Journal of Mathematics and Statistics

Original Research Paper

Canonical Correlation-based Tests for the Agreement of Sensory Panelists

[1]Marcela C Rocha, [2]Eric B Ferreira and [3]Daniel F Ferreira

[1]Federal Institute of Education, Science and Technology of South of Minas Gerais, Machado, Brazil
[2]Department of Statistics, Federal University of Alfenas, Alfenas, Brazil
[3]Department of Statistics, Federal University of Lavras, Lavras, Brazil

Article history
Received: 09-10-2019
Revised: 22-11-2019
Accepted: 08-01-2020

Corresponding Author:
Eric B Ferreira
Department of Statistics, Federal University of Alfenas, Alfenas, Brazil
Email: eric.ferreira@unifal-mg.edu.br

Abstract: The reliability of the results of sensory analysis is directly linked to the performance of panel of assessors what, in general, means the ability of judges to identify small differences between products, the replicability of their ratings for the same product and the panel consonance. The panel consonance - usually called unidimensionality - can be understood as the agreement between the judges, thus it reflects the degree of training. Several methods have been proposed for assessing panel unidimensionality but always checking one attribute at a time. We proposed a generalization of the unidimensionality concept based on canonical correlation and enabling to consider several attributes simultaneously.

Keywords: Sensory Assessors, Multivariate Agreement, R Software, Monte Carlo Simulation, Power, Type I Error Rate

Introduction

The search for improving the quality of products and their suitability for the consumer market is constant and, therefore, sensory analysis has been increasingly used in several industry segments, both in product development and in quality control. Its purpose is to interpret, measure and understand the human responses to properties of a product perceived by the senses Hummer (1998) Martens (1999) Dutcosky (2011).

The reliability of the results of sensory analysis is directly linked to the performance of panel that, in general, results from the ability of judges to identify small differences between products, the replicability of their ratings for the same product and the panel consonance Amorim *et al*. (2010).

The panel consonance can be understood as the agreement between the judges, thus it reflects the degree of training. In a trained panel, the judges score the same product in the same way. Thus, the development of techniques to assess the level of agreement among the judges is essential to improve reliability of the sensory evaluation.

Therefore there is an increasing number of studies on that topic. Bi (2003) and Latreille *et al*. (2006) propose methods for assessing agreement between judges based on mixed linear models. Pinto *et al*. (2014) shows Cronbach's alpha index as an alternative for evaluating the consonance. Furthermore, in several studies, the consonance of the sensory panel is measured through its unidimensionality.

The concept of unidimensionality was inserted in the context of sensory analysis by Dijksterhuis (1995), which proposed a method based on Principal Component Analysis (PCA) to evaluate the consonance of a panel, for a given attribute.

Considering an experiment with n products, p judges and q attributes; the data may be arranged in a $\boldsymbol{X}$ hypermatrix of size $n \times q \times p$, with entries x_{ijk}; with $i = 1, 2,\ldots,n$, $j = 1,2,\ldots,q$ and $k = 1,2,\ldots,p$. The hypermatrix $\boldsymbol{X}$ consists of q slices $n \times p$, i.e., taking a matrix $\boldsymbol{X_j}$ is setting up all the scores for the j-th attribute.

Dijksterhuis (1995) states that if a PCA on $\boldsymbol{X_j}$ results in a great first eigenvalue or high variance accounted for the first component then the panel should be considered unidimensional, i.e., there is a high degree of agreement between the judges for the single attribute. The comparison between the first dimension with the others, for the j-th attribute, can be made by:

$$C_j = \frac{\lambda_1^2}{\sum_{2}^{p} \lambda_i^2}$$

where, λ_i^2 is the *i*-th eigenvalue of the matrix $X_j^\top X_j$ Dijksterhuis (1995).

Based on the concept of unidimensionality of a sensory panel, Fernandes (2012) proposed a Monte Carlo Eigenvalues Test (MCET) to infer about the equality of the last *p*-1 last eigenvalues, that is, the null hypothesis established by the author was H_0: $\lambda_2 = \lambda_3 = \ldots = \lambda_p = 0$. According to the author, under the null, all variability should be contained in the first principal component, i.e., the first eigenvalue. Thus, the sum of the *p* eigenvalues must be equal to the first eigenvalue and thus the null hypothesis can be rewritten as $H_0 : \lambda_1 = \sum_{i=1}^{p} \lambda_i$ and, thus, the test statistic is given by:

$$z_c = \frac{\hat{\lambda}_1 - \sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{p} \lambda_i \sqrt{\frac{2}{(n-1)}}} \dot{\sim} N(0,1), \tag{1}$$

which can be used to the test H_0 with significance α.

Alternatively, the test can be done in an empirical distribution, generated as follows. *B* samples are generated from a random variable $X_{p\times 1}$, which are scores vectors from *p* judges regarding *n* products, such that $X_{p\times 1} \sim N(0, \Sigma)$ and Σ represents an equicorrelated structure.

Both the observed and simulated vectors should go under PCA to determine the first eigenvalue ($\hat{\lambda}_1 e \hat{\lambda}_{1\ell}$, respectively, with $\ell \in \{1,\ldots,B\}$) and the test statistic given in (1). Then, an empirical p-value can be computed as p-value = $\frac{\sum_{l=1}^{B} I\left(z_{c\ell} \le z_c\right)}{B}$, where *I* is the indicator function.

Amorim *et al*. (2010) proposed a Monte Carlo Undimensionality Test (MCUT). According to the authors, the variance accounted for the first principal component $\left(\rho_1^2\right)$ is estimated by:

$$R_1^2 = \frac{1+(p-1)\hat{\rho}}{p} \times 100\% \tag{2}$$

where, *p* is the number of panelists and ρ is the equicorrelation between them.

To test the hypothesis $H_0 : \rho_1^2 \ge \rho_{10}^2$, where ρ_1^2 is estimated by the expression (2) and ρ_{10}^2 is a value previously established, one must generate *B* Monte Carlo samples under H_0 of a random variable $X_{p\times 1}$ such that $X_{p\times 1} \sim N(0, \Sigma)$ and:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \rho \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} = \rho,$$

with $\sigma^2 = 1$ without loss of generality.

