Journal of Classification
https://doi.org/10.1007/s00357-023-09434-2

E-ReMI: Extended Maximal Interaction Two-mode Clustering

Zaheer Ahmed[1] · Alberto Cassese[1,2] · Gerard van Breukelen[1,3] · Jan Schepers[1]

Accepted: 17 February 2023

Abstract
In this paper, we present E-ReMI, a new method for studying two-way interaction in row by column (i.e., two-mode) data. E-ReMI is based on a probabilistic two-mode clustering model that yields a two-mode partition of the data with maximal interaction between row and column clusters. The proposed model extends REMAXINT by allowing for unequal cluster sizes for the row clusters, thus introducing more flexibility in the model. In the manuscript, we use a conditional classification likelihood approach to derive the maximum likelihood estimates of the model parameters. We further introduce a test statistic for testing the null hypothesis of no interaction, discuss its properties and propose an algorithm to obtain its distribution under this null hypothesis. Free software to apply the methods described in this paper is developed in the R language. We assess the performance of the new method and compare it with competing methodologies through a simulation study. Finally, we present an application of the methodology using data from a study of person by situation interaction.

Keywords Bicluster interaction effect parameters · Penalized classification maximum likelihood · Likelihood ratio · Monte Carlo sampling

1 Introduction

This paper addresses the analysis of two-way two-mode data (Carroll & Arabie, 1980) that can be arranged in an $I \times J$ real-valued data matrix $\boldsymbol{D}$, with elements d_{ij} ($i = 1, \ldots, J$, $j = 1, \ldots, J$), and in which the rows pertain to one of the two modes and the columns to the other. Specifically, we consider the case in which both modes are considered as categorical predictor variables and d_{ij} is a real-valued outcome value for the combination of row i and column j. Such data abound in many research settings. An example is that of contextualized personality research, where a set of I persons is measured on some behavior of interest in J different situations. Other examples include the study of micro-array data in genome research

✉ Jan Schepers
jan.schepers@maastrichtuniversity.nl

[1] Department of Methodology and Statistics, Graduate School of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

[2] Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy

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

Published online: 10 May 2023

where DNA expression level is obtained for a set of I genes under J different conditions; Measurement studies, where a set of I units (e.g., tissues, subjects) is measured on some characteristic of interest by a set of J measurement methods (e.g., questionnaires, raters); Research in cognitive psychology, where the response time of I participants is recorded on J test items or stimuli; and consumer research, where a preference rating is obtained for a set of I customers on J different products. Typically, one of the two modes has a large number of elements (e.g., persons) which are a random sample from a population of interest.

A question of scientific interest in these studies is whether there is an interaction between the two modes and, if so, what is its nature. For instance, in studies on aggression it may be of interest to observe how aggressively subjects react in a set of situations. In these studies, in order to understand if and how specific contexts provoke aggression, it is crucial to know whether context effects on aggression are equal for all individuals or not (Geiser et al., 2015; Mischel & Shoda, 1995, 1998; Shoda et al., 2015, 2013). Likewise, if in a study of agreement between measurement methods (Choudhary & Nagaraja, 2017) it appears that two (or more) methods yield different measurements, it is important to know 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).

When, for each combination of a row and column, only one observation is measured, classical methods like ANOVA cannot be used. Thus, in the literature, authors have introduced methods to reduce the number of interaction parameters under study to summarize the massive amount of information in $\mathbf{D}$ or to ensure that some degrees of freedom are available to estimate residual variance (see e.g., (Tukey, 1949; Mandel, 1971; Corsten & Denis, 1990; Denis & Gower, 1994; Gauch, 2006; Post & Bondell, 2013; Franck et al., 2013; Forkman & Piepho, 2014; Alin & Kurt, 2006; Van Mechelen et al., 2004; Madeira & Oliveira, 2004)). Some of these methods focus only on hypothesis testing for interaction (for relatively small data matrices). Shenaravi & Kharrati-Kopaei (2018) review a number of these, further studying different methods for combining them into a single test, and provide an accompanying R package that is available on CRAN (https://CRAN.R-project.org/package=combinIT). However, as explained in the previous paragraph, it is often of major interest to additionally understand what that interaction looks like, especially when dealing with larger data matrices. An example can be found in Piepho (1997), who considers a parsimonious representation of the two modes in terms of latent factor-analytic components capturing as much of the row by column interaction sum of squares as possible. However, with this type of approach, it is challenging to interpret this interaction if more than two latent components are needed to fit the data adequately.

Approaches that are typically more easily interpretable focus on simultaneously clustering the rows and columns (i.e., biclustering/two-mode clustering). For example, Govaert & Nadif (2013) proposed a two-mode clustering method using a mixture approach for modelling the row clusters. This method estimates, among others, the population mean of each combination of a row and a column cluster (i.e., bicluster), which is the joint effect of row and column cluster main effects and of a row by column cluster interaction. This is a block mixture model (Govaert & Nadif, 2003), a class of probabilistic biclustering approaches that includes various methods suitable for specific data types, such as continuous, binary or count (Govaert & Nadif, 2018). Common to models in this class is the assumption that there exists a partition $\mathcal{R}$ into P row clusters $R_1, ..., R_P$ on the row set and a partition $\mathcal{C}$ into Q column clusters $C_1, ..., C_Q$ on the column set, such that the random variables d_{ij} are conditionally independent given $\mathcal{R}$ and $\mathcal{C}$. This implies that these models assume absence of row (e.g., subject) effects within row clusters and, likewise, absence of column (e.g., situation) effects within column clusters. Hence, if there are individual row (column) main effects, the row (column) partitions may tend to capture those effects so as to explain the associations between observations of any two columns (resp. between observations of any two rows) that are implied by those effects (see Section 5 for an illustration). In short, a block mixture model does not yield row and column partitions that maximize interaction but rather the sum of row and column main effects plus row by column interaction. In contrast, REMAXINT (Ahmed et al., 2021) assumes that the random variables d_{ij} are conditionally independent, not only given $\mathcal{R}$ and $\mathcal{C}$, but also the individual row and column main effects. Therefore, this method yields a two-mode partition of $\mathbf{D}$ in which the row and column clusters explain the associations between observations of any two columns (resp. between observations of any two rows) over and above the associations that are implied by row/column main effects. REMAXINT extends maximal interaction two-mode clustering ((Schepers et al., 2017), MAXINT), by allowing the rows to be random instead of fixed.

Other biclustering approaches that are based on interaction concepts include the methods proposed in Cheng & Church (2000) and Cho et al. (2004), which are widely applied to gene expression data. These methods look for biclusters with minimal within-bicluster interaction, that is, minimal mean sum-squared residue (Yu et al., 2021). Specifically, row and column main effects are bicluster-specific and within-bicluster row by column interaction is negligible. This implies that the total row by column interaction sum of squares is represented by the between-bicluster differences with regard to the bicluster means plus the between-bicluster differences with regard to the bicluster-specific main effects. In contrast, in MAXINT (and REMAXINT), row (resp. column) main effects are not specific to column (resp. row) clusters. The total row by column interaction sum of squares is therefore represented more easily by between-bicluster differences with regard to the bicluster means (i.e., biclusters with large between-bicluster interaction). A thorough discussion comparing biclustering methods based on interaction concepts can be found in Schepers et al. (2017).

In this paper, we focus on the objective of finding biclusters with large between-bicluster interaction. Specifically, we propose E-ReMI (Extended-ReMaxInt), a new two-mode clustering method that relaxes the unlikely assumption of equal population cluster sizes that is (implicitly) made in REMAXINT. E-ReMI is a model-based clustering method (Bock, 1996; Banfield & Raftery, 1993) that assumes interaction between row and column clusters in the sense that all row by column interactions are identical within each of the PQ pairs of row and column clusters (i.e., within each bicluster). In other words, it is assumed that any substantial row by column interaction in the data matrix ($\mathbf{D}$) is attributable to a row by column cluster interaction. Our proposed method yields simultaneous partitions of the rows and columns of $\mathbf{D}$ such that a conditional likelihood criterion is maximized. The row and column clusters are not determined by differences between row and column main effects, respectively, but only by row by column interaction effects. Furthermore, row main effects are considered random, and row cluster sizes are allowed to vary between clusters and are unknown parameters to be estimated. Additionally, we introduce a pivotal likelihood ratio test, based on E-ReMI and Monte Carlo sampling, to test the null hypothesis of no interaction between the row and column clusters. Finally, in order to make the methodology of this paper available to a large public, we implement this method in the free software R. This implementation not only includes the newly proposed method but also codes for REMAXINT, which was previously only available in Matlab.

The remainder of this article is organized as follows: In Section 2 we formulate the statistical model and consider parameter estimation using a conditional likelihood approach. In Section 3, we propose a method for statistical inference on the interaction effect parameters that is based on a Monte Carlo scheme. Section 4 investigates, by means of a simulation study, the impact of relaxing the assumption of equal row cluster sizes on statistical power and parameter recovery. This study also includes a comparison, in terms of power, to some of the tests reviewed in Shenaravi & Kharrati-Kopaei (2018). Section 5 studies this impact on a real data set from a study in personality psychology. Finally, Section 6 is dedicated to a discussion and some final remarks.

2 Method

2.1 Model Formulation

We consider a two-mode partitioning of an $I \times J$ two-mode real-valued data matrix $\mathbf{D}$ with elements d_{ij} $(i = 1, \ldots, I, j =, 1, \ldots, J)$ to capture the gist of row by column interaction that is included in these data. We assume that the elements of the row set $R = \{1, \ldots, I\}$ are sampled independently from a population that includes P subpopulations with relative sizes ω_p $(0 \leq \omega_p \leq 1; p = 1, \ldots, P$ and $\sum_{p=1}^{P} \omega_p = 1)$. This results in a latent random P-partition $\mathcal{R} = \{R_1, ..., R_P\}$ of R that is characterized by the random P-dimensional cluster indicator vectors $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_I$, where, for every i, $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iP})^T$, with $Z_{ip} = 1$ for $i \in R_p$ and $Z_{ip} = 0$ for $i \notin R_p$. This implies that $\boldsymbol{Z}_i$ is distributed according to a Multinomial distribution that consists of one draw on P categories with probabilities $\omega_1, \ldots, \omega_P$ (i.e., $\boldsymbol{Z}_i \sim Mult_P(1, \boldsymbol{\omega})$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_P)^T$).

Furthermore, we assume that there is a latent fixed Q-partition $\mathcal{C} = \{C_1, ..., C_Q\}$ of the column set $C = \{1, \ldots, J\}$ that is characterized by the binary indicator matrix $\mathbf{K}$, such that an element $k_{jq} = 1$ for $j \in C_q$ and $k_{jq} = 0$ for $j \notin C_q$ $(j = 1, \ldots, J, q = 1, \ldots, Q)$, with $|C_q| = \sum_{j=1}^{J} k_{jq}$.

The row and column clusters are characterized by two features:

1. The clusters must be jointly exhaustive, i.e. $\bigcup_p R_p = R$ and $\bigcup_q C_q = C$, for rows and columns, respectively;
2. The clusters must be mutually exclusive, i.e. $R_p \bigcap R_{p\prime} = \phi$ $(\forall p \neq p\prime)$, and $C_q \bigcap C_{q\prime} = \phi$ $(\forall q \neq q\prime)$, for rows and columns, respectively.

A bicluster $R_p \times C_q$ is the Cartesian product of row cluster R_p and column cluster C_q and we denote a two-mode partition as $\mathcal{R} \times \mathcal{C} = \{R_p \times C_q; p = 1, \ldots, P, q = 1, \ldots, Q\}$.

Let $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_I$ denote a random sample of size I, where $\boldsymbol{D}_i$ is a J-dimensional vector with probability density function $f(\boldsymbol{d}_i)$ on $\mathbb{R}^J$. We use $\boldsymbol{D} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_I)^T$ to represent the entire sample. We denote by $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_I)^T$ an observed random sample, where $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{iJ})^T$ is the realisation of the random vector $\boldsymbol{D}_i$. Likewise, $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iP})^T$ is the realisation of the random vector $\boldsymbol{Z}_i$, with $|R_p| = \sum_{i=1}^{I} z_{ip}$. Given a two-mode partition $\mathcal{R} \times \mathcal{C}$ of $\boldsymbol{D}$, each of its elements D_{ij}, is assumed to be described by the following linear framework:

$$D_{ij} = \mu + \alpha_i + \beta_j + \gamma_{pq} + \epsilon_{ij}, \tag{1}$$

for $i \in R_p$, $j \in C_q$, $p = 1, \ldots, P$, $q = 1, \ldots, Q$, and where μ is the overall mean, α_i is a random main effect of row i, β_j is a fixed main effect of column j and γ_{pq} is a fixed interaction effect associated to bicluster $R_p \times C_q$. We assume known numbers of row and column clusters P and Q, respectively (we comment on this last assumption in the discussion). We assume that the row main effects α_i are random (with $E(\alpha_i) = 0$ and $\sigma^2_{\alpha_i} > 0$, $\forall i$) and that the column main effects β_j are fixed (with $\sum_{j=1}^{J} \beta_j = 0$). Furthermore, we impose identifiability constraints $\sum_{p=1}^{P} \omega_p \gamma_{pq} = 0$ $(q = 1, \ldots, Q)$ and $\sum_{q=1}^{Q} |C_q| \gamma_{pq} = 0$ $(p = 1, \ldots, P)$ on the interaction effect parameters. Lastly, ϵ_{ij} represents the random error term. The error terms are assumed to be $i.i.d.$ across rows and columns, with mean zero and variance σ^2.

