Visualizing Cluster Results

Using Package FlexClust and Friendsd

Friedrich Leisch

University of Munich

useR!, Rennes, 10.7.2009

Acknowledgements & Apology

- Sara Dolnicar (University of Wollongong)

- Theresa Scharl, Ingo Voglhuber (Vienna University of Technology)

- Paul Murrell, Deepayan Sarkar (R Core)

Acknowledgements & Apology

- Sara Dolnicar (University of Wollongong)
- Theresa Scharl, Ingo Voglhuber (Vienna University of Technology)
- Paul Murrell, Deepayan Sarkar (R Core)

Apology: Microarray data only in Theresa's talk (try time-shift back to Wednesday).

Partitioning Clustering – KCCA

K-centroid cluster algorithms:

- data set $\mathcal{X}_N = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, set of centroids $C_K = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$
- distance measure $d(\mathbf{x}, \mathbf{y})$
- centroid $\mathbf{c}$ closest to $\mathbf{x}$:

$$c(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{c} \in C_K} d(\mathbf{x}, \mathbf{c})$$

- Most KCCA algorithms try to find a set of centroids C_K for fixed K such that the average distance

$$D(\mathcal{X}_N, C_K) = \frac{1}{N} \sum_{n=1}^{N} d(\mathbf{x}_n, c(\mathbf{x}_n)) \rightarrow \min_{C_K},$$

  of each point to the closest centroid is minimal.
- Optimization algorithm not important for rest of talk

Example: Australian Volunteers

Survey among 1415 Australian adults about which organziations they would consider to volunteer for, main motivation to volunteer, actual volunteering, image of organizations, . . .

We use a block of 19 binary questions ("applies", "does not apply") about motivations to volunteer: "I want to meet people", "I have no one else", "I want to set an example", . . .

Organziations investigated: Red Cross, Surf Life Savers, Rural Fire Service, Parents Association, . . .

Our analyses show that there is both competition between organizations with similar profiles as well as complimentary effects (idividuals volunteering for more than one arganization, in most cases with very different profiles).

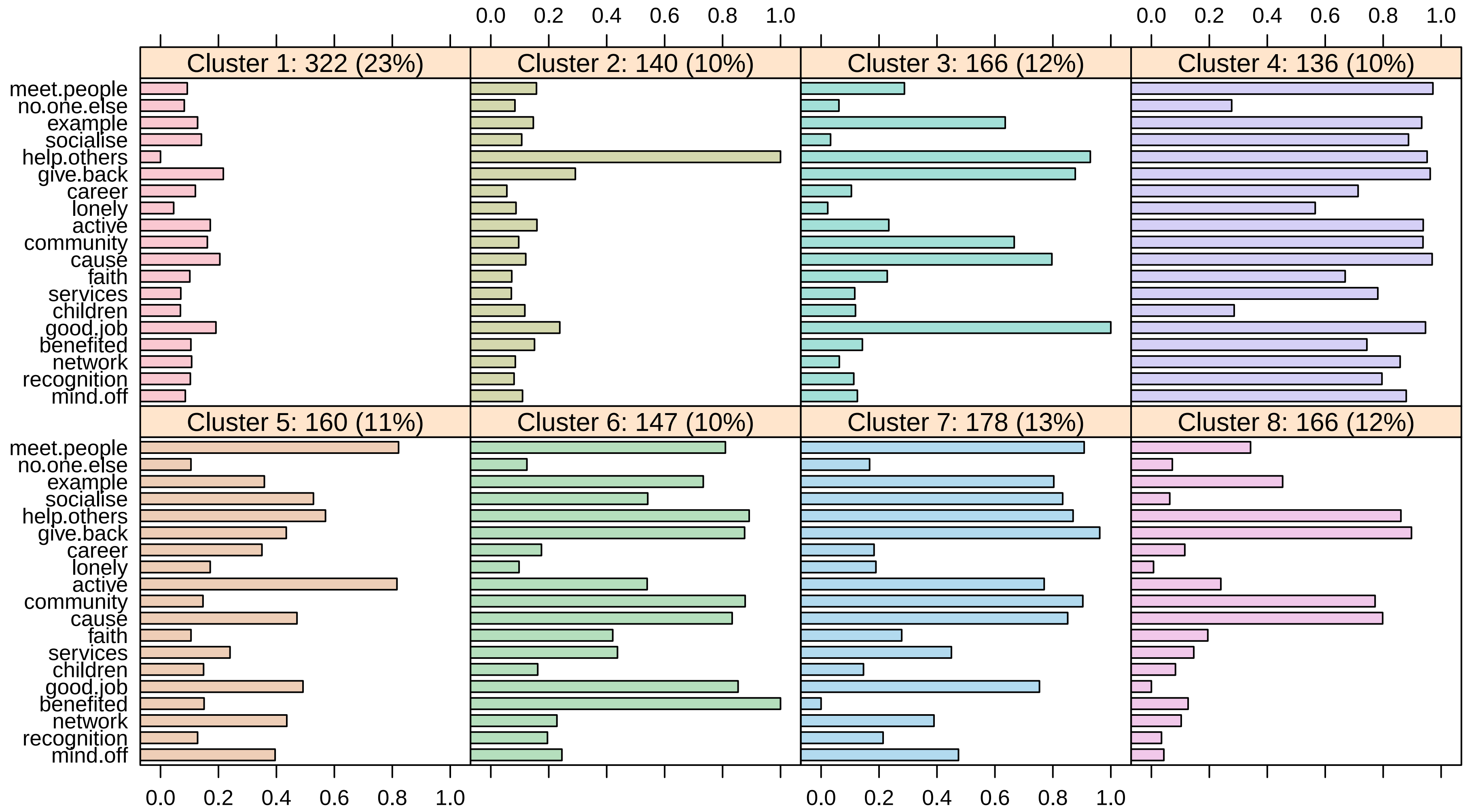
8 Volunteer Clusters

	Cl.1	Cl.2	Cl.3	Cl.4	Cl.5	Cl.6	Cl.7	Cl.8	Total
meet.people	9.24	15.77	28.77	97.18	82.17	80.97	90.81	34.23	49.47
no.one.else	8.21	8.34	6.16	27.72	10.51	12.47	16.75	7.23	11.38
example	12.80	14.68	63.56	93.30	35.84	73.33	80.32	45.30	47.63
socialise	14.12	10.68	3.28	88.74	52.79	54.17	83.39	6.35	35.83
help.others	0.00	100.00	92.93	95.17	56.95	89.18	86.98	86.12	66.78
give.back	21.68	29.18	87.73	96.24	43.41	87.60	96.18	89.77	63.75
career	12.04	5.54	10.46	71.34	35.01	17.49	18.29	11.54	20.57
lonely	4.55	8.71	2.30	56.55	17.16	9.76	18.93	0.75	13.14
active	17.17	15.93	23.38	93.82	81.61	53.97	77.00	23.98	44.88
community	16.17	9.64	66.66	93.75	14.67	87.80	90.34	77.21	52.72
cause	20.49	12.07	79.66	96.91	47.11	83.31	85.12	79.82	58.66
faith	10.12	7.24	22.84	66.89	10.52	42.10	27.83	19.47	24.03
services	7.00	7.10	11.63	78.15	23.99	43.72	44.98	14.64	25.23
children	6.88	11.76	11.88	28.58	14.86	16.20	14.64	8.31	13.00
good.job	19.14	23.81	100.00	94.61	49.18	85.35	75.38	0.00	51.80
benefited	10.49	15.09	14.26	74.37	15.04	100.00	0.00	12.68	26.29
network	10.75	8.47	6.29	85.86	43.59	22.83	38.98	10.28	25.94
recognition	10.30	8.03	11.29	79.59	12.80	19.56	21.40	3.49	18.73
mind.off	8.56	10.95	12.53	87.96	39.55	24.55	47.43	4.31	26.57

8 Volunteer Clusters

	Cl.1	Cl.2	Cl.3	Cl.4	Cl.5	Cl.6	Cl.7	Cl.8	Total
meet.people	9	16	29	97	82	81	91	34	49
no.one.else	8	8	6	28	11	12	17	7	11
example	13	15	64	93	36	73	80	45	48
socialise	14	11	3	89	53	54	83	6	36
help.others	0	100	93	95	57	89	87	86	67
give.back	22	29	88	96	43	88	96	90	64
career	12	6	10	71	35	17	18	12	21
lonely	5	9	2	57	17	10	19	1	13
active	17	16	23	94	82	54	77	24	45
community	16	10	67	94	15	88	90	77	53
cause	20	12	80	97	47	83	85	80	59
faith	10	7	23	67	11	42	28	19	24
services	7	7	12	78	24	44	45	15	25
children	7	12	12	29	15	16	15	8	13
good.job	19	24	100	95	49	85	75	0	52
benefited	10	15	14	74	15	100	0	13	26
network	11	8	6	86	44	23	39	10	26
recognition	10	8	11	80	13	20	21	3	19
mind.off	9	11	13	88	40	25	47	4	27

8 Volunteer Clusters



8 Volunteer Clusters

We cannot easily test for differences between clusters, because they were constructed to be different.

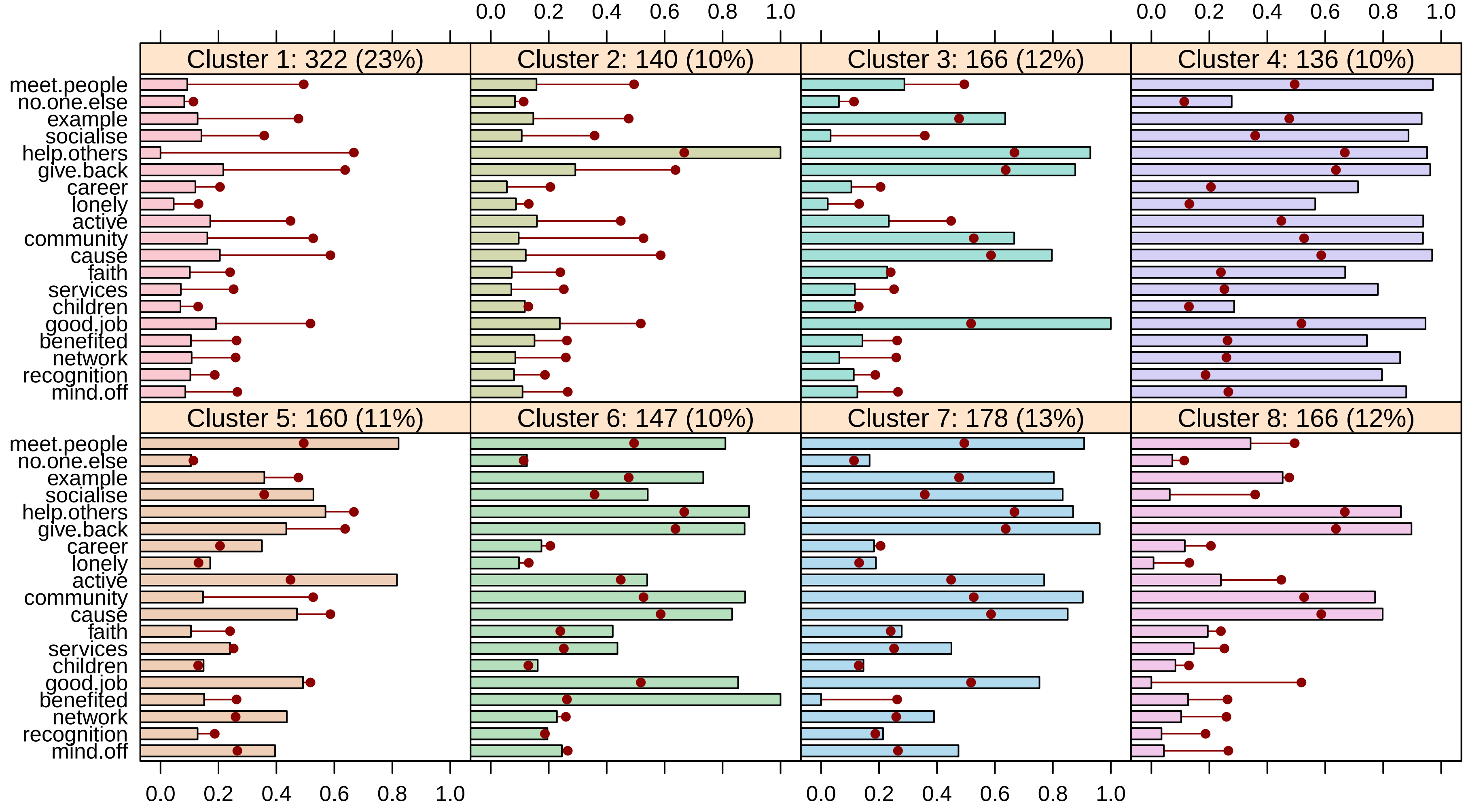
8 Volunteer Clusters

We cannot easily test for differences between clusters, because they were constructed to be different.

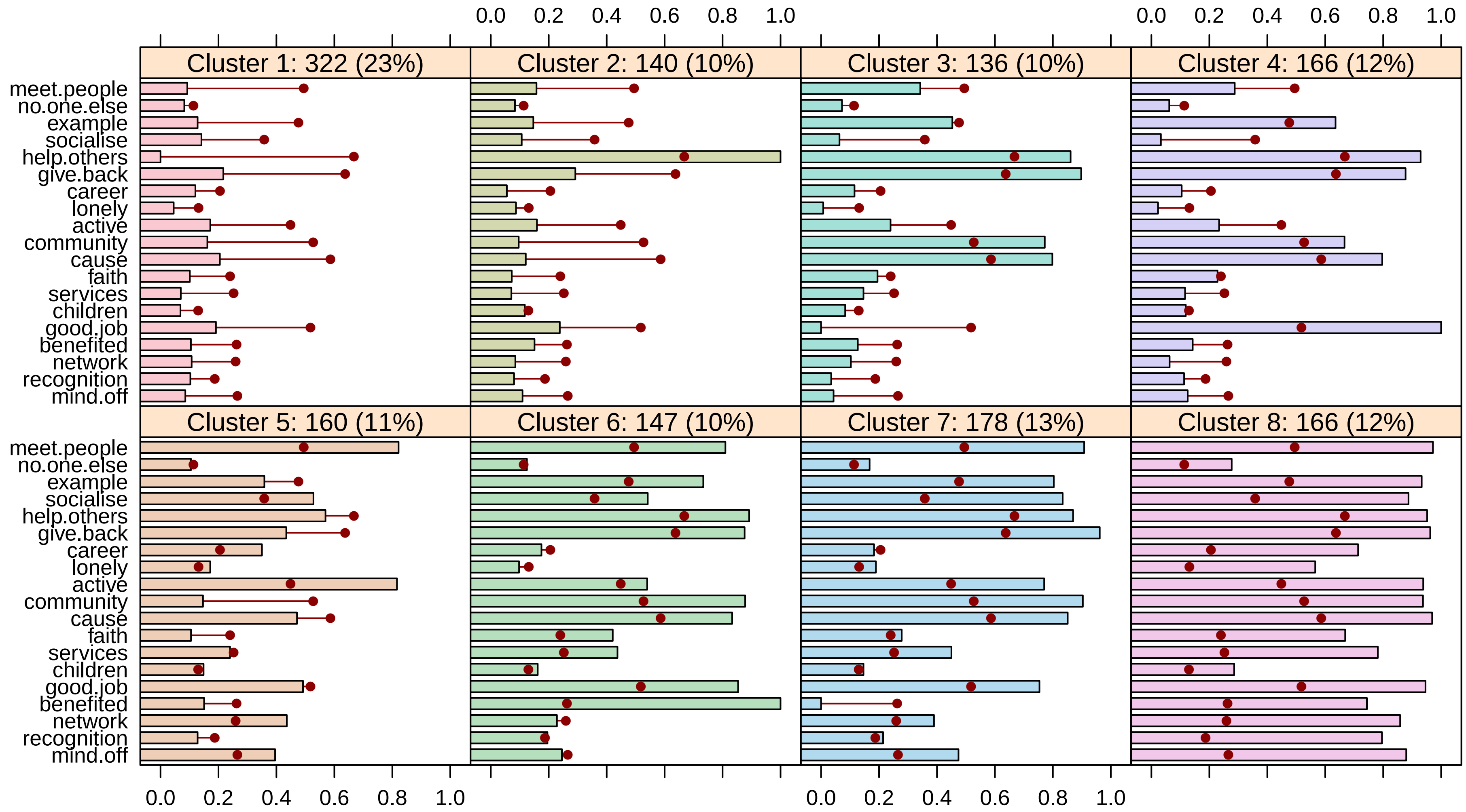
Improved presentation of results (following advice that is only around for a few decades):

