Received: 31 October 2022 | Revised: 27 July 2023 | Accepted: 31 July 2023

DOI: 10.1002/cem.3514

SPECIAL ISSUE - RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

# On the complementary nature of ANOVA simultaneous component analysis (ASCA+) and Tucker3 tensor decompositions on designed multi-way datasets

**Farnoosh Koleini**[1] | **Siewert Hugelier**[2] | **Mahsa Akbari Lakeh**[3] | **Hamid Abdollahi**[3] | **José Camacho**[4] | **Paul J. Gemperline**[1]

[1]Department of Chemistry, East Carolina University, Greenville, North Carolina, USA

[2]Department of Physiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran

[4]Department of Signal Theory, Network, and Communication, Research Centre for Information and Communication Technologies (CITIC-UGR), University of Granada, Granada, Spain

**Correspondence**
Paul J. Gemperline, Department of Chemistry, East Carolina University, Greenville, NC, 27858, USA.
Email: gemperlinep@ecu.edu

**Abstract**

The complementary nature of analysis of variance (ANOVA) Simultaneous Component Analysis (ASCA+) and Tucker3 tensor decompositions is demonstrated on designed datasets. We show how ASCA+ can be used to (a) identify statistically sufficient Tucker3 models; (b) identify statistically important triads making their interpretation easier; and (c) eliminate non-significant triads making visualization and interpretation simpler. For multivariate datasets with an experimental design of at least two factors, the data matrix can be folded into a multi-way tensor. ASCA+ can be used on the unfolded matrix, and Tucker3 modeling can be used on the folded matrix (tensor). Two novel strategies are reported to determine the statistical significance of Tucker3 models using a previously published dataset. A statistically sufficient model was created by adding factors to the Tucker3 model in a stepwise manner until no ASCA+ detectable structure was observed in the residuals. Bootstrap analysis of the Tucker3 model residuals was used to determine confidence intervals for the loadings and the individual elements of the core matrix and showed that 21 out of 63 core values of the $3 \times 7 \times 3$ model were not significant at the 95% confidence level. Exploiting the mutual orthogonality of the 63 triads of the Tucker3 model, these 21 factors (triads) were removed from the model. An ASCA+ backward elimination strategy is reported to further simplify the Tucker3 $3 \times 7 \times 3$ model to 36 core values and associated triads. ASCA+ was also used to identify individual factors (triads) with selective responses on experimental factors A, B, or interactions, A × B, for improved model visualization and interpretation.

## 1 | INTRODUCTION

Tensor decompositions were invented by Hitchcock in 1927, and the multiway model was invented by Cattell in 1944. These ideas received little attention until Tucker's work in the 1960s and Carroll, Chang, and Harshman's work in

1970, which all appeared in the psychometrics literature. Tensor decompositions were reportedly used for the first time in the field of chemometrics by Appellof and Davidson in 1981, and have since then grown in popularity[1,2] across various disciplines including signal processing, computer vision, data mining, graph analysis, neuroscience, and more. Additionally, there are numerous software packages that can be used to work with tensors.[2] Recently, bootstrap methods for obtaining uncertainty estimates in the form of confidence intervals for all parameters resulting from tensor decompositions (CANDECOM/PARAFAC or Tucker3) have been developed.[3]

In designed experiments where a multivariate dataset is generated, the design of the experiment as well as the relationship between the different variables should be considered, as both are interesting and can help to understand the system under study and the underlying variation in the dataset. ANOVA simultaneous component analysis (ASCA) was introduced as an exploratory tool for the analysis of multivariate datasets with an underlying experimental design and to quantify the statistical significance of the experimental factors by determining *p*-values through the different permutations and bootstrap methods.[4]

High dimensional datasets with an underlying experimental design of at least two factors in multiple levels can be folded into a tensor form which can be analyzed by a Tucker3 tensor decomposition if at least one of the factors has common samples or subjects. Therefore, ASCA and Tucker3 are complementary to each other, as they can be used to analyze the same kind of datasets. We use this novel combination of ASCA+ and Tucker3 models to revisit the Tucker3 analysis published in an original report[5] and illustrate how this combination of ASCA+ and tensor decompositions can be used to gain insights into the statistical significance of various factors and loadings in the tensor decomposition.

## 1.1 | Experimental methods

Eastern North Carolina's Pamlico River is a significant commercial source of blue crabs (*Callinectes sapidus*). In 1986, there was a cause for concern as the appearance of diseased crabs with lesions of 5 to 25 mm penetrating the carapace of the crabs was observed. Interestingly. diseased crabs were being caught in greater numbers near a phosphate strip mine.[5] A similar issue was discovered in the Saint Johns River near a phosphate strip mine in Florida.[6] At that time, the operator of the mine had a permit to discharge up to 20 ppm fluoride into the river, which was mixed with large quantities of groundwater pumped from the perimeter of the strip mine to depressurize the aquifer.[6] In a study by Gemperline et al., it was hypothesized that environmental stress because of this discharge weakened the organism so that its normal immunological response was unable to ward off opportunistic infection by chitinoclastic bacteria. Knowing that fluoride ions can form water-soluble complexes with many minerals that are insoluble at normal river pH, a study of trace elements in crab tissue samples was conducted.[5]

In October and November of 1989, gill, muscle, and hepatopancreas tissue samples were taken from 16 blue crabs in each of three groups: Albemarle, diseased Pamlico, and non-diseased Pamlico (48 crabs in total; equal samples for each group) to study whether trace element levels might be associated with the occurrence of the disease. Twenty-eight elements including Ag, Al, As, Be, Ca, Cd, Co, Cr, Cu, Fe, K, Li, Mg, Mn, Mo, Na, Ni, P, Pb, Se, Si, Sn, Ti, Tl, U, V, Y, and Zn were measured in the digested tissue samples by inductively coupled plasma atomic emission spectroscopy (ICP-AES).[5] The dataset was arranged into a three-way array of 48 individuals $\times$ 25 elements $\times$ 3 tissue samples. The elements, Tl, Be, and Y were excluded as the concentrations of these elements were at or below the detection limit. In the original paper, a three-mode principal component analysis (PCA) analysis was used to construct a Tucker3 model of rank $4 \times 5 \times 2$ orthogonal basis vectors and was used to visualize clusters of elements and crabs. In a subsequent paper, a three-mode mixture method of clustering analysis was performed[7] and confirmed the existence of the clusters that were only "visually" observed in the original report.[5]

