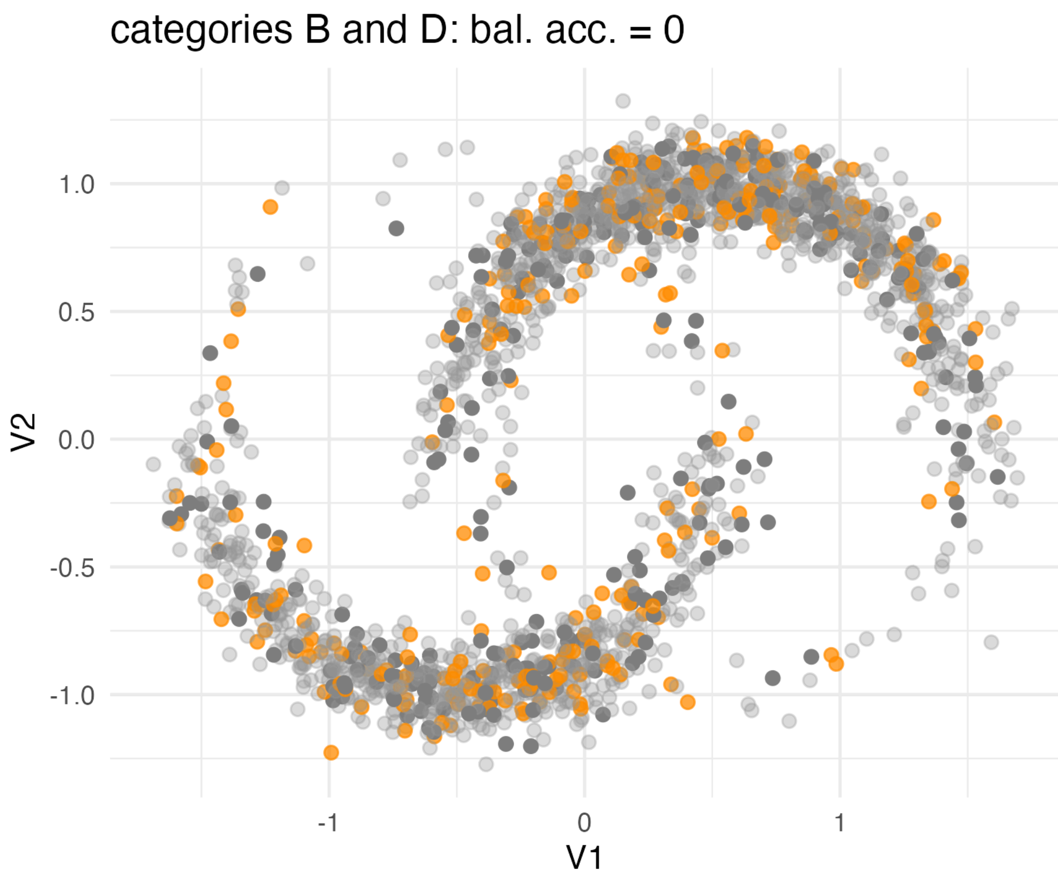
for the general categorical variable j with q_j you compute the quantities for $\frac{q_j(q_j-1)}{2}$ pairs



$$\Delta_{int} = \begin{pmatrix} 0 & 0.94 & 0.4 & 0.39 \\ 0.94 & 0 & 0.54 & 0.55 \\ 0.4 & 0.54 & 0 & \cdot \\ 0.39 & 0.55 & \cdot & 0 \end{pmatrix}$$

Building Δ_j^{int}

for the general categorical variable j with q_j you compute the quantities for $\frac{q_j(q_j-1)}{2}$ pairs



$$\Delta_{int} = \begin{pmatrix} 0 & 0.94 & 0.4 & 0.39 \\ 0.94 & 0 & 0.54 & 0.55 \\ 0.4 & 0.54 & 0 & 0 \\ 0.39 & 0.55 & 0 & 0 \end{pmatrix}$$

Just one-way interaction?

Just one-way interaction?

