Advances in Data Analysis and Classification (2021) 15:987–1013
https://doi.org/10.1007/s11634-021-00441-y

REGULAR ARTICLE

REMAXINT: a two-mode clustering-based method for statistical inference on two-way interaction

Zaheer Ahmed[1] · Alberto Cassese[1] · Gerard van Breukelen[1,2] · Jan Schepers[1]

Received: 12 October 2020 / Revised: 18 March 2021 / Accepted: 31 March 2021 /
Published online: 27 April 2021

Abstract

We present a novel method, REMAXINT, that captures the gist of two-way interaction in row by column (i.e., two-mode) data, with one observation per cell. REMAXINT is a probabilistic two-mode clustering model that yields two-mode partitions with maximal interaction between row and column clusters. For estimation of the parameters of REMAXINT, we maximize a conditional classification likelihood in which the random row (or column) main effects are conditioned out. For testing the null hypothesis of no interaction between row and column clusters, we propose a $max - F$ test statistic and discuss its properties. We develop a Monte Carlo approach to obtain its sampling distribution under the null hypothesis. We evaluate the performance of the method through simulation studies. Specifically, for selected values of data size and (true) numbers of clusters, we obtain critical values of the $max - F$ statistic, determine empirical Type I error rate of the proposed inferential procedure and study its power to reject the null hypothesis. Next, we show that the novel method is useful in a variety of applications by presenting two empirical case studies and end with some concluding remarks.

Keywords Two-mode clustering · Conditional classification likelihood · Interaction effect parameters · $max - F$ test

Mathematics Subject Classification 62-07: Data analysis · 62F03: Hypothesis testing · 62F40: Bootstrap, jackknife and other resampling methods · 62H30: Classification and discrimination; cluster analysis · 62J10: Analysis of variance and covariance

✉ Jan Schepers
jan.schepers@maastrichtuniversity.nl

[1] Department of Methodology and Statistics, Graduate school of Psychology and Neuroscience, Maastricht University, P.O. Box 616, 6200 Maastricht, MD, The Netherlands

[2] Department of Methodology and Statistics, School CAPHRI, Care and Public Health Institute, Maastricht University, Maastricht, The Netherlands

1 Introduction

Many studies in psychology, agriculture, bioinformatics, social networking and marketing, among others, yield two-way two-mode data (Caroll and Arabie 1980). Typically, two-way two-mode data can be represented in an I by J data matrix $\mathbf{D} = \{d_{ij}\}$. Often the rows and columns of this data matrix constitute the levels of two categorical predictors X and Y, say, with I and J categories, respectively, and cell entries denote the observed values of a single quantitative dependent variable D. Such data matrices are collected, for instance, in contextualized personality research, where a set of I individuals (labeled by $i = 1, \ldots, I$) is measured on some behavior of interest D in J different situations (labeled by $j = 1, \ldots, J$). Other examples include the study of micro-array data in genome research where DNA expression level D is obtained for I genes under J different conditions, agriculture studies where crop yield per hectare D is recorded for crops of I different genotypes at J different locations, and consumer research where a preference rating D is obtained for I customers on J products.

A question of scientific interest is whether there is an interaction between X and Y, and, if so, understanding the nature of this interaction. For instance, studying *person by situation* interaction is a critical challenge for researchers in contextualized personality psychology (Geiser et al. 2015; Mischel and Shoda 1995, 1998). This challenge implies finding out whether the situation effect is the same for all individuals and, if not, in what way the situation effect differs between individuals. For example, individuals may be characterized through distinctive sensitivities to specific types of frustrating situations, such as responding more aggressively as a result of being let down by others versus as a result of being narcissistically offended. Such individual-specific response patterns across situations are referred to as *behavioral signatures*, are considered central to the study of personality (Shoda et al. 2013, 2015), and imply the existence of a person by situation interaction. Another example is in the field of measurement (Choudhary and Nagaraja 2017). Specifically, in the study of agreement between measurement methods administered to a set of objects, it is important to determine, in case two (or more) methods (e.g., questionnaires, raters, consumers, etc.) yield different measurements (e.g., intelligence, blood pressure, etc.), whether that difference is attributable only to an additive constant that differs between methods (i.e., no interaction between object and method) or whether that method effect depends on the object being measured (i.e., *object by method* interaction). In the latter case it is of interest to understand the nature of that object by method interaction. A third example is in the case of experimental studies in which each study unit (e.g., patient, lab animal, etc.) is repeatedly observed on an outcome variable across different experimental conditions. A scientific question of importance is whether condition effects are equal for all study units or not. For instance, arguments in favor of the need for personalized medicine (Hamburg and Collins 2010; Collins and Varmus 2015) are based on the assumption that *patient by treatment* interaction (i.e., treatment effect heterogeneity) exists (Rothwell 1995) and should be taken into account in the assignment of patients to treatments. Additional examples of research problems pertaining to the study of interaction in two-mode data are discussed in Schepers et al. (2017).

In these examples there are two issues that preclude a traditional (mixed) two-way ANOVA approach to the study of interaction. The first is that two-mode data typically imply a single replication per cell of **D**. In that case, any row by column interaction is in the model fully confounded with residual effects, ruling out the possibility of statistical hypothesis testing on the presence of row by column interaction. Furthermore, two-mode data typically include a large number of elements for at least one of the two modes (e.g., subjects) in which case the two-way ANOVA model is not very useful if one wishes to understand the nature of the row by column interaction. The amount of information is simply overwhelming and some means to describe it succinctly is needed. To address these issues, several methods have been proposed in the literature (see e.g., Tukey 1949; Mandel 1971; Corsten and Denis 1990; Denis and Gower 1994; Post and Bondell 2013; Franck et al. 2013; Forkman and Piepho 2014). These methods have in common that a specific type of structure is imposed on the row by column interaction (Alin and Kurt 2006). Another approach recently proposed is maximal interaction two-mode clustering (Schepers et al. 2017), which belongs to the more general class of two-mode clustering or biclustering methods (Van Mechelen et al. 2004; Madeira and Oliveira 2004).

Maximal interaction two-mode clustering yields simultaneous partitions of the rows and columns of **D** such that in each bicluster the set of observed pairwise interactions between rows and columns is optimally approximated by a single bicluster interaction parameter. It is based on a probabilistic model, the parameters of which are estimated by the maximum likelihood method. There are two issues not yet addressed by this methodology. Firstly, in many applications the probabilistic model may be unrealistic since it assumes fixed row (e.g., subject) and column (e.g., situation) main effects. Often the elements of at least one of the two modes are a random sample from a population. For instance, in psychology studies, the set of subjects that are included in a study is typically a random sample from some population of individuals. Secondly, no procedure that allows for generalizability across replications, either with the same subjects or with a new sample of subjects, has been discussed yet. In this paper, we extend maximal interaction two-mode clustering to address these two issues. Firstly, we relax the fixed effects assumption for one of the two modes (e.g., subjects) and show how maximum likelihood estimators of the interaction effect parameters are obtained under this relaxed model. Secondly, we develop a method to perform statistical inference on the interaction effect parameters. Specifically, we develop an omnibus statistical test for the null hypothesis of no interaction between row and column clusters. A strong feature of the test statistic used to perform this test is that it is pivotal with respect to the parameters that are not of interest, as we will show in Sect. 2.2.

The remainder of this article is organized as follows: In Sect. 2 we relax the fixed effects assumption for one of the two modes and discuss estimation of the parameters of interest. Furthermore, in Sect. 3 we propose a method for statistical inference on the interaction effect parameters that is based on a Monte Carlo approach. Section 4 discusses the design and results of a simulation study in which two aspects of the newly proposed inferential test are studied: empirical Type I error rate and statistical power. Subsequently, Sect. 5 presents applications of the proposed method on two real case studies. Finally, Sect. 6 includes concluding remarks and possible directions for future research.

2 Methods

2.1 Maximal interaction two-mode clustering model with fixed and random effects

Recently, a statistical model for maximal interaction two-mode clustering was proposed (Schepers et al. 2017) which we will refer to as the MAXINT model. This development extended earlier work by Bock (1980), who proposed a maximal interaction clustering criterion. The latter, however, lacked a statistical justification in terms of a stochastic model that describes a generating process for the observable data. Making the stochastic model explicit is valuable as it clarifies the conditions under which a proposed clustering method can be useful (Bock 1996; Banfield and Raftery 1993).

The MAXINT model assumes a simultaneous partition of the row set R into P row clusters R_p $(p = 1, \ldots, P)$ and the column set C into Q column clusters C_q $(q = 1, \ldots, Q)$. Furthermore, $\bigcup_p R_p = R$ (i.e., row clusters are jointly exhaustive) and $R_p \bigcap R_{p\prime} = \phi$ for all $p \neq p\prime$ (i.e., row clusters are mutually exclusive). Likewise, $\bigcup_q C_q = C$ and $C_q \bigcap C_{q\prime} = \phi$ for all $q \neq q\prime$. A bicluster $R_p \times C_q$ is defined as the Cartesian product of row cluster R_p and column cluster C_q. Formally, the model is expressed as:

$$d_{ij} = \mu + \alpha_i + \beta_j + \gamma_{pq} + \epsilon_{ij} \qquad i \in R_p, j \in C_q, p = 1, \ldots, P, q = 1, \ldots, Q. \tag{1}$$

According to this model, an observed data value d_{ij} can be decomposed as a sum of several terms: an overall mean (μ), fixed (additive) main effects of, respectively, row i (represented as α_i) and column j (represented as β_j), a fixed interaction effect γ_{pq} associated to the bicluster $R_p \times C_q$, and a residual term ϵ_{ij}. The residual terms ϵ_{ij} represent the only stochastic component of the MAXINT model and are assumed to be iid normally distributed random variables with mean zero and variance σ^2. The interaction effect γ_{pq} is constant for all observations (ij) that belong to the same bicluster $R_p \times C_q$. Interaction is thus assumed to occur between row and column clusters, but not between rows and columns within the same bicluster.