## 1.2 | ANOVA simultaneous component analysis

The dataset considered in this paper follows a three-factor nested design with subjects (crabs-Factor C) nested in a disease state/region (Factor A). Three tissue types, muscle, hepatopancreas, and gill were sampled from each crab (Factor B). In this paper, we use a recently published ASCA+[8,9] as implemented in the MEDA toolbox that performs permutation analysis of unbalanced nested designs to determine the statistical significance of experimental factors and their interactions.[10] ASCA is particularly useful for determining the significance of one or more factors in designed experiments by separating the variance attributable to the effects of experimental factors, typically a treatment or an

experimental condition, and their interactions.[11] In a typical nested ANOVA (also known as hierarchical ANOVA), the values of individuals (in our case, blue crabs, Factor C) are found in combination with only one value of the higher-level factor (Factor A, disease state/region). The lower-level subgroupings must be treated as random effects variables, meaning they are random samples of a larger set of possible subgroups.[12]

Summarizing the experimental design of this dataset gives the following:

1. Factor A: Disease state/region, three levels (Diseased Pamlico, Healthy Pamlico, and Albemarle control); a fixed factor (A) that measures the variance over different disease state/regions.
2. Factor B: Tissue type, three levels (gill, hepatopancreas, and muscle); a fixed factor (B) that measures the variance over different tissues
3. Factor C: subject factor (crabs) nested in Factor A, disease state/region, noted in the remaining of the paper as **C**(**A**), a random factor that measures the inter-subject variance nested in Factor A.
4. Interaction A×B: noted as **AB**, which represents the extent to which regions cause a differential evolution over the tissue between the Diseased Pamlico, Healthy Pamlico, and Albemarle control groups.

In matrix notation, the $n \times m$ dataset $\mathbf{X}$ of measurements can be decomposed as follows using ASCA+:

$$\mathbf{X} = \mathbf{1}\mathbf{m}^{\mathrm{T}} + \mathbf{A} + \mathbf{B} + \mathbf{C}(\mathbf{A}) + \mathbf{AB} + \mathbf{R} \tag{1}$$

where **1** is a vector of ones of suitable length, $m$ represents the overall mean, and **A**, **B**, and **C**(**A**) represent the factor or effect matrices, **AB** the interaction matrix, and **R** the residual matrix. In this paper, we use the technique referred to as ASCA+[9] as implemented in the MEDA toolbox[10] to account for the study's unbalanced data. In ASCA+, the original ASCA methodology is extended to unbalanced designs by using general linear models (GLMs) to estimate the effect matrices, instead of the classical ANOVA estimators based on differences in means.[8,9]

Simultaneous component analysis (SCA) was then performed on the individual effect matrices to model and visualize the variability of each effect. In SCA, the different samples are modeled using PCA. Each of the matrices resulting from the ANOVA partitioning is decomposed as

$$\mathbf{X}_i = \mathbf{T}_i \mathbf{P}_i^{\mathrm{T}} + \mathbf{R} \tag{2}$$

where $\mathbf{T}_i$ and $\mathbf{P}_i^{\mathrm{T}}$ are the scores and loadings for the $i^{\mathrm{th}}$ partition, respectively, where a partition, $I$, represents an experimental factor or interaction, and $\mathbf{R}_i$ is the corresponding residual matrix. Factor A (disease state/region) and Factor B (tissue) in this study have three levels each, so the dimensionality of the PCA visualizations of the effect matrices **A** and **B** were constrained to rank two. Rank four PCA models were used to visualize the interaction matrix, **AB**.

ASCA is a supervised method where external knowledge about the experimental design is used and, as such, needs proper validation. To validate the significance of each factor/interaction matrix, unconstrained permutations on the original observations was used. This provides an approximate test[13] with better properties that other alternatives or even exact tests.[8] Permutation tests were performed by using 10,000 randomizations, where the $p$-value of the test is defined as the fraction of the permutations for which the employed metric was better than the unpermuted one. An effect is considered significant if its $p$-value is smaller than an appropriate significance threshold. In this work, residuals and tensors with $p$-values less than 0.05 were considered to be significant and have ASCA+ detectable structure at the 95% confidence level.[4,14]

Outlier detection is important when dealing with problems such as hypothesis testing, goodness of fit tests, regression, or classification techniques. In this study, 95% confidence ellipsoids of the mean centered and scaled original data were calculated for each experimental factor in ASCA, according to Zwanenburg et al.[4] Objects that were outside of the 95% confidence interval using Hoteling's $T^2$ distribution were considered to be outliers. Details are provided in Section 2.

