Food Quality and Preference 44 (2015) 36–43

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Product selection for liking studies: The sensory informed design

Brian C. Franczak [a,*], Ryan P. Browne [a], Paul D. McNicholas [a], Christopher J. Findlay [b]

[a] Department of Mathematics & Statistics, McMaster University, Hamilton, Ontario L8S 4L8, Canada
[b] Compusense Inc., Guelph, Ontario N1H 1C5, Canada

ARTICLE INFO

Article history:
Received 2 September 2014
Accepted 24 February 2015
Available online 7 March 2015

Keywords:
Consumer response
Incomplete-block designs
Imputation
Liking studies
Nested designs
Parsimonious Gaussian mixture models

ABSTRACT

Liking studies are designed to ascertain consumers likes and dislikes on a variety of products. However, it can be undesirable to construct liking studies where each panelist evaluates every target product. In such cases, an incomplete-block design, where each panelist evaluates only a subset of the target products, can be used. These incomplete blocks are often balanced, so that all pairs occur the same number of times. While desirable in many situations, balanced incomplete blocks have the disadvantage that, by their nature, they cannot favor placing dissimilar products next to one another. A novel incomplete-block design is introduced that utilizes the target product's sensory profile to allocate products to each panelist so that they are, in general, as dissimilar as possible while also ensuring position balance. The resulting design is called a sensory informed design (SID). Herein, details on the formulation of SIDs are given. Data arising from these SIDs are analyzed using a simultaneous clustering and imputation approach, and the results are discussed.

1. Introduction

Experimental design is the cornerstone of sensory analysis. The development of Latin square and complete-block designs for sensory and consumer evaluation date back at least as far as Ferris (1957). However, the majority of literary contributions in this area are based on the earlier work of Williams (1949) and the more recent paper by MacFie et al. (1989). Since MacFie et al. (1989), work on developing other types of experimental designs for sensory analysis has flourished. Wakeling and MacFie (1995) extended the results of Williams (1949) to situations where only a subset of treatments can be provided to each experimental unit. Ball (1997) developed incomplete-block designs that are balanced for carry-over effects. Deppe et al. (2001) provided a procedure for constructing nested incomplete-block designs. Kunert and Sailer (2007) discussed the development of generalized Youden designs, where the experimental units are randomized.

In sensory analysis, the experimental units are either consumers or sensory assessors and the treatments are food products. Evaluating consumer preference, i.e., likes and dislikes, is done through a liking study. Formally, liking studies are comprised of p products and n panelists who are asked to rate each product using either a hedonic or line scale. Ofttimes, due to limited resources, time constraints, or to prevent the onset of fatigue, researchers do not ask the panelists to rate the entire set of products. Instead, they present each panelist a subset of k target products.

There are a number of papers discussing the analysis of liking studies that are either complete or incomplete-block designs (see Bastian et al., 2010; Bower and Whitten, 2000; Gilbert et al., 1996; Harker et al., 2008; Lange et al., 2002; Voorpostel et al., 2014, for examples). In a complete-block design, all treatments are applied to every experimental unit the same number of times. Therefore, when the number of target products is too large, a liking study in the form of a complete-block design arises when one subset of k products is assigned to each panelist. An incomplete-block design also utilizes subsets of the target products; however, these subsets change for each panelist. In a balanced incomplete-block design, all products appear the same number of times and all pairs of products appear the same number of times.

Researchers typically desire that one, or each, subset be "representative" of the set of all target products. As such, they usually rely on the recommendation of trained assessors (see Hersleth et al., 2005, for example). However, it is possible that the assessors could be unintentionally subjective and struggle to agree on what qualifies as a truly representative subset. Herein, we say a subset of products is representative of the set of all target products if its elements are as dissimilar as possible, where dissimilarity is determined using the Euclidean distance measure. Formally, using

* Corresponding author. Tel.: +1 (905) 525 9140x23421.
E-mail addresses: bfrancza@math.mcmaster.ca (B.C. Franczak), rbrowne@math.mcmaster.ca (R.P. Browne), mcnicholas@math.mcmaster.ca (P.D. McNicholas), cfindlay@compusense.com (C.J. Findlay).

http://dx.doi.org/10.1016/j.foodqual.2015.02.015
0950-3293/

their sensory profile, we calculate the Euclidean distance between each target product and argue that the subset of $k > 1$ target products that maximize the Euclidean distance best represents the set of all target products. We form a sensory informed design (SID) by maximizing distance between each target product while also maintaining overall position balance, so that each product occurs the same number of times in each position (or as close to the same number of times as the total number of panelists allows). Of course, we also require that no product is presented to the same panelist more than once.

The remainder of the paper is outlined as follows. In Section 2, we formulate an SID. In Section 3, we review a model-based approach developed for analyzing data with missing values. In Section 4, we apply this model-based approach to two SIDs collected at Compusense Inc., and we conclude with a summary and suggestions for future work (Section 5).

2. The sensory informed design

An SID assumes that consumers have more difficultly discriminating between similar products compared to dissimilar products and requires a sensory profile for the target products. A sensory profile is a $d \times p$ matrix, where d is the number of attributes, constructed by trained assessors who objectively measure each product's attributes using an unstructured line scale at "0" and "100". A score of '0' indicates low intensity and a score of '100' indicates high intensity. Typically, each product will be evaluated multiple times. As such, we use the average attribute scores. Table 1 displays the average scores (i.e., the sensory profile) for the first ten attributes and five products of twelve white breads, denoted $A, \ldots, L$ (cf. Browne et al., 2013). Note that there were 42 attributes in total.