Thus, the *p*-value for this test is given by *p*-value = $\frac{\sum_{\ell=1}^{B} I\left(R_{1\ell}^2 \le R_1^2\right)}{B}$, where *I* is the indicator function.

In addition to the tests proposed by Fernandes (2012) and Amorim *et al*. (2010), in this work we address two parametric bootstrap tests proposed by Gebert (2010) and the Fujikoshi test proposed by Ferreira (2011) which are essentially tests for retaining principal components and are inserted in the sensory context by Fernandes (2012). It is worth noting that the tests are presented in the same way they have been proposed, i.e., to infer about the variation accounted for the *k* first principal components $\left(\rho_k^2\right)$. However, during the generalization we consider just variance explained by the first principal component, i.e., the particular case where *k* = 1.

The Fujikoshi test (FUJI) has been proposed to assess the proportion of the variance explained by k first principal components of the sample which can be calculated by:

$$R_k^2 = \frac{\sum_{i=1}^{k} \hat{\lambda}_i}{\sum_{i=1}^{p} \hat{\lambda}_i} \tag{3}$$

where $R_k^2 \in [0,1]$ and $\hat{\lambda}_i$ are the eigenvalues of the sample covariance matrix ***S***.

According to Ferreira (2011), the distribution of R_k^2 can be known, using the results given by Fujikoshi (1980), which states that $\sqrt{n-1}\left(R_k^2 - \rho_k^2\right) \dot{\sim} N\left(0, \omega^2\right)$, where ρ_k^2 is the proportion of variance explained by the *k* first principal components and ω^2 is estimated by the expression:

$$\hat{\omega}^2 = \sqrt{\frac{2tr\left(S^2\right)}{\left[tr\left(S\right)\right]^2}\left[\left(\rho_k^2\right)^2 - 2\hat{\beta}\rho_k^2 + \hat{\beta}\right]}$$

and:

$$\hat{\beta} = \frac{\sum_{i=1}^{k} \hat{\lambda}_i^2}{\sum_{i=1}^{p} \hat{\lambda}_i^2}.$$

Based on this result, the test was proposed for the null $H_0 : \rho_k^2 \geq \rho_{k0}^2$ and test statistic was given by the expression:

$$z_c = \frac{R_k^2 - \rho_{k0}^2}{\frac{\omega_0}{\sqrt{n-1}}}$$

where:

$$\omega_0 = \sqrt{\frac{2tr\left(S^2\right)}{\left[tr\left(S\right)\right]^2}\left[\left(\rho_{k0}^2\right)^2 - 2\hat{\beta}\rho_{k0}^2 + \hat{\beta}\right]} \; and \; z_c \sim N(0,1).$$

The construction of the parametric bootstrap tests proposed by Gebert (2010) depend on a normal distribution with *p* variables following:

1. The mean should be the zero vector, that is, μ_b = [0…0].
2. The covariance matrix (Σ_b) should satisfy the condition $\rho_k^2 = \rho_{k0}^2$.

Considering a random sample in $\mathbb{R}^p$ and covariance matrix Σ, the proportion of the variance explained by *k* first principal components $\left(\rho_k^2\right)$ has estimator $\left(R_k^2\right)$ obtained by expression (3).

The matrix Σ_b can be constructed from the spectral decomposition $\Sigma_b = \hat{P}\Lambda_b\hat{P}^\top$, where $\hat{P}^\top$ is the matrix of eigenvectors of *S*, which is the sample covariance matrix and Λ_b is a diagonal matrix defined for the proportion of the variance explained by *k* first principal components is equal to ρ_{k0}^2 Gebert (2010) Gebert and Ferreira (2013).

Given the parameters μ_b and Σ_b, *B* random samples of size *n* are generated from a normal distribution with *p* variables with these parameters so that each bootstrap sample are under the assumption $H_0 : \rho_k^2 \geq \rho_{k0}^2$, $\ell \in \{1,2,\ldots,B\}$.

In the first bootstrap test, known as Parametric Bootstrap Test Based on the Proportion of the Variance Explained by *k* First Principal Components (*PBr*), $R_{k\ell}^2$ is calculated by the expression (3) from ℓth bootstrap sample.

Thus, its p-value is defined as $\frac{\sum_{\ell=1}^{B} I\left(R_{k\ell}^2 \leq R_k^2\right)}{B}$, where *I* is the indicator function.

In the second bootstrap test, named Parametric Bootstrap Test Based on $z_{c\ell}$ (*PBz*), the test statistic is defined by:

$$z_{c\ell} = \frac{R_{k\ell}^2 - \rho_{k0}^2}{\frac{\omega_{0\ell}}{\sqrt{n}}} \quad (4)$$

where, $z_{c\ell} \sim N(0,1)$ and $R_{k\ell}^2$ is calculated according to expression (3) for the ℓth bootstrap sample.

In the same way, the p-value of this test is given by the proportion of cases where the calculated values for $z_{c\ell}$ are below the value of z_c (expression (4) calculated for the original sample), i.e., $\frac{\sum_{\ell=1}^{B} I\left(z_{c\ell} \leq z_c\right)}{B}$, where *I* is the indicator function.

Considering that the use of the tests mentioned above is restricted for one attribute, to assess agreement among the judges, the aim of this study is to generalize such tests. In addition, the objective was to evaluate the performance of generalized tests in terms of power and type I error, via Monte Carlo simulation and recommend the test with better performance.

The generalization proposed here is to use the canonical correlation matrix as input for the tests. Thus, it is possible to infer about the panel agreement considering all attributes simultaneously.

The following section presents the proposed method; the next one describes the simulation study; then the results and discussion section brings the power curves, where the tests can be compared.

On the Proposed Method

Consider the evaluation of n products, for *p* judges that use *q* attributes (variables).