- Add reference lines/points
- Sort variables by content:
  1. sort clusters by mean
  2. sort variables by hierarchical clustering
- Highlight important points

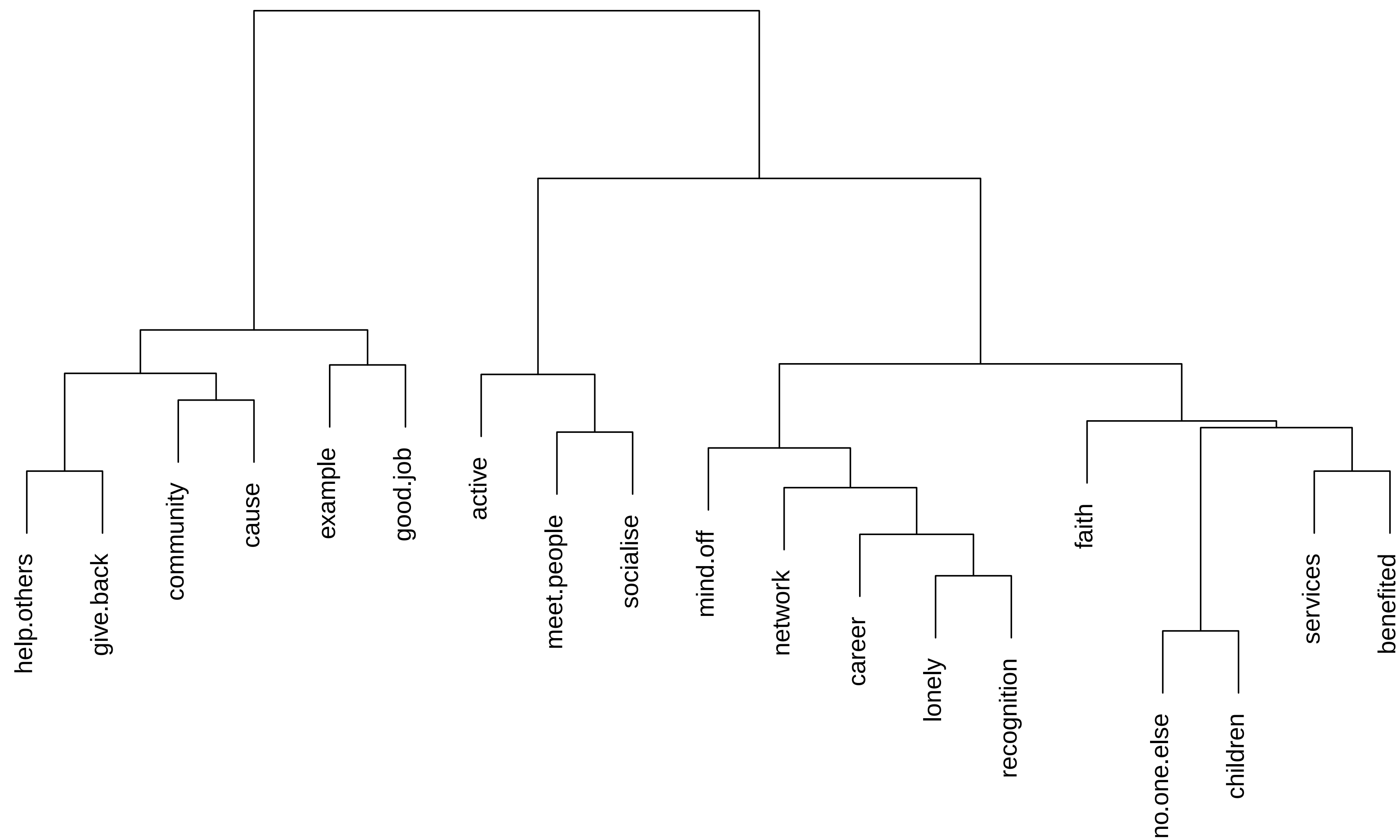
Add reference points



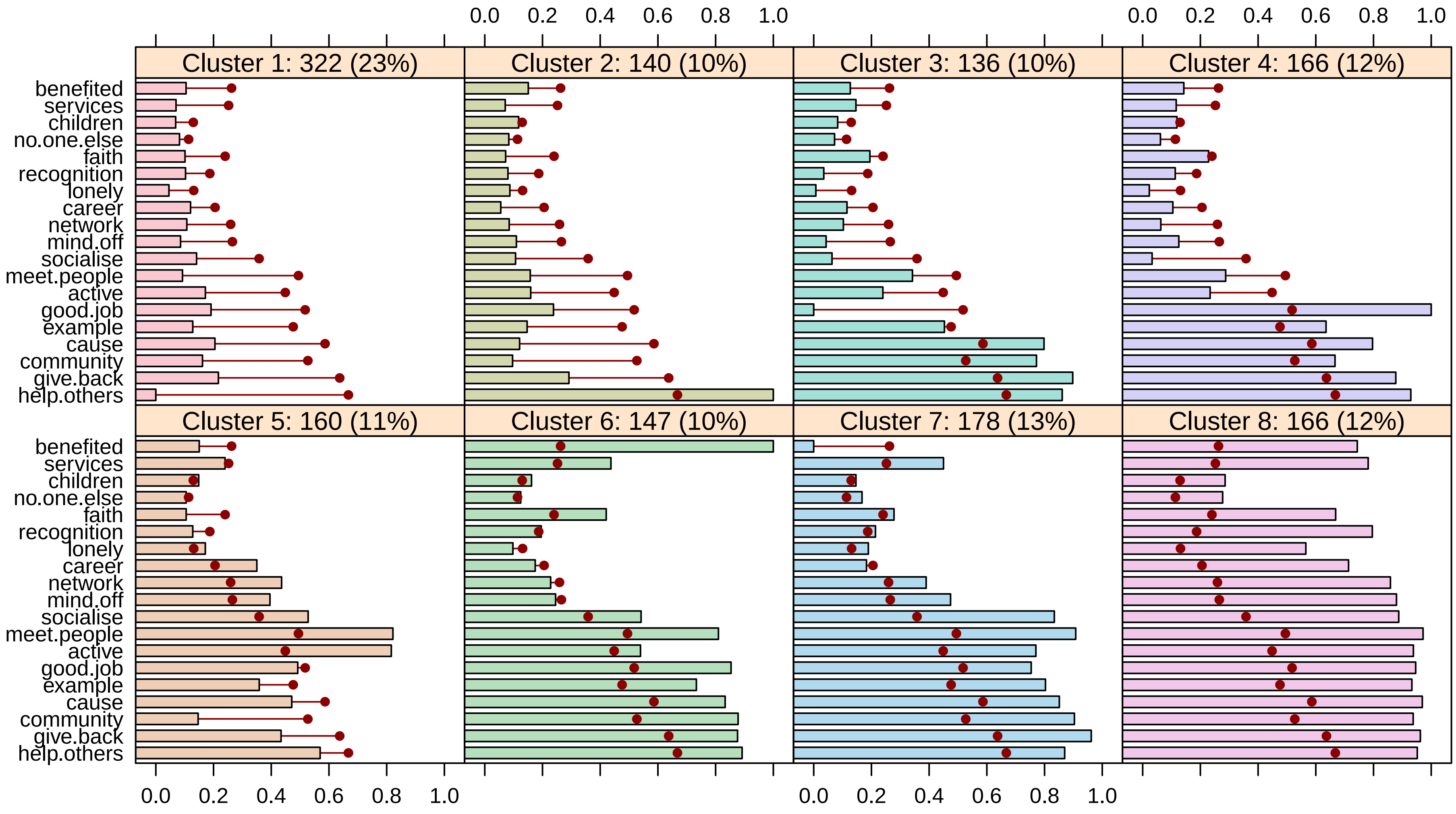
Sort Clusters



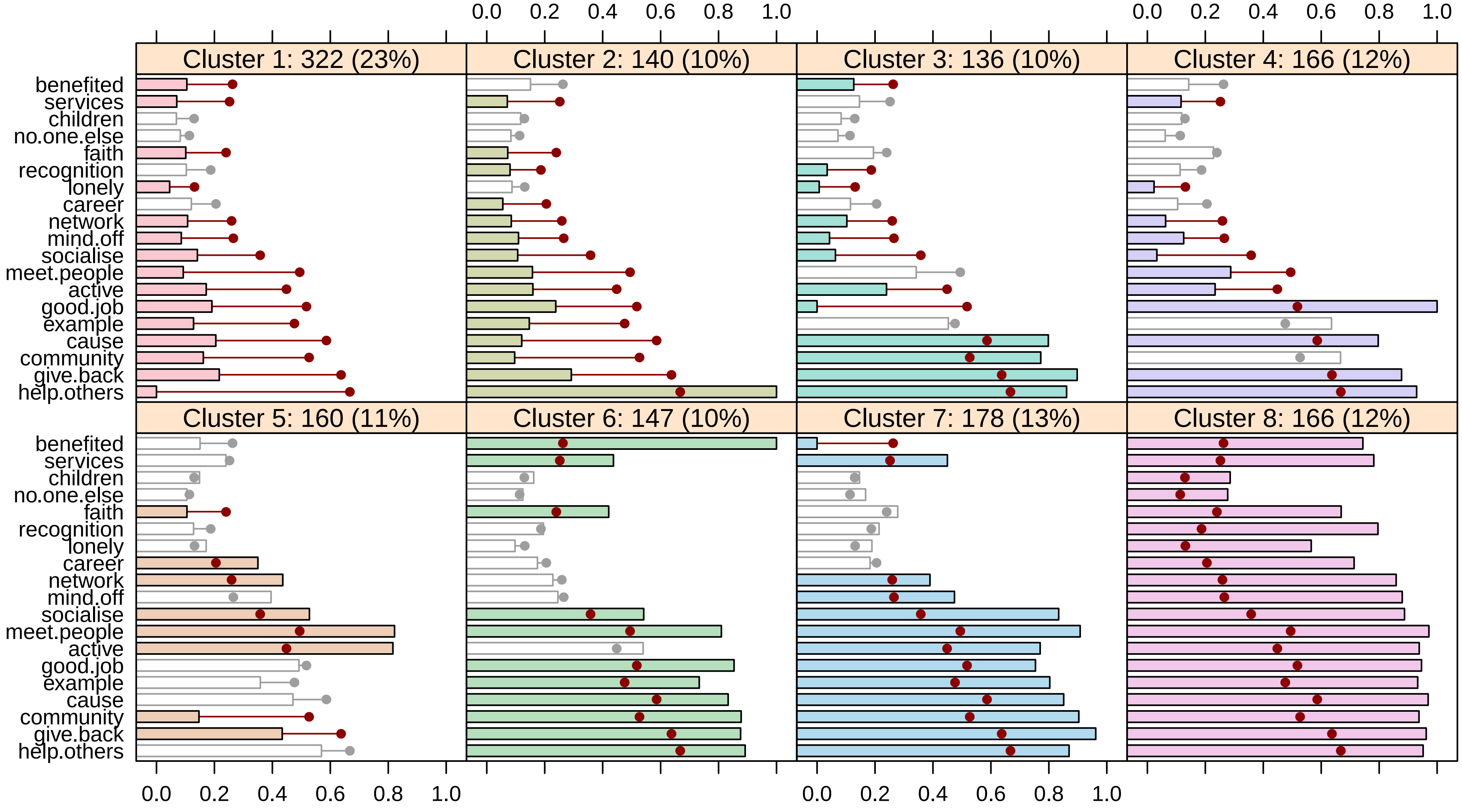
Sort Variables



Sort Variables



Highlight Important Points



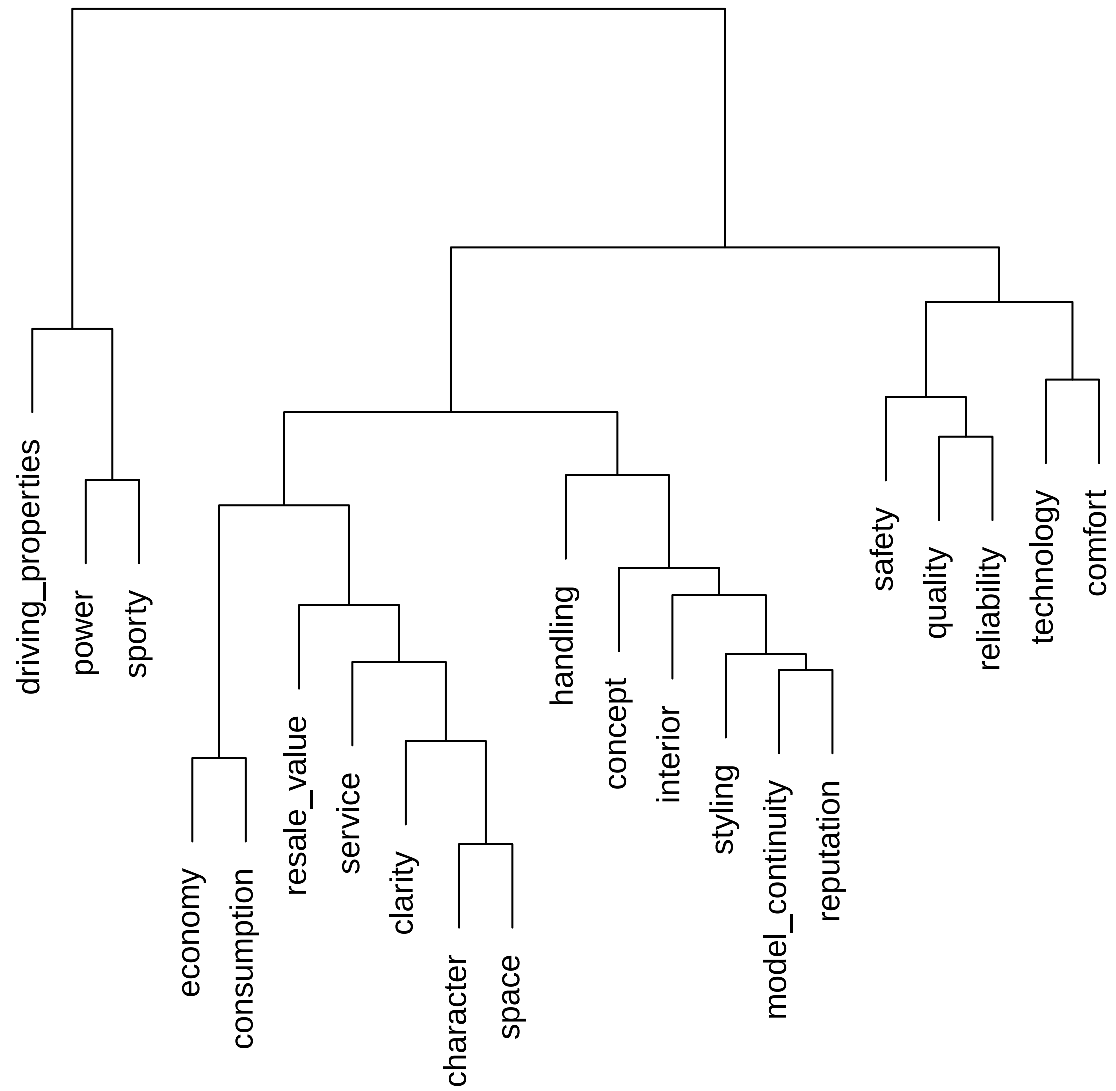
Example: Car Data

A German manufacturer of premium cars asked recent customers about main motivation to buy one of their cars. Binary data, respondents were asked to check properties like sporty, power, interior, safety, quality, resale value, etc.