## 1.3 | ASCA+ analysis of the Tucker3 residuals

ASCA+ was used in a novel way to determine the significance of the Tucker3 models. The alternating least squares algorithm TuckerALS with orthogonality constraints was used to construct Tucker3 models,[4] where three matrices of

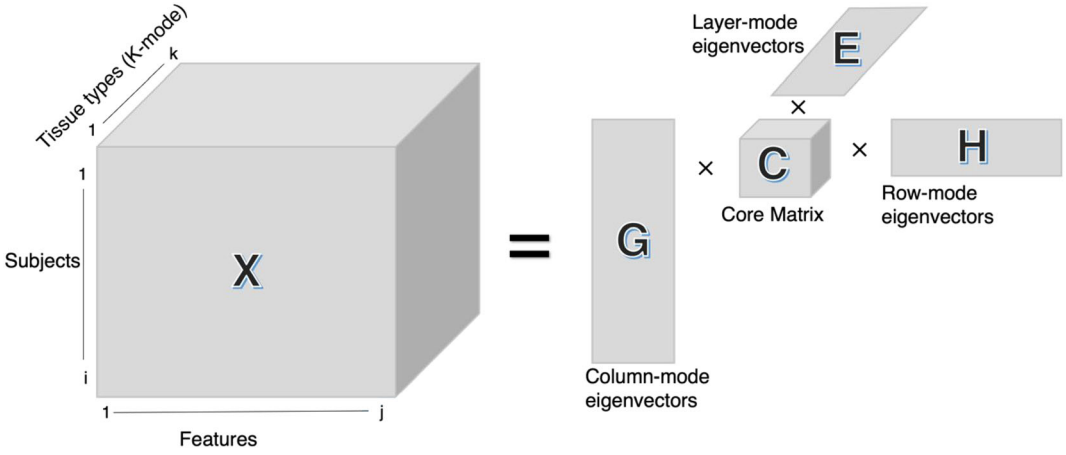


**FIGURE 1** Diagram of the Tucker3 model of the dataset.

eigenvectors are computed (orthonormal loading vectors), one for each dimension in the original data table (see Figure 1). Equation (3) shows the Tucker3 model for a three-way array **X**, where $x_{ijk}$ are the individual values of the tensor; $I$, $J$, and $K$ represent the original dimension of the tensor (in this case, $48 \times 25 \times 3$); $P$, $Q$, and $R$ represent the number of factors selected for eigenvectors **G**, **H** and **E** (in this case, $3 \times 7 \times 3$) and $c_{pqr}$ is an element of the core matrix, **C**, a $3 \times 7 \times 3$ tensor. The sum of the squared core values are analogous to eigenvalues in two-way PCA, equal to the total variance explained by the model.[15,16]

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} c_{pqr} \left( g_{ip} h_{jq} e_{kr} \right) + \varepsilon_{ijk} \tag{3}$$

The total variance can be partitioned into two parts according to Equation (4).

$$SS_{\text{total}} = SS_{\text{fit}} + SS_{\text{residual}} \tag{4}$$

where $SS_{\text{fit}}$ is the sum of squares explained by the three-mode model, and $SS_{\text{residual}}$ is the residual sum of squares. In the previous work,[5] for Tucker3 analysis, four factors in the first mode ($P = 4$), five factors in the second mode ($Q = 5$), and two factors in the third mode were used ($R = 2$). The original model selection was accomplished by comparing the variance explained by models of different complexity, preserving 70.66% of the variance in the original dataset. However, to determine if this model adequately explains all the experimental factors and their interactions in this dataset, ASCA+ analysis was performed on the Tucker3 residuals. Surprisingly, the ASCA+ results showed that the main factors, A, B, and the interactions, AB, were statistically significant in the $4 \times 5 \times 2$ model residuals, indicating that an insufficient number of factors were selected. Details are discussed in Section 2.

## 1.4 | Bootstrap analysis

There are three major strategies for performing bootstrap analysis: the parametric bootstrap, resampling of residuals, and resampling cases or whole data points.[17] In this work, we used the resampling of residuals approach. In this approach, the Tucker3 model is estimated using the original data, and then bootstrap samples are obtained by resampling the residuals with replacement and adding them back to the model estimated values. This strategy assumes that the model is correct, and the distribution of the residuals is consistent from individual to individual. This is different from the strategy of Kiers[3] which assumes that the entities in the first mode are a random sample from a population of such entities. The Kiers method uses resampling of cases (rows) from **X** with replacement to produce "pseudo populations", $\mathbf{X}_b$. In the case of the blue crab data, resampling rows of **X** would disrupt the experimental design, for

example, the original structure of the three different disease state/region populations represented in the designed dataset. Instead, in this work, the Tucker3 residuals are resampled with replacement and added to the estimated Tucker3 model. By resampling the residuals, it is presumed that the Tucker3 model used is adequate and that the distribution of residuals from individual cases or objects is the same.

### 1.5 | Software

ASCA+ analysis was performed using MEDA toolbox (https://github.com/josecamachop/MEDA-Toolbox), and Tucker3 modeling was computed with MATLAB software written at ECU. The Tucker3 code is available from the corresponding author upon request.

## 2 | RESULTS AND DISCUSSION

### 2.1 | ASCA analysis on the raw data

The Mahalanobis distance method was used to detect outlier samples based on ellipsoids constructed at the 95% confidence interval using the ASCA score plots for each main factor and the interactions by back-projecting the ASCA residuals on the loadings of the ASCA model.[14] The score plots of ASCA on the mean-centered and scaled dataset with their 95% confidence intervals ellipsoids are shown in Figure 2. When the Mahalanobis distances and the sample probability densities based on Hoteling's $T^2$ were calculated, rows 17, 36, and 128 appear to be outlier objects in both Factor A (disease state/region) and Factor B (tissue), see Figure 2. These results are summarized in Table 1, listing objects with a probability density of less than 0.05. Only outliers common to both Factor A and Factor B were selected (see bold face entries in Table 1).

ASCA+ as described in Section 1.1 was used to assess the significance of the underlying factors and their interactions. Although the underlying experimental design was unbalanced because of outlier removal, the ParGLM function in the MEDA toolbox is able to accommodate unbalanced designs using GLMs to estimate the effect matrices.[8–10] The amount of sum of squares, degrees of freedom, $F$ ratios, and $p$-values for the different factors the blue crab dataset is reported in Table 2. The total variance preserved was 81.85%.