We assume that the residuals ϵ_{ij} in (1) are Normally distributed. Then, conditionally on $i \in R_p$ and $j \in C_q$,

$$D_{ij}|\alpha_i \sim N(\mu + \alpha_i + \beta_j + \gamma_{pq}, \sigma^2),$$

where the distribution function $F(\alpha_i)$ of the random effects α_i does not need to be specified (see below). The random-partition model E-ReMI considers the data $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_I$ as incomplete since the associated row cluster labels $z_1, \ldots, z_I$ are unobserved (i.e., missing). Note that the model assumes that, conditionally on z_i and α_i, the univariate random variables D_{ij} $(j = 1, \ldots, J)$ are statistically independent (i.e., local stochastic independence). Let $g_{ij}(d_{ij}|\alpha_i, \mu, \beta_j, \gamma_{pq})$ denote the density function of D_{ij}, conditionally on $i \in R_p$ and $j \in C_q$. We then have

$$\begin{aligned} f_i(\boldsymbol{d_i}, z_i|\mu, \alpha_i, \boldsymbol{\xi}) &= \prod_{p=1}^{P} \left(\omega_p \prod_{q=1}^{Q} \prod_{j=1}^{J} g_{ij}(d_{ij}|\alpha_i, \mu, \beta_j, \gamma_{pq})^{k_{jq}} \right)^{z_{ip}} \qquad (2) \\ &= \prod_{p=1}^{P} \left(\omega_p \prod_{q=1}^{Q} \prod_{j=1}^{J} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(d_{ij} - \mu - \alpha_i - \beta_j - \gamma_{pq})^2}{\sigma^2} \right) \right)^{k_{jq}} \right)^{z_{ip}} \end{aligned}$$

where $\boldsymbol{\xi} = (\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{K})$ is the vector of unknown parameters, and where

$$\begin{aligned} \boldsymbol{\phi} &= (\beta_1, \ldots, \beta_J, \sigma^2), \\ \boldsymbol{\theta} &= (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P), \\ \boldsymbol{\theta}_p &= (\omega_p, \gamma_{p1}, \ldots, \gamma_{pQ}). \end{aligned}$$

2.2 Conditional Classification Likelihood

In this section we will introduce some notation to shorten and simplify the interpretation of the equations. For the convenience of the reader the introduced notation is listed in Table 1.

Table 1 Summary of the notation introduced in this section to shorten and simplify the equations

Symbol	Definition
W	$W = \exp\left(\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip} \log \omega_p\right)$
V	$V = \exp\left(\frac{-IJ}{2} \log 2\pi\sigma^2\right)$
A_i	$A_i = (\bar{d}_{i.} - \mu - \alpha_i)$
B_{ij}	$B_{ij} = (d_{ij} - \bar{d}_{i.} - \beta_j - \gamma_{pq})$
U	$U = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}B_{ij}A_i$
A	$A = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}A_i^2$
B	$B = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}B_{ij}^2$
H	$H = \frac{IJ}{2}\left(\log\left(\frac{IJ}{2\pi}\right) - 1\right)$
$(dc)_{ij}$	$d_{ij} - \bar{d}_{i.} - \bar{d}_{.j} + \bar{d}_{..}$ (doubly centered data)

For estimation of the parameters of interest, E-ReMI makes use of a combination of two strategies, namely, conditional likelihood and classification likelihood. We first discuss the conditional likelihood approach.

Conditional likelihood is a well-known approach in modern psychometrics (Andersen, 1973; Fischer & Molenaar, 1995) and biostatistics (Anderson & Senthilselvan, 1980) that involves the elimination of so-called nuisance parameters from the likelihood function. The nuisance parameters may pertain to fixed or random effects ((Verbeke & Molenberghs, 2000), sec.13.5). If the nuisance parameters pertain to random effects, the main advantage of this approach is that no distributional assumption is needed with respect to those random effects (Verbeke et al., 2001). Here, we treat the random effects α_i in (2) as nuisance parameters and, rather than maximizing the joint likelihood of all random variables, estimation of the parameters of interest is achieved by maximizing the conditional likelihood given the sufficient statistics ($\bar{d}_{i\cdot}$'s) for these nuisance parameters (α_i's). We only assume that the latter are i.i.d. with finite variance and zero expectation (the latter for identifiability of μ in the model). This leads us to formulate the following theorem.

Theorem *For the I independent joint realizations $(\boldsymbol{d}_1, z_1), \ldots, (\boldsymbol{d}_I, z_I)$, each with density (2), the statistic $\bar{d}_{i\cdot} = \frac{1}{J}\sum_j^J d_{ij}$ $(i = 1, \ldots, I)$ is sufficient for $\mu + \alpha_i$ $(i = 1, \ldots, I)$ under the following constraints: $\sum_{q=1}^{Q} |C_q|\gamma_{pq} = 0$ $(p = 1, \ldots, P)$, where $|C_q|$ indicates the cardinality of column cluster C_q*

Proof We prove sufficiency of $\bar{d}_{i\cdot}$ $(i = 1, \ldots, I)$ for $\mu + \alpha_i$ $(i = 1, \ldots, I)$ by showing that the joint density of the data matrix $\boldsymbol{d}$ and the row partition z

$$f(\boldsymbol{d}, z|\mu, \boldsymbol{\alpha}, \boldsymbol{\xi}) = \prod_{i=1}^{I} f_i(\boldsymbol{d_i}, z_i|\alpha_i, \boldsymbol{\xi}), \tag{3}$$

can be factored such that it satisfies Fisher's factorization theorem (Fisher, 1922; Neyman, 1935; Rice, 2007), where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)$ is the vector of the random row main effects and $z = (z_1, \ldots, z_I)^T$ is the realized vector of the latent random row cluster membership indicator vectors. For our purposes, we rewrite (3) as:

$$f(\boldsymbol{d}, z|\mu, \boldsymbol{\alpha}, \boldsymbol{\xi}) = WV \exp\left(-\frac{1}{2\sigma^2}\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}(d_{ij} - \mu - \alpha_i - \beta_j - \gamma_{pq})^2\right),$$

where $W = \exp\left(\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}\log\omega_p\right)$ and $V = \exp\left(\frac{-IJ}{2}\log 2\pi\sigma^2\right)$. Note that W and V do not depend on the data nor on the main and interaction effect parameters. We then add and subtract $\bar{d}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} d_{ij}$ into the last term of our expression and expand the square, obtaining

$$f(\boldsymbol{d}, z|\mu, \boldsymbol{\alpha}, \boldsymbol{\xi}) = WV \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}(B_{ij}^2 + A_i^2)\right) + 2U\right), \tag{4}$$

where $B_{ij}^2 = (d_{ij} - \bar{d}_{i\cdot} - \beta_j - \gamma_{pq})^2$ and $A_i^2 = (\bar{d}_{i\cdot} - \mu - \alpha_i)^2$ are the square terms, and U is a sum of cross product terms that can be written as

$$U = \sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}A_i\sum_{q=1}^{Q}|C_q|(-\gamma_{pq}). \tag{5}$$

We show how to get this result in Appendix A. Under the set of identifiability constraints $\sum_{q=1}^{Q} |C_q|(-\gamma_{pq}) = 0$ $(p = 1, \ldots, P)$, it follows that $U = 0$. Therefore, (4) reduces to the simplified

$$f(\boldsymbol{d}, z|\mu, \boldsymbol{\alpha}, \boldsymbol{\xi}) = \exp\left(-\frac{1}{2\sigma^2}A\right) WV \exp\left(-\frac{1}{2\sigma^2}B\right), \tag{6}$$

where $B = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}B_{ij}^2$ and $A = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}A_i^2$. The joint density in (6) satisfies Fisher's factorization theorem since W and V do not depend on the observed data, A is a function of $\mu + \alpha_i$ and depends on the observed data only through the statistics $\bar{d}_{i\cdot}$, while B is not a function of $\mu + \alpha_i$. □

Estimation of the vector of unknown parameters $\boldsymbol{\xi} = (\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{K})$, conditional on the sufficient statistics $\bar{d}_{i\cdot}$, implies maximizing the conditional likelihood obtained from (6) by leaving out the first factor. Specifically,

$$CL(z, \boldsymbol{\xi}; \boldsymbol{d}) = WV \exp\left(-\frac{1}{2\sigma^2}B\right) = \prod_{i=1}^{I} f_i(\boldsymbol{d_i}, z_i|\bar{d}_{i\cdot}, \boldsymbol{\xi}), \tag{7}$$

is maximized with respect to $\boldsymbol{\xi}$ and the unobserved row cluster labels $z_1, \ldots, z_I$. Since the latter are treated as parameters to be estimated along with $\boldsymbol{\xi}$, this is referred to as a classification likelihood approach (Scott & Symons, 1971; Symons, 1981; Bock, 1996; Govaert & Nadif, 2013). No restrictions are put on z (resp. $\mathbf{K}$) other than that $\sum_{p=1}^{P} z_{ip} = 1$ $i = 1, \ldots, I$ (resp. $\sum_{q=1}^{Q} k_{jq} = 1$ $j = 1, \ldots, J$) and $\sum_{i=1}^{I} z_{ip} \geq 1$ $p = 1, \ldots, P$ (resp. $\sum_{j=1}^{J} k_{jq} \geq 1$ $q = 1, \ldots, Q$).

After applying a logarithmic transformation to (7), the equation for the conditional classification log-likelihood becomes:

$$\ell(\boldsymbol{\xi}) = \log W + \log V - \frac{1}{2\sigma^2}B. \tag{8}$$

The unknown parameter σ^2 is involved in the kernel of this equation (and in V) and, thus, has to be accounted for in the maximization. We replace it with its maximizer, obtained by partial differentiation of the equation above. This yields the following result

$$\sigma^2 = \frac{\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}(d_{ij} - \bar{d}_{i\cdot} - \beta_j - \gamma_{pq})^2}{IJ} = \frac{B}{IJ}.$$

After plugging this expression in the equation above, and after some simplifications, we obtain the following criterion to maximize (see Appendix B):

$$\log CL = \log W - \frac{IJ}{2}\log(B) + H,$$

where $H = \frac{IJ}{2}\left(\log\left(\frac{IJ}{2\pi}\right) - 1\right)$ is an additive constant. This criterion is fully written as

$$\log CL = \left(\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}\log\omega_p\right) - \frac{IJ}{2}\log\left(\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}(d_{ij} - \bar{d}_{i\cdot} - \beta_j - \gamma_{pq})^2\right) + H \tag{9}$$

where ω_p $(p = 1, \ldots, P)$, β_j $(j = 1, \ldots, J)$, γ_{pq} $(p = 1, \ldots, P, q = 1, \ldots, Q)$, z_{ip} $(i = 1, \ldots, I, p = 1, \ldots, P)$ and k_{jq} $(j = 1, \ldots, J, q = 1, \ldots, Q)$ are to be estimated, subject to the constraints given in this section. This is a penalized classification likelihood criterion (Bryant, 1991; Celeux & Govaert, 1992), where the first term is a penalty term that penalizes row partitions that imply higher levels of unpredictability (Shannon, 1948), that is, solutions with more balanced cluster sizes.

2.3 Model fitting

Maximizing (9) with respect to the unknown parameters $\boldsymbol{z}$, $\boldsymbol{\omega}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{K}$ is a mixed continuous-combinatorial optimization problem because B and W involve the unknown cluster membership indicators z_{ip} and k_{jq} on top of the unknown row cluster sizes ω_p and the fixed effects β_j and γ_{pq}. For this optimization problem, there are currently no routines available in statistical software. We therefore developed an iterative procedure that is presented in this subsection. To explain this iterative procedure, we will first discuss parameter estimation of ω_p, β_j, and γ_{pq} for a given two-mode partition $\mathcal{R} \times \mathcal{C}$, that is, for given matrices z and $\mathbf{K}$. Subsequently, we discuss how to maximize the criterion given in (9) with respect to the two-mode partition $\mathcal{R} \times \mathcal{C}$.

2.3.1 Parameter estimation given an arbitrary two-mode partition

For a fixed number of row and column clusters, P and Q, and a given arbitrary two-mode partition $\mathcal{R} \times \mathcal{C}$, we obtain the maximum likelihood estimates of ω_p, β_j and γ_{pq} by maximizing (9) under the identifiability constraints $\sum_{p=1}^{P} \omega_p = 1$, $\sum_{j=1}^{J} \beta_j = 0$, $\sum_{p=1}^{P} \omega_p \gamma_{pq} = 0$ $(q = 1, \ldots, Q)$ and $\sum_{q=1}^{Q} |C_q| \gamma_{pq} = 0$ $(p = 1, \ldots, P)$. This leads to the following estimates (see Appendix C) for the parameters of interest:

$$\begin{aligned}
\hat{\omega}_p &= \frac{|R_p|}{I}, \\
\hat{\beta}_j &= \bar{d}_{.j} - \bar{d}_{..}, \\
\hat{\gamma}_{pq} &= \frac{1}{|R_p||C_q|} \sum_{i=1}^{I} \sum_{j=1}^{J} z_{ip} k_{jq} \left(d_{ij} - \bar{d}_{i.} - \bar{d}_{.j} + \bar{d}_{..} \right),
\end{aligned}$$

for $j = 1, \ldots, J$, $p = 1, \ldots, P$ and $q = 1, \ldots, Q$

2.3.2 Estimation of two-mode partition

In the previous subsection, we discussed parameter estimation of ω_p, β_j and γ_{pq}, given an arbitrary two-mode partition $\mathcal{R} \times \mathcal{C}$. The challenge of finding the best fitting E-ReMI model for a data set at hand is completed by addressing the estimation of z and $\mathbf{K}$.

Based on the estimates of ω_p, β_j and γ_{pq}, finding the best fitting configuration (i.e, the configuration that maximizes $\ell(\boldsymbol{\xi})$), given P and Q, comes down to maximizing the following clustering criterion[1]:

$$CC = \sum_{p=1}^{P} \sum_{i=1}^{I} z_{ip} \log \hat{\omega}_p - \frac{IJ}{2} \log \left(\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i=1}^{I} \sum_{j=1}^{J} z_{ip} k_{jq} (d_{ij} - \bar{d}_{i.} - \hat{\beta}_j - \hat{\gamma}_{pq})^2 \right),$$

[1] This criterion is the likelihood Criterion (9) in which the parameters ω_p, β_j and γ_{pq} are replaced by their maximum likelihood estimates and the constant H has been dropped.

with respect to z and $\mathbf{K}$. This is a combinatorial optimization problem, for which it is not feasible to apply a procedure that guarantees finding the global maximizer. Instead, we use a greedy optimization algorithm that starts from some initial configuration $(z^0, \mathbf{K}^0)$ and deterministically searches through the solution space as long as neighbouring configurations with a better likelihood value can be found.