The sensory profile allows us to place each product into a "product space." For an SID, the product space is defined by the Euclidean distance measure. Consider the space constructed by calculating the Euclidean distance between two of the twelve white bread's attributes: color intensity of crust (whole loaf) and color intensity of crumb (Fig. 1). The pair of products with the largest Euclidean distance are considered the most dissimilar, whereas the pair of products with the smallest Euclidean distance are considered the most similar. In Fig. 1, products J and K are the most dissimilar, and products F and I are the most similar.

In an SID, each panelist is given products so that consecutive products are as dissimilar as possible. Fig. 2 illustrates a "greedy" product selection process for one panelist. In this example, four out of twelve white breads are being selected, creating a 12-present-4 design, denoted ${}_{12}P_4$. The first step in constructing an SID is to randomly assign one of the products to a panelist.

Table 1
Mean scores for the first ten attributes and five products resulting from a sensory analysis of a white bread data set.

Attribute	Product				
	A	B	C	D	E
Color Intensity of Crust (Whole Loaf)	57	58	63	48	71
Glossiness of Crust (Whole Loaf)	9	14	15	15	15
Visual Roughness of Crust (Whole Loaf)	9	17	27	15	16
Color Intensity of Crumb	18	28	21	20	21
Cell Uniformity (Crumb)	61	67	71	66	68
Cell Size	20	17	12	14	13
Overall Aroma	30	35	33	31	35
Grain Aroma	8	12	7	8	11
White Flour Aroma	17	16	17	16	16
Yeasty Fermented Aroma	8	12	7	8	11

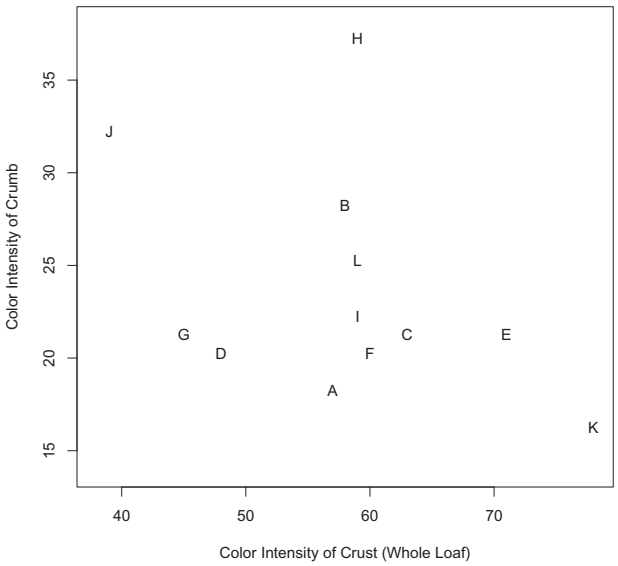

Fig. 1. Color intensity of crumb versus color intensity of crust (whole loaf) for 12 white breads ($A,\ldots,L$).

Panel 1 of Fig. 2 shows that product A is randomly selected as the first product this panelist will evaluate. Product J is then selected because it is the most dissimilar to product A (Panel 2). Now, because there are at least two products selected (A and J), a centroid is calculated. Product K is then selected as it is the most dissimilar from the centroid of products A and J (Panel 3). Finally, product H is selected as it is most dissimilar from the centroid of products A, J and K (Panel 4). Panel 5 gives the shape formed by the selected products, and Panel 6 gives the order that each product will appear to the panelist.

Fig. 2 illustrates a "greedy" selection process that would always be made if the SID was not constructed to adhere to position balance. For example, suppose there are $n = 396$ panelists and we wish to construct a ${}_{12}P_6$ design. To ensure position balance, we require that each product appears $r = 33$ times (where $r = n/p$) in each position, while not letting any panelist rate any one product more than once. Note that, in practice, it will not always be the case that an SID is perfectly balanced. Misuse of the rating apparatus, a non-divisible sample size or neglect on behalf of the panelists are a few of the possible scenarios that could lead to some products being evaluated more than others.

2.1. Formulation

Consider an $n \times k$ matrix where the rows correspond to panelists and the columns represent the order that each product is evaluated (Table 2). From a set of p products, we want to find n subsets of k products such that every subset contains unique elements and, between all subsets, every product appears the same number of times. We allocate products in a column-wise fashion, i.e., we will assign n products to the first position, then we will assign n products to the second position, and so on until all positions are filled.

