Clustering consumers based on product discrimination in check-all-that-apply (CATA) data

J. C. Castura[1*], M. Meyners [2], P. Varela[3], T. Næs[3,4]

[1]Compusense Inc., 255 Speedvale Av. W., Guelph, Ontario, N1H 1C5, Canada

[2]Procter & Gamble Service GmbH, 65824 Schwalbach am Taunus, Germany

[3] Nofima AS, Osloveien 1, P.O. Box 210, N-1431 Ås, Norway

[4] Dept. of Food Science, Faculty of Sciences, University of Copenhagen, Rolighetsvej 30, 1958 Fredriksberg, Copenhagen, Denmark.

[*] jcastura@compusense.com

Abstract

Consumers can be clustered based on their product-related check-all-that-apply (CATA) responses. We identify two paradoxes that can occur if these clusters are derived from conventional similarity coefficients. The first paradox is that clustering similar consumers can nullify within-cluster sensory differentiation of products. The second paradox is that consumers who check many attributes yet disagree can be clustered together, whereas consumers who check fewer attributes without disagreement can be split into different clusters. After illustrating these paradoxes with toy data sets, we propose "b-cluster analysis", in which consumers are clustered according to how they differentiate products. We define performance metrics to compare cluster analysis solutions. By design, b-cluster analysis is expected to give different results than CLUSCATA, since the objective of CLUSCATA is to cluster consumers who characterize products similarly, not according to how they differentiate products. We apply b-cluster analysis to the same toy data sets and show that the identified paradoxes do not occur. Then we apply both b-cluster analysis and CLUSCATA to a real consumer data set. We find that the b-cluster analysis solutions have better within-cluster sensory differentiation, better sensory discrimination, and less redundant clusters than CLUSCATA solutions. To investigate the sensitivity of b-cluster analysis to the initial (random) cluster membership allocations, we obtained 10,000 two-cluster solutions, each initialized with a different random partitioning of consumers. The best solution, which retains the most sensory differentiation, was observed in 21.4% of the runs. As a best practice, we recommend running b-cluster analysis several times and choosing the best solution. The proposed b-cluster analysis approach can be extended to other types of sensometric data and may have applications in other fields.

Keywords: cluster analysis; unsupervised classification; binary data; sensory evaluation; consumer testing; agreement

1. Introduction

Check-all-that-apply (CATA; Ares & Jaeger, 2015; Meyners & Castura, 2014) data are often collected in product testing with consumers. The researcher provides a list of attributes (e.g. words, phrases, pictures, emojis); consumers are instructed to check all the attributes that describe each sample they evaluate. The CATA question format is popular in part because it allows untrained consumers to rapidly describe the products under study, which may be commercial products, prototypes, cultivars, concepts, environments, or some other experimental condition or factor of interest.

Aggregated consumer CATA data can yield useful sensory profiles of products. One reason for this is the existence of a linear relationship between CATA attribute citation rates and attribute intensities (Jaeger et al., 2020b; Ares et al., 2015; Bruzzone et al., 2012; Ares et al., 2010). However, this finding does not imply that CATA responses are absolute judgments related to the presence or absence of particular stimuli. One consumer might endorse an attribute for a product that another consumer does not for various reasons. The two consumers may differ in their sensitivities for the attribute, their understanding of attribute meaning, their perception of an attribute's relevance to the evaluated samples, or their susceptibility to contextual effects, such as psychological biases (Meyners & Castura, 2014). They might have approached the CATA task with different cognitive strategies (Galler et al., 2020) or differ in their visual attention to the options provided (Antúnez et al., 2016). It is also possible that the samples they evaluated were dissimilar due to product variability. CATA responses might differ due to differences in perception or in how the perceptions are described. Consumers might check an attribute when the perceived intensity exceeds a certain threshold which is consumer- and product category-specific (Jaeger et al., 2020a; Vidal et al., 2018), so that two consumers with similar perceptions but different thresholds might respond differently. One consumer might have a low elicitation threshold for endorsing an attribute, leading to a high citation rate, whereas another consumer with a similar sensory experience might have a high elicitation threshold for endorsing an attribute, leading to a low citation rate. Due to such response bias, the feature that we will focus on to cluster consumers is the product contrasts within each consumer's CATA responses.

The objective of this paper is to introduce a novel approach for clustering consumers based on their CATA data. We call this approach "b-cluster analysis". We use toy and real data sets to demonstrate its ability to resolve certain paradoxes and to retain the sensory differentiation in the solution. This clustering algorithm is inspired by, but different from, the quick-transfer stage of the k-means algorithm proposed by Hartigan and Wong (1979). Ours is not the first proposal for conducting unsupervised classification of consumers based on their CATA responses. Llobell et al. (2019a) proposed CLUSCATA, which focuses on the similarity in how consumers *characterize* products. Vigneau et al. (2022) proposed clustering consumers based on *associations* between CATA and liking responses. The unsupervised classification procedure that we propose in this paper is different because it focuses on how consumers *differentiate* products—a different type of similarity that, as far as we know, has not been considered previously. These cluster analyses have different objectives and, as will be shown, produce different clusters. Cluster analysis is not always needed, nor always useful, but when it is done, it is important to select a clustering method that aligns with the study objectives.

In Section 2, we identify two paradoxes that may arise if conventional similarity coefficients are used to quantify similarity of consumers based on their CATA data. We illustrate these paradoxes using toy data sets. The paradoxes motivate b-cluster analysis. The reasons for this name and the algorithm are given in Section 3. A key feature that differentiates b-cluster analysis from other methods is that it maximizes the retained sensory differentiation in the solution. In Section 4, we show that b-cluster analysis resolves the paradoxes that we identify. Then we apply b-cluster analysis and CLUSCATA to a real CATA data set and compare results. Discussion and conclusions follow.

2. Paradoxes in clustering solutions based on conventional similarity measures

Suppose that consumers are clustered based on the similarity of their CATA responses for various products. Here, it would be natural to make certain assumptions, such as that

(i) members within each cluster will describe the products in a similar way, but differently than members of another cluster;
(ii) each cluster member will differentiate the products in a manner that is similar to how other cluster members differentiate the products; and
(iii) consumers who disagree about how products differ will tend to be allocated to different clusters.

In this section, we identify two paradoxes, which we describe using simple examples. These paradoxes demonstrate that even if assumption (i) holds, assumptions (ii) and (iii) might not hold if cluster analysis is performed using conventional similarity measures. These paradoxes motivate our proposition of b-cluster analysis in Section 3.

2.1. Paradox 1: Clustering similar consumers can nullify within-cluster sensory differentiation

Table 1 shows a toy CATA data set with three consumers (C1, C2, C3), four products (P1 through P4), and two attributes. If two consumers check an attribute for the same product, we call this "elicitation agreement" (denoted c_{11}). If two consumers do not check a given attribute for the same product, we call this "non-elicitation agreement" (denoted c_{00}). Disagreement occurs if the attribute is checked for a product by only the first consumer (denoted c_{10}) or only the second consumer (denoted c_{01}).

Table 1. Toy CATA data for illustrating Paradox 1, consisting of three consumers (C1, C2, C3), four products (P1, P2, P3, P4), and two attributes. Elicitation counts are shown for each combination of two consumers.

	Attribute 1						**Attribute 2**					
	C1	C2	C3	C1+C2	C1+C3	C2+C3	C1	C2	C3	C1+C2	C1+C3	C2+C3
P1	1	0	1	1	2	1	1	0	0	1	1	0
P2	1	0	1	1	2	1	1	0	0	1	1	0
P3	1	0	1	1	2	1	0	0	1	0	1	1
P4	1	0	1	1	2	1	0	1	1	1	1	2

Conventional similarity coefficients quantify the similarities between pairs of consumers using counts from all attributes. Table 2 shows various similarity coefficients for each pair of consumers in Table 1. The similarity coefficients s_{Och} (Ochiai coefficient; Ochiai, 1957), s_{Jac} (Jaccard coefficient; Jaccard, 1912) and s_{DS} (Dice-Sørenson coefficient; Dice, 1945; Sørenson, 1948) were proposed originally to quantify species overlap between two geographical areas; all of these coefficients exclude from their calculations the species that are absent in both locations. Consequently, they consider agreement only from mutual presence of a species in both areas; the mutual absence of a species in both areas is ignored completely. The similarity coefficient s_{SM} (simple matching; Zubin, 1938) considers mutual presence and mutual absence of some characteristic as providing equal evidence of agreement; the average Hamming distance (Hamming, 1950) is not shown but is equal to $1 - s_{SM}$.

Table 2. Similarity coefficients (s) across both attributes are shown for pairs of consumers in Table 1. In each row, the pair of consumers having the largest similarity is shown in bold. [s_{Och}: Ochiai coefficient; s_{Jac}: Jaccard coefficient s_{DS}: Dice-Sørenson coefficient; s_{SM}: simple matching coefficient.]

Measure	**Formula**	$\boldsymbol{s}$(C1, C2)	$\boldsymbol{s}$(C1, C3)	$\boldsymbol{s}$(C2, C3)
s_{Och}	$\frac{c_{11}}{\sqrt{(c_{11}+c_{10})(c_{11}+c_{01})}}$	0.00	**0.67**	0.41
s_{Jac}	$\frac{c_{11}}{c_{11}+c_{10}+c_{01}}$	0.00	**0.50**	0.17
s_{DS}	$\frac{2c_{11}}{2c_{11}+c_{10}+c_{01}}$	0.00	**0.67**	0.29
s_{SM}	$\frac{c_{11}+c_{00}}{c_{11}+c_{10}+c_{01}+c_{00}}$	0.13	**0.50**	0.38

Table 2 indicates that C1 and C3 have the largest similarity coefficients, indicating that these two consumers should be clustered together. We find it reasonable that the number of elicitations in their aggregated data is identical for all products within Attribute 1 (every product has two citations; see Table 1) since no consumer differentiated products on this attribute. However, we find it unreasonable that the number of elicitations in their aggregated data is also identical for all products within Attribute 2 (every product has one citation) since individually the consumers had differentiated the products on this attribute. Paradox 1 is that clustering the two most similar consumers produces a cluster in which all sensory differentiation of products is nullified.

2.2. Paradox 2: Consumers with real disagreements are considered as being similar because they check many attributes

Table 3 shows a toy CATA data set with three assessors (C4, C5, C6), five products (P5 through P9), and two attributes. Elicitation rates for C4, C5, and C6 are 70%, 30%, and 80%, respectively. C4 and C6 have high levels of elicitation agreement; however, C4 indicates that both attributes describe P5

but not P9, and that attribute 4 describes P8, whereas C6 indicates that both attributes describe P8 and P9 but not P5. These inversions indicate real disagreements between C4 and C6. No such inversions exist in the responses from C5 and C6. This means that differences in responses from C5 and C6 could be due to differences in their elicitation thresholds for the attributes instead of any real disagreement. Table 4 shows the four standard similarity coefficients across attributes for each pair of consumers in Table 3.

Table 3. Toy CATA data to illustrate Paradox 2, consisting of three consumers (C4, C5, C6), five products (P5, P6, P7, P8, P9), and two attributes. Elicitation counts are shown for each combination of two consumers.

	Attribute 3						**Attribute 4**					
	C4	C5	C6	C4+C5	C4+C6	C5+C6	C4	C5	C6	C4+C5	C4+C6	C5+C6
P5	1	0	0	1	1	0	1	0	0	1	1	0
P6	1	0	1	1	2	1	1	0	1	1	2	1
P7	1	0	1	1	2	1	1	0	1	1	2	1
P8	1	0	1	1	2	1	0	1	1	1	2	2
P9	0	1	1	1	1	2	0	1	1	1	1	2

Table 4. Similarity coefficients (s) are shown for consumer pairs from Table 3. In each row, the pair of consumers having the largest similarity is shown in bold. (Ties occur in the last row.) [s_{Och}: Ochiai coefficient; s_{Jac}: Jaccard coefficient s_{DS}: Dice-Sørenson coefficient; s_{SM}: simple matching coefficient.]

Measure	$\boldsymbol{s}$(C4, C5)	$\boldsymbol{s}$(C4, C6)	$\boldsymbol{s}$(C5, C6)
s_{Och}	0.00	**0.67**	0.61
s_{Jac}	0.00	**0.50**	0.38
s_{DS}	0.00	**0.67**	0.55
s_{SM}	0.00	**0.50**	**0.50**

For the reasons given above, C5 and C6 might be considered to be the most similar pair. However, according to the first three similarity coefficients in Table 4, C4 and C6 are most similar. The similarity coefficient s_{SM} indicates that it would be equally justifiable to cluster C4 and C6 as it would be to cluster C5 and C6. Paradox 2 is that consumers with low elicitation thresholds and real disagreement are considered to be as or even more similar than consumers whose lack of concordance does not necessarily arise from any real disagreement.

3. Methods

3.1. Notation

Before introducing b-cluster analysis, we give notation. CATA data from I consumers on J products for M attributes are organized into a three-way array $\underline{\mathbf{X}}$. Each datum in $\underline{\mathbf{X}}$ has the value 0 (not checked) or 1 (checked). A "group" (denoted g) refers to any group of N consumers regardless of the reason that they are grouped. A "cluster" is a specific group in which consumers are grouped together due to their data and specified criteria. The term "singleton cluster" refers to a cluster that consists of only one consumer.

3.2. b-cluster analysis

The goal of b-cluster analysis is to cluster consumers to maximize within-cluster sensory differentiation of products. The name for this cluster analysis was selected because b and c are regularly used to denote the number of consumers who check an attribute for the first and not the second product, and who check the second and not the first product, respectively, e.g. in 2×2 tables leading to McNemar's test (McNemar, 1947). The next section shows how what we call the b-measure is obtained from the difference in these counts ($b - c$), where $b - c$ is embedded in the name b-cluster analysis.