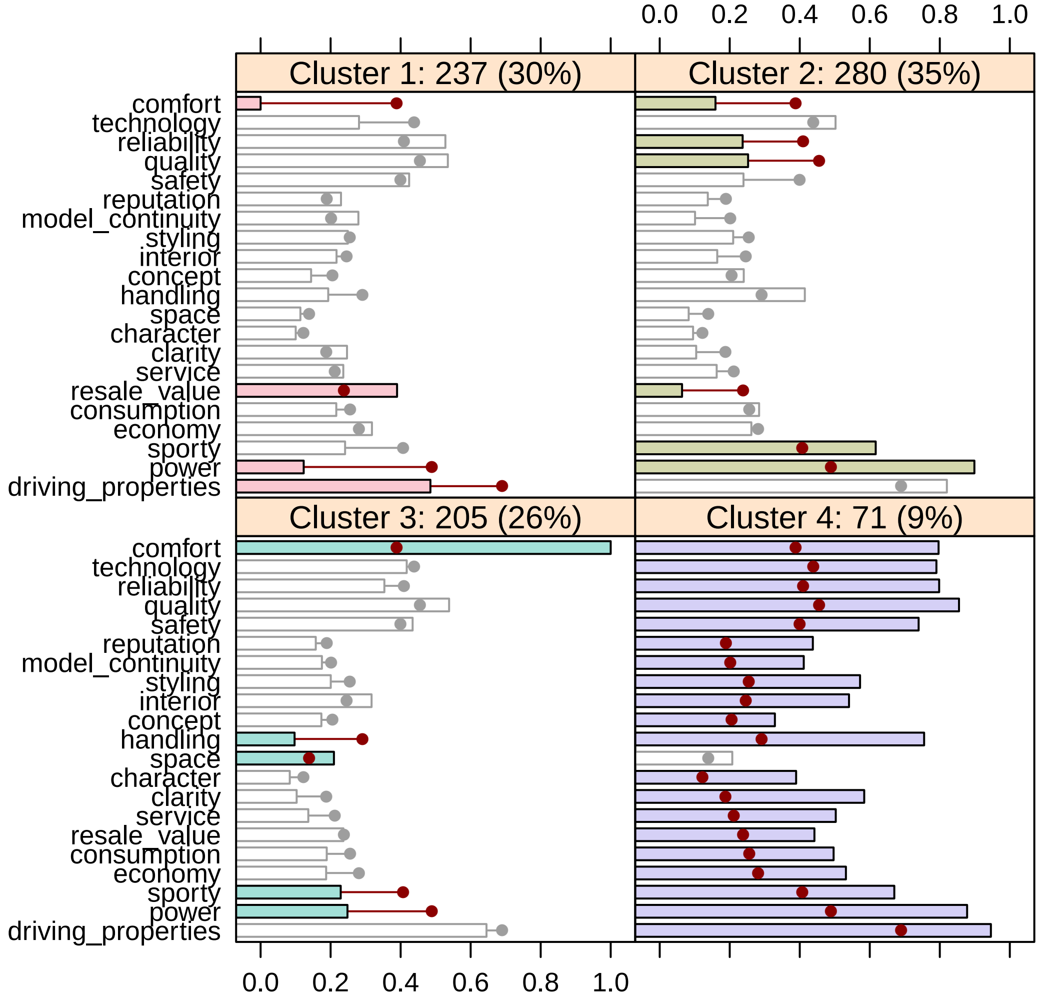
Exploration of the data shows no natural groups, a segmentation of the market imposes a partition on the data (which is absolutely fine for the purpose).

Here: hierarchical clustering of variables, partition with 4 clusters from neural gas algorithm for customers.