The construction of an SID is as follows. In the first position, randomly allocate one product to each panelist such that every product appears the same number of times. In the second position, randomly select a panelist such that priority is given to a panelists whose first position contains a product that is very similar to the other products. Assign the randomly selected panelist a second product, j, that belongs to the set of remaining products,

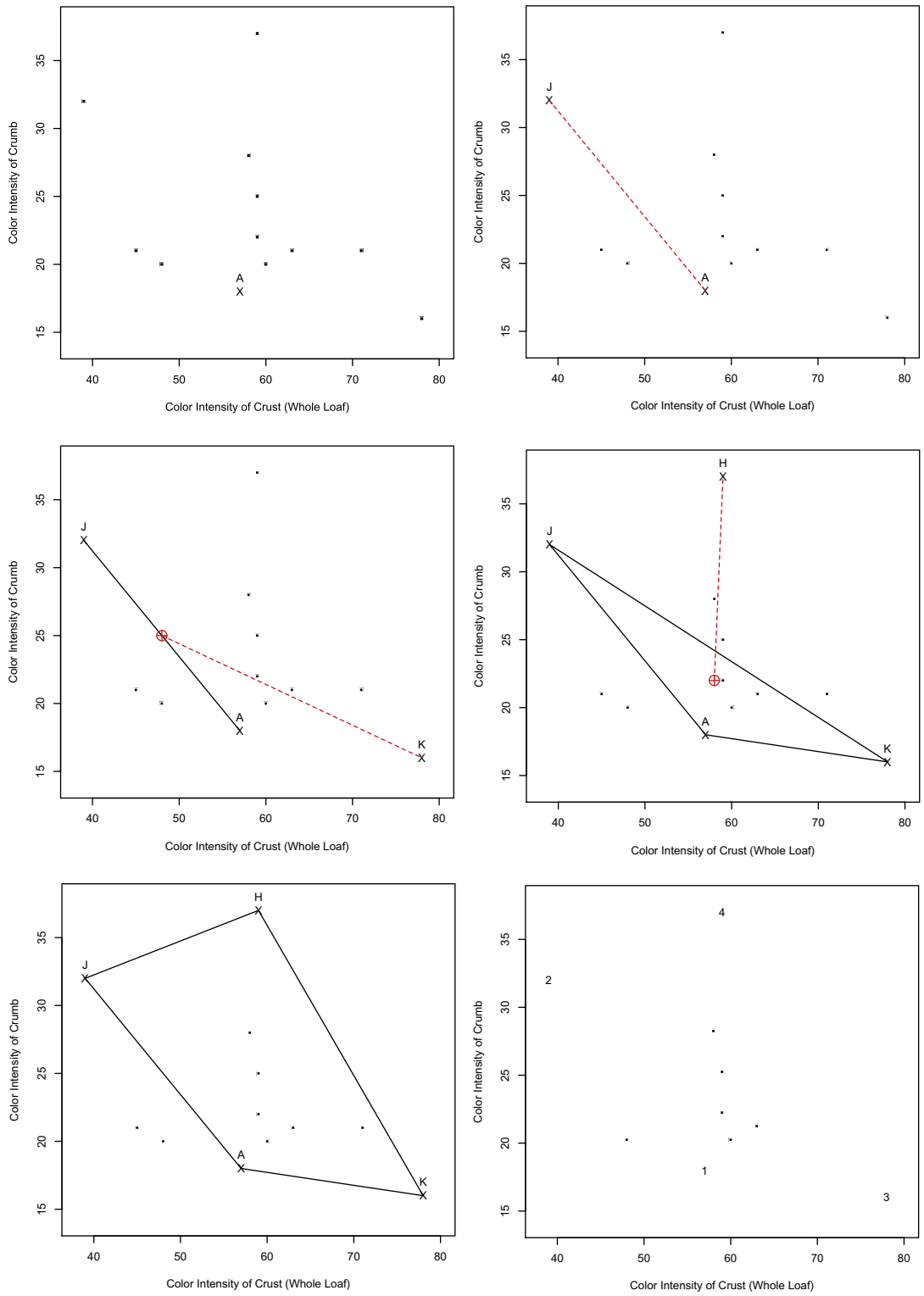

Fig. 2. A demonstration of the SID's product selection process for one panelist. The symbol ⊕ marks the centroid of the selected points and the broken red line segment connects a selected product (or products' centroid) to the most dissimilar product. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

which is as dissimilar as possible to the product in the selected panelist's first position. Repeat this step n times until every panelist's second position is filled. Again, each product is allocated such that it appears the same number of times in the second position. Then, for each position $k \geqslant 3$, randomly select a panelist (with priority given to the panelist's whose first $k-1$ positions contain products that are similar) and assign a product to their kth position that is as dissimilar as possible to the centroid of the first k products assigned to them. This procedure can be summarized, via pseudo-code, as follows:

Initialization:
1. Assign each product a number from 1 to p, where p is the total number of products
2. Compute the Euclidean distance between each target product using their sensory profile.
Note: Herein we will refer to this matrix as `dist`.
Construction:
Create a vector of length p where element is equal to r
Note: Herein we will refer to this vector as `ct.prod`
for l in 1:k
for i in 1:n
if $l = 1$ Randomly assign panelist i one product such that position l is position balanced.
if $l \geqslant 2$ {
Step 1. Randomly select a panelist, with priority given to a panelist whose first position contains a product that is very similar to the other products.
Step 2. Use `dist` to determine which product, j, is the most dissimilar to the product in position 1, or to the centroid of positions 1:l − 1 if l > 2, of the panelist selected in Step 1.
Step 3. Assign product, j, found in Step 2 to position l of the panelist selected in Step 1.
}
Subtract 1 from element j of `ct.prod` where j corresponds to the chosen product.
if any element of `ct.prod` equals 0, no longer assign that product.
end for i in 1 : n
end for l in 1 : k

3. Analyzing the resulting data

3.1. Imputation then cluster analysis

SIDs will contain missing values by design. Little and Rubin (2002) discuss a number of procedures suitable for analyzing data with missing values. These procedures can be categorized according to whether they perform single or multiple imputation (Rubin, 1978). Formally, single imputation procedures replace each missing observation with a single value whereas multiple imputation procedures replace each missing observation with $m \geqslant 2$ values. Notably, multiple imputation shares the same advantages as single imputation while rectifying its shortcomings (Little and Rubin, 2002). After imputation has been carried out, the data can be clustered.