3.2.1. b-measure

If every consumer evaluates every product using a CATA question, the responses for each pair of products (j and j') on any attribute m can be treated as matched binary data (Meyners, Castura & Carr, 2013). Consumer CATA results may be organized as shown in Table 5. The counts n_{11}, n_{10} (commonly "*b*" in the above-mentioned table for McNemar's test), n_{01} ("*c*"), and n_{00} indicate the number of consumers who checked attribute m for both products, only product j, only product j', and neither product. Upper case N refers to the total number of consumers, i.e. $n_{11} + n_{10} + n_{01} + n_{00}$. Lower case n refers to the consumers who differentiate this pair of products on this attribute, i.e. $n = n_{10} + n_{01}$.[1]

Table 5. General data structure for matched CATA data for a pair of products (j and j'); n_{11}, n_{10}, n_{01}, and n_{00} indicate the number of consumers who checked an attribute for both products, only product j, only product j', and neither product, respectively.

	checked for j'	not checked for j'	Row totals
checked for j	n_{11}	n_{10}	$n_{11} + n_{10}$
not checked for j	n_{01}	n_{00}	$n_{01} + n_{00}$
Column totals	$n_{11} + n_{01}$	$n_{10} + n_{00}$	N

The null hypothesis assumes that the attribute under consideration is cited with equal probability for products j and j'. The alternative hypothesis states that the underlying probabilities of citation are

[1] Whereas $n_{\#\#}$ refers to responses for a matched pair of products, $c_{\#\#}$ in Section 2 refers to responses from a matched pair of consumers.

unequal. McNemar (1947) shows that under the null hypothesis, $(n_{11}+n_{10})/N = (n_{11}+n_{01})/N$ (i.e., $n_{10} = n_{01}$), the squared test statistic

$$Z^2 = \frac{(n_{10}-n_{01})^2}{n_{10}+n_{01}} \quad (1)$$

is asymptotically distributed as a χ^2-distributed random variable with one degree of freedom. Since Z^2 can be calculated using data from only the n consumers on the off-diagonal, the effective sample size is n, not N. We set $Z^2 = 0$ if $n = 0$. We use the symbol Δ to refer to $n_{10} - n_{01}$ (or $b - c$). The numerator in Eq. (1) has the range $0 \leq \Delta^2 \leq n^2 \leq N^2$. The squared test statistic Z^2 has the range $0 \leq Z^2 \leq n \leq N$, and quantifies sensory differentiation regardless of n.

Since n is often small, rather than using the χ^2-distribution to approximate the null distribution, we test whether j and j' are cited with equal underlying probabilities (with the alternative being that they are unequal) using a two-tailed binomial test with n_{10} successes, sample size n, and probability ½ (Meyners et al., 2013). This test will be applied to evaluate the quality of a cluster analysis solution (Section 3.3). We also apply one-tailed binomial tests to evaluate the direction of the difference, if a difference exists (i.e., either $j < j'$ or $j > j'$). For consistency, we conduct two-tailed tests at the 95% confidence level and each one-tailed test at the 97.5% confidence level. A significant test result indicates sensory discrimination.

We quantify the total sensory differentiation across J products and M attributes by any group of N consumers ($N \leq I$) via

$$b = \sum_{m=1}^{M} \sum_{j=2}^{J} \sum_{j'=1}^{j-1} Z_{jj'm}^2 \quad (2)$$

where $Z_{jj'm}^2$ is the squared test statistic from Eq. (1) for products j and j' on attribute m. For each of M attributes, we obtain the measure $\sum_{j=2}^{J} \sum_{j'=1}^{j-1} Z_{jj'm}^2$ from $J(J-1)/2$ tables, each organized as shown in Table 5. We do not compare j with itself, and once j and j' are compared, it is redundant to compare j' and j as this would give the same value. Thus, b is the sum of squared test statistics from a total of $MJ(J-1)/2$ tables. Details on the b-measure, including the range of possible values in any given data set, are provided in Appendix A.1.

3.2.2. b-cluster analysis via iterative ascent

The b-cluster analysis algorithm is a criterion-based, non-hierarchical procedure. The objective function that quantifies the quality of this solution is

$$B_G = \sum_{g=1}^{G} b_g \quad (3)$$

where b_g is the b-measure for cluster g and B_G is the sensory differentiation retained in the G-cluster solution. For a given G, a solution with the larger B_G is a better solution. It begins with an initialization stage. In this stage, the I consumers are each allocated to one of G groups. The initial cluster memberships can be specified by the researcher or determined by random assignment.

Next, we initialize $\mathbf{V}$, a "calculated values matrix" with I rows and G columns; its elements are b-measures for candidate clusters. Specifically, if i is currently a member of cluster g, then the element $v_{i,g}$ is the b-measure for cluster g if it loses i; otherwise $v_{i,g}$ is the b-measure for cluster g

if it acquires i. We also initialize $\mathbf{U}$, an "update matrix" with I rows and $G-1$ columns; each of its elements quantifies how B_G changes if a candidate transfer is completed. Transferring i to the candidate cluster g_{cand} affects two clusters: it affects the current cluster g_{curr}, which after losing i is denoted $g_{curr(-i)}$, and it affects the candidate cluster g_{cand}, which after acquiring i is denoted $g_{cand(+i)}$. Each column of $\mathbf{U}$ corresponds to an index (c) of the $G-1$ candidate groups, where matrix elements

$$u_{i,c_{g_{cand}}} = b_{g_{cand(+i)}} - b_{g_{cand}} + b_{g_{curr(-i)}} - b_{g_{curr}} \quad (4)$$

quantify the change in B_G if i is transferred from g_{curr} to g_{cand}. The values $b_{g_{cand(+i)}}$ and $b_{g_{curr(-i)}}$ are taken from $\mathbf{V}$, whereas $b_{g_{cand}}$ and $b_{g_{curr}}$ are the current b-measures of the clusters specified. The initialization stage ends once b_g, B_G, $\mathbf{V}$, and $\mathbf{U}$ are calculated based on the initial cluster memberships.

Next, we begin the transfer stage. At each iteration, the maximum value in matrix $\mathbf{U}$, $\max(u_{i,c_{g_{cand}}})$, is determined and the corresponding consumer i is transferred to g_{cand}. If multiple potential transfers yield the same improvement (i.e. the largest value in $\mathbf{U}$ occurs more than once), then one of these transfers is made at random. After a transfer, it is only necessary to update two columns and one row in $\mathbf{V}$, which requires $2(I-1)+G$ rather than IG calculations. The number of calculations (of Eq. (4)) that must be made to update $\mathbf{U}$ depends on the state of the current solution. Specifically, members of the two clusters that have acquired and lost i, respectively, each require $G-1$ calculations; other consumers each require only two calculations for the clusters that acquired and lost *i*, respectively. Our experience indicates that calculating only the necessary values of $\mathbf{V}$ and $\mathbf{U}$ gives advantages for larger G, but provides no efficiencies when $G=2$. B_G is maximized by transferring consumers iteratively until $\max\left(u_{i,c_{g_{cand}}}\right) < 0$. The transfer stage ends if this condition is reached because every possible transfer erodes the quality of the solution. To avoid the potential of an infinite loop, we also stop the transfer stage if both $\max\left(u_{i,c_{g_{cand}}}\right) = 0$ and the variance in B_G over the most recent five iterations is smaller than a pre-specified limit (we used $e^{-8} \approx 0.0003355$).

In the completion stage, the algorithm returns the final cluster memberships for the fixed number of clusters (G). The algorithm can, and often does, reach a local maximum where it cannot increase B_G from the current solution, even though we know that a better solution exists, because we have found one having a larger B_G. For this reason, we advise running the algorithm multiple times with different initial cluster memberships to determine whether a better solution exists. The solution that achieves the largest B_G is considered to be the best G-cluster solution.

The largest possible B_G occurs for the trivial solution without clustering, where $G=I$ (Appendix A.1.4). The percentage of sensory differentiation retained in a G-cluster solution,

$$\%B_G = 100(B_G/B_I)\% \quad (5)$$

measures the quality of a b-cluster analysis solution. A higher percentage indicates a better-quality solution. In what follows, if G is understood implicitly then B_G and $\%B_G$ may be denoted B and $\%B$ for ease of notation.

3.2.3. Determining the number of clusters (G)

B_G tends to decline when G gets smaller and the number of consumers per cluster becomes larger. More sensory differentiation is nullified because heterogeneous consumers cannot be split across more clusters. To determine the number of clusters, we attempt to balance several objectives. For parsimony, we want G to be small. For quality, we want B_G to be high. For robustness and relevance, we want the number of consumers per cluster to be large since overly small clusters sizes are commercially meaningless. Robustness and relevance objectives can be met by keeping G small, or by interpreting only clusters with approximately 50 or more consumers.

To determine the number of clusters, we calculate the percentage change in B_G in a G-cluster solution vs. in B_K in a K-cluster solution ($K = G + 1$) via

$$\Delta B_{K\to G}\% = 100\left(1 - \frac{B_G}{B_K}\right)\% \quad (6)$$

A larger $\Delta B_{K\to G}\%$ value indicates relatively more sensory differentiation is eroded when the number of clusters is reduced from K to G. It also indicates that relatively more sensory differentiation can be reclaimed by increasing the number of clusters from G to K. Values of $\Delta B_{K\to G}\%$ are plotted to evaluate how the quality of the b-cluster analysis solutions evolves. First, we consider $\Delta B_{2\to 1}\%$, then $\Delta B_{3\to 2}\%$, and $\Delta B_{4\to 3}\%$. We prefer the smaller solution (with G clusters) if it retains a relatively high proportion of sensory differentiation in the larger solution (with K clusters). We prefer to decrease the number of clusters whenever we can do so without substantially eroding the sensory differentiation retained.

3.2.4. Evaluating the initialization effect

In this study, we conducted b-cluster analysis with $G = 2, \ldots, 5$ clusters with 500 runs each, using different random initial allocations of consumers to clusters. These results were used to determine the number of clusters to retain. We selected a solution with two clusters. Then, to understand the initialization effect (i.e. the variation in the solution due to random initializations of the algorithm), and to search more exhaustively for a better solution, we re-ran a two-cluster b-cluster analysis 10,000 times.

The best solution (with the largest $\%B$) in 500 runs was also the best solution in 10,000 runs. Then we used two approaches to investigate the agreement between the best solution and all other solutions found in 10,000 runs. First, we considered the raw agreement in cluster memberships between the best solution and each other solution. Second, we considered agreement in cluster-wise product configurations. To obtain this second type of agreement for two different solution, we calculated a $J \times M$ matrix of citation proportions for each cluster in each solution. Then we calculated the RV coefficient (Robert & Escoufier, 1976) for every possible pairing of clusters. The best pairing of clusters was considered to be the one with the highest average RV. The raw agreement in cluster memberships and average RV of the two solutions were plotted to review the initialization effect.

3.3. Evaluating the quality of a CATA cluster analysis solution by additional measures

To evaluate the quality of a cluster analysis solution, we use the following measures, which are based on hypothesis test results.

3.3.1. Within-cluster product discrimination (D_g) — The percentage of the $MJ(J-1)/2$ product comparisons across attributes that are significantly discriminated (at alpha level 5%) by consumers in cluster g is calculated via

$$D_g = 100\left(\frac{s_g}{MJ(J-1)/2}\right)\% \quad (7)$$

where s_g is the count of product pairs that are significantly different. A solution in which D_g is higher in more clusters is considered to be better. Further details are provided in Appendix A.2.1.

3.3.2. Between-cluster non-redundancy ($NR_{gg'}$) — If two clusters mostly discriminate product pairs in the same way, then they provide redundant information. If two groups discriminate product pairs in different ways (e.g., elicitation rates for P1 are significantly higher than P2 in g and significantly lower than P2 in g'), then there is value in keeping these two groups separate. $NR_{gg'}$ is the percentage of non-redundant discriminating test results in groups g and g' and is calculated via

$$NR_{g.g'} = 100\left(\frac{S_{g\veebar g'}}{MJ(J-1)}\right)\% \quad (8)$$

where $S_{g\veebar g'}$ is the count of one-tailed tests on which we obtain a signficant result for either g or g' (but not both; i.e. "xor"). A solution in which between-cluster non-redundancy is high for all pairs of clusters is considered to be better. Further details are provided in Appendix A.2.2.

3.3.3. Overall diversity (Div_G) — The percentage of unique directionally discriminating test results that are observed across the solution is calculated via

$$Div_G = 100\left(\frac{S_G}{MJ(J-1)}\right)\% \quad (9)$$

where S_G is the count of one-tailed tests on which we obtain a significant result for at least one of the G clusters. A solution in which overall diversity is higher is considered to be better. Further details are provided in Appendix A.2.3.