Let $\mathbf{X} = [\mathbf{X}_{con}\mathbf{X}_{cat}]$ be a mixed data matrix with n observations described by Q_d continuous and Q_c categorical variables, respectively, and let $\mathbf{x}_i = [\mathbf{x}_{i_{con}}\mathbf{x}_{i_{cat}}]$ the i^{th} observation

We build upon the following results

result A

The distribution of $\mathbf{x}_i$ can be written as

$$f(\mathbf{x}_{i_{con}}, \mathbf{x}_{i_{cat}}) = f(\mathbf{x}_{i_{con}})f(\mathbf{x}_{i_{cat}} \mid \mathbf{x}_{i_{con}})$$

result B

According to Hennig *et al.* (2019)[1], starting from an arbitrary distance $d(x, c_k)$ from a prototype, it possible to construct a probabilistic clustering model[2] as $f(x; c_k, s_k) = g(c_k, s_k)exp(-s_k d(x, c_k))$, where s_k is a concentration parameter.

Then if $f(\cdot) \sim N(\mathbf{c}_k, \mathbf{S}_k)$, and $d(\cdot, \cdot)$ the Mahalanobis distance, it results that

$$f(\mathbf{x}_i; \mathbf{c}_k, \mathbf{S}_k) = g(\mathbf{c}_k, \mathbf{S}_k) \exp\left[-d(\mathbf{x}_i, \mathbf{c}_k,)\right]$$

where $g(\mathbf{c}_k, \mathbf{S}_k)$ is a normalization constant to make sure $f(\mathbf{x}_i; \mathbf{c}_k, \mathbf{S}_k)$ is a density function.

Just one-way interaction?

result C

the dissimilarity between $\mathbf{x}_{i_{con}}$ and a generic cluster k with center $\mathbf{c}_k$ and covariance matrix $\mathbf{S}_k$ is[1]

$$d(\mathbf{x}_{i_{con}}, \mathbf{c}_k) = \log(M_k f(\mathbf{x}_{i_{con}}; \mathbf{c}_k, \mathbf{S}_k)^{-1})$$

- $f(\cdot)$ a symmetric probability density function
- M_k is the maximum of the density function
- that is, if $\mathbf{x}_{i_{con}} = \mathbf{c}_k$ then $d(\mathbf{x}_{i_{con}}, \mathbf{c}_k) = 0$.

replace $\mathbf{c}_k$ with a generic observation i', the above becomes

$$d(\mathbf{x}_{i_{con}}, \mathbf{x}_{i'_{con}}) = \log(M f(\mathbf{x}_{i_{con}}; \mathbf{x}_{i'_{con}}, \mathbf{S})^{-1})$$

where M is the maximum of the density function.

Just one-way interaction…

Using the result C, it can be shown that

$$d(\mathbf{x}_{i_{con}}, \mathbf{x}_{i'_{con}}) = \frac{1}{2}(\mathbf{x}_{i_{con}} - \mathbf{x}_{i'_{con}})\mathbf{S}^{-1}(\mathbf{x}_{i_{con}} - \mathbf{x}_{i'_{con}})^{\mathsf{T}}$$

categorical analogue

consider the Q_{cat}-dimensional categorical vector $\mathbf{x}_{i_{cat}}$, it results, for its j^{th} element, that

$$p(x_{ij_{cat}} = a | x_{i'j_{cat}} = b) = \left[\delta^j(a,b)\right]^{-1}$$

where a and b are two general categories of the jth variable

For the whole vector it results

$$p(\mathbf{x}_{ij_{cat}}; \mathbf{x}_{i'j_{cat}}) = \prod_{j=1}^{Q_c} p(x_{ij_{cat}} = a_j; x_{i'j_{cat}} = b_j) = \prod_{j=1}^{Q_c} \left[\delta^j(a_j, b_j)\right]^{-1}$$

Just one-way interaction… (wrap-up)

using the previous results, it is possible to define the dissimilarity between two mixed-data observations

$\mathbf{x}_i$ and $\mathbf{x}_{i'}$

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{1}{2}(\mathbf{x}_{i_{con}} - \mathbf{x}_{i'_{con}})\mathbf{S}^{-1}(\mathbf{x}_{i_{con}} - \mathbf{x}_{i'_{con}})^{\mathsf{T}} - \log(p(\mathbf{x}_{i_{cat}}|\mathbf{x}_{i_{con}}; \mathbf{x}_{i'_{cat}}, \mathbf{x}_{i'_{con}}))$$

that takes into account correlations, associations and cross-type interactions and is a equivalent to the genral entry of the previously defined