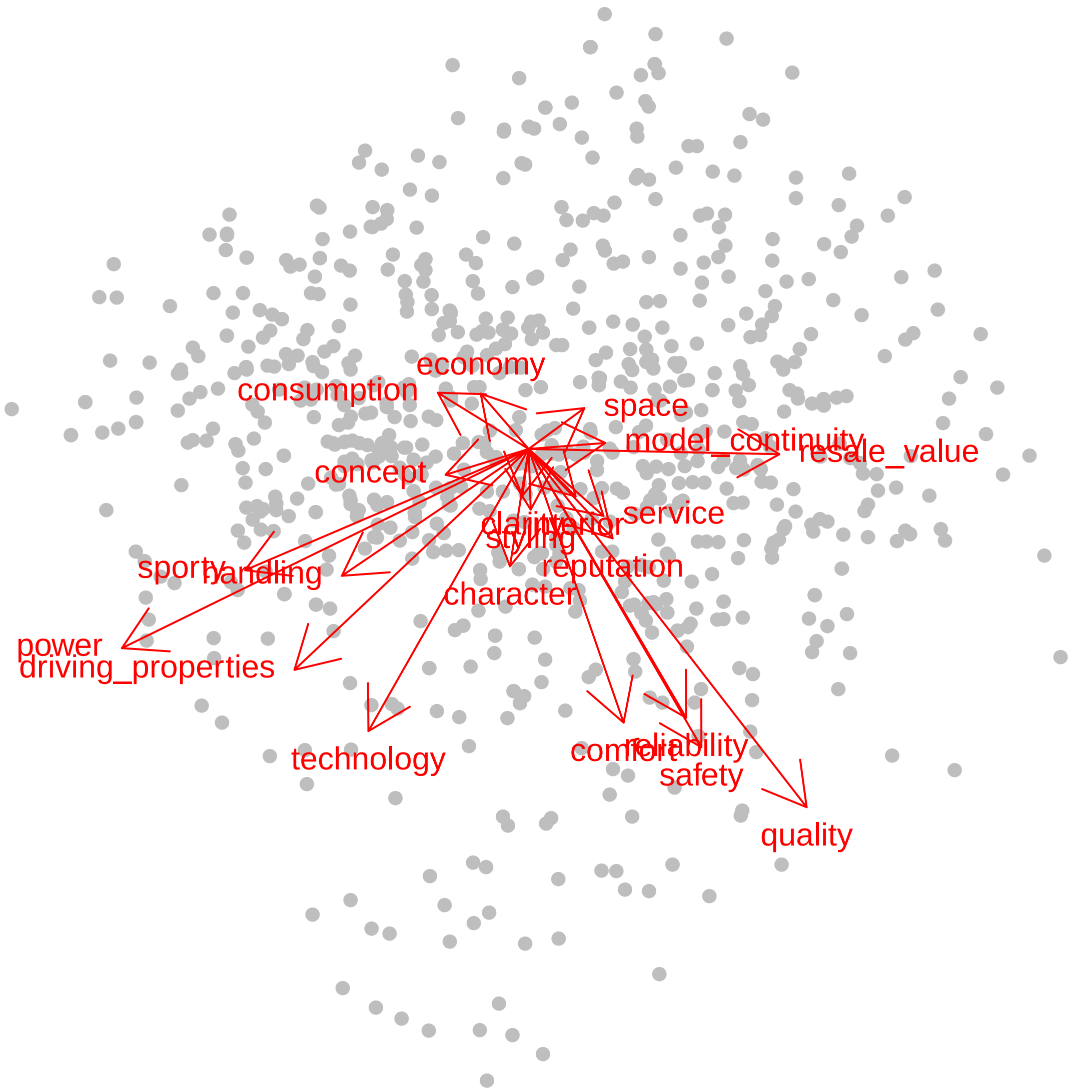
Example: Car Data



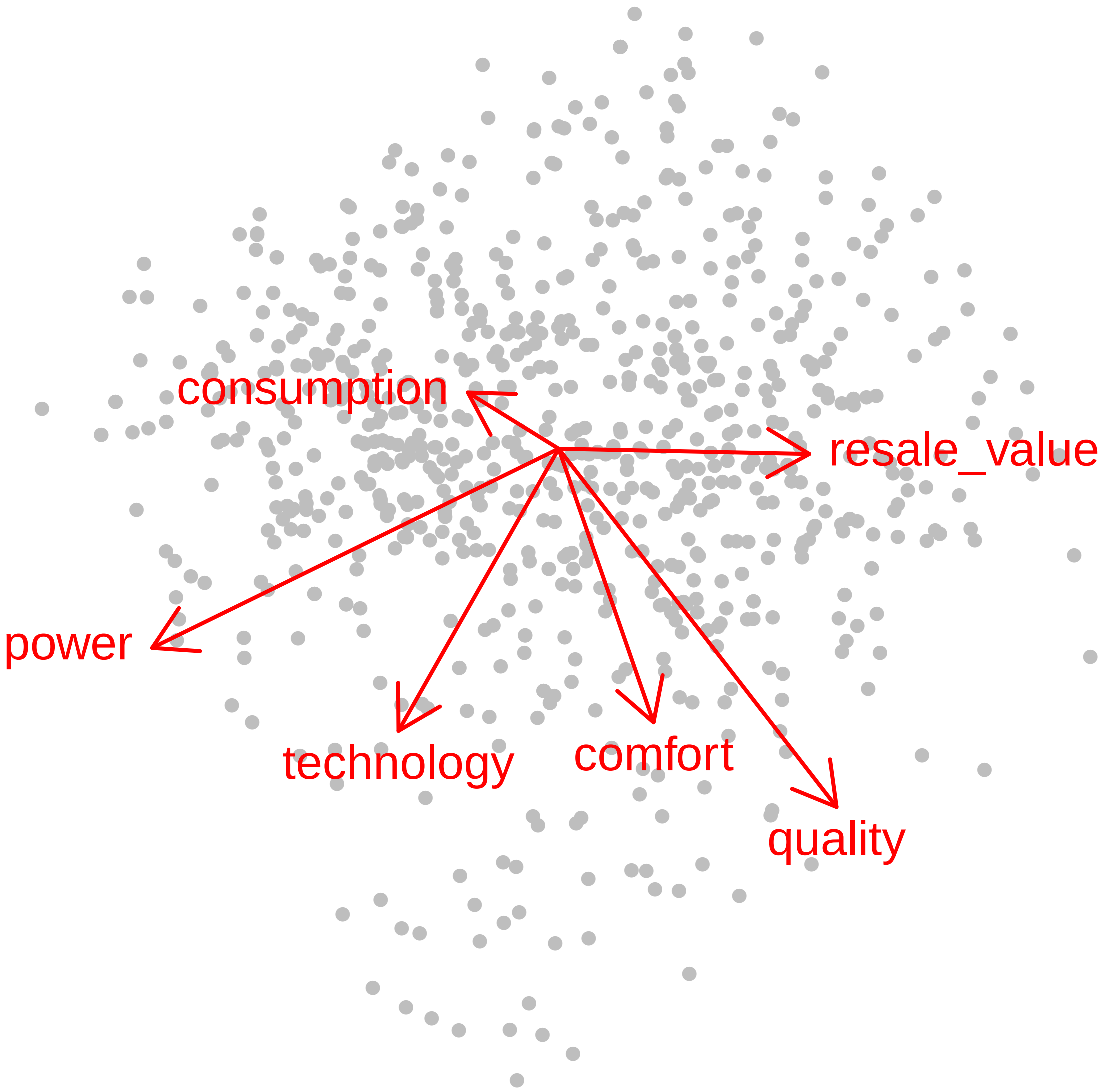
Example: Car Data



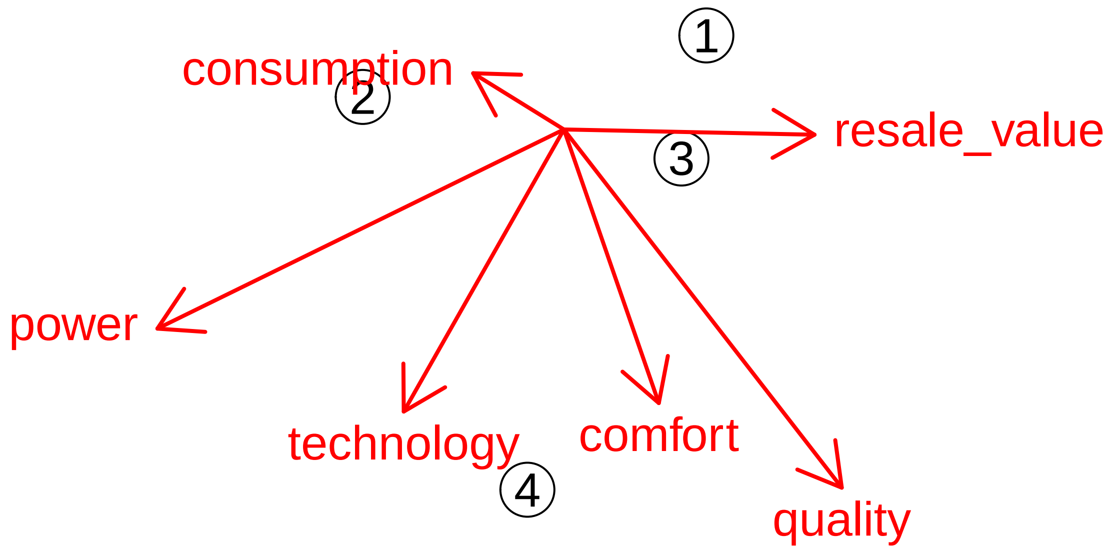
PCA Projection



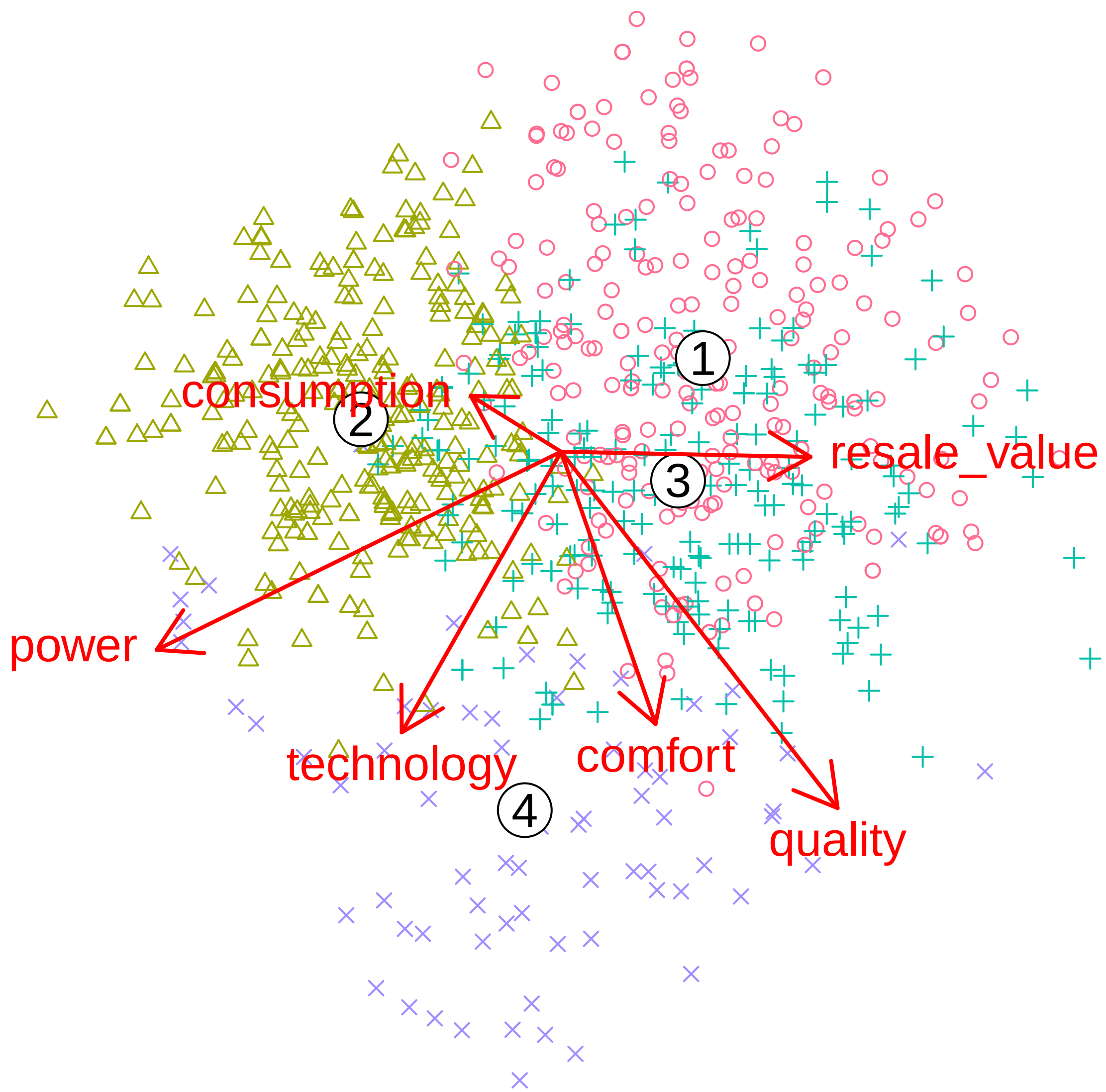
PCA Projection



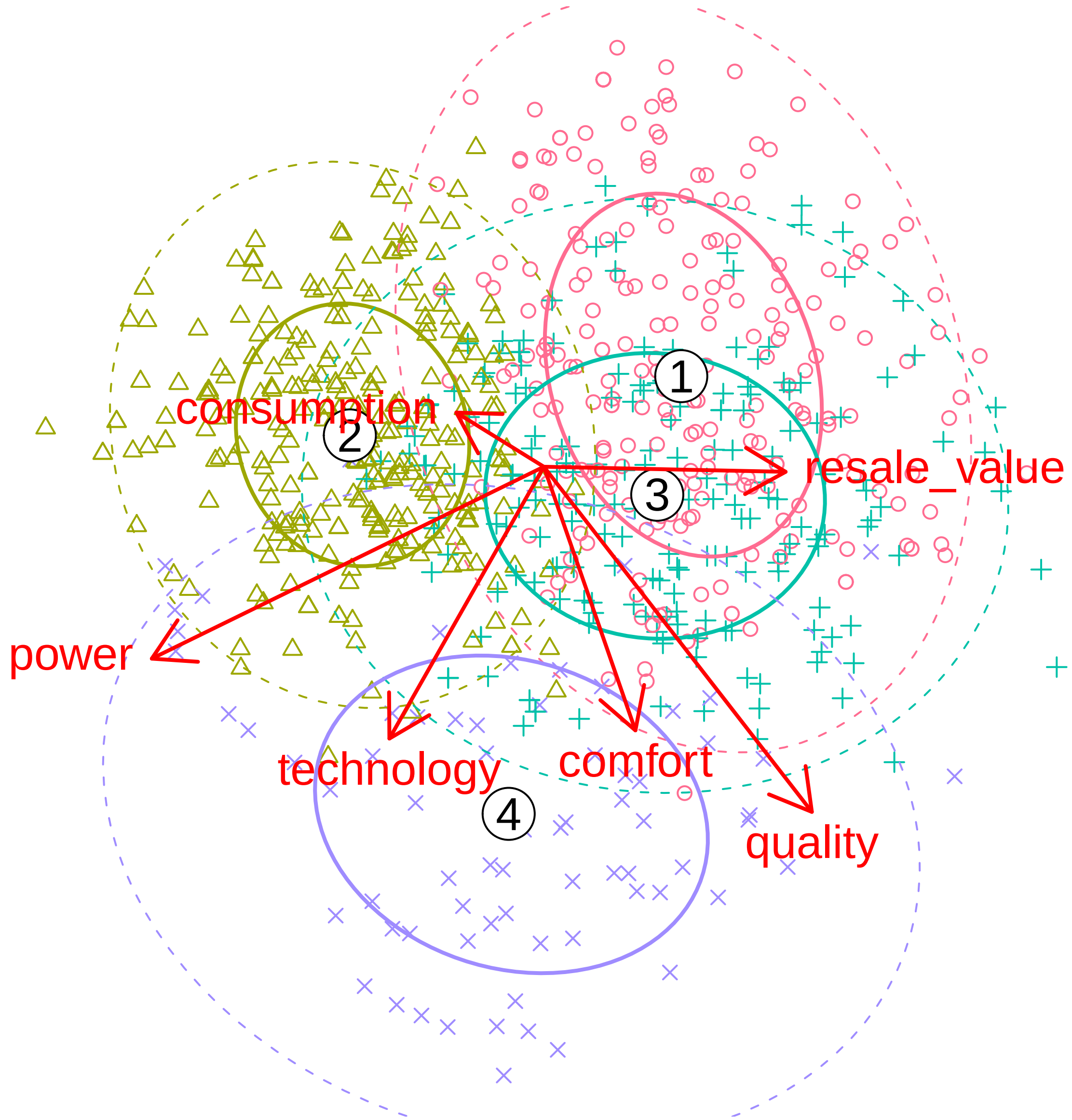
PCA Projection



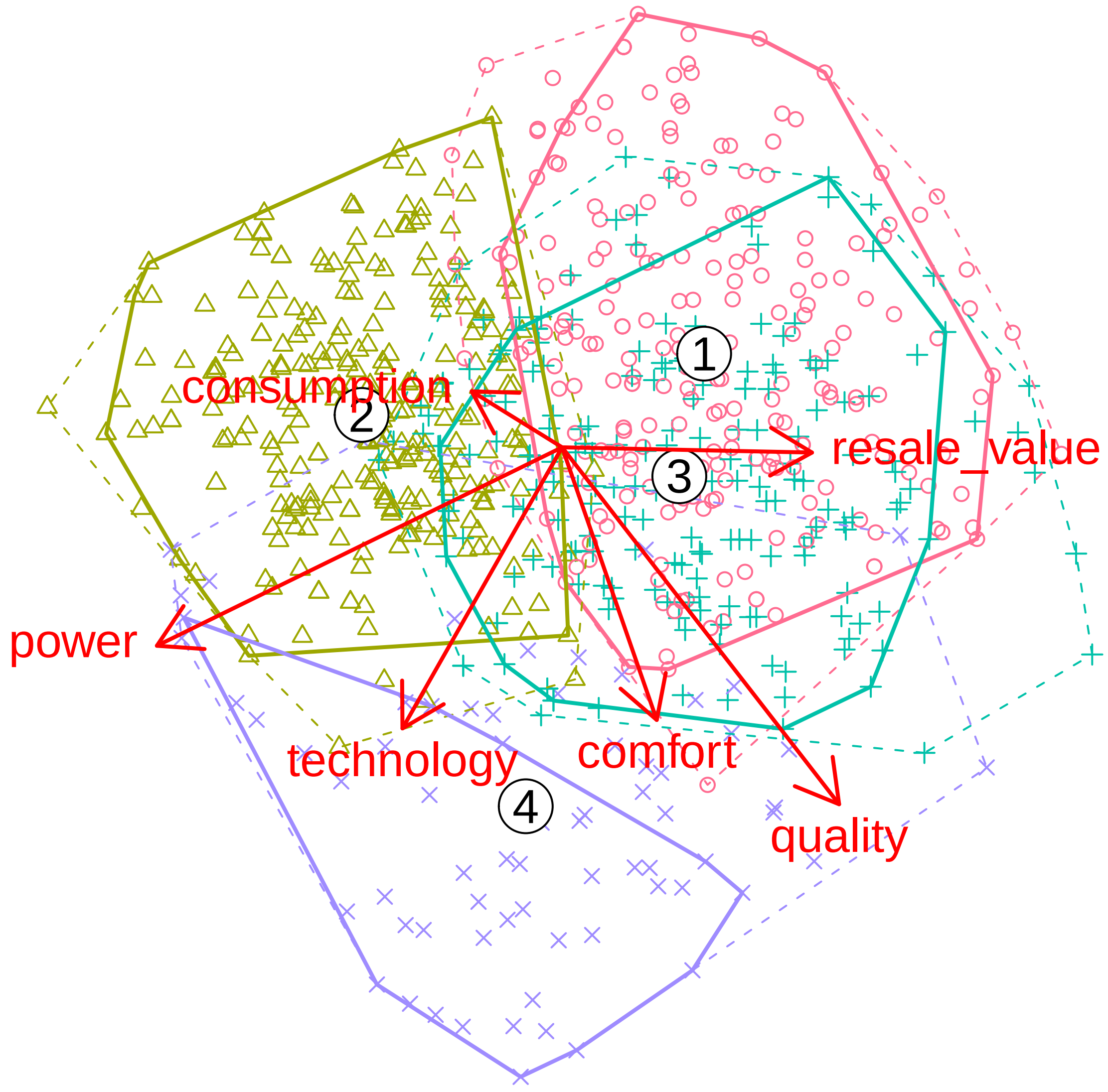
PCA Projection



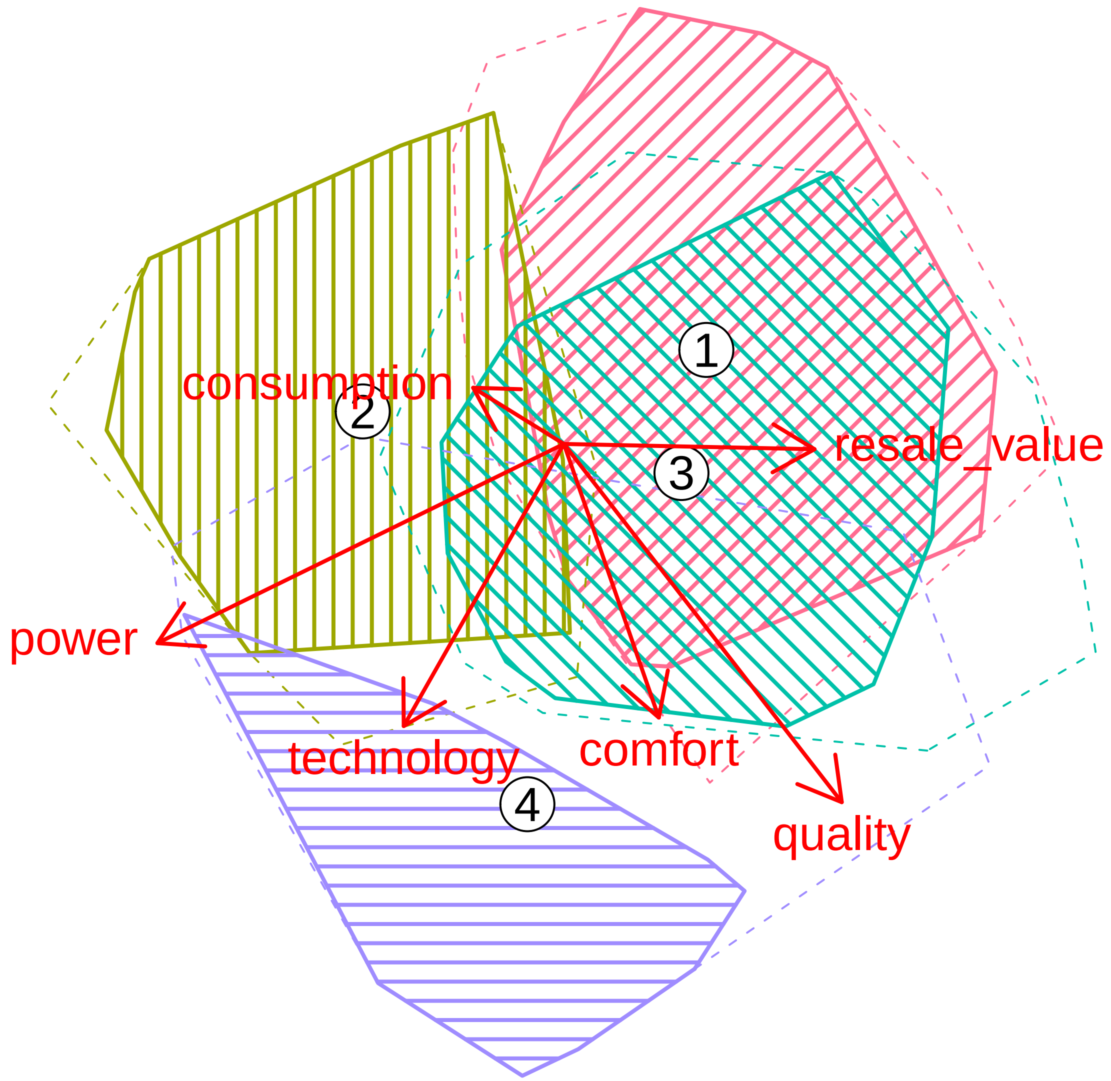
PCA Projection



PCA Projection



PCA Projection



Convex Cluster Hulls

The main problem for 2-dimensional generalizations of boxplots is that there is no total ordering of $\mathbb{R}^2$ (or higher).

For data partitioned using a centroid-based cluster algorithm there is a natural total ordering for each point in a cluster: The distance $d(\mathbf{x}, c(\mathbf{x}))$ of the point to its respective cluster centroid. Let A_k be the set of points in cluster k, and

$$m_k = \text{median}\{d(\mathbf{x}_n, \mathbf{c}_k) | \mathbf{x}_n \in A_k\}$$

inner area: all data points where $d(\mathbf{x}_n, \mathbf{c}_k) \leq m_k$

outer area: all data points where $d(\mathbf{x}_n, \mathbf{c}_k) \leq 2.5 m_k$

Shadow Values

Second-closest centroid to $\mathbf{x}$:

$$\tilde{c}(\mathbf{x}) = \underset{\mathbf{c} \in C_K \setminus \{c(\mathbf{x})\}}{\text{argmin}} \; d(x, \mathbf{c})$$

Shadow value:

$$s(\mathbf{x}) = \frac{2d(\mathbf{x}, c(\mathbf{x}))}{d(x, c(\mathbf{x})) + d(\mathbf{x}, \tilde{c}(\mathbf{x}))} \in [0, 1]$$

$s(x) = 0$: centroid

$s(x) = 1$: on border of clusters

Neighborhood Graph

Let

$$A_{ij} = \left\{\mathbf{x}_n \,|\, c(\mathbf{x}_n) = \mathbf{c}_i,\, \tilde{c}(\mathbf{x}_n) = \mathbf{c}_j\right\}$$

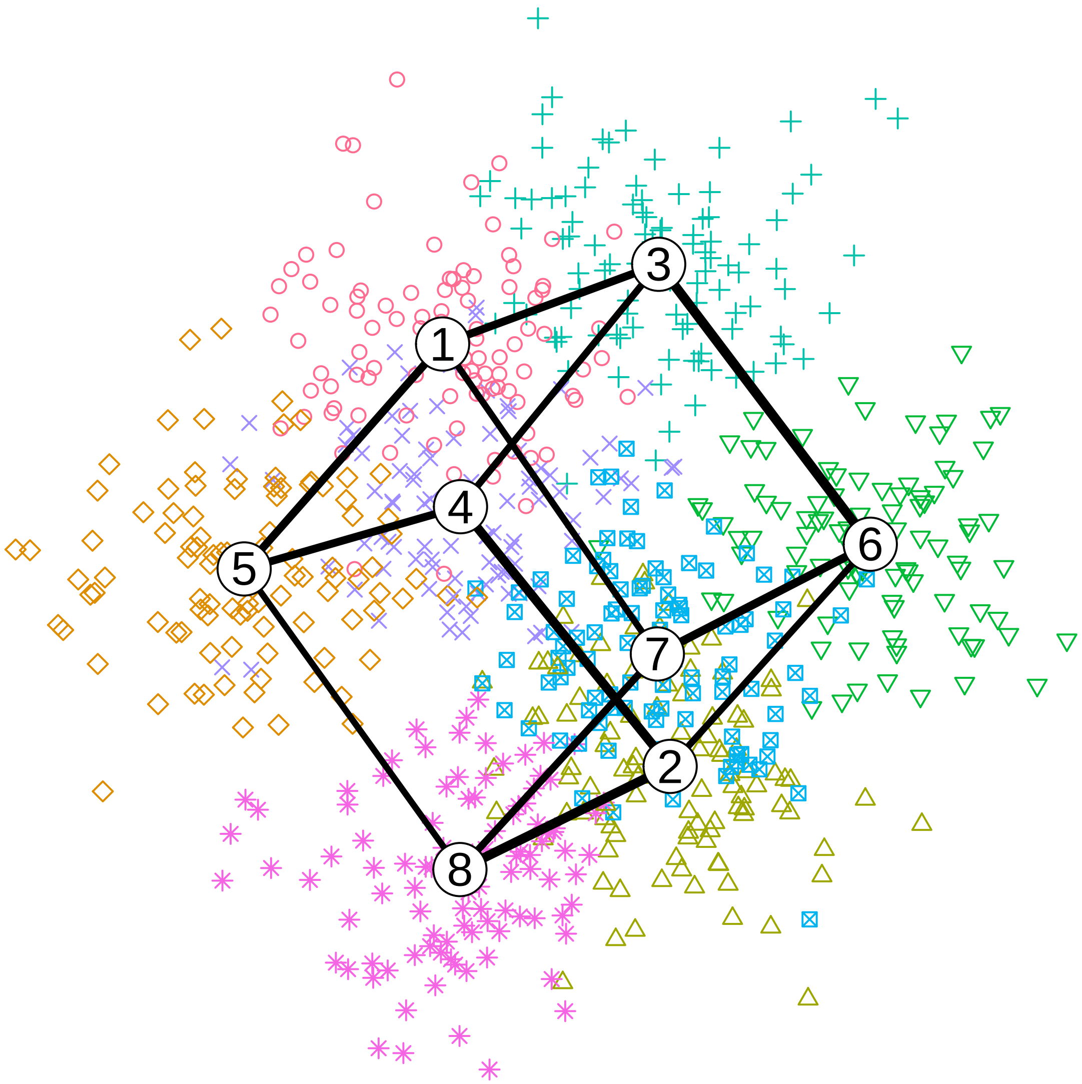
be the set of all points where $\mathbf{c}_i$ is the closest centroid and $\mathbf{c}_j$ is second-closest.