In order to explain the details of this estimation algorithm, it is useful to rewrite the optimization problem as follows:

$$CC = \sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}\log(\sum_{i=1}^{I}\frac{z_{ip}}{I}) - \frac{IJ}{2}\log\left(\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}((dc)_{ij} - \hat{\gamma}_{pq})^2\right), \quad (10)$$

where $(dc)_{ij} = d_{ij} - \bar{d}_{i\cdot} - \bar{d}_{\cdot j} + \bar{d}_{\cdot\cdot}$ is a transformation (double centering, see Table 1) of the observed data that can be computed once, before the start of the algorithm. The algorithm generates increasing values of the (log)likelihood by iterating between obtaining updated estimates of the row and column cluster membership indicators and the interaction effect parameters, respectively. It is described in more detail in Algorithm 1. In order to increase the probability of finding the global maximum, one may run the algorithm M times using independently generated random initial configurations $(z^0, \mathbf{K}^0)$ and retain the configuration that yields the highest value of (10).

Note that the assumption of equal row cluster sizes can be imposed by setting $\hat{\omega}_p = \frac{1}{P}$. This implies that the term $\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}\log(\sum_{i=1}^{I}\frac{z_{ip}}{I})$ in (10) reduces to the constant $-I\log(P)$. Let $(CC)^*$ denote this constrained criterion. Maximizing $(CC)^*$ is equivalent to fitting a REMAXINT model (Ahmed et al., 2021) to the data at hand.

3 Hypothesis testing

In this section we are concerned with testing the null hypothesis that all interaction effect parameters in (1) are equal to zero. Formally,

$$\begin{aligned} &H_0 : \gamma_{pq} = 0 && (\text{for} \quad p = 1, \ldots, P, q = 1, \ldots, Q) \\ &H_1 : \exists(p, q) \text{ s.t.} && \gamma_{pq} \neq 0, \end{aligned}$$

for a fixed number of row and column clusters P and Q, respectively. We propose a likelihood ratio test for testing this null hypothesis. This implies obtaining the conditional likelihood of the data under the null hypothesis. In the next subsection, we first obtain this expression and then we propose a new test statistic based on the ratio of two conditional likelihoods, that is, the conditional likelihoods under the alternative and null hypotheses.

3.1 Test statistic

Under the null hypothesis, all interaction effect parameters in (1) are equal to zero (i.e., $\gamma_{pq} = 0$, $p = 1, \ldots, P$, $q = 1, \ldots, Q$). This implies that (1) reduces to

$$d_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad (11)$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Note that (11) implies that there is no partitioning in the data generating mechanism assumed by the reduced model. In order to obtain the conditional likelihood expression for this model, we start by denoting the reduced parameter

Algorithm 1 Algorithm to find E-ReMI local maximum.

```
Input: $d_{ij} (i = 1, \dots, I, j = 1, \dots, J), P, Q$
Output: $\mathbf{z}, \mathbf{K}$ (Estimated row and column cluster memberships)

function LocalMax($\mathbf{d}, P, Q$)
    $(dc)_{ij} \leftarrow (d_{ij} - \bar{d}_{i\cdot} - \bar{d}_{\cdot j} + \bar{d}_{\cdot\cdot})$
    $n \leftarrow 0$
    Randomly generate a starting configuration $(z^n, \mathbf{K}^n)$
    Compute $\hat{\gamma}^n_{pq}$ for that configuration
    $CC^{(0)} \leftarrow$ the value of (10) for this initial configuration
    $\zeta \leftarrow$ Any arbitrary value larger than 0
    while $\zeta > 0$ do
        Keep $\mathbf{K}^n$ fixed and find update $z^{n+1}$ as follows:
        for $(i = 1, \dots, I)$ do
            if $i$ not in a singleton cluster then
                for $(p = 1, \dots, P)$ do
                    define candidate $z$ by setting $z_{ip} \leftarrow 1$ and $z_{ip^*} \leftarrow 0 (p^* \neq p)$
                    update $\hat{\gamma}_{pq}$ and evaluate (10)
                end for
                choose candidate $z$ that maximizes (10) as $z^{n+1}$
            end if
        end for
        Keep $z^{n+1}$ fixed and find update $\mathbf{K}^{n+1}$ as follows:
        for $(j = 1, \dots, J)$ do
            if $j$ not in a singleton cluster then
                for $(q = 1, \dots, Q)$ do
                    define candidate $\mathbf{K}$ by setting $k_{jq} \leftarrow 1$ and $k_{jq^*} \leftarrow 0 (q^* \neq q)$
                end for
            end if
            choose candidate $\mathbf{K}$ that maximizes (10) as $\mathbf{K}^{n+1}$
        end for
        compute $\hat{\gamma}^{n+1}_{pq}$ based on $z^{n+1}$ and $\mathbf{K}^{n+1}$
        compute value of (10) for $z^{n+1}$, $\mathbf{K}^{n+1}$ and $\hat{\gamma}^{n+1}_{pq}$
        $CC^{(n+1)} \leftarrow$ the value of (10) for the current configuration;
        $\zeta \leftarrow CC^{(n+1)} - CC^{(n)}$;
        $n \leftarrow n + 1$
    end while
    return $(\mathbf{z}, \mathbf{K})$
end function
```

vector as $\boldsymbol{\xi}^0 = \boldsymbol{\phi} = (\mu, \beta_1, \dots, \beta_J, \sigma^2)$, that does not include any parameters that pertain to a bipartition $\mathcal{R} \times \mathcal{C}$ (since there are no clusters). The density (conditional on $\boldsymbol{\alpha}$) of the entire sample, can be written as:

$$f(\boldsymbol{d}|\boldsymbol{\alpha}, \boldsymbol{\xi}^0) = \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(d_{ij} - \mu - \alpha_i - \beta_j)^2}{\sigma^2} \right) \right).$$

We next factorize this joint density by applying the same steps as in Section 2.2 and finally obtain the conditional log-likelihood for the model under the null hypothesis:

$$\ell(\boldsymbol{\xi}^0) = \log V - \frac{1}{2\sigma^2} B,$$

where V and B are defined as in Section 2.2 but with $\gamma_{pq} = 0$ ($p = 1, \ldots, P, q = 1, \ldots, Q$). Replacing σ^2 by its maximizer yields, after some simplifications, the following criterion to maximize w.r.t. the β_j's

$$\log CL^0 = -\frac{IJ}{2} \log \left(\sum_{i=1}^{I} \sum_{j=1}^{J} (d_{ij} - \bar{d}_{i\cdot} - \beta_j)^2 \right) + \frac{IJ}{2} \left(\log \left(\frac{IJ}{2\pi} \right) - 1 \right).$$

It can be shown that the conditional m.l. estimate of β_j is $\hat{\beta}_j = \bar{d}_{\cdot j} - \bar{d}_{\cdot\cdot}$.

We propose the following test statistic to test the null hypothesis of no interaction:

$$\lambda_{LR} = \ell(\hat{\boldsymbol{\xi}}) - \ell(\hat{\boldsymbol{\xi}}^0),$$

where $\ell(\hat{\boldsymbol{\xi}})$ and $\ell(\hat{\boldsymbol{\xi}}^0)$ are the logarithms of the maximized conditional likelihood functions for the alternative and null model, respectively. After some simplifications, this yields the following form of the test statistic:

$$\begin{aligned} \lambda_{LR} = {} & \sum_{p=1}^{P} \sum_{i=1}^{I} z_{ip} \log \hat{\omega}_p \\ & + \frac{IJ}{2} \left(\log \left(\sum_{i=1}^{I} \sum_{j=1}^{J} (dc)_{ij}{}^2 \right) - \log \left(\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i=1}^{I} \sum_{j=1}^{J} z_{ip} k_{jq} ((dc)_{ij} - \hat{\gamma}_{pq})^2 \right) \right), \end{aligned} \tag{12}$$

where in practice z_{ik} and k_{jq} are replaced by their estimated values $\hat{z}_{ik}$ and $\hat{k}_{jq}$ as returned by Algorithm 1. Note that the difference between the logarithms in the parentheses (second line) is a difference between log-squared-residuals that is necessarily non-negative and is expected to grow if I and/or J is increased. The penalty term $\sum_{p=1}^{P} \sum_{i=1}^{I} z_{ip} \log \hat{\omega}_p = \sum_{p=1}^{P} |R_p| \log \hat{\omega}_p$ is necessarily negative (for $P \geq 2$), does not depend on J, and becomes smaller (i.e., larger in absolute value) if I increases.

It is worth noting that if one imposes an assumption of equal row cluster sizes by setting $\hat{\omega}_p = \frac{1}{P}$, this penalty term becomes a constant $E = -I \log(P)$. Let this constrained version of the test statistic be denoted as λ^*_{LR}:

$$\begin{aligned} \lambda^*_{LR} &= E + \frac{IJ}{2} \left(\log \left(\sum_{i=1}^{I} \sum_{j=1}^{J} (dc)_{ij}{}^2 \right) - \log \left(\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i=1}^{I} \sum_{j=1}^{J} z_{ip} k_{jq} ((dc)_{ij} - \hat{\gamma}_{pq})^2 \right) \right), \\ &= E + \frac{IJ}{2} \left(\log \left(\frac{\sum_{i=1}^{I} \sum_{j=1}^{J} (dc)_{ij}{}^2}{\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i=1}^{I} \sum_{j=1}^{J} z_{ip} k_{jq} ((dc)_{ij} - \hat{\gamma}_{pq})^2} \right) \right). \end{aligned} \tag{13}$$

Note that, for a given data set, the numerator of the term in the logarithm is a constant. Therefore, (13) is maximized by minimizing the denominator of the term in the logarithm. Furthermore, minimizing that term is equivalent to maximizing $\sum_{i=1}^{I} \sum_{j=1}^{J} |R_p||C_q|(\hat{\gamma}_{pq})^2$, since

$$\sum_{i=1}^{I} \sum_{j=1}^{J} (dc)_{ij}{}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} |R_p||C_q|(\hat{\gamma}_{pq})^2 + \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{i=1}^{I} \sum_{j=1}^{J} z_{ip} k_{jq} ((dc)_{ij} - \hat{\gamma}_{pq})^2.$$

This implies that the maximizer of (13) is the maximizer of

$$\text{max-}F = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}|R_p||C_q|(\hat{\gamma}_{pq})^2}{\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}((dc)_{ij}-\hat{\gamma}_{pq})^2},$$

which is the test statistic used to test for interaction in REMAXINT (Ahmed et al., 2021). We therefore refer to a test based on (13) as REMAXINT test.

It is important to emphasize that $\boldsymbol{z}$ and $\boldsymbol{K}$ are not known, but must be estimated from the data by maximizing (9). As a result, the sampling distribution of λ_{LR} is not known and must be obtained by Monte Carlo simulations, just like related statistics defined on clustering approaches (Ahmed et al., 2021; Bock, 1996). We will elaborate on the Monte Carlo procedure in the next subsection, but first it is useful to discuss some properties of λ_{LR}.

Property 1 *The distribution of λ_{LR} under H_0 and H_1 does not depend on the value of the unknown residual variance σ^2.*

Proof Consider the following transformation of the data: $d'_{ij} = m \cdot d_{ij}$, that is, multiplication by a constant factor m. This transformation implies the residual variance of the transformed data to be $\sigma'^2 = m^2 \cdot \sigma^2$. This transformation further implies $(dc)'_{ij} = m \cdot (dc)_{ij}$ and $\hat{\gamma}'_{pq} = m \cdot \hat{\gamma}_{pq}$. Since all terms within the squares of λ_{LR} are multiplied by this factor m, the arguments of the logarithms in the second and third term of (12) are multiplied by m^2 and cancel out of the equation. □

Property 2 *The distribution of λ_{LR} under H_0 and H_1 does not depend on the values of the unknown parameters $\mu, \alpha_i (i = 1, \ldots, I)$ and $\beta_j (j = 1, \ldots, J)$.*

Proof This follows from $(dc)_{ij}$ being only a function of γ_{pq} and ϵ_{ij}, which has been shown in Ahmed et al. (2021). □

3.2 Computational Procedure

In order to test the null hypothesis, it is possible to draw from the true null distribution of λ_{LR} rather than using a bootstrap approach (e.g. (McLachlan & Peel, 1997; Hennig & Lin, 2015)). Specifically, in order to obtain the sampling distribution of the test statistic λ_{LR} under the null hypothesis of no interaction, we propose a three steps Monte Carlo scheme (for fixed numbers of row and column clusters P and Q, respectively),

- Step 1: Generate a data matrix $\mathbf{D}^{(\text{sim})}$ of size $I \times J$ such that each cell (ij) contains a single observation $D_{ij}^{(\text{sim})} \sim \mathcal{N}(\mu^{(\text{sim})} + \alpha_i^{(\text{sim})} + \beta_j^{(\text{sim})}, \sigma^{2(\text{sim})})$, which is the model under the null hypothesis of no interaction. We discuss briefly how to set these parameters below.
- Step 2: Fit the model based on (1) to the generated data matrix $\mathbf{D}^{(\text{sim})}$ using Algorithm 1 to obtain estimates $\hat{\boldsymbol{z}}, \hat{\mathbf{K}}, \hat{\boldsymbol{\omega}}$, and $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_P)^T$ - with elements $\hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}_{p1}, \ldots, \hat{\gamma}_{pQ})^T$ - and compute λ_{LR}.
- Step 3: Repeat Step 1 and Step 2 L times. This yields the set of Monte Carlo values $\lambda_{LR}^{(l)}$ $(l = 1, \ldots, L)$.

If L is sufficiently large, the empirical distribution of $\lambda_{LR}^{(l)}$ approaches the sampling distribution of λ_{LR} under the null hypothesis. Importantly, in Step 1, **Property 1** implies that one may choose any arbitrary value, other than 0, for $\sigma^{2(\text{sim})}$. Furthermore, **Property 2** implies that one can set, without loss of generality, $\mu^{(\text{sim})} = 0$, $\alpha_i^{(\text{sim})} = 0$ $(i = 1, \ldots, I)$, $\beta_j^{(\text{sim})} = 0$ $(j = 1, \ldots, J)$.

4 Simulation Study

In this section, we report an evaluation of the proposed methodology in terms of several criteria. In the following subsections, we first present the design of the simulation studies and then discuss the results, focusing on Type-I error rate and power of the likelihood ratio test of interaction performed through λ_{LR}. Subsequently, we focus on evaluating E-ReMI estimates in terms of parameter recovery.