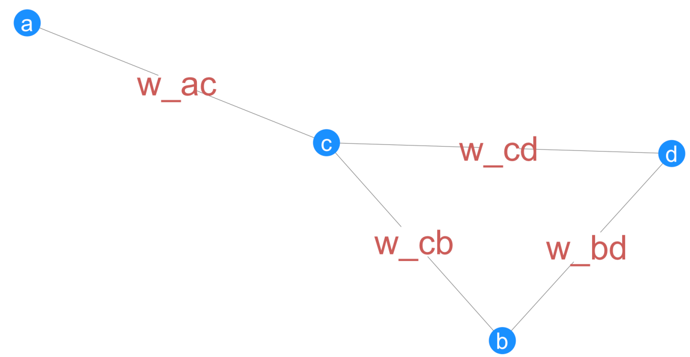
$$\mathbf{D}_{mix}^{(int)} = \mathbf{D}_{mah} + \mathbf{D}_{cat}^{(int)}.$$

spectral clustering in a nutshell

Spectral clustering: a graph partitioning problem

Graph representation

a graph representation of the data matrix $\mathbf{X}$: the aim is then to cut it into K groups (clusters)



the affinity matrix $\mathbf{A}$

the elements $\mathbf{w_{ij}}$ of $\mathbf{A}$ are high (low) if i and j are in the same (different) groups

.	a	b	c	d
a	0	0	w_ac	0
b	0	0	w_cb	w_bd
c	w_ca	w_cb	0	w_cd
d	0	w_db	w_dc	0

Spectral clustering: making the graph easy to cut

An approximated solution to the graph partitioning problem:

- spectral decomposition of the graph Laplacian matrix, that is a normalized version of the affinity matrix $\mathbf{A}$:

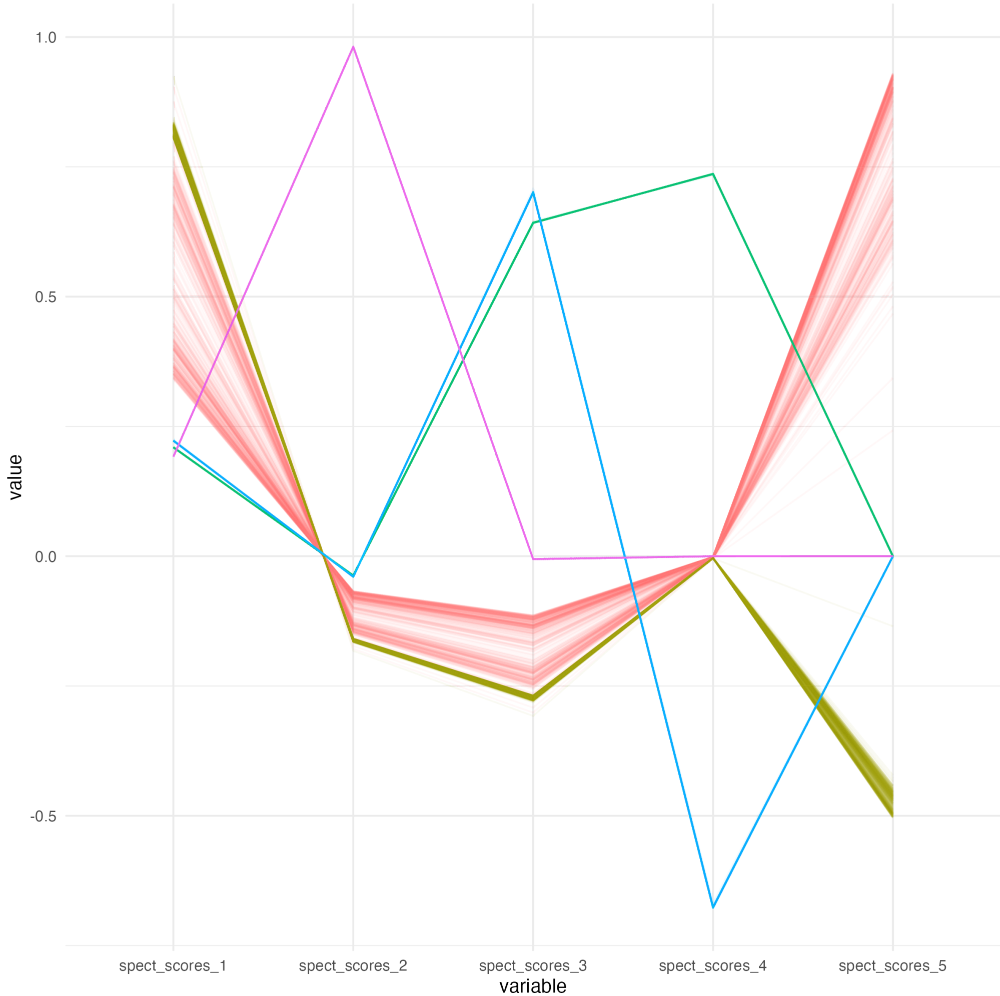
$$\mathbf{L} = \mathbf{D}_r^{-1/2} \underbrace{exp(-\mathbf{D}^2(2\sigma^2)^{-1})}_{\text{affinity matrix } \mathbf{A}} \mathbf{D}_r^{-1/2} = \mathbf{Q}\Lambda\mathbf{Q}^\mathsf{T}$$