Assume that $X_{ij}^\top = \left[X_{ij1}, X_{ij2}, ..., X_{ijq}\right]$ is a vector of dimensional *q* scored by *j*th judge regarding the *i*th product, with $i \in \{1,2,\ldots,n\}$ and $j \in \{1,2,\ldots, p\}$. Thus, the matrix *X* ($n \times pq$) with the panel scores can be conceived as:

$$X = \begin{bmatrix} X_{11}^\top & X_{12}^\top & \cdots & X_{1j}^\top & \cdots & X_{1p}^\top \\ X_{21}^\top & X_{22}^\top & \cdots & X_{2j}^\top & \cdots & X_{2p}^\top \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{i1}^\top & X_{i2}^\top & \cdots & X_{ij}^\top & \cdots & X_{ip}^\top \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n1}^\top & X_{n2}^\top & \cdots & X_{nj}^\top & \cdots & X_{np}^\top \end{bmatrix}. \quad (5)$$

From the matrix (5) we want to test the hypothesis H_0: The panel is unidimensional.

In order to reduce the dimension of *X* so that it is possible to apply the tests for unidimensionality of sensory panels, the sample covariance matrix (*S*) is replaced by the canonical correlation matrix (*R*). So, the

initial step consisting in constructing the R matrix is described below and is common to all tests.

The generalizations are proposed for the cases where the number of attributes evaluated is greater than one ($q > 1$) and consisting initially in reducing the size for the one-dimensional case, that is, the column of the data matrix dimension is reduced from pq to p. The proposed adaptation is due to the fact that the application of generalized tests in this study only be possible for a fixed attribute. Therefore, the advantage offered by the generalization of the testing is to analyze simultaneously all sensory attributes.

Given two matrices X_j and $X_{j'}$, that contains the scores of two different judges (say j and j'), linear combinations $Y_j = a_j^\top X_j$ and $Y_{j'} = b_{j'}^\top X_{j'}$ are sought such that the sample correlation $r_{Y_j, Y_{j'}}$ is maximized. This maximum is easily obtained using the theory of canonical correlations. If this process is repeated for all pairs of judges, the sample correlation matrix R can be expressed as:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}.$$

Given matrix R, each of the proposed test is finalized by one of the tests for the unidimensionality of sensory panels described above.

It is worth noting that the parametric bootstrap test based on $R_{k\ell}^2$ (PBr), parametric bootstrap test based on $z_{c\ell}$ (Bpz) and Fujikoshi test infer about the proportion of the variation explained by the first principal component. Thus, generalizing the tests involve testing the hypothesis H_0: $\rho_1 \geq \rho_0$ or, equivalently, H_0: $\rho_k^2 \geq \rho_{k0}^2$.

The Monte Carlo Eigenvalues Test infers about the equality of the last eigenvalues. Thus, considering λ_1, λ_2,…, λ_p eigenvalues of ρ, estimated by the respective eigenvalues of R a null hypothesis can be expressed by $H_0 : \lambda_{p+1} = \lambda_{p+2} = \ldots = \lambda_{pq} = 0$.

The steps described earlier in this section are shown in Figure 1.

It is noteworthy that, although there are differences between pairs of hypotheses presented for generalizations, the non-rejection of H_0 in both cases involves the finding that the panel can be unidimensional.

In any case, the idea is that once rejected the null hypothesis of multivariate unidimensionality, unidimensionality tests to be applied for each attribute. Thus, it is possible to see an analogy between the proposed procedure and analysis of variance F test followed by a multiple comparison test.

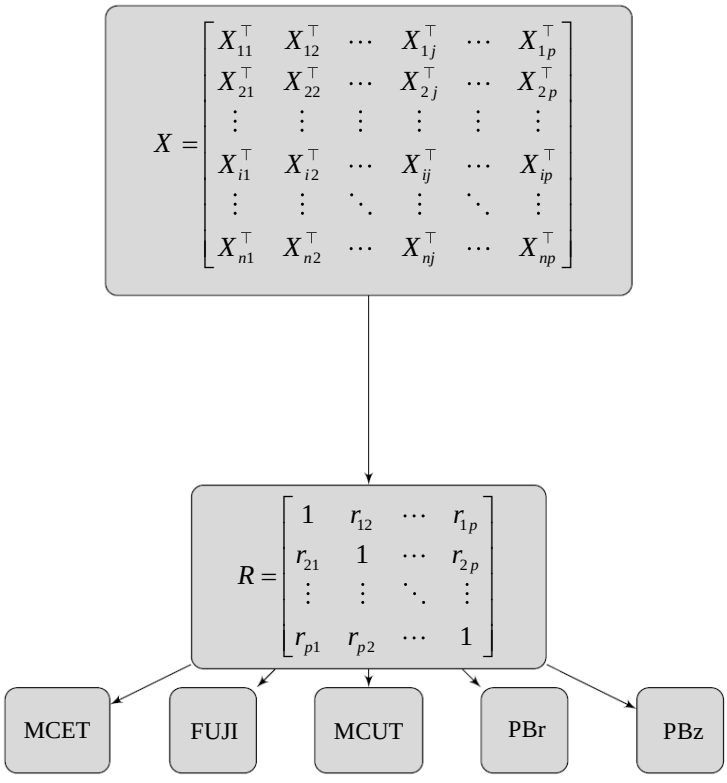

Fig. 1: Representation of the steps of generalization of the tests

Simulation Experiment

Simulations experiments were carried out to evaluate and compare the tests regarding their performance in terms of power and type I error rate. For that reason, routines were written in R code (R CORE TEAM, 2014), both for the implementation of testing and for validation and comparison, which were performed by Monte Carlo simulation.

For the computer simulation, we considered 216 situations, generated from the combination of the number of attributes $q \in \{2,5,10,20\}$, number of products $n \in \{5,10,15,20\}$, number of assessors $p \in \{2,5,10,15\}$ and correlations between assessors $r \in \{0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$. The scenarios were established so that $n > q$ and $n \geq p$.

The degree of training or proportion of the variation accounted for the first principal component $\left(\rho_1^2\right)$, used for constructing graphs, indicates the level of agreement of the sensory panel and varies according to the configuration scenario because it is calculated based on the number of judges (p) and the correlation between them (ρ). As results presented by Fernandes (2012), can be obtained by the expression:

$$\rho_1^2 = \frac{\rho(p-1)+1}{p}$$

For each of the scenarios were simulated N = 1000 Monte Carlo samples. Furthermore, in all experiments, samples were simulated populations with multivariate parameter Σ_q so that the degree of training of raters were confined to the range between 0 and 1.