Inspection of the table shows that tissue (Factor B) is the largest effect accounting for more than 43.6% of the modeled variation, whereas the disease state/region (Factor A) is a much smaller effect (11.3% of the modeled variance). Moreover, it is important to note that 18.15% of the total variance is not explained by the ASCA+ model and corresponds to the response differences among the different replicates. In general, the ASCA+ results reported in Table 2

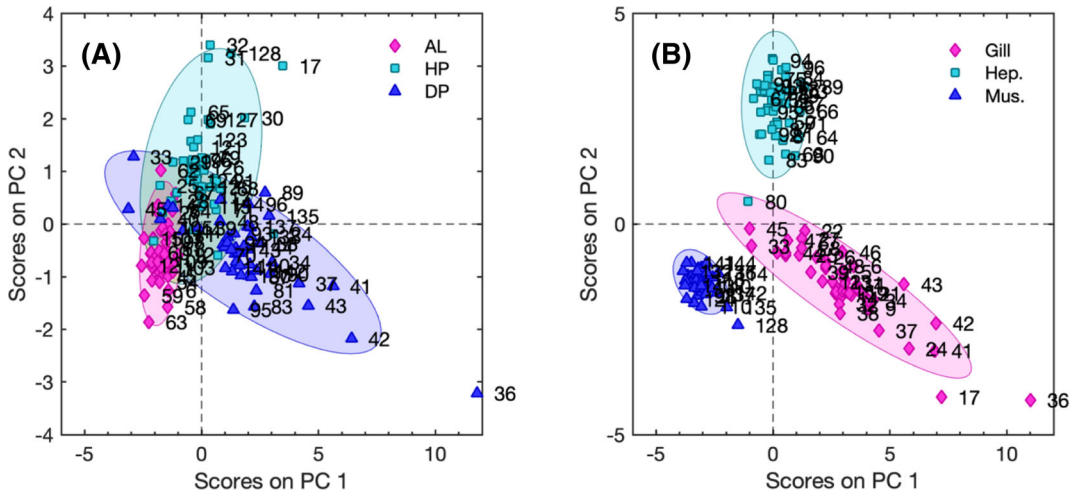


FIGURE 2 Score plots of ASCA on the raw mean centered and scaled data. Left: score plot for factor A, (disease state/region) $\mathbf{X}_a$, right: score plot for Factor B (tissue), $\mathbf{X}_b$.

**TABLE 1** Outliers detected using the Mahalanobis distance and probability density (only samples with a probability density < 0.05 are shown) from ASCA score plots of factor A (disease state/region), and Factor B (tissue).

| Factor A (disease state/region) | | | Factor B (tissue) | | |
|---|---|---|---|---|---|
| **Sample number** | **Mahalanobis distance** | **Probability density** | **Sample number** | **Mahalanobis distance** | **Probability density** |
| **17** | **3.5541** | **0.0041** | **17** | **3.1955** | **0.0108** |
| **36** | **3.7672** | **0.0023** | 31 | 2.7199 | 0.0346 |
| 43 | 2.7484 | 0.0324 | 32 | 2.9414 | 0.0205 |
| 80 | 3.8919 | 0.0016 | **36** | **4.3383** | **0.0004** |
| **128** | **4.1756** | **0.0007** | **128** | **2.1696** | **0.0037** |
| 135 | 2.7250 | 0.0342 | | | |

**TABLE 2** ASCA+ sum of squares, degrees of freedom, *F* ratios, and *p*-values (10,000 permutations) for the different effects in the mean centered and scaled blue crab dataset with outliers removed.

| Source | SSQ | df | *F*-ratio | *p*-value |
|---|---|---|---|---|
| Disease state/Region, A | 379 | 2 | 17.1 | 0.0001 |
| Tissue, B | 1,461 | 2 | 100.9 | 0.0001 |
| Individuals C(A) | 466 | 42 | 1.5 | 1.0000 |
| Interaction A × B | 456 | 4 | 15.7 | 0.0001 |
| Residuals | 608 | 84 | | |
| Total | 3350 | 135 | | |

indicate that all the factors and the interactions are large. To determine whether these differences were statistically significant, permutation tests with 10,000 randomizations were performed. As shown in Table 2, both the main effects and their interaction are statistically significant ($p < 0.05$), indicating that the interaction between disease state/region and tissue type is statistically significant such that the concentration of the trace elements in the various tissues is dependent on the population measured.

## 2.2 | ASCA+ analysis of the Tucker3 residuals

Tucker3 models were constructed using the TuckerALS method as noted in the method section. The core matrix values associated with each triad of eigenvectors represents the total variance explained by the corresponding triad (see Figure 1). For Tucker3 models, the total variance can be partitioned into two parts: the sum of squares explained by the model and the residual sum of squares. In the original paper, a 4 × 5 × 2 Tucker3 model was used to explain 70.66% of the variance in the dataset. When ASCA+ was performed on the residual matrix, it was determined that there was ASCA+ detectable structure in the residuals, that is, the residuals still contained structure that could be associated with the main factors and their interactions. Table 3 shows the amount of sum of squares, degrees of freedom, *F* ratios, and *p*-values for the different factors of the blue crab 4 × 5 × 2 Tucker3 residual model dataset.

This indicates that the 4 × 5 × 2 model does not sufficiently explain the main factors and interactions. Figure 3 shows the resulting score plots obtained by ASCA+ on the Tucker3 residuals. This figure clearly shows that there is still a certain degree of separation between clusters of populations in the residuals, a result that was not expected.

