Detecting the Guttman effect with the help of ordinal correspondence analysis in synchrotron X-ray diffraction data analysis

Claude Manté, Sophie S. Cornu, D. Borschneck, C. Mocuta, R. van den Bogaert

ORIGINAL RESEARCH ARTICLE

Detecting the Guttman effect with the help of Ordinal Correspondence Analysis in synchrotron X-ray diffraction data analysis.

C. Manté[a], S. Cornu[b], D. Borschneck[b], C. Mocuta[c] and R. van den Bogaert[b]

[a]Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU,IRD, MIO,UM 110, Campus de Luminy, Case 901, F13288 Marseille Cedex 09, France, (ORCID:0000 − 0002 − 7268 − 9789) ; [b]Aix Marseille Université, CNRS, IRD, INRA, Coll France, CEREGE, Aix-en-Provence, France ; [c]Synchrotron SOLEIL, L'Orme des Merisiers, Saint-Aubin, BP 48, Gif-sur-Yvette 91192, France.

ARTICLE HISTORY
Compiled September 18, 2020

ABSTRACT
We propose a method for detecting a Guttman effect in a complete disjunctive table **U** with Q questions. Since such an investigation is a nonsense when the Q variables are independent, we reuse a previous unpublished work about the chi-squared independence test for Burt's tables.

Then, we introduce a two-steps method consisting in plugging the first singular vector from a preliminary Correspondence Analysis (CA) of **U** as a score x into a subsequent singly-ordered Ordinal Correspondence Analysis (OCA) of **U**. OCA mainly consists in completing x by a sequence of orthogonal polynomials superseding the classical factors of CA. As a consequence, in presence of a pure Guttman effect, we should in principle have that the second singular vector coincide with the polynomial of degree 2, *etc.* The hybrid decomposition of the Pearson chi-squared statistics (resulting from OCA) used in association with permutation tests makes possible to reveal such relationships, *i.e.* the presence of a Guttman effect in the structure of **U**, and to determine its degree - with an accuracy depending on the signal to noise ratio.

The proposed method is successively tested on artificial data (more or less noisy), a well-known benchmark, and synchrotron X-ray diffraction data of soil samples.

KEYWORDS
Ordinal Correspondence Analysis; Detrended Correspondence Analysis; Randomization; Eigenvalues; Orthogonal polynomials; Synchrotron X-rays diffraction

1. Introduction

The Guttman effect (also named arch or horseshoe effect) is frequently met in displays resulting from Correspondence Analysis (CA) [14, 34], or other multivariate methods. The characteristic of this phenomenon is that the second and sometimes higher factors have a strong nonlinear relationship with the first factor. Theoretical models [14, 18] show that in this case the k^{th} factor is an orthogonal polynomial of degree k in the first factor.

CONTACT C. Manté. Email: claude.mante@gmail.com

This phenomenon is an avatar of the scalogram analysis introduced in Psychometry by L. Guttman (see Section 2), as Benzécri [14, 17] noticed. Such a structure is of paramount importance in Psychometry, because it corresponds to a latent (and desired) general factor (of intelligence, *etc.*), but it is considered as a nuisance in Ecology, where it often corresponds to some well-known structure (depth or temperature gradient, *etc.*). Thus, while psychometricians considered scalogram analysis as an important tool, Hill and Gauch [40] surprisingly claimed that "the arch effect is simply a mathematical artifact, corresponding to no real structure in the data" and tried to remove it.

Hill and Gauch [40] therefore proposed an heuristic "detrending-by-segment" algorithm, giving rise to the so-called "Detrended Correspondence Analysis" (DCA). However DCA is a controversial method [41], because of the frequent instability resulting from its *ad hoc* detrending procedure. More recently, Ter Braak [55] proposed an alternative "detrending-by-polynomials" method, which does not seem to work much better than the original DCA [42]. However, postulating the existence of such nonlinear relationships between principal axes before erasing them is far from being innocent: if such a structure is absent from the data, this can lead to artifacts (think to the Slutsky-Yule effect in time series analysis), or loss of information [18].

So, it seems that preliminary questions to answer to are: "Is there really a Guttman effect in the data? What is the order of this phenomenon (degree of the polynomial)?" In this paper, we tackle these questions by combining CA with Ordinal Correspondence Analysis (OCA) to produce tests and graphical tools designed for this purpose. The proposed method is successively tested on artificial data, a well-known benchmark (Chinese vases data from [24]), and a dataset synchrotron X-ray diffraction pattern obtained on soil features.

2. Scalogram Analysis

This term was coined in 1944 by Louis Guttman [35], as "a procedure for testing the hypothesis that a universe of qualitative data is a **scale** for a given population of people".

Definition 2.1. [35]. The universe of content is said to be scalable for the population if it is possible to rank the people from high to low in such a fashion that from a person's rank alone we can reproduce his response to each of the items in a simple fashion.

Guttman's method for elaborating such a scale consisted in ranking people thanks to weights assigned to the categories associated with each question. Adding up these weights, one obtains a score for each person, depending on her opinion (typically: favorable/unfavorable). In a (possible) second step, categories could be combined, and people be ranked again, giving rise to the scale and the sorted table of observations. In addition to the scale, which typically ranks people from unfavorable to favorable, Guttman defined the **intensity**, which codes the strength of opinions, and noticed that, plotting the intensity against the scale, one generally obtains a more or less parabolic curve [35, 36], which could be reasonably fitted by a polynomial of degree 2. Furthermore, analyzing the sorted table through Principal Components Analysis (PCA), Guttman observed that when the universe of content is scalable, the n^{th} component looks like a polynomial of degree n.

A bit later, Benzécri [14, 17] showed that scalograms can be easily built from a contingency table $\mathbf{T}$ by ranking its I rows and J columns along the first factor of the Correspondence Analysis of $\mathbf{T}$ (if the universe of content is scalable, of course). More precisely, he demonstrated [14, pp. 192-196.] that the factors $\left\{\varphi_k^J : k \geq 1\right\}$ associated with the questions, issued from CA of a perfect scalogram, converge towards the family of Legendre polynomial when the number of questions becomes infinite. In addition, Benzécri demonstrated that in the case of normal correspondences, the factors converge towards Hermite polynomials. For an illustration, see Benzécri and coll. [14, pp. 481-486]; see also [18].

3. Variants of Correspondence Analysis (CA)

3.1. Simple CA

Consider some frequency table $\mathbf{T}$ of size $I \times J$ and of grand total N, where I (resp. J) denotes the modalities of a single nominal variable (a question). Let's denote $\mathbf{P} := \mathbf{T}/N$ the associated probability table, $P_i := \sum_{j \leq J} P_{i,j}$ (resp. $P_j := \sum_{i \leq I} P_{i,j}$), and $\mathbf{P}_I := (P_1, \cdots, P_i, \cdots, P_I)$ (resp. $\mathbf{P}_J := (P_1, \cdots, P_j, \cdots, P_J)$) the marginal column (resp row) profiles. The aim of simple CA is to highlight the ways $\mathbf{P}$ differs from the $I \times J$ matrix $\mathbf{P}_I \otimes \mathbf{P}_J$ of general entry $P_i\, P_j$ (independence of the rows and columns). Practically, it consists in performing the Generalized Singular Value Decomposition [34] of the matrix Θ of general entry $\theta_{i,i} := \frac{P_{i,j}}{P_i\, P_j}$, giving rise to a system of singular values and singular vectors $\left(\lambda_m; \varphi_m^I, \varphi_m^J\right) : 0 \leq m \leq M^* := \min\left(I-1, J-1\right) - 1$, with the trivial factor $\left(\lambda_o; \varphi_0^I, \varphi_0^J\right) = \left(0, \mathbf{1}^I, \mathbf{1}^J\right)$. The singular vectors are centered and normed:

$$\forall\ (m,p) \in M^* \times M^*,\ \sum_{i \leq I} P_i\, \varphi_{m,i}^I\, \varphi_{p,i}^I = \sum_{j \leq J} P_j\, \varphi_{m,j}^J\, \varphi_{p,j}^J = \delta_m^p \tag{1}$$

were $\delta_m^p := \begin{cases} 1\, if\, m = p \\ 0\, if\, m \neq p \end{cases}$ is the usual Dirac symbol.
In addition, they fulfill

$$\forall\ (m,p) \in M^* \times M^*,\ \sum_{(i,j) \in I \times J} P_{i,j}\, \varphi_{m,i}^I\, \varphi_{p,j}^J = \lambda_m \delta_m^p. \tag{2}$$

One obtains this way a first decomposition of the Pearson chi-squared statistics X^2 along the singular vectors:

$$\frac{X^2}{n} = \sum_{m=1}^{M^*} \lambda_m^2. \tag{3}$$

Remark 1. Other kinds of tables (similarity measures, ratings, *etc*) can be submitted to CA [14]; they are not considered in this study, which focuses exclusively on true

contingency and indicator tables.

3.2. A natural extension: Multiple Correspondence Analysis (MCA)

The above contingency table $\mathbf{T}$ can be constructed from the complete disjunctive table $\mathbf{U} = (\mathbf{U}_1 \mid \mathbf{U}_2) \in N \times (I + J)$ with N rows and $(I + J)$ columns, assigning to the r^{th} individual the row r of $\mathbf{U}$ obtained by concatenation of the pair of logical vectors describing this individual. It is well-known [15, 16, 43] that the CA of $\mathbf{U}$ is equivalent (with slightly different eigenvalues and eigenvectors [5, Section 6.2]) to the CA of the Burt table $\mathbf{B_U} := \mathbf{U}^t\mathbf{U}$ or to the CA of $\mathbf{T}$, which is the sub-table $\mathbf{U}_1^t\mathbf{U}_2$ of $\mathbf{B_U}$. Consequently, classical CA can be straightforwardly generalized to the analysis of $Q > 2$ questions (MCA). MCA consists in analyzing the complete disjunctive table $\mathbf{U} = (\mathbf{U}_1 \mid \mathbf{U}_2 \mid \cdots \mid \mathbf{U}_Q) \in N \times \left|\mathbf{W}^{(Q)}\right|$ of width $\left|\mathbf{W}^{(Q)}\right| := \sum_{q=1}^{Q} w_q$, where w_q denote the number of modalities (width) of the q^{ith} question. We will denote $\mathbf{W}^{(Q)} = \mathbf{W}_1 \oplus \mathbf{W}_2 \oplus \cdots \mathbf{W}_Q$ (where $\oplus$ is the concatenation operation) the direct sum of all the possible answers to the Q questions.

Any Burt's table issued from an experience can be considered as a realization of a multinomial distribution, whose parameters consist in some probability matrix $\mathbf{P}_{\mathbf{W}^{(Q)},\mathbf{W}^{(Q)}}$ equipped with a special blocks structure exclusively depending on $(w_1, \cdots, w_Q)$.

Due to the special structure of $\mathbf{U}$, issues from the associated MCA have several characteristics [43]:

- the rank of the analysis is less than $M^{\mathbf{U}} := \left|\mathbf{W}^{(Q)}\right| - Q \leq M^*$
- the total variance is $\frac{\left|\mathbf{W}^{(Q)}\right|}{Q} - 1$
- the part of variance associated with the q^{th} question is $\frac{w_q - 1}{Q}$.

Remark 2. Lebart and Saporta [44] reported that the foundations of MCA were also laid by Guttman, in 1941!

4. Significance of eigenvalues in PCA, CA and MCA

4.1. Inference about eigenvalues in PCA, in connection with the bootstrap approach

Theoretically, the distribution of the sample covariance matrix of a random vector of size J obeying $\mathcal{N}(0, \Sigma)$ is known: it is a Wishart distribution, whose eigenvalues are also known [4]. When $\Sigma = \mathbb{I}_J$, the expression of these eigenvalues is less complicated [43] but it is still very complex and, above all, this case is quite unrealistic, with very little practical utility. Consequently, researchers used instead simulations, or resampling methods such as the bootstrap.