3.4. Comparison with CLUSCATA

We compare b-cluster analysis with the CLUSCATA method (Llobell et al., 2019a). CLUSCATA is run using the recommended procedure: it starts with a hierarchical algorithm to determine the number of clusters, and then uses these cluster memberships to initialize a k-means algorithm that optimizes the solution (Llobell et al., 2019a). In the first step of CLUSCATA, each consumer's CATA data are arranged in a matrix with J products in rows and the M attributes in columns. The binary cosine similarity coefficient (Llobell et al., 2019a) is calculated for each pair of consumers, which is the same as the Ochiai similarity coefficient (s_{Och}; see Table 2) calculated on each consumer's vectorized

multivariate CATA data in the way shown in Section 2. It is equivalent to normalizing each consumer's data by dividing the 0 and 1 values by the Frobenius norm (square root of sum of all squared entries based on each consumer's own data, which weights consumers with different citation rates more equally), then obtaining the trace of the cross-product of the normed matrices (i.e. if $\mathbf{A}$ is the normed matrix, then we obtain the sum of the diagonal elements of the cross-product $\mathbf{A}^{\mathrm{T}}\mathbf{A}$). Similarity coefficients are organized into an $I \times I$ similarity matrix. At each iteration, all candidate groups are evaluated for possible merger. Singular values (Mardia et al., 1979) are determined for the submatrix of similarity coefficients belonging to members of the potential new group. The two groups for which the respective first singular value (λ_1) explains the largest percentage of variability are merged. The procedure continues until all consumers are members of a single group. The number of clusters is chosen based on an adaptation of Hartigan's index (Llobell, 2020; Hartigan, 1975). The solution is optimized using a k-means algorithm that permits many cluster memberships to be updated at each iteration until cluster memberships are finalized. CLUSCATA is deterministic so only needs to be run once for a given data set. The reason is that for a given input (data set), the hierarchical cluster analysis always produces the same output (cluster memberships). This output then initializes the k-means algorithm, which also converges in a deterministic manner to produce the final cluster memberships. The k-means CLUSCATA algorithm also allows the possibility of a "noise cluster" for absorbing consumers who do not fit any cluster well (Llobell et al., 2019a; Dave, 1991). For simplicity, we report only results without a noise cluster.

CLUSCATA aims to achieve high within-cluster homogeneity (H_g), so we will also evaluate clusters using this measure. For any group (g) comprised of I_g consumers, the homogeneity index $H_g = 100\left(\lambda_1^{(g)}/I_g\right)\%$, where $\lambda_1^{(g)}$ is the first singular value of the $I_g \times I_g$ submatrix of the similarity coefficients between all pairs of the I_g consumers. H_g is constrained between $1/I_g$ and 1 since $1 \leq \lambda_1^{(g)} \leq I_g$. Larger values of H_g indicate greater homogeneity in how consumers characterize, rather than differentiate, products. For G groups comprised of $I = \sum_{g=1}^{G} I_g$ consumers, the weighted average of these H_g indices gives the overall homogeneity $H_G = \left(\sum_{g=1}^{G} \lambda_1^{(g)}\right)/I$ (Llobell et al., 2019a).

3.5. Graphical evaluation of a CATA cluster analysis solution

Following each cluster analysis, results within each cluster are visualized using multiple-response correspondence analysis (MR-CA; Mahieu et al., 2021). MR-CA builds on the bootstrap-driven approach of Loughin and Scherer (1998) which accounts for dependencies between attributes in consumers' product-related responses. For this reason, they consider MR-CA more appropriate than conventional correspondence analysis (CA; Greenacre, 2007) for investigating associations between products (rows) and attributes (columns) in the $J \times M$ matrix of CATA citation rates. For details and a comparison of MR-CA and CA we refer to Mahieu et al. (2021). We report the number of significant dimensions based on the multiple-response chi-squared test (χ^2_{MR}; Mahieu et al., 2021). We display products in principal coordinates (analogous to PCA scores) and attributes in standard coordinates (unit vectors). When products and attributes are displayed jointly, the uncertainty of the products is indicated by 95% confidence ellipses that are obtained from the total bootstrap procedure (Cadoret & Husson, 2013).

3.6. The "strawberry data"

The methods in this paper are demonstrated using CATA data from a consumer sensory test of six "products", which are strawberry cultivars: (1) Festival, (2) Yvahé, (3) Yurí, (4) Guenoa, (5) L20.1, (6) K31.5 (Ares & Jaeger, 2013). Consumers evaluated a sample of each strawberry cultivar in sequential monadic presentation format according to a complete-block design that balanced both carry-over and presentation order effects. Consumers described the strawberry samples using 16 CATA attributes. The test was performed in Spanish; the English translations of the evaluated attributes were *sweet, sour, strawberry flavor, strawberry odor, flavorsome, tasteless, red color, irregular shape, regular shape, small, big, firm, hard, soft, juicy, dry*. Complete data were obtained from 114 consumers. The strawberry CATA data from this study have also been analyzed elsewhere (e.g. Meyners & Castura, 2014; Meyners & Hasted, 2021a, 2021b; and Bi & Kuesten, 2021). Llobell et al. (2019a) demonstrated the CLUSCATA method for clustering consumers using this data set, which is publicly available in the R package `ClustBlock` (Llobell, Vigneau, Cariou & Qannari, 2020).

3.7. Software

Data analyses were conducted in R 4.1.1 (R Core Team, 2021) and using functions provided in the R package `cata` (Castura, 2021). We completed b-cluster analyses with $G = \{2, \ldots, 5\}$ clusters with 500 random starts without benchmarking runtimes. Later, we ran b-cluster analyses with two clusters; total running time to complete these 10,000 analyses was approximately 153 h on a virtual machine running Windows 10 Pro (version 21H1, build 19043.1237) with dedicated resources of one core (Intel® Xeon® Silver 4114 CPU, 2.19 GHz) and 10 GB RAM hosted in Hyper-V on a Dell PowerEdge T440 tower server. The average time to complete one such analysis was 55 s. CLUSCATA was conducted using the R package `ClustBlock` version 2.3.1 (Llobell et al., 2020) using the CLUSCATA algorithm (Section 3.4). To visualize CATA results per cluster for the best solutions, we conducted MR-CA and visualized results using the R package `MultiResponseR` (Mahieu, 2021). The adjusted Rand index (ARI; Hubert & Arabie, 1985) is a chance-corrected comparison of the agreement of clustering results that ranges from 0 (no agreement beyond chance) to 1 (perfect agreement). ARI was calculated using the R package `mcclust` (Fritsch, 2012).

4. Results

4.1. Revisiting Paradoxes 1 and 2 with b-cluster analysis

In Table 6, we show all calculations for b-cluster analysis of the toy data sets from Tables 1 and 3. For the "paradox 1" results in Table 1, a two-cluster solution maximizes B_G by placing C1 in one cluster and C2 and C3 together in another cluster. For the "paradox 2" results in Table 3, a two-cluster solution maximizes B_G by placing C4 in one cluster and C5 and C6 together in another cluster. Calculations for the best solutions are provided (e-Component Suppl. Tables S1 and S2). In both cases, the two consumers who are clustered together satisfy the expectations stated at the beginning of Section 2. Recall that the conventional similarity coefficients in Tables 2 and 4 do not cluster consumers in this manner. They cluster C1 and C3, which nullifies the sensory differentiation for this pair of consumers (see Table 6), and cluster C4 and C6, who both have high elicitation rates, but also real disagreement.

Table 6. Results from b-cluster analysis with $G = 2$ for results in Table 1 (columns 1 & 2) and Table 3 (columns 3 & 4). Results are whole numbers as shown (not rounded). The solutions for each data set are ordered best (largest B_G) to worst. (Calculations for the best solutions are shown in Suppl. Tables S1 & S2.)

	Table 1		Table 3
***b*(C1) + *b*(C2,C3)**	**4 + 7 = 11**	***b*(C4) + *b*(C5,C6)**	**10 + 18 = 28**
b(C3) + b(C1,C2)	4 + 3 = 7	b(C5) + b(C4,C6)	10 + 12 = 22
b(C2) + b(C1,C3)	3 + 0 = 3	b(C6) + b(C4,C5)	8 + 0 = 8

4.2. b-cluster analysis of the strawberry data

If each consumer is placed in a singleton cluster ($G = 114$), then no sensory differentiation is lost and $B_{G=114} = 9230$. If all consumers are in one big cluster ($G = 1$), then $B_{G=1} = 1150$. The implication is that a solution without clustering retains only 12.5% (=1150/9230) of the sensory differentiation in this data set. We calculated the best solution from b-cluster analysis based on 500 random starts specifying $G = 2, \ldots, 5$. Fig. 1 shows the variation in $\Delta B_{K \to G}\%$ over the last four changes in the number of clusters. With so few clusters, sensory differentiation is eroded with each reduction in the number of clusters (each $\Delta B_{K \to G} > 0\%$). At the far left, we see that the absolute and relative loss in sensory differentiation is largest when the number of clusters is reduced from two to one. Using a two-cluster solution instead of a solution without clustering provides a fairly large benefit ($\Delta B_{2 \to 1}\% = 46.5\%$). A solution with three clusters does not bring the same size of benefit ($\Delta B_{3 \to 2}\% = 22.0\%$), but whether the additional complexity of the solution is justified or useful depends on the objectives of a project and its analysis. For brevity, only the two-cluster solution will be characterized in Section 4.6, whereas the four-cluster solution will only be discussed briefly in Section 4.4 to allow comparison with the results from CLUSCATA reported by Llobell et al. (2019a).

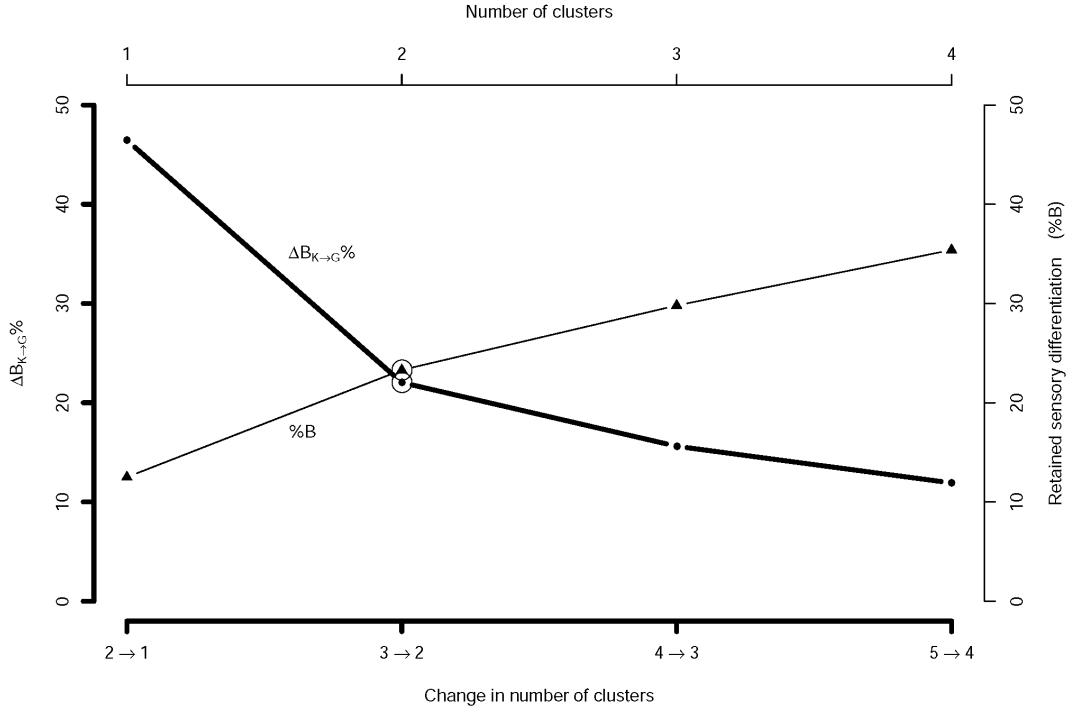


Fig. 1. Relative change in sensory differentiation ($\Delta B_{K\rightarrow G}\%$, heavy line; primary x, y axes on the left, bottom) and percentage of sensory differentiation retained ($\%B$; light line; secondary x, y axes on the top, right) for b-cluster analysis of the strawberry data. The two-cluster solution is emphasized. It retains 23.3% of the total sensory differentiation vs. 12.5% in the solution without clustering.

Next, we conducted b-cluster analysis with $G = 2$ and 10,000 random starts. We chose the best two-cluster solution, which was identical to the best solution that we obtained from only 500 random starts. On average, it took 62.5 iterations from a random start to reach a solution; the number of iterations required were 54 and 69 at the first and third quartiles, respectively.

The best solution was observed in 21.4% of the 10,000 runs; however, 266 other unique solutions were observed. Some of these solutions differ markedly from the best solution. The top six solutions were observed in more than half (55.6%) of the runs; these solutions retained at least 23.2% of the sensory differentiation, which was almost the same as the best solution. The raw agreement in cluster membership between the best solution and each of the five next-best solutions was 82% or higher (e-Component Suppl. Fig. S1), and product configurations for the best-matched cluster pairings were similar to the best solution (average $RV > 0.97$; see e-Component Suppl. Fig. S2). So, if by chance the best solution was not observed, a near-best suboptimal solution would have strongly resembled solution that we used. Regardless, we recommend running b-cluster analysis 50 to 100 times with random starting cluster membership allocations to avoid settling for a suboptimal solution.

Although we chose a two-cluster solution, we noticed that solutions with more than two clusters were more sensitive to the initial (random) cluster membership allocations. For example, the best solutions with three, four, and five clusters were only observed in 10, 2, and 1 out of 500 runs,

respectively. If we recreate Fig. 1 using the average B_G values from 500 runs instead of the best B_G values, the plot characteristics are similar (in particular, $\Delta B_{K \to G}\%$ drops much more sharply for $2 \to 1$ than for any other $K \to G$) and would lead us to select the same two-cluster solution. However, if settling on a larger G, then using more (random) initial configurations is recommended to get (closer) to an optimum solution.