3.2. Simultaneous imputation and cluster analysis

Another approach to analyzing data with missing observations is given in Browne et al. (2013). In this paper, the authors expand upon the work of Ghahramani and Hinton (1997) and McNicholas and Murphy (2008) to develop a Gaussian mixture model-based approach that simultaneously performs imputation and cluster analysis. Formally, a Gaussian mixture model has density

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi_p\left(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right), \tag{1}$$

where $\pi_g > 0$ are the mixing proportions, subject to $\sum_{g=1}^{G} \pi_g = 1$, and $\phi_p\left(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$ is the density of the p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance $\boldsymbol{\Sigma}_g$. It is clear from (1) that a GMM is a convex linear combination of multivariate Gaussian densities and, accordingly, is well suited for performing cluster analysis (see Bouveyron and Brunet, 2012; Bouveyron et al., 2007; Celeux and Govaert, 1995; Fraley and Raftery, 2002; McLachlan and Basford, 1988; McNicholas et al., 2010, for examples). Note that the use of a mixture model for cluster analysis is known as model-based clustering.

The model in (1) has $Gp(p+1)/2$ free parameters in the covariance matrices alone. Therefore, it becomes highly parameterized as p grows, even for relatively small values of p (see Fig. 3). To overcome this issue, constraints can be imposed on the component covariance matrices $\boldsymbol{\Sigma}_g$ to introduce parsimony. Another way to introduce parsimony is to consider a mixture of factor analyzers model, which assumes that each component can be represented in an underlying low-dimensional ($q \ll p$) latent factor space (see Ghahramani and Hinton, 1997). McNicholas and Murphy (2008) combined these two approaches by constraining the covariance matrices in a mixture of factor analyzers model. Browne et al. (2013) use one of the models introduced by McNicholas and Murphy (2008); this model uses a mixture of factor analyzers model with common factor loadings and its density is

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi_p\left(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g\right), \tag{2}$$

where $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{\Psi}_g$ is a $p \times p$ diagonal matrix with positive diagonal entries (cf. McNicholas and Murphy, 2008, 2010). Browne et al. (2013) extend this model to

Table 2
A hypothetical position matrix for the white bread data where each row corresponds to a panelist and the columns represent the order that he or she will evaluate their subset of products.

Panelist	Position					
	1	2	3	4	5	6
1	A	J	M	L	B	H
2	C	K	J	B	D	A
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	D	E	J	F	M	I

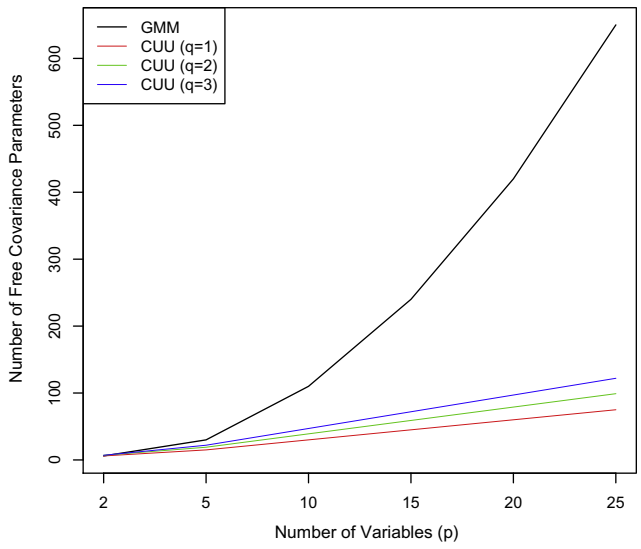

Fig. 3. The number of free parameters in the covariance structure of a two-component Gaussian mixture model (GMM; black line) and a CUU model with $q = 1$ (red line), $q = 2$ (green line), and $q = 3$ (blue line) latent factors for different values of p. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

account for missing data, developing a partial expectation–maximization (PEM) algorithm specifically for the sort of data that will arise from an SID. Details on the model, the algorithm, and its computational efficiency, are given by Browne et al., 2013.

From a practical viewpoint, the model in (2) lets components share covariances while allowing variances to differ from component to component. For our purposes, a component can be considered synonymous with a cluster. Furthermore, the model in (2) has a mere $pq - q(q-1)/2 + Gp$ parameters in its component covariance matrices (see Fig. 3). Following McNicholas and Murphy (2008), we refer to the model in (2) as the CUU model.

Since the work of Dasgupta and Raftery (1998), it has been common practice to use the Bayesian Information Criterion (BIC; Schwarz, 1978) to choose the number of components G and, if applicable, the number of latent factors q. The BIC is given by

$$\text{BIC} = 2l(\mathbf{x}|\hat{\vartheta}) - \rho \log n,$$

where $l\left(\mathbf{x}|\hat{\vartheta}\right)$ is the log-likelihood, $\hat{\vartheta}$ is the MLE of ϑ, ρ is the number of free parameters, and n is the number of observations. We use the BIC to chose the values of G and q in our analyses, noting that is has previously been used for the models of McNicholas and Murphy (2008).