For the generation of multivariate samples of this study, it should be noted that there is a correlation between sensory attributes (variables) and among the panelists. The covariance between attributes and between judges were fixed by the composition procedure of the data, which follows the linear model:

$$Y_{ijm} = \mu + \tau_j + \beta_m + e_{ijm},$$

where, Y_{ijm} is the observation of the jth variable for the product i, for the judge m, μ is a common constant to all observations (zero without loss of generality), τ_j is the effect of the variable (attribute) j, β_m is the random effect of the judge m and e_{ijm} is the random error associated with Y_{ijm}. The composition of each observation is illustrated in Fig. 2 and explained below.

It was initially generated a hyper zero matrix of size $n \times q \times p$. Thus, the constant $m = 0$ was assigned to all observations.

Then was added the effect of attribute (τ_j) to each of the $n \times p$ slices of the hypermatrix.

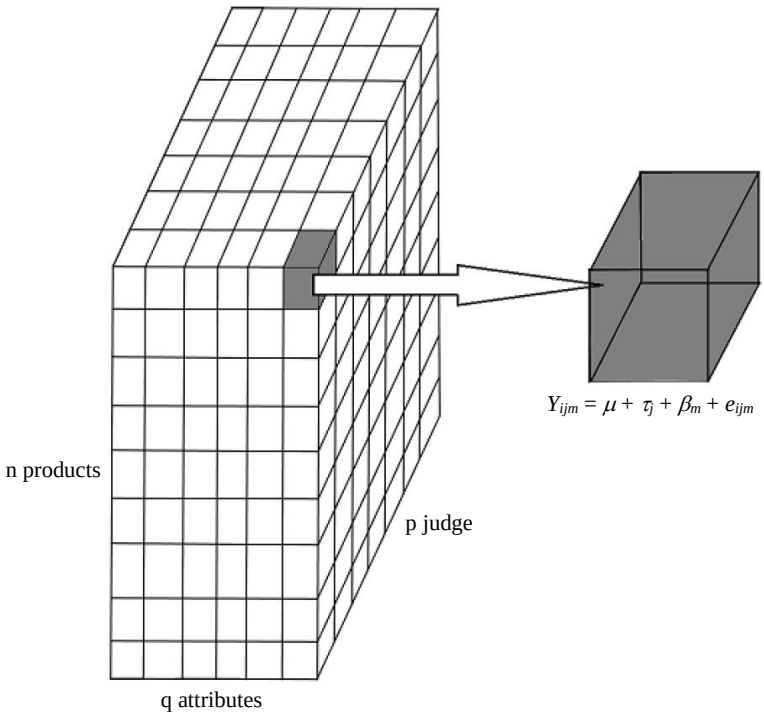

Fig. 2: On the left, the sample, right, an observation Y_{ijm}

For that, q variables were drawn from a distribution $N_q(\mu, \Sigma_q)$, where $\mu = 0$ the mean vector of attributes and Σ_q the covariance matrix between the attributes. In this study, the correlation between the sensory attributes was set at zero.

Similarly, for each of the slices $n \times q$ from the hypermatrix, the effect of the judge m (β_m) was added to all the marks awarded by the same. Thus, for the draw of the effects of judge (settling the quality of the desired training, ρ), was used a distribution $N_p(0, \Sigma_p)$, where:

$$\Sigma_p = \sigma^2 R = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \rho \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}.$$

Note that R is equicorrelated and the correlation (which indicates the quality of training) assumes the values mentioned above $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$. $\sigma^2 = 1$ was set without loss of generality. Finally, the random errors were drawn from a standard normal distribution and added to each of the observations.

In each sample generated were performed all generalized tests, setting up the ideal correlation of $\rho_0 = 0.75$. Were calculated percentages of rejection of the null hypothesis for the $N = 1{,}000$ samples for the nominal significance level $\alpha = 5\%$ and thus the estimator of the Power Function (PF), that is, the proportion of rejections of H_0 is given by:

$$PF = \frac{\sum_{\ell=1}^{N} I\left(p - value_{\ell} \leq \alpha\right)}{N} \quad (6)$$

where, $p\text{-}value_\ell$ is the p-value of ℓth Monte Carlo sample simulated and I is the indicator function. The significance α was set to 5% for all cases.

Results and Discussion

The following are the results of Monte Carlo simulation which brings the graphs for the performance of the Monte Carlo eigenvalues test, Fujikoshi test, parametric bootstrap test based on $R^2_{k\ell}$, parametric bootstrap test based on $z_{c\ell}$ and Monte Carlo undimensionality test, generalized in this work.

In the figures, the dotted line parallel to the ordinate at the point where the degree of training corresponds to the correlation between the judges equal to 0.75 is the separation between the regions under H_1 and under H_0 and the horizontal dotted line determines the level of significance ($\alpha = 0.05$).

To verify the existence of differences between the nominal level of significance adopted (α) and type I error rates we calculated the exact confidence interval for proportion, with 99% of confidence whose extremes are represented by the dashed line parallel to the abscissa.

Although some figures have been omitted due to the space limitation, the detailed description of the behavior of the tests is shown below.

Evaluating $q = 2$ attributes and $p = 2$ judges, Fujikoshi test do not reject the null hypothesis for all levels of training set, while the Monte Carlo undimensionality test and bootstrap parametric tests (PBr and PBz) reject forever. As we can see in Fig. 3, this result is recurrent, even with the increased number of products.