The iterative ascent algorithm anticipated the possibility of ties, but we found that every run at every G ended without a tie, such that any additional transfer would reduce B_G and erode the quality of the solution.

4.3. CLUSCATA results

The strawberry data were also submitted to CLUSCATA cluster analysis. Based on Hartigan's index, we chose a two-cluster solution, even though a four-cluster solution based on this data set was described in an earlier publication (Llobell, 2019). By design, and unlike b-cluster analysis, the CLUSCATA algorithm is deterministic, i.e. a particular data set always produces the same cluster analysis solution (Section 3.4).

4.4. Comparison of solutions

Table 7 shows association matrices for the two- and four-cluster solutions from CLUSCATA (rows) and b-cluster analysis (columns). Counts on the main diagonal indicate the cluster allocation agreement between the cluster allocations from the two cluster analysis methods. The results show that the two cluster analysis methods produced different results. The association matrix for the respective two-cluster solutions (top left) indicates that any agreement between the methods might be due to chance alone ($ARI = 0.00$). The association matrix for the respective four-cluster solutions (bottom right) also indicates that any agreement might be due to chance alone ($ARI = 0.00$). Cluster memberships from the best two-cluster solution and the best four-cluster solution are not strongly similar within b-cluster analysis ($ARI = 0.30$) nor within CLUSCATA ($ARI = 0.33$). This seems reasonable since there is no reason to believe, a priori, that the best two-cluster solution should be obtained by simply merging clusters from the best four-cluster solution.

Table 7. Association matrices showing the two- and four-cluster solutions from both b-cluster analysis and CLUSCATA. Total counts of consumers in the respective clusters are given in the margins.

		b-cluster analysis $G = 2$		*b-cluster analysis* $G = 4$				
		$g1$	$g2$	$g1$	$g2$	$g3$	$g4$	*Total*
CLUSCATA	$g1$	40	32					*72*
$G = 2$	$g2$	20	22					*42*
	$g1$			10	9	4	9	*32*
CLUSCATA	$g2$			11	14	7	5	*37*
$G = 4$	$g3$			6	9	11	7	*33*
	$g4$			1	6	4	1	*12*
	Total	*60*	*54*	*28*	*38*	*26*	*22*	114

In Table 8, we consider the quality of the b-cluster analysis and CLUSCATA solutions based on the two- and four-cluster solutions. We also show results for the solution without clustering, which provides a baseline.

Table 8. Summary of results from b-cluster analysis and CLUSCATA for two- and four-group solutions along with a summary of results without clustering. Cluster order is the same as shown in Table 7. [I_g: cluster size; b_g: sensory differentiation (Section 3.2.1); $\bar{x}_g$: elicitation rate; Signif. Dim.: number of significant dimensions as determined by the χ^2_{MR} test (Section 3.5); H_g: homogeneity (Section 3.4); H_G: overall homogeneity (Section 3.4); D_g: within-group sensory discrimination (Section 3.3.1); $\%B_G$: sensory differentiation retained (Section 3.2.2); $min\left(NR_{gg'}\right)$: minimum non-redundancy (Section 3.3.2); $\mathrm{mean}(RV_{gg'})$: average RV between clusters (Section 3.2.4); Div_G: overall diversity (Section 3.3.3).]

	$G = 1$	$G = 2$		$G = 4$	
	No clustering	b-cluster analysis	CLUSCATA	b-cluster analysis	CLUSCATA
I_g	114	60, 54	72, 42	28, 38, 26, 22	32, 37, 33, 12
b_g	1150	934, 1214	880, 865	789, 1144, 785, 547	632, 733, 595, 308
$\bar{x}_g$	0.29	0.29, 0.29	0.30, 0.27	0.29, 0.30, 0.30, 0.29	0.32, 0.31, 0.29, 0.19
Signif. Dim.	5	4, 5	4, 5	4, 5, 5, 4	5, 4, 5, 4
H_g	37%	39%, 40%	41%, 37%	41%, 43%, 42%, 42%	45%, 44%, 38%, 42%
H_G	37%	39%	39%	42%	42%
D_g	34%	28%, 39%	28%, 22%	24%, 40%, 26%, 13%	18%, 23%, 18%, 0%
$\%B_G$	12.5%	23.3%	18.9%	35.4%	24.6%
$\min\left(NR_{gg'}\right)$	n/a	26%	20%	18%	8%
$\mathrm{mean}(RV_{gg'})$	n/a	0.31	0.56	0.45	0.63
Div_G	17%	30%	24%	40%	25%

First, we consider the two-cluster solutions shown in Table 8. Cluster $g2$ from b-cluster analysis matches CLUSCATA $g2$ better ($RV = 0.74$) than CLUSCATA $g1$ ($RV = 0.58$), but b-cluster analysis $g1$ matches both $g1$ and $g2$ from CLUSCATA equally well ($RV = 0.48$). The two $g1$ clusters are larger in size (60 and 72) than the two $g2$ clusters (54 and 42). Although CLUSCATA $g1$ is larger and retains more sensory differentiation than b-cluster analysis $g1$, these two clusters are similar in within-cluster sensory discrimination. Compared with CLUSCATA $g2$, the b-cluster analysis $g2$ discriminates the cultivars better (D_g is 77% higher; 39/22=1.77) and retains 40% more sensory differentiation (b_g; 1214/865=1.40). In b-cluster analysis, $g2$ discriminates at least one cultivar pair on every attribute, whereas $g1$ fails to discriminate any cultivar pairs on *big* or *strawberry flavour*. In CLUSCATA, $g1$ fails to discriminate any cultivar pairs on *firm* and *dry*; $g2$ fails to discriminate any cultivar pairs on *irregular shape* or *strawberry flavour*. The two CLUSCATA clusters characterize the cultivars in a more redundant manner than the two clusters from b-cluster analysis (based on $\min\left(NR_{gg'}\right)$ and $\mathrm{mean}(RV_{gg'})$). Overall, the solution from b-cluster analysis captures more sensory diversity (Div_G) than the CLUSCATA solution (25%; 30/24=1.25). Both two-cluster solutions will be explored further in Section 4.6.

Since the aim of the paper is methodological, we also show the respective four-cluster solutions in Table 8 even though these clusters are probably too small to be commercially relevant. Both cluster analysis methods split consumers into four clusters with similar homogeneities. The b-cluster analysis solution retains more sensory differentiation (B_g) and is more discriminating (D_g). It has 33% more diversity than the two-cluster solution (40/30=1.33) and more than twice the diversity of the solution without clustering (40/17=2.35). Collectively, the diversity of the four-cluster CLUSCATA solution (25%) does not even achieve the diversity of the b-cluster analysis solution with two clusters (30%), let alone that of its four-cluster b-cluster analysis solution (40%). One reason is that in b-cluster analysis, every cluster achieves at least 18% non-redundancy with some other cluster, which is larger than every pair of CLUSCATA clusters. The multivariate correlation in the product configuration is relatively high for some pairs of clusters in CLUSCATA ($RV_{g3,g4} = 0.87$,$RV_{g2,g3} = 0.67$, $RV_{g2,g4} = 0.65$). Cluster $g2$ from the four-cluster b-cluster analysis solution captures similar information as both CLUSCATA $g3$ ($RV = 0.85$) and CLUSCATA $g4$ ($RV = 0.69$), and has some overlap with CLUSCATA $g2$ ($RV = 0.59$). Taken together, b-cluster analysis yields a more diverse, more discriminating, and less redundant solution than does CLUSCATA.

Finally, the best solutions in Table 8 were compared with solutions obtained by randomly allocating consumers to clusters. Specifically, we generated 10^6 two-cluster solutions each with (i) group sizes identical to the best b-cluster analysis solution, (ii) group sizes identical to the best CLUSCATA solution, and (iii) without restriction on group sizes, respectively. The B_G for the best two-cluster solutions from both b-cluster analysis and CLUSCATA were higher than all of the randomly generated two-cluster solutions. We also generated the same number of four-cluster solutions in an analogous manner. Again, the B_G for the best four-cluster solutions from both b-cluster analysis and CLUSCATA were higher than all of the randomly generated four-cluster solutions. These results illustrate the effectiveness of these cluster analysis algorithms, and especially b-cluster analysis, to identify cluster memberships that retain sensory differentiation.

4.5. Revisiting the paradoxes in the strawberry data

In Section 2, we used toy data sets to show that conventional similarity coefficients tend to group consumers with high overall citation rates and produce clusters that might not differentiate products. We also showed that the clusters obtained from b-cluster analysis differentiate products (Section 4.1). But we do not find compelling evidence that CLUSCATA clustered heavy-checking consumers together. Citation rates are similar in most of the clusters. The 30 consumers (26%) with the highest citation rates are allocated to clusters more evenly in b-cluster analysis (split 15:15) than in the CLUSCATA solution (22:8). Ten of these 30 consumers had checked every product for at least one attribute, and they were split 6:4 by b-cluster analysis and 10:0 by CLUSCATA. However, these outcomes seem unremarkable because they are roughly proportional to their respective cluster sizes.

4.6. Cluster-wise characterization of products in the two-cluster solutions

In both two-cluster solutions, the six strawberry cultivars are well discriminated (Section 4.4). The MR-CA biplots for Dimensions 1 vs. 2 and Dimensions 3 vs. 4 (Figs. 2-5) show the strawberry cultivars overlaid with their 95% confidence ellipses. The first component in each plot is associated with strawberry ripeness. Subsequent components are related to other factors, including shape, size, and

texture. In the two-cluster solution from b-cluster analysis, the confidence ellipses overlap more for $g1$ (Fig. 2) and less for $g2$ (Fig. 3), consistent with the finding that they are less and more discriminating, respectively (see D_g in Table 8). In the CLUSCATA two-cluster solution, the confidence ellipses overlap more for $g1$ (Fig. 4) than for $g2$ (Fig. 5), even though the within-cluster sensory discrimination was higher for $g1$ than for $g2$ (Table 8).

The MR-CA biplots show similarities and differences between $g1$ (Fig. 2) and $g2$ (Fig. 3) from b-cluster analysis. Within each of these clusters, the first two MR-CA components extract 80% of the total inertia (total sum of squares) and the four components shown extract more than 97% of the total inertia in both cases. Both of these clusters characterize Yvahé strawberries as relatively *small* and *irregularly shaped*, Guenoa as relatively *soft*, K31.5 as relatively *sour* and *dry*, and Yurí as relatively *hard*, *sour*, and *dry*, and *tasteless*. Differences also exist; for example, $g1$ characterizes Festival strawberries as relatively *flavoursome* and *sweet* with *strawberry odor*, whereas $g2$ characterizes these strawberries as relatively *firm*, *sour*, *dry*, *hard*, and *tasteless*. Yurí strawberries are characterized relatively more often as *dry* by $g1$ than by $g2$. Cluster $g1$ characterizes Festival strawberries as ripe and Yurí and K31.5 as not ripe, whereas $g2$ characterizes Guenoa and L20.1 strawberries as ripe and Festival strawberries as not ripe. Overall, we find that $g1$ and $g2$ both differentiate the cultivars, but each cluster does so using different attributes.

In CLUSCATA clusters $g1$ and $g2$, MR-CA extracts 82% and 63% of total inertia in the first two components, respectively, and more than 95% of the total inertia in the four components shown (Figs. 4-5). Both CLUSCATA clusters characterize Festival strawberries as relatively *sour*, Yvahé as *small*, Yurí as relatively *hard* and *tasteless*, Guenoa as relatively *juicy* and *soft*, L20.1 as *big*, and K31.5 as relatively *sour* and having a *regular shape*. The clusters differ in the following attributes used to characterize the cultivars. Cluster $g1$ describes K31.5 as *dry*, whereas $g2$ does not. Cluster $g2$ characterizes Yvahé as relatively *juicy* and *regular shaped* and separates it from Yurí, which is characterized as relatively *firm* and *dry*; $g1$ does not discriminate these cultivars quite as well and finds both of these cultivars to be relatively *small* and *irregular shaped*. Cluster $g2$ also discriminates more strongly between Guenoa (*soft*) and Festival (*small*) cultivars, whereas $g1$ is less discriminating of these cultivars, and does not discriminate them all in the first plane. Both clusters characterize Yurí and K31.5 as having characteristics associated with lack of ripeness, each with a different focus. Festival is characterized as ripe by $g1$, but not by $g2$. Although $g2$ has lower homogeneity, fewer consumers, and lower within-cluster sensory discrimination than $g1$ (Table 8), it retains more sensory differentiation per cluster member. This might be one reason why the MR-CA plots show that $g2$ (Fig. 5) separates the strawberry cultivars better than $g1$ (Fig. 4).