MAXINT finds row and column partitions by maximizing a specific likelihood function (see Eq. 3) based on Model (1), in which the following identifiability constraints are imposed on the fixed effect parameters:

$$\sum_{i=1}^{I} \alpha_i = 0, \tag{2a}$$

$$\sum_{j=1}^{J} \beta_j = 0, \tag{2b}$$

$$\sum_{p=1}^{P} \sum_{i \in R_p} \gamma_{pq} = \sum_{p=1}^{P} |R_p| \gamma_{pq} = 0, \tag{2c}$$

$$\sum_{q=1}^{Q}\sum_{j\in C_q}\gamma_{pq}=\sum_{q=1}^{Q}|C_q|\gamma_{pq}=0, \tag{2d}$$

where $|R_p|$ and $|C_q|$ denote the cluster cardinalities of row cluster R_p and column cluster C_q, respectively. Let us denote the vector of row main effects as $\boldsymbol{\alpha}=\{\alpha_i\}$ $(i=1,\ldots,I)$, the vector of column main effects as $\boldsymbol{\beta}=\left\{\beta_j\right\}$ $(j=1,\ldots,J)$, the matrix of interaction effects as $\boldsymbol{\gamma}=\left\{\gamma_{pq}\right\}$ $(p=1,\ldots,P$ and $q=1,\ldots,Q)$ and the vector of location parameters $\boldsymbol{\theta}=\{\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}\}$. Furthermore, let $\mathcal{R}$ and $\mathcal{C}$ represent row and column partitions, respectively, that is, $\mathcal{R}=\{R_1,\ldots,R_P\}$ and $\mathcal{C}=\left\{C_1,\ldots,C_Q\right\}$.

Schepers et al. (2017) obtained estimators of the interaction effect parameters γ_{pq} by maximizing the *classification likelihood* of model (1), that is, maximizing

$$L(\mathbf{D}|\mathcal{R},\mathcal{C},\boldsymbol{\theta},\sigma^2)=\prod_{p=1}^{P}\prod_{q=1}^{Q}\prod_{i\in R_p}\prod_{j\in C_q}\frac{1}{\sqrt{2\pi\sigma^2}}$$
$$\exp\left(-\frac{1}{2}\frac{(d_{ij}-\mu-\alpha_i-\beta_j-\gamma_{pq})^2}{\sigma^2}\right). \tag{3}$$

The MAXINT model assumes all row (α_i) and column (β_j) effects to be fixed and estimation proceeds by maximizing classification likelihood (3) across all possible two-mode partitions $\mathcal{R}\times\mathcal{C}=(R_p\times C_q;\quad p=1,\ldots,P,\quad q=1,\ldots,Q)$. However, this fixed effects assumption is often unrealistic. For instance, in contextualized personality research, rows of the observed data matrix $\mathbf{D}=\left\{d_{ij}\right\}$ may refer to persons, which are typically a random sample from a population of interest, whereas columns may refer to situations or conditions, for which a fixed effects assumption is appropriate in most studies. Therefore, we propose a generalization of model (1), coined REMAXINT, in which the main effects of one of the two modes, say, α_i for the rows, are allowed to be either fixed or random. To prevent any misunderstanding, the rows are assumed to be randomly drawn from a population that consists of P fixed (but unknown) biclusters. The interaction effect parameters γ_{pq} that pertain to these clusters are therefore fixed too. For estimation, we will be using a conditional likelihood approach in which the random effects are treated as nuisance parameters. These are parameters that are not of scientific interest but, nonetheless, have to be accounted for in the estimation of the parameters of interest. Conditional likelihood is well-known in modern psychometrics (Andersen 1973; Fischer and Molenaar 1995) and biostatistics (Anderson and Senthilselvan 1980). The main advantage of this approach is that no distributional assumption is needed with respect to the nuisance parameters (Verbeke et al. 2001). In REMAXINT, it implies that we make no assumptions on the random effects (α_i), except that these are *iid* with finite variance and zero expectation (the latter for identifiability of μ in the model). However, the constraints 2b–2d are kept on the fixed effects parameters. Furthermore, REMAXINT can be used for the purpose of capturing the gist of row by column interaction in a data set at hand, but by itself is not sufficient for inferring whether there is statistical evidence of such interaction.

Therefore we develop an omnibus statistical test, based on REMAXINT, for the null hypothesis of no interaction between row and column clusters.

2.2 Model estimation

In this subsection, we derive *conditional classification maximum likelihood* estimators of the interaction effect parameters of the REMAXINT model. The conditional likelihood approach implies conditioning on a sufficient statistic for nuisance parameters. As a result, the conditional likelihood does not depend on those nuisance parameters. To obtain this conditional likelihood, we factorized likelihood (3) by focusing on $\overline{d}_{i.} = \frac{1}{J}\sum_{j=1}^{J} d_{ij}$ as sufficient statistic for nuisance parameter α_i. As a starting point, $\overline{d}_{i.}$ was added and subtracted into likelihood (3):

$$
\begin{aligned}
&L(\mathbf{D}|\mathcal{R}, \mathcal{C}, \boldsymbol{\theta}, \sigma^2)\\
&\quad= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^{I\cdot J} \prod_{p=1}^{P}\prod_{q=1}^{Q}\prod_{i\in R_p}\prod_{j\in C_q}\\
&\qquad \exp\Big(-\frac{1}{2}\frac{((d_{ij}-\overline{d}_{i.}-\beta_j-\gamma_{pq})+(\overline{d}_{i.}-\mu-\alpha_i))^2}{\sigma^2}\Big).
\end{aligned}
$$

Expanding the squares the following expression is obtained:

$$
\begin{aligned}
L(\mathbf{D}|\mathcal{R}, \mathcal{C}, \boldsymbol{\theta}, \sigma^2) =&\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^{I\cdot J}\\
&\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\beta_j-\gamma_{pq})^2\right.\right.\\
&\left.\left.+\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(\overline{d}_{i.}-\mu-\alpha_i)^2+2\cdot U\right)\right),
\end{aligned}
$$

where U is the sum of the cross product terms, that is, $U = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\beta_j-\gamma_{pq})A_i$, with $A_i = \overline{d}_{i.}-\mu-\alpha_i$. After rewriting U as $\sum_{p=1}^{P}\sum_{i\in R_p}A_i\sum_{q=1}^{Q}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\beta_j-\gamma_{pq})$, it is straightforward to show that $U = 0$, since $\sum_{q=1}^{Q}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}) = J(\overline{d}_{i.}-\overline{d}_{i.}) = 0$, $\sum_{q=1}^{Q}\sum_{j\in C_q}(\beta_j) = 0$ [see Eq. (2b)] and $\sum_{q=1}^{Q}\sum_{j\in C_q}(\gamma_{pq}) = 0$ [see Eq. (2d)].

Therefore likelihood (3) can be written as

$$
L(\mathbf{D}|\mathcal{R}, \mathcal{C}, \boldsymbol{\theta}, \sigma^2) = \exp\left(-\frac{1}{2\sigma^2}\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\beta_j-\gamma_{pq})^2\right)
$$

$$(\frac{1}{\sqrt{2\pi\sigma^2}})^{I\cdot J} \exp\left(-\frac{J}{2\sigma^2}\sum_{p=1}^{P}\sum_{i\in R_p}(\overline{d}_{i.}-\mu-\alpha_i)^2\right), \quad (4)$$

which satisfies the Fisher–Neyman factorization theorem (Fisher 1922; Neyman 1935). Specifically, the second factor is the likelihood of $\overline{d}_{i.}$, which is a function of α_i and depends on d_{ij} only through the sufficient statistic $\overline{d}_{i.}$. As a clustering criterion, we will maximize the first factor, which does not include α_i,

$$CL = \exp\left(-\frac{1}{2\sigma^2}\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\beta_j-\gamma_{pq})^2\right) \quad (5)$$

for the unknown parameters β_j, γ_{pq} and $\mathcal{R}$ and $\mathcal{C}$. This implies a classification likelihood approach to the clustering problem (see e.g., Scott and Symons 1971; Symons 1981; Bock 1996). For the problem at hand, this then is a maximal conditional classification likelihood problem. For a given two-mode partition $\mathcal{R}\times\mathcal{C}$, maximum likelihood (m.l.) estimation of the unknown parameters β_j, and γ_{pq} amounts to minimizing the quadratic criterion:

$$S := \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\beta_j-\gamma_{pq})^2, \quad (6)$$

subject to identification constraints (2b–2d).

At this point, it is convenient to define the following statistics:

- $\widetilde{\beta}_j = \overline{d}_{.j} - \overline{d}_{..}$,
- $\widetilde{\gamma}_{pq} = \frac{1}{|R_p||C_q|}\cdot\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\overline{d}_{.j}+\overline{d}_{..})$

for $j = 1, ..., J$, $p = 1, ..., P$ and $q = 1, ..., Q$, where $\overline{d}_{.j} = \frac{1}{I}\sum_{i=1}^{I} d_{ij}$ and $\overline{d}_{..} = \frac{1}{I\cdot J}\sum_{i=1}^{I}\sum_{j=1}^{J} d_{ij}$.

Proposition 1 *For the REMAXINT model the m.l. estimators of the unknown main and interaction effects are given by*

$$\widehat{\beta}_j = \widetilde{\beta}_j \quad \textit{and} \quad \widehat{\gamma}_{pq} = \widetilde{\gamma}_{pq}$$

for $j = 1, ..., J$, $p = 1, ..., P$ *and* $q = 1, ..., Q$.