Practically, in the large sample case ($N \gg J$), the sample covariance $\widehat{\Sigma}$ can be considered as a good estimate of Σ and one can accept that $\lambda_k(\Sigma) \approx \lambda_k\left(\widehat{\Sigma}\right)$, but things change in the case of high dimension, when $\frac{J}{N} \underset{N\to\infty}{\longmapsto} \gamma > 0$. Then, if γ is not close to zero, the standard estimate $\lambda_1\left(\widehat{\Sigma}\right)$ of the first eigenvalue $\lambda_1(\Sigma)$ overestimates

it, and the bootstrap estimate of the bias $\lambda_1(\Sigma) - \lambda_1\left(\widehat{\Sigma}\right)$ is itself highly biased [31]! Indeed, according to El Karoui and Purdom [31], the bootstrap completely changes the geometry of the dataset by re-weighting the observations, giving rise to an important bias. In the same high-dimensional setting, Hendrikse et al. [38] considered an iterative bootstrap approach to diminish the bias $\left\|\lambda(\Sigma) - \lambda\left(\widehat{\Sigma}\right)\right\|$, but obtained better results with another method, also based on the Marcenko-Pastur theorem [39]. Consequently, bootstrap methods do not work well for obtaining good estimates of the eigenvalues in PCA, except when $N \gg J$. Since CA can be considered as a special case of PCA, the situation is similar for the eigenvalues of CA. For instance, studying two different textual datasets, Alvarez et al. [1, 2] found that bootstrap eigenvalues estimates were highly positively biased.

To sum up, the bootstrap is much better-suited for studying the stability of principal axes [1–3, 43] than for estimating eigenvalues or testing their significance. Consequently, we will give in Section 6 preference to randomization methods for testing the significance of eigenvalues.

4.2. Inference about eigenvalues in CA and MCA

4.2.1. The independence trace test

Let's remind first the relationships between the eigenvalues $\left\{\lambda_i : 1 \leq i \leq M^{\mathbf{U}}\right\}$ issued from the CA of the binary table $\mathbf{U} = (\mathbf{U}_1 \mid \mathbf{U}_2)$, those issued from the CA of $\mathbf{T} = \mathbf{U}_1^t\,\mathbf{U}_2$: $\left\{(2\lambda_i - 1)^2 : 1 \leq i \leq M^*\right\}$ and those issued from the Burt table $\mathbf{B_U} := \mathbf{U}^t\mathbf{U}$ [15, 16, 34, 43]: $\left\{\lambda_i^2 : 1 \leq i \leq M^{\mathbf{U}}\right\}$. Thus, the eigenvalues issued from the analysis of $\mathbf{U}$ or $\mathbf{B_U}$ can be obtained from those of $\mathbf{T}$ and all the eigenvectors too, up to simple symmetries [34, pp. 130-133]. In classical (binary) CA, the unique rigorous test (trace test) is based on the Pearson chi-squared statistics X^2 defined by (3) which is the trace of the operator associated with $\mathbf{B_U}$. Under the hypothesis of independence of the columns and rows of $\mathbf{T}$, X^2 asymptotically obeys $\chi^2\left((I-1)(J-1)\right)$ [14]. It is possible to build a similar trace test in the general case of Q questions, based on the set $\Lambda := 1 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\mathbf{M}^{\mathbf{U}}} \geq 0$ of non-trivial eigenvalues issued from the CA of $\mathbf{U} = (\mathbf{U}_1 \mid U_2 \mid \cdots \mid \mathbf{U}_Q)$. This has been done by the first author, in an unpublished document [48].

Notice first that any Burt's table $\mathbf{B_U}$ issued from an experience with N individuals can be considered as a realization of a multinomial distribution, whose parameters depend on some unknown probability matrix $\mathbf{P}_{\mathbf{W}^{(Q)},\mathbf{W}^{(Q)}}$, whose estimation $\hat{\mathbf{P}}_{\mathbf{W}^{(Q)},\mathbf{W}^{(Q)}}$ by empirical proportions is the maximum likelihood one. Consequently, under the classical hypothesis of independent sampling, the Pearson statistics

$$\sum_{i=1}^{|W^{(Q)}|}\sum_{j=1}^{|W^{(Q)}|}\frac{\left(B_{U_{i,j}} - N\,P_{i,j}\right)^2}{N\,P_{i,j}}$$

tightly associate with the trace of the CA of $\mathbf{B_U}$ should (naively) obey $\chi^2_{DF(Q,W^{(Q)})}$ (asymptotically), where $DF\left(Q, W^{(Q)}\right)$ denotes the number of free parameters of the considered space of Burt's tables. But, due to the special blocks structure of such tables, this is a bit more complicated.

Proposition 4.1. *[48] Under (**H**), the distribution of* $N\left(\sum_{m=1}^{M}\lambda_m^2-\frac{W-Q}{Q^2}\right)$ *obeys* $\chi^2\left(\Gamma_Q-(Q-1)W+\frac{Q(Q-1)}{2}\right)$, *where* $\Gamma_Q:=\sum_{1\leq q_1<q_2\leq Q}w_{q_1}w_{q_2}$.

Remark 3. The term $\frac{W-Q}{Q^2}$ stems from the contribution of all the diagonal block matrices of $\hat{P}$ to the trace of CA: it's a natural correction of the total inertia. Consider now the probability $\hat{P}$ associated with B_U, and two questions i and j. The test proposed in [48] was based on the following modified probability:

$$\widetilde{P}_{q_i\,q_j}:=\begin{cases}\hat{P}_{q_i\,q_j} & if\ i\neq j\\ \hat{P}_{q_i}\,\hat{P}_{q_j} & if\ i=j\end{cases}.$$

But Manté [48] didn't analyze $\widetilde{P}$, while Greenacre went further with *Joint Correspondence Analysis* [19, 23], analyzing only the off-diagonal part of $\hat{P}$ (or $\widetilde{P}$ as well).

Remark 4. One can find in the literature [10, 45] an apparently different formulation for the number of d.f. in the independence test: $\left(\left(-Q+\sum_{q\leq Q}w^q\right)^2-\sum_{q\leq Q}\left(w^q-1\right)^2\right)/2$; this value comes from a paper of Bekker and de Leeuw [11], and indeed matches with ours.

4.2.2. Confidence intervals

Benzécri [16] highlighted the typical value $\bar{\lambda}:=1/Q$, as the "average eigenvalue" issued from the MCA of $\mathbf{U}$. Consequently, he proposed to discard all the eigenvalues smaller than $\bar{\lambda}$ and to supersede the classical part of variance λ^2 apportioned to each eigenspace by

$$\rho(\lambda):=\left(\frac{Q}{(Q-1)}\left(\lambda-\bar{\lambda}\right)\right)^2. \tag{4}$$

More recently, Ben Hammou and Saporta [12], [13] reported that under (**H**), $\bar{\lambda}$ is the unique non-trivial eigenvalue issued from the **theoretical** MCA of $\mathbf{U}$, with multiplicity Q. They also showed that, under the same hypothesis, the dispersion of non-trivial eigenvalues around $\bar{\lambda}$ is given by

$$\mathfrak{S}^2:=\frac{1}{Q^2N\,M^{\mathbf{U}}}\sum_{i\neq j}(w_i-1)(w_j-1).$$

In addition, they showed that

$$\sqrt{N}\left(\frac{1}{Q}-\lambda_k\left(\widehat{\Sigma}\right)\right)\underset{N\to\infty}{\longrightarrow}\mathcal{N}(0,\mathfrak{S})$$

and reported that this convergence in distribution is very slow for largest and smallest eigenvalues. Thus the confidence interval $\left[\bar{\lambda}-2\,\mathfrak{S},\bar{\lambda}+2\,\mathfrak{S}\right]$ should contain about 95% of the eigenvalues in the case of pairwise independence of the questions (see Figures 2, 4, 6, 8 and 10).

5. Ordinal Correspondence Analysis (OCA)

Suppose now the categories I and/or J are ordered: this is not taken into account by CA. That is why Beh [6] developed Ordinal Correspondence Analysis (OCA), resulting in a decomposition of the Pearson chi-squared statistics X^2 different from (3). In [6], this author proposed two variants of OCA, corresponding to the cases of singly-ordered or doubly-ordered contingency tables. We will focus on the first one, which gives rise to the **hybrid decomposition** [8, 9] of X^2. Since both Beh's methods require to compute orthogonal polynomials, we will focus in the next section on this important topic.

5.1. Computation of orthogonal polynomials

Classically [25, 27], one consider the real line equipped with a nonnegative (absolutely continuous, discrete, or a mixture of these categories) measure λ, such that all its moments μ_k defined by

$$\mu_k := \int_{\mathbb{R}} t^k \, d\lambda(t)$$

exist. Then, there exist an Hilbertian basis $\{b_0(t), b_1(t), \cdots, b_k(t), \cdots\}$ of $\mathbb{L}^2_\lambda$ consisting of orthogonal polynomials. That is how classical orthogonal polynomials (Legendre, Chebyshev, Laguerre, Hermite, Krawtchouk, *etc*) are defined, with respect to various measures.

Despite of its apparent simplicity, the construction of such bases is delicate. One could try to use the Gram-Schmidt method, but it is lengthy and the orthogonality between the $b_k(t)$ rapidly deteriorates [32]. Consequently, for computing discrete orthogonal polynomials, Emerson [32] used the Christoffel-Darboux method (named Stieltjes procedure by Gautschi [25]). Contrary to the Gram-Schmidt one, it only applies to polynomials; it is based on the following three-term recurrence relation (in the notations of Beh [7]):

$$b_k(j) = S_k \left((s(j) - T_k) \, b_{k-1}(j) - V_k \, b_{k-2}(j) \right) \tag{5}$$

where $s(j)$ is the value of the score associated with the j^{th} modality of the ordinal variable. Notice that, in the Emerson's terminology, $s(j)$ is the j^{th} sampled abscissa x_j. The coefficients in (5) are given by

$$\begin{cases} T_k := \sum_{j=1}^{J} \lambda_j \, s(j) \, b^2_{k-1}(j) \\ V_k := \sum_{j=1}^{J} \lambda_j \, s(j) \, b_{k-1}(j) \, b_{k-2}(j) \\ S_{k:} := \sqrt{-T_k^2 - V_k^2 + \sum_{j=1}^{J} \lambda_j \, s^2(j) \, b^2_{k-1}(j)} \end{cases} \tag{6}$$

where $\lambda_j = P_j$ is the weight assigned to the position $x_j := s(j)$ in the interval $[s(1), s(J)]$. Clearly, one can infer from formulas (5,6) that Emerson's polynomials are totally data dependent. Since any linear transformation of the score does not change the values of the orthogonal polynomials [7], we can indeed suppose that the common support of all scores is $[s(1), s(J)]$ (fixed). Then, **changing of score merely changes**

the position of the abscissas "sampled" in this interval. Notice now that formulas (6) are discrete approximations of integrals corresponding to true moments, and that these formulas can be seen (roughly speaking) as quadrature rules [27] for computing the coefficients involved in (5). This fact explains that "when comparing different types of scores, most of them will give similar results" [7]. Furthermore, if the discretization associated with the pairs $\{(\lambda_1, s(1)), \cdots, (\lambda_j, s(j)), \cdots, (\lambda_J, s(J))\}$ has been ill-designed, the construction of $\{b_0(t), b_1(t), \cdots, b_k(t), \cdots\}$ will more or less severely fail.
From another side, the recurrence formula (5) itself can exhibit some kind of "pseudostability", even in well-known theoretical condition [26, 27], particularly if the sampled points are equally, or nearly equally, spaced. This is the case of the "natural score" proposed by Beh [7]. In such cases, the accuracy of the $b_k(\bullet)$ computed from (5) may severely deteriorate as k approaches J. Fortunately, this is not very handicapping for us, since high degree $b_k(\bullet)$ are of no practical importance because of their very high variance estimation (Rayner and Best [51] recommended to rule out polynomials of degree > 4). For more insights on orthogonal polynomials, see the nice paper of Gautschi [27].