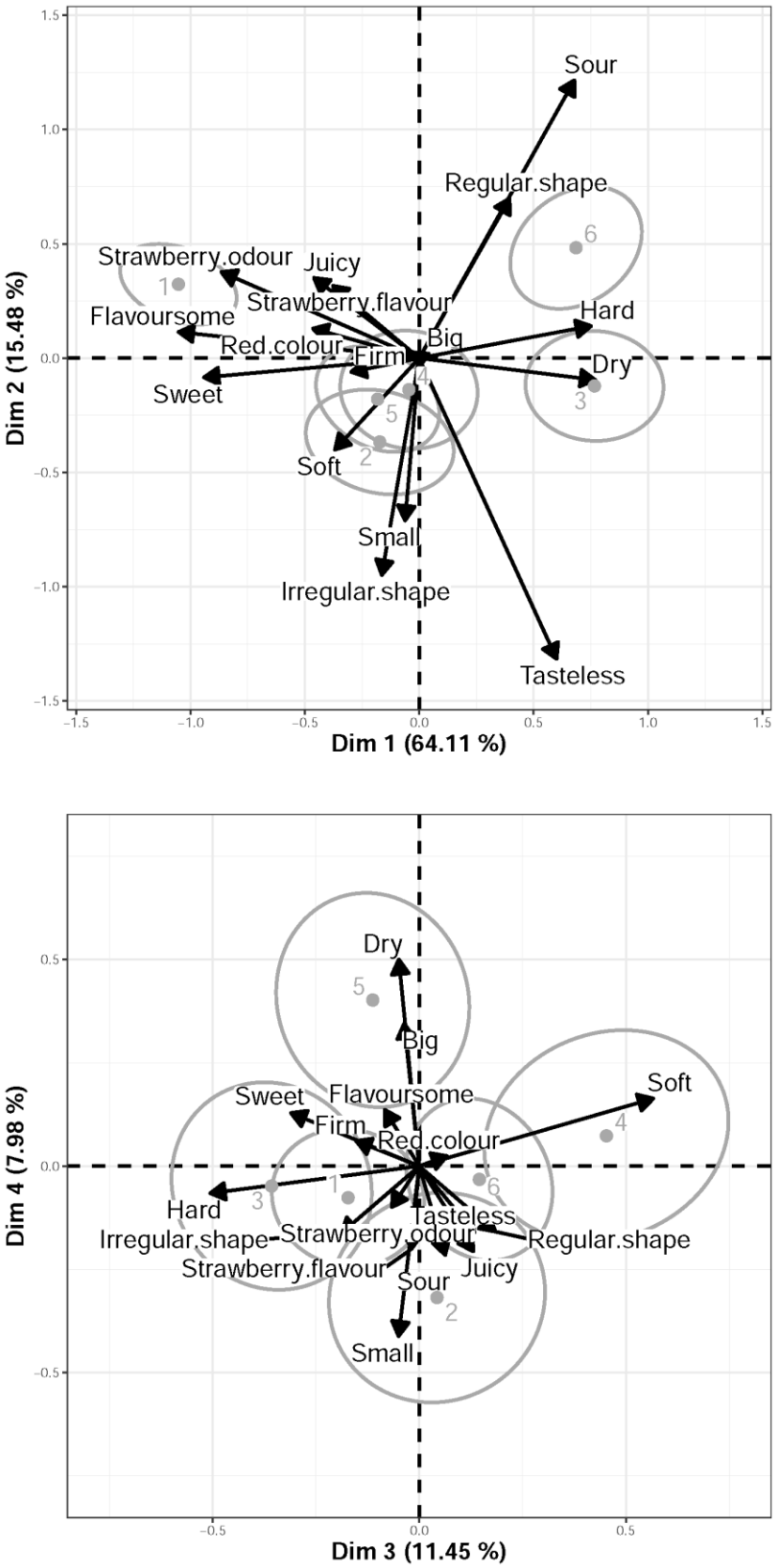


*Fig. 2. MR-CA biplots for cluster g*1 *from b-cluster analysis with 95% confidence ellipses for the strawberry cultivars (1) Festival, (2) Yvahé, (3) Yurí, (4) Guenoa, (5) L20.1, (6) K31.5.*

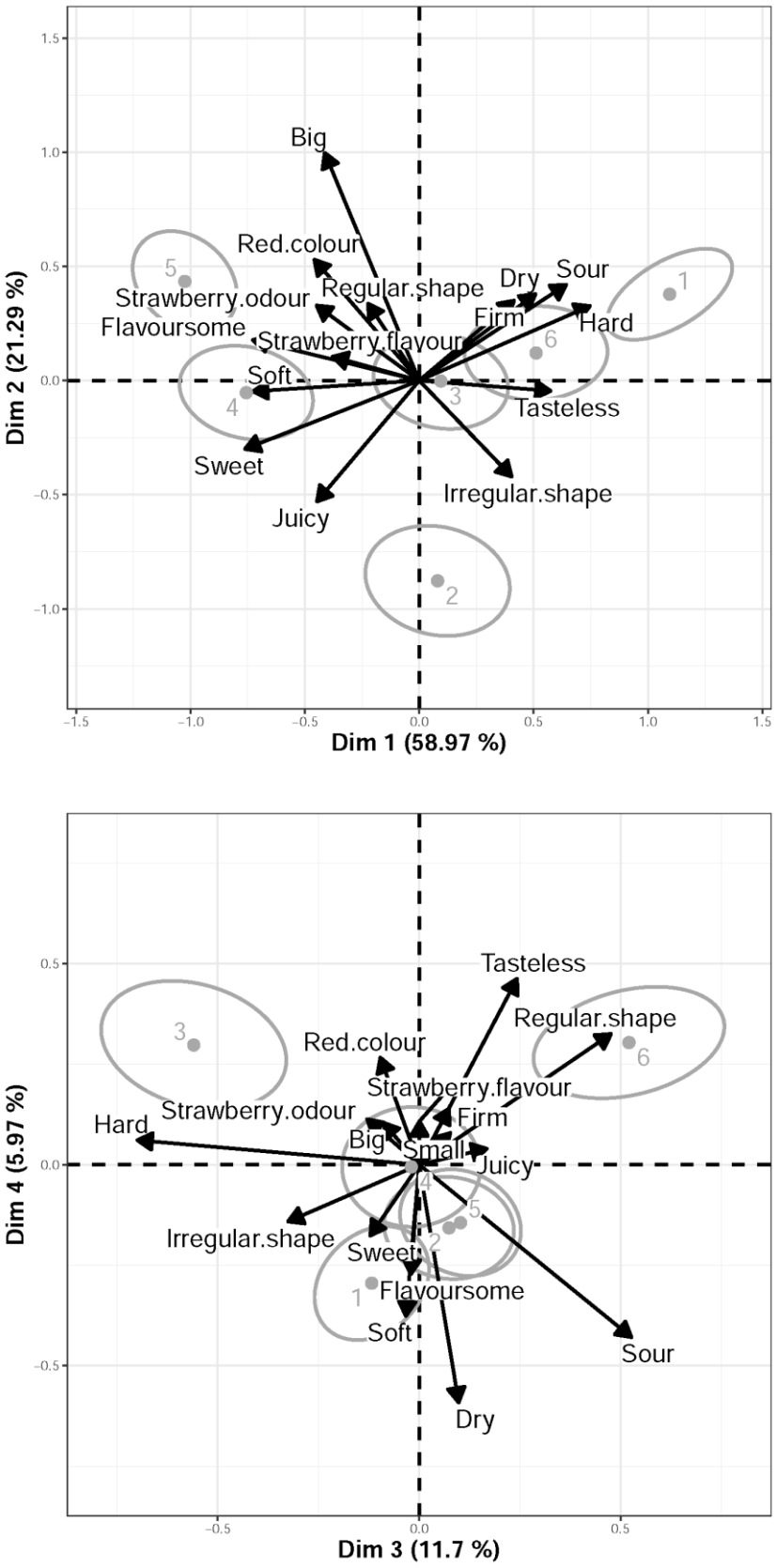


Fig. 3. MR-CA biplots for cluster g2 from b-cluster analysis with 95% confidence ellipses for the strawberry cultivars. (1) Festival, (2) Yvahé, (3) Yurí, (4) Guenoa, (5) L20.1, (6) K31.5.

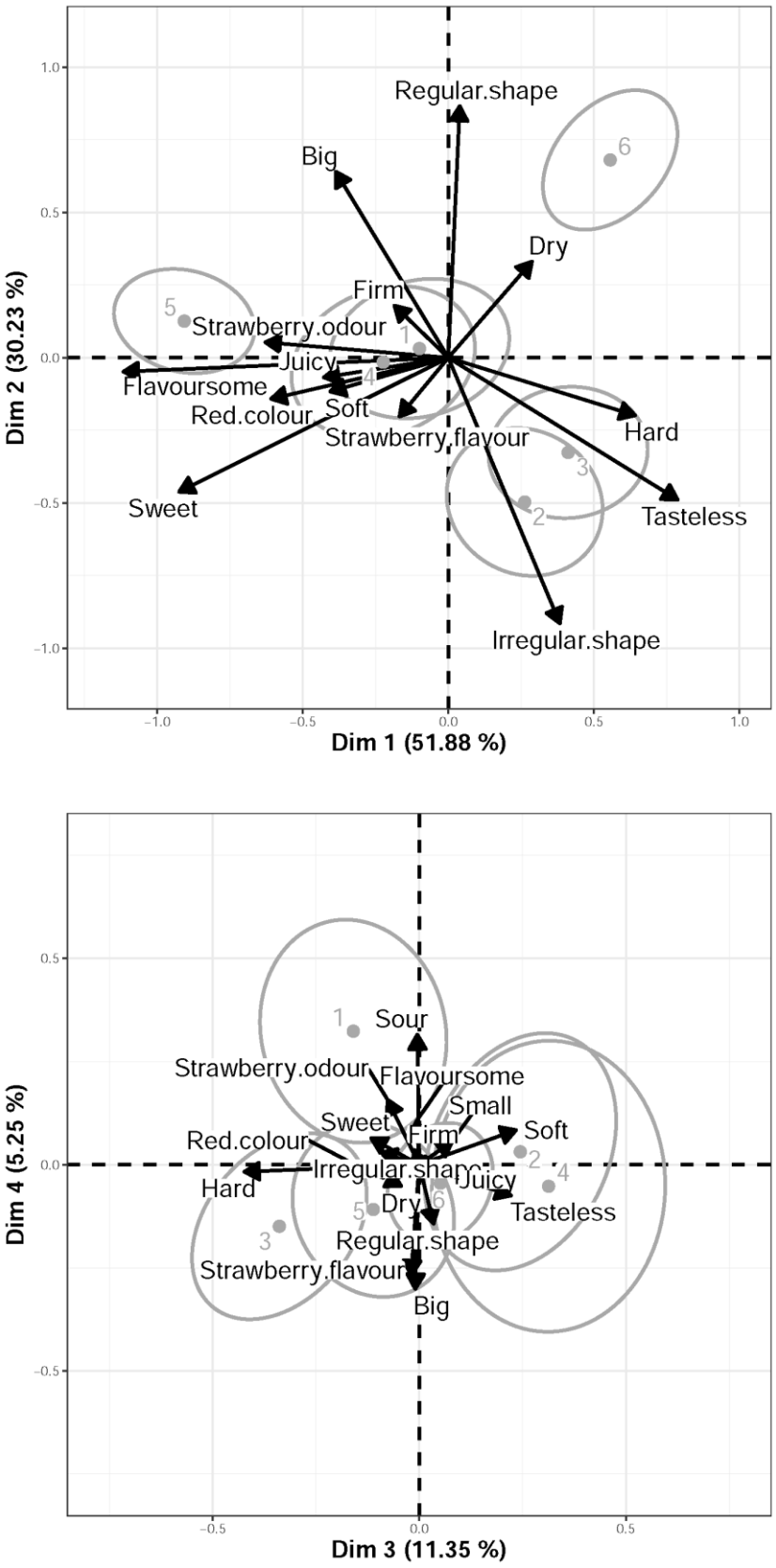


*Fig. 4. MR-CA biplots for cluster g*1 *from CLUSCATA, with 95% confidence ellipses for the strawberry cultivars. (1) Festival, (2) Yvahé, (3) Yurí, (4) Guenoa, (5) L20.1, (6) K31.5.*

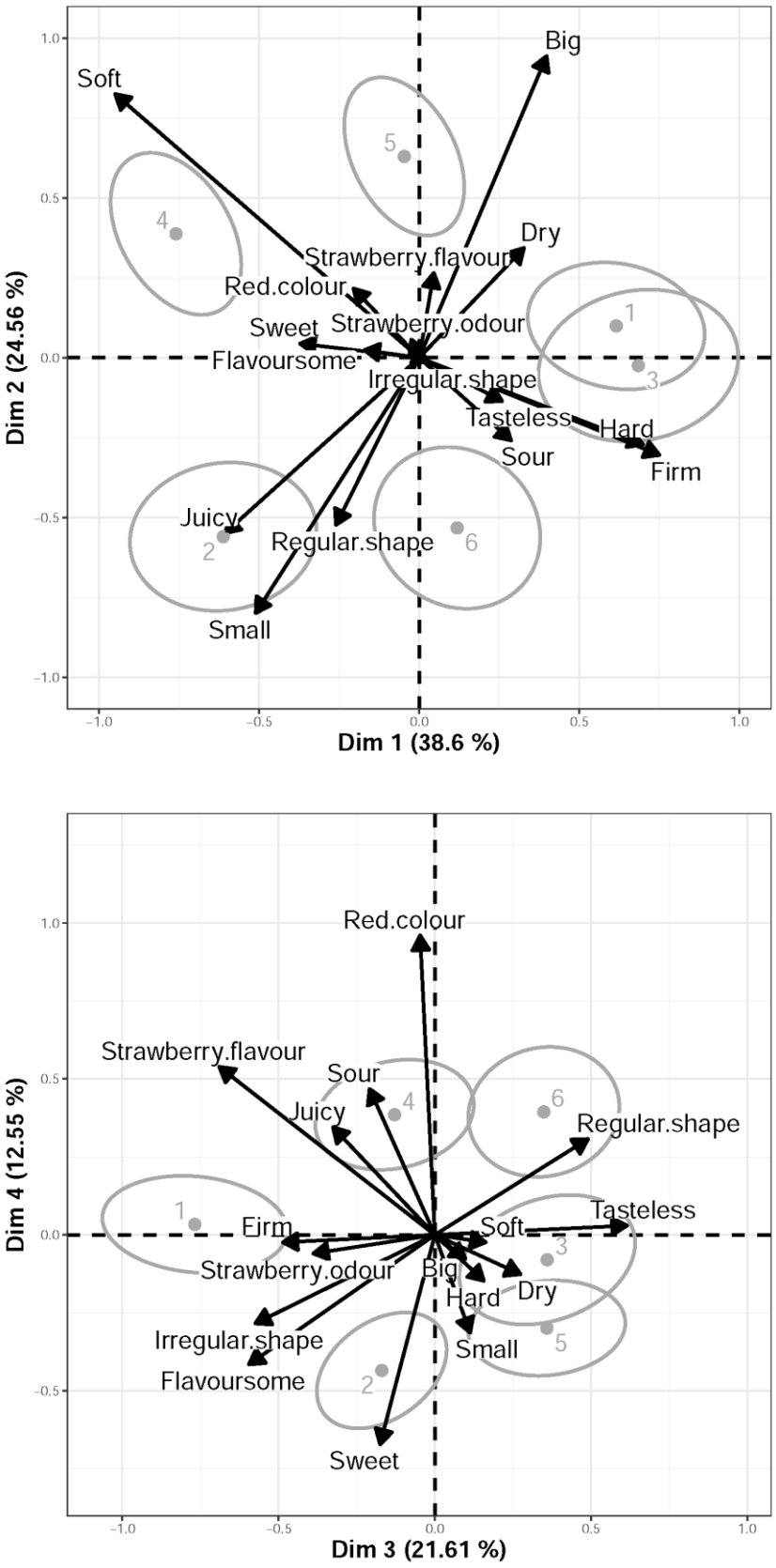


Fig. 5. MR-CA biplots for cluster $g2$ from CLUSCATA, with 95% confidence ellipses for the strawberry cultivars. (1) Festival, (2) Yvahé, (3) Yurí, (4) Guenoa, (5) L20.1, (6) K31.5.