4. Applications

4.1. Two SIDs

We consider two liking studies, undertaken by Compusense Inc., where SIDs were used: one for twelve white breads and one for sixteen brown breads. Based on an SID, panelists rated the products using the nine-point Hedonic scale where a score of (1) indicates an extreme dislike, a score of (5) indicates no preference, and a score (9) indicates an extreme liking. In Sections 4.2 and 4.3, we discuss the data and present results obtained from fitting the CUU model to each SID. The parameters of the CUU model are estimated using the PEM algorithm and the BIC is used to select the numbers of components and factors. Note that the CUU model was fitted using the `CUUimpute` function, available in the `sensory` package (Franczak et al., 2014) for R (Core Team, 2014).

4.2. White bread data

In total, 420 panelists rated six white breads in a ${}_{12}P_6$ SID; for illustration, Table 3 displays the liking scores given by the first six panelists. We fitted the CUU model to the white bread data for $G = 1, \ldots, 4$ components and $q = 1, \ldots, 6$ latent factors. Table 4 gives the BIC values for each CUU model fitted to the white bread data. The CUU model with $G = 2$ components and $q = 1$ latent factor (BIC $= -5080.66$) was selected as the best fitting model.

Table 3
The first six rows of the white bread data.

Panelist	Product											
	A	B	C	D	E	F	G	H	I	J	K	L
1	9		8	6				9			4	8
2	3		8		7		8	7	8			
3		8	6	7					6	9	7	
4			5	4		6		4	3	6		
5			7	7			8	7	6		8	
6				8			3	4	8		7	7

At convergence, the component membership probabilities $\hat{z}_{ig}$ are used to calculate the maximum *a posteriori* (MAP) probabilities of group membership for each panelist. Specifically, $\text{MAP}\{\hat{z}_{ig}\} = 1$ if $\max_g\{\hat{z}_{ig}\}$ occurs in component g and $\text{MAP}\{\hat{z}_{ig}\} = 0$ otherwise. We calculate the mean liking scores for each product based on the MAP probabilities (Fig. 4). While bread *J*, the only ciabatta-style bread in this study, seems to polarize the two groups, they could be largely interpreted as high scorers and low scorers (Fig. 4). The problem of different panelists using the scale in different ways will be discussed further in Section 5.

When constructing the ${}_{12}P_6$ SID for the white bread data, we nested ${}_{12}P_3$ and ${}_{12}P_4$ designs within it. This allows for some validation of the clustering results obtained for the ${}_{12}P_6$ design. The nested designs are not difficult to produce because of the SID's column-wise allocation feature (cf. Section 2.1), which ensures that each nested design is position balanced. Note that column-wise allocation effectively means that we will have nested ${}_pP_k$ SIDs for $k \geqslant 2$.

For both the ${}_{12}P_3$ and ${}_{12}P_4$ designs, the best-fitting CUU models (BIC $= -2506.67$ and BIC $= -3405.32$, respectively) have $G = 2$ components and $q = 1$ latent factor. Comparing the classification rates between the ${}_{12}P_3, {}_{12}P_4$, and ${}_{12}P_6$ SIDs (Table 5) — the classification rate is the number of classification agreements divided by the number of observations — the results indicate agreement between the classifications obtained for each design based on the CUU model, and a closer inspection of each comparison is given in Table 6.

When we use classification rates to compare two classification procedures, it is instructive to consider the nature of the points that are classified into different groups by the two procedures. Model-based clustering approaches, like the CUU model, facilitate this by way of the *a posteriori* probabilities of component membership $\hat{z}_{ig}$. Specifically, if $\hat{z}_{ig} = 1$ then there is certainty, under the fitted model, that the ith panelist belongs to the gth component (or cluster). A value $\hat{z}_{ig} \approx 0.5$, in a $G = 2$ component model, indicates that component membership is effectively a coin toss for observation i under the fitted model. In a $G = 2$ component model, we can think of a point moving nearer to a boundary between the two components as values of $\hat{z}_{ig}$ decrease from 1 towards 0.5 under the fitted model.

Table 6 shows that 80 observations were misclassified between the MAP probabilities obtained for the ${}_{12}P_3$ and ${}_{12}P_4$ designs. Of these 80 observations, 48 (60%) had $\hat{z}_{ig} \in [0.5, 0.9)$. Of the other 340 observations, just 108 (31.76%) had probabilities of group membership in this range. Therefore, observations classified into different components under the ${}_{12}P_3$ and ${}_{12}P_4$ designs are roughly twice as likely to be near component boundaries.

For the ${}_{12}P_3$ vs. ${}_{12}P_6$ comparison, of the 111 misclassified observations, 63 (56.76%) had $\hat{z}_{ig} \in [0.5, 0.9)$. Of the other 309

Table 4
The BIC values for each CUU model fitted to the white bread data, with the largest value highlighted in bold face.