5.2. The OCA procedure

From now, we will suppose that the ordered set of categories is J, while the row set I is merely nominal. We will denote $\{\mathfrak{P}_p,\ 1 \leq p \leq J-1\}$ the system of column orthonormal polynomials (in the terminology of Beh [9]), sampled on $\{1, \ldots, J\}$, which play in OCA the role that principal axes play in CA; for an example, see Figure 3. They only depend on the marginal distribution $\mathbf{P}_J$ and on some **user-assigned score** s (*i.e.* a positive monotone function on J, which codes the ordinal structure of modalities [7]). Like in CA, the rank of the analysis is $M^* = \min(I-1, J-1) - 1$. These polynomials are characterized by the relationship:

$$\forall\ (m,p) \in M^* \times M^*,\ \sum_{j \leq J} P_j\, \mathfrak{P}_{m,j}\, \mathfrak{P}_{p,j} = \delta_m^p \tag{7}$$

which is quite similar to (1), but note that they are **not** orthogonal to the factors issued from CA (in other words, there is no relationship like (2)). Denoting Φ the $I \times M^*$ matrix of non-trivial singular vectors issued from CA, and $\mathfrak{P}_*$ the matrix of column orthonormal polynomial with the first (trivial, constant) column vector omitted, Beh [8, 9] considered the matrix of interactions

$$\mathbf{Z} := \Phi^t\, \mathbf{P}\, \mathfrak{P}_* \in M^* \times (J-1)$$

associated to the following decomposition of X^2:

$$\frac{X^2}{n} = \sum_{f=1}^{M^*} \sum_{p=1}^{J-1} Z_{f,p}^2 \tag{8}$$

where $Z_{f,p}^2$ (square of the entry of $\mathbf{Z}$ of row f and columm p) measures the intensity of the relationship between the polynomial of degree p , $\mathfrak{P}_p$, and the f^{th} factor φ_f^J (hybrid moment; for further details, see [8]).

The information bore by φ_f^J can be partitioned [9] in function of the polynomials:

$$\lambda_f^2 = \frac{1}{n}\sum_{p=1}^{J-1} Z_{f,p}^2. \tag{9}$$

We can measure the overall information bore by $\mathfrak{P}_p$ by the positive number μ_p:

$$\mu_p := \frac{1}{n}\sum_{f=1}^{M^*} Z_{f,p}^2 \tag{10}$$

obtaining this way a supplementary partition of X^2:

$$\frac{X^2}{n} = \sum_{p=1}^{J-1} \mu_p. \tag{11}$$

Notice that Formula (9) shows that $Z_{f,p}^2/n\lambda_f^2$ is the relative contribution of $\mathfrak{P}_p$ to the variance of φ_f^J.

Definition 5.1. Let us fix some $0 < \rho \leq 1$; we will write out that $\varphi_f^J \overset{\rho}{\approx} \mathfrak{P}_k$ if $\left(Z_{f,k}^2/n\lambda_f^2\right) \geq \rho$.

5.3. Choosing the score

As we saw in Section 5.1, the choice on the score in OCA can have an influence on the polynomial basis used in the analysis and consequently on its results.
This topic has been investigated from other points of view by Beh [7], Sarnacchiaro *et al.* [53], either for classical *a priori* scores, or *a posteriori* scores issued from a preliminary multivariate analysis (CA in the case of Beh [7], NSCA in the case of Sarnacchiaro *et al.* [53]). More precisely, Beh [7] noticed that, in the case of a doubly-ordered table (DOCA in the terminology of Lombardo and Beh [45]) "the correlation between two scoring schemes is equivalent to the correlation of their associated first non-trivial orthogonal polynomials"; consequently, the similarity between the results associated with various *a priori* scoring schemes can be roughly predicted. In addition, Beh [7, pp. 419-421] suggested to use singular vectors resulting from a preliminary classical CA as scores for DOCA; we will indeed adopt this strategy.

5.4. Significance tests in OCA

In the singly-ordered case, OCA gives rise to several significance tests since, asymptotically, each column component $n\,\mu_p \sim \chi^2\,(M^*)$ **because the user-assigned score** s **is not estimated** [52]. For the same reason, in the doubly-ordered case, a three-level battery of tests is available [6, 51] : at each (m, p) cell level, at each column component level, at each row component level, and of course at the global level (see [49] for an application in marine ecology).

But here, the situation will be different, because the chosen score will highly depend on the data! Consequently, we will have to perform permutation tests.

Remark 5. A similar approach for processing ordinal data with the help of discrete orthogonal polynomials was adopted by Haberman [37] in the setting of Log-linear models (see also Ben Hamou and Saporta [13]).

5.5. Extension to MCA

Beh and Lombardo [10], Lombardo and Beh [45], Lombardo and Meulman [46] proposed an extension of MCA to ordinal variables (OMCA). It generalizes the singly-ordered CA proposed by Beh [8], where the columns (say) of the studied table are ordered while the columns are not, to the case where the columns consist of $Q \geq 2$ blocks associated to ordinal variables (in other words, the complete disjunctive table $\mathbf{U}$ equipped with an order for each block). In OMCA, to the q^{th} block is associated a family $\mathbb{P}^q := \left\{\mathfrak{P}_0^q, \cdots, \mathfrak{P}_{w_q-1}^q\right\}$ of orthogonal polynomials associated with some *a priori* score, and each individual (row) belongs to the direct sum space generated by the orthogonal basis $\oplus_{q=1}^{Q}\mathbb{P}^q$. These authors stress that the position of the column categories in OMCA and classical MCA are the same, while the position of the individuals in OMCA greatly differ from those issued from MCA.

6. The proposed method

Since simple CA is a special case of MCA (see Section 4.2.1), we directly considered some complete disjunctive table $\mathbf{U} = (\mathbf{U}_1 \mid \mathbf{U}_2 \mid \cdots \mid \mathbf{U}_Q) \in N \times \left|\mathbf{W}^{(Q)}\right|$. The first factor issued from MCA of $\mathbf{U}$ is $\varphi_1 = \left(\varphi_{1,1}, \ldots, \varphi_{1,|\mathbf{W}^{(Q)}|}\right)$; let us denote π the permutation of $\mathbf{W}$ such that $\varphi_{1,\pi(1)} \leq \varphi_{1,\pi(2)} \leq \cdots \leq \varphi_{1,\pi\left(|\mathbf{W}^{(Q)}|\right)}$, and consider the **sorted** first factor

$$\left(x_1, \ldots, x_{|\mathbf{W}^{(Q)}|}\right) = \left(\varphi_{1,\pi(1)}, \varphi_{1,\pi(2)}, \ldots, \varphi_{1,\pi\left(|\mathbf{W}^{(Q)}|\right)}\right) := \pi \odot \varphi_1. \tag{12}$$

Suppose now the Guttman effect is present; we should have: $\varphi_{2,\pi(w)} = \pi \odot \varphi_2(w) = \mathfrak{P}_2\left(x_w\right)$, where $\mathfrak{P}_2$ is some polynomial of degree 2 such that $\sum_{w \leq \left|\mathbf{W}^{(Q)}\right|} x_w \, P_w \, \varphi_{2,\pi(w)} = 0$ while $\sum_{w \leq |W^{(Q)}|} x_w \, P_w \, x_w = 1$ and $\sum_{w \leq \left|\mathbf{W}^{(Q)}\right|} \varphi_{2,\pi(w)} \, P_w \, \varphi_{2,\pi(w)} = 1$, because of (1). Afterwards, if the scalogram associated with $\mathbf{U}$ is perfect, we will have $\pi \odot \varphi_3 = \mathfrak{P}_3$ for some degree 3 polynomial, *etc.* This is exactly the construction of Emerson [32], which is the basis of Beh's method [6]! Thus, the hybrid decomposition of X^2 will enable us to infer the order of the Guttman effect. Indeed, since in both cases (CA and OCA) the orthogonalization process starts with $\mathfrak{P}_0 = C_0$ and $\mathfrak{P}_1(x) = C_{1,0} + C_{1,1}x$ (with suitable constants), if the next factors are polynomials too, they should match with Emerson's polynomials. Thus, the proposed method consists in

(1) performing MCA of $\mathbf{U}$, obtaining φ_1 and sorting its coordinates, changing the

same way the order of the variables in the table (individuals can be processed the same way): $\mathbf{U} \to \pi \odot \mathbf{U}$

(2) perform OCA of the singly-ordered table $\pi \odot \mathbf{U}$, using the variable x defined in (12) as the score

(3) investigate the similarity between $\pi \odot \varphi_k$ and $\mathfrak{P}_k$, for each $k \geq 2$.

Remark 6. It is well-known that (M)CA can be considered as an optimal scaling method [54] (see also Section 2 and [33]): this is even the "Dutch approach" of multivariate analysis [47, 50]! We will follow this approach, using the optimal scaling of the columns resulting from MCA of $\mathbf{U}$ as a score for OCA of this **singly-ordered** multiple indicator table; thus, we consider a **unique** family $\left\{\mathfrak{P}_0, \cdots, \mathfrak{P}_{|W^{(Q)}|}\right\}$ of polynomials. One should obtain quite different results with OMCA (see Section 5.5), but our goal is different too...

In reference to Equation 9, we now use Definition 5.1 to lay a first definition of the order of the Guttman effect.

Definition 6.1. The Guttman effect is of order $K \geq 1$ for some fixed $0 < \rho \leq 1$ if

$$K = \underset{k \leq M^U}{\arg\max} \left(\pi \odot \varphi_k \overset{\rho}{\approx} \mathfrak{P}_k \right).$$

In reference to Equations (3, 8, 11), we laid supplementary definition requiring some inference.

Definition 6.2. The interaction $Z_{f,\,k}$ is **strongly significant** if

$$Z_{f,\,k} \overset{\alpha}{\neq} 0 \wedge \left(\lambda_f \overset{\alpha}{\neq} 0 \wedge \mu_k \overset{\alpha}{\neq} 0 \right)$$

where the expression $Z_{f,\,k} \overset{\alpha}{\neq} 0$ (resp. $\lambda_f \overset{\alpha}{\neq} 0$ or $\mu_k \overset{\alpha}{\neq} 0$) means that the interaction between $\pi \odot \varphi_f$ and $\mathfrak{P}_k$ (resp. the role of this eigenvalue or this polynomial) is statistically significant at some fixed level $1 - \alpha$.

Definition 6.3. The Guttman effect is **strongly** of order K if

$$K = \underset{k \leq M^U}{\arg\max} \left(Z_{k,\,k} \overset{\alpha}{\neq} 0 \wedge \left(\lambda_k \overset{\alpha}{\neq} 0 \wedge \mu_k \overset{\alpha}{\neq} 0 \right) \right).$$

Remark 7. Since the chosen score is highly dependent on the data, the tests proposed in [6, 8] are inapplicable in our case: we will have instead to perform the permutation tests detailed hereunder.

6.1. Randomization

Permutation tests (or sometimes cross-validation [21, 29]) are frequently used in Multivariate Analysis [22, 28, 30], because of the complexity of the distributions involved (see Section 4). Here, we build from the original table $\mathbf{U}$ a convenient number K of random tables $\mathbf{U}^\tau := \left(\mathbf{U}_1^{\tau_1} \mid \mathbf{U}_2^{\tau_2} \mid \cdots \mid \mathbf{U}_Q^{\tau_Q} \right) \in N \times \left| \mathbf{W}^{(Q)} \right|$, such that the block

$\mathbf{U}_q^{\tau_q}$ is obtained from $\mathbf{U}_q$ by permuting all the w_q columns of **each row** r with some random permutation $\tau_q(r)$. The random matrix $\mathbf{U}^\tau$ is similar to $\mathbf{U}$: it has the same dimensions, the same grand total and the same marginal probability $\frac{1}{NQ}\mathbf{1}^N$, and it is a complete disjunctive table. But relationships between the Q variables are completely destroyed.