Proof After adding and subtracting $\widetilde{\beta}_j$ and $\widetilde{\gamma}_{pq}$, the squared-error residual sum S can be decomposed as follows:

$$S = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}\Big((d_{ij}-\overline{d}_{i.}-\widetilde{\beta}_j-\widetilde{\gamma}_{pq})+(\widetilde{\beta}_j-\beta_j)+(\widetilde{\gamma}_{pq}-\gamma_{pq})\Big)^2$$

$$= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}\left(d_{ij} - \overline{d}_{i.} - \widetilde{\beta}_j - \widetilde{\gamma}_{pq}\right)^2$$

$$+ \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}\left((\widetilde{\beta}_j - \beta_j)^2 + (\widetilde{\gamma}_{pq} - \gamma_{pq})^2\right) + 2\cdot V$$

where V is a sum of cross-product terms that equals 0 (see “Appendix 1”). Since the first sum does not depend on β_j and γ_{pq} and the second sum is always non-negative, S is minimized if and only if $\beta_j = \widetilde{\beta}_j$ and $\gamma_{pq} = \widetilde{\gamma}_{pq}$, as asserted. □

Substituting the maximum likelihood estimators $\widehat{\beta}_j = \overline{d}_{.j} - \overline{d}_{..}$ and $\widehat{\gamma}_{pq}$ in CL (5) we obtain

$$\widetilde{CL}(\mathcal{R},\mathcal{C}) = \exp\left(-\frac{1}{2\sigma^2}\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij} - \overline{d}_{i.} - \overline{d}_{.j} + \overline{d}_{..} - \widehat{\gamma}_{pq})^2\right). \quad (7)$$

The maximum likelihood solution is then found by maximizing (7) over all possible two-mode partitions $\mathcal{R}\times\mathcal{C}$, or equivalently, *minimizing* the following clustering criterion:

$$\widetilde{S}(\mathcal{R},\mathcal{C}) = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d^*_{ij} - \widehat{\gamma}_{pq})^2, \quad (8)$$

where $d^*_{ij} = d_{ij} - \overline{d}_{i.} - \overline{d}_{.j} - \overline{d}_{..}$. These results based on considering a conditional likelihood approach are important in two ways. Firstly, they show how to obtain maximum likelihood estimates of the interaction effect parameters for either of two different model assumptions: (i) row and column effects are fixed or (ii) row effects are random and column effects are fixed. As the method proposed in Schepers et al. (2017) was suitable only for model assumption (i), this result extends the scope of data sets that are eligible for maximal interaction two-mode clustering. Note that applications with fixed row effects and random column effects can also be addressed by making model assumption (ii) and analyzing the transpose of **D**. Secondly, these results show that, computationally, the associated optimization problem is equivalent to the one addressed in Schepers et al. (2017). As a consequence, maximum likelihood estimates of γ_{pq} in (5) can be obtained by applying the numerical optimization algorithm described in Schepers et al. (2017) and Schepers and Hofmans (2009), the latter of which also provides open-source software. We describe this algorithm (Algorithm 1) in Sect. 3.2.

3 Statistical test

In this Section, we propose a statistical test to test the null hypothesis of all interaction effect parameters in the REMAXINT model to be equal to zero. Specifically, we test the null hypothesis,

$$
\begin{aligned}
&H_0 : \forall p, q \quad \gamma_{pq} = 0 && (9)\\
&\text{versus}\\
&H_1 : \exists\, p, q \quad \text{such that} \quad \gamma_{pq} \neq 0, && (10)
\end{aligned}
$$

for a fixed number of row and column clusters, P and Q.

We first introduce a new test statistic based on a decomposition of the variance, we show that maximizing CL (5) is equivalent to maximizing this statistic and we describe its properties. We then propose a procedure to obtain the distribution of the test statistic under the null hypothesis and to perform the statistical test.

3.1 Test statistic

From this subsection on, we refer to $iTSS = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(d_{ij}^{*}\right)^2$ as the total observed interaction sum of squares. Furthermore, for any arbitrary two-mode partition $\mathcal{R} \times \mathcal{C} = (R_p \times C_q; \quad p = 1, \ldots, P, \quad q = 1, \ldots, Q)$, we refer to:

$$iESS = \sum_{p=1}^{P} \sum_{q=1}^{Q} |R_p||C_q| \left(\widehat{\gamma}_{pq}\right)^2 \tag{11}$$

$$iRSS = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i \in R_p} \sum_{j \in C_q} \left(d_{ij}^{*} - \widehat{\gamma}_{pq}\right)^2, \tag{12}$$

where (11) defines the explained interaction sum of squares and (12) the residual interaction sum of squares. Note that for any arbitrary two-mode partition $\mathcal{R} \times \mathcal{C}$ the following decomposition of the total observed interaction sum of squares $iTSS$ holds:

$$iTSS = iESS + iRSS.$$

This follows by noting that

$$
\begin{aligned}
iTSS = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(d_{ij}^{*}\right)^2 &= \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i \in R_p} \sum_{j \in C_q} \left(d_{ij}^{*} + \widehat{\gamma}_{pq} - \widehat{\gamma}_{pq}\right)^2 \\
&= \sum_{p=1}^{P} \sum_{q=1}^{Q} |R_p||C_q| \left(\widehat{\gamma}_{pq}\right)^2 \\
&\quad + \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i \in R_p} \sum_{j \in C_q} \left(d_{ij}^{*} - \widehat{\gamma}_{pq}\right)^2, \qquad (13)
\end{aligned}
$$

where the equality between the last two lines is due to the cross product term being equal to zero.

Let F denote the ratio between iESS and iRSS, that is,

$$F(\mathcal{R}, \mathcal{C}) = \frac{\sum_{p=1}^{P} \sum_{q=1}^{Q} |R_p||C_q| \left(\widehat{\gamma}_{pq}\right)^2}{\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i \in R_p} \sum_{j \in C_q} \left(d_{ij}^{*} - \widehat{\gamma}_{pq}\right)^2} = \frac{iESS}{iRSS}. \tag{14}$$

Proposition 2 *Maximizing the conditional classification likelihood in Eq.* (5) *is equivalent to maximizing the value of* F *across all possible two-mode partitions* $\mathcal{R} \times \mathcal{C}$ *for fixed* P *and* Q*.*

Proof $iTSS$ does not depend on $\mathcal{R} \times \mathcal{C}$. Therefore, it follows that maximizing $iESS$ implies minimizing $iRSS$, which is equivalent to minimizing $\widetilde{S}$ (8). From decomposition (13) and clustering criterion (8) it follows that maximizing CL (5) is equivalent to maximizing the value of F across all possible two-mode partitions $\mathcal{R} \times \mathcal{C}$ for fixed P and Q, as asserted. □

Let F^{*} denote the maximized value of F, that is $F^{*} = \max_{\mathcal{R}\times\mathcal{C}} F(\mathcal{R}, \mathcal{C})$. Larger observed values of F^{*} indicate stronger interaction between the row and column partitions in the sample. We propose to use F^{*} as a test statistic to test the null hypothesis of all interaction effect parameters in the REMAXINT model to be equal to zero. This statistic is computationally similar to the standard F-test statistic used in two-way ANOVA. However, the REMAXINT model estimation implies searching through the entire model space of possible two-mode partitions. As such, F^{*} can be considered a new $max - F$ test statistic designed for the study of two-way interaction. For finite sample sizes, the sampling distribution of $max - F$ and related statistics defined on clustering approaches is not known and must be computed by simulations (Bock 1996). We describe how to obtain the sampling distribution via Monte Carlo simulations in Sect. 3.2, but first we focus on two properties of F^{*} that are appealing for the specification of this computational task.

Property 1 *The distribution of* F^{*} *does not depend on the value of the unknown variance* σ^2*.*

Proof This can be shown by considering the following transformation of the data: $d_{ij}^{\prime} = c \times d_{ij}$, that is, multiplication by a constant factor c. This transformation implies the residual variance of the transformed data to be $\sigma^{2^{\prime}} = c^2 \times \sigma^2$. This transformation further implies $d_{ij}^{*^{\prime}} = c \times d_{ij}^{*}$ and $\widehat{\gamma}_{pq}^{\prime} = c \times \widehat{\gamma}_{pq}$. Thus, since all terms within the squares of F^{*} are multiplied by the same factor c, both the numerator and denominator are multiplied by c^2 and, therefore, they cancel out of Eq. (14). This implies that F^{*} does not depend on the value of the unknown residual variance, as asserted. □

Property 2 *The distribution of* F^{*} *does not depend on the size of the row and column main effects nor on the overall mean* (μ)*.*

Proof Starting from $d^*_{ij} = d_{ij} - \overline{d}_{i.} - \overline{d}_{.j} + \overline{d}_{..}$, rewriting $\overline{d}_{i.}$ and $\overline{d}_{.j}$ in terms of d_{ij} and replacing d_{ij} with Eq. (1) leads to

$$\begin{aligned} d^*_{ij} &= d_{ij} - \overline{d}_{i.} - \overline{d}_{.j} + \overline{d}_{..} \\ &= d_{ij} - \frac{1}{J}\sum_j d_{ij} - \frac{1}{I}\sum_i d_{ij} + \frac{1}{I \times J}\sum_i\sum_j d_{ij} \\ &= \gamma_{pq} + \epsilon_{ij} - \frac{1}{J}\sum_j \epsilon_{ij} - \frac{1}{I}\sum_i \epsilon_{ij} + \frac{1}{I \times J}\sum_i\sum_j \epsilon_{ij} \\ &= \gamma_{pq} + \epsilon_{ij} - \overline{\epsilon}_{i.} - \overline{\epsilon}_{.j} + \overline{\epsilon}_{..}, \end{aligned}$$

where ϵ_{ij} indicates the residual term for row i and column j as defined in Model (1). Details on the steps between the second and third line are provided in "Appendix 2". This final expression does not depend on α_i, β_j nor on μ, as asserted. □

Hence, F^* is pivotal with respect to μ, α_i, β_j and σ^2.