To find the Tucker3 model that on the one hand uses a sufficient number of factors in each mode to explain all the variation in the dataset and, on the other hand, is as parsimonious as possible, a grid search strategy was employed for all possible combinations of Tucker3 models with 1 to 10 factors for *P*, 1 to 10 factors for *Q*, and 1 to 3 factors for *R*. Models that did not meet the Kruskal rank criterion[18] for uniqueness were skipped. For each combination of factors, the resulting residual matrix was tested for significance using ASCA+ with 10,000 permutations. Using this approach, we concluded that the most parsimonious model that explained all the contributions of the experimental factors in the

**TABLE 3** ASCA+ sum of squares, degrees of freedom, *F* ratios, and *p*-values (10,000 permutations) for the different effects of the residuals of the 4 × 5 × 2 Tucker3 model of the blue crab dataset with outliers removed.

| Source | SSQ | df | F | *p*-value |
|---|---|---|---|---|
| Mean | 105.6 | 1 | | |
| Disease state/Region, A | 150.4 | 2 | 10.7 | 0.0001 |
| Tissue, B | 24.8 | 2 | 3.3 | 0.0135 |
| Individuals C(A) | 294.1 | 42 | 1.9 | 0.1051 |
| Interaction A × B | 93.3 | 4 | 6.3 | 0.0001 |
| Residuals | 311.6 | 84 | | |
| Total | 982.9 | 135 | | |

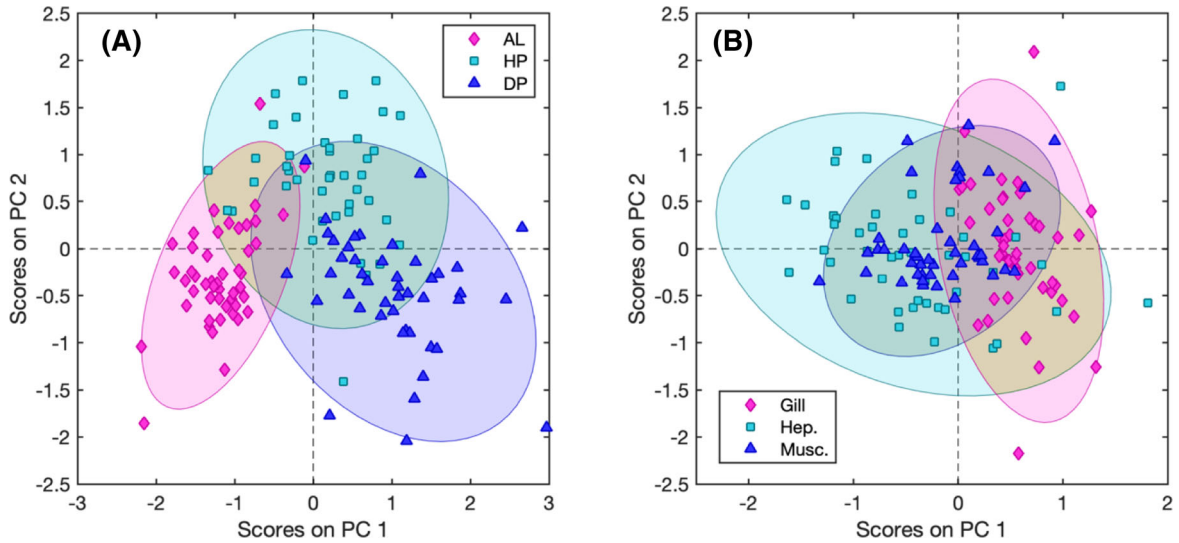


**FIGURE 3** Score plots of ASCA+ on the 4 × 5 × 2 Tucker3 model residuals. Left: score plot for Factor A, (disease state/region) $\mathbf{X}_a$, right: score plot for Factor B (tissue), $\mathbf{X}_b$.

dataset was the 3 × 7 × 3 Tucker3 model with residual variance of 21.97%, which is similar compared to the explained variance by ASCA+ (18.15%). The results obtained by ASCA+ analysis of the 3 × 7 × 3 Tucker3 residuals show that the *p*-value for factors A, B, and A × B were larger than 0.05 ($p = 0.883$, $p = 1$, and $p = 0.953$, respectively), indicating there was not any ASCA detectable structure in the residuals.

Figure 4 shows a plot of the residuals by element and by tissue type from the 3 × 7 × 3 Tucker3 compared to 4 × 5 × 2. In the 3 × 7 × 3 Tucker3 model (bottom panel), the distribution of the residuals for each variable in all three tissue types is symmetrical with a mean of zero, whereas in the 4 × 5 × 2 Tucker3 model (top panel), the distribution of the residuals still has structure (some are non-symmetrical) and many of the means are not zero.

In summary, the 3 × 7 × 3 model explains 78.03% of the dataset's variance and the ASCA+ model explains 81.85%. This difference is small; thus, it is unlikely that there is sufficient variation remaining in the ASCA residual matrix or the 3 × 7 × 3 model residuals that could be relevant for interpretation.

## 2.3 | Bootstrap analysis

A bootstrap analysis (10,000 randomizations) was performed to determine the significance of the Tucker3 core values. Our bootstrap method consisted of resampling the model residuals with replacement and adding them to the model estimated dataset. It is well-known that a sign ambiguity and ordering ambiguity exist in the core values and in corresponding triads of eigenvectors or loadings in Tucker3 models.[1] We also observed this ambiguity in the bootstrap