The hybrid decomposition of $\pi \odot \mathbf{U}$ simultaneously give rise to the spectra $\Lambda := (\lambda_1, \cdots, \lambda_{M^\cup})$, the vector of moments $\Xi := \left(\mu_1, \cdots, \mu_{|\mathbf{W}^{(Q)}|\text{-}1}\right)$ and the table of interactions $\Theta := \left(Z_{f,p},\ 1 \leq f \leq M^{\mathbf{U}},\ 1 \leq p \leq \left|\mathbf{W}^{(Q)}\right| - 1\right)$ associated with the sorted data. Consider now the K randomized tables $\{\pi_1 \odot \mathbf{U}_1^\tau, \cdots, \pi_K \odot \mathbf{U}_K^\tau\}$, associated with the hypothesis (**H**) of pairwise independence of the variables. Since the analysis of each one of these tables ($\pi_r \odot \mathbf{U}_r^\tau$, say) give rise to some (Λ_r, Ξ_r, Θ_r), the inference on eigenvalues (resp. moments, interactions) will consist in comparing each λ_f (resp. μ_p, $Z_{f,p}$) with the distribution of the randomized analogues, recorded in the series $\{(\Lambda_r, \Xi_r, \Theta_r) : r \leq K\}$. Then, we will be able to draw box-plots of these quantities and/or decide with respect to some fixed threshold α, whether or not each interaction $Z_{f,p}$ is significant, and whether or not the eigenvalue λ_f (resp. moment μ_p) is significant.

After a few trials, we systematically fixed the number of permutations to $K = 100$, and α to 0.1.

7. Application to artificial data

First, we tested the method on artificial data, either totally random or presenting by construction the Guttman effect. The latter dataset consisted of three functions supported by $[-a, a]$ with $a = 1.23758$, corrupted by a uniformly distributed noise of increasing level σ. For each one of these functions, $f(x)$ say, we calculated $\bar{f} := \max_{x\in[-a,a]} f(x)$, $\underline{f} := \min_{x\in[-a,a]} f(x)$ and divided $\mathbb{R}$ into a family $\mathscr{F}$ of seven disjointed intervals:

$$\mathscr{F} = \left\{]-\infty, \underline{f} + \tfrac{(\bar{f}-\underline{f})}{3}], \cdots,]\bar{f} - \tfrac{(\bar{f}-\underline{f})}{3}, +\infty[\right\}.$$

Then, we generated the data

$$\{y_k = f(x_k) + \mathcal{D}^\sigma,\ 1 \leq k \leq 200\} \tag{13}$$

where $\mathcal{D}^\sigma$ denotes the uniform distribution on $\left[-\frac{\sigma(\bar{f}-\underline{f})}{6}, \frac{\sigma(\bar{f}-\underline{f})}{6}\right]$, with $\sigma \in \{0, 0.1, 0.5, 1, 1.5, 3\}$ and x_k denotes the k^{th} Chebyshev point on $[-a, a]$. The chosen functions were:

$$\begin{cases} \Theta(x) := 20 \arcsin\left(\frac{x}{a}\right) \\ f(x) := (5(x+1)-3)(5(x+1)-6)(5(x+1)-9)/5 \\ g(x) := -50\, x \exp\left(-\sqrt{a+x}\right) \end{cases}$$

and the obtained data are plotted on Figure 1.

[Figure 1 about here.]

Each y_k was then assigned to the right interval of $\mathscr{F}$ (logical coding), giving rise to some binary table, and the complete table $\mathbf{U} = (\mathbf{U}_1 \mid \mathbf{U}_2 \mid \mathbf{U}_3) \in 200 \times 21$ was submitted to CA. The 21 characters associated with the three functions and the seven intervals were labeled $\{\Theta_1, \cdots, \Theta_7\}$, $\{f_1, \cdots, f_7\}$ and $\{g_1, \cdots, g_7\}$.

7.1. Analysis of a random dataset

It consisted in a purely random table $\mathbf{U} = (\mathbf{U}_1 \mid \mathbf{U}_2 \mid \mathbf{U}_3) \in 200 \times 21$. At the level 0.9, the trace test of Section 4.2.1 didn't indicate any structure in these data. This is confirmed by Figure 2: none of the eigenvalues (except the fifth one, by chance) was significant, and all of them were situated in the confidence band given by [12] (see Section 4.2.2). More precisely, on the left upper panel of Figure 2, one can find:

- the plot of eigenvalues (black stars and dashed curve)
- the typical value $\bar{\lambda}$ and the confidence band given by [12, 13] (gray horizontal lines)
- the box-plot of eigenvalues issued from randomization, and the associated quantiles of order 0.9 (green curve).

On the right panel, the reader will find the corresponding information (when it is available) for the polynomials: while none of the eigenvalues seemed significant, the coefficients of polynomials of degree 1 and 3 were slightly above the corresponding quantiles of order 0.9.

[Figure 2 about here.]

In addition, Figure 2 shows that, even if a lot of interactions were significant at the prescribed level, the Guttman effect was absent, since no interaction was compatible with Definition 6.2, for $\alpha = 0.1$.

7.2. Analyses of the functional datasets

We analyzed the six datasets associated with $\sigma \in \{0, 0.1, 0.5, 1, 1.5, 3\}$ but, for sake of brevity, we will only detail two cases; the complete results are summarized in Table 1.

Consider first a moderately noisy dataset, generated according to formula (13) with $\sigma = 0.1$ (see Figure 1). We plotted on Figure 3 the six first Emerson's polynomials associated with the first eigenvector of CA as a score (21 characters: 3 variables, seven intervals). Formulas corresponding to the four first ones are: $0.999886x - 0.000162283$, $2.33877x^2 + 2.40305x - 2.33969$, $2.95361x^3 + 5.83345x^2 - 0.619064x - 2.79967$ and $8.49825x^4 + 18.6544x^3 - 3.41594x^2 - 13.9222x + 3.5578$.

[Figure 3 about here.]

According to the trace test, ($\boldsymbol{H}$) was clearly rejected while, according to the permutation tests, 5 eigenvalues and 5 polynomials were significant (see the upper panels of Figure 4).

[Figure 4 about here.]

Furthermore, one can see on the lower right panel of Figure 4 that a Guttman effect could be detected, which was strongly of order 3 (black cells on the diagonal of the table of interactions). The first principal plane is represented on Figure 5; notice that both variables and individuals are projected on a common parabola.

[Figure 5 about here.]

Consider now a much noisier dataset, with $\sigma = 3$ (see Figure 1). According to the trace test, $(\boldsymbol{H})$ was accepted; nevertheless, contrary to the purely random case (see Figure 4), a single factor and a single polynomial could be detected on the upper panels of Figure 6; this is confirmed on the lower right panel of the figure.

[Figure 6 about here.]

The first principal plane is represented on Figure 7. Notice that variables don't exhibit any particular structure, while **individuals are projected yet on some "fuzzy parabola"**. So, no Guttman effect could be detected by the tests in this case. Nevertheless, thanks to CA, one can suspect its reality, although the functional nature of the data has been blurred by the noise.

The main results are displayed on Table 1; one can see that the Guttman effect was of order 3 (for $\rho = 0.4$), except for very noisy data; the strong Guttman order was similar, with higher fluctuations.

[Figure 7 about here.]

When a stronger similarity was demanded ($\rho = 0.8$) the order of the Guttman effect decreased, but it was still unveiled, except for data contaminated by some high level of noise.

[Table 1 about here.]

8. A toy dataset: the archaic Chinese vases data

We will exemplify a data set analyzed by Benzécri and his collaborators [14, pp. 323-325], after Elisseeff [24] who used Guttman's permutation methods. The original dataset [14, 24] was a contingency table: 17 types of vases ("large Yeou") described by 8 binary characters. Notice there were indeed **112 vases** assigned to these 17 types. Eliminating four characters, Benzécri and his collaborators [14, pp. 323-325] built a sub-table resulting in a perfect scalogram. We built from this sub-table the complete disjunctive table $\mathbf{U} = (\mathbf{U}_1 \mid \mathbf{U}_2 \mid \mathbf{U}_3 \mid \mathbf{U}_4) \in 112 \times 8$, and analyzed it. According to the trace test, $(\boldsymbol{H})$ was rejected (p-value: 10^{-5}).

The reader can see on the upper panels of Figure 8 that, according to the permutation tests, a single eigenvalue and two polynomials were statistically significant at the level 0.9, while a single interaction was strongly significant (right lower panel of Figure 8). $\mathbf{U}$ could be represented by a perfect scalogram, and the detected Guttman effect was of order two (see the right panel of Figure 9).

[Figure 8 about here.]

Remark 8. We plotted on the left panel of Figure 9 similarities between all the sorted factors $\pi \odot \varphi_f$ and $\mathfrak{P}_k$ (with 0.1 as a threshold), and similarities between the pairs $(\pi \odot \varphi_k, \mathfrak{P}_k)$, on the right panel. Both figures show a clear Gutmann effect of order two, but it is worth noting (see the left panel of this figure) that further factors were similar to others Beh's polynomials... More precisely, while the similarity between $\pi \odot \varphi_2$ and $\mathfrak{P}_2$ was 0.984977, we found that between $\pi \odot \varphi_3$ and $\mathfrak{P}_4$ it was 0.564397; $\pi \odot \varphi_3$ and $\mathfrak{P}_5$ corresponded to 0.380075, while the similarity between $\pi \odot \varphi_4$ and $\mathfrak{P}_6$ was 0.98357!

[Figure 9 about here.]

In conclusion, these archaic Chinese vases presented a Guttman effect of order two, but not statistically significant (strongly of order one)...

9. Application to the synchrotron X-ray diffraction dataset

Lessivage is among the most widespread processes in soils and it has been described in many soils types. This process is defined as a substantial vertical transfer of fine particles from a horizon, called eluviated horizon to another horizon referred to illuviated horizon. It was experimentally simulated in lab (i.e. simulation on a sequence of rainfall events) to identify the processes and factors responsible for it. The lab experiment setup consists in a sequence of 30 rainfalls on undisturbed and unsaturated soil columns of decimeter size. As smectites were described as especially sensitive to eluviation, an experiment, designed to simulate illuviation, consisted in columns made of two overlaid soil monoliths, the upper one contained smectite while the lower one did not (for more information, see [20]). Thin sections were made in the lower monoliths for different amounts of rain and different rainfall intensities. Localizing and determining the mineralogical composition of these thin sections by mapping them with a focused X-ray beam (for lateral resolution) would allow to locate structural changes due to lessivage, thanks to the presence of eluviated smectite. X-ray diffraction (XRD) is used for this purpose. For technical reasons (need of micron-sized collimated intense X-ray beams), analyzing them with a conventional lab X-ray diffraction device is impossible in our case and synchrotron X-ray diffraction had to be considered. However, with that technique, the main peaks classically used for clay identification were not recorded, and the relative intensities of the different diffraction peaks are meaningless. Indeed, with a probed sample volume of $10 \times 10 \times 30$ μ m^3 and "large" sizes of crystallites (within the micron), the sample is very far from what a powder used in XRD experiments would be. Consequently, depending on the size, number and orientation of the crystallites present in the illuminated volume, the diffracted signal is, in the most favorable case, a spotty one approximating the ring- shape (2θ = constant) of a diffraction peak. Even if using an area detector (like in our experiment) and consider a random orientation of crystallites, it is still possible that no crystallite will diffract in the detector (i.e. missing Bragg peak) or, in a more favorable case, to detect intensity originating only from few crystallites (i.e. only few spots on the area detector). Consequently, the corresponding detected scattered intensity is not representative anymore to be used in a structure refinement procedure. This can be summarized as follows: if a Bragg peak is not detected, it does not mean the corresponding lattice does not exist (possibly no crystallites oriented in Bragg condition). If a Bragg peak is detected, the corresponding inter-reticular distance is present and fulfill Bragg law (but the corresponding detected intensity is meaningless, since not proportional to the quantity of the corresponding crystalline phase in the investigated volume, even after structure factor correction). The low number of crystallites in the probed sample volume (due to the large grain sizes) is the origin of both of these issues.