Although the D_g results (Table 8) may give the impression that discrimination in clusters from b-cluster analysis is uniformly better, this is not the case. The CLUSCATA clusters discriminate fewer cultivar pairs and attributes overall, but they successfully discriminate every cultivar pair on at least one attribute. Not so in b-cluster analysis: $g1$ fails to discriminate Guenoa from Yvahé and neither cluster discriminates Guenoa from L20.1. In the MR-CA biplot from both types of cluster analysis, the 95% confidence ellipses of the cultivars are well separated in at least one MR-CA plane (Figs. 2-5), suggesting that every cluster achieves multivariate discrimination. Here we also note that the attribute vector lengths in all of these plots are roughly proportional to the number of cultivar pairs discriminated (Figs. 2-5).

In the first MR-CA plane, *big* is important for CLUSCATA $g1$ and more closely associated with ripeness attributes (Fig. 4) than b-cluster analysis $g1$ (its best-matching cluster if clusters were paired exclusively; see Section 4.4), in which interpretation of *big* provides almost no value (Fig. 2). Only in the second plane do we find *big* associated with *dry* and the L20.1 cultivar in b-cluster analysis. In the first MR-CA plane, CLUSCATA $g2$ (Fig. 5) separates all but the Festival and Yurí cultivars based mainly on visual (shape, size) and textural (*soft*, *firm*, *hard*, *juicy*) attributes. It was previously noted that cluster $g2$ from b-cluster analysis was especially discriminating; Fig. 3 shows that nearly all of the attribute vectors for this cluster have similar lengths in the first MR-CA plane such that together they discriminate the six strawberry cultivars.

5. Discussion

We initially implemented b-cluster analysis using a hierarchical clustering algorithm (available in Castura, 2021), in which consumers each start in a singleton cluster, then pairs of clusters are merged to maximize the retained sensory differentiation at each iteration. These results are not presented here because the non-hierarchical iterative ascent algorithm, which we propose, produces much better clustering solutions. For example, in 10,000 runs, the best two-cluster solution from the iterative ascent algorithm was realized in 21.4% of runs and retained 23.3% of sensory differentiation. By contrast, the best two-cluster solution from the hierarchical algorithm retained only 20.5% of sensory differentiation and was observed in only 1.8% of runs. One reason that the best hierarchical cluster analysis solution is found infrequently is that when equally good mergers are possible, one of these mergers is made randomly. Once merged, consumers are never split again, so each merger has long-ranging implications. Then, when we initialized the iterative ascent algorithm with the final cluster memberships from each of the solutions obtained by the hierarchical algorithm, we found that the best iterative ascent solution was reached in only 2.0% of the runs, and it was not reachable if using the cluster memberships from the best hierarchical solution as a starting point (e-Component Suppl. Table S3). This means that there was often no path from the best hierarchical cluster analysis solutions to the best solution using the b-cluster analysis algorithm. A limitation of hierarchical cluster analysis is that it assumes nesting of cluster analysis solutions across varying numbers of clusters, i.e. the best two-cluster solution is found by merging two clusters from the best three-cluster solution, which in turn is found by merging two clusters from the best four-cluster solution, and so on. The assumption might hold if the clustering structure is very clear, but we see no reason to believe this assumption would hold generally. It constrains the range of solutions that can be obtained, which apparently leads to the relatively poor performance

that we observed. These findings are mentioned because they may have broader implications for the routine use of cluster analysis in sensory evaluation.

Improvement of b-cluster analysis might be achieved by algorithmic refinements inspired by simulated annealing (Kirkpatrick, Gelatt & Vecchi, 1983) or some other adaptation of the Metropolis-Hastings algorithm (Metropolis et al., 1953). The iterative ascent algorithm in Section 3.2 often yields a solution that, based on its B_G value relative to the other B_G values observed, is known to be a local maximum, not a global maximum. We tried slight algorithmic changes that did not yield a solution that was better than the best solution that we reported, but we cannot rule out the possiblity that a better solution exists. More work is needed to investigate whether algorithmic refinements can achieve a better outcome or a similar outcome at lower computational cost. Based on the insights that we have so far, we recommend to use computing resources for runs from random starts rather than more complex algorithms.

Our study of the initialization effect was always done using the same 114 consumers. Thus, the results tell us how reliably the iterative ascent algorithm converged to the best solution for this data set. The recommended number of runs that is likely to lead to the best solution at least once is determined based on these results. Obviously, this initialization study investigates only the sensitivity of the algorithm to the random initial memberships, not the reproducibility of the study results with new consumers.

Other authors have grappled with challenges related to paradoxes described in Section 2. Llobell et al. (2019a) identified that CLUSCATA agreement tends to decline as the number of attributes increases. Llobell et al. (2019b) proposed running Cochran's Q test per attribute (Meyners et al., 2013; Cochran, 1950) across all consumers, then setting non-discriminating attributes aside when conducting the cluster analysis. Dropping non-discriminating attributes might also seem to resolve Paradox 1 (Section 2.1) since dropping Attribute 1, which none of the consumers differentiated in the toy data set, would resolve the paradox. However, in another data set, it is possible that subgroups of consumers will differentiate products systematically yet in contradictory ways, such that the attribute is non-discriminating if all consumers are pooled. If this attribute was dropped, then an attribute that is relevant to the clustering structure will be lost. Also, significance of discrimination is usually a function of sample size, so the approach might drop attributes due to lack of statistical significance from an underpowered study even when the differences in that attribute might be consumer-meaningful. Llobell et al. (2019a) also sought to reduce the outsized influence of heavy-checking consumers on a CLUSCATA solution. They partially address the problem by normalizing each consumer's responses (Section 3.4). In b-cluster analysis, the situation is different: the most influential consumers are those who are the most differentiating, and the most differentiating consumers of all are those who check as close as possible to half of the products for every attribute (Appendix A.1). Since Paradox 2 (Section 2.2) deals conceptually with what constitutes disagreement, it also concerns what constitutes agreement. Llobell et al. (2019a) discuss why elicitation agreement (denoted c_{11}) is more relevant than non-elicitation agreement (denoted c_{00}), and suggest alternative approaches that might be explored, such as using the Faith index (Faith, 1983). In b-cluster analysis, agreement is defined differently (Section 3.2): elicitation agreements for two consumers might be discounted entirely if they occur on attributes that do not differentiate the products.

When working on b-cluster analysis, we discovered another method that also seems to resolve the paradoxes that we described in Section 2. It is based on the φ-coefficient, which for binary data is identical to Pearson's product-moment correlation coefficient and Spearman's rank correlation. Meyners et al. (2013) use the φ-coefficient to investigate the association between two binary variables. If we instead measure the association between two consumers via the φ-coefficient, then we can construct an $I \times I$ table in which entries are $1 - \varphi$ dissimilarities between pairs of consumers. Cluster analysis can then be conducted on the dissimilarity matrix. For brevity, the results are not presented here, but association matrices for cluster memberships based on $1 - \varphi$ dissimilarities vs. cluster memberships from b-cluster analysis as well as from CLUSCATA indicate that very different solutions are obtained. It is also possible that a sum of Cochran's Q test statistics (Meyners et al., 2013; Cochran, 1950) can be used in place of the b-measure in Eq. (2) for a different cluster analysis (Castura, 2021). The mixture of latent trait models with common slope parameters for multivariate binary data (Tang, Browne & McNicholas, 2015), which also incorporates elicitation thresholds into the model but in a different way, might be adapted to cluster consumers on their multivariate CATA data. Research into these and other clustering approaches could prove useful. Additionally, the b-cluster analysis approach as we have presented it here could be extended to analyze other types of data, including continuous, ordinal, and ranking data.

In this paper, we have discussed the clusters as if they were related mainly to the consumers' perception of the samples. But strawberries are an agricultural product, and the fruits of these plants are inherently variable. Even plants with similar genetics and growing conditions may yield strawberries that differ in sensory properties due to the timing of the harvest (ripeness), the location on the plant, the soil micro-conditions, and other factors. It is possible that the consumers were clustered according to the particular sensory characteristics of the strawberry samples that were presented to them for evaluation by the researchers. For example, in b-cluster analysis, cluster $g1$ characterizes the Festival cultivar as *flavoursome* and *sweet*, whereas $g2$ characterizes this cultivar as *tasteless* and *sour* (Section 4.6). If Festival samples were highly variable, some *flavoursome* and *sweet* and others *tasteless* and *sour*, then the sample that the researcher served to a particular consumer would strongly influence the consumer's perceptions as well as the cluster membership allocation. If different consumers evaluate samples of a particular product that are non-equivalent, then cluster memberships will be driven by these within-product sample differences rather than, or in addition to, true differences in consumer perception. Perhaps this is not always a major concern if the purpose of the study is to size a market opportunity. However, if the researcher will make associations between the cluster membership and other types of consumer data (e.g. demographic, socio-economic, attitudinal, behavioural factors) then some or all of these "insights" might be spurious. Previously, it has been pointed out that serving order, which is known and thus easily investigated, can influence hedonic scores and subsequent preference cluster memberships (Hottenstein, Taylor & Carr, 2008). If the cluster memberships are influenced by serving non-equivalent samples of the same product to consumers, then these sample-to-sample differences are often unknown. Without appropriate replication or potential measures suitable to either measure or improve homogeneity of the samples, it is impossible to determine whether the consumer differences observed are due to consumer perception or sample heterogeneity. Here, we note that our goal is merely to present the b-cluster analysis method for finding consumer clusters, not to determine why the clusters that we find exist. The concerns we raise regarding variability of samples extends to any cluster analysis, not only b-cluster analysis. In many sensory applications, products

are manufactured rather than agricultural; samples of such products can often be assumed to be less variable than the samples of the strawberry cultivars in this data set. Future work should thus focus on applying the b-cluster analysis in cases where samples of a particular product or treatment are relatively similar to one another and sample sets have varying degrees of product-to-product differences.

6. Conclusions

In this paper, we identified two paradoxes that can occur when grouping consumers based on conventional similarity coefficients. The first paradox is that grouping the most similar consumers has the potential of nullifying within-group sensory differentiation. The second paradox is that consumers who check many attributes yet have real disagreements are identified as being similar. The CLUSCATA method is based on one such similarity coefficient.

Then we proposed a new approach for clustering consumers based on their CATA data, which we call b-cluster analysis. Its clustering strategy aims to maximize the retained sensory differentiation. We show that the paradoxes that we identified do not occur in b-cluster analysis. We submitted a real CATA data set to both CLUSCATA and b-cluster analysis and found that b-cluster analysis performed better in terms of within-group sensory discrimination and non-redundancy of the groups. This result makes sense: CLUSCATA focuses on *similarity in characterization*, as do many cluster analyses. By contrast, b-cluster analysis focuses on *similarity in product differentiation*, which seems relevant yet, as far as we know, has not been considered previously. This shows a point that might not be broadly appreciated, which is that a cluster analysis solution is often influenced not only by what data are submitted to clustering (e.g. which attributes and products are chosen for the study), but also which clustering algorithm is used.

Appendix A.1

This appendix gives properties of the b-measure when there are J products and one consumer evaluates one attribute (A.1.1), when N consumers evaluate one attribute (A.1.2), when N consumers evaluate M attributes (A.1.3), and when I consumers are each allocated to one of G clusters (A.1.4). In what follows, let $1_{J\in\{odd\}} = 1$ if J is odd and 0 otherwise.