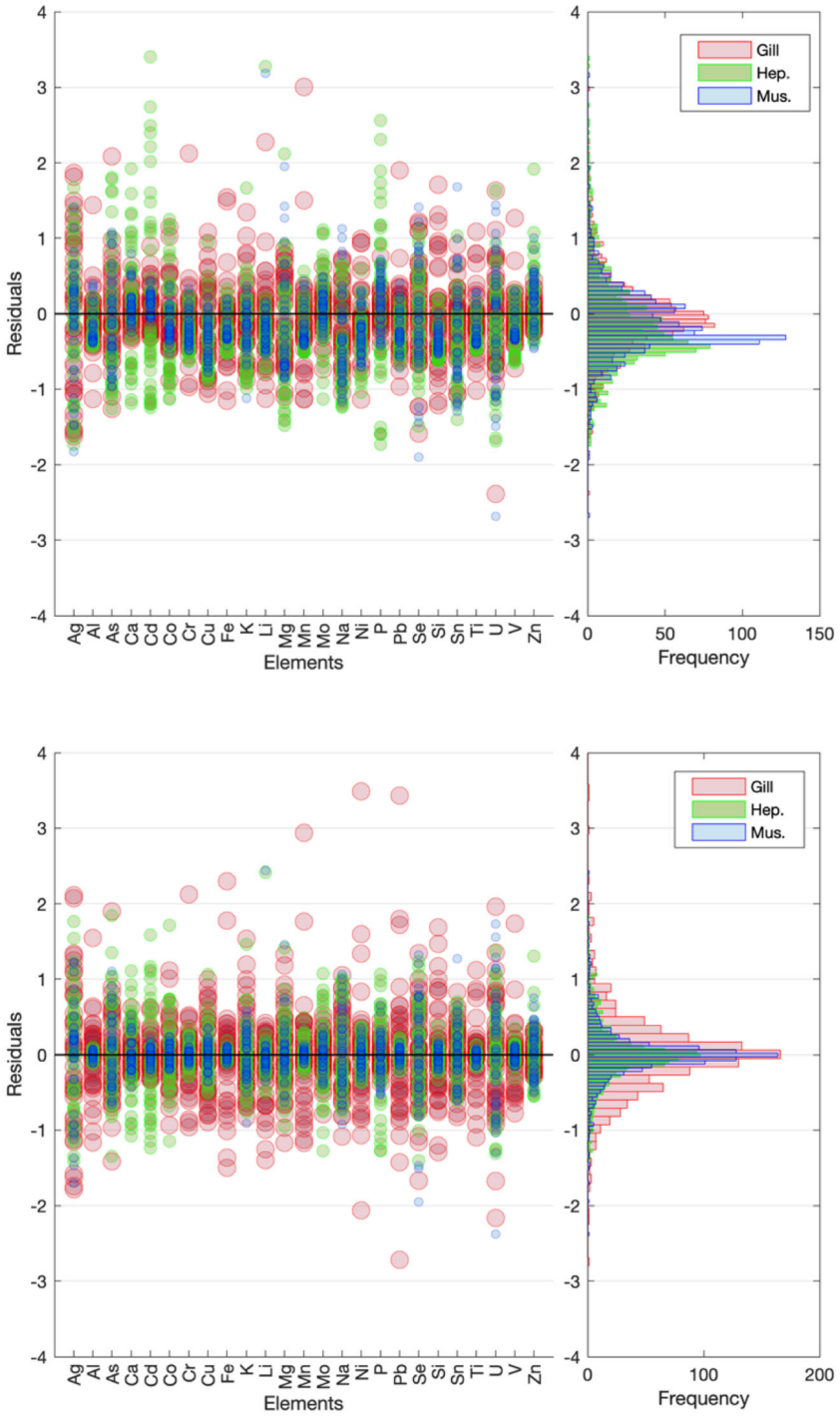


**FIGURE 4** Distribution of the residuals for each variable in all three tissue types, top: 4 × 5 × 2 Tucker3 model, bottom: 3 × 7 × 3 Tucker3 model.

analysis used in this study. To correct for the shuffling of core values and columns of eigenvectors in the bootstrap models and to speed up the calculations, we used the initial non-bootstrapped solution as a reference model and the starting point for the TuckerALS algorithm. The resulting Tucker3 models of bootstrap samples were sorted to match the reference model according to the following procedures. First, correlation analysis was used to determine whether the loadings were in the same order as the reference loadings for each of the three modes, starting with **G**, followed by **H**, and then **E**. Simultaneously, the corresponding core values were reordered to match. Additionally, we maintained the sign parity of each combination of four values, $c_{ijk} \times \mathbf{g}_i \times \mathbf{h}_j \times \mathbf{e}_k$, by systematically cycling through the full set of

core values and flipping the sign $c_{ijk}$ of the bootstrap model when necessary to match the reference model, followed by flipping the sign of $\mathbf{g}_i$. In this manner, we ensure that the original solution's order and algebraic sign are matched in the model of the bootstrap sample, $\mathbf{X}_b$, without having to implement an extensive bookkeeping strategy. As an example, Figure 5 shows the bootstrap distribution for core value $c_{212}$ before and after correction.

Confidence intervals of the resulting bootstrap loading matrices, **G, H, E,** and the core values were then computed. The 95% confidence interval was determined by sorting the bootstrap objects and identifying the upper 2.5% and lower 2.5% of the distribution. As an example, Figure 6 shows the distribution of the core values of the first three factors. The cyan region of the histogram lines defines the 95% confidence interval, and the solid line represents the value of the core element in the reference model. Examination of the distributions shown in Figure 6 represents $c_{111}$ (left panel) and $c_{211}$ (middle panel), and reveals that these two core elements are statistically significant at the 95% confidence level, as the value of 0 is not included in the interval. On the other hand, the histogram of the distribution of $c_{311}$ (right panel) clearly illustrates that the estimated value of the core matrix is not significantly different from zero and. Therefore, does not significantly contribute to the Tucker3 model. This analysis was systematically carried out for all core values. Table 4 shows that 21 of the 63 core values are not statistically significant in the $3 \times 7 \times 3$ Tucker3 model.