Therefore the herunder preliminary analysis, consisting in investigating the feasibility of identifying the minerals of interest by synchrotron X-ray diffraction mapping of soil features, was unavoidable. More precisely, if each mineral was associated with a specific group of "coding diffraction angles", the minerals could be considered as

independent: the detection of angles associated with some mineral S, say, would bore no information about the possible detection of $S' \neq S$. But it is not the case: many angles are paired with several minerals! Moreover, as it will be shown later in this paper, there is co-localization / coexistence of phases on particular lateral positions on the sample slab. That's why it was relevant to analyze the relationship between angles and minerals.

9.1. The coding used

The following list of minerals was selected: {Feldspars, Goethite, Illite, Kaolinite, Maghemite, Quartz, Smectite}. All the minerals were characterized (in 0/1) by 228 diffraction angles, each angle making possible to detect one or several minerals. Each mineral S was thus associated with a binary vector of length 228.

Notice that the masses (number of characteristic angles) of these 7 vectors were very different from each other (see Table 2), and that each one of the minerals is associated with a smaller number of exclusive angles. Consequently we split each S into S^+ (detection of S) and the "anti-mineral" S^- (non-detection of S) in order that all the minerals, described by both these variables, had the same weight: 228. In addition, any S, characterized by the pair (S^+, S^-), will have $\frac{1}{Q} = \frac{1}{7}$ as part of variance (see Section 3.2). So, no mineral was favored in the CA of the resulting 228×14 binary table.

[Table 2 about here.]

9.2. Results of the tests

We displayed on Figure 10 issues from the analysis of this table, which are rather similar to Figure 6. According to the trace test, the variables seemed independent, while according to the permutation tests, two polynomials (of degree 1 and 13) and two factors were significant. These factors were accounted for 23% and 18% of the total variance, respectively, or 80% and 15% if we adopt the Benzécri's correction for percentages of inertia (4) based on the three first eigenvalues (see Figure 10). A number of interactions were retained by the permutation test, but only $Z_{1,1}$ and $Z_{1,13}$ were strongly significant. It is worth noting here, in connection with the considerations of Section 5.1, that the orthonormality relationship (7) between Beh's polynomials gradually deteriorated for degrees greater than 11; more precisely, $\sum_{j \leq J} P_j \, \mathfrak{P}_{13,j} \, \mathfrak{P}_{p,j} \approx 3.\,10^{-5}$ for $p \leq 3$. So, we cannot be sure that the three last polynomials were determined with a satisfactory precision...

[Figure 10 about here.]

Thus, there was no Guttman effect in this case, and only the first factor was undoubtedly significant. It is displayed on Figure 11, together with the second factor (not significant). Interestingly, none of the "anti-minerals" (except Feldspars-) seems to play an important part in the analysis. We can distinguish along the first factor six clusters of variables: Goethite+, Quartz+, {Kaolinite+, Feldspars-}, Feldspars+, {Maghemite+, Illite+}, {Smectite+, Goethite-, Illite-, Maghemite-, Quartz-, Smectite-} and Feldspars+.

In conclusion, only this common structure seems meaningful, while the remaining variability is noise.

Remark 9. The fact that Smectite+ and Smectite- are very close to each other along

factor 1 means on the one hand that this mineral is rather hard to detect for us and, on the other hand that the presence or absence of Smectite does not depend much on the other minerals of interest. This is perhaps due to the fact that it possesses a single exclusive angle (see Table 2), but Maghemite shares the same property.

[Figure 11 about here.]

10. Conclusion

We propose a method for detecting the existence of a Guttman effect in a complete disjunctive table, with a given level of confidence.

Firstly, we recall results from the literature about the significance of eigenvalues resulting from the MCA of such tables, and about the χ^2 independence test for associated Burt's tables. Thanks to the randomization method proposed afterward, we are able to test the presence of this phenomenon and its order K (*i.e.* the maximum degree of the significant polynomials). The data could in this case be approximately represented by a parametric curve in $\mathbb{R}^K$ and, as a consequence, the original table could be roughly reconstituted (filtered) from the first K components of either CA or OCA, thanks to the reconstitution formulas (it seems that JCA works better than MCA for data reconstruction [19]; nevertheless superseding MCA by JCA in the proposed method is not straightforward).

The original table could be this way split into a table associated with the Guttman effect, and a residual. When the Guttman effect is associated with some gradient, its influence could be eliminated by considering only the residual table. This method is related to the "detrending-by-polynomials" approach of DCA, but in this case the existence and the order of the phenomenon could be tested (not postulated), and the polynomials used for a possible detrending would be directly associated with the Guttman effect, instead of being arbitrary. At last, while the importance of some component is classically measured by the corresponding eigenvalue, in the case of a Guttman effect of order K, it would be consistent to measure the importance of the first component by the sum of the K first eigenvalues.

With a view to future work, it would be judicious to improve the computation of the orthogonal polynomials. This problem has been tackled by Gautschi [25, 27], who proposed to supersede the Stieltjes procedure (5,6) by the modified Chebyshev algorithm. Roughly speaking, it consists in the orthogonalization (in $\mathbb{L}^2_\lambda$) of some standard family of orthogonal polynomials (Legendre, Chebyshev, Laguerre, Hermite, Krawtchouk, *etc*), much better conditioned, thanks to another recurrence relation found by Chebyshev in 1859 [27].

The method detailed in this work has been implemented in a *Mathematica* [56] package, available from the first author (work in progress).

Acknowledgements

This research was conducted in the framework of the Agriped project (ANR-10-BLANC-605) supported by the French National Research Agency (ANR).

We acknowledge C. Lelay (URSol INRA Orléans,France) for the thin sections production. Synchrotron SOLEIL (France) is acknowledged for allocating beamtime for the local-probe X-ray diffraction experiments.

We thank the referees for their stimulating comments and bibliographic contributions, as well as J.-P. Durbec for helpful discussions.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] R. Alvarez, M. Bécue, J. J. Lanero and O. Valencia, *Results stability in textual analysis: its application to the study of the Spanish investiture speeches*, 6^{es} Journées internationales d'Analyse Statistique des Données Textuelles (2002), pp. 1-12.

[2] R. Alvarez, M. Bécue and O. Valencia, *Etude de la stabilité des valeurs propres de l'AFC d'un tableau lexical au moyen de procédures de rééchantillonnage*, 7^{es} Journées internationales d'Analyse Statistique des Données Textuelles (2004), pp. 42-51.

[3] R. Alvarez, M. Bécue and O. Valencia, *Partial bootstrap in CA: correction of the coordinates. Application to textual data*, 8^{es} Journées internationales d'Analyse Statistique des Données Textuelles (2006), pp. 43-53.

[4] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York, 1958.

[5] A. Baccini, *Etude comparative des représentations graphiques en analyses factorielles des correspondances simples et multiples*,Toulouse, Université Paul Sabatier, Laboratoire de Statistique et Probabilités, Rapport No 02-84 (1984), pp. 1-64.

[6] E. J. Beh, *Simple Correspondence Analysis of ordinal cross-classifications using orthogonal polynomials*, Biom. J. (1997), 39, 5, pp. 589-613.

[7] E. J. Beh, *A comparative study of Scores for Correspondence Analysis with ordered categories* (1998), Biom. J. 40, 4, pp.413-429.

[8] E. J. Beh,*Partitioning Chi-squared statistics for singly ordered two-way contingency tables*, Aust. N. Z. J. Stat. (2001), 43, 3 pp. 327-333.

[9] E. J. Beh, *Simple Correspondence Analysis of Nominal-Ordinal contingency tables*, Journal of Applied Mathematics and Decision Sciences (2008), pp. 1-17.

[10] E. J. Beh and R. Lombardo, *Correspondence Analysis Theory, Practice and New Strategies*, Wiley Series in Probability and Statistics, Chichester, United Kingdom, 2014.

[11] P. Bekker and J. de Leeuw, *Relations between variants of non-linear Principal Components Analysis*, In: Component and Correspondence Analysis - dimension reduction by functional approximation, J. L. A. van Rijckevorsel and J. de Leeuw Editors, John Wiley & Sons, Chichester, United Kingdom, pp. 1-31, 1988.

[12] S. Ben Hammou and G. Saporta, *Sur la normalité asymptotique des valeurs propres en ACM sous l'hypothèse d'indépendance des variables*, Revue de Statistique Appliquée (1998), 46, 3, pp. 21-35.

[13] S. Ben Hammou and G. Saporta, *On the connection between the distribution of eigenvalues in Multiple Correspondence Analysis and Log-Linear models*, In: Proceedings of CARME2003, Barcelona, pp. 41-79, 2003.

[14] J.-P. Benzécri et coll., *L'Analyse des Données, Vol.2*, Dunod, Paris, 1973.

[15] J.-P. Benzécri, *Sur l'analyse des tableaux binaires associés à une correspondance multiple [BIN. MULT.]*, Les cahiers de l'analyse des données (1977), 2, 1, pp. 55-71.

[16] J.-P. Benzécri, *Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [BIN. MULT.]*, Les cahiers de l'analyse des données (1979), 4, 3, pp. 377-378.

[17] J.-P. Benzécri, *Histoire et préhistoire de l'Analyse Des Données*, Bordas,Paris, 1982.

[18] S. Camiz, *The Guttman effect: its interpretation and a new redressing method*, Data Analysis Bulletin (2005), 5, pp. 7-34.

[19] S. Camiz and G. C. Gomes, *Joint Correspondence Analysis Versus Multiple Correspondence Analysis: A Solution to an Undetected Problem*, In: Classification and Data Mining, Studies in Classification, Data Analysis and Knowledge Organization, GIUSTI A. et al. (eds.), Berlin, Springer, pp. 11-18.

[20] S. Cornu, L. Quénard, I. Cousin and A. Samouëlian, *Experimental approach of lessivage: quantification and mechanisms*, Geoderma (2014), 213, pp. 357-370.

[21] G. Diana and C. Tommasi, *Cross-validation methods in principal component analysis: a comparison*, Statistical Methods & Applications (2002), 11, pp. 71-82.

[22] §. Dray, *On the number of principal components: a test of dimensionality based on measurements of similarity between matrices*, Computational Statistics & Data Analysis (2008), 52, pp. 2228-2237.

[23] M. J. Greenacre, *Correspondence analysis of multivariate categorical data by weighted least squares*, Biometrika (1988), 75, 3, pp. 457-467.

[24] V. Elisseeff, *Possibilité du scalogramme dans l'étude des bronzes chinois archaïques*, Mathématiques et Sciences Humaines (1965), 11, pp. 1-10.

[25] W. Gauschi, *On generating orthogonal polynomials* , SIAM J. Sci. Stat. Comput. (1982), 3, 3, 289-317.

[26] W. Gauschi, *Is the recurrence relation for orthogonal polynomials alwaws stable?* , BIT (1993), 277-284.

[27] W. Gauschi, *Orthogonal polynomials: applications and computation*, Acta Numerica (1996), 45-119.

[28] J. Josse, J. Pagès and F. Husson, *Testing the significance of the RV coefficient*, Computational Statistics & Data Analysis (2008), 53, pp. 82-91.

[29] J. Josse and F. Husson, *Selecting the number of components in principal component analysis using cross-validation approximations*, Computational Statistics & Data Analysis (2012), 56, pp. 1869-1879.

[30] J. Josse and S. Holmes, *Measuring multivariate association and beyond*, Statistics Surveys (2016), 10, pp. 132–167.

[31] N. El Karoui and E. Purdom, *The bootstrap, covariance matrices and PCA in moderate and high-dimensions*, https://arxiv.org/abs/1608.00948v1 (2016).

[32] P. L. Emerson, *Numerical construction of orthogonal polynomials from a general recurrence formula*, Biometrics (1968), 24, 3, pp. 695-701.

[33] J. C. Gower, *Fisher's optimal scores and Multiple Correspondence Analysis*, Biometrics (1990), 46, pp. 947-961.