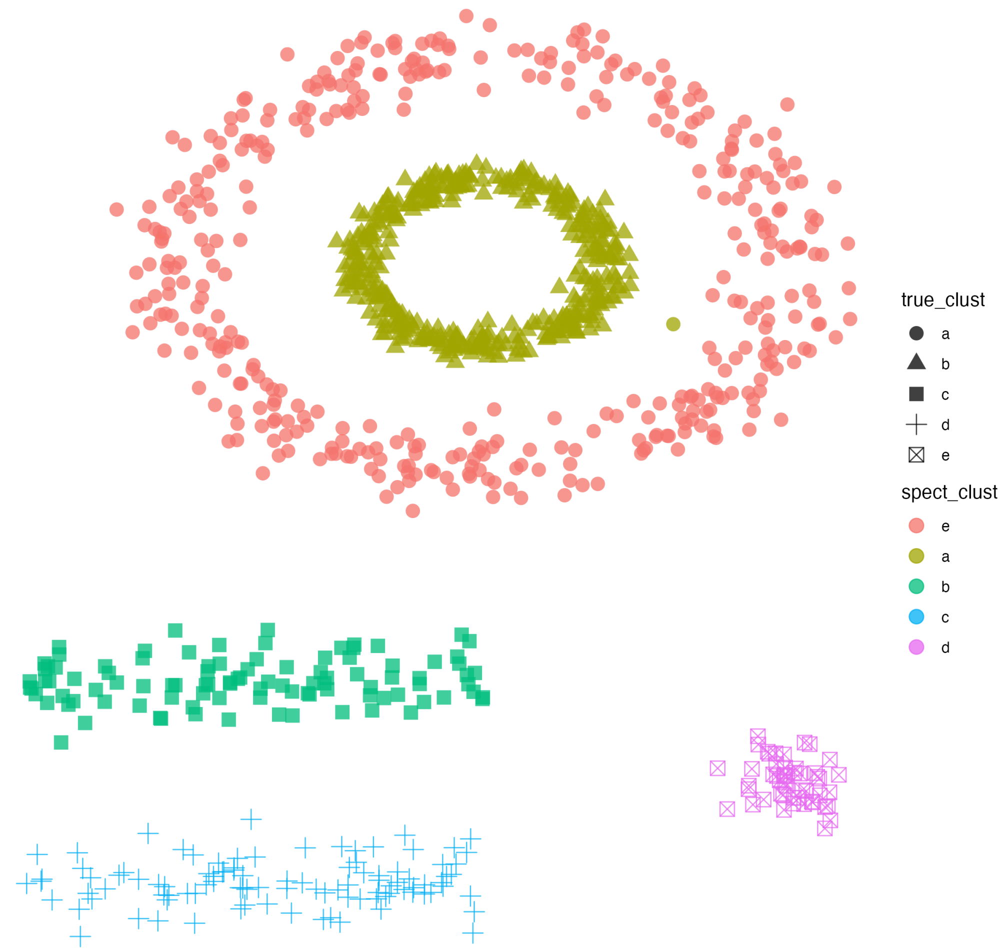
- $\mathbf{D}$ be the $n \times n$ matrix of pairwise distances
- the σ parameter dictates the number of neighbors each observation is linked to (rule of thumb: median distance to the 20th nearest neighbor)
- diagonal terms of $\mathbf{A}$ are set to zero: $a_{ii} = 0$, $i = 1, \dots, n$
- $\mathbf{D}_r = diag(\mathbf{r})$, $\mathbf{r} = \mathbf{A1}$ and $\mathbf{1}$ is an n-dimensional vector of 1's