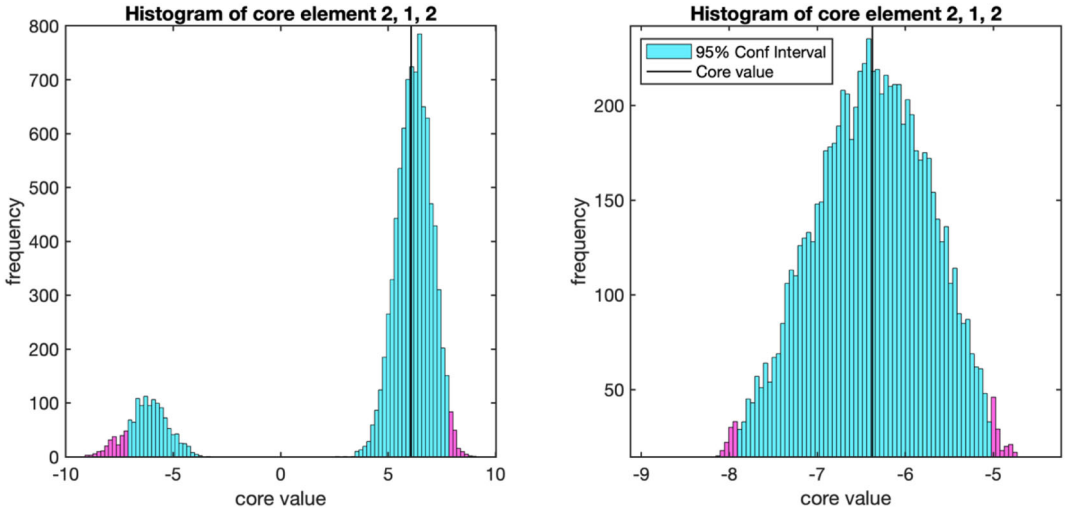


**FIGURE 5** The bootstrap distribution for core value $c_{212}$ before (left panel) and after (right panel) sign flipping correction. Cyan areas indicate core element values within the 95% confidence interval, magenta areas indicate core element values outside the 95% confidence interval, and the solid line indicates the value of the core element of the original reference model (before bootstrapping).

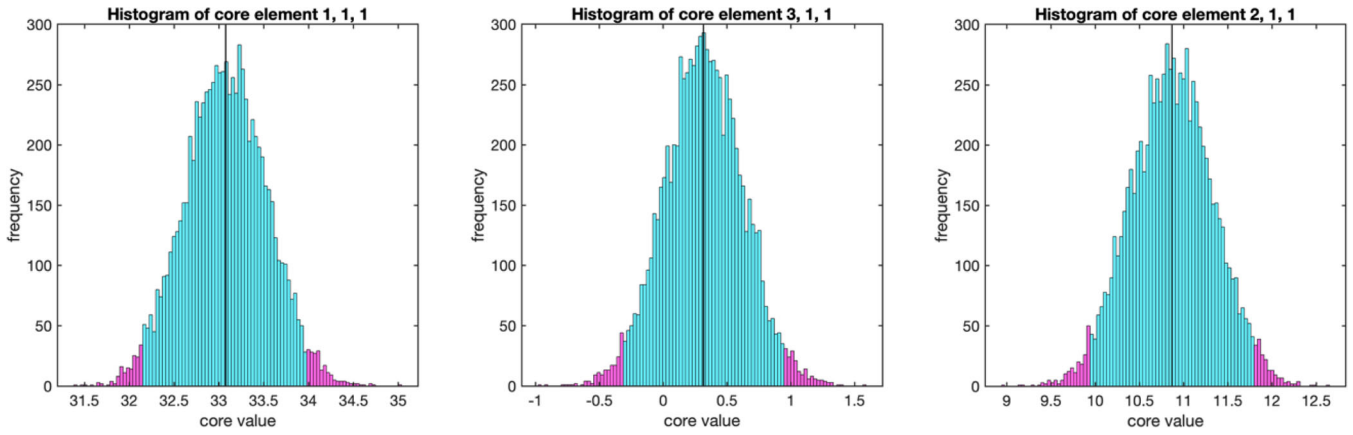


**FIGURE 6** Distribution histograms (frequency vs core element value) for the null hypothesis obtained by bootstrap analysis of selected core values. The cyan region is inside the 95% confidence interval, and the magenta region is outside the 95% confidence interval. The solid line shows the core value of the reference model.

**TABLE 4** Statistically insignificant core values determined by bootstrap analysis. $H_0$: $c_{ijk} = 0$. The 21 core values are sorted smallest to largest (out of 63) and are statistically not different from 0 (95% confidence level).

| Core value ($c_{ijk}$) | Explained variance (%) | $p = 1 - \alpha$, reject $H_0$: $c_{ijk} = 0$ | Core value ($c_{ijk}$) | Explained variance (%) | $p = 1 - \alpha$, reject $H_0$: $c_{ijk} = 0$ |
|---|---|---|---|---|---|
| 1, 3, 2 | 0.0348 | 0.06 | 1, 5, 1 | 0.0013 | 0.23 |
| 3, 3, 2 | 0.0302 | 0.08 | 3, 1, 1 | 0.0007 | 0.06 |
| 3, 7, 3 | 0.0226 | 0.09 | 2, 7, 1 | 0.0004 | 0.36 |
| 2, 5, 3 | 0.0180 | 0.30 | 1, 7, 1 | 0.0003 | 0.40 |
| 3, 6, 1 | 0.0123 | 0.26 | 2, 7, 3 | 0.0001 | 0.34 |
| 2, 4, 2 | 0.0074 | 0.17 | 2, 3, 3 | 0.0000 | 0.44 |
| 1, 6, 2 | 0.0071 | 0.28 | 1, 7, 3 | 0.0000 | 0.35 |
| 3, 2, 3 | 0.0065 | 0.19 | 2, 6, 2 | 0.0000 | 0.39 |
| 3, 6, 2 | 0.0050 | 0.40 | 3, 1, 2 | 0.0000 | 0.48 |
| 1, 5, 1 | 0.0039 | 0.22 | 3, 7, 1 | 0.0013 | 0.49 |
| 3, 1, 1 | 0.0026 | 0.15 | | | |