3.2 Computational procedure

It is important to note that in the current approach, the clustering of the data is considered primarily as a tool to make inference on the interaction effects.

To test the null hypothesis in Eq. 9, it is possible to draw from the true null distribution of F^* rather than using a bootstrap approach (see e.g., McLachlan and Peel 1997; Hennig and Lin 2015). Note that under the null hypothesis, there is no partitioning in the data generating mechanism (i.e., $\gamma_{pq} = 0 \ \forall p, q$). Specifically, we propose the following three steps Monte Carlo (MC) approach to obtain the sampling distribution of F^*:

- Step 1: Generate a simulated data matrix from the null model. Specifically, generate $\mathbf{D}^{(sim)}$ of size $I \times J$ such that each cell (ij) contains a single observation $d^{(sim)}_{ij} \sim \mathcal{N}(\mu^{(sim)} + \alpha^{(sim)}_i + \beta^{(sim)}_j, \ \sigma^{2(sim)})$. We discuss and specify $\mu^{(sim)}$ $\alpha^{(sim)}_i$, $\beta^{(sim)}_j$ and $\sigma^{2(sim)}$ in Sect. 4.1.1.
- Step 2: Fit the model with P row clusters and Q column clusters to $\mathbf{D}^{sim}$ to estimate the parameters of interest and compute the value of the corresponding test statistic. Specifically, apply REMAXINT to estimate γ_{pq} and compute F^*.
- Step 3: Repeat steps (1) and (2) K times, to compose the distribution of the test statistic under the null hypothesis. Each value computed in step (2) is a value from this distribution. Specifically, we obtain the following set of values $\boldsymbol{\xi} = \{F^*_{(b)}\}$, with $(b = 1, \ldots, K)$.

If K is sufficiently large, then the empirical distribution of $F^*_{(b)}$ approaches the sampling distribution of F^* under the null hypothesis. Step 2 implies that any of the $F^*_{(b)}$ is yielded by a search through the complete model space of two-mode partitions and, as such, is in line with the guidelines of Lipkovich et al. (2017). Specifically, in order to fit REMAXINT we implemented the following two steps. First, use Algorithm 1 to find a local maximum. Second, in order to increase the probability of finding the global maximum, this process is repeated $M = 20$ times with different independently generated random starting solutions. The highest maximum out of all $M = 20$ runs is kept as the final solution, yielding REMAXINT estimates of γ_{pq}.

Algorithm 1 Algorithm to find REMAXINT local maximum of (7) with respect to $\mathcal{R}$ and $\mathcal{C}$.

```
Set a randomly generated bicluster configuration $\mathcal{R} \times \mathcal{C}$ and compute $\widehat{\gamma}_{pq}$ for that solution
repeat
    Keep $\mathcal{C}$ and $\widehat{\gamma}_{pq}$ fixed and update row partition $\mathcal{R}$ such that for each row assignment clustering criterion (8) is minimized;
    Keep $\mathcal{R}$ and $\mathcal{C}$ fixed and update $\widehat{\gamma}_{pq}$;
    Keep $\mathcal{R}$ and $\widehat{\gamma}_{pq}$ fixed and update column partition $\mathcal{C}$ such that for each column assignment clustering criterion (8) is minimized;
    Keep $\mathcal{R}$ and $\mathcal{C}$ fixed and update $\widehat{\gamma}_{pq}$;
    Compute $\zeta$ as the difference between the value of clustering criterion (8) in the previous iteration and the current iteration;
until $\zeta > 0$
```

4 Simulation study

In this section, we report an evaluation of the proposed methodology in terms of two statistical criteria: Type-I error rate and power. In the following subsections, we first present the design of the studies and then discuss the results.

4.1 Design

In this subsection we discuss the design of three simulation studies. The first study is used to establish critical values of test statistic F^* as a function of two completely crossed experimental factors: *size* $(I \times J)$ of the data set and *complexity of analysis* (P, Q). The critical values are found following the three steps in the preceding section, with $K = 10{,}000$. The second simulation study investigates to what extent these critical values are subject to sampling errors being based on finite K in the first simulation. Furthermore, this study is used to verify that these critical values do not depend on chosen values of the parameters that are not of interest (μ, α, β and σ). The third simulation study is to assess the power of test statistic F^* to detect row by column interaction. In each of these three simulation studies, the two fully crossed design factors, *size* and *complexity of analysis*, were varied across a range of values:

(i) *size* $(I \times J)$ of the data, at 6 levels: 20×20, 30×30, 40×40, 50×10, 50×30, 100×20;
(ii) *complexity of analysis* (P, Q), at 4 levels: (2, 2), (3, 3), (4, 4), (5, 5).

4.1.1 Determination of critical values

In order to determine critical values for the proposed test statistic F^* we applied the three-step Monte Carlo approach as described in Sect. 3. Specifically, we generated $K = 10{,}000$ independent data sets without any row by column interaction for each level of the design factor *size*. That is, data were generated from the null model such that $d_{ij}^{(sim)} \sim \mathcal{N}(\mu^{(sim)} + \alpha_i^{(sim)} + \beta_j^{(sim)}, \sigma^{2(sim)})$. Without loss of generality, we set $\sigma^{2(sim)} = 1$ and, likewise, $\mu^{(sim)} + \alpha_i^{(sim)} + \beta_j^{(sim)} = 0$, since F^* is pivotal with respect to these parameters.

On each generated data set, we applied REMAXINT for each level of *complexity of analysis*. For any combination of *size* and *complexity of analysis* this yields $K = 10{,}000$ simulated Monte Carlo (MC) test statistic values $\boldsymbol{\xi} = \{F^*_{(b)}\}$, $(b = 1, \ldots, K)$, of which the distribution approaches the sampling distribution of F^* under the null hypothesis if K is sufficiently large (Efron 1982; Chernick 2011). From the simulated distribution of F^* we then obtain critical values F^*_α for any significance level α by finding the $(100 - \alpha)$th quantile of the empirical distribution.

4.1.2 Type-I error rate

We used a second simulation study to investigate (a) the extent to which sampling error affects establishing critical values if $K = 10{,}000$ and (b) to verify that the inferential procedure is not sensitive to choices of values of $\mu^{(sim)}$, $\alpha^{(sim)}$, $\beta^{(sim)}$, (which were set to 0 in study 1) and $\sigma^{2(sim)}$ (which was set to 1 in study 1). Specifically, for each level of *size*, we generated independent data sets from the null model such that $d_{ij} \sim \mathcal{N}(\mu + \alpha_i + \beta_j, \sigma^2)$, where $\mu \sim \mathcal{U}(0, 1)$, $\alpha_i \sim \mathcal{N}(0, 1)$ and $\beta_j \sim \mathcal{N}(0, 1)$. In order to reduce the computational burden we now generated 5000 data sets as opposed to the 10,000 we used to determine critical values. Furthermore, compared to simulation study 1, we chose $\sigma^2 = 2$ to show empirically that F^* is pivotal with respect to this variance.

Each data set was then analyzed by applying REMAXINT for each level of *complexity of analysis*. Every single analysis yields an observed value of the test statistic $F^*_{(obs)}$, which may be compared to the critical value for that combination of *size* and *complexity of analysis* as was obtained in the first simulation study in Sect. 4.1.1. Specifically, if $F^*_{(obs)} > F^*_\alpha$ the decision is to reject the null hypothesis (i.e., no row by column interaction) in favour of the alternative hypothesis that there is some interaction. For each combination of *size* and *complexity of analysis*, the proportion of $F^*_{(obs)}$ values (out of all 5000 data sets for that combination of *size* and *complexity of analysis*) that fall in the rejection region corresponds to the empirical Type I error rate or empirical significance level of the newly proposed test of interaction. If study 2 yields empirical Type-I error rates close to the nominal level α, we may safely conclude that (a) choosing $K = 10{,}000$ is sufficient for accurately establishing critical values, and (b) that the proposed method does not depend on the choice of $\mu^{(sim)}$, $\alpha^{(sim)}$, $\beta^{(sim)}$ and $\sigma^{2(sim)}$ for generating MC samples.

4.1.3 Power of the test

This study was designed to evaluate the power of the REMAXINT test of interaction. To this end, data sets were generated such that cell entries $d_{ij} \sim \mathcal{N}(\mu_{pq}, \sigma_\epsilon^2)$, where the bicluster means μ_{pq} were drawn from a standard uniform distribution, that is, $\mu_{pq} \sim \mathcal{U}(0, 1)$. Furthermore, each row (resp. column) was randomly and independently assigned to one of the row (resp. column) clusters with probability $1/P$ (resp. $1/Q$). If this yielded a row (resp. column) partition with empty clusters, this procedure was repeated until a partition without empty row (resp. column) clusters was obtained. Having the row and column cluster memberships assigned allows to reparametrize the bicluster means as μ_{pq}=$\mu + \alpha_i + \beta_j + \gamma_{pq}$ (imposing constraints 2a-2d) and choose σ_ϵ^2 such that

$$\omega = \frac{\sum_{p=1}^{P} \sum_{q=1}^{Q} |R_p||C_q| \left(\gamma_{pq}\right)^2}{\sum_{p=1}^{P} \sum_{q=1}^{Q} |R_p||C_q| \left(\gamma_{pq}\right)^2 + I \cdot J \cdot \sigma_\epsilon^2} = \frac{1}{16}. \tag{15}$$

That is, we made sure that the ratio of interaction variance to the sum of interaction variance and residual variance was small.