Cluster similarity:

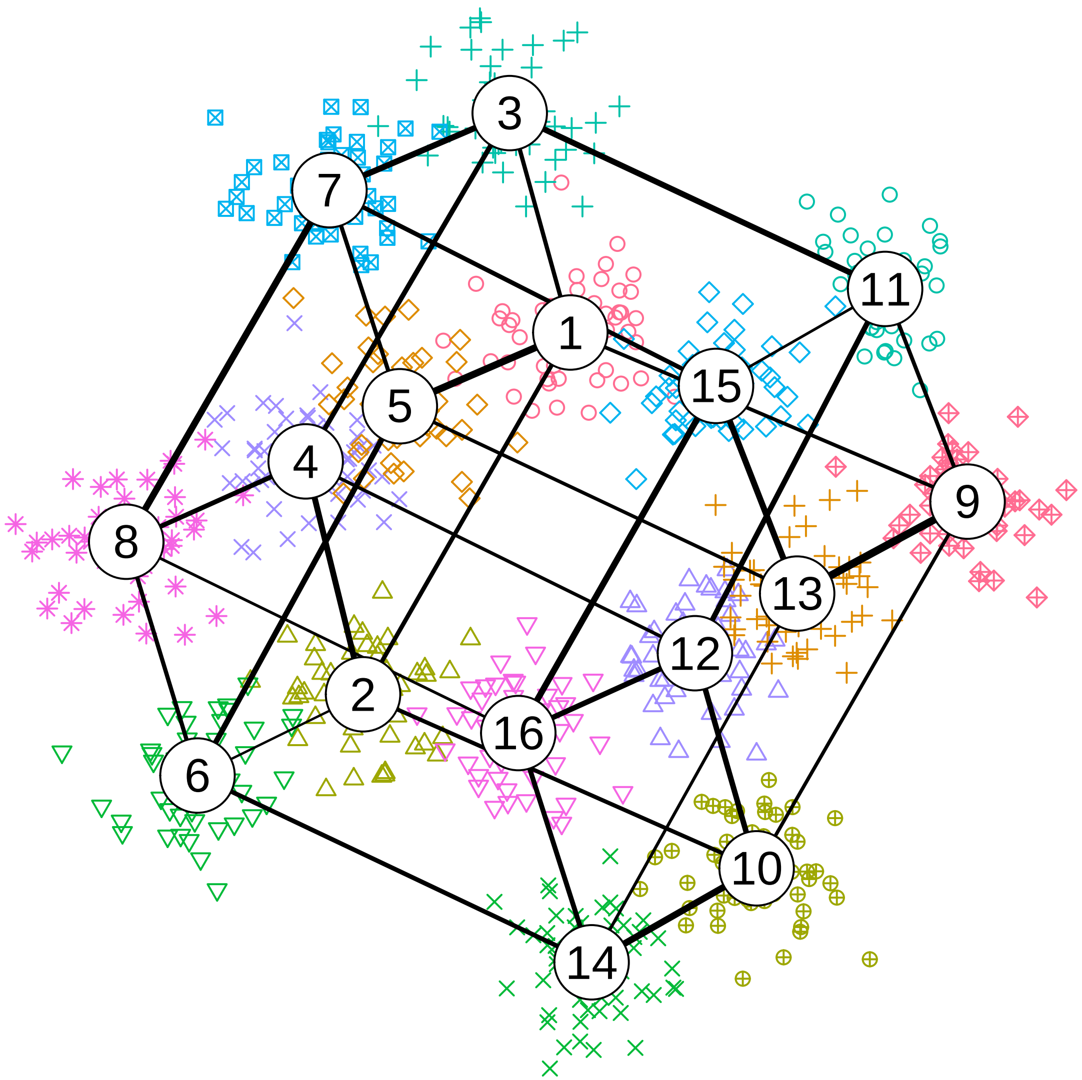
$$s_{ij} = \begin{cases} |A_i|^{-1} \sum_{x \in |A_{ij}|} s(\mathbf{x}), & A_{ij} \neq \emptyset \\ 0, & A_{ij} = \emptyset \end{cases}$$

Use $s_{ij} + s_{ji}$ for thickness of line connecting $\mathbf{c}_i$ and $\mathbf{c}_j$.