A.1.1. b-measure for one consumer on one attribute — Each consumer can only give one of two possible responses to each product: not checked (0) or checked (1). Whenever $J > 2$, one or more products will have the same response, called a tie. In general, a consumer who checks J_1 products and does not check $J - J_1$ products differentiates

$$\sum_{j=2}^{J}\sum_{j'=1}^{j-1}|\Delta|_{jj'} = J_1(J - J_1) \qquad \text{(A.1.1)}$$

product pairs. The result of Eq. (A.1.1) is also the b-measure for this consumer and attribute. If $J_1 = 0$ or $J_1 = J$, then the consumer responds the same way to all products and differentiates none of the $J(J-1)/2$ product pairs. The consumer can differentiate at most $(J^2 - 1_{J\in\{odd\}})/4$ pairs, which is realized when the consumer checks as close as possible to half of the products.

A.1.2. b-measure for N consumers on one attribute — The non-discriminating result $b = 0$ can only be obtained if every $\Delta = 0$, in which case every $Z^2 = 0$. (We define Δ in Section 3.2.1, and note

there that $Z^2 = 0$ when $n = 0$.) This outcome occurs due either to consumer disagreement or to consumer non-differentiation. If N consumers all differentiate a product pair in the same way, then for this product pair $Z^2 = N$. If none of the N consumers differentiates the products, then $Z^2 = 0$ (Section 3.2.1). If N consumers differentiate one product from the other $J - 1$ products in an identical manner, then $b = N(J - 1)$. If N consumers differentiate $(J + 1_{J\in\{odd\}})/2$ products from the other $(J - 1_{J\in\{odd\}})/2$ products (i.e. as close as possible to half of the products) in an identical manner, then $b = N(J^2 - 1_{J\in\{odd\}})/4$, which is the largest possible value.

A.1.3. b-measure for N consumers on M attributes — The smallest b-measure for M attributes is $b = 0$, which is only realized if $b = 0$ for every attribute. The maximum value is obtained only if the maximum value is obtained for all M attributes. Thus, the range of possible outcomes is

$$0 \le b \le NM(J^2 - 1_{J\in\{odd\}})/4. \qquad \text{(A.1.2)}$$

A.1.4. Sum of b-measures for I consumers split into G *groups* — If I consumers are split into G groups of size I_g, $g = 1, \dots, G$, then each group's b-measure, denoted b_g, has the range indicated in Eq. (A.1.2), where $N = I_g$. Since $\sum_{g=1}^{G} I_g = I$, the sum of b-measures across the G groups has the range

$$0 \le \sum_{g=1}^{G} b_g \le IM(J^2 - 1_{J\in\{odd\}})/4 \qquad \text{(A.1.3)}$$

regardless of G or how the I consumers are split. The maximum value is realized if the I_g consumers in every group agree perfectly and the numbers of elicitation and non-elicitations differ at most by 1 for every consumer and every attribute. This would be an extraordinary occurrence if it ever happens in practice.

Although the range of $\sum_{g=1}^{G} b_g$ is the same regardless of the number of clusters, the sum

$$B_G = \sum_{g=1}^{G} b_g = \sum_{g=1}^{G} \sum_{m=1}^{M} \sum_{j=2}^{J} \sum_{j'=1}^{j-1} Z_{gjj'm}^2 \qquad \text{(A.1.4)}$$

will tend to be smaller when G is small.

When $G = I$, each consumer forms a singleton cluster and there is no within-cluster disagreement. In this case, the numerator of Eq. (1) is either 0 or 1, so $|\Delta|_{gjj'm} = \Delta_{gjj'm}^2 = n_{gjj'm}$, with $n_{gjj'm}$ being the quantity $n = n_{10} + n_{01}$ defined earlier for attribute m and products j and j' in cluster g. For any data set, the sum $B_G = \sum_{g=1}^{G} b_g$ is maximized for $G = I$ since only here is $B_G = \sum_{g=1}^{G} \sum_{m=1}^{M} \sum_{j=2}^{J} \sum_{j'=1}^{j-1} n_{gjj'm}$. The reason is that no consumers are grouped, so every differentiating response contributes to B_G (since no differentiating responses cancel each other out). Whenever $G < I$, some consumers are grouped. If different consumers contribute differentiating responses that disagree with other consumers who are in the same group g, then the inequalities in $Z_{gjj'm}^2 \le n_{gjj'm} \le N_{gjj'm}$ are strict. This shows why we expect B_G to decrease as the number of clusters decreases and the number of consumers per cluster increases.

Appendix A.2

This appendix describes how the quantities presented in Section 3.3 are calculated. Some properties are also given.

A.2.1. Within-cluster product discrimination (D_g) — The percentage of product comparisons that are discriminated by consumers in group g is given in Eq. (7). The dominator, $MJ(J-1)/2$, indicates the number of tests conducted. It is redundant to test both $j \neq j'$ and $j' \neq j$, and unnecessary to test $j \neq j$. Thus, for each of M attributes, there are $J(J-1)/2$ paired comparisons. The smallest possible result is $D_g = 0\%$. Although the largest possible result is $D_g = 100\%$, this result cannot be obtained when consumers agree perfectly. The reason is that if all N consumers in group g agree perfectly and in how they differentiate products and N is large enough to detect differences, then D_g is fully determined by the number of products. The reason is that for binary data, consumers who agree perfectly can discriminate at most $S_g = M(J^2 - 1_{J\in\{odd\}})/4$ product pairs, thus

$$D_g = \begin{cases} \frac{(J+1)}{2J} & \text{if J is odd} \\ \frac{J}{2(J-1)} & \text{if J is even.} \end{cases} \quad \text{(A.2.1)}$$

For the strawberry data ($J = 6$), the largest possible value for perfectly agreeing consumers is $D_g = 60\%$. D_g can only be higher if there is disagreement in how the consumers differentiate products, yet the effective sample size (n) remains relatively high to provide reasonable power for the tests within groups. Of course, D_g can be lower if there is disagreement in how the consumers differentiate products or if statistical power is low since too few consumers differentiate the products. The implication is that it is not necessarily possible to maximize both within-cluster product discrimination and within-cluster agreement simultaneously.

A.2.2. Between-cluster non-redundancy ($NR_{gg'}$) — The percentage of non-redundant discriminating test results in groups g and g' is given in Eq. (8). The denominator, $MJ(J-1)$, indicates the number of test outcomes evaluated. Both $j < j'$ and $j' > j$ are tested, but j is not tested against itself. Thus, for each of M attributes, there are $J(J-1)$ paired comparisons. Since $NR_{gg'}$ cannot be larger than $D_g + D_{g'}$, possible values range from 0% to 100%. $NR_{gg'}$ is calculated for all $G(G-1)/2$ group comparisons, but only the smallest $NR_{gg'}$ is reported. The reason is that a low $NR_{gg'}$ value may indicate that a solution has two clusters that provide redundant information.

A.2.3. Overall diversity or coverage (Div_G) — The percentage of unique directionally discriminating test results that are observed across the solution is given in Eq. (9). As above, the denominator, $MJ(J-1)$, indicates the number of test outcomes evaluated. It is obvious that a group can discriminate either $j > j'$ and $j < j'$, not both. So the largest possible Div_G is 100% if $G \geq 2$, but only 50% if $G = 1$. The smallest possible Div_G is 0% but it cannot be smaller than $\max(D_g)/2$ taken across all *G* groups.

Acknowledgements

We appreciate the reviewers whose feedback helped to improve this manuscript. Authors TN and PV acknowledge financial support from the Research Council of Norway and the Norwegian Fund for Research Fees for Agricultural Products (FFL) through the project “FoodForFuture” (Project number 314318; 2021-2024).

References

Antúnez, L., Ares, G., Giménez, A., & Jaeger, S.R. (2016). Do individual differences in visual attention to CATA questions affect sensory product characterization? A case study with plain crackers. *Food Quality and Preference*, *48*, 185-194. https://doi.org/10.1016/j.foodqual.2015.09.009

Ares, G., Antúnez, L., Bruzzone, F., Vidal, L., Giménez, A., Pineau, B., ... & Jaeger, S.R. (2015). Comparison of sensory product profiles generated by trained assessors and consumers using CATA questions: Four case studies with complex and/or similar samples. *Food Quality and Preference*, *45*, 75-86. https://doi.org/10.1016/j.foodqual.2015.05.007

Ares, G., Barreiro, C., Deliza, R., Giménez, A.N.A., & Gambaro, A. (2010). Application of a check-all-that-apply question to the development of chocolate milk desserts. *Journal of Sensory Studies*, *25*, 67-86. https://doi.org/10.1111/j.1745-459X.2010.00290.x

Ares, G., & Jaeger, S.R. (2013). Check-all-that-apply questions: Influence of attribute order on sensory product characterization. *Food Quality and Preference*, *28,* 141-153. https://doi.org/10.1016/j.foodqual.2012.08.016

Ares, G., & Jaeger, S.R. (2015). Check-all-that-apply (CATA) questions with consumers in practice: Experimental considerations and impact on outcome. In: J. Delarue, J.B. Lawlor and M. Rogeaux (eds.): *Rapid Sensory Profiling Techniques* (pp. 227-245). Woodhead Publishing.

Bi, J., & Kuesten, C. (2021). Commentary on Meyners and Hasted (2021): On the applicability of ANOVA models for CATA data, Food Quality and Preference, 92. *Food Quality and Preference*, *95*, 104340. https://doi.org/10.1016/j.foodqual.2021.104340

Bruzzone, F., Ares, G., & Giménez, A. (2012). Consumers' texture perception of milk desserts. II–Comparison with trained assessors' data. *Journal of Texture Studies*, *43*, 214-226. https://doi.org/10.1111/j.1745-4603.2011.00332.x

Cadoret, M., & Husson, F. (2013). Construction and evaluation of confidence ellipses applied at sensory data. *Food Quality and Preference*, *28*, 106-115. https://doi.org/10.1016/j.foodqual.2012.09.005

Castura, J.C. (2021). `cata`: Analysis of Check-All-that-Apply (CATA) data. *R Package Version 0.0.10.3*. https://CRAN.R-project.org/package=cata

Cochran, W.G. (1950). The comparison of percentages in matched samples. *Biometrika*, *37*, 256-266. https://doi.org/10.2307/2332378

Dave, R.N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, *12*, 657-664. https://doi.org/10.1016/0167-8655(91)90002-4

Dice, L.R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26, 297–302. https://doi.org/10.2307/1932409

Faith, D.P. (1983). Asymmetric binary similarity measures. *Oecologia*, *57*, 287-290. https://doi.org/10.1007/BF00377169

Fritsch, A. (2012). mcclust: Process an MCMC Sample of Clusterings. R package version 1.0. https://CRAN.R-project.org/package=mcclust

Galler, M., Næs, T., Almli, V.L., & Varela, P. (2020). How children approach a CATA test influences the outcome. Insights on ticking styles from two case studies with 6–9-year old children. *Food Quality and Preference*, *86*, 104009. https://doi.org/10.1016/j.foodqual.2020.104009

Greenacre, M. (2007). *Correspondence Analysis in Practice*, 2nd Ed. London: Chapman & Hall/CRC.

Hamming, R.W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, *29*, 147-160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Hartigan, J.A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons, Inc.

Hartigan, J.A., & Wong, M.A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*, 100-108. https://doi.org/10.2307/2346830

Hottenstein, A.W., Taylor, R., & Carr, B.T. (2008). Preference segments: A deeper understanding of consumer acceptance or a serving order effect? *Food Quality and Preference*, *19*, 711-718. https://doi.org/10.1016/j.foodqual.2008.04.004

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193-218. https://doi.org/10.1007/BF01908075

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*, 37-50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jaeger, S.R., Beresford, M.K., Lo, K.R., Hunter, D.C., Chheang, S.L., & Ares, G. (2020a). What does it mean to check-all-that-apply? Four case studies with beverages. *Food Quality and Preference*, *80*, 103794. https://doi.org/10.1016/j.foodqual.2019.103794

Jaeger, S.R., Chheang, S.L., Jin, D., Roigard, C.M., & Ares, G. (2020b). Check-all-that-apply (CATA) questions: Sensory term citation frequency reflects rated term intensity and applicability. *Food Quality and Preference*, *86*, 103986. https://doi.org/10.1016/j.foodqual.2020.103986

Kirkpatrick, S., Gelatt, C.D., & Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671-680. https://doi.org/10.1126/science.220.4598.671

Llobell, F. (2020). *Classification de tableaux de données, applications en analyse sensorielle* [Doctoral dissertation]. Oniris, INRAE, Nantes, France.

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., & Qannari, E.M. (2019a). A new approach for the analysis of data and the clustering of subjects in a CATA experiment. *Food Quality and Preference 72*, 31-39. https://doi.org/10.1016/j.foodqual.2018.09.006

Llobell, F., Giacalone, D., Labenne, A., & Qannari, E.M. (2019b). Assessment of the agreement and cluster analysis of the respondents in a CATA experiment. *Food Quality and Preference*, *77*, 184-190. https://doi.org/10.1016/j.foodqual.2019.05.017

Llobell, F., Vigneau, E., Cariou, V., & Qannari, E.M. (2020). `ClustBlock`: Clustering of Datasets. R package version 2.3.1. https://CRAN.R-project.org/package=ClustBlock

Loughin, T.M., & Scherer, P.N. (1998). Testing for association in contingency tables with multiple column responses. *Biometrics*, *54*, 630–637. https://doi.org/10.2307/3109769

Mahieu, B. (2021, May 2). `MultiResponseR`: Analysis of multiple-response contingency data. R package version 1.0.0. https://github.com/MahieuB/MultiResponseR

Mahieu, B., Schlich, P., Visalli, M., & Cardot, H. (2021). A multiple-response chi-square framework for the analysis of Free-Comment and Check-All-That-Apply data. *Food Quality and Preference*, *93*, 104256. https://doi.org/10.1016/j.foodqual.2021.104256

Mardia, K.V., Kent, J.T., Bibby, J.M. (1979). *Multivariate Analysis*. London: Academic Press.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153–157. https://doi.org/10.1007/BF02295996

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*, 1087-1092. https://doi.org/10.1063/1.1699114

Meyners, M., & Castura, J.C. (2014). Check-all-that-apply questions. In: P. Varela and G. Ares (eds.): *Novel Techniques in Sensory Characterization and Consumer Profiling* (pp. 271-306). Boca Raton, FL: CRC Press.