# of components	# of latent factors					
	1	2	3	4	5	6
1	−5273.09	−5312.00	−5351.97	−5395.05	−5430.51	−5461.38
2	**−5080.66**	−5129.73	−5175.32	−5214.10	−5252.54	−5286.03
3	−5113.55	−5159.21	−5206.01	−5242.94	−5280.20	−5312.35
4	−5153.24	−5227.30	−5208.32	−5300.32	−5348.52	−5359.25

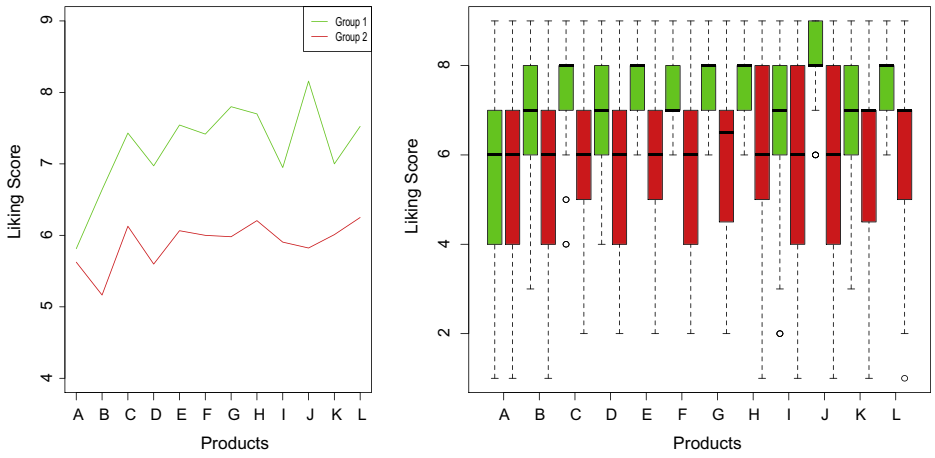

Fig. 4. The average liking scores (left) and box plots (right) for the two clusters of panelists found by the best CUU model for the white bread SID. Note that color code given in Panel 1 is consistent for both plots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Classification rates between the MAP classifications obtained from the best-fitting CUU mixture for each white bread SID.

${}_{12}P_3$ vs. ${}_{12}P_4$	${}_{12}P_3$ vs. ${}_{12}P_6$	${}_{12}P_4$ vs. ${}_{12}P_6$
0.810	0.736	0.831

Table 6
Cross-tabulation of true versus predicted (i.e., MAP) classifications for the three pairs of white bread SIDs.

	${}_{12}P_3$ vs. ${}_{12}P_4$		${}_{12}P_3$ vs. ${}_{12}P_6$		${}_{12}P_4$ vs. ${}_{12}P_6$	
	A	B	A	B	A	B
A	136	44	136	44	197	51
B	36	204	67	173	20	152

observations, 79 (25.57%) had $\hat{z}_{ig}$ values in that range. For the ${}_{12}P_4$ vs. ${}_{12}P_6$ comparison, 42 of 71 (51.15%) misclassified observations had $\hat{z}_{ig} \in [0.5, 0.9)$. In all, 87 of the other 349 (24.93%) observations had $\hat{z}_{ig} \in [0.5, 0.9)$. Therefore, for both the ${}_{12}P_3$ and ${}_{12}P_6$ and ${}_{12}P_4$ and ${}_{12}P_6$ designs, observations classified into different components are more than twice as likely to be near component boundaries. This, taken together with the result of the ${}_{12}P_3$ and ${}_{12}P_6$ comparison, provides a nice validation of our clustering results based on the ${}_{12}P_6$ SID. Furthermore, they suggest that we could have obtained very similar results had a ${}_{12}P_3$ or ${}_{12}P_4$ SID been used. That three of 12 breads might well have sufficed here is, in itself, an interesting result.

4.3. Brown bread data

The brown bread data are larger than the white bread data: there are more panelists and more breads. Specifically, we have an SID where 570 panelists evaluated 6 brown breads in a ${}_{16}P_6$ design. While the total number of breads has increased from 12 to 16, the number presented has remained the same so that 62.5% of the values are missing by design. For illustration, Table 7 shows the liking scores given by the first six panelists. The CUU model was fitted to the brown bread SID data for $G = 1, \ldots, 4$ components and $q = 1, \ldots, 6$ latent factors. The best fitting CUU model (BIC: -7735.94) had $G = 2$ components and $q = 1$ latent factor. The average liking scores for each product for each of the two components, as determined by MAP classifications, are displayed in Fig. 5.

Table 7
The first six rows of the brown bread data. Each panelist evaluated six brown breads using the nine-point Hedonic scale.

Panelist	Product														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	8	6	4						4	6						9
2	8		9						6			7		7		5
3			7		7		6	8				8			7	
4		9				9	6		4			4				8
5	7					6					3			5	8	8
6		9			8		7	8		7			8			

This time, the identified components do not appear to represent either high or low scorers. There is a noticeable contrast in the liking scores between bread B, which was a very seedy bread. Bread N, a hearth bread, is also clearly polarizing. There are also some breads that neither group of panelists particularly liked, e.g., bread D, and others that both groups liked similarly, e.g., bread C. Perhaps the most convincing argument, however, that these results do not reflect usage of the scale is that the average scores for the two clusters are very close (5.75 vs. 5.56).