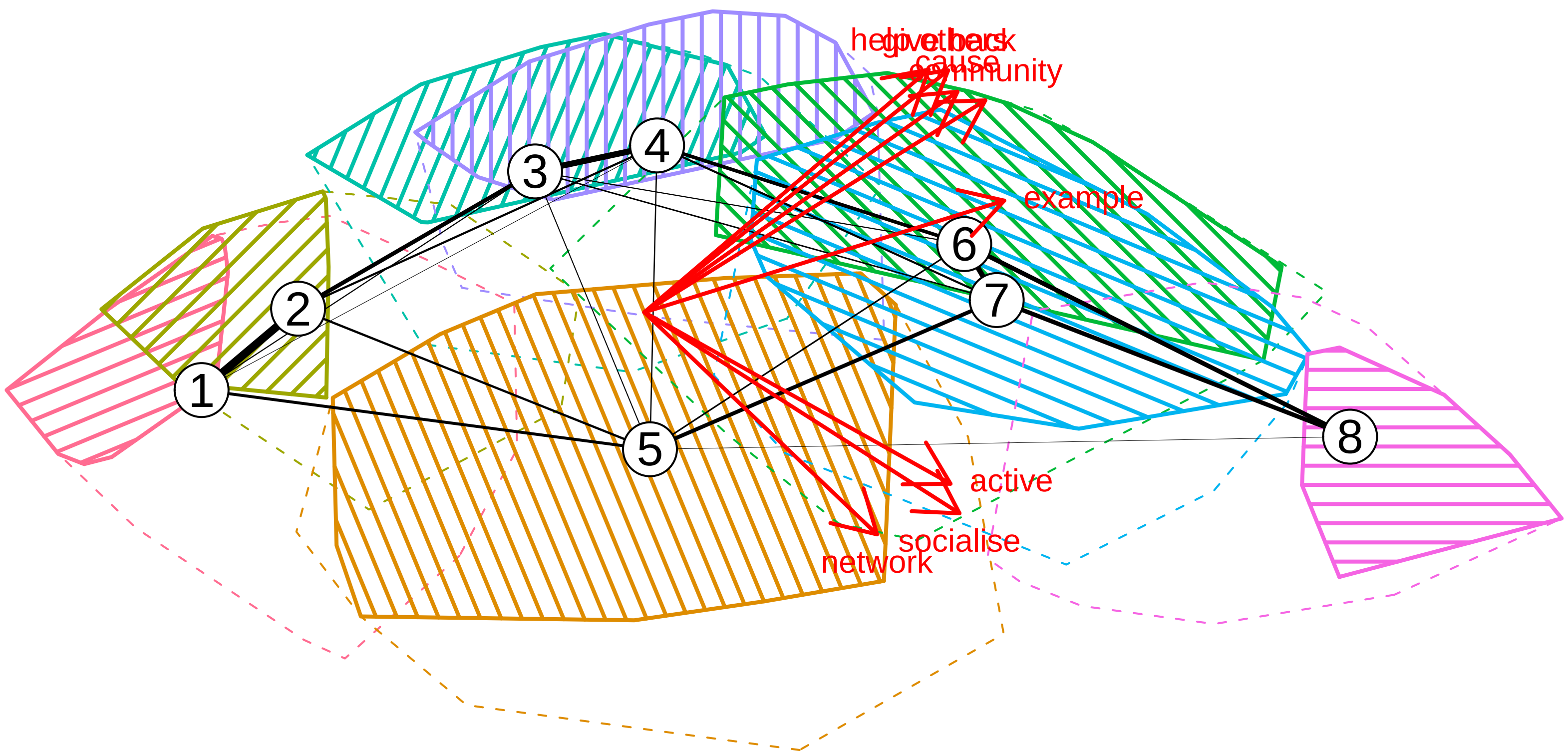
Neighborhood Graph



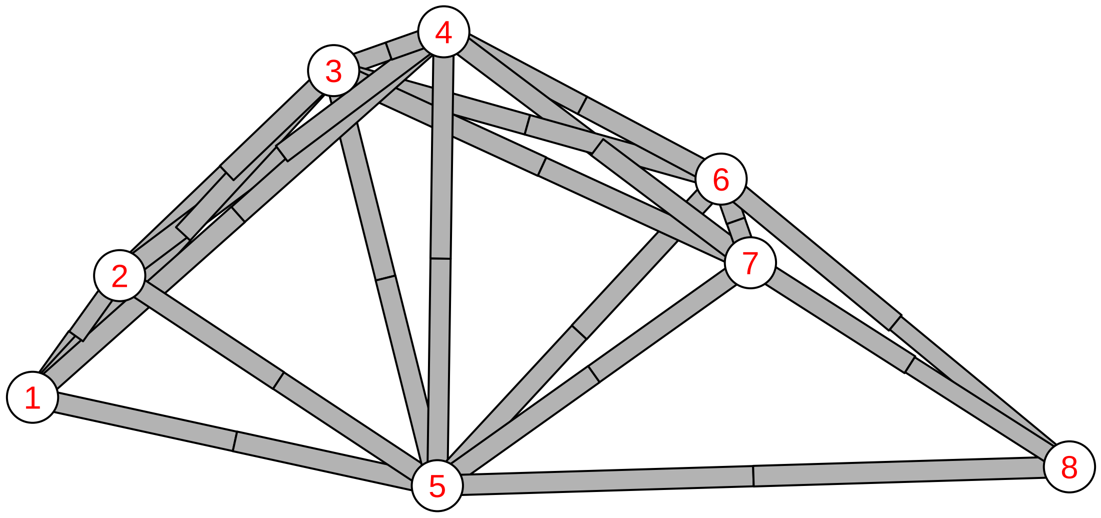
Neighborhood Graph



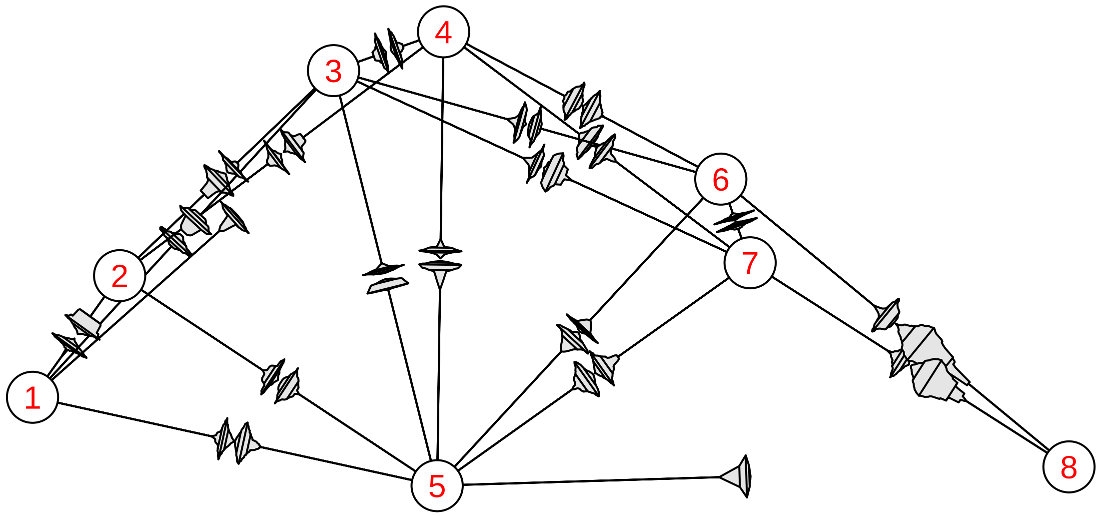
Volunteer PCA-Projection



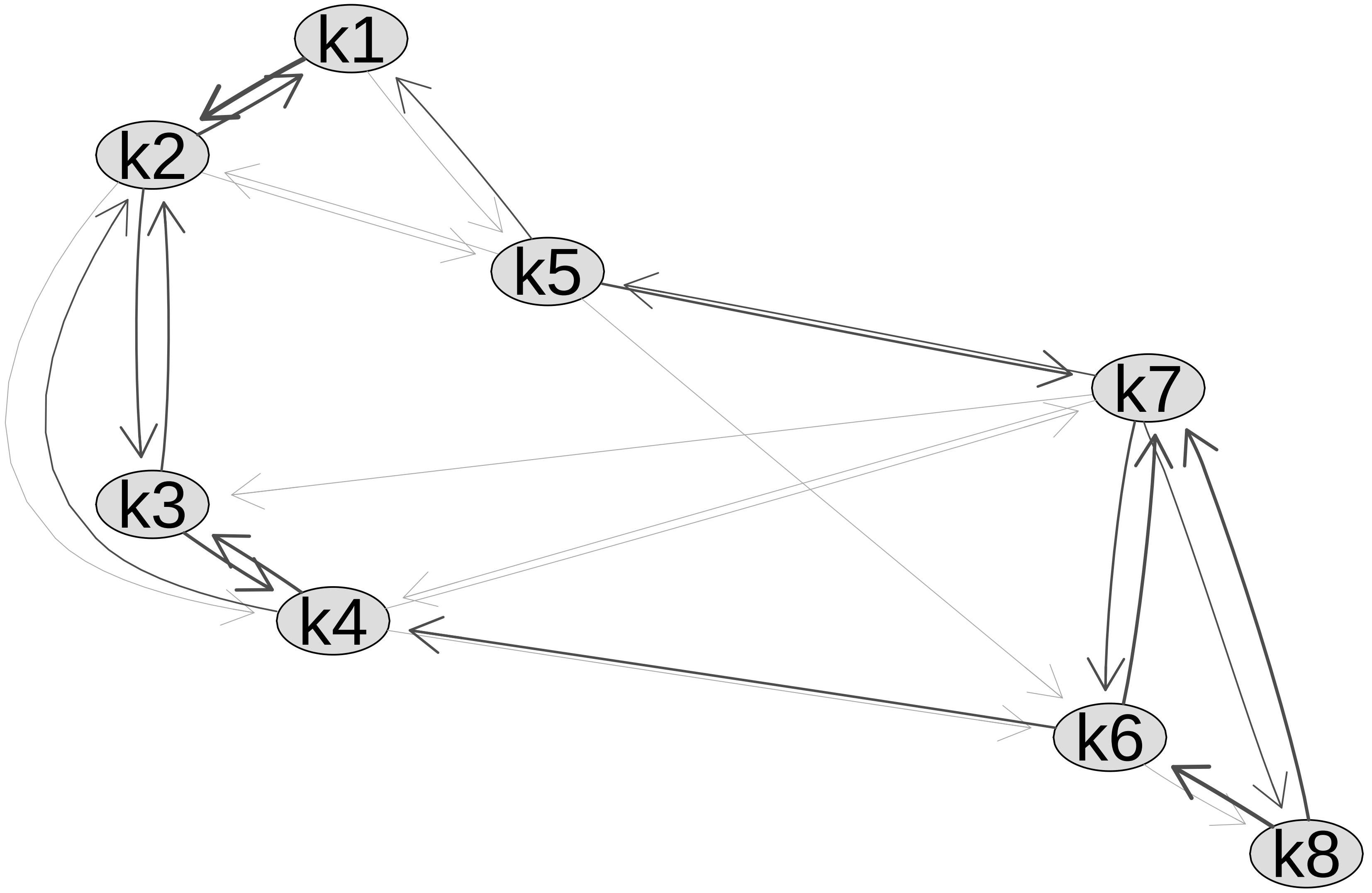
Volunteer PCA-Projection



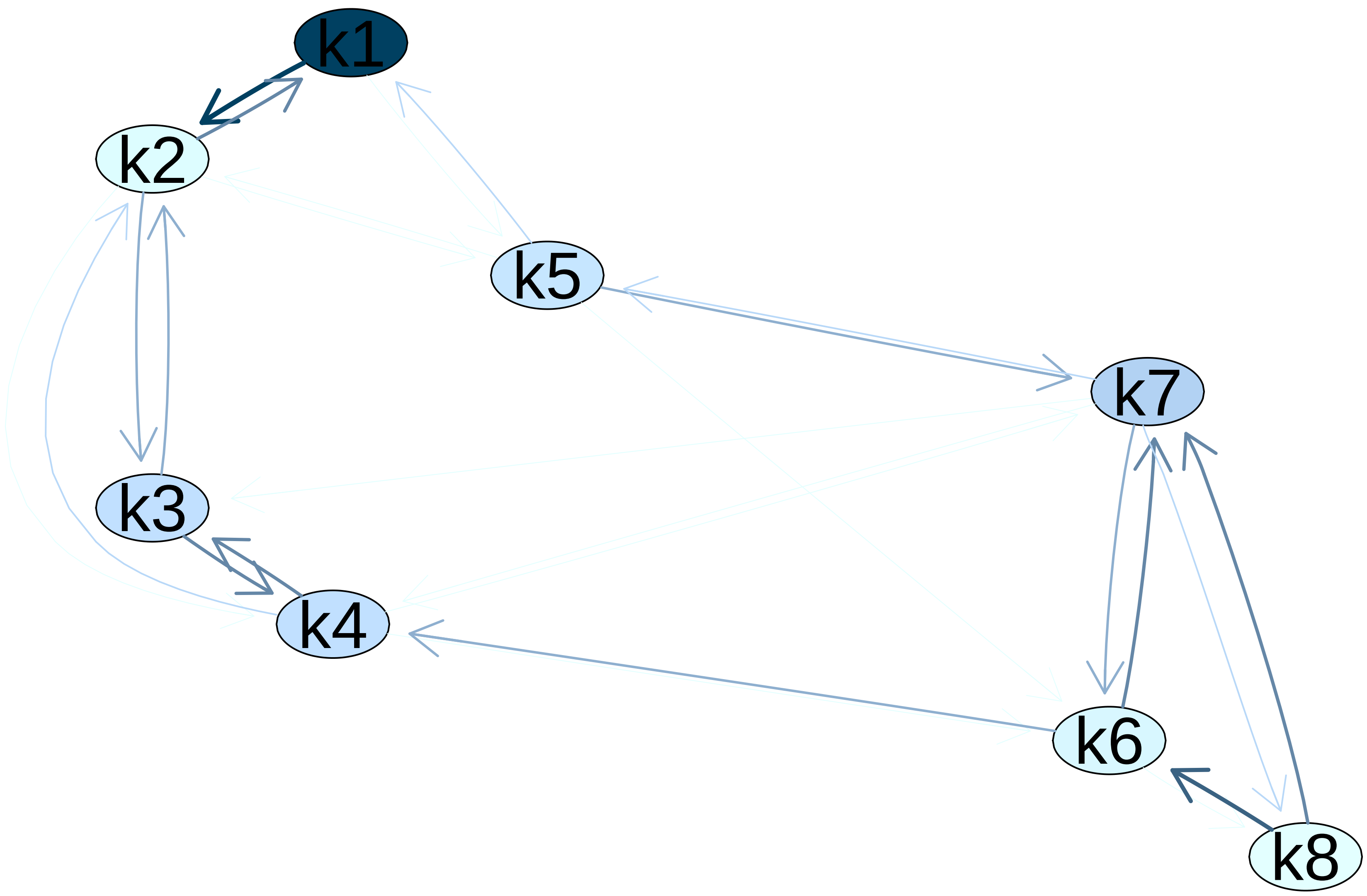
Volunteer PCA-Projection



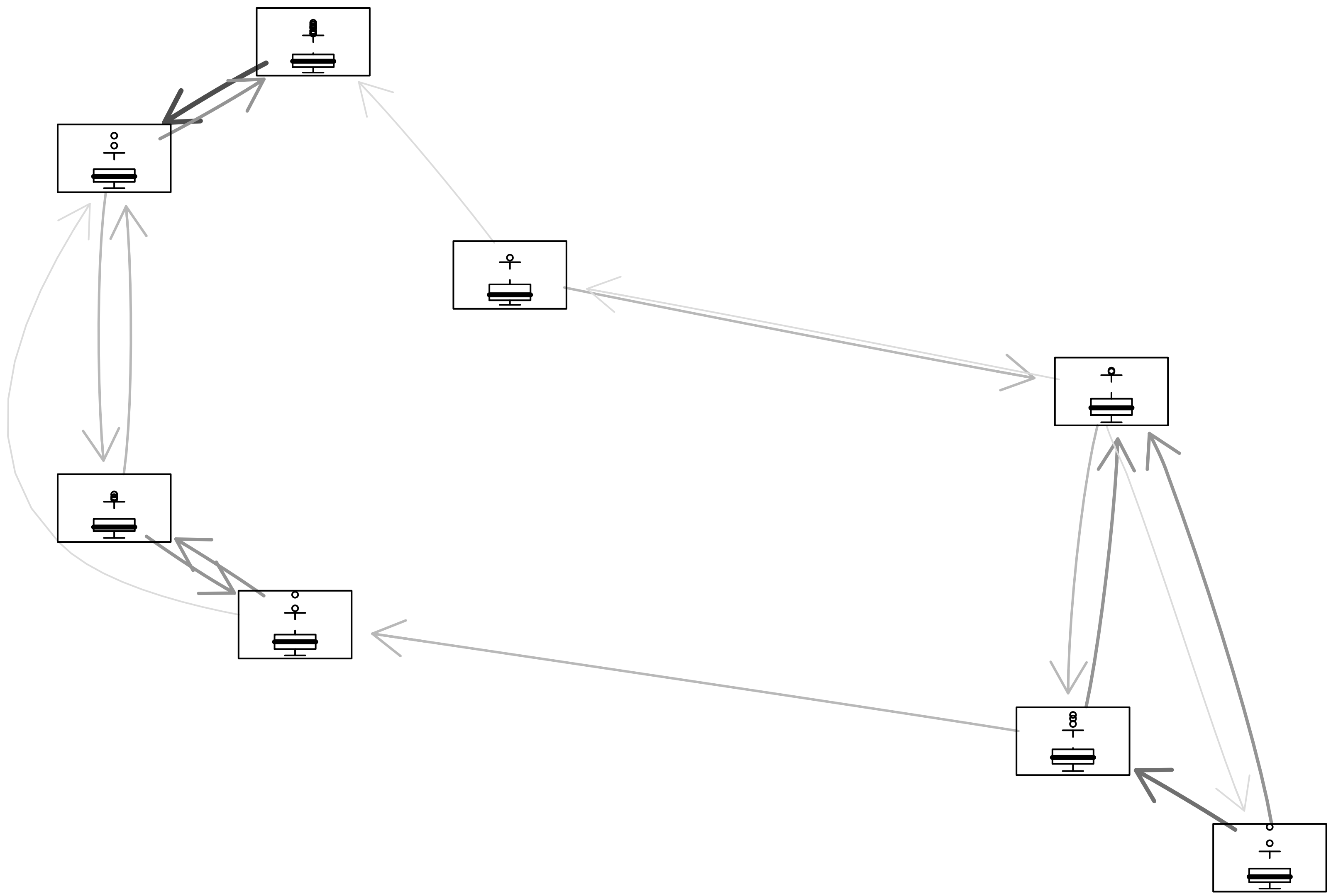
Nonlinear Arrangement



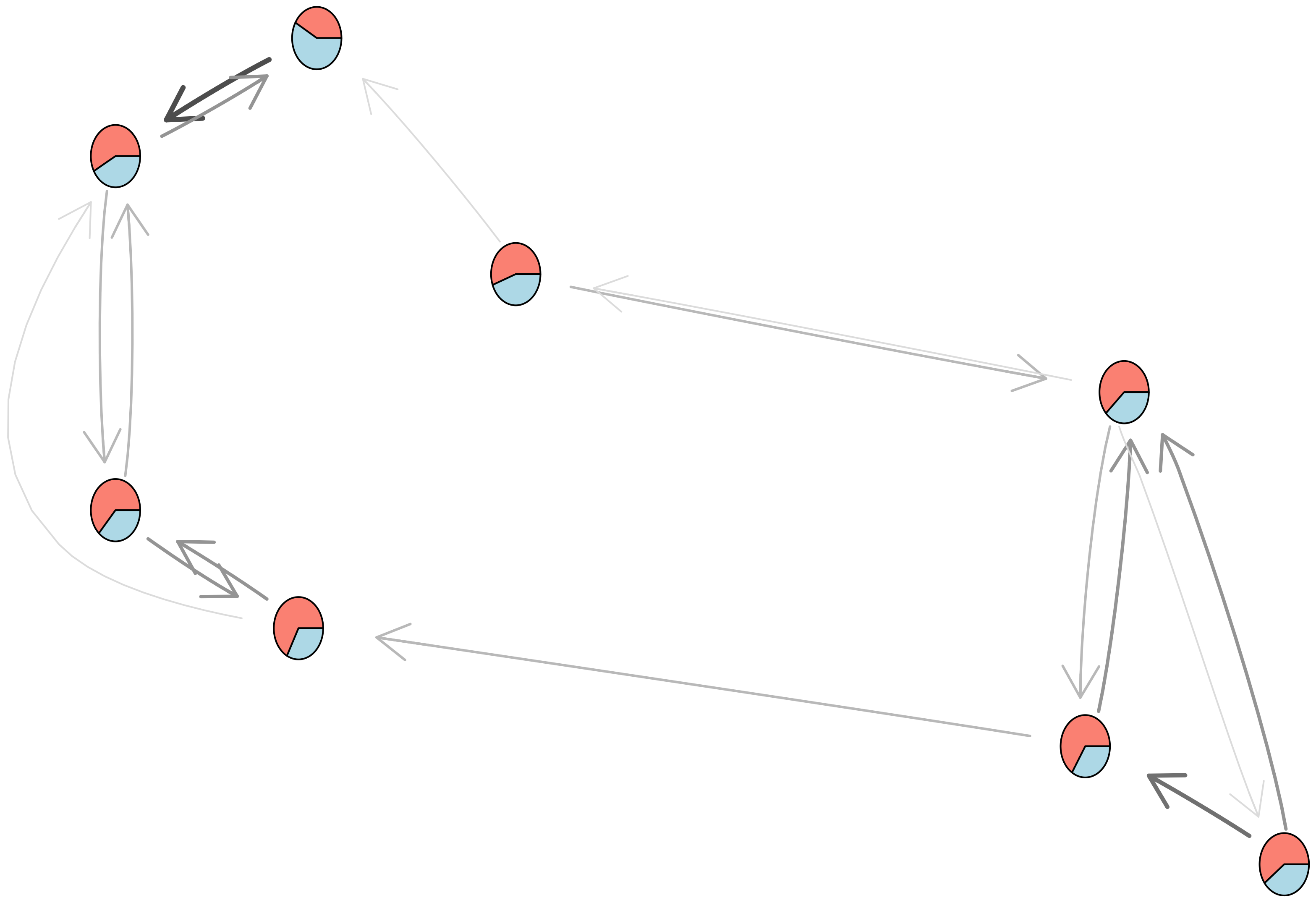
Cluster Size



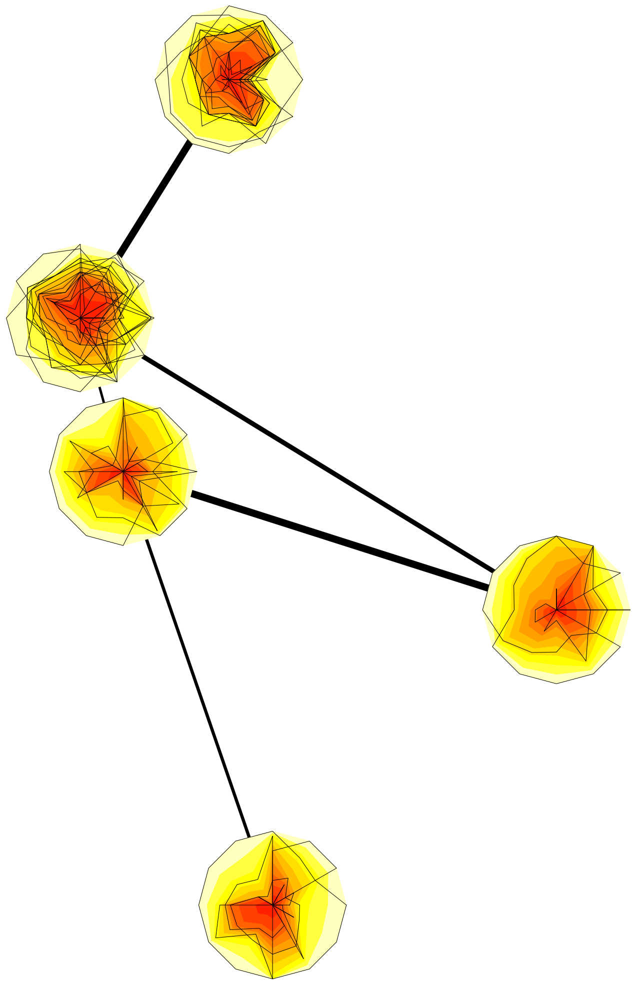
Age Distribution



Gender Distribution



Nice to Look at



Model-based Ordinal Clustering