Meyners, M., Castura, J.C., & Carr, B.T. (2013). Existing and new approaches for the analysis of CATA data. *Food Quality and Preference*, *30*, 309-319. https://doi.org/10.1016/j.foodqual.2013.06.010

Meyners, M., & Hasted, A. (2021a). On the applicability of ANOVA models for CATA data. *Food Quality and Preference*, *92*, 104219. https://doi.org/10.1016/j.foodqual.2021.104219

Meyners, M., & Hasted, A. (2021b). Reply to Bi and Kuesten: ANOVA outperforms logistic regression for the analysis of CATA data. *Food Quality and Preference*, 104339. https://doi.org/10.1016/j.foodqual.2021.104339

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of Japanese Society of Scientific Fisheries*, *22*, 526-530. https://doi.org/10.2331/suisan.22.531

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robert, P., & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *25*, 257-265.

https://doi.org/10.2307/2347233

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Det Kongelige Danske Videnskabernes Selskab*, *5*, 1–34.

Tang, Y., Browne, R.P., & McNicholas, P.D. (2015). Model based clustering of high-dimensional binary data. *Computational Statistics & Data Analysis*, *87*, 84-101. https://doi.org/10.1016/j.csda.2014.12.009

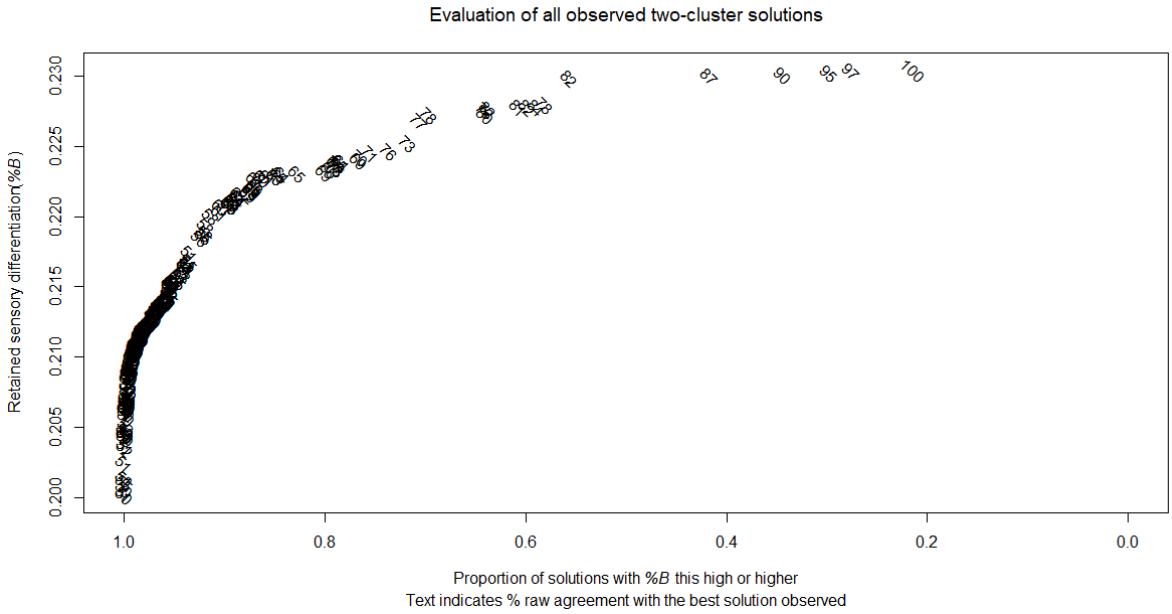
Vidal, L., Ares, G., Hedderley, D. I., Meyners, M., & Jaeger, S. R. (2018). Comparison of rate-all-that-apply (RATA) and check-all-that-apply (CATA) questions across seven consumer studies. *Food Quality and Preference*, *67*, 49-58. https://doi.org/10.1016/j.foodqual.2016.12.013

Vigneau, E., Cariou, V., Giacalone, D., Berget, I., & Llobell, F. (2022). Combining hedonic information and CATA description for consumer segmentation. *Food Quality and Preference*, *95*, 104358. https://doi.org/10.1016/j.foodqual.2021.104358

Zubin, J. (1938). A technique for measuring like-mindedness. *The Journal of Abnormal and Social Psychology*, *33*, 508. https://doi.org/10.1037/h0055441

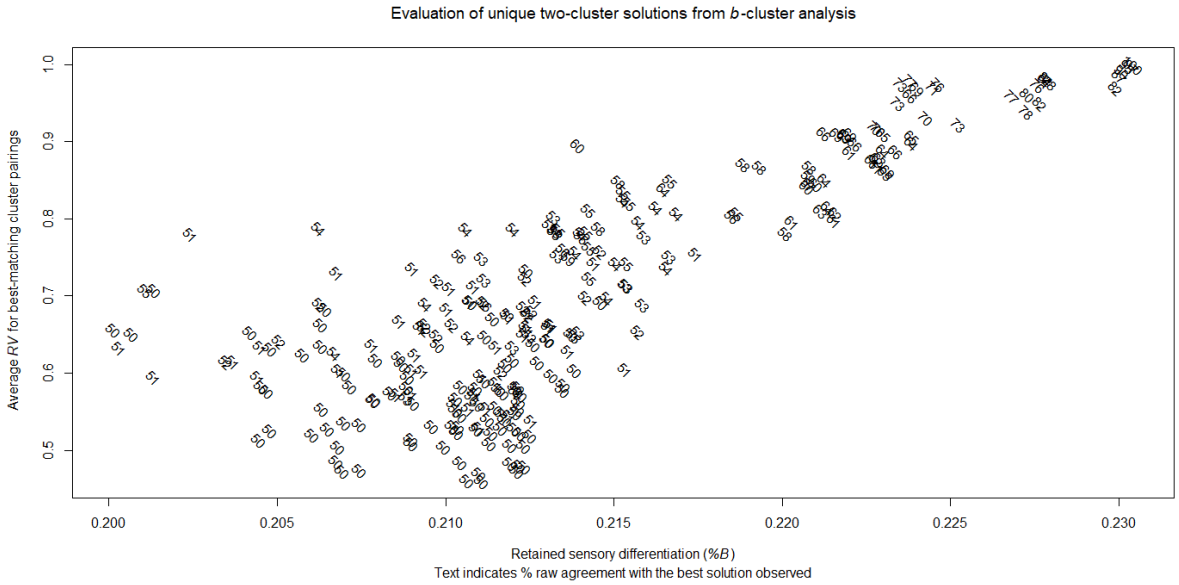
e-Component

Suppl. Fig. S1. Sensory differentiation retained (%B; y axis) vs. cumulative proportion of 10000 b-cluster analyses that retain at least the sensory differentiation indicated (x axis). Raw agreement between each unique solution vs. the best two-cluster solution (top right, indicated by "100") is indicated numerically for each result.



e-Component

Suppl. Fig. S2. The RV coefficient for the best solution's two clusters and the closest matching two clusters from unique solutions (y axis) is plotted against the sensory differentiation retained (x axis). The best solution vs. itself is also included ($RV = 1$; top right).



e-Component

Suppl. Table S1. Calculations for the best solution shown in Table 6 for raw data from Table 1.

Cluster	Attribute	Sample Pair	n_{10}	n_{01}	Z^2	b_g
C1	1	P1, P2	1	1	0	
C1	1	P1, P3	1	1	0	
C1	1	P1, P4	1	1	0	
C1	1	P2, P3	1	1	0	
C1	1	P2, P4	1	1	0	
C1	1	P3, P4	1	1	0	
C1	2	P1, P2	1	1	0	
C1	2	P1, P3	1	0	1	
C1	2	P1, P4	1	0	1	
C1	2	P2, P3	1	0	1	
C1	2	P2, P4	1	0	1	
C1	2	P3, P4	0	0	0	
***b*(C1)**						4
C2,C3	1	P1, P2	1	1	0	
C2,C3	1	P1, P3	1	1	0	
C2,C3	1	P1, P4	1	1	0	
C2,C3	1	P2, P3	1	1	0	
C2,C3	1	P2, P4	1	1	0	
C2,C3	1	P3, P4	1	1	0	
C2,C3	2	P1, P2	0	0	0	
C2,C3	2	P1, P3	0	1	1	
C2,C3	2	P1, P4	0	2	2	
C2,C3	2	P2, P3	0	1	1	
C2,C3	2	P2, P4	0	2	2	
C2,C3	2	P3, P4	0	1	1	
***b*(C2,C3)**						7
***b*(C1) + *b*(C2,C3)**						11

e-Component

Suppl. Table S2. Calculations for the best solution shown in Table 6 for raw data from Table 3.

Cluster	Attribute	Sample Pair	n_{10}	n_{01}	Z^2	b_g
C4	3	P5, P6	0	0	0	
C4	3	P5, P7	0	0	0	
C4	3	P5, P8	0	0	0	
C4	3	P5, P9	1	0	1	
C4	3	P6, P7	0	0	0	
C4	3	P6, P8	0	0	0	
C4	3	P6, P9	1	0	1	
C4	3	P7, P8	0	0	0	
C4	3	P7, P9	1	0	1	
C4	3	P8, P9	1	0	1	
C4	4	P5, P6	0	0	0	
C4	4	P5, P7	0	0	0	
C4	4	P5, P8	1	0	1	
C4	4	P5, P9	1	0	1	
C4	4	P6, P7	0	0	0	
C4	4	P6, P8	1	0	1	
C4	4	P6, P9	1	0	1	
C4	4	P7, P8	1	0	1	
C4	4	P7, P9	1	0	1	
C4	4	P8, P9	0	0	0	
***b*(C4)**						10
C5,C6	3	P5, P6	0	1	1	
C5,C6	3	P5, P7	0	1	1	
C5,C6	3	P5, P8	0	1	1	
C5,C6	3	P5, P9	0	2	2	
C5,C6	3	P6, P7	0	0	0	
C5,C6	3	P6, P8	0	0	0	
C5,C6	3	P6, P9	0	1	1	
C5,C6	3	P7, P8	0	0	0	
C5,C6	3	P7, P9	0	1	1	
C5,C6	3	P8, P9	0	1	1	
C5,C6	4	P5, P6	0	1	1	
C5,C6	4	P5, P7	0	1	1	
C5,C6	4	P5, P8	0	2	2	
C5,C6	4	P5, P9	0	2	2	
C5,C6	4	P6, P7	0	0	0	
C5,C6	4	P6, P8	0	1	1	
C5,C6	4	P6, P9	0	1	1	
C5,C6	4	P7, P8	0	1	1	
C5,C6	4	P7, P9	0	1	1	
C5,C6	4	P8, P9	0	0	0	
***b*(C5, C6)**						18
***b*(C4)+ *b*(C5, C6)**						28

e-Component

Suppl. Table S3. The hierarchical agglomerative b-cluster analysis algorithm yielded 39 unique solutions after 10,000 runs. The final cluster memberships from each of these unique solution were used as starting point for 10 runs of the non-hierarchical iterative ascent b-cluster analysis algorithm (described in Section 3.2). From each starting point, the 10 runs converged on the same solution. The sensory differentiation retained is reported. The hierarchical solution and the non-hierarchical solutions that retain the most sensory differentiation are emphasized (in bold italics). The best solution overall was only achieved when starting from the cluster memberships produced by one (the 13th best) of the 39 solutions from the hierarchical agglomerative algorithm.

Hierarchical solution			*After b-cluster analysis*
Rank	$\mathbf{B}_{G=2}$	Observed Frequency /10000	$\mathbf{B}_{G=2}$
1	***1888.986***	182	1988.593
2	1878.476	224	2142.815
3	1848.182	206	2142.815
4	1846.482	222	2057.662
5	1839.096	235	2086.377
6	1838.545	201	2086.377
7	1837.604	213	2093.550
8	1836.788	429	2142.815
9	1834.564	235	2142.815
10	1830.030	211	2145.912
11	1824.112	199	2142.815
12	1821.981	223	2118.120
13	1817.530	198	***2147.665***
14	1814.625	193	2142.815
15	1807.171	815	2086.377
16	1804.362	198	2142.815
17	1801.815	382	2019.062
18	1796.695	409	2143.668
19	1793.182	198	2145.912
20	1792.536	213	2142.815

Hierarchical solution			*After b-cluster analysis*
Rank	$\mathbf{B}_{G=2}$	Observed Frequency /10000	$\mathbf{B}_{G=2}$
21	1791.568	207	2121.715
22	1788.838	185	2036.398
23	1787.853	442	2142.815
24	1785.445	443	2057.662
25	1783.379	241	2143.668
26	1782.202	200	1978.892
27	1779.341	205	2145.912
28	1774.063	213	2121.715
29	1766.942	194	2123.195
30	1765.732	214	2145.912
31	1763.094	184	2142.815
32	1759.773	190	2142.815
33	1757.844	227	2078.950
34	1746.985	199	2142.815
35	1739.508	200	2078.950
36	1728.529	211	1944.106
37	1726.192	435	2118.120
38	1721.318	234	2142.815
39	1699.873	190	2123.195