

UCLA

Department of Statistics Papers

Title

Homogeneity Analysis of Pavings

Permalink

<https://escholarship.org/uc/item/6wq739zf>

Author

Jan de Leeuw

Publication Date

2011-10-25

Homogeneity Analysis of Pavings

Jan de Leeuw
UCLA Statistics

<http://gifi.stat.ucla.edu/pub/roll-call.pdf>

- A *paving* is a set of subsets of a *carrier*, a set of *objects*. These subsets are the *data*.
- Each subset is a *characteristic*. Characteristics cover objects, objects *belong to* characteristics.
- A characteristic can be *coded* as an *indicator function*, i.e. a function of the set of objects to $\{0,1\}$.
- The *carrier* and the number of characteristics are not necessarily finite.
- Objects have *profiles*.

$$O : \Lambda \longrightarrow 2^{\Omega} \quad \text{and} \quad \mathcal{O} = O(\Lambda)$$

$$L : \Omega \longrightarrow 2^{\Lambda}$$

$$\Lambda = \{1, 2, 3, 4\}$$

$$\Omega = \{\text{John}, \text{Paul}, \text{George}, \text{Ringo}, \text{Yoko}, \text{Linda}\}$$

$$O(1) = \{\text{John}, \text{Paul}, \text{George}, \text{Ringo}\}$$

$$O(2) = \{\text{John}, \text{Paul}\}$$

$$O(3) = \{\text{John}, \text{Yoko}\}$$

$$O(4) = \{\text{Ringo}\}$$

$$L(\text{John}) = \{1, 2, 3\}$$

$$L(\text{Paul}) = \{1, 2\}$$

$$L(\text{George}) = \{1\}$$

$$L(\text{Ringo}) = \{1, 4\}$$

$$L(\text{Yoko}) = \{3\}$$

$$L(\text{Linda}) = \emptyset$$

John	1	1	1	0
Paul	1	1	0	0
George	1	0	0	0
Ringo	1	0	0	1
Yoko	0	0	1	0
Linda	0	0	0	0

- legislators and votes
- students and items
- animals and morphology
- plants and transects
- artefacts and graves
- interviewees and attitude questions

- A *representation* is a mapping of the carrier into a *representation space*. Often metric, often Euclidean.
- In this space a notion of the *size* of a subset is defined. Often a gauge or norm. Often a function of the distances.
- representation

$$\phi : \Omega \rightarrow \mathcal{X}$$

For each characteristic we define its *homogeneity*

$$\mathbf{hom}(\phi, O) = \frac{\mathbf{size}(\phi(O))}{\mathbf{size}(\phi(\Omega))}$$

and we aggregate homogeneities over characteristics

$$\mathbf{hom}(\phi) = \frac{\mathbf{ave}\{\mathbf{size}(\phi(O)) \mid O \in \mathcal{O}\}}{\mathbf{size}(\phi(\Omega))}$$

Usually $\text{size}()$ is an *outer measure*, i.e.

- $i)$ $\text{size}(X) \geq 0 \quad \forall X \in \mathcal{X}$
- $ii)$ if $X \subseteq Y$ then $\text{size}(X) \leq \text{size}(Y) \quad \forall X, Y \in \mathcal{X}$
- $iii)$ $\text{size}(X \cup Y) \leq \text{size}(X) + \text{size}(Y) \quad \forall X, Y \in \mathcal{X}$

And usually $\text{ave}()$ satisfies (for all sequences of reals)

- $i)$ $\min(X) \leq \text{ave}(X) \leq \max(X)$
- $ii)$ if $X \leq Y$ then $\text{ave}(X) \leq \text{ave}(Y)$

Representation on the line: possible sizes

- the range
- sum of squares around the mean
- mean deviation around mean or median
- Gini's mean difference
- other robust measures of scale

Representation space is Euclidean: possible sizes

- length of the MST
- size of convex hull
- size of Weber star
- size of Gifi star
- length of traveling salesman tour
- size of smallest enclosing sphere/ellipsoid

Algorithm

Consider the problem of maximizing

$$\lambda(x) = \frac{\alpha(x)}{\beta(x)}$$

where both numerator and denominator are gauges (homogeneous convex functions). We use Robert's algorithm

$$y^{(k)} \in \partial\alpha(x^{(k)}) \quad x^{(k+1)} \in \partial\beta^\circ(y^{(k)})$$

Polar and Subdifferential

$$\eta^\circ(y) = \inf\{\mu \geq 0 \mid x'y \leq \mu\eta(x) \ \forall x\}$$

$$\eta^\circ(y) = \max_{x \neq 0} \frac{x'y}{\eta(x)}$$

$$\partial\eta(x) = \{y \mid \eta(z) \geq \eta(x) + y'(z - x) \ \forall z\}$$

Dedoublement

- For each characteristic we can add its complement to the data
$$O \in \mathcal{O} \iff \Omega - O \in \mathcal{O}$$
- This means we do not only try make the size of the representation of the characteristic small, but also the size of its complement.
- Generalization of this: missing data.
- Related to this: homogeneity analysis of partitionings (see below).

Order Theorems

- If we use dedoublement and the least squares size measure, then order is recovered for *monotone* items (Guttman, 1941)
- If we do *not* use dedoublement and the least squares size measure, then order is recovered for *single-peaked* items (Mosteller, 1942).
- Does this generalize to other size measures ?

Further Developments

- Fuzzy Pavings. These are mappings of the objects into $[0,1]$ instead of $\{0,1\}$.
- Partitions. This generalizes dedoublement to more than two subsets.
- Fuzzy Partitions. For each object we use a partition of unity.

Example from ADA

- House and Senate 2000.
- With and without dedoublement.
- Using Gifi stars (i.e. size is squared distance to the centroid).