4.1 Design

In this subsection we discuss the design of three simulation studies. The first study is used to establish critical values of the test statistic λ_{LR} as a function of two completely crossed experimental factors: size $(I \times J)$ of the data set and number of clusters (P^*, Q^*) as assumed in the data analysis. The critical values are found following the three steps in the computational procedure described in Section 3.2, with $L = 5000$ and $M = 20$ number of random starts. The second simulation study investigates to what extent these critical values are subject to sampling errors, since the sampling distribution is generated based on a finite L. The third simulation study is to assess the power of the test statistic λ_{LR} to detect row by column interaction. In the first two simulation studies the design factors size of the data and number of clusters were varied across a range of values:

(i) size $(I \times J)$ of the data, at 6 levels: 20×20, 40×20, 50×30, 30×50, 100×20, 200×30;
(ii) number of clusters (P^*, Q^*), at 5 levels: $(2, 2)$, $(3, 2)$, $(4, 2)$, $(3, 3)$, $(4, 4)$.

In the third simulation study, there are two additional design factors, which are, the true number of clusters used for data generation (P, Q) and equality of expected row cluster sizes. In this simulation study the design factors are not fully crossed.

4.1.1 Critical Values

In order to determine critical values of the proposed test statistic λ_{LR} we applied the three steps Monte Carlo scheme described in Section 3.2. Specifically, we generated $L = 5000$ independent data sets without any row by column interaction for each level of the design factor size $(I \times J)$ of the data. That is, data were generated from the null model such that $D_{ij}^{(\text{sim})} \sim \mathcal{N}(\mu^{(\text{sim})} + \alpha_i^{(\text{sim})} + \beta_j^{(\text{sim})}, \sigma^{2(\text{sim})})$. Without loss of generality, we set $\sigma^{2(\text{sim})} = 1$ and, likewise, $\mu^{(\text{sim})} + \alpha_i^{(\text{sim})} + \beta_j^{(\text{sim})} = 0$, since λ_{LR} is pivotal with respect to these parameters (see Section 3.2).

On each generated data set, we applied E-ReMI for each level of number of clusters assumed in the analysis. For any combination of size and number of clusters this yields $L = 5000$ simulated Monte Carlo test statistic values $\boldsymbol{\lambda}_{LR} = \{\lambda_{LR}^{(l)}; l = 1, \ldots, L\}$, whose distribution approaches the sampling distribution of λ_{LR} under the null hypothesis, if L is sufficiently large (Efron, 1982; Chernick, 2011). From this simulated empirical distribution of λ_{LR} we then obtain critical values $\lambda_{LR}^{(\alpha)}$ for any significance level α by finding its $100(1-\alpha)$th quantile.

4.1.2 Type-I Error Rate

In a second simulation study we investigate to which extent sampling error affects establishing critical values if $L = 5000$. Specifically, for each level of size of the data, we generated a new set of 5000 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 $\beta_j \sim \mathcal{N}(0, 1)$ rather than set all of them to 0 as in the first simulation study. Furthermore, given **Property 1** of the test statistic λ_{LR} (see Section 3.1), we set the error variance arbitrarily at $\sigma^2 = 7$.

Each data set was then analyzed by applying E-ReMI for each level of number of clusters. Every single analysis yields an observed value of the test statistic $\lambda_{LR}^{(\mathrm{obs})}$, which may be compared to the critical value for that combination of size and number of clusters as was obtained in the first simulation study in Section 4.1.1. Specifically, if $\lambda_{LR}^{(\mathrm{obs})} > \lambda_{LR}^{(\alpha)}$ 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 number of clusters, the proportion of $\lambda_{LR}^{(\mathrm{obs})}$ values (out of all 5000 data sets for that combination of size and number of clusters) that fall in the rejection region corresponds to the empirical Type-I error rate of the E-ReMI interaction test. If this study yields empirical Type-I error rates close to the nominal level $\alpha = 0.05$, we may conclude that choosing $L = 5000$ is sufficient for accurately establishing critical values.

4.1.3 Power and Parameter Recovery

For this study, data sets were generated with a true underlying two-mode clustering structure for the row by column interaction. Four design factors were varied in this study:

(i) size $(I \times J)$ of the data, at 6 levels: 20×20, 40×20, 30×50, 50×30, 100×20, 200×30;
(ii) equality of expected row cluster sizes, at 2 levels: unequal versus equal;
(iii) true number of clusters (P, Q), at 3 levels: (2, 2), (3, 3), (4, 4);
(iv) number of clusters for the analysis (P^*, Q^*), at 3 levels: one level where $(P^*, Q^*) = (P, Q)$ and two other levels where the value of (P^*, Q^*) implies a misspecification compared to (P, Q) (for details, see Figs. 1, 2, 3, 4, 5 and 6).

To generate a data set of size $I \times J$, with true number of clusters equal to (P, Q), and with unequal expected row cluster sizes, we used the following procedure. First, randomly generate row and column partition matrices $\mathbf{z}$ and $\mathbf{K}$. Specifically, an $I \times P$ row partition matrix $\mathbf{z}$ with unequal row cluster sizes (in expectation) is generated by randomly assigning each row to a row cluster R_p with probability ω_p, where $\omega_1 = 0.7$ and $\omega_p = (1 - \omega_1)/(P - 1)$, $(\forall p \neq 1)$. This means one may expect one large row cluster that includes 70% of the rows, while the remaining 30% of the rows are distributed evenly across the remaining row clusters. Since this is true only in expectation, it is possible that this step yields empty row clusters, in which case it is repeated until a $\mathbf{z}$ without empty clusters is generated. We further study the case of equal row cluster sizes (in expectation), where the $I \times P$ row partition matrix $\mathbf{z}$ is generated by randomly assigning each row to a row cluster R_p with probability $\omega_p = 1/P$, $(\forall p)$, and this step is, if necessary, repeated until a $\mathbf{z}$ without empty clusters is generated. Likewise, under each scenario, a $J \times Q$ column partition matrix $\mathbf{K}$ is generated by randomly assigning each column to a column cluster C_q with probability $1/Q$ (if necessary repeated until a $\mathbf{K}$ without empty clusters is obtained).

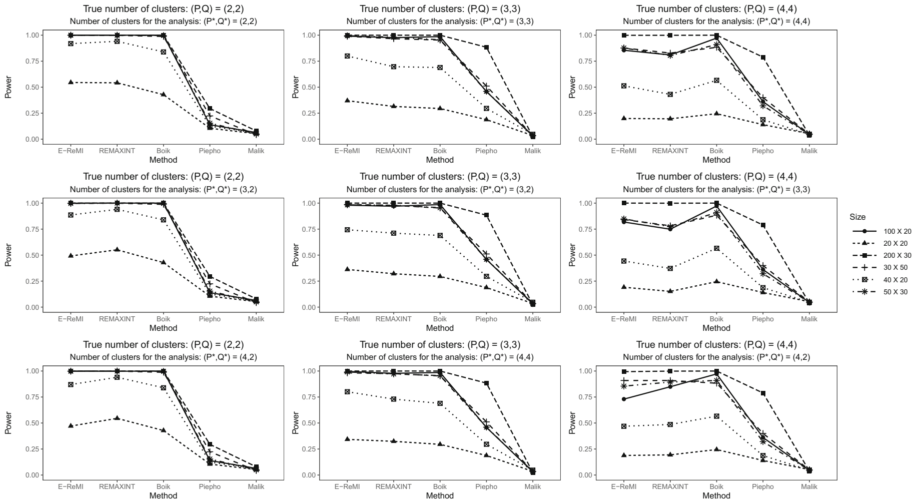

Fig. 1 Empirical power as a function of method (x-axis) and size of the data (curves), for data generated with unequal expected row cluster sizes. Column subfigures refer to true number of clusters set to (2, 2), (3, 3) and (4, 4), for the first, second and third column, respectively. Subfigures in the first row are obtained when the number of clusters for the analysis coincides with the true number of clusters (i.e., no misspecification), while the second and third row correspond to misspecification of the number of clusters for the analysis (see subfigure headings for details)

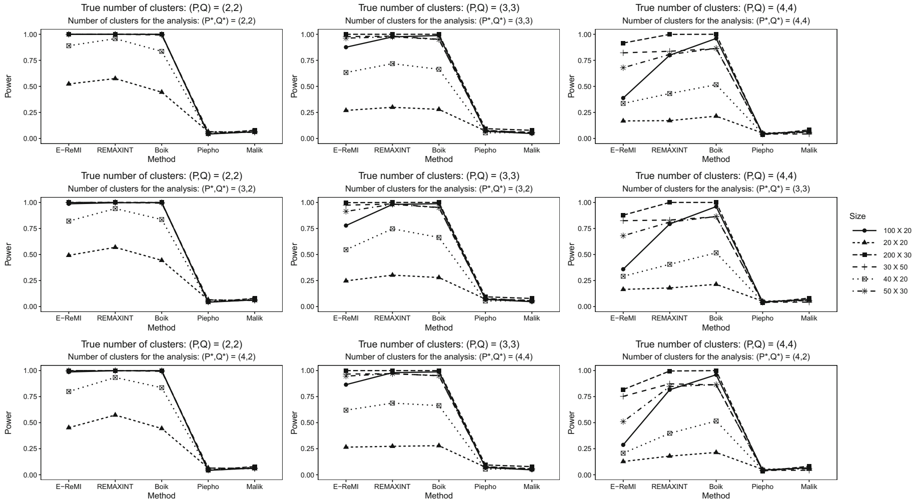

Fig. 2 Empirical power as a function of method (x-axis) and size of the data (curves), for data generated with equal expected row cluster sizes. Column subfigures refer to true number of clusters set to (2, 2), (3, 3) and (4, 4), for the first, second and third column, respectively. Subfigures in the first row are obtained when the number of clusters for the analysis coincides with the true number of clusters (i.e., no misspecification), while the second and third row correspond to misspecification of the number of clusters for the analysis (see subfigure headings for details)

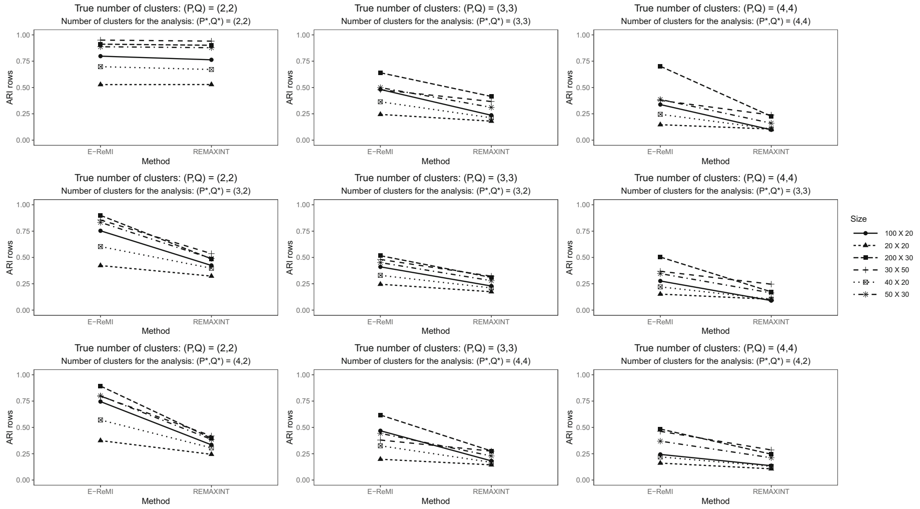

Fig. 3 Mean ARI for row clusters as a function of method (x-axis) and size of the data (curves), for data generated with unequal expected row cluster sizes. Column subfigures refer to true number of clusters set to $(2, 2)$, $(3, 3)$ and $(4, 4)$, for the first, second and third column, respectively. Subfigures in the first row are obtained when the number of clusters for the analysis coincides with the true number of clusters (i.e., no misspecification), while the second and third row correspond to misspecification of the number of clusters for the analysis (see subfigure headings for details)

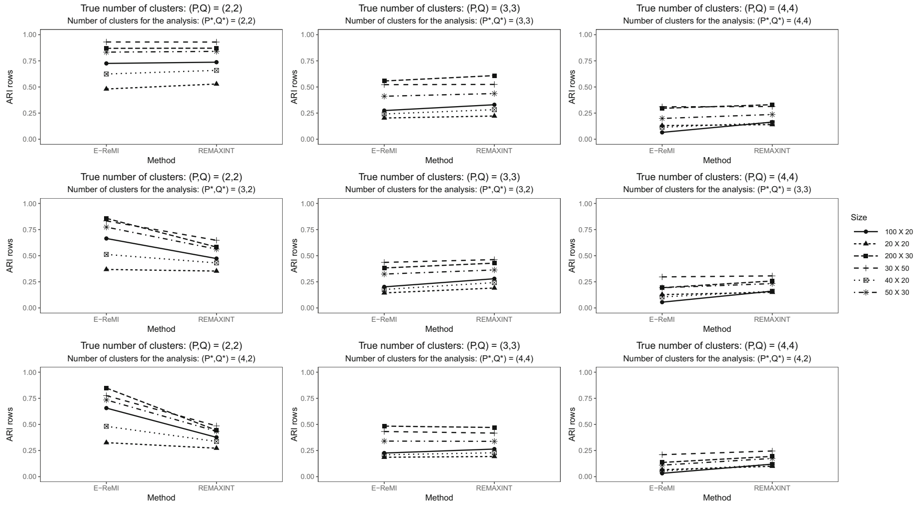

Fig. 4 Mean ARI for row clusters as a function of method (x-axis) and size of the data (curves), for data generated with equal expected row cluster sizes. Column subfigures refer to true number of clusters set to $(2, 2)$, $(3, 3)$ and $(4, 4)$, for the first, second and third column, respectively. Subfigures in the first row are obtained when the number of clusters for the analysis coincides with the true number of clusters (i.e., no misspecification), while the second and third row correspond to misspecification of the number of clusters for the analysis (see subfigure headings for details)