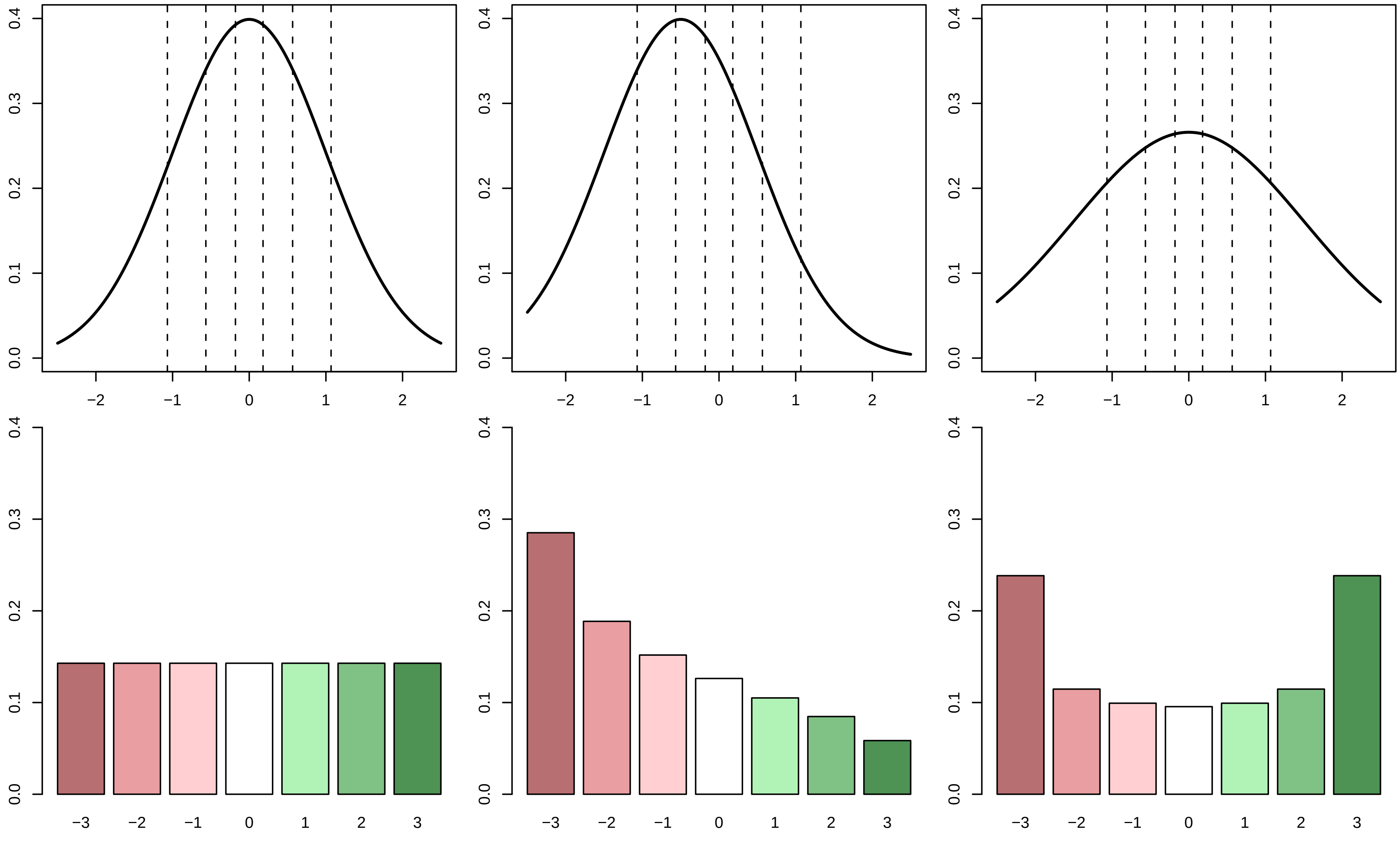
Survey data for two fastfood chains (McDonald's, Subway) from 715 respondents on 10 items (yummy, fattening, greasy, fast, . . .).

We are interested in capturing scale usage heterogeneity under the assumption that different involvement with a brand may provoke different scale usage.

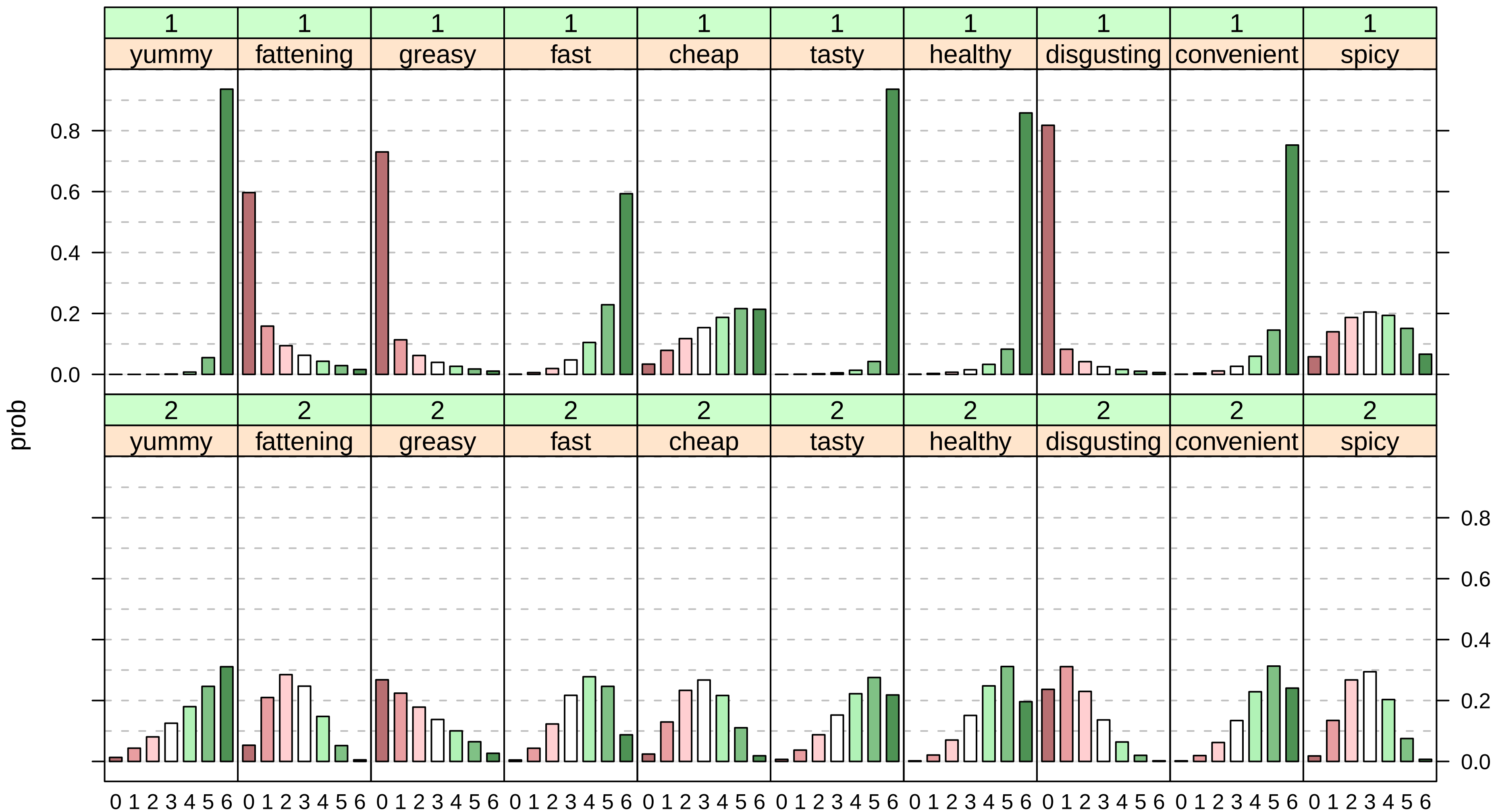
Finite mixture model: For each item and group we estimate mean and standard deviation of a latent Gaussian. Assumption of independence between items given group membership, estimation by EM.

For a 3-component model we get 30 means and 30 standard deviations.