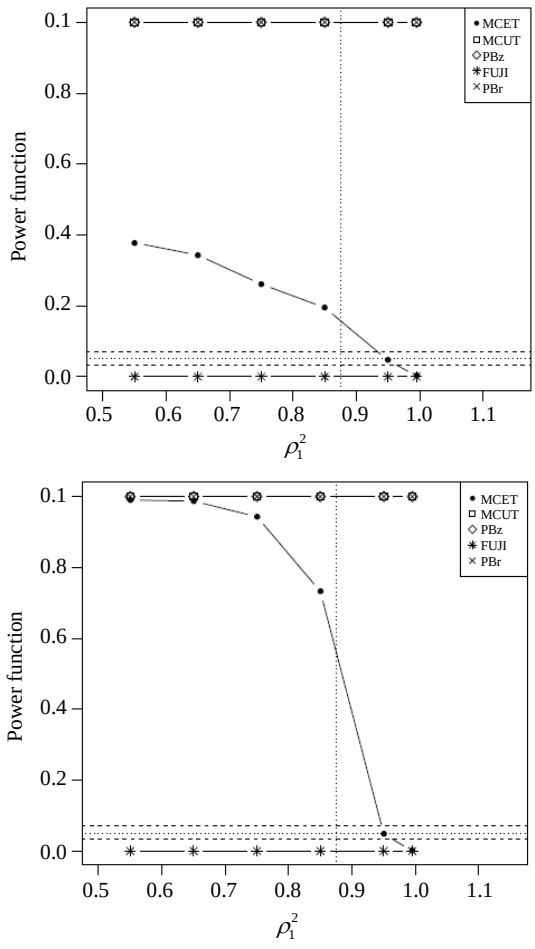

Fig. 3: Power of tests MCET, FUJI, MCUT, PBr e PBz, considering $q = 2$ attributes, $p = 2$ judges, $n = 5$ products (left) and $n = 20$ products (right)

In these scenarios, the Monte Carlo eigenvalues test presents low power (37:8%) analysing n = 5 products and not control type I error rates in the region close to the level of training taken as reference, but its power grows with the increasing number of products (n), exceeding 85% for n = 20. Furthermore, the Monte Carlo eigenvalues test takes control of Type I error rates were considered when ten or more products.

Thus, analyzing the scenarios with two attributes, the power of Monte Carlo eigenvalues test improves with increasing number of products (n). Can also be noted that, as shown in Fig. 4, the control of Type I error rates of the Monte Carlo test of the eigenvalues decreases with increasing the number of attributes (q), while the performance of other tests remains unchanged.

For the minimum number of assessors (p = 2), only the MCET differentiated the simulated scenarios under H_0 of the scenarios under H_1. When evaluating the performance of tests for an attribute, Gebert (2010) obtained different results, as it considered satisfactory performance of the parametric bootstrap tests when considered p = 2 products. Amorim *et al*. (2010), to propose and evaluate the performance of the MCUT to one sensory attribute, states the test was more liberal to a small number of judges, but in general, obtained performed well.

From Fig. 5 it can be seen that, if compared with the cases where p = 5, the FUJI, MCUT and parametric bootstrap tests showed improvement in their performance.

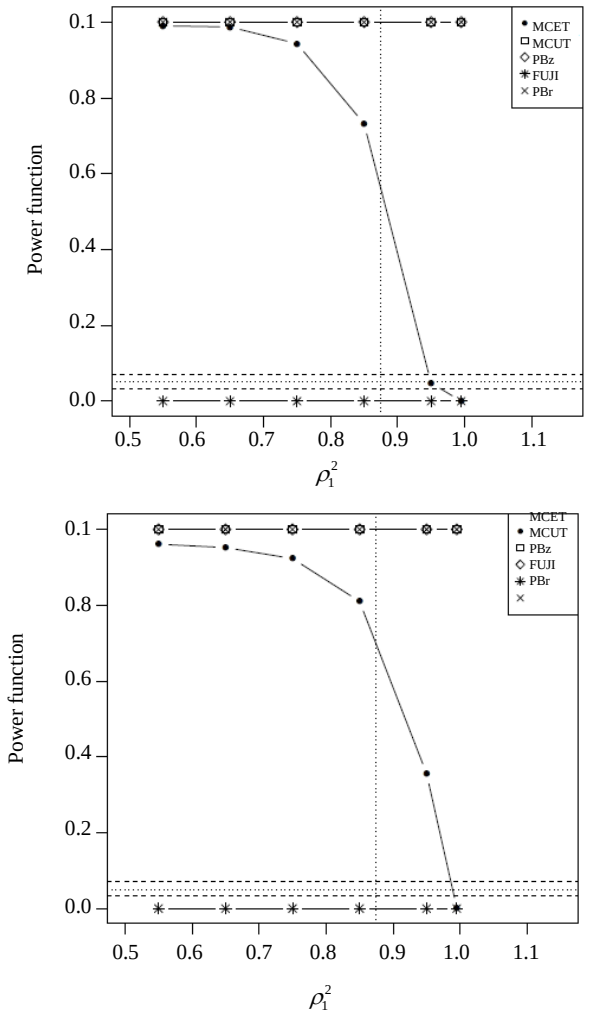

Fig. 4: Power of tests MCET, FUJI, MCUT, PBr e PBz, considering q = 10 attributes (left) and q = 15 attributes (right), n = 20 products and p = 2 judges.

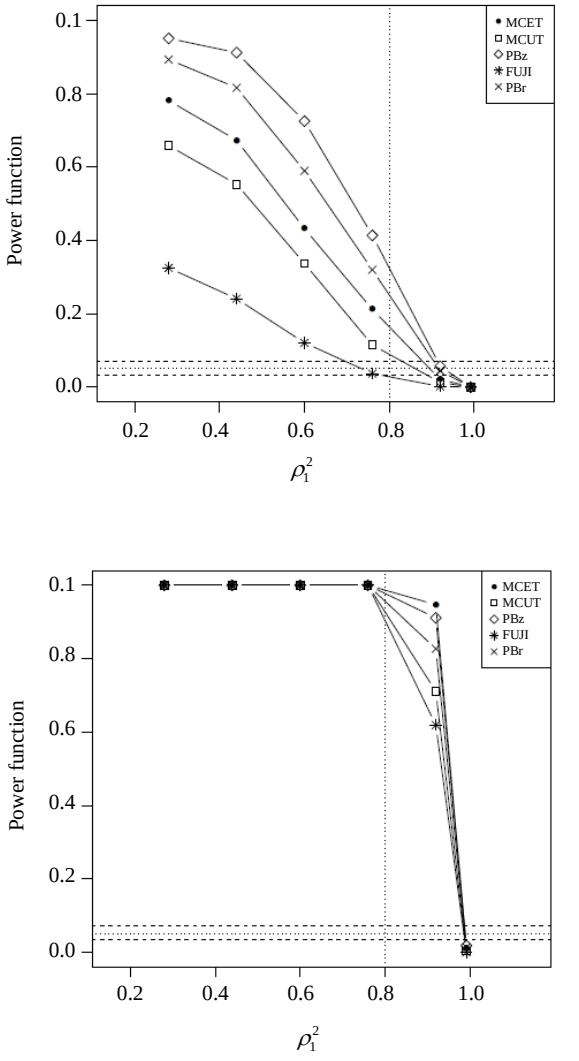

Fig. 5: Power of tests MCET, FUJI, MCUT, PBr e PBz, considering q = 2 attributes, p = 5 judges, n = 5 products (left) and n = 20 products (right)