- the spectral clustering of the n original objects is a K-means applied on the rows of the matrix $\tilde{\mathbf{Q}}$, containing the first K columns of $\mathbf{Q}$

Spectral clustering: solution and performance

SC works well, particularly in case of non-convex and overlapping clusters[1]





a toy experiment

toy experiment: data generation

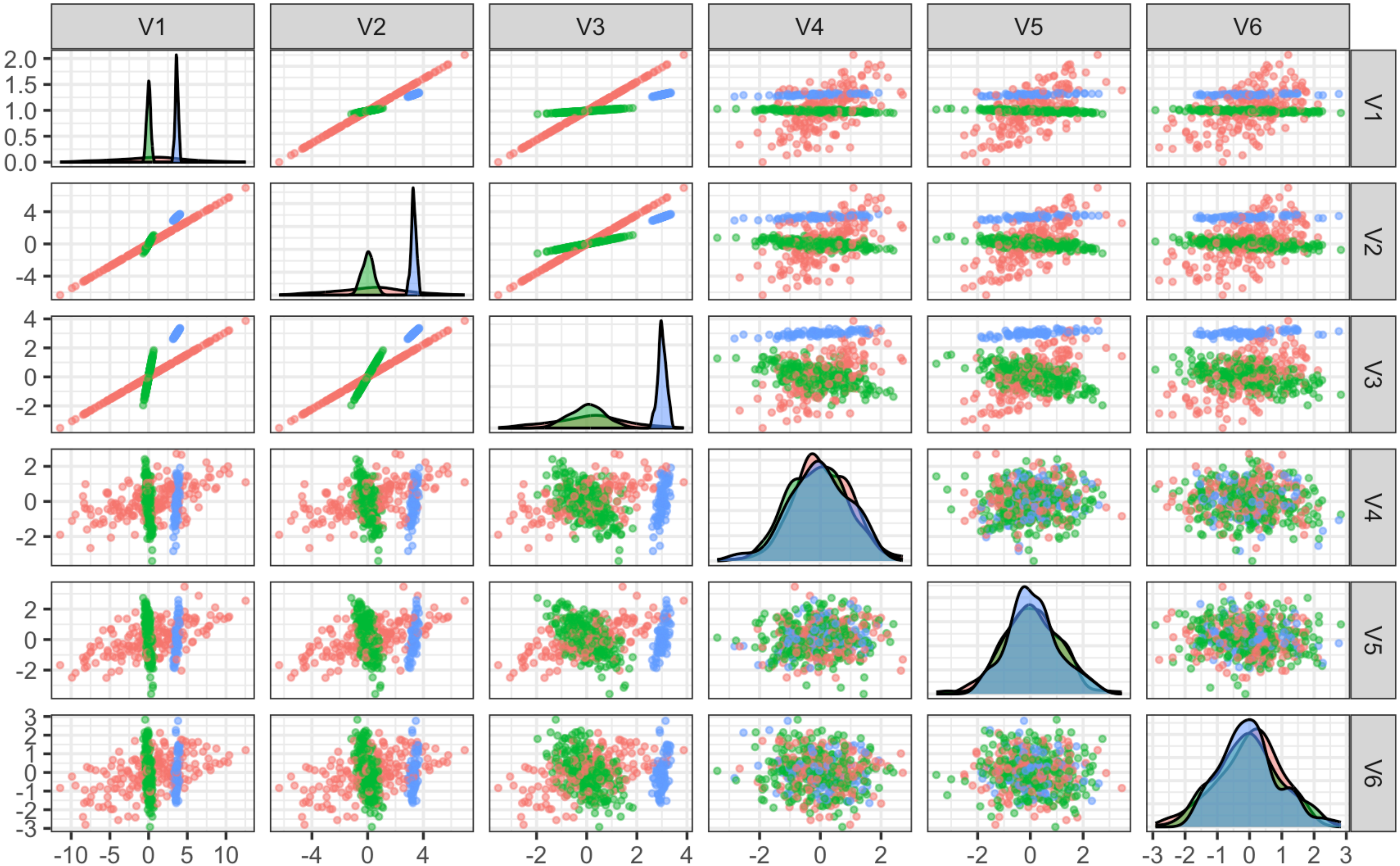
is $\delta^{j}_{int}(a,b)$ of help?