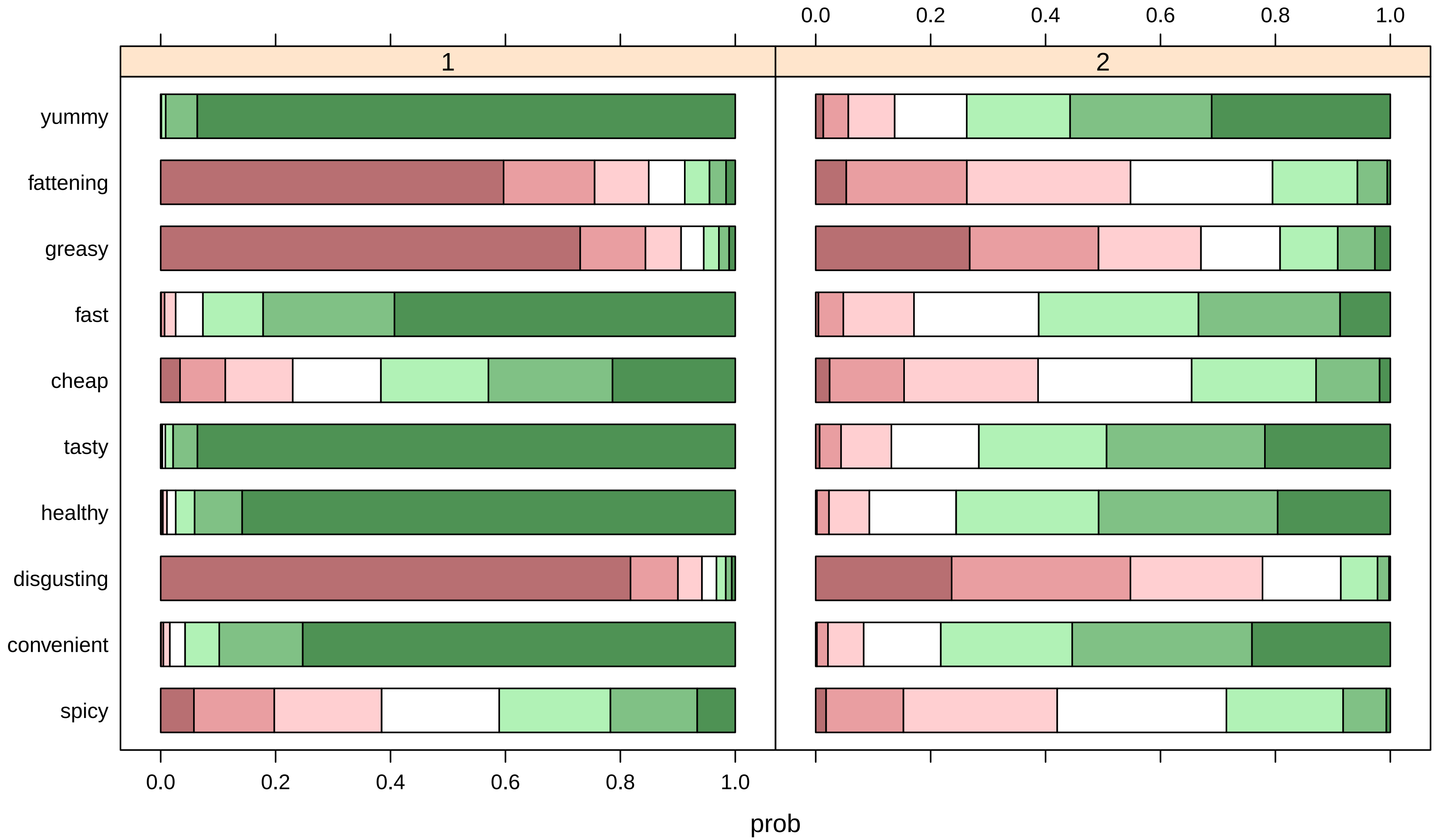
Model-based Ordinal Clustering



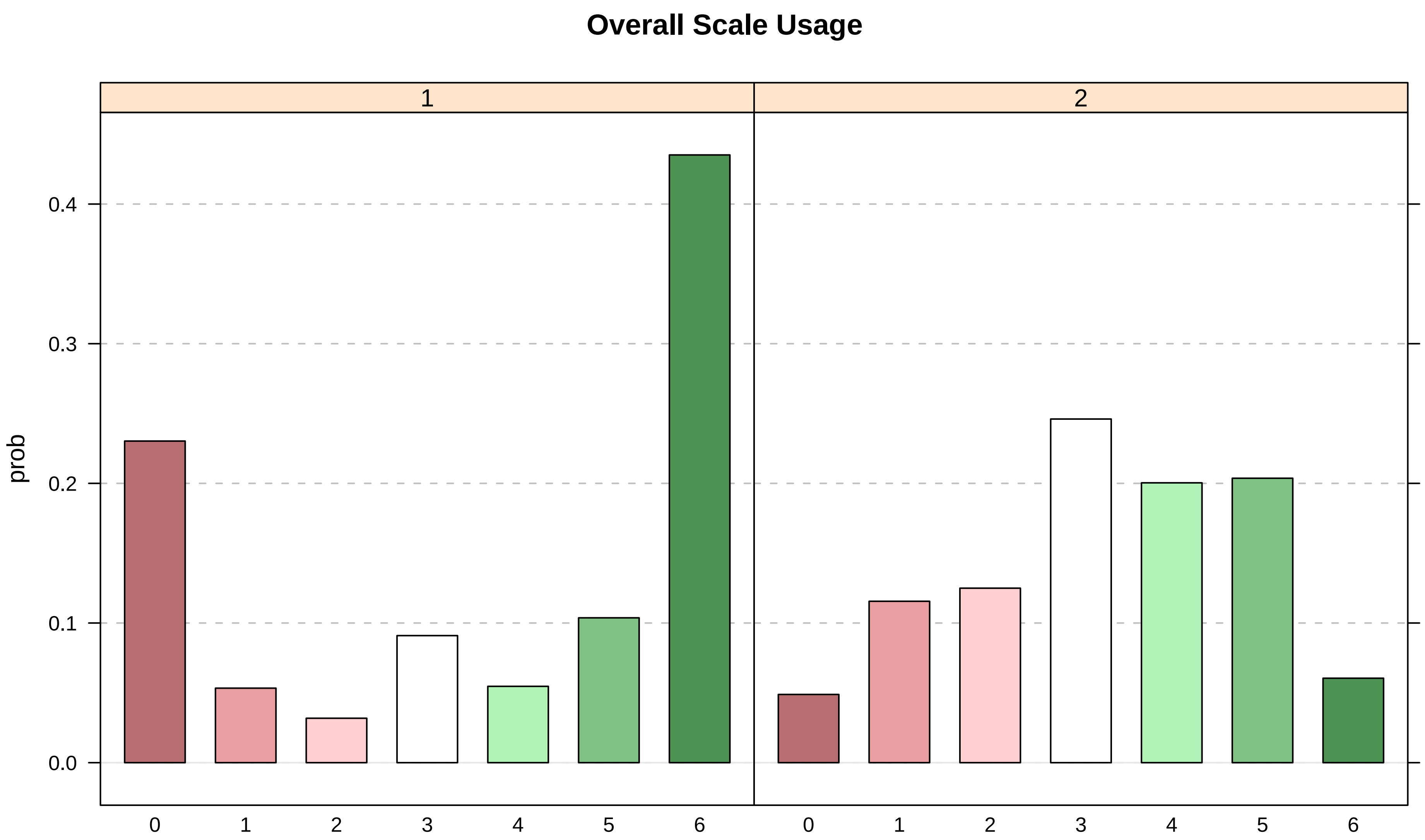
Subway: 2 Components



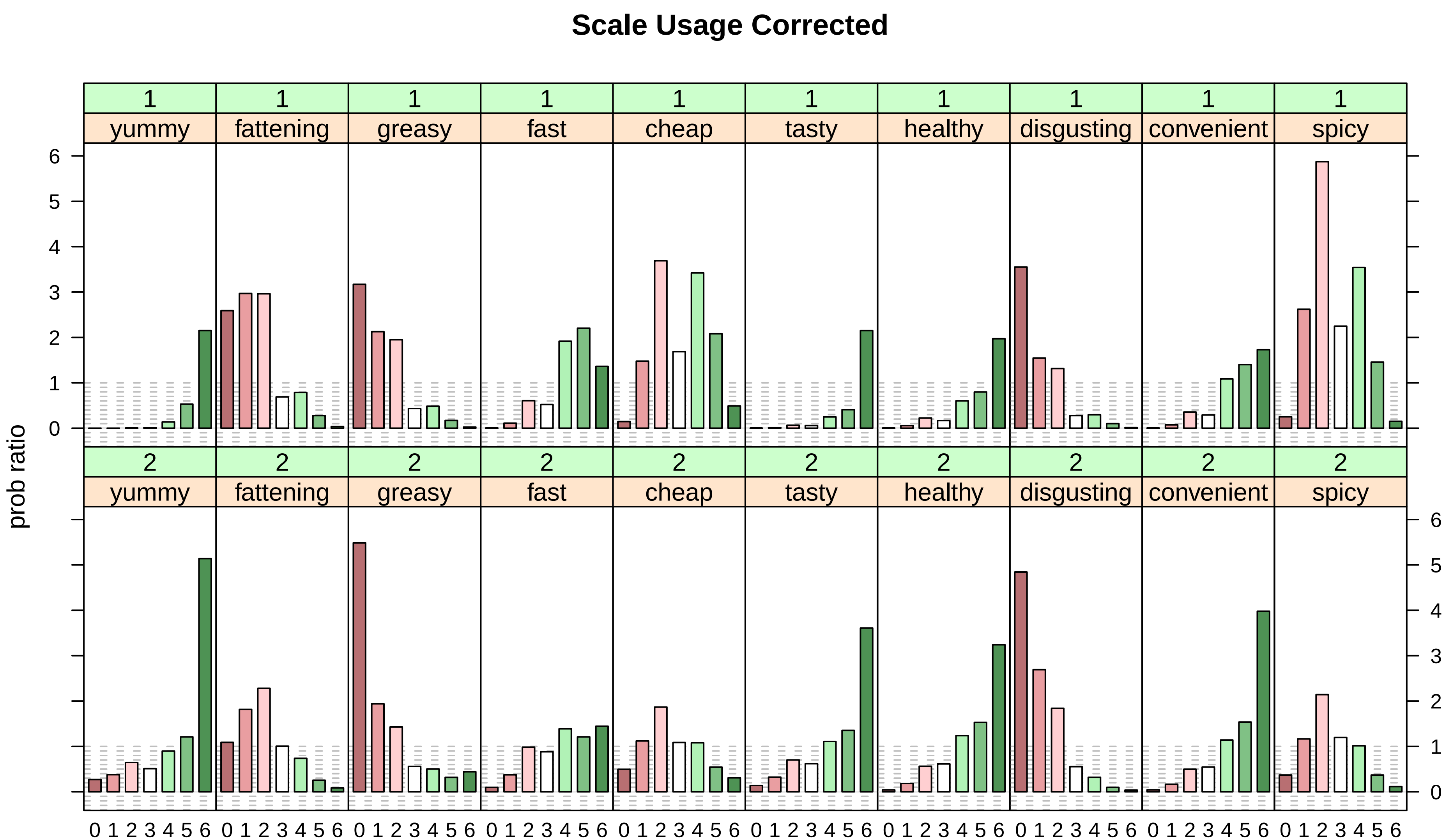
Subway: 2 Components



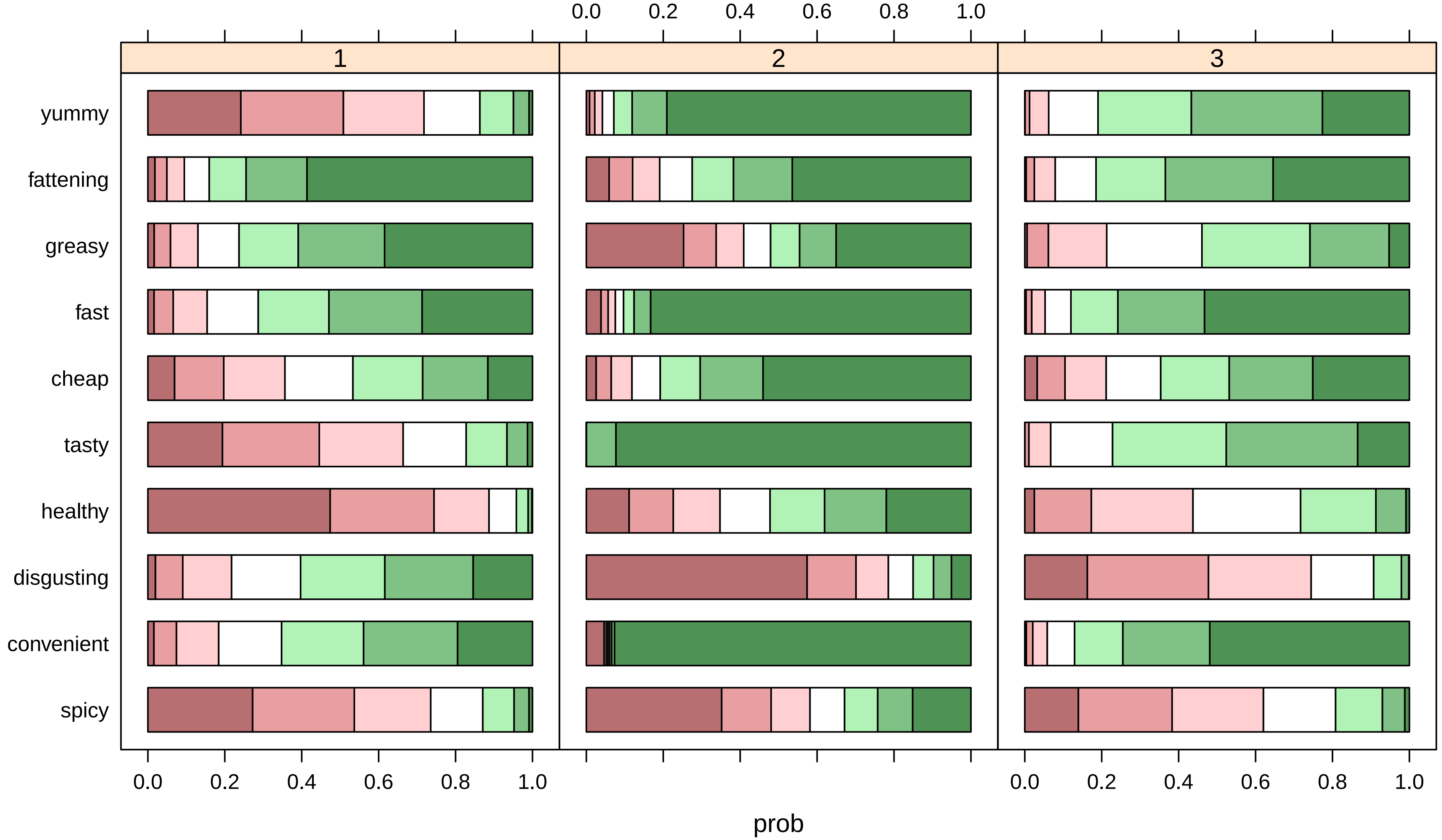
Subway: 2 Components



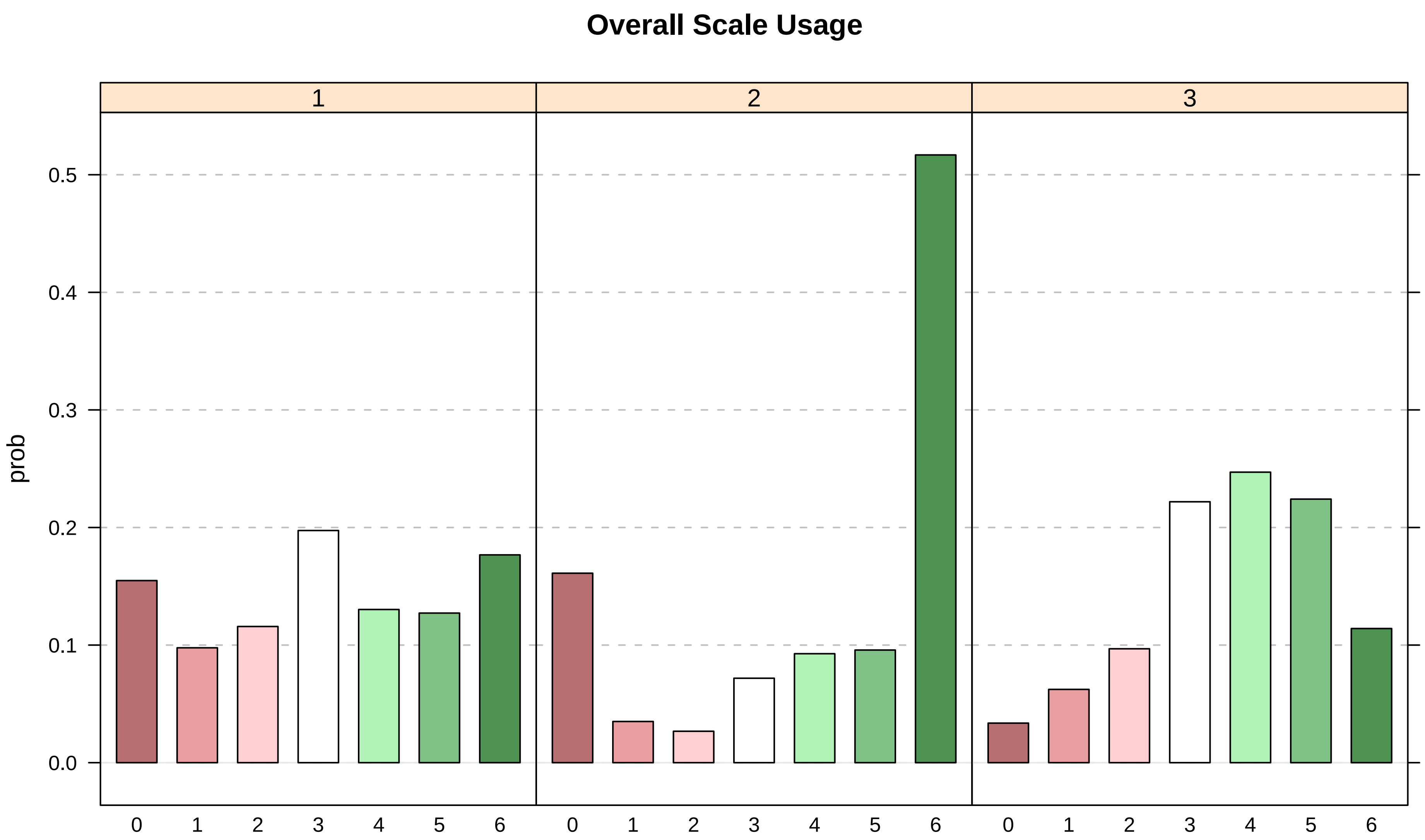
Subway: 2 Components



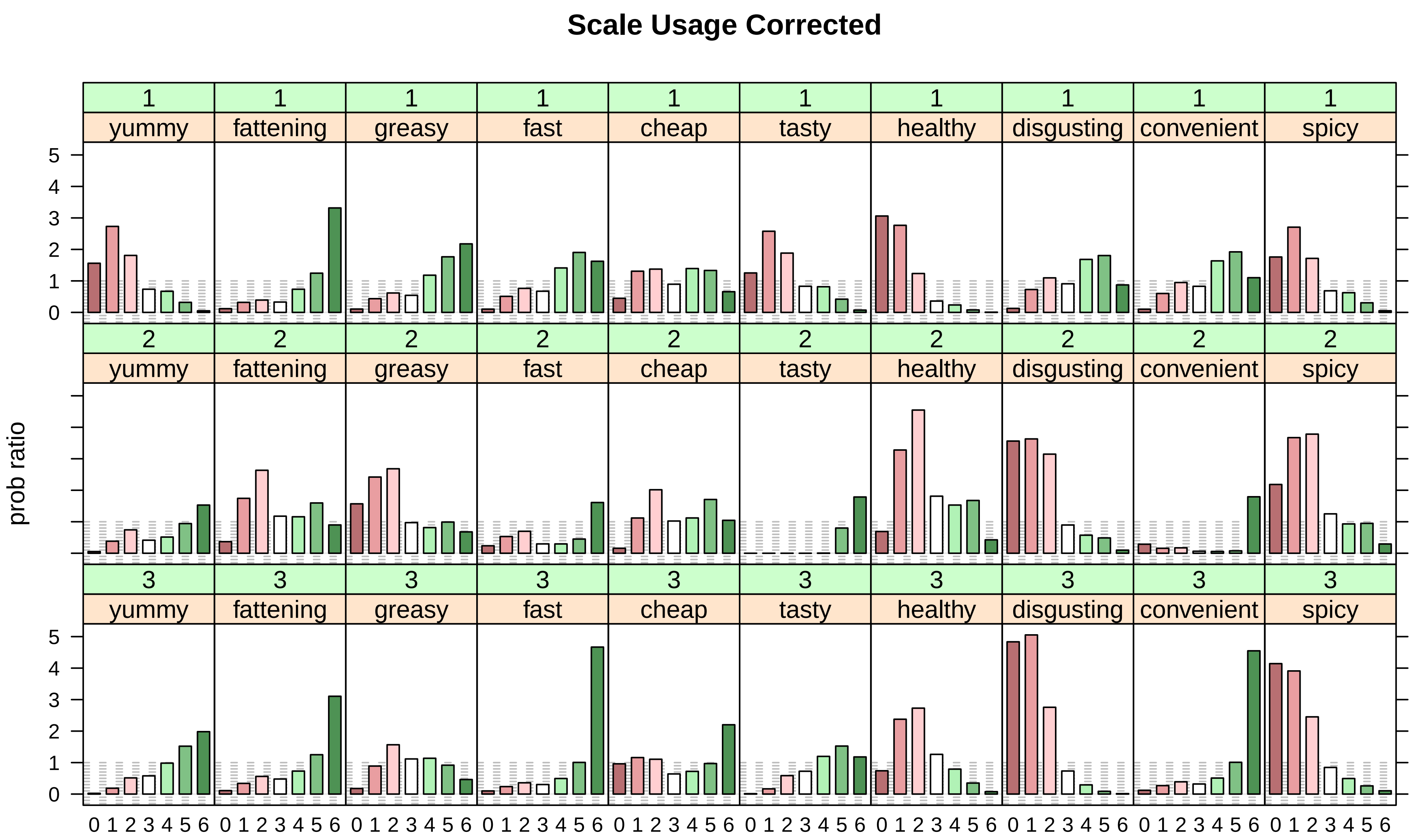
McDonald's: 3 Components



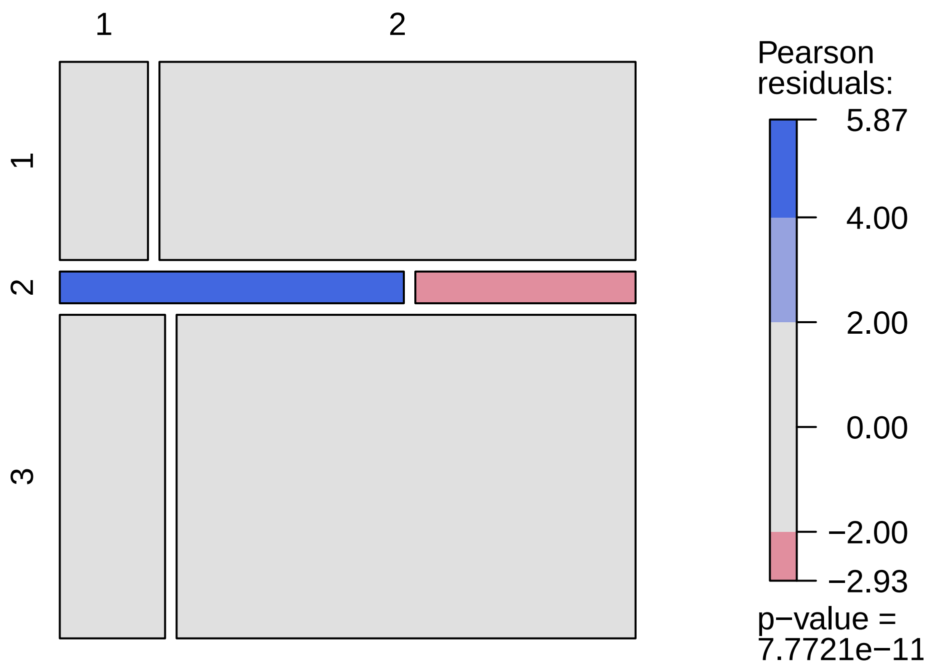
McDonald's: 3 Components



McDonald's: 3 Components



Crosstabulation of Clusters



Only a very small group has extreme response style for both brands, otherwise almost independence.

Software, Papers, Outlook

Packages are available from CRAN and/or R-Forge:

flexclust: KCCA clustering for arbitrary distances, shaded barcharts, projections, convex hulls, static neighborhood graphs, . . .

gcExplorer: Interactive neighborhood graphs with links to Gene Ontology.

symbols: Grid versions of `boxplot()`, `symbols()`, `stars()`, . . .

flexmix: Finite mixture models. Model based clustering for various distributions and mixtures of generalized linear models.

Public availability of features shown in this talk depends on release version of packages.

Papers available at

`http://www.statistik.lmu.de/~leisch`

Big 2do: Redo most with iPlots Extreme . . .