[34] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press, Orlando, 1984.

[35] L. Guttman, *The Cornell technique for Scale and Intensity Analysis*, Educational and Psychological Measurement (1947), pp. 247-279.

[36] L. Guttman and E. A. Suchman, *Intensity and a Zero Point for Attitude Analysis*, American Sociological Review (1947), 12, 1, pp. 57-67.

[37] S. J. Haberman, *Log-Linear Models for Frequency Tables with Ordered Classifications*, Biometrics (1974), 30, 4, pp. 589-600.

[38] A. Hendrickse, L. Spreeuwers and R. Veldhuis, *A bootstrap approach to eigenvalues correction* (2009), N^{inth} IEEE International Conference on Data Mining, https://doi.org/10.1109/ICDM.2009.111.

[39] A. Hendrikse, R. Veldhuis and L. Spreeuwers, *Smooth eigenvalue correction*, EURASIP Journal on Advances in Signal Processing (2013), pp. 1-16.

[40] M. O. Hill and H. G. Gauch, *Detrended Correspondence Analysis: an improved ordination technique*, Vegetatio (1980), 42, pp. 47-58.

[41] D. A. Jackson and K. M. Somers, *Putting Things in Order: The Ups and Downs of Detrended Correspondence Analysis*, The American Naturalist (1991), 137, 5 , pp. 704-712.

[42] R. G. Knox, Effects of detrending and rescaling on correspondence analysis: solution stability and accuracy, Vegetatio (1989), 83, pp. 129-136.
[43] L. Lebart, A. Morineau and M. Piron, *Statistique exploratoire multidimensionnelle*, Dunod, Paris, 1995.
[44] L. Lebart and G. Saporta, *Historical Elements of Correspondence Analysis and Multiple Correspondence Analysis*, in: Visualization and Verbalization of Data, J. Blasius and M. Greenacre eds, CRC Press, Chapman & al., New York, 2014.
[45] R. Lombardo and E. Beh, *Simple and multiple correspondence analysis for ordinal-scale variables using orthogonal polynomials*, Journal of Applied Statistics (2010), 37, 12, pp. 2101-2116.
[46] Lombardo R. and Meulman J. , *Multiple correspondence analysis via polynomial transformations of ordered categorical variables*, Journal of Classification (2010), 27, 191- 216.
[47] P. Mair and J. de Leeuw, *A general framework for Multivariate Analysis with Optimal Scaling: the R package "aspect"*, Journal of Statistical Software (2010), 32, 9, pp. 1-23.
[48] C. Manté, *Etude par l'Analyse des Données de la mémoire d'un champ météorologique*, Thèse de 3 ème cycle, Université Pierre et Marie Curie, Paris, 1981.
[49] C. Manté, G. Bernard, P. Bonhomme and D. Nerini, *Application of ordinal correspondence analysis for submerged aquatic vegetation monitoring*, Journal of Applied Statistics (2013), 40, 8, pp. 1619-1638.
[50] G. Michailidis and J. de Leeuw, *The Gifi system of Descriptive Multivariate Analysis*, Statistical Science (1998), 13, 4, pp. 307-336.
[51] J.C.W. Rayner and D.J. Best, *Smooth extensions of Pearson's product moment correlation and Spearman's Rho*, Statist. Probab. Lett. 30 (1996), pp. 171-177.
[52] J.C.W. Rayner and D.J. Best, *Analysis of singly ordered two-way contingency tables*, Journal of Applied Mathematics and Decision Sciences (2000), pp. 83-98.
[53] P. Sarnacchiaro, A. D'Ambra and L. D'Ambra, *CATANOVA for ordinal variables using orthogonal polynomials with different scoring methods*, Journal of Applied Statistics (2016), 43, 13, pp. 2490-2502.
[54] M. Tenenhaus and F. Young, *An analysis and synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity analysis and other methods for quantifying categorical multivariate data*, Psychometrika (1985), 50, 1, pp. 91-119.
[55] C.J.F. Ter Braak, *CANOCO - a FORTRAN Program for Canonical Community Ordination by [Partial] [Detrended] [Canonical] Correspondence Analysis, Principal Components Analysis and Redundancy Analysis (Version 2.1)*, Agriculture Mathematics Group, Wageningen (1987).
[56] Wolfram Research, Inc., Mathematica, Version 12.1, Champaign, IL (2020).

Figure captions

Figure 1 : Three functions with added noise.

Figure 2 : Random data. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel: no interaction was strongly significant (see Definition 6.2).

Figure 3 : Moderately noisy data: $\sigma = 0.1$. Plot of the five first column orthonormal polynomials, in accordance with Emerson's paper [32]. A abscissas correspond to the 21 values of the score (first factor of the preliminary CA.

Figure 4 : Moderately noisy data: $\sigma = 0.1$. Same structure as in Figure 2. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel (in black): strongly significant interactions (see Definition 6.2). The right lower panel indicates the existence of a strong Guttman effect of order 3.

Figure 5 : Moderately noisy data: first plane of CA (the 21 variables are plotted in green/gray).

Figure 6 : Very noisy data: $\sigma = 3$. Same structure as Figure 2. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel (in black): strongly significant interactions (see Definition 6.2).

Figure 7 : Very noisy data: first plane of CA. (the 21 variables are plotted in green/gray).

Figure 8 : The archaic Chinese vases data. Same structure as Figure 2. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel (in black): strongly significant interactions (see Definition 6.2).

Figure 9 : The archaic Chinese vases data: representation of the table of similarities between the sorted factors $\pi \odot \varphi_f$ and the $\mathfrak{P}_k$. On the left panel, black cells correspond to similarities greater than the fixed threshold (0.1). On the right panel, we plotted the values extracted from the diagonal of this table.

Figure 10 : The X-rays diffraction data. Same structure as Figure 2. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel (in black): strongly significant interactions (see Definition 6.2).

Figure 11 : Coordinates of the minerals on the first plane of CA . Detection variables (+) are plotted with a yellow background; non-detection ones (-) are plotted vertically, with a blue background. Remember that the second dimension is not significant.

Table 1. Artificial data: summary of the obtained results.

σ	Guttman effect order	Strong order	Significant λ_k	Significant μ_k
	$\rho = 0.4$ ($\rho = 0.8$)	($\alpha = 0.1$)	($\alpha = 0.1$)	($\alpha = 0.1$)
0	3 (3)	3	7	5
0.1	3 (3)	3	7	6
0.5	3 (2)	2	6	6
1	5 (3)	5	6	5
1.5	3 (2)	3	4	3
3	1 (1)	1	1	1
Rnd	1 (1)	0	0	0

Table 2. The minerals characteristics

Mineral:	Feldspars	Goethite	Illite	Kaolinite	Maghemite	Quartz	Smectite
Angles:	148	14	17	52	5	12	6
Exclusive:	117	7	11	30	1	7	1

Figure 1. Three functions with added noise

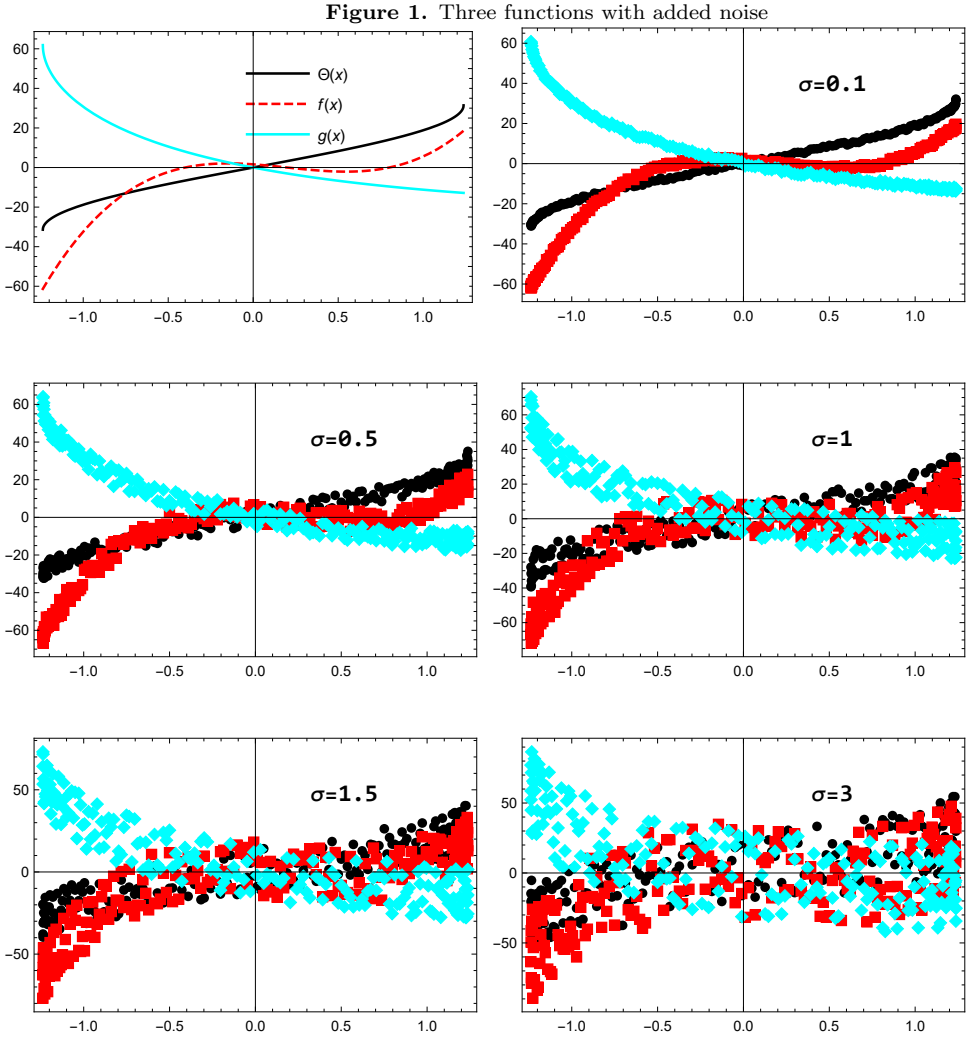

Figure 2. Random data. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel: no interaction was strongly significant (see Definition 6.2).

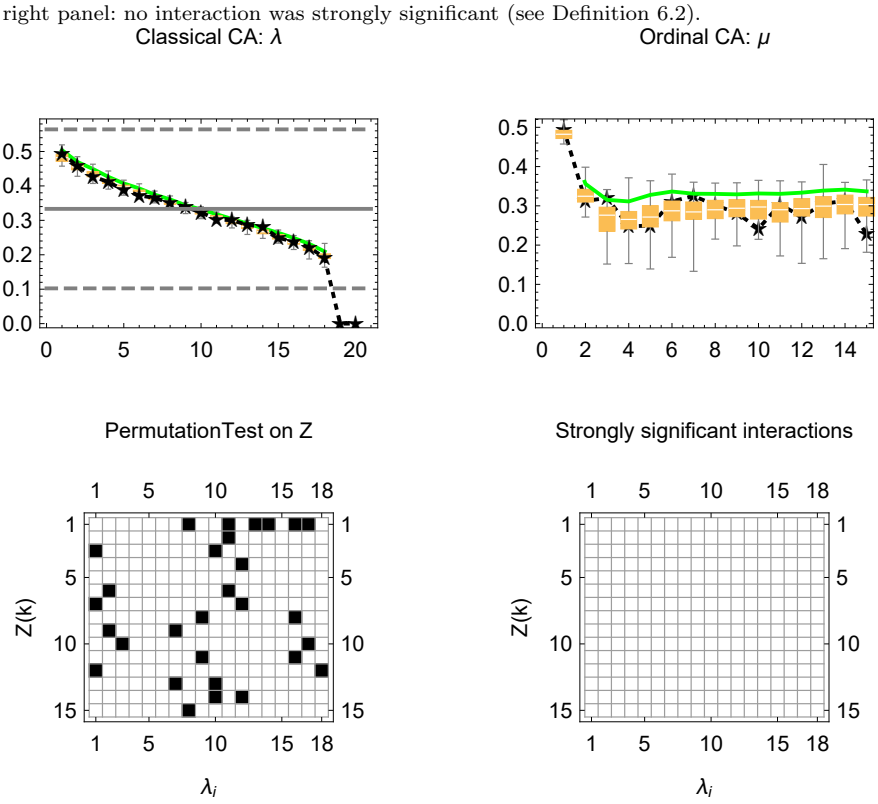

Figure 3. Moderately noisy data: $\sigma = 0.1$. Plot of the five first column orthonormal polynomials, in accordance with Emerson's paper [32]. abscissas correspond to the 21 values of the score (first factor of the preliminary CA.