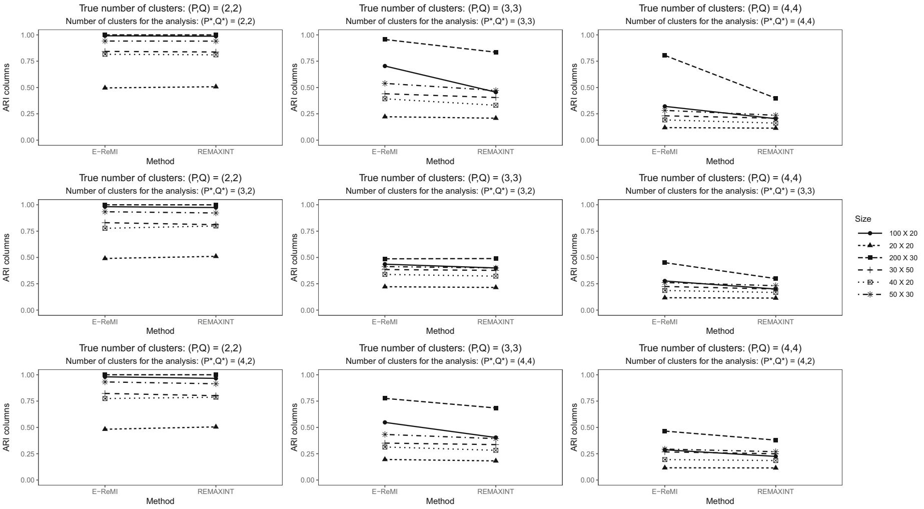

Fig. 5 Mean ARI for column clusters as a function of method (x-axis) and size of the data (curves), for data generated with unequal expected row cluster sizes. Column subfigures refer to true number of clusters set to $(2, 2)$, $(3, 3)$ and $(4, 4)$, for the first, second and third column, respectively. Subfigures in the first row are obtained when the number of clusters for the analysis coincides with the true number of clusters (i.e., no misspecification), while the second and third row correspond to misspecification of the number of clusters for the analysis (see subfigure headings for details)

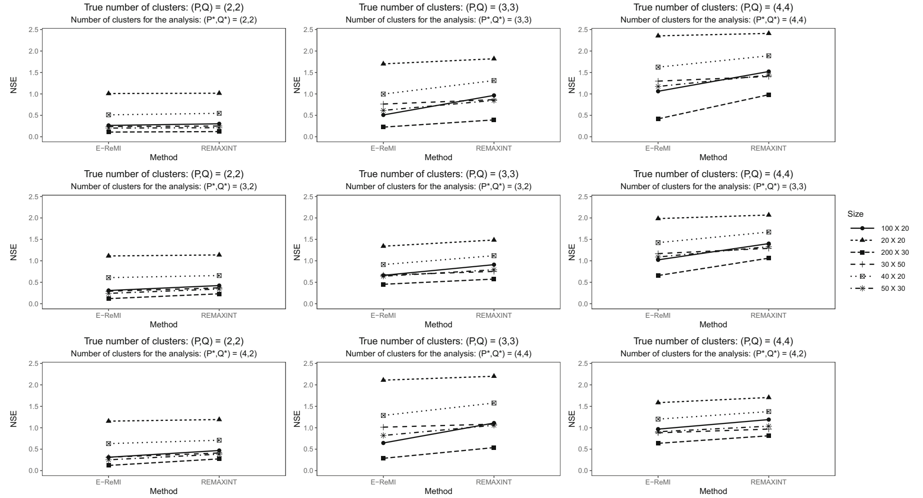

Fig. 6 Mean NSE as a function of method (x-axis) and size of the data (curves), for data generated with unequal expected row cluster sizes. Column subfigures refer to true number of clusters set to $(2, 2)$, $(3, 3)$ and $(4, 4)$, for the first, second and third column, respectively. Subfigures in the first row are obtained when the number of clusters for the analysis coincides with the true number of clusters (i.e., no misspecification), while the second and third row correspond to misspecification of the number of clusters for the analysis (see subfigure headings for details)

Next, each level of true number of clusters, we constructed a fixed $P \times Q$ matrix $\mathbf{\Gamma}$ of true interaction parameter values γ_{pq} as

$$\mathbf{\Gamma} = \begin{pmatrix} \frac{-1}{I\omega_1|C_1|} & \frac{+1}{I\omega_1|C_2|} \\ \frac{+1}{I\omega_2|C_1|} & \frac{-1}{I\omega_2|C_2|} \end{pmatrix}, \qquad \text{if } (P, Q) = (2, 2),$$

$$\mathbf{\Gamma} = \begin{pmatrix} \frac{-1}{I\omega_1|C_1|} & 0 & \frac{+1}{I\omega_1|C_3|} \\ 0 & 0 & 0 \\ \frac{+1}{I\omega_3|C_1|} & 0 & \frac{-1}{I\omega_3|C_3|} \end{pmatrix}, \qquad \text{if } (P, Q) = (3, 3),$$

$$\mathbf{\Gamma} = \begin{pmatrix} \frac{-1}{I\omega_1|C_1|} & \frac{-1}{I\omega_1|C_2|} & \frac{+1}{I\omega_1|C_3|} & \frac{+1}{I\omega_1|C_4|} \\ \frac{+1}{I\omega_2|C_1|} & \frac{+1}{I\omega_2|C_2|} & \frac{-1}{I\omega_2|C_3|} & \frac{-1}{I\omega_2|C_4|} \\ \frac{-1}{I\omega_3|C_1|} & \frac{+1}{I\omega_3|C_2|} & \frac{-1}{I\omega_3|C_3|} & \frac{+1}{I\omega_3|C_4|} \\ \frac{+1}{I\omega_4|C_1|} & \frac{-1}{I\omega_4|C_2|} & \frac{+1}{I\omega_4|C_3|} & \frac{-1}{I\omega_4|C_4|} \end{pmatrix}, \qquad \text{if } (P, Q) = (4, 4),$$

where $I\omega_p$ $(p = 1, \ldots, P)$ denotes the expected cluster cardinality of row cluster R_p $(p = 1, \ldots, P)$ and $|C_q|$ $(q = 1, \ldots, Q)$ denotes the cluster cardinality of column cluster C_q $(q = 1, \ldots, Q)$ as implied by the generated column partition matrix $\mathbf{K}$. Therefore, each of these $\mathbf{\Gamma}$s guarantees that the constraints are met, that is

$$\sum_{p=1}^{P} \omega_p \gamma_{pq} = 0,$$

for $q = 1, \ldots, Q$, and

$$\sum_{q=1}^{Q} |C_q| \gamma_{pq} = 0,$$

for $p = 1, \ldots, P$.

Subsequently, an $I \times J$ matrix $\mathbf{T}$, with elements t_{ij} $(i = 1, \ldots, I,\ j = 1, \ldots, J)$, of true interaction effects is obtained as $\mathbf{T} = \mathbf{z}\mathbf{\Gamma}\mathbf{K}^T$ and each element d_{ij} of the observed data matrix $\mathbf{D}$ is generated as $D_{ij} \sim \mathcal{N}(t_{ij}, \sigma_\epsilon^2)$. The value of σ_ϵ^2 is chosen to obtain a specific effect size η^2 that is defined as the ratio of interaction variance to the sum of interaction variance and error variance. Specifically,

$$\eta^2 = \frac{||\mathbf{T}||^2}{||\mathbf{T}||^2 + IJ\sigma_\epsilon^2} = \frac{\sum_{p=1}^{P} \sum_{q=1}^{Q} |R_p||C_q|\gamma_{pq}^2}{\left(\sum_{p=1}^{P} \sum_{q=1}^{Q} |R_p||C_q|\gamma_{pq}^2\right) + IJ\sigma_\epsilon^2}, \tag{14}$$

which was set to $\eta^2 = 0.10$.

For each combination of size, equality of expected row cluster sizes and true number of clusters we generated 1000 simulated data sets. Each data set $\mathbf{D}$ was then analyzed three times using the newly proposed E-ReMI method, each of which setting the number of clusters for the analysis equal to a specific value of (P^*, Q^*) (for details, see Figs. 1–6). This yielded for each data set and each value of (P^*, Q^*) an observed value of the test statistic $\lambda_{LR}^{(\text{obs})}$, which, as in the second simulation study, was compared to the corresponding critical value $\lambda_{LR}^{(\alpha)}$ for that combination of size and number of clusters (P^*, Q^*). The proportion of observed values (out of all 1000 data sets for that combination) that fall in the rejection region corresponds to the empirical power of the E-ReMI test of interaction.

Furthermore, to study cluster recovery performance of E-ReMI, we examined the extent to which the optimal partitions, as obtained from the data analyses, resemble the true partitions underlying the data. Specifically, we measured the agreement between the true underlying partition of the set of rows ($\mathbf{z}$) and the estimated row partition ($\hat{z}$) making use of the adjusted Rand index (ARI), see (Hubert & Arabie (1985)). This index is 1 if the two partitions are identical and 0 if the two partitions do not correspond more than expected by chance, and its minimal value can be smaller than 0 (Chacón & Rastrojo, 2022). Furthermore, the index is insensitive to permutations of the cluster labels. We measured the agreement between the true and estimated column partitions (i.e., $\mathbf{K}$ and $\hat{\mathbf{K}}$) in the same way. Finally, we examined the extent to which the true interaction parameters are recovered by computing the normalized squared error (NSE):

$$NSE = \frac{||\mathbf{T} - \hat{\mathbf{T}}||^2}{||\mathbf{T}||^2} = \frac{||\mathbf{T} - \hat{z}\hat{\mathbf{\Gamma}}\hat{\mathbf{K}}^T||^2}{||\mathbf{T}||^2}.$$

Note that NSE is a decreasing function of both cluster recovery and the extent to which the estimated interaction effect parameters $\hat{\gamma}_{pq}$ resemble the true interaction parameters γ_{pq}. Larger values of NSE indicate stronger disagreement.

For the purpose of method comparison, each data set was also fitted by a model that assumes equal expected row cluster sizes. This was achieved by using the constrained criterion $(CC)^*$, as explained in Section 2.3. From a modeling perspective, this is equivalent to fitting the REMAXINT model discussed in Ahmed et al. (2021). In order to test the hypothesis of no interaction based on this model, we used the constrained test statistic λ^*_{LR}, as explained in Section 3.1. Furthermore, we tested each simulated data set for the presence of interaction by the methods developed in Boik (1993); Malik et al. (2016); Piepho (1994), using the released R-package *CombinIT* (Shenaravi & Kharrati-Kopaei, 2018). This package includes various tests for the presence of interaction in two-mode data. We selected those that are (computationally) able to handle the data sizes included in this simulation study. *Boik* is based on a test statistic that involves the singular values of the (scaled) non-additivity matrix (i.e., the matrix of least squares residuals from fitting a two-way additive model to $\mathbf{D}$). *Malik* partitions that set of residuals (non-additivity values) into three clusters and is based on the idea that in the case of interaction one may expect to see at least one cluster of residuals that are positive, one cluster of residuals that are negative, and possibly one cluster with residuals close to zero. The test statistic is an F-like test statistic for a two-way model that includes row and column main effects and a cluster effect. *Piepho* tests for interaction by checking equality of variances between rows. This is based on the fact that if there is no interaction, then the expected row variances are all equal. For further details the reader is referred to the reference manual of *combinIT* on the CRAN website and Shenaravi & Kharrati-Kopaei (2018).

We study empirical power for E-REMI, REMAXINT, Boik, Malik and Piepho. For the estimated solutions obtained by REMAXINT, we also computed the three recovery measures (i.e., ARI for rows, ARI for columns and NSE) and compared these results to those obtained by E-ReMI. A similar comparison between E-ReMI and Boik, Malik and Piepho, respectively, is not possible because these methods are not based on an underlying two-mode clustering model and *CombinIT* only focuses on hypothesis testing for interaction.

Table 2 Critical values $\lambda_{LR}^{(\alpha)}$ for $\alpha = 0.05$ for each combination of size and number of clusters

Size	Number of clusters ($\mathbf{P^*}$, $\mathbf{Q^*}$)				
	(2, 2)	(3, 2)	(4, 2)	(3, 3)	(4, 4)
20 × 20	15.029	14.542	12.844	25.043	34.636
40 × 20	13.875	12.932	10.476	22.349	29.991
50 × 30	18.562	17.975	15.288	28.986	39.144
30 × 50	29.511	29.213	27.725	47.065	62.392
100 × 20	13.101	11.655	8.157	20.167	27.509
200 × 30	17.206	15.738	11.909	26.005	35.052

4.2 Results

4.2.1 Critical Values

Table 2 shows critical values $\lambda_{LR}^{(\alpha)}$ for each combination of size of the data and number of clusters, at a nominal significance level α= 0.05. These critical values were obtained by applying the three steps Monte Carlo scheme described in Section 3.2.

Inspection of Table 2 shows that, for each level of size, the null distribution of λ_{LR} is shifted towards the right for increasing number of column clusters. This is an expected result because more parameters are estimated from the observed data and, thus, more chance capitalization resulting in higher λ_{LR} values by chance. In contrast, increasing the number of row clusters appears to result in null distributions that are shifted towards the left, despite implying a larger number of estimated parameters. This shift must be attributed to the penalty term in (12), which is affected by the number of estimated row clusters. Furthermore, if the number of columns J is fixed, then for each level of number of clusters, the null distribution of λ_{LR} is shifted to the left as the number of rows I increases. This is due to the penalty term decreasing faster, as I increases, than the difference in log-squared-residuals (see (12)), leading to smaller test statistic values. Finally, comparing 50 × 30 to 20 × 20 and 40 × 20 shows that increasing the number of columns J shifts the null distribution of λ_{LR} towards the right, despite a larger number of rows I, which we have seen shifts the null distribution to the left. This may be explained by the fact that J does not affect the penalty term in (12) but only the difference in log-squared-residuals.

4.2.2 Type-I Error Rate

Table 3 shows the proportion of significant test results for all combinations of size and number of clusters. Inspection of this table reveals that the proportion of significant test results for each design is close to the nominal significance level $\alpha = 0.05$. Specifically, only one of the empirical Type-I error rates is outside the interval [0.0403; 0.0597], which is the interval centered at the nominal significance level (i.e., 0.05) and with radius equal to the (normal approximation of the) margin of error for a population proportion of 0.05, with $L = 5000$ trials, and corrected for the number of tests, i.e. cells in the table, using Bonferroni for familywise error rate of 5%. The accuracy of the empirical Type-I error rates suggests that generating $L = 5000$ Monte Carlo data sets using the three steps Monte Carlo scheme described in Section 3.2 is a reasonable choice for approximating the null distribution of λ_{LR}.