- three continuous signal, three continuous of Gaussian noise, three categorical variables (one noise)
- different dependence structures and location shifts in the signal continuous across the groups defined by the signal categorical variables.
- the continuous signal variables are generated conditionally on the signal categorical variables

$$X_j = \beta_{0,hq} + \sum_{k>j}^{6} \beta_{1,hq} X_k, \qquad j = 1, 2, 3$$

where $\beta_{1,hq}$ takes different values depending on the observed categories of the two signal categorical variables, h and q, respectively.

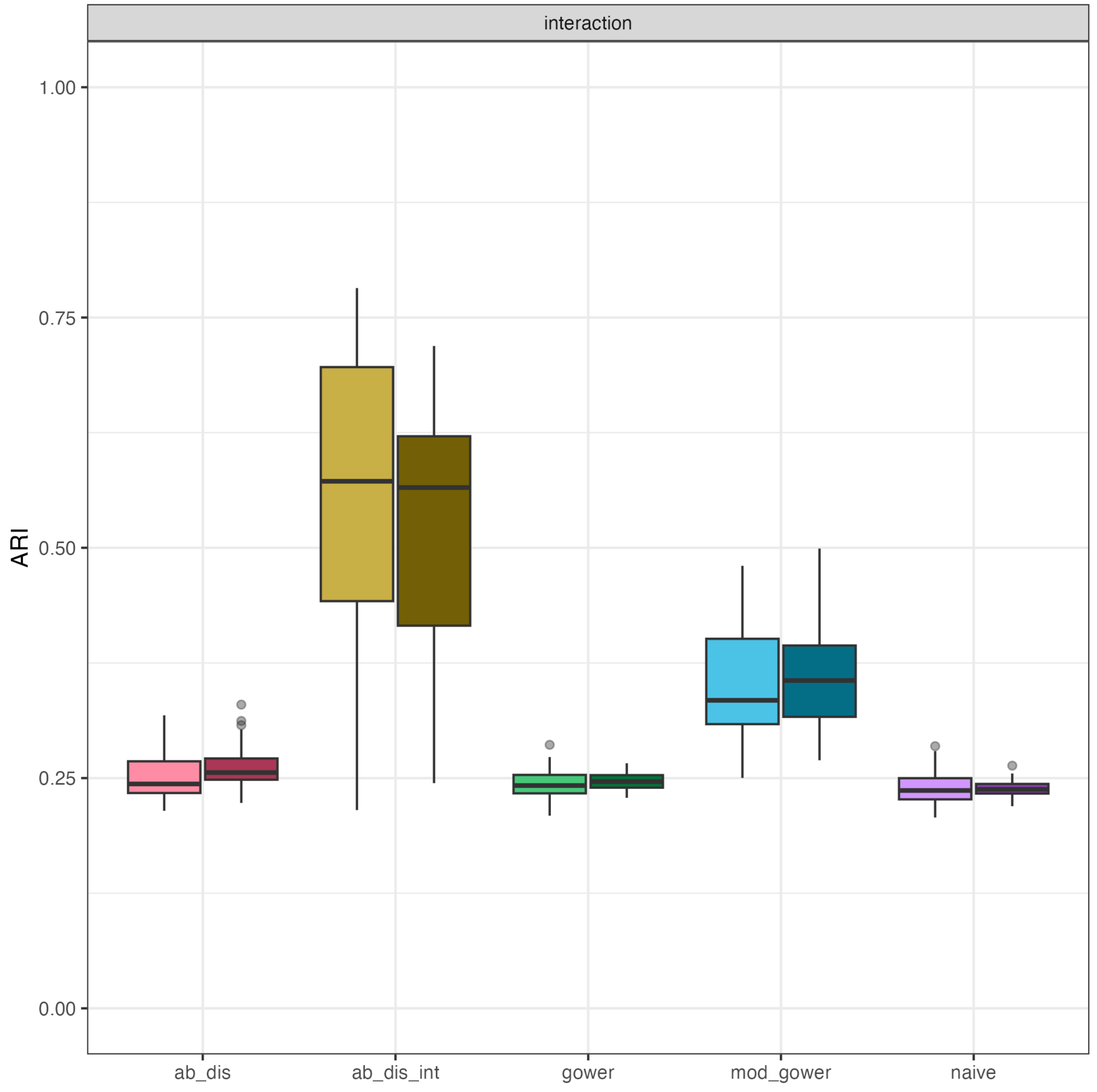
toy experiment



toy experiment: compared methods

- gower dissimilarity: a straightforward option
- naive [1]
- modified gower[2]
- association-based approach[3]: with and without interaction

toy experiment



Final considerations and future work

- association-based measures aim to go beyond match/mismatch of categories
- when the signal is limited to few variables, retrieving information from cont/cat interaction may be useful
- measuring interactions via non-parametric approach NN-based approach is suitable for non-convex/oddly shaped clusters
- computationally demanding (but it can be made bearable)
- π_{nn} tuning, regularization of δ_{int}'s to reduce variability

an R package to compute distances: manydist!

manydist on CRAN (stable(-ish) version)

manydist on GitHub version (for the latest updates)

general features

The `manydist` package is designed around a modular distance-based workflow.

Component	Function / Interface	Role
Core distance constructor	`mdist()`	Computes a pairwise dissimilarity matrix for mixed-type data: different indipendence-based and association based are specified, fully customizable
Recipe step	`step_mdist()`	Preprocessing step that computes an `mdist` dissimilarity matrix within a `recipe`, for seamless integration in learning pipelines.
k-nearest neighbors	`mdist_knn()`	Model specification for k-NN using a precomputed `mdist` dissimilarity matrix as input. Supports `tune`