Again, we consider nested ${}_{16}P_3$ and ${}_{16}P_4$ SIDs. The CUU model was fitted to each nested design for $G = 1, \ldots, 4$ components and $q = 1, \ldots, 6$ latent factors. In both cases, the best fitting CUU models have $G = 2$ components and $q = 1$ latent factor, with BIC values of -3915.65 and -5307.41, respectively. The classification rates between the MAP classifications obtained from the best fitting CUU models for all three designs are given in Table 8. There appears to be good agreement between the classifications for the ${}_{16}P_4$ and ${}_{16}P_6$ designs—looking at the *a posteriori* $\hat{z}_{ig}$ values for the classification disagreements will shed further light on this. However, comparisons involving the ${}_{16}P_3$ designs do not indicate as good a classification agreement. A cross tabulation of the three classification comparisons is given in Table 9.

For the ${}_{16}P_4$ vs. ${}_{16}P_6$ comparison, 94 of 156 (60.26%) misclassified panelists had $\hat{z}_{ig} \in [0.5, 0.9)$. Of the other 414 panelists, only 128 (30.92%) had $\hat{z}_{ig}$ values in that range. Therefore, observations classified into different components under the ${}_{12}P_4$ and ${}_{12}P_6$ designs are roughly twice as likely to be near component boundaries. This suggests that we do indeed have good classification agreement between the ${}_{16}P_4$ and ${}_{16}P_6$ SIDs for the brown bread data. Furthermore, this also tells us that we could have obtained similar results had a ${}_{16}P_4$ SID been used for these data.

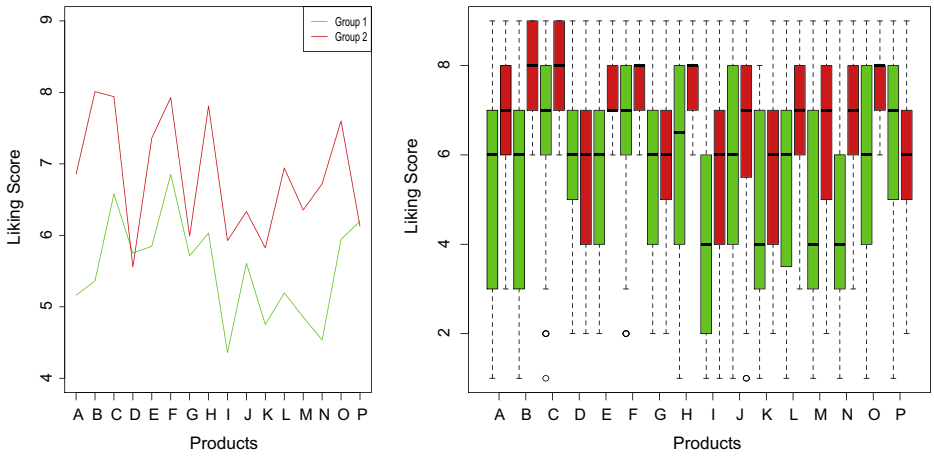

Fig. 5. The average liking scores (left) and box plots (right) for the two clusters of panelists found by the best CUU model for the brown bread SID. Note that the color code given in Panel 1 is consistent for both plots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
Classification rates between the MAP classifications obtained from the best-fitting CUU mixture for each brown bread SID.

${}_{16}P_3$ vs. ${}_{16}P_4$	${}_{16}P_3$ vs. ${}_{16}P_6$	${}_{16}P_4$ vs. ${}_{16}P_6$
0.595	0.595	0.726

Table 9
Cross-tabulation of true versus predicted (i.e., MAP) classifications for the three pairs of brown bread SIDs.

	${}_{16}P_3$ vs. ${}_{16}P_4$		${}_{16}P_3$ vs. ${}_{16}P_6$		${}_{16}P_4$ vs. ${}_{16}P_6$	
	A	B	A	B	A	B
A	92	51	120	23	192	106
B	180	247	208	219	50	222

Comparing the ${}_{16}P_3$ vs. ${}_{16}P_4$ and the ${}_{16}P_3$ vs. ${}_{16}P_6$ SIDs tells a different story. For the ${}_{16}P_3$ vs. ${}_{16}P_4$ comparison, 107 of the 231 (46.32%) panelists had $\hat{z}_{ig} \in [0.5, 0.9)$. Of the other 339 observations, 159 (46.90%) had a probability of group membership in this range. Therefore, misclassified observations are no more likely to be near the cluster boundary than observations that the classifications agree on. For the ${}_{16}P_3$ vs. ${}_{16}P_6$ comparison, 108 of the 231 (46.75%) panelists had $\hat{z}_{ig} \in [0.5, 0.9)$. Of the other 339 observations, 118 (34.81%) had a probability of group membership in this range. Therefore, misclassified observations for these two comparisons are only a little more likely to be near cluster boundaries than observations that the classifications agree on. These comparisons, together with the classification agreement summary in Table 8, tells us that there is relatively poor classification agreement between the ${}_{16}P_3$ and ${}_{16}P_4$, and between the ${}_{16}P_3$ and ${}_{16}P_6$ SIDs for the brown bread data. Furthermore, these results suggest that while a ${}_{12}P_3$ SID might have sufficed for the white bread data, a ${}_{16}P_3$ SID may not have sufficed for the brown bread data.