Table 3 Empirical Type-I error rate for nominal $\alpha = 0.05$ as a function of size of the data and number of clusters

Size	Number of clusters $(\mathbf{P^*}, \mathbf{Q^*})$				
	(2, 2)	(3, 2)	(4, 2)	(3, 3)	(4, 4)
20 × 20	.0424	.0564	.0562	.0596	.0520
40 × 20	.0536	.0502	.0420	.0464	.0566
50 × 30	.0550	.0440	.0442	.0526	.0512
30 × 50	.0418	.0492	.0504	.0492	.0542
100 × 20	.0492	.0512	.0450	.0518	.0488
200 × 30	.0510	.0608	.0492	.0496	.0426

4.2.3 Power and Parameter Recovery

Power

Figure 1 shows empirical power as a function of method (x-axis) and size of the data (curves), for data generated with unequal expected row cluster sizes. Subfigures in each column refer to true number of clusters set to (2, 2), (3, 3) and (4, 4), for the first, second and third column, respectively. Subfigures in the first row are obtained when the number of clusters for the analysis coincides with the true number of clusters (i.e., no misspecification), while the second and third row correspond to misspecification of the number of clusters for the analysis (see subfigure headings for details). Note that the methods *Boik*, *Piepho* and *Malik* are not based on a two-mode clustering model, and hence their power depends only on the generated data and not on the number of clusters for the analysis. The results of these methods therefore do not change across the different rows of Fig. 1, but only across columns. Overall, empirical power decreases if the number of observations (i.e., I and/or J) decreases (comparison between curves within a subfigure) and/or the true number of clusters (P, Q) increases (comparison across subfigures of the first row). There is a decrease, but it is not dramatic, in power for E-ReMI and REMAXINT when the number of clusters for the analysis does not match the true number of clusters (comparison across rows of the subfigures within each column). This suggests that, when using E-ReMI or REMAXINT as a test for interaction a small under/over-fitting does not have serious consequences for power (see Section 6 for a discussion on setting the number of clusters for the analysis). Comparing the different methods, it stands out that Malik performs the worst, followed by Piepho, which shows the second worst performance in terms of power. Comparing the remaining three methods when the true number of clusters is set to (2, 2), REMAXINT seems to perform slightly better than E-ReMI, which, in turn, tends to perform better than Boik. Instead, when the true number of clusters is set to (3, 3) E-ReMI overall has the best performance, followed by REMAXINT and then Boik. Given the choice of data generation mechanism (with one row cluster comprising 70% of cases in expectation), this is not surprising as increasing the number of row clusters leads to a higher level of inequality of the expected row cluster sizes. Lastly, Boik becomes the best method, followed by E-ReMI, once the true number of clusters increases to (4, 4).

Figure 2 is similar to Fig. 1 and shows empirical power as a function of method (x-axis) and size of the data (curves), but for data generated with equal expected row cluster sizes (same subfigure structure). Similarly as in the previous scenario, empirical power decreases if the number of observations decreases (comparison between curves within a subfigure) and/or the true number of clusters (P, Q) increases (comparison across subfigures of the first row). However in this case, it seems that increasing the number of columns has a stronger effect than increasing the number of rows, as, when size is set to 30 × 50 results are equal or better than when size is set to 50 × 30. A possible explanation is that since E-ReMI is too flexible, it requires more columns so as not to overfit the data with respect to the row clusters. REMAXINT, on the other hand, correctly assumes equal row cluster sizes and thus is less affected by the number of columns. As in the unequal row cluster sizes case, there is a small decrease in power for E-ReMI and REMAXINT when the number of clusters for the analysis does not match the true number of clusters (comparison across rows of the subfigures within each column), with E-ReMI being less robust to this misspecification. Comparing the different methods, Malik and Piepho have clearly the worst performance. Comparing the remaining three methods reveals that when the true number of clusters is set to (2, 2) and to (3, 3) REMAXINT seems to perform better than E-ReMI and Boik, which have a similar performance. When increasing the true number of clusters to (4, 4) Boik is clearly the best performing method in terms of power.

Parameter recovery

Figures 3–6 present the results in terms of means, across all 1000 data sets per condition, of ARI for rows, ARI for columns and NSE. Similarly to the figures for power, different columns refer to different true number of clusters, while different rows to different number of clusters for the analysis (see subfigure headings for details). For studying parameter recovery, the comparison is possible only between E-ReMI and REMAXINT, as they are the only two methods that yield estimated row and column partitions with corresponding bicluster interaction effect parameters.

Figure 3 presents the means across all 1000 data sets per condition of ARI for row clusters as a function of method (x-axis) and size of the data (curves), for data generated with unequal expected row cluster sizes. Overall, mean ARI is higher (i.e., better performance) for larger data sizes. Specifically, it is the highest when size of the data is set to 200 × 30, followed by 30 × 50 and then by 50 × 30. Moreover, for fixed number of columns and increasing number of rows (i.e., comparing 100 × 20, 40 × 20 and 20 × 20), the performance increases as the number of rows increases. Lastly, comparing the cases 100 × 20, 50 × 30 and 30 × 50, the latter has the overall best performance, while 100 × 20 has the worst, despite having a larger number of total observations (i.e. 2000 as compared to 1500). This is as expected, since a larger number of columns implies more information to estimate the row clusters. Subfigures in the top row show the results when there is no missspecification of the number of clusters for the analysis, that is, when they coincide with the true number of clusters. Focusing on these figures, it can be seen that the two methods perform equally well in the scenario with true number of clusters set to (2, 2), while E-ReMI performs better in the (3, 3) and (4, 4) cases. In case of misspecification of the number of clusters for the analysis, E-ReMI performs always better than REMAXINT. This result is particularly interesting in the (2, 2) case, as the performance between the two methods was similar under correct specification of the number of clusters. This increased comparative performance can be explained by the fact that the misspecified models in this case imply an overfitting of the number of row clusters. It is likely that E-ReMI is capable of classifying correctly most of the observations, by creating additional small clusters for observations that (randomly) differentiate from the two main clusters. REMAXINT, on the other hand, because of the implicit assumption of equal row cluster sizes, is encouraged more strongly to yield surplus row clusters containing a substantial number of rows.

Figure 4 presents the means of ARI for row clusters, for data generated with equal expected row cluster sizes. Overall, we see a slightly better performance of REMAXINT in all scenarios but those where the true number of clusters is equal to (2, 2) and the number of clusters for the analysis are misspecified, where E-ReMI has clearly a better performance. These results can again be explained by the flexibility of E-ReMI with respect to yielding row clusters with unequal row cluster sizes that may (partially) make up for the fact that an excess number of row clusters is fitted to these data.

Figure 5 presents the means of ARI for column clusters, for data generated with unequal expected row cluster sizes. Overall, mean ARI is higher (i.e., better performance) for larger data sizes. Specifically, it is the highest when size of the data is set to 200 × 30, followed by 100 × 20 and then by 50 × 30. As opposed to ARI for rows, increasing the number of rows has a stronger effect on ARI for columns than increasing the number of columns. This can be seen by 50 × 30 performing better than 30 × 50 (it was the opposite in ARI for rows) and by 100 × 20 performing better than those two (which was not the case for ARI for rows). This is as expected, since a larger number of rows imply more information to estimate the column clusters. The performance of the methods decreases for higher values of true number of clusters (comparison across columns), and it is negatively affected by misspecification of the number of clusters for the analysis when true number of clusters is set to (3, 3) and (4, 4) (comparison across rows of the middle and rightmost columns). Note that when true number of clusters is set to (2, 2), both misspecifications imply overfitting in terms of the number of row clusters, whereas when true number of clusters is set to (4, 4) both misspecifications imply underfitting in terms of the number of column clusters. When true number of clusters is set to (2, 2) the two methods perform equally well, while when it is set to (3, 3) or (4, 4), the two methods perform equally well in most cases, except for data sizes with a large number of rows in which E-ReMI has a better performance. This suggests that estimation of the column partitions benefits substantially from a correct specification of the model if that sample is sufficiently large (i.e., 100 rows or more). Results when data are generated with equal row cluster sizes are very similar, but now the two methods perform always equally well (results not shown). Note that in this case, the sampling mechanism for the rows as assumed by REMAXINT is correct.

Figure 6 presents the means of NSE, for data generated with unequal expected row cluster sizes. The outcome measure NSE is the most general parameter recovery performance measure out of the three considered in this study, since it takes into account the quality of the estimated row clustering, of the estimated column clustering and of the estimated interaction effect parameters. Bearing in mind that lower values of NSE imply better performance, mean NSE is the lowest (and thus the best) for size set to 200 × 30, i.e. the largest data size, and is very similar for sizes set to 100 × 20, 50 × 30 and 30 × 50. Since 100 × 20 has a larger data size than the other two cases, but it is also the most asymmetrical case, this suggests that more symmetrical data sets tend to perform better in terms of NSE. Also in this case, increasing the true number of clusters has a detrimental effect on the performance of the methods under study. Interestingly, misspecification of the number of clusters for the analysis does not lead to a clear trend with results that are sometimes not affected, sometimes positively affected (better performance of the methods) and sometimes negatively affected (worse performance of the methods). Lastly, E-ReMI tends to perform at least as well as REMAXINT, and, very often, better. Results when data are generated with equal row cluster sizes are very similar, but the two methods perform always equally well (results not shown).

Summarizing the results in this subsection, increasing size of the data leads to a better performance, with some performance criteria more affected by an increase in the number of rows and some others by an increase in the number of columns. Increasing the complexity of the clustering structure, that is, when true number of clusters increases, leads to a worse performance. Misspecification of the number of clusters for the analysis generally has a detrimental effect on the performance of the methods, but for most performance criteria (in most scenarios) this effect is minimal. As could be expected, in terms of power, E-ReMI tends to perform better than REMAXINT when data are generated with unequal expected row cluster sizes whereas, in a few cases, the performance of REMAXINT is better than that of E-ReMI when data are generated with equal expected row cluster sizes. Overall, in terms of parameter recovery, E-ReMI performs better than REMAXINT when data are generated with unequal expected row cluster sizes and both methods perform equally well when data are generated with equal expected row cluster sizes. Lastly, when it comes to power, Boik's method is a good choice as its performance is always very good and it is more robust to increased complexity of the data. However, this method was not designed to facilitate interpreting the row by column interaction in a data set at hand. When interest is in understanding that interaction, REMAXINT/E-ReMI are recommended because they yield an estimated two-mode clustering structure and corresponding interaction effect parameter estimates.

5 Application to a Study of Altruism Behavior

In this section, we analyze data from a real case study on altruism in order to infer whether there is statistical evidence of interaction and to study what it looks like. The 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 & Shoda 1995, 1998).

The data in question were collected in a study by Quintiens (1999) and were more recently reanalyzed in Schepers & Van Mechelen (2011), Schepers et al. (2017) and Ahmed et al. (2021). A group of $I = 102$ participants was presented with a set of $J = 16$ hypothetical situations, each describing an emergency situation in which a victim could possibly be helped. Each participant was asked to indicate, for each situation, 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).

In order to infer whether there is evidence of interaction between person and situation clusters, E-ReMI and REMAXINT interaction tests were both applied to the 102×16 data matrix of help ratings. For REMAXINT, this implied analyzing the data set at hand by maximizing constrained criterion $(CC)^*$, see Section 2.3, and testing the hypothesis based on the constrained test statistic λ^*_{LR}, see Section 3.1. For both methods, we analyzed the data assuming $(P, Q) = (3, 3)$, since Ahmed et al. (2021) suggested this choice for REMAXINT, using a post-hoc analysis. In order to obtain critical values for λ_{LR} and λ^*_{LR}, we employed the procedure described in Section 3.2 for a significance level $\alpha = 0.05$, and using Algorithm 1. Each analysis yields an observed value $\lambda^{(\text{obs})}_{LR}$ (for E-ReMI) and $\lambda^{*(\text{obs})}_{LR}$ (for REMAXINT) that was based on $M = 500$ random starts to reduce the possibility of finding a locally optimal solution. For each test, p-values were computed as $P(\lambda^{(l)}_{LR} > \lambda^{(\text{obs})}_{LR})$ (for E-ReMI) and $P(\lambda^{*(l)}_{LR} > \lambda^{*(\text{obs})}_{LR})$ (for REMAXINT). The results are shown in Table 4.

Table 4 Test results of E-ReMI and REMAXINT applied to help data assuming $(P, Q) = (3, 3)$

	Method	
	E-ReMI	REMAXINT
Critical value for $\alpha = 0.05$	19.732	−9.22
Test statistic	40.292	18.40
P-value	0.0000	0.0000

For both methods, the observed value of the test statistic falls in the rejection region. In fact, for both tests, the observed test statistic value is larger than any of the simulated values obtained under the null hypothesis, implying empirical p-values that are equal to zero. Thus, based on the E-ReMI and REMAXINT interaction tests assuming $(P, Q) = (3, 3)$, there is strong empirical evidence to conclude that there is an interaction between person cluster and situation cluster on willingness to help.

We now turn to studying what the gist of the interaction in these data looks like. For this purpose, Fig. 7 shows bicluster means associated to the estimated person and situation partitions yielded by E-ReMI and REMAXINT (middle and right panel, respectively), assuming $P = 3$ person clusters and $Q = 3$ situation clusters. Furthermore, this figure also shows the bicluster means associated to the estimated person and situation partitions yielded by a Gaussian latent block mixture model (Govaert & Nadif, 2003) assuming the same number of person and situation clusters (left panel). The block mixture model parameter estimates can be seen as a summary of the data. Hence, if there is substantial interaction in some data set, one may expect to see it exhibited in a suitable summary of those data. The latent block mixture model was estimated using the R package *blockcluster* (Bhatia et al., 2017) allowing unequal person and situation cluster sizes and assuming homogeneous residual variance across biclusters.