leave-one-variable-out	`lovo_mdist()`	LOVO dissimilarity computations to assess by-variable contributions

forthcoming features

Component	Function / Interface	Role
Partitioning around medoids	`mdist_pam()`	Model specification for PAM clustering based on `mdist` dissimilarities.
spectral clustering	`mdist_spectral()`	Model specification for spectral clustering based on `mdist` dissimilarities.

and more to come…

main references

Hennig, C., C. Viroli, and L. Anderlucci (2019). “Quantile-based clustering”. In: *Electronic Journal of Statistics* 13.2, pp. 4849 - 4883. DOI: 10.1214/19-EJS1640. URL: https://doi.org/10.1214/19-EJS1640.

Le, S. Q. and T. B. Ho (2005a). “An association-based dissimilarity measure for categorical data”. In: *Pattern Recognition Letters* 26.16, pp. 2549-2557.

Mbuga, F. and C. Tortora (2021a). “Spectral Clustering of Mixed-Type Data”. In: *Stats* 5.1, pp. 1-11.

Murugesan, N., I. Cho, and C. Tortora (2021). “Benchmarking in cluster analysis: a study on spectral clustering, DBSCAN, and K-Means”. In: *Conference of the International Federation of Classification Societies*. Springer. , pp. 175-185.

Velden, M. van de, A. Iodice D’Enza, A. Markos, et al. (2024). “A general framework for implementing distances for categorical variables”. In: *Pattern Recognition* 153, p. 110547.

Velden, M. van de, A. Iodice D’Enza, A. Markos, et al. (2025a). “A general framework for unbiased mixed-variables distances”. In: *second round review: Journal of Computational and Graphical Statistics*.

Acknowledgments

The publication was produced with funding from the Italian Ministry of University and Research under the Call for Proposals related to the scrolling of the final rankings of the PRIN 2022 call.

Latent variable models and dimensionality reduction methods for complex data (PI Prof. Paolo Giordani)

Project No. 20224CRB9E, CUP C53C24000730006

Grant Assignment Decree No. 1401 adopted on 18.9.2024 by MUR

you can check this presentation at

https://alfonsoiodicede.github.io/Workshop_Bologna_26_pres_SC_mix.html