According to Fig. 5 and considering a small number of attributes (q = 2), the power presented by the parametric bootstrap tests based on $R^2_{k\ell}$ and parametric bootstrap based on $z_{c\ell}$ (89.27% and 95.04%, respectively) exceeded the power of tests on eigenvalues Monte Carlo (78.10%) and MCUT (65.80%). Considering 5 judges and the minimum number of attributes (q = 2), although it has shown a low power when n = 5 (below 35%), the FUJI showed an improvement over the power with the increasing number of products and in such cases, control better Type I error rates in all situations, unlike what happened with the other tests.

In general, for $p \geq 5$, the FUJI was slightly less liberal than the other tests on some of the simulated scenarios. In tests performed for one attribute, according to Fernandes (2012), the FUJI was more liberal than the others. The performance of the FUJI contrasts with the results of Gebert (2010) that when comparing the tests for one attribute, said the FUJI had underperformed to bootstrap tests.

Furthermore, as shown in Fig. 6, with the increase in the number of attributes (q), the MCET, MCUT, FUJI, PBr and PBz start to have power curves increasingly similar to each other, while the FUJI continues slightly lower than the other tests in relationship to power, but looks better control the type I error rate.

Power curves of the two bootstrap parametric tests, MCET, MCUT and FUJI approaches to increase the number of assessors. A similar result was found by Fernandes (2012), which compared the performance of the tests for q = 1 attribute.

Figure 7 shows results for q = 2 and p = 10. In these cases, the curves of power of tests are closer to each other. In general, the tests have high power, but do not control the type I error rate in the region near the critical point.

Considering further p = 15 and $q \geq 10$, it is noted from Fig. 10 that, with increasing the number of attributes, the bootstrap tests did not control the Type I error rate in any scenario and the MCUT is liberal when $n = p$.

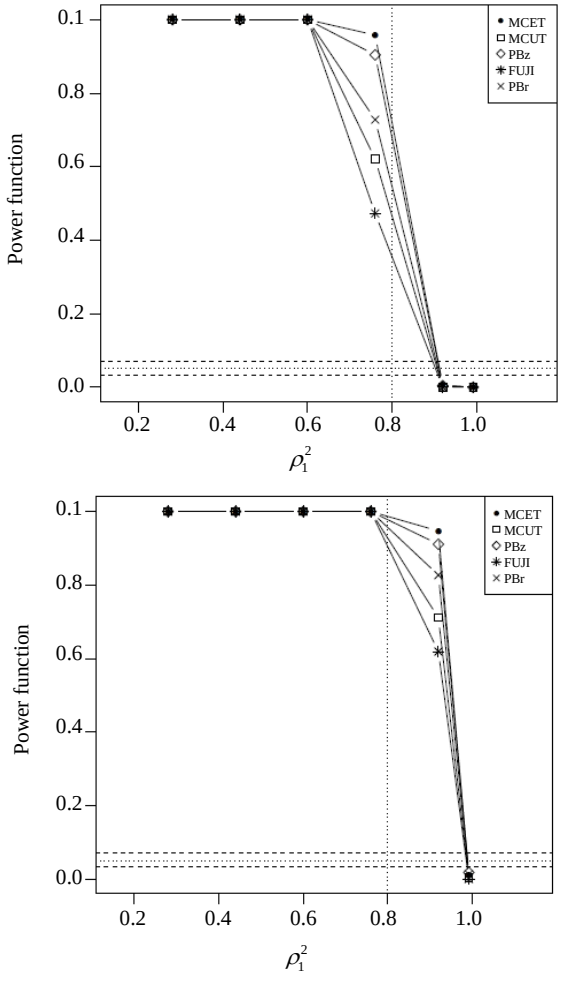

Fig. 6: Power of tests MCET, Fuji, MCUT, Bp e Bpz, considering q = 5 attributes (left) and q = 15 attributes (right), n = 20 products and p = 5 judges

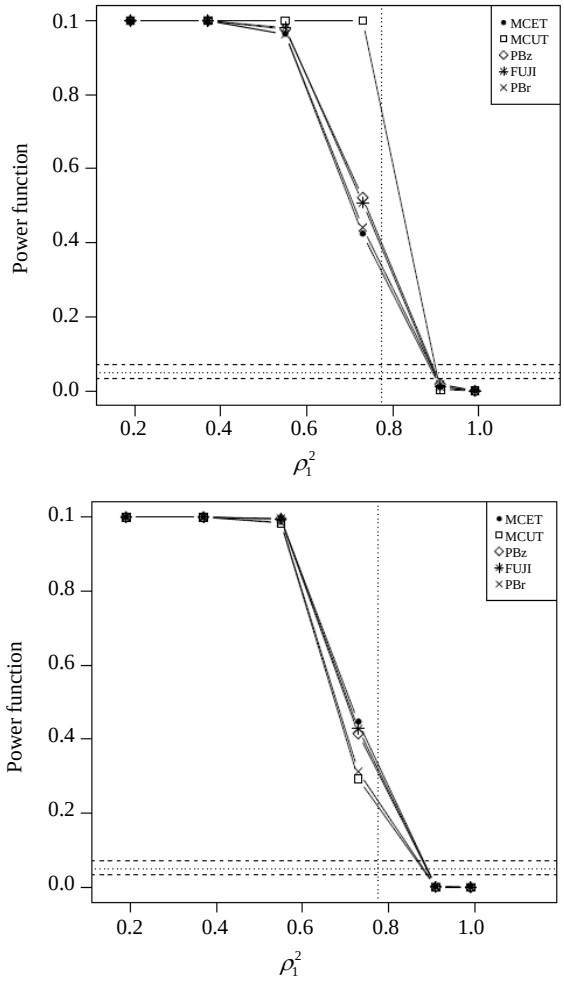

Fig. 7: Power of tests MCET, Fuji, MCUT, Bp e Bpz, considering q = 2 attributes, p = 10 judges, n = 10 products (left) e n = 15 products (right)

In scenarios where $p \geq 5$ increasing the number of products (N) resulted in a improves the behavior of all the tests with increased power and increased control in type I error rate, the result was also obtained by Fernandes (2012) for $q = 1$.