Compared to the latent block mixture model, REMAXINT and E-ReMI yield person and situation clusters that represent a stronger degree of person cluster by situation cluster interaction. This difference implies a substantially different type of conclusion. Notably, for the latent block mixture model, a ranking of the person clusters in terms of average willingness to help is consistent across all three situation clusters. This implies that members of one person cluster are on average more willing to help than members of another person cluster, regardless of the situation. In contrast, for REMAXINT and E-ReMI that ranking of the person clusters depends on the situation cluster. Specifically, the latter two methods yield person clusters with members that, in some situations, are on average more willing to help than members of another person cluster, but not in other situations. This difference between, on the one hand, the solution yielded by the latent block mixture model and, on the other the hand, the solutions yielded by REMAXINT and E-ReMI is due to the presence of individual person and situation main effects in these data. The latent block mixture model

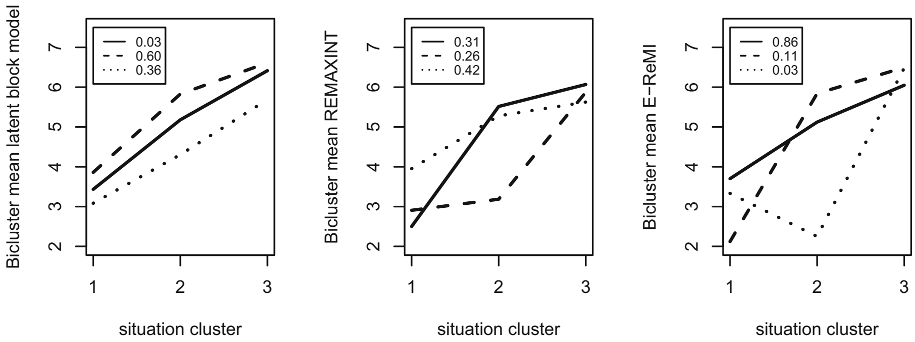

Fig. 7 Bicluster means for latent block mixture model (left panel), REMAXINT (middle panel) and E-ReMI (right panel) applied to altruism data. Situation clusters are represented on the x-axis. Person clusters are represented by the curves. The legends include, for each method, the estimated person cluster sizes. Since REMAXINT does not estimate population person cluster sizes, the person cluster sizes for this method were obtained by calculating the relative person cluster cardinalities of the estimated person clusters (i.e., $\frac{|R_p|}{102}$)

yields person and situation clusters that capture person by situation interaction as well as person and situation main effects. Remember that the latent block mixture model does not allow for row differences within a given row cluster, and likewise not for column differences within a given column cluster. Stated differently, it assumes stochastic independence between observations within the same row and column cluster. In contrast, REMAXINT and E-ReMI assume local independence given $\mathcal{R}$ and $\mathcal{C}$ and the individual person and situation main effects, implying that person and situation clusters as yielded by these methods are not affected by those main effects. It is important to note that this does not imply that the REMAXINT and E-ReMI person clusters (resp. situation clusters) are necessarily independent of the person (resp. situation) main effects in the data. For instance, it can be seen in Fig. 7 that the situation clusters yielded by REMAXINT differ in terms of average willingness to help. The same applies to the situation clusters yielded by E-ReMI. However, this association simply happens to be a characteristic of these data and may be (almost) absent in other applications.

Choosing between a latent block mixture approach or a maximal interaction clustering approach must be based on the type of question that one wishes to address with the chosen method. A latent block mixture approach will yield parameter estimates that are a good summary of the data, and will thus yield row and column partitions that also capture row and column main effects. A maximal interaction clustering approach is more useful if one wishes to describe the gist of the row by column interaction. As the nature of person by situation interaction is of central interest in contextualized personality psychology, and not so much the possible person and situation main effects, one may argue that a maximal interaction clustering approach is the more interesting choice for this application. Using a latent block mixture approach may, compared to a maximal interaction clustering approach, require the extraction of much more row and column clusters to capture the observed correlations between columns and rows, respectively. Furthermore, compared to E-ReMI, REMAXINT tends towards yielding person clusters that are more equal in size. According to E-ReMI, there is one large cluster and two much smaller ones whereas the relative cluster cardinalities of the row clusters yielded by REMAXINT are closer to each other. A likelihood ratio test suggests that E-ReMI provides a better fit to these data than REMAXINT ($-2(\ell(\boldsymbol{\xi}^0)-\ell(\boldsymbol{\xi})) = 43.784, df = 2, p = 0.000$). Note that the person cluster by situation cluster interaction captured by E-ReMI is structurally different from the one captured by REMAXINT. For REMAXINT, a ranking of the situation clusters in terms of average willingness to help is equal across all three person clusters. This implies that, consistently across person clusters, situations of one situation cluster elicit on average more willingness to help than situations of another situation cluster. In contrast, for E-ReMI this ranking of the situation clusters depends on the person cluster: Situations of situation cluster 2 elicit on average more willingness to help than situations of situation cluster 1 for members of two of the three person clusters, but the opposite holds for members of the other person cluster. In contextualized personality psychology, this difference in structure is of theoretical importance, as the consistent ranking of situation clusters as yielded by REMAXINT suggests the possible existence of a latent (one-dimensional) force that stems from the situations and that plays a key role in determining the behavior of all persons (Van Mechelen, 2009). In this application, an example of that force could be the level of compassion as induced by a situation. However, since E-ReMI yields a structure that does not show a consistent ranking of the situation clusters, and fits the altruism data better than REMAXINT, it appears that willingness to help cannot be explained by such a one-dimensional underlying situational force.

6 Discussion

In this paper, we presented E-ReMI, a method based on a probabilistic two-mode clustering model that yields two-mode partitions of the data with maximal interaction between row and column clusters. Specifically, this paper extends existing work (Ahmed et al., 2021), in several ways. First, in the specification of the model, we relaxed the assumption of equal cluster size on the random rows, thus allowing for unequal cluster size of the row clusters. Moreover, we introduced a new testing procedure for the null hypothesis of no interaction. This includes a test statistic, its properties and an algorithm to obtain the distribution of the test statistic under the null hypothesis. Finally, we developed software for all the methods presented in this paper (i.e., estimating E-ReMI and REMAXINT models and Monte Carlo sampling for hypothesis testing). In order to make these methods available to a large group of users, we implemented our codes in the free software R. In summary, using the proposed method, users will be able to test the presence of interaction between rows and columns. They will also obtain a two-mode partition of the data set based on maximal interaction, as well as estimates of the model parameters, i.e. interaction effects, row cluster weights and main effects.

We assessed the performance of the proposed method by means of a simulation study and a real-life application. We further compared the proposed method with other competing methods, wherever possible. Specifically, in the simulation studies, we studied the performance of E-ReMI in terms of Type-I error rate, power and parameter/partition recovery. Empirical Type-I error rates showed good performance, as their deviation from the nominal significance level was not more than what is expected by chance. In terms of power, as predictable, we observed lower power of the methods considered here when size of the data decreases or the number of row/column clusters increases. Misspecification of the number of clusters for the analysis generally has a detrimental effect on the performance of the methods, but for most performance criteria (in most scenarios) this effect is minimal. Boik's method has always a good performance and it is more robust to increased complexity of the data as compared to E-ReMI and REMAXINT, however it is not designed to facilitate interpreting the row by column interaction found. E-ReMI tends to perform better than REMAXINT under unequal expected row cluster sizes, whereas in a few cases REMAXINT performs better than E-ReMI under equal expected row cluster sizes. As for parameter/partition recovery E-ReMI tends to perform better than REMAXINT under unequal expected row cluster sizes and both methods perform equally well when data are generated with equal expected row cluster sizes. In an analysis of a real data set on altruism, we found strong empirical evidence for person by situation interaction based on REMAXINT and E-ReMI hypothesis testing. For these data, we further discussed the parameter estimates yielded by REMAXINT, E-ReMI and a Gaussian latent block mixture model and highlighted the different underlying structures that are uncovered by the methods, including their implications for important substantive research questions in contextualized personality psychology.

Several extensions of the work reported in this paper are possible. Firstly, a limitation of the current approach is the assumption of normality on the residual terms ϵ_{ij} of the probabilistic model of E-ReMI. It is not clear how the method behaves in case of a violation of this assumption, which may be likely in applications. Therefore, it would be interesting to perform a study on the robustness of E-ReMI to violations of the normality assumption. This study is beyond the scope of this paper and it is currently under investigation. Secondly, just like REMAXINT, the newly proposed method E-ReMI requires that the number of row and column clusters are fixed a priori. Ahmed et al. (2021) proposed a post analysis approach to select optimal values for these parameters, but other approaches are certainly possible. For example, this can be done developing a method where P and Q are parameters to be estimated (see e.g., (Miller & Harrison, 2018)). Thirdly, it would be interesting to generalize the model such that it allows to perform two-mode maximal interaction clustering for binary or count response data. This can be done using a logistic and Poisson regression framework, respectively, but may necessitate using a marginal likelihood approach with a distributional assumption on the random row effects α_i instead of a conditional likelihood approach. Finally, for some applications it may be more suitable to fit a model with a random effects assumption for the columns too (and assuming a latent random Q-partition of the column set). It is to be studied whether a conditional likelihood approach that conditions on sufficient statistics for the row main effects as well as sufficient statistics for the column main effects, is an appropriate choice as a method for estimating the model parameters of interest.

Data Availability R codes for generating simulated data as described in Section 4 are available in DataverseNL at the following url https://doi.org/10.34894/PWQHEC. We do not own the rights for the person by situation data used in Section 5.

Code Availability R codes for analyzing data as described in Sections 2 and 3 are available in DataverseNL at the following url https://doi.org/10.34894/PWQHEC.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A: Cross-product Term *U*

In this appendix, we show how to derive (5). The cross product term U in (4) can be rewritten as

$$\begin{aligned} U &= \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{q=1}^{Q}\sum_{j=1}^{J} z_{ip}k_{jq}A_iB_{ij} \\ &= \sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}A_i \sum_{q=1}^{Q}\sum_{j=1}^{J} k_{jq}(d_{ij} - \bar{d}_{i\cdot} - \beta_j - \gamma_{pq}) \\ &= \sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}A_i \sum_{q=1}^{Q} |C_q|(-\gamma_{pq}). \end{aligned}$$

The second line is obtained by replacing B_{ij} by its definition, i.e. $B_{ij} = d_{ij} - \bar{d}_{i\cdot} - \beta_j - \gamma_{pq}$, see Table 1, and by taking the elements that do not depend on j out of its summation. The third line is obtained by noting that $\sum_{q=1}^{Q}\sum_{j=1}^{J}k_{jq}(d_{ij}-\bar{d}_{i\cdot})=0$ and $\sum_{q=1}^{Q}\sum_{j=1}^{J}k_{jq}\beta_j=\sum_{j=1}^{J}\beta_j=0$, since each column j is assigned to one and only one cluster, and the summation is across all k_{jq}. Additionally, $\sum_{q=1}^{Q}\sum_{j=1}^{J}k_{jq}(-\gamma_{pq})=\sum_{q=1}^{Q}|C_q|(-\gamma_{pq})$, since γ_{pq} does not depend on j, and $|C_q|$ is the cardinality of column cluster C_q.

Appendix B: Conditional Likelihood

In this appendix, we show how to derive (9). The conditional classification log-likelihood in (8) can be rewritten as:

$$\begin{aligned}\ell(\boldsymbol{\xi}) &= \log W + \log V - \frac{1}{2\sigma^2}B\\ &= \log W - \frac{IJ}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}B\end{aligned}$$

By replacing V by its definition, i.e. $V=\exp\left(\frac{-IJ}{2}\log 2\pi\sigma^2\right)$, see Table 1. The optimization problem entails maximizing this function w.r.t. ω_p, β_j and γ_{pq}. This problem is equivalent to maximizing the conditional classification log-likelihood $\log CL$, obtained by replacing σ^2 in $\ell(\boldsymbol{\xi})$ with its likelihood maximizer $\frac{B}{IJ}$ obtained by partial differentiation of $\ell(\boldsymbol{\xi})$ w.r.t. σ^2. This leads to

$$\begin{aligned}\log CL &= \log W - \frac{IJ}{2}\log\frac{2\pi}{IJ}B - \frac{IJ}{2}\\ &= \log W - \frac{IJ}{2}\log(B) + H,\end{aligned}$$

where $H=\frac{IJ}{2}\left(\log\left(\frac{IJ}{2\pi}\right)-1\right)$ is an additive constant.

Appendix C: Maximum Likelihood Estimation

In this appendix we derive the estimates of the parameters ω_p, β_j and γ_{pq}, by maximizing the conditional classification log-likelihood in (9). We first rewrite the full expression as

$$\log CL = \left(\sum_{p=1}^{P}\sum_{i=1}^{I}z_{ip}\log\omega_p\right) - \frac{IJ}{2}\log\left(\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J}z_{ip}k_{jq}(d_{ij}-\bar{d}_{i\cdot}-\beta_j-\gamma_{pq})^2\right) + H,$$

where H is an additive constant. In order to obtain the estimate of ω_p, it is sufficient to note that only $\left(\sum_{p=1}^{P}\sum_{i=1}^{I}z_{ip}\log\omega_p\right)$ depends on ω_p, and that this is equivalent to finding the maximum likelihood estimate of a multinomial distribution. Thus $\hat{\omega}_p=\frac{|R_p|}{I}$, $p=1,\ldots,P$.

Next we apply the Lagrange multiplier method to find the estimate of β_j and γ_{pq}. Focusing on the terms that depend on β_j and γ_{pq} the Lagrangian function is written as

$$\mathcal{L} = -\frac{IJ}{2}\log\left(\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}(d_{ij}-\bar{d}_{i\cdot}-\beta_j-\gamma_{pq})^2\right)$$
$$-\lambda_\beta\sum_{j=1}^{J}\beta_j-\lambda_{\gamma_q}\sum_{p=1}^{P}|R_p|\gamma_{pq}-\lambda_{\gamma_p}\sum_{q=1}^{Q}|C_q|\gamma_{pq}, \tag{15}$$

where λ_β, λ_{γ_q} and λ_{γ_p} are the Lagrange multipliers associated to the equality constraints on β_j and γ_{pq}.