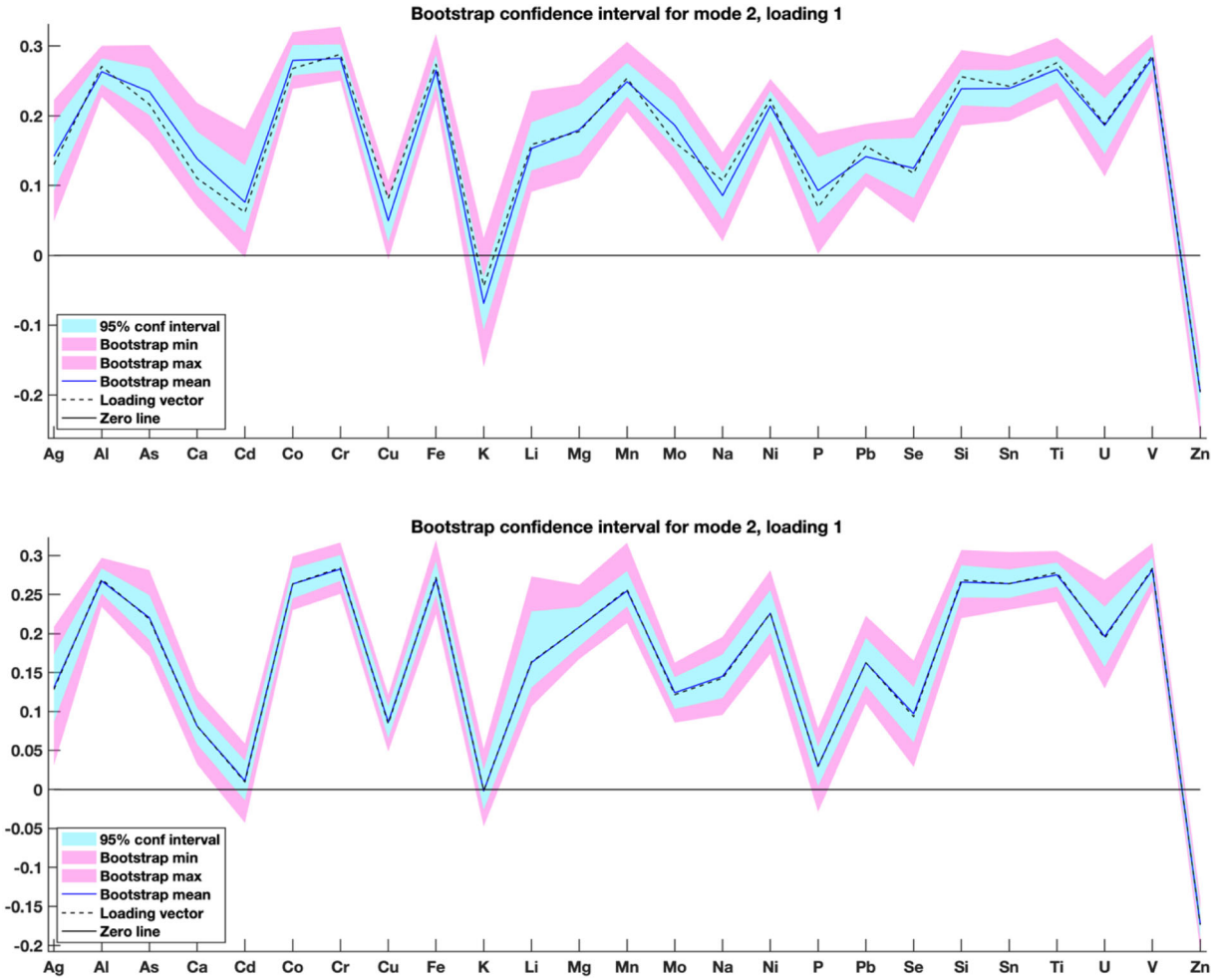


**FIGURE 7** Bootstrap confidence intervals for eigenvector $\mathbf{h}_1$, top: $4 \times 5 \times 2$ Tucker3 model, bottom: $3 \times 7 \times 3$ Tucker3 model.

Wanting to create a more parsimonious Tucker3 model, we sought to constrain small, non-significant core values to zero. However, applying this strategy, we observed that constraining even the smallest core value to zero completely changes the model. This can be explained because the TuckerALS algorithm uses orthogonality constraints, and thus, the core matrix must be three-way orthogonal. This guarantees a mathematically unique tensor decomposition, analogous to two-way PCA. When we constrain one of those values to zero, the orthogonality constraints must be relaxed such that the core matrix is no longer orthogonal. This changes the model's eigenvectors and their interpretation. However, noting that each of the 63 individual tensors obtained from the 63 combinations of triads are mutually orthogonal, we are justified in excluding the 21 non-significant core values and their associated triads (factors) from visualization and interpretation, giving a more parsimonious or simpler model containing only 42 triads out of 63 of the $3 \times 7 \times 3$ model. These removed core values account for only 0.13% of the dataset variance; thus, the variance modeled by the $3 \times 7 \times 3$ model is decreased from 78.03% to 77.90%.

## 2.4 | Interpretation of the model loadings

When comparing the $3 \times 7 \times 3$ Tucker3 model with $4 \times 5 \times 2$ model, the bootstrap confidence intervals for $\mathbf{h}_1$ are narrower with the $3 \times 7 \times 3$ model. This is because the residuals for the $3 \times 7 \times 3$ model are smaller with no ASCA detectable structure left in them, and thus, it is to be expected that the $4 \times 5 \times 2$ residuals give larger confidence intervals compared to the $3 \times 7 \times 3$ residuals (Figure 7). Eigenvectors associated with small core values are computed with greater uncertainty in the $4 \times 5 \times 2$ model, as can be seen in the confidence intervals. The shape of the first loading vector for the $4 \times 5 \times 2$ model compared to the $3 \times 7 \times 3$ model is slightly different, although the differences do not