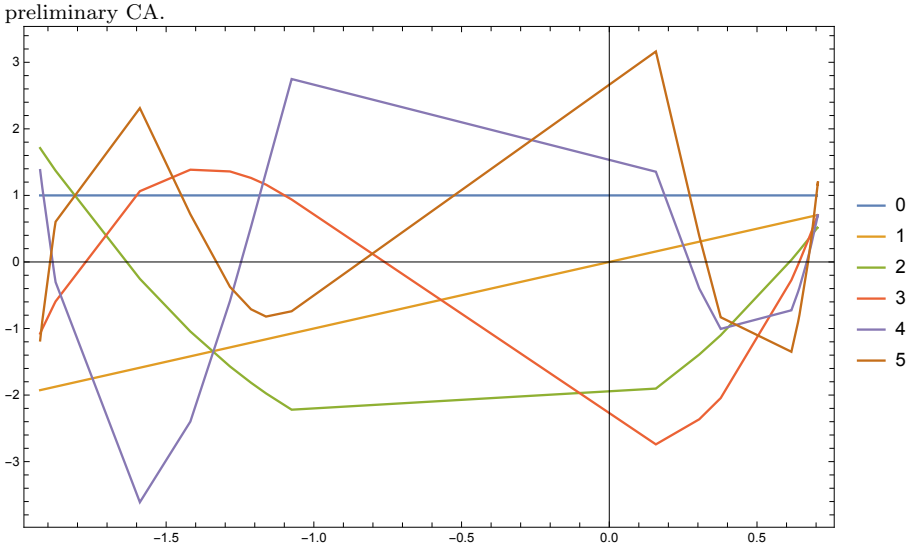

Figure 4. Moderately noisy data: $\sigma = 0.1$. Same structure as in Figure 2. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel (in black): strongly significant interactions (see Definition 6.2). The right lower panel indicates the existence of a strong Guttman effect of order 3.

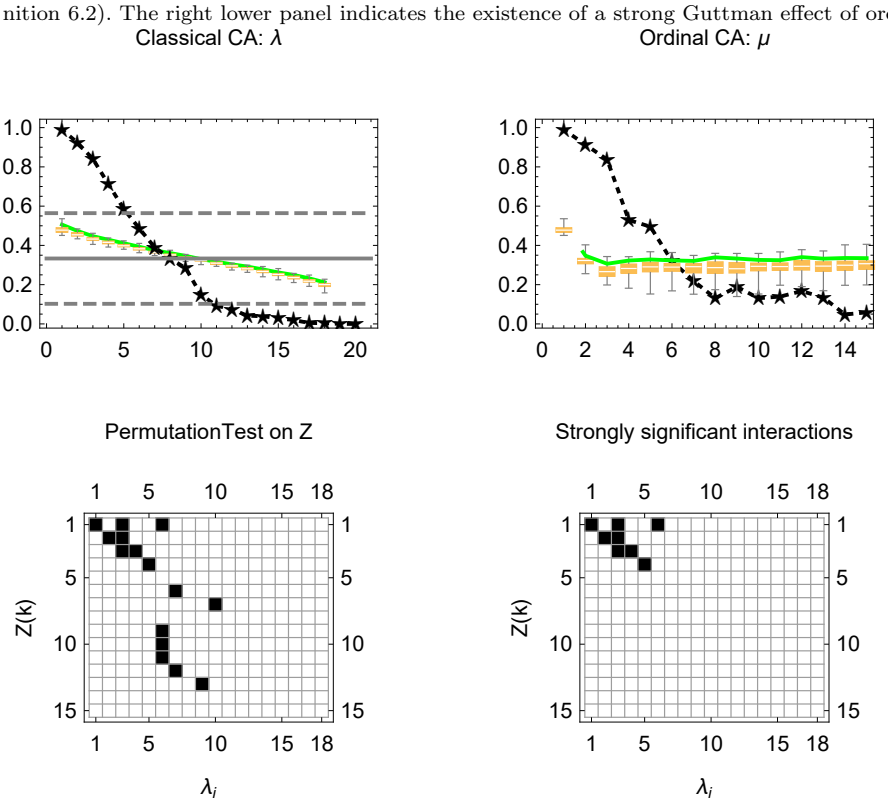

Figure 5. Moderately noisy data: first plane of CA. (the 21 variables are plotted in green/gray).

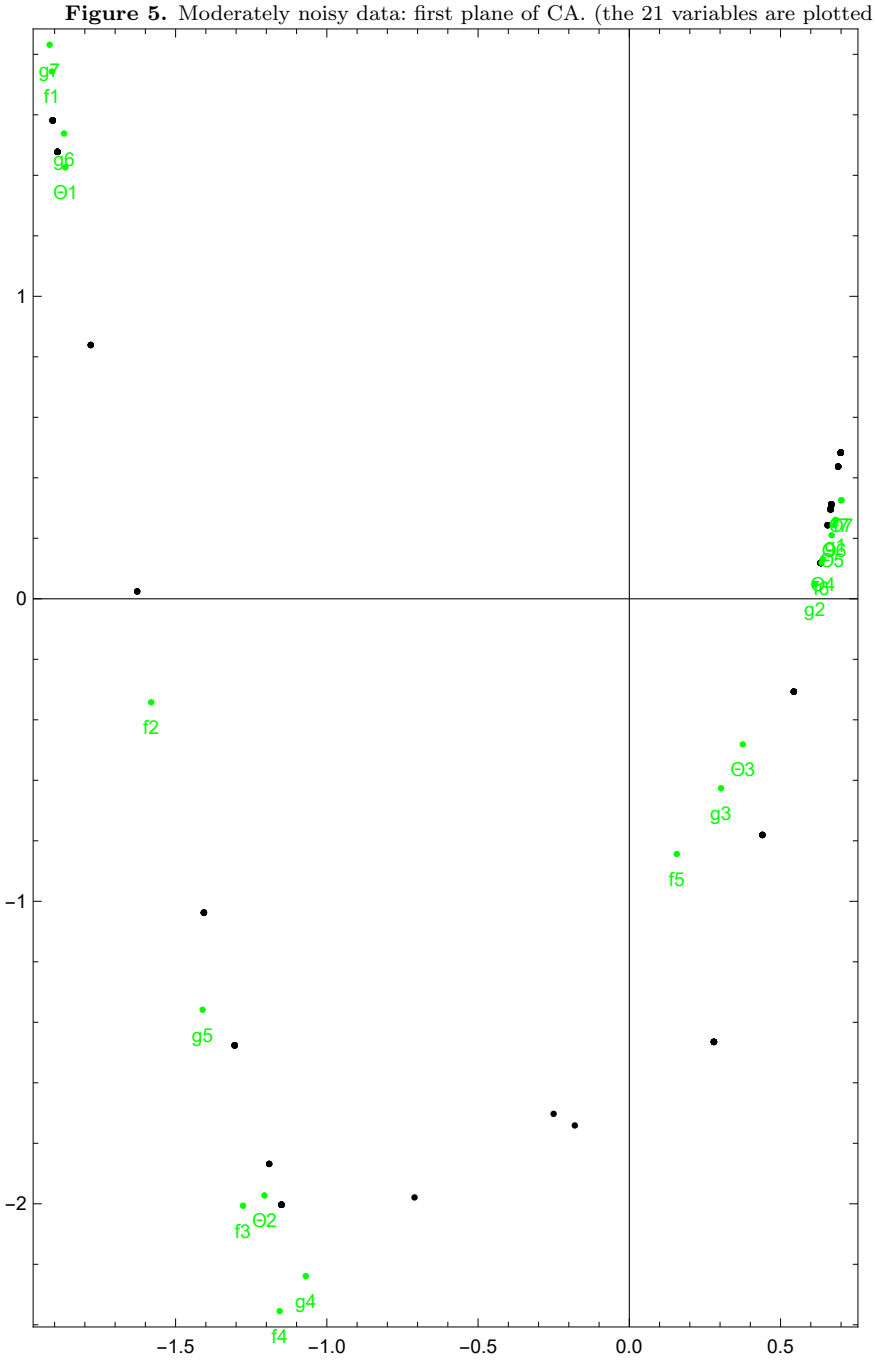

Figure 6. Very noisy data: $\sigma = 3$. Same structure as Figure 2. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel (in black): strongly significant interactions (see Definition 6.2).

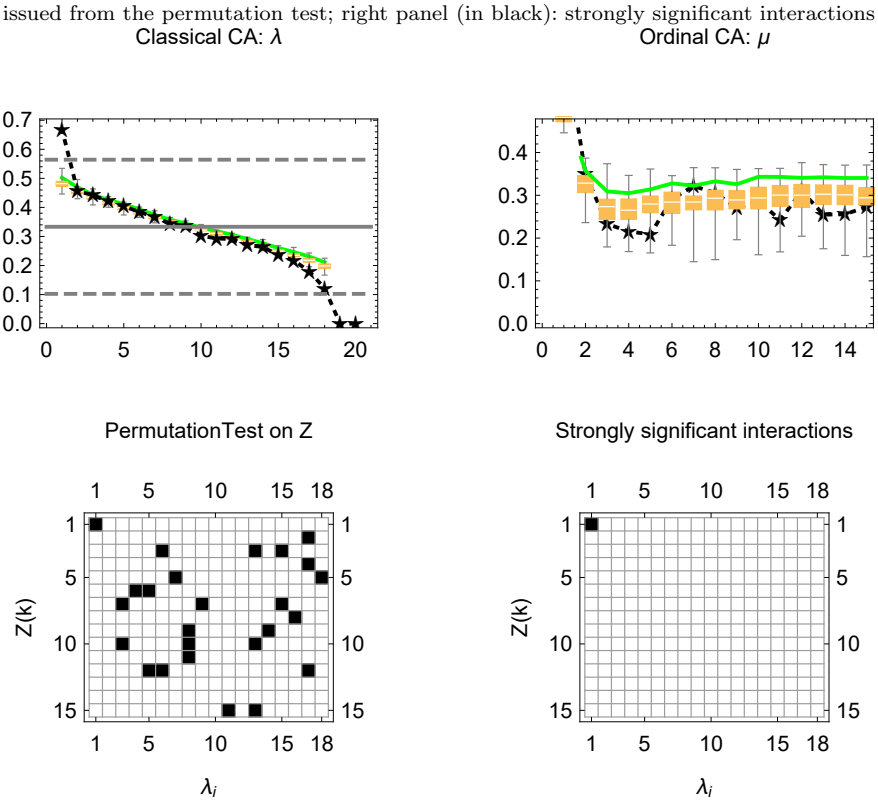

Figure 7. Very noisy data: first plane of CA. (the 21 variables are plotted in green/gray).