Focusing on finding the maximum of the β_j $(j = 1, \ldots, J)$ parameters, the first derivative is:

$$\frac{\partial\mathcal{L}}{\partial\beta_j} = \frac{1}{\sigma^2}\left(\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}\left(d_{ij}-\bar{d}_{i\cdot}-\beta_j-\gamma_{pq}\right)\right)-\lambda_\beta.$$

Equating this equation to 0, and solving for β_j leads to the following result,

$$\begin{aligned}
&\frac{\partial\mathcal{L}}{\partial\beta_j} = \frac{1}{\sigma^2}\left(\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}\left(d_{ij}-\bar{d}_{i\cdot}-\beta_j-\gamma_{pq}\right)\right)-\lambda_\beta = 0\\
&\iff \sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}d_{ij}-\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}\bar{d}_{i\cdot}-\beta_j\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}-\sum_{p=1}^{P}\gamma_{pq}\sum_{i=1}^{I} z_{ip}-\sigma^2\lambda_\beta = 0\\
&\iff \beta_j = \bar{d}_{\cdot j}-\bar{d}_{\cdot\cdot}-\frac{1}{I}\sum_{p=1}^{P}|R_p|\gamma_{pq}-\frac{1}{I}\lambda_\beta\sigma^2,
\end{aligned}$$

where the second line is obtained by applying the sum operators to each term within the brackets and multiplying each side of the equation by σ^2. The third line is obtained by noting that $\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}d_{ij} = I\bar{d}_{\cdot j}$, $\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip}\bar{d}_{i\cdot} = I\bar{d}_{\cdot\cdot}$, $\sum_{p=1}^{P}\sum_{i=1}^{I} z_{ip} = I$, $\sum_{i=1}^{I} z_{ip} = |R_p|$ (for a specific p), and by solving for β_j.

In order to find the Lagrange multiplier λ_β, the next step is to apply a sum operator over j to both sides of the equations leading to

$$\begin{aligned}
\sum_{j=1}^{J}\beta_j &= \sum_{j=1}^{J}\bar{d}_{\cdot j}-\sum_{j=1}^{J}\bar{d}_{\cdot\cdot}-\frac{1}{I}\sum_{j=1}^{J}\sum_{p=1}^{P}|R_p|\gamma_{pq}-\frac{1}{I}\sum_{j=1}^{J}\lambda_\beta\sigma^2\\
\iff \frac{J}{I}\lambda_\beta\sigma^2 &= -\frac{J}{I}\sum_{p=1}^{P}|R_p|\gamma_{pq}\\
\iff \lambda_\beta &= -\frac{1}{\sigma^2}\sum_{p=1}^{P}|R_p|\gamma_{pq},
\end{aligned}$$

with the second line resulting from $\sum_{j=1}^{J}\beta_j = 0$ (constraint) and $\sum_{j=1}^{J}\bar{d}_{\cdot j}-\sum_{j=1}^{J}\bar{d}_{\cdot\cdot} = 0$, and the final result obtained by solving for λ_β. This result, together with the equation obtained for β_j, leads to

$$\hat{\beta}_j = \bar{d}_{.j} - \bar{d}_{..},$$

since, after replacing λ_β in the derivative, the last two terms cancel out.

The estimate of γ_{pq} can be obtained similarly. First, we compute the first derivative of (15) w.r.t. γ_{pq}

$$\frac{\partial \mathcal{L}}{\partial \gamma_{pq}} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}\left(d_{ij} - \bar{d}_{i\cdot} - \beta_j - \gamma_{pq}\right)\right) - \lambda_{\gamma_q}|R_p| - \lambda_{\gamma_p}|C_q|.$$

Equating this equation to zero, and solving for γ_{pq} leads to the following result,

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \gamma_{pq}} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}\left(d_{ij} - \bar{d}_{i\cdot} - \bar{d}_{.j} + \bar{d}_{..} - \gamma_{pq}\right)\right) - \lambda_{\gamma_q}|R_p| - \lambda_{\gamma_p}|C_q| = 0\\
&\iff \gamma_{pq}|R_p||C_q| = \left(\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}\left(d_{ij} - \bar{d}_{i\cdot} - \bar{d}_{.j} + \bar{d}_{..}\right)\right) - \lambda_{\gamma_q}|R_p|\sigma^2 - \lambda_{\gamma_p}|C_q|\sigma^2,
\end{aligned}$$

where in the first line β_j is replaced by its estimate, and the second line is obtained by solving for γ_{pq} and noting that $\sum_{i=1}^{I}\sum_{j=1}^{J} z_{ip}k_{jq}\gamma_{pq} = \gamma_{pq}\sum_{i=1}^{I} z_{ip}\sum_{j=1}^{J} k_{jq}$ and $\gamma_{pq}\sum_{i=1}^{I} z_{ip}\sum_{j=1}^{J} k_{jq} = \gamma_{pq}|R_p||C_q|$. In order to find the Lagrange multiplier λ_{γ_q}, the next step is to apply a sum operator over p to both sides of the equations leading to

$$\begin{aligned}
&\sum_{p=1}^{P}\gamma_{pq}|R_p||C_q| = \left(\sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{p=1}^{P} z_{ip}k_{jq}\left(d_{ij} - \bar{d}_{i\cdot} - \bar{d}_{.j} + \bar{d}_{..}\right)\right) - \sum_{p=1}^{P}\lambda_{\gamma_q}|R_p|\sigma^2 - \sum_{p=1}^{P}\lambda_{\gamma_p}|C_q|\sigma^2\\
&\iff -I\lambda_{\gamma_q}\sigma^2 - |C_q|\sigma^2\sum_{p=1}^{P}\lambda_{\gamma_p} = 0\\
&\iff \lambda_{\gamma_q} = -\frac{|C_q|}{I}\sum_{p=1}^{P}\lambda_{\gamma_p},
\end{aligned}$$

with the second line resulting from $\sum_{p=1}^{P}\gamma_{pq}|R_p||C_q| = 0$ (constraint), $\sum_{i=1}^{I}\sum_{p=1}^{P} z_{ip}\left(d_{ij} - \bar{d}_{i\cdot}\right) = 0$, $\sum_{j=1}^{J} k_{jq}\left(\bar{d}_{..} - \bar{d}_{.j}\right) = 0$ and $\sum_{p=1}^{P}|R_p| = I$, and the final result obtained by solving for λ_{γ_q}. Similarly, one can obtain $\lambda_{\gamma_p} = -\frac{|R_p|}{J}\sum_{q=1}^{Q}\lambda_{\gamma_q}$. Putting these results on the Lagrange multipliers together with the equation obtained for γ_{pq}, leads to

$$\begin{aligned}
&\hat{\gamma}_{pq}|R_p||C_q| = \left(\sum_{j=1}^{J}\sum_{i=1}^{I} z_{ip}k_{jq}(dc)_{ij}\right) + \frac{|C_q||R_p|\sigma^2}{I}\sum_{p=1}^{P}\lambda_{\gamma_p} + \frac{|R_p||C_q|\sigma^2}{J}\sum_{q=1}^{Q}\lambda_{\gamma_q}\\
&\iff \hat{\gamma}_{pq}|R_p||C_q| = \left(\sum_{j=1}^{J}\sum_{i=1}^{I} z_{ip}k_{jq}(dc)_{ij}\right) + \frac{|R_p||C_q|\sigma^2}{IJ}\left(J\sum_{p=1}^{P}\lambda_{\gamma_p} + I\sum_{q=1}^{Q}\lambda_{\gamma_q}\right)\\
&\iff \hat{\gamma}_{pq} = \frac{1}{|R_p||C_q|}\sum_{j=1}^{J}\sum_{i=1}^{I} z_{ip}k_{jq}\left(d_{ij} - \bar{d}_{i\cdot} - \bar{d}_{.j} + \bar{d}_{..}\right),
\end{aligned}$$

The second line of the equation is obtained by collecting common terms. Last line is obtained by noting that $J\sum_{p=1}^{P}\lambda_{\gamma_p} + I\sum_{q=1}^{Q}\lambda_{\gamma_q} = 0$. This last result is derived as follows

$$\begin{aligned} J\sum_{p=1}^{P}\lambda_{\gamma_p} + I\sum_{q=1}^{Q}\lambda_{\gamma_q} &= -J\sum_{p=1}^{P}\frac{|R_p|}{J}\sum_{q=1}^{Q}\lambda_{\gamma_q} + I\sum_{q=1}^{Q}\lambda_{\gamma_q} \\ &= I\sum_{q=1}^{Q}\lambda_{\gamma_q} - I\sum_{q=1}^{Q}\lambda_{\gamma_q} = 0, \end{aligned}$$

with the first equality obtained by replacing λ_{γ_p} with $-\frac{|R_p|}{J}\sum_{q=1}^{Q}\lambda_{\gamma_q}$ and the second noting that $\sum_{p=1}^{P}|R_p| = I$.

References

Ahmed, Z., Cassese, A., van Breukelen, G., & Schepers, J. (2021). Remaxint: a two-mode clustering-based method for statistical inference on two-way interaction. *Advances in Data Analysis and Classification, 15*, 987–1013.

Alin, A., & Kurt, S. (2006). Testing non-additivity (interaction) in two-way anova tables with no replication. *Statistical Methods in Medical Research, 15*(1), 63–85.

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26*(1), 31–44.

Anderson, J. A., & Senthilselvan, A. (1980). Smooth estimates for the hazard function. *Journal of the Royal Statistical Society: Series B (Methodological), 42*(3), 322–327.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49*, 803–821.

Bhatia, P., Iovleff, S., & Govaert, G. (2017). blockcluster: An r package for model-based co-clustering. *Journal of Statistical Software, 76*, 1–24.

Bock, H.-H. (1996). Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis, 23*(1), 5–28.

Boik, R. J. (1993). Testing additivity in two-way classifications with no replications: the locally best invariant test. *Journal of Applied Statistics, 20*(1), 41–55.

Bryant, P. G. (1991). Large-sample results for optimization-based clustering methods. *Journal of Classification, 8*, 31–44.

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology, 31*, 607–649.

Celeux, G., & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis, 14*, 315–332.

Chacón, J., & Rastrojo, A. (2022). Minimum adjusted rand index for two clusterings of a given size. *Advances in Data Analysis and Classification*.

Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. In *Proc 8th International conference on Intelligent Systems for Molecular Biology* (pp. 93–103).

Chernick, M. R. (2011). *Bootstrap methods: A guide for practitioners and researchers.*, vol. 619 of Wiley Series in Probability and Statistics. John Wiley & Sons.

Cho, H., Dhillon, I. S., Guan, A., & Sra, S. (2004). Minimum sum-squared residue co-clustering of gene expression data. In *Proc. 4th SIAM International conference on Knowledge Discovery and Data Mining* (pp. 124–125).

Choudhary, P. K., & Nagaraja, H. N. (2017). *Measuring Agreement: Models, Methods, and Applications*, vol. 34 of Wiley Series in Probability and Statistics. John Wiley & Sons.

Corsten, L. C. A., & Denis, J. B. (1990). Structuring interaction in two-way tables by clustering. *Biometrics, 46*, 207–215.

Denis, J. B., & Gower, J. C. (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. Philadelphia, Pa: Society for Industrial and Applied Mathematics.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch Models: Foundations, Recent developments, and Applications*. New York: Springer-Verlag.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society serie 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, C. T., Nielsen, D. M., & Osborne, J. A. (2013). A method for detecting hidden additivity in two-factor unreplicated experiments. *Computational Statistics & Data Analysis, 67*, 95–104.

Gauch, H. G. (2006). Statistical analysis of yield trials by AMMI and GGE. *Crop Science, 46*, 1488–1500.

Geiser, C., Litson, K., Bishop, J., Keller, B. T., Burns, G. L., 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. *Psychological Methods, 20*, 165–192.

Govaert, G., & Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, *36*(2), 463–473. Biometrics.

Govaert, G., & Nadif, M. (2013). *Co-clustering: Models, Algorithms and Applications*. FOCUS Series. Chichester, UK: Wiley.

Govaert, G., & Nadif, M. (2018). Mutual information, phi-squared and model-based co-clustering for contingency tables. *Advances in Data Analysis and Classification, 12*, 455–488.

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

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 1*(1), 24–45.

Malik, W. A., Möhring, J., & Piepho, H. P. (2016). A clustering-based test for nonadditivity in an unreplicated two-way layout. *Communications in Statistics - Simulation and Computation, 45*(2), 660–670.

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

McLachlan, G. J., & Peel, D. (1997). On a resampling approach to choosing the number of components in normal mixture models. In L. Billard & N. Fisher (Eds.), *Computing Science and Statistics* (Vol. 28, pp. 260–266). Fairfax Station, Virgina: Interface Foundation of North America.

Miller, J. W., & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association, 113*(521), 340–356. PMID: 29983475.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246–268.

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology, 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. (1994). On tests for interaction in a nonreplicated two-way layout. *Australian and New Zeland Journal of Statistics, 36*(3), 363–369.

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

Post, J. B., & Bondell, H. D. (2013). Factor selection and structural identification in the interaction anova model. *Biometrics, 69*(1), 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.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis* (3rd ed.). Belmont, CA: Duxbury Press.

Schepers, J., Bock, H.-H., & Van Mechelen, I. (2017). Maximal interaction two-mode clustering. *Journal of Classification, 34*, 49–75.

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. *Psychological Methods, 16*, 361–371.

Scott, A. J., & Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics, 27*, 387–397.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.

Shenaravi, Z., & Kharrati-Kopaei, M. (2018). A method for testing additivity in unreplicated two-way layouts based on combining multiple interaction tests. *International Statistical Review, 86*, 469–487.

Shoda, Y., Wilson, N. L., Chen, J., Gilmore, A. K., & Smith, R. E. (2013). Cognitive-affective processing system analysis of intra-individual dynamics in collaborative therapeutic assessment: Translating basic theory and research into clinical applications. *Journal of Personality, 81*, 554–1568.

Shoda, Y., Wilson, N. L., Whitsett, D. D., Lee-Dussud, J., & Zayas, V. (2015). The person as a cognitive affective processing system: Quantitative idiography as an integral component of cumulative science. Personality processes and individal differencesIn M. Mikulincer & P. Shaver (Eds.), *APA Handbook of Personality and Social Psychology* (Vol. 4, pp. 491–513). Washington: American Psychological Association APA.

Symons, M. J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics, 37*(1), 35–43.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics, 5*(3), 232–242.

Van Mechelen, I. (2009). A royal road to understanding the mechanisms underlying person-in-context behavior. *Journal of Research in Personality, 43*, 179–186.

Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2004). Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research, 13*, 363–394.

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Verbeke, G., Spiessens, B., & Lesaffre, E. (2001). Conditional linear mixed models. *The American Statistician, 55*(1), 25–34.

Yu, X., Yu, G., Wang, J., & Domeniconi, C. (2021). Co-clustering ensembles based on multiple relevance measures. *IEEE Transactions on Knowledge and Data Engineering, 33*(04), 1389–1400.