In this third simulation study, the *true* number of row and column clusters $(P^{(t)}, Q^{(t)})$ is an additional design factor varied at four levels: (2, 2), (3, 3), (4, 4), (5, 5), which was fully crossed with *size* and *complexity of analysis*.

For each combination of *size* and *true* number of row and column clusters, we independently generated 5000 simulated data sets. We analyzed each of these data sets at all 5 levels of *complexity of analysis*, by applying the REMAXINT model. This yields for each data set and each level of *complexity of analysis* an observed value of test statistic $F^*_{(obs)}$, which, as in simulation study 2, is compared to the corresponding critical value F^*_α for that combination of *size* and *complexity of analysis*. The proportion of observed values (out of all 5000 data sets for that combination of *size*, *true* number of row and column clusters and *complexity of analysis*) that fall in the rejection region corresponds to the empirical power of the REMAXINT test of interaction.

4.2 Results

4.2.1 Simulation results for critical values

At nominal significance level $\alpha = 0.05$, Table 1 shows critical values F^*_α, for each combination of *size* and *complexity of analysis*. These critical values were obtained by applying the MC approach, as described in Sect. 4.1.1.

Inspection of Table 1 shows that, for each level of *size*, the null distribution of F^* is shifted towards the right for higher levels of *complexity of analysis*. This is an expected result as higher levels of *complexity of analysis* imply models with more parameters to fit the observed data and thus higher F^*-values observed by chance. Furthermore, note that the levels of *size* constitute a partially ordered set with respect to the relation $\leq$ (i.e., $(I, J) \leq (I', J')$ if $I \leq I'$ and $J \leq J'$). Table 1 shows that, for each level of *complexity of analysis*, the null distribution of F^* is shifted towards the right for

Table 1 Critical values F_{α}^{*} corresponding to $\alpha = 0.05$ for each combination of *size* and *complexity of analysis*

Size	*Complexity of Analysis (P,Q)*			
	(2, 2)	(3, 3)	(4, 4)	(5, 5)
20 × 20	0.1440	0.2504	0.3535	0.4660
30 × 30	0.0890	0.1478	0.2000	0.2524
40 × 40	0.0635	0.1037	0.1375	0.1703
50 × 10	0.1569	0.2715	0.3859	0.5126
50 × 30	0.0670	0.1096	0.1459	0.1811
100 × 20	0.0648	0.1048	0.1400	0.1751

the smallest level of any pair of levels (I, J) and (I', J') that are comparable (i.e., $(I, J) \leq (I', J')$ or $(I', J') \leq (I, J)$). This again is an expected result since, for a given level of *complexity of analysis*, larger data sets imply relatively fewer parameters as compared to the total number of observations and thus less chance capitalization. Finally, the *size* effect on the critical values of F^* cannot only be explained by the total number of observations implied by the levels of *size*, as becomes clear by comparing *size* levels 40 × 40 and 100 × 20: critical values in the first condition, with fewer number of total observations, are systematically (i.e., across all levels of *complexity of analysis*) smaller than those in the second. This suggests that the amount of symmetry in terms of data size plays a role as well. Specifically, more symmetrical data sizes imply smaller critical values. This result makes sense too, because for a fixed total number of observations the probability of obtaining small clusters (e.g., singleton) is larger for less symmetrical data sizes. Smaller clusters are more likely to yield more extreme F^* values.

4.2.2 Simulation results for Type-I error rate

Table 2 shows the proportion of significant test results for all combinations of *size* and *complexity of analysis*. This table shows that the proportions of significant test results are close to the nominal significance level α, which is 0.05 in this case. This suggests that generating $K = 10{,}000$ samples in the MC procedure is a reasonable choice for approximating the null distribution of F^*. As expected, observed deviations from the nominal level do not appear to be systematically related to *size* nor *complexity of analysis*. Furthermore, the empirical Type-I error rates are close to the nominal level, confirming that it does not matter what choices are made for μ, α, β and σ^2 in the process of generating MC data.

4.2.3 Simulation results for power of the test

Figure 1 shows empirical power as a function of *size*, *true* numbers of row and column clusters $(P^{(t)}, Q^{(t)})$ and *complexity of analysis* (P, Q). These results show that empirical power increases for increasing levels of *size*, as expected. Furthermore, the results suggest that the most parsimonious level of *complexity of analysis* $(i.e, (P, Q) = (2, 2))$ tends to outperform higher levels of *complexity of analysis* for

Table 2 Type-I error rate corresponding to $\alpha = 0.05$ for different data *sizes* and *complexity of analysis*

Size	*Complexity of Analysis (P,Q)*			
	(2, 2)	(3, 3)	(4, 4)	(5, 5)
20 × 20	0.0572	0.0534	0.0566	0.0520
30 × 30	0.0502	0.0512	0.0502	0.0500
40 × 40	0.0498	0.0532	0.0520	0.0510
50 × 10	0.0508	0.0488	0.0504	0.0506
50 × 30	0.0518	0.0526	0.0542	0.0520
100 × 20	0.0518	0.0540	0.0518	0.0540

all data *sizes*. However, this effect decreases for increasing levels of true complexity $(P^{(t)}, Q^{(t)})$. The fact that even for higher levels of *true* number of row and column clusters empirical power does not appear to benefit (much) from increasing the level of *complexity of analysis* allows the formulation of straightforward advice for practitioners intending to use this method: If one does not have strong prior hypotheses about the *true* number of row and column clusters, then apply this test of interaction for *complexity of analysis* equal to (2, 2). Furthermore, comparing *size* levels 40 × 40 and 100 × 20, the results are quite similar. It thus appears that empirical power is not only a function of the total number of observations but rather that the degree of symmetry of the data size also figures in.

4.3 Effect of data size and of symmetry

In order to study in more detail the effect of data size and of symmetry, we ran an additional simulation study in which we varied the size and degree of symmetry of the data size. It is known that to preserve the same level of statistical power, traditional F-tests require larger sample sizes by a factor k^2 if the size of the tested effect (expressed as a ratio of true to total variance) changes by a factor $\frac{1}{k}$ (Cohen 1992). In order to check whether this property carries over to the REMAXINT test of interaction, we decreased the interaction effect size ω in (15). Specifically, for all data sets generated in this additional study, ω was decreased by a factor 2 as compared to simulation study 3 (i.e., $\omega = 1/32$ vs $\omega = 1/16$). Furthermore, we chose data size 100 × 20 as a reference and specified three additional levels of *size*. Compared to data size 100 × 20, these three additional size levels imply an increase in terms of the total number of observations by a factor of 4. However, these data sizes are of varying degrees of data size symmetry:

- data size 100 × 80 is more symmetrical than data size 100 × 20
- data size 200 × 40 has the same degree of symmetry as data size 100 × 20
- data size 400 × 20 is less symmetrical than data size 100 × 20.

Figure 2 shows that the beneficial effect of increasing the total number of observations on empirical power is stronger the more the resulting data size is symmetrical [compare panels (b) and (c) to panel (d)]. It appears that if a larger total number of observations is achieved by introducing more asymmetry in the resulting data size, then the associated power increase is limited [compare panel (d) to panel (a)]. Finally, we also note that the empirical power values shown in panel (c) of Fig. 2 are very similar to the results shown

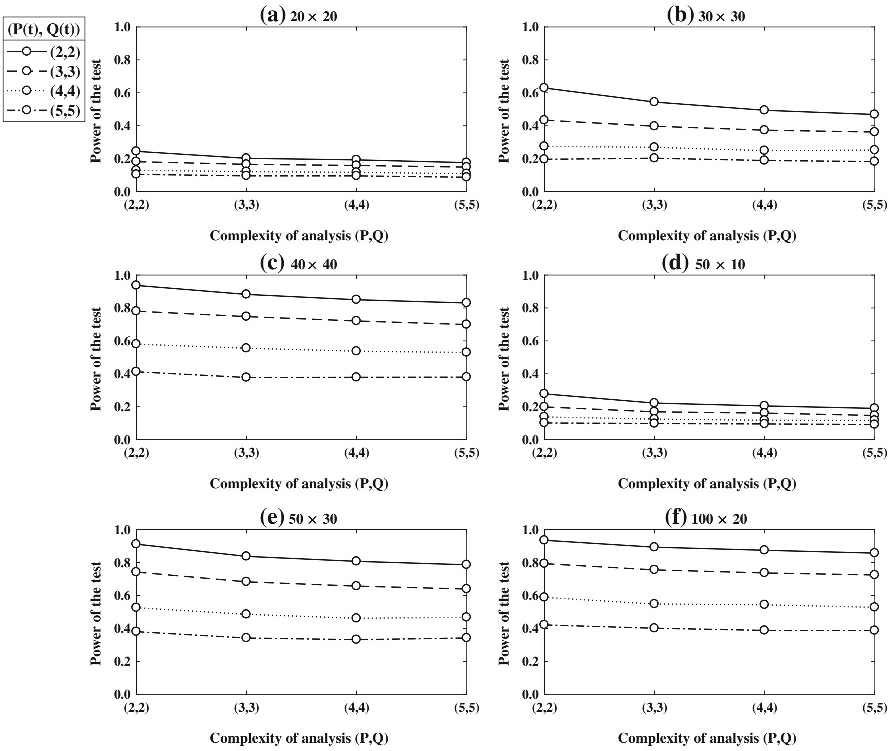

Fig. 1 Empirical power as a function of *size* (panel), *true* number of row and column clusters (curves) and *complexity of analysis* (horizontal axis)

in panel (f) of Fig. 1, where the total number of observations is 4 times as small, but the effect size is 2 times as large. This suggests that, in terms of how required sample size relates to effect size, this REMAXINT based test of interaction indeed behaves in the same way as traditional two-way ANOVA F-tests do (if degree of symmetry in data size is kept constant).