5. Discussion

We have introduced an approach for generating effective SIDs and illustrated that verifiable clustering results can be obtained therefrom. These SIDs strive to make the products considered by each panelist as different as possible, based on the sensory profiles, while also maintaining position balance. The fact that nested designs can be easily produced makes possible a verification step that is relatively rare in cluster analyses. We wish to emphasize that the work herein should be viewed as sort of blueprint or paradigm for how incomplete-block designs should be constructed and analyzed when used in liking studies. Depending on the application, various steps could be changed along the way. For example, simple Euclidean distance was used to find the distance between products, whereas, in another study, a weighted distance might be desirable.

In the white bread analysis, we observed that the clusters generally corresponded to higher scorers versus lower scorers. There are a few things that could be done about this. For one, each cluster could be investigated for possible heterogeneity therein, leading to a sort of hierarchical clustering procedure. Another approach would be to collect data as ranks, rather than on a hedonic scale. This would, of course, bring its own problems, including having to deal with ties. One thing that we cannot do is standardize panelists' results within the SID framework because doing so would require that every panelist saw the same products. In fact, a nice illustration of why standardization will not work here can be seen by looking at the nested designs.

Acknowledgements

This work was supported by a grant-in-aid from Compusense Inc. and by a Collaborative Research and Development grant from the Natural Sciences and Engineering Research Council of Canada.

References

Ball, R. D. (1997). Incomplete block designs for the minimisation of order and carry-over effects in sensory analysis. *Food Quality and Preference, 8*(2), 111–118.
Bastian, S. E., Collins, C., & Johnson, T. E. (2010). Understanding consumer preferences for shiraz wine and cheddar cheese pairings. *Food Quality and Preference* (7), 668–678. Eighth Pangborn Sensory Science Symposium – Wine Special Issue.
Bouveyron, C., & Brunet, C. (2012). Simultaneous model-based clustering and visualization in the fisher discriminative subspace. *Statistics and Computing, 22*(1), 301–324.
Bouveyron, C., Girard, S., & Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis, 52*(1), 502–519.
Bower, J., & Whitten, R. (2000). Sensory characteristics and consumer liking for cereal bar snack foods. *Journal of Sensory Studies, 15*(3), 327–345.
Browne, R. P., McNicholas, P. D. & Findlay, C. J. (2013). A partial EM algorithm for clustering white breads. arXiv:1302.6625.
Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition, 28*, 781–793.
Core Team, R. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Dasgupta, A., & Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association, 93*, 294–302.

Deppe, C., Carpenter, R., & Jones, B. (2001). Nested incomplete block designs in sensory testing: Construction strategies. *Food Quality and Preference, 12*(5), 281–290.

Ferris, G. (1957). A modified latin square design for taste-testing. *Journal of Food Science, 22*(3), 251–258.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*(458), 611–631.

Franczak, B. C., Browne, R. P., & McNicholas, P. D. (2014). *Sensory: Cluster-Wise Imputation via a PEM algorithm*. R package version 1.0.

Ghahramani, Z. & G.E. Hinton (1997). *The EM algorithm for factor analyzers*. Technical Report CRG-TR-96-1, University of Toronto, Toronto.

Gilbert, J., Young, H., Ball, R., & Murray, S. H. (1996). Volatile flavor compounds affecting consumer acceptability of kiwifruit. *Journal of Sensory Studies, 11*(3), 247–259.

Harker, F. R., Kupferman, E. M., Marin, A. B., Gunson, F. A., & Triggs, C. M. (2008). Eating quality standards for apples based on consumer preferences. *Postharvest Biology and Technology, 50*(1), 70–78.

Hersleth, M., Ilseng, M. A., Martens, M., & Naes, T. (2005). Perception of cheese: A comparison of quality scoring, descriptive analysis and consumer responses. *Journal of Food Quality, 28*(4), 333–349.

Kunert, J., & Sailer, O. (2007). Randomization of neighbour balanced generalized Youden designs. *Journal of Statistical Planning and Inference, 137*(6), 2045–2055.

Lange, C., Martin, C., Chabanet, C., Combris, P., & Issanchou, S. (2002). Impact of the information provided to consumers on their willingness to pay for champagne: comparison with hedonic scores. *Food Quality and Preference, 13*(78), 597–608.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.

MacFie, H. J., Bratchell, N., Greenhoff, K., & Vallis, L. V. (1989). Designs to balance the effect of order of presentation and first-order carry-over effects in hall tests. *Journal of Sensory Studies, 4*(2), 129–148.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker Inc.

McNicholas, P. D., & Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing, 18*, 285–296.

McNicholas, P. D., & Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics, 26*(21), 2705–2712.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., & Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis, 54*(3), 711–723.

Rubin, D. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Section on Survey Research Methods* (p. 20). The Association.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Voorpostel, C. R., Dutra, M. B. d. L., & Bolini, H. M. A. (2014). Sensory profile and drivers of liking for grape nectar among smoker and nonsmoker consumers. *Food Science and Technology (Campinas), 34*(1), 164–173.

Wakeling, I. N., & MacFie, H. J. (1995). Designing consumer trials balanced for first and higher orders of carry-over effect when only a subset of k samples from t may be tested. *Food Quality and Preference, 6*(4), 299–308. Second Sensometrics Meeting.

Williams, E. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry, 2*(2), 149–168.