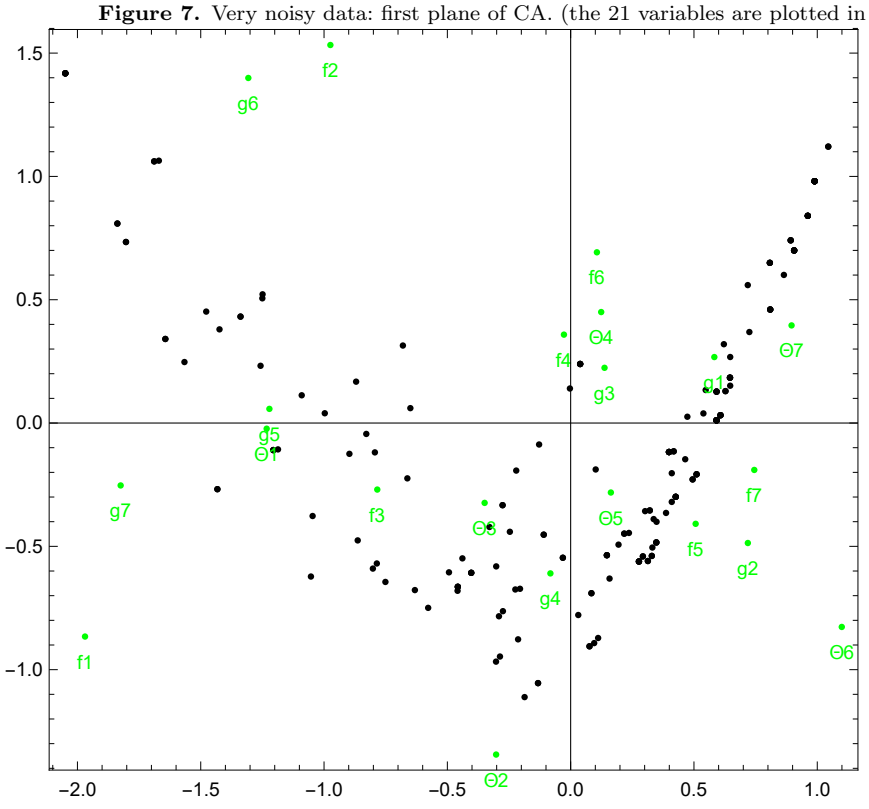

Figure 8. The archaic Chinese vases data. Same structure as Figure 2. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel (in black): strongly significant interactions (see Definition 6.2).

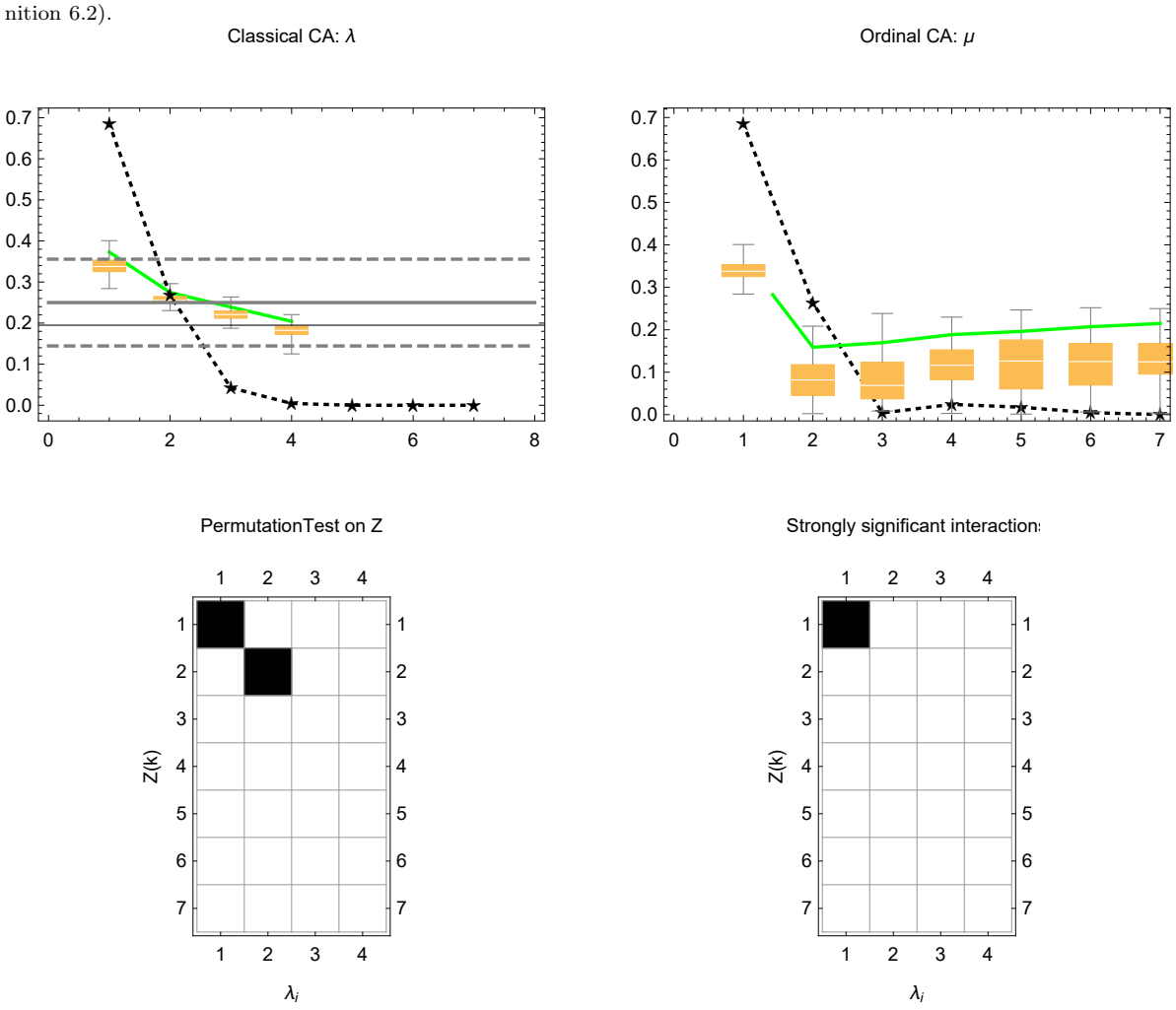

Figure 9. The archaic Chinese vases data: representation of the table of similarities between the sorted factors $\pi \odot \varphi_f$ and the $\mathfrak{P}_k$. On the left panel, black cells correspond to similarities greater than the fixed threshold (0.1). On the right panel, we plotted the values extracted from the diagonal of this table.

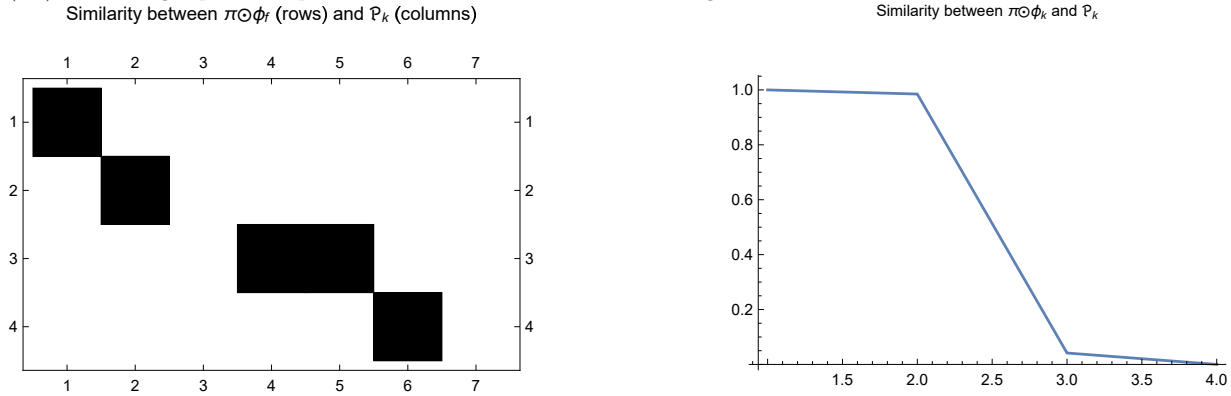

Figure 10. The X-rays diffraction data. Same structure as Figure 2. Upper plots: statistical significance of eigenvalues and polynomials coordinates. Lower plots, left panel: complete table of significant interactions (in black) issued from the permutation test; right panel (in black): strongly significant interactions (see Definition 6.2).

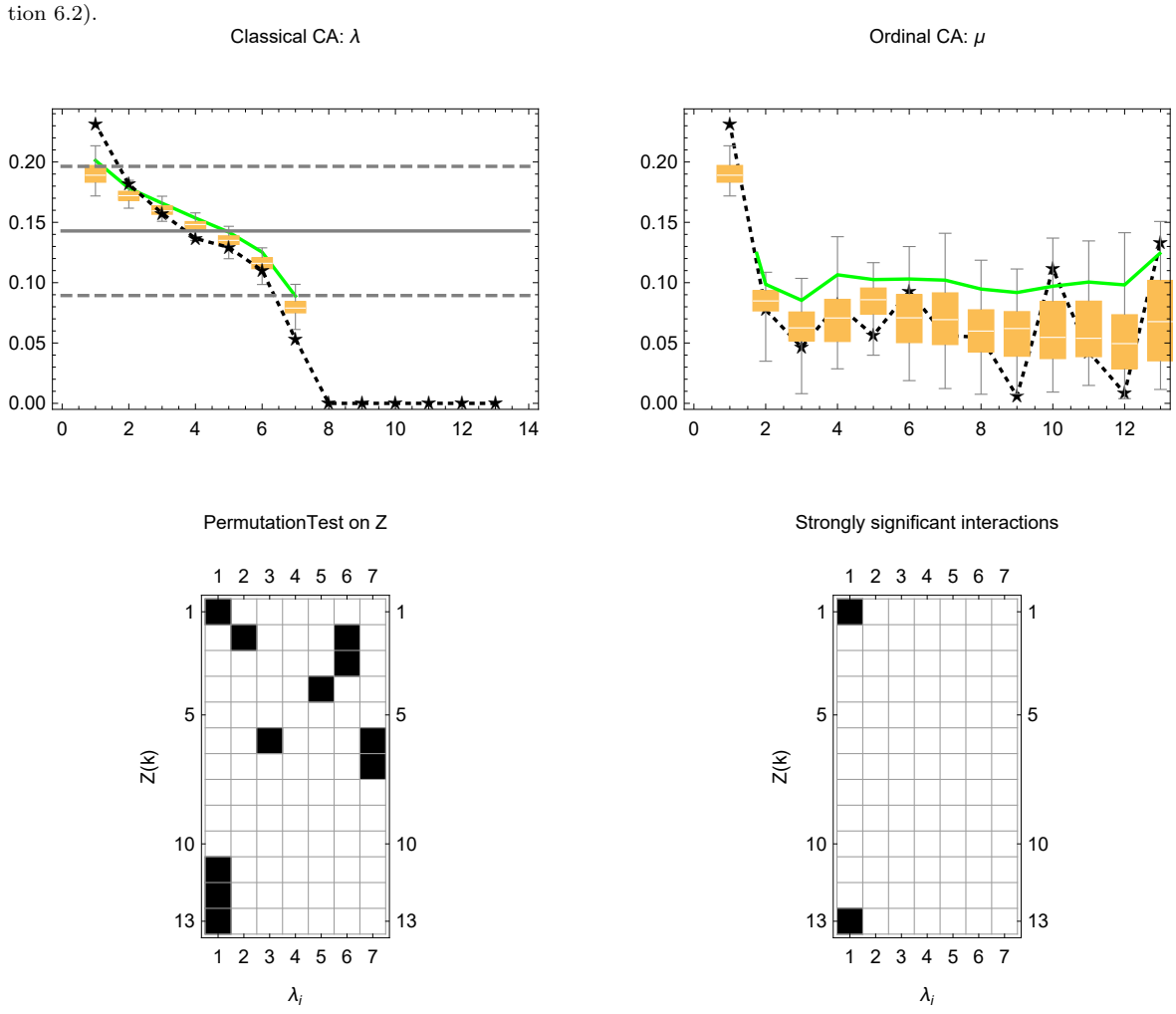

Figure 11. Coordinates of the minerals on the first plane of CA . Detection variables (+) are plotted horizontally; non-detection ones (-) are plotted vertically. Remember that the second dimension is not significant.