5 Application to case studies

The previous section evaluated the performance of the REMAXINT interaction test using simulated data sets. In this section, we apply the method on two case studies. The first application stems from a study of *person by situation interaction*, one of the key questions addressed by researchers in contextualized personality psychology (Geiser et al. 2015; Mischel and Shoda 1995, 1998). The second application pertains to a study of *genotype by location interaction*, which is of substantial interest in agriculture (Corsten and Denis 1990; Piepho 1997, 1999; Gauch 2006; Forkman and Piepho 2014). The first application is an example of a study in which one of the modes (i.e., persons) is a random sample from a population of interest. In the second application,

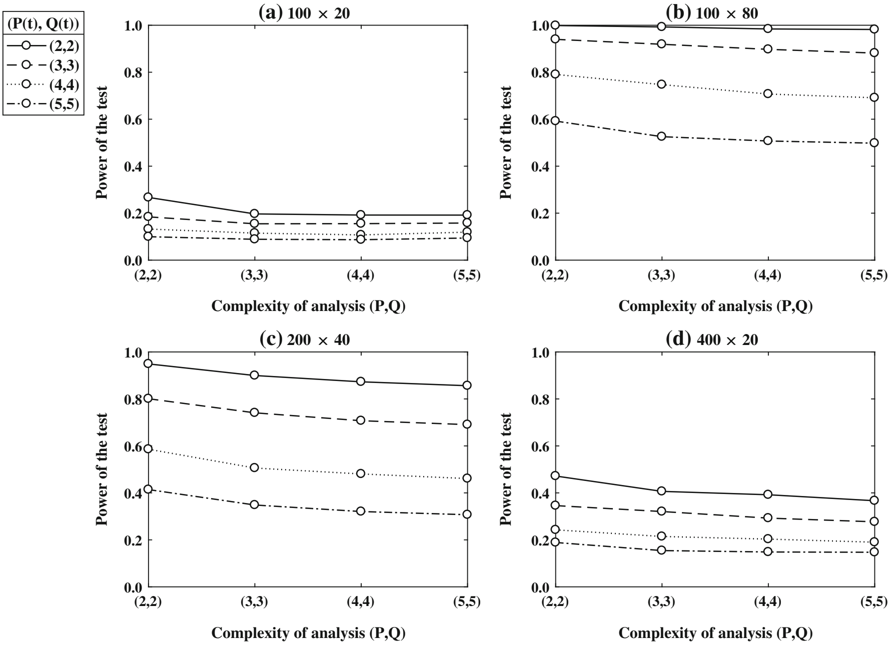

Fig. 2 Additional simulation study: empirical power as a function of *size* and symmetry (panels), *true* number of row and column clusters (curves) and *complexity of analysis* (horizontal axis)

on the other hand, it is more typical to assume that both modes (i.e., genotypes and locations) represent fixed effects. As discussed in Sect. 2.2, the REMAXINT model is appropriate for both of these model assumptions.

5.1 Person by situation case study

In this case study, we applied our method to infer whether there is evidence of *person by situation interaction* in a study on altruism. The altruism data were collected in a study by Quintiens (1999) and were more recently reanalyzed in Schepers and Van Mechelen (2011) and Schepers et al. (2017). A group of $I = 102$ participants was presented with a set of $J = 16$ hypothetical situations, each describing an emergency in which a victim could possibly be helped by a person. Each participant was asked to indicate for each emergency to what degree they would be willing to help the victim. Ratings were given on a 7-point scale from 1 (*definitely not*) to 7 (*definitely yes*).

The REMAXINT interaction test was applied to the resulting 102 × 16 data matrix of help ratings in order to infer whether there is evidence of interaction between person and situation clusters. Specifically, complexity of analysis (P, Q) was varied at five levels: (2, 2), (3, 3), (4, 4), (5, 5) and (6, 6), each of which yields an observed value $F^*_{(obs)}$ of test statistic F^*. In order to obtain critical values we employed the procedure described in Sect. 4.1.1 for a significance level $\alpha = 0.05$. For each test, p values were computed as $P(F^*_b > F^*_{(obs)})$. The results are shown in Table 3.

Table 3 Results of REMAXINT interaction test applied to help data for five levels of *complexity of analysis*

	Complexity of Analysis (P,Q)				
	(2, 2)	(3, 3)	(4, 4)	(5, 5)	(6, 6)
Critical value (F^*_α)	0.0774	0.1269	0.1719	0.2163	0.2625
Test statistic ($F^*_{(obs)}$)	0.1021	0.1577	0.2282	0.2857	0.3615
p value	0.0000	0.0000	0.0000	0.0000	0.0000

We note that $F^*_{(obs)}$ falls within the rejection region for each level of *complexity of analysis*. In fact, for each (P, Q), the $F^*_{(obs)}$ value is larger than any of the simulated values $F^*_{(b)}$ obtained under the null distribution, implying p values that are equal to zero. Altogether this is strong evidence in favour of the alternative hypothesis that person clusters differ with respect to the effect that these 16 situations have on their willingness to help. A complete analysis requires, in addition, a consideration of which model is most useful (i.e., selecting a level of (P, Q) in some optimal way) followed by an interpretation of the nature of the associated interaction. With respect to model selection in the framework of two-mode partitioning problems, various procedures for choosing appropriate numbers of clusters P and Q have been proposed and evaluated in the literature (see e.g., Ceulemans and Kiers 2006; Schepers et al. 2008; Wilderjans et al. 2013). Some of these procedures may potentially be useful for the model selection problem at hand, possibly upon a suitable adaptation. However, this problem is beyond the scope of this paper.

Figure 3 shows a plot of the explained interaction sum of squares [$iESS$, see Eq. (11)] in the 102 × 16 data matrix of help ratings, for each *complexity of analysis*. Model complexity was quantified as the number of freely estimated interaction effect parameters $(P - 1)(Q - 1)$ in the model. Based on this graph, and favouring simpler models for the sake of interpretation, we selected the model with $(P, Q) = (3, 3)$ (i.e., three person clusters and three situation clusters) for further elaboration. Figure 4 shows a plot of $\widehat{\gamma}_{pq}$ for each combination of person cluster p $(p = 1, 2, 3)$ and situation cluster q $(q = 1, 2, 3)$.

A substantive interpretation of the situation clusters was obtained by comparing them on external ratings from independent raters. Specifically, each situation was rated by a set of 56 raters in terms of the extent to which each of 22 characteristics applies. For each situation s and each characteristic f, these ratings were then averaged across the 56 raters, yielding values AvR_{sf} $(s = 1, \ldots, 16$ and $f = 1, \ldots, 22)$. Next, for each f, an ANOVA on AvR_{sf} was performed with situation cluster membership as factor. The strongest associations were found between situation cluster membership and the following two characteristics: *frequency of occurrence* $(R^2 = .35)$ and the extent to which a situation *requires skills to get resolved* $(R^2 = .34)$. The situation cluster labels in Fig. 4 include descriptions characterizing them with respect to these two situation characteristics. Specifically, compared to the situations in situation clusters 1 and 3, the situations in situation cluster 2 are, on average, considered to occur more frequently. Likewise, compared to the situations in situation clusters 1 and 2, the situations in

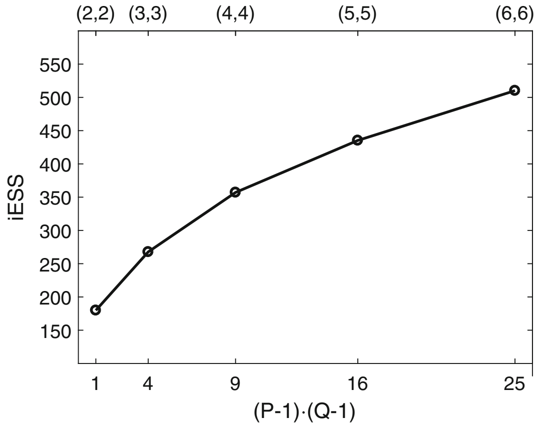

Fig. 3 Plot of explained interaction sum of squares ($iESS$) in altruism data, for each model

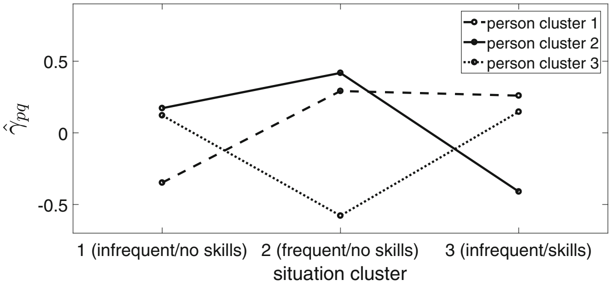

Fig. 4 Plot of $\widehat{\gamma}_{pq}$ for each combination of person cluster p ($p = 1, ..., P$) and situation cluster q ($q = 1, ..., Q$)

situation cluster 3 are, on average, considered to require a higher skill level in order to be resolved.

It appears from Fig. 4 that the largest amount of non-additivity between the person clusters is with respect to situation cluster 2. That is, it is mostly in these situations (i.e., situations that occur frequently and do not require high skill levels to get resolved) that individuals from different person clusters respond differentially. In contrast, the smallest degree of non-additivity is observed for situation cluster 1 (i.e., the estimated interaction effect parameters tend to be closer to zero in that situation cluster). Furthermore, this graph also helps characterizing the person clusters in terms of their sensitivity profiles to situations. For instance, those who are in person cluster 2 are, compared to the other person clusters, less inclined to help when they are in situations that require a higher skill level.