Increasing the number of attributes (q) results in an increase of type I error rates and this growth was more detrimental to the performance of the parametric bootstrap tests. However, increasing the number of products (n) improves the control of Type I error rate, as in the simulations performed by Gebert (2010), Amorim *et al*. (2010) and Fernandes (2012) which evaluated the performance of tests for an attribute.

It is worth noting that the performance of the tests is best whenever $n \geq pq$, since they have higher power and better control of type I error rates. However, this situation is little used in practice in sensory analysis, because it is quite common for the number of attributes and/or the number of panelists products exceeds the number of products.

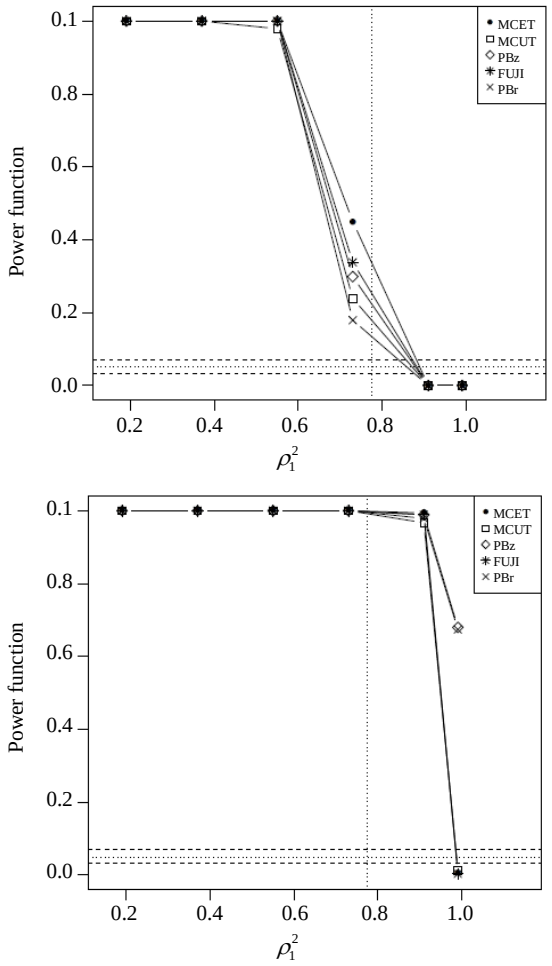

Fig. 8: Power of tests MCET, Fuji, MCUT, Bp e Bpz, considering $q = 2$ attributes (left) and $q = 15$ attributes (right), $n = 20$ products and $p = 10$ judges

The FUJI and MCET have their power curves almost coincident when $p = 10$ and further show more similar with increasing the number of attributes (q) and the number of products (n), in this situation, the two bootstrap parametric tests and MCUT are more liberal insofar as it increases the number of attributes (as shown in Fig. 8). Furthermore, all tests show a slight improvement in the control of type I error rate with increasing of n.

In Figure 9 the results of computer simulation are shown in the scenarios composed of two attributes and fifteen assessors. Note that tests have high power (100% in most scenarios simulated under H_1) and show similar behavior to each other. Furthermore, the proximity of its power curve is proportional to number of products (n).

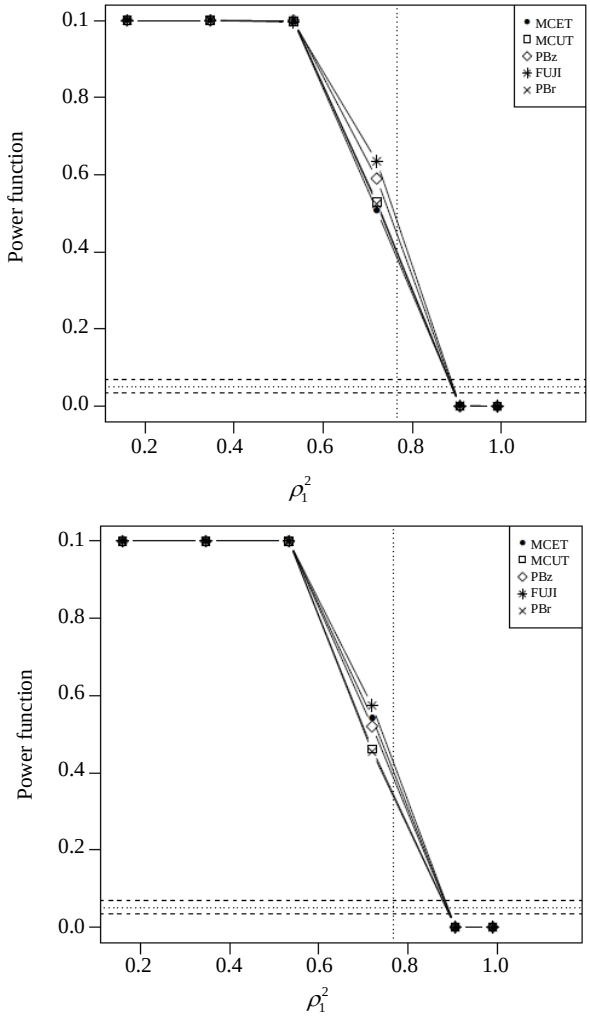

Fig. 9: Power of tests MCET, Fuji, MCUT, Bp e Bpz, considering $q = 2$ attributes, $p = 15$ judges, $n = 15$ products (left) and $n = 20$ products (right)

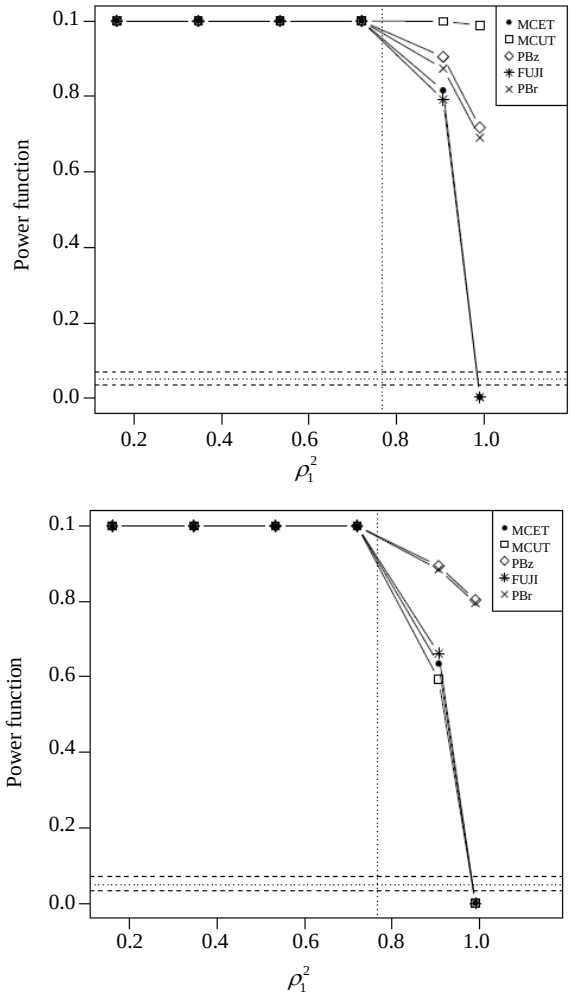

Fig. 10: Power of tests MCET, Fuji, MCUT, Bp e Bpz, considering q = 10 attributes, p = 15 judges, n = 15 products (left) and n = 20 products (right)

Final Remarks

Given the need to evaluate the unidimensionality of sensory panels considering all sensory attributes simultaneously, were proposed generalizations of five tests for retention of principal components available in the literature.

The Monte Carlo eigenvalues test was higher than the other tests when the number of judges was minimal (p = 2) because it was the only one that differentiated the simulated situations under H_0 those simulated under H_1. Thus, other tests are not recommended for a small number of judges.

For large samples, especially with the increasing number of attributes (q), the Monte Carlo eigenvalues test and Fujikoshi test are more recommended than the parametric bootstrap tests and the Monte Carlo test for unidimensionality because have high power and better control the type I error rate.

In general, the proposed generalizations are liberal in the vicinity of the critical point, particularly for a large number of attributes. However, the control of the Type I error rate is better as increases number of products.

Author's Contributions

Marcela C Rocha: She was the one who effectively wrote the monography and participated of the manuscript writing.

Eric B Ferreira: Proposed the research problem, designed the simulation study and participated of the manuscript writing.

Daniel F Ferreira: Proposed the theoretical solution of the problem, checked the program and participated of the manuscript writing.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Amorim, I.S., E.B. Ferreira, R.R. Lima and R.G.F.A. Pereira, 2010. Monte carlo based test for inferring about the unidimensionality of a brazilian coffee sensory panel. Food Quality Preference, 21: 319-323. DOI: 10.1016/j.foodqual.2009.08.018

Bi, J., 2003. Agreement and reliability assessments for performance of sensory descriptive panel. J. Sensory Stud., 18: 61-76. DOI: 10.1111/j.1745-459X.2003.tb00373.x

Dijksterhuis, G., 1995. Assessing panel consonance. Food Quality Preference, 6: 7-14. DOI: 10.1016/0950-3293(94)P4207-M

Dutcosky, S., 2011. Analise Sensorial de Alimentos. 3rd Edn., Champagnat, Curitiba, ISBN-10: 857292244X, pp: 426.

Fernandes, F.M.O., 2012. Proposta de um Teste Monte Carlo para unidimensionalidade de pain´eis sensoriais. Doutorado em Estatıstica e Experimentacao Agropecuaria, Universidade Federal de Lavras, Lavras, MG.

Ferreira, D.F., 2011. Estatıstica Multivariada. 2nd Edn., UFLA, Lavras, pp: 676.

Fujikoshi, Y., 1980. Asymptotic expansions for the distributions of the sample roots under nonnormality. Biometrika, 67: 45-51. DOI: 10.1093/biomet/67.1.45

Gebert, D., 2010. Proposta de testes bootstrap para inferir sobre o n´umero de componentes principais retidos. Doutorado em Estatıstica e Experimentacao Agropecuaria, Universidade Federal de Lavras, Lavras, MG.

Gebert, D. and D. Ferreira, 2013. Parametric bootstrap tests for determining the number of principal components. J. Stat. Theory Pract., 8: 674-691. DOI: 10.1080/15598608.2013.828337

Hummer, S., 1998. Application of multivariate analysis of variance. Food Quality Preference, 9: 83-85. DOI: 10.1016/S0950-3293(97)00034-7

Latreille, J., E. Mauger, L. Ambroisine, M. Tenenhaus and M. Vincent *et al*., 2006. Measurement of the reliability of sensory panel performances. Food Quality Preference, 55: 365-369. DOI: 10.1016/j.foodqual.2005.04.010

Martens, M.A., 1999. A philosophy for sensory science. Food Quality Preference, 10: 233-234. DOI: 10.1016/S0950-3293(99)00024-5

Pinto, F.S.T., E.M. Qannari and F.S. Fogliatto, 2014. A method for panelists consistency assessment in sensory evaluations based on the cronbachs alpha coefficient. Food Quality Preference, 32: 41-47. DOI: 10.1016/j.foodqual.2013.06.006

R Core Team, 2014. R: A language and environment for statistical computing. Vienna, Austria. R Core Team.