5.2 Genotype versus location case study

In genotype studies, plant breeders are typically interested in studying the behavior of genotypes at different locations. An issue of particular interest is whether there is a *genotype by location interaction*. For instance, it is important for plant breeders to

Table 4 Results of REMAXINT interaction test applied to genotype data for five levels of *complexity of analysis*

	Complexity of Analysis (P,Q)				
	(2, 2)	(3, 3)	(4, 4)	(5, 5)	(6, 6)
Critical value (F^*_α)	0.3355	0.6409	1.0117	1.5014	2.1843
Test statistic ($F^*_{(obs)}$)	0.3202	0.6879	1.0868	1.3758	1.9200
P-value	0.0804	0.0183	0.0165	0.1658	0.2652

know if the performance of genotypes in terms of crop yield differs across locations. The genotype data we used was published in Corsten and Denis (1990) and consists of average yield of corn in kg per acre for $I = 20$ genotypes in $J = 7$ locations.

The REMAXINT test was applied to the resulting 20×7 data matrix of corn yield to infer whether there is evidence of interaction between genotype and location clusters. Again, *complexity of analysis* was varied at five levels: (2, 2), (3, 3), (4, 4), (5, 5) and (6, 6), each of which yields an observed value $F^*_{(obs)}$ of test statistic F^*. Critical values were obtained by the procedure described in Sect. 4.1.1 for a significance level $\alpha = 0.05$. The results are shown in Table 4.

We observe that the value of test statistic F^* falls in the rejection region for only two levels of *complexity of analysis* (i.e., (3, 3) and (4, 4)). Furthermore, a (conservative) Bonferroni correction would result in none of the five test results being significant at a familywise Type I error rate of 5%. One may note that this correction could have been avoided by following the practical guideline to test only at *complexity of analysis* equal to (2, 2). This would not have led to a different conclusion, as the *p* value of that test is larger than 0.05. Thus, in this study there is no unambiguous empirical evidence to conclude that the performance of these 20 genotypes in terms of crop yield differs across the set of 7 locations under investigation.

6 Discussion

In this paper, we have presented REMAXINT, a method for studying interaction based on two-mode clustering which, unlike classical ANOVA, allows testing interaction in a two-mode data matrix if there is only one observation per cell. The methodological contribution of this manuscript is two-fold. Firstly, we extended the maximal interaction two-mode clustering model (MAXINT) that assumes fixed row and column main effects to include random row or column main effects. This extension is relevant, since many studies are better characterized by one of the two modes being random rather than fixed. Secondly, we introduced a new test statistic with appealing properties and developed a parametric re-sampling based procedure to perform hypothesis testing on the interaction effect parameters. In order to evaluate the performance of the presented method, we used simulations and real-life data sets. In the simulation studies, we studied the performance of the REMAXINT based test in terms of Type-I error rate and power. Results showed that empirical Type I error rate was always close to the nominal level, that is, in general deviated by not more than 0.005. Empirical power of the REMAXINT test of interaction substantially decreased as *complexity of analysis* increased. The simulation results also showed that statistical power of the proposed test is not only a function of the total number of observations. Specifically, we found that power was higher for (almost) square data sizes (i.e., $I \approx J$) as compared to more rectangular sizes (e.g., $I > J$). Since in our simulation studies only the case $P = Q$ was considered, and it may be conceivable that the results are different if $P > Q$ or $P < Q$, we ran some additional simulation studies (results not shown) with $P \neq Q$ to check whether the optimal scenario is the case where $I/J = P/Q$. The results indicate that the phenomenon is more complicated than that and requires further investigation, which is beyond the scope of this paper. However, symmetric data sizes appear to be optimal even if $P \neq Q$. A speculative explanation is that the case $I = J$ maximizes precision of the sufficient statistics for the many nuisance parameters (α's, β's) that are conditioned out/estimated, and thereby improves the precision for the γ's (of which there are far fewer). Note that in applications with a random sample of I rows and a fixed set of J columns, the dimensions of the data set will typically be such that $I > J$. Therefore, the case $P > Q$ may occur more frequently than $P < Q$. Furthermore, an interesting conclusion of the power study is that the lowest level of *complexity of analysis*, (2, 2), gave the highest power to detect interaction, even if the *true* number of row and column clusters was larger than (2, 2), which is important for practitioners to be aware of. The method developed can be applied to address relevant research questions in a variety of applications. In this paper, we presented two such applications on real-life data sets.

The proposed method only requires users to set the number of row and column clusters, and the number of MC samples needed to establish the critical F^*-value for their data matrix size and assumed number of clusters. There is no need to choose any other parameter or to look for some optimal setting of tuning parameters. In this paper, we showed that 10,000 MC samples are sufficient, while in the applications (Sect. 5) we selected the number of row and column clusters, P and Q, based on a 'post analysis' approach, as is commonly done. However, other approaches are possible. For instance, one could develop a method that treats P and Q as parameters to be estimated. This can be done, for example, in a Bayesian framework using mixture models with a prior on the number of mixture components, see Miller and Harrison (2018). Several other extensions of the REMAXINT method are also worth investigating. For example, a formulation that allows to perform two-mode maximal interaction clustering for binary response data (e.g., consumers that are willing to buy a product or not, testees that correctly respond to test items or not) would be useful. This can be done by considering a logistic model framework. Additionally, the current method implicitly assumes that the probability to be assigned to a cluster is the same for all clusters. This assumption can be relaxed by introducing a vector of mixing proportions, which leads to a different likelihood and, thus, to a different optimization problem (Symons 1981). Lastly, an alternative way to obtain a sampling distribution of the test statistic would be by using a non-parametric resampling procedure. This alternative approach would be useful if the assumption of normality on the residuals is dubious.

Appendix 1

In order to show that the sum of cross product terms V equals 0, we make use of the fact that the sum of $\widetilde{\gamma}_{pq}$ over $i = 1, ..., I$ is 0:

$$\begin{aligned}\sum_{p=1}^{P}\sum_{i\in R_p}\widetilde{\gamma}_{pq} &= \sum_{p=1}^{P}|R_p|\Big(\frac{1}{|R_p|\cdot|C_q|}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\overline{d}_{.j}+\overline{d}_{..})\Big)\\ &= \frac{1}{|C_q|}\sum_{p=1}^{P}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\overline{d}_{.j}+\overline{d}_{..})\\ &= \frac{1}{|C_q|}\sum_{j\in C_q}(I\cdot\overline{d}_{.j}-I\cdot\overline{d}_{..}-I\cdot\overline{d}_{.j}+I\cdot\overline{d}_{..})\\ &= 0.\end{aligned}$$

For V we have

$$\begin{aligned}V &= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\widetilde{\beta}_j-\widetilde{\gamma}_{pq})\cdot[(\widetilde{\beta}_j-\beta_j)+(\widetilde{\gamma}_{pq}-\gamma_{pq})]\\ &\quad+\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}[(\widetilde{\beta}_j-\beta_j)\cdot(\widetilde{\gamma}_{pq}-\gamma_{pq})]\\ &= \underbrace{\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\widetilde{\beta}_j-\widetilde{\gamma}_{pq})\cdot(\widetilde{\beta}_j-\beta_j)}_{T_1}\\ &\quad+\underbrace{\sum_{p=1}^{P}\sum_{q=1}^{Q}(\widetilde{\gamma}_{pq}-\gamma_{pq})\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij}-\overline{d}_{i.}-\widetilde{\beta}_j-\widetilde{\gamma}_{pq})}_{T_2}\\ &\quad+\underbrace{\sum_{q=1}^{Q}\sum_{j\in C_q}(\widetilde{\beta}_j-\beta_j)\sum_{p=1}^{P}\sum_{i\in R_p}(\widetilde{\gamma}_{pq}-\gamma_{pq})}_{T_3}.\end{aligned}$$

The sum T_3 is 0 because of the fact that $\sum_{p=1}^{P}\sum_{i\in R_p}\widetilde{\gamma}_{pq} = 0$ and identification constraint (2c). Furthermore, the sum T_2 is 0 since the deviations

$$\begin{aligned}\Delta_{ij} &:= d_{ij} - \overline{d}_{i.} - \widetilde{\beta}_j - \widetilde{\gamma}_{pq}\\ &= d_{ij} - \overline{d}_{i.} - \overline{d}_{.j} + \overline{d}_{..} - \widetilde{\gamma}_{pq},\end{aligned}$$

sum to 0 over $i \in R_p$, $j \in C_q$:

$$\begin{aligned}\sum_{i\in R_p}\sum_{j\in C_q}\Delta_{ij} &= \Big(\sum_{i\in R_p}\sum_{j\in C_q}(d_{ij} - \overline{d}_{i.} - \overline{d}_{.j} + \overline{d}_{..})\Big) - (|R_p|\cdot|C_q|\cdot\widetilde{\gamma}_{pq})\\ &= |R_p|\cdot|C_q|\cdot\widetilde{\gamma}_{pq} - |R_p|\cdot|C_q|\cdot\widetilde{\gamma}_{pq}\\ &= 0.\end{aligned}$$

Finally, the sum T_1 is, considering the fact that $\sum_{p=1}^{P}\sum_{i\in R_p}\widetilde{\gamma}_{pq} = 0$, given by

$$\begin{aligned}&\sum\nolimits_{p=1}^{P}\sum\nolimits_{q=1}^{Q}\sum\nolimits_{i\in R_p}\sum\nolimits_{j\in C_q}(d_{ij} - \overline{d}_{i.} - \widetilde{\beta}_j - \widetilde{\gamma}_{pq})\cdot(\widetilde{\beta}_j - \beta_j)\\ &= I\cdot\sum\nolimits_{q=1}^{Q}\sum\nolimits_{j\in C_q}(\overline{d}_{.j} - \overline{d}_{..} - \overline{d}_{.j} + \overline{d}_{..} - 0)\cdot(\widetilde{\beta}_j - \beta_j)\\ &= 0.\end{aligned}$$

Hence, V is a sum of zero sums and equals 0.

Appendix 2

Here we show that

$$\begin{aligned}&d_{ij} - \frac{1}{J}\sum_j d_{ij} - \frac{1}{I}\sum_i d_{ij} + \frac{1}{I\times J}\sum_i\sum_j d_{ij}\\ &= \gamma_{pq} + \epsilon_{ij} - \frac{1}{J}\sum_j \epsilon_{ij} - \frac{1}{I}\sum_i \epsilon_{ij} + \frac{1}{I\times J}\sum_i\sum_j \epsilon_{ij}.\end{aligned}$$

where γ_{pq} is the interaction effect associated to bicluster $R_p \times C_q$ and ϵ_{ij} indicates the residual term for row i and column j as defined in Model (1).

We replace all d_{ij} terms on the left side with Eq. (1): $d_{ij} = \mu + \alpha_i + \beta_j + \gamma_{pq} + \epsilon_{ij}$. We then focus on rewriting each of the components of the expression involving μ, α_i, β_j, γ_{pq} and ϵ_{ij}, solve and then add everything back together.

Overall mean μ:

$$\mu - \frac{1}{J}\sum_j \mu - \frac{1}{I}\sum_i \mu + \frac{1}{I\times J}\sum_i\sum_j \mu = \mu - \frac{J}{J}\mu - \frac{I}{I}\mu + \frac{I\times J}{I\times J}\mu = 0.$$

Row main effect α_i:

$$\alpha_i - \frac{1}{J}\sum_j \alpha_i - \frac{1}{I}\sum_i \alpha_i + \frac{1}{I \times J}\sum_i \sum_j \alpha_i$$
$$= \alpha_i - \frac{J}{J}\alpha_i - \frac{1}{I}\sum_i \alpha_i + \frac{J}{I \times J}\sum_i \alpha_i = 0.$$

Column main effect β_j:

$$\beta_j - \frac{1}{J}\sum_j \beta_j - \frac{1}{I}\sum_i \beta_j + \frac{1}{I \times J}\sum_i \sum_j \beta_j$$
$$= \beta_j - \frac{1}{J}\sum_j \beta_j - \frac{I}{I}\beta_j + \frac{I}{I \times J}\sum_j \beta_j = 0.$$

Row cluster by column cluster interaction γ_{pq}:

$$\gamma_{pq} - \frac{1}{J}\sum_j \gamma_{pq} - \frac{1}{I}\sum_i \gamma_{pq} + \frac{1}{I \times J}\sum_i \sum_j \gamma_{pq} = \gamma_{pq},$$

with the second, third and fourth term on the left side being zero because of constraints (2b) and (2c). No simplification can be made for the ϵ_{ij} terms. In summary we are left with only γ_{pq} and (averages of) the residuals terms ϵ_{ij}, which is what we wanted to show.

References

Alin A, Kurt S (2006) Testing non-additivity (interaction) in two-way ANOVA tables with no replication. Stat Methods Med Res 15:63–85

Andersen EB (1973) Conditional inference for multiple-choice questionnaires. Br J Math Stat Psychol 26:31–44

Anderson JA, Senthilselvan A (1980) Smooth estimates for the hazard function. J R Stat Soc Ser B (Methodol) 42:322–327

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49:803–821

Bock H-H (1980) Simultaneous clustering of objects and variables. In: Tomassone R, Amirchhay M, Néel D (eds) Analyse de données et informatique. Cours de la Commission des Communautés Européennes à Fontainebleau, 19-30 Mars 1979. Institut National de Recherche en Informatique et en Automatique (INRIA), Le Chesnay, pp 187–203

Bock H-H (1996) Probabilistic models in cluster analysis. Comput Stat Data Anal 23:5–28

Caroll JD, Arabie P (1980) Multidimensional scaling. Ann Rev Psychol 31:607–649

Ceulemans E, Kiers HAL (2006) Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method. Br J Math Stat Psychol 59:133–150

Chernick MR (2011) Bootstrap methods: a guide for practitioners and researchers, vol 619. Wiley, New York

Choudhary PK, Nagaraja HN (2017) Measuring agreement: models, methods, and applications, vol 34. Wiley, New York

Cohen J (1992) A power primer. Psychol Bull 112:155

Collins FS, Varmus H (2015) A new initiative on precision medicine. N Engl J Med 372:793–795

Corsten LCA, Denis JB (1990) Structuring interaction in two-way tables by clustering. Biometrics 46:207–215

Denis JB, Gower JC (1994) Biadditive model. Letter to the editor. Biometrics 50:310–311

Efron B (1982) The Jackknife, the bootstrap and other resampling plans, no. 38 in Regional Conference Series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia

Fischer GH, Molenaar IW (1995) Rasch models: foundations, recent developments, and applications. Springer, New York

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc Ser A 222:309–368

Forkman J, Piepho H-P (2014) Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. Biometrics 70:639–647

Franck CT, Nielsen DM, Osborne JA (2013) A method for detecting hidden additivity in two-factor unreplicated experiments. Comput Stat Data Anal 67:95–104

Gauch HG (2006) Statistical analysis of yield trials by AMMI and GGE. Crop Sci 46:1488–1500

Geiser C, Litson K, Bishop J, Keller BT, Burns GL, Servera M, Shiffman S (2015) Analyzing person, situation and person x situation interaction effects: latent state-trait models for the combination of random and fixed situations. Psychol Methods 20:165–192

Hamburg MA, Collins FS (2010) The path to personalized medicine. N Engl J Med 363:301–304

Hennig C, Lin C-J (2015) Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. Stat Comput 25:821–833

Lipkovich I, Dmitrienko A, D'Agostino BR Sr (2017) Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med 36:136–196

Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform 1:24–45

Mandel J (1971) A new analysis of variance model for non-additive data. Technometrics 13:1–18

McLachlan GJ, Peel D (1997) On a resampling approach to choosing the number of components in normal mixture models. In: Billard L, Fisher N (eds) Computing science and statistics, vol 28. Interface Foundation of North America, Fairfax Station, pp 260–266

Miller JW, Harrison MT (2018) Mixture models with a prior on the number of components. J Am Stat Assoc 113:340–356

Mischel W, Shoda Y (1995) A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. Psychol Rev 102:246–268

Mischel W, Shoda Y (1998) Reconciling processing dynamics and personality dispositions. Ann Rev Psychol 49:229–258

Neyman J (1935) Su un teorema concernente le cosiddette statistiche sufficienti. Giornale dell'Istituto Italiano degli Attuari 6:320–334

Piepho H-P (1997) Analyzing genotype-environment data by mixed models with multiplicative terms. Biometrics 53:761–766

Piepho H-P (1999) Fitting a regression model for genotype by environment data on heading dates in grasses by methods for nonlinear mixed models. Biometrics 55:1120–1128

Post JB, Bondell HD (2013) Factor selection and structural identification in the interaction ANOVA model. Biometrics 69:70–79

Quintiens G (1999) Een interactionistische benadering van individuele verschillen in helpen en laten helpen [An interactionist approach to individual differences in helping and allowing to help], Unpublished master's thesis. KULeuven, Belgium

Rothwell PM (1995) Can overall results of clinical trials be applied to all patients? The Lancet 345:1616–1619

Schepers J, Bock H-H, Van Mechelen I (2017) Maximal interaction two-mode clustering. J Classif 34:49–75

Schepers J, Ceulemans E, Van Mechelen I (2008) Selecting among multi-mode partitioning models of different complexities: a comparison of four model selection criteria. J Classif 25:67

Schepers J, Hofmans J (2009) TwoMP: a MATLAB graphical user interface for two-mode partitioning. Behav Res Methods 41:507–514

Schepers J, Van Mechelen I (2011) A two-mode clustering method to capture the nature of the dominant interaction pattern in large profile data matrices. Psychol Methods 16:361–371

Scott AJ, Symons MJ (1971) Clustering methods Based on likelihood ratio criteria. Biometrics 27:387–397

Shoda Y, Wilson NL, Chen J, Gilmore AK, Smith RE (2013) Cognitive-affective processing system analysis of intra-individual dynamics in collaborative therapeutic assessment: translating basic theory and research into clinical applications. J Personal 81:554–1568

Shoda Y, Wilson NL, Whitsett DD, Lee-Dussud J, Zayas V (2015) The person as a cognitive affective processing system: quantitative idiography as an integral component of cumulative science. In: Mikulincer M, Shaver P (eds) APA handbook of personality and social psychology: vol 4. Personality processes and individual differences, vol 4. American Psychological Association APA, Washington, pp 491–513

Symons MJ (1981) Clustering criteria and multivariate normal mixtures. Biometrics 37:35–43

Tukey JW (1949) One degree of freedom for non-additivity. Biometrics 5:232–242

Van Mechelen I, Bock H-H, De Boeck P (2004) Two-mode clustering methods: a structured overview. Stat Methods Med Res 13:363–394

Verbeke G, Spiessens B, Lesaffre E (2001) Conditional linear mixed models. Am Stat 55:25–34

Wilderjans TF, Ceulemens E, Meers K (2013) CHull: a generic convex hull based model selection method. Behav Res Methods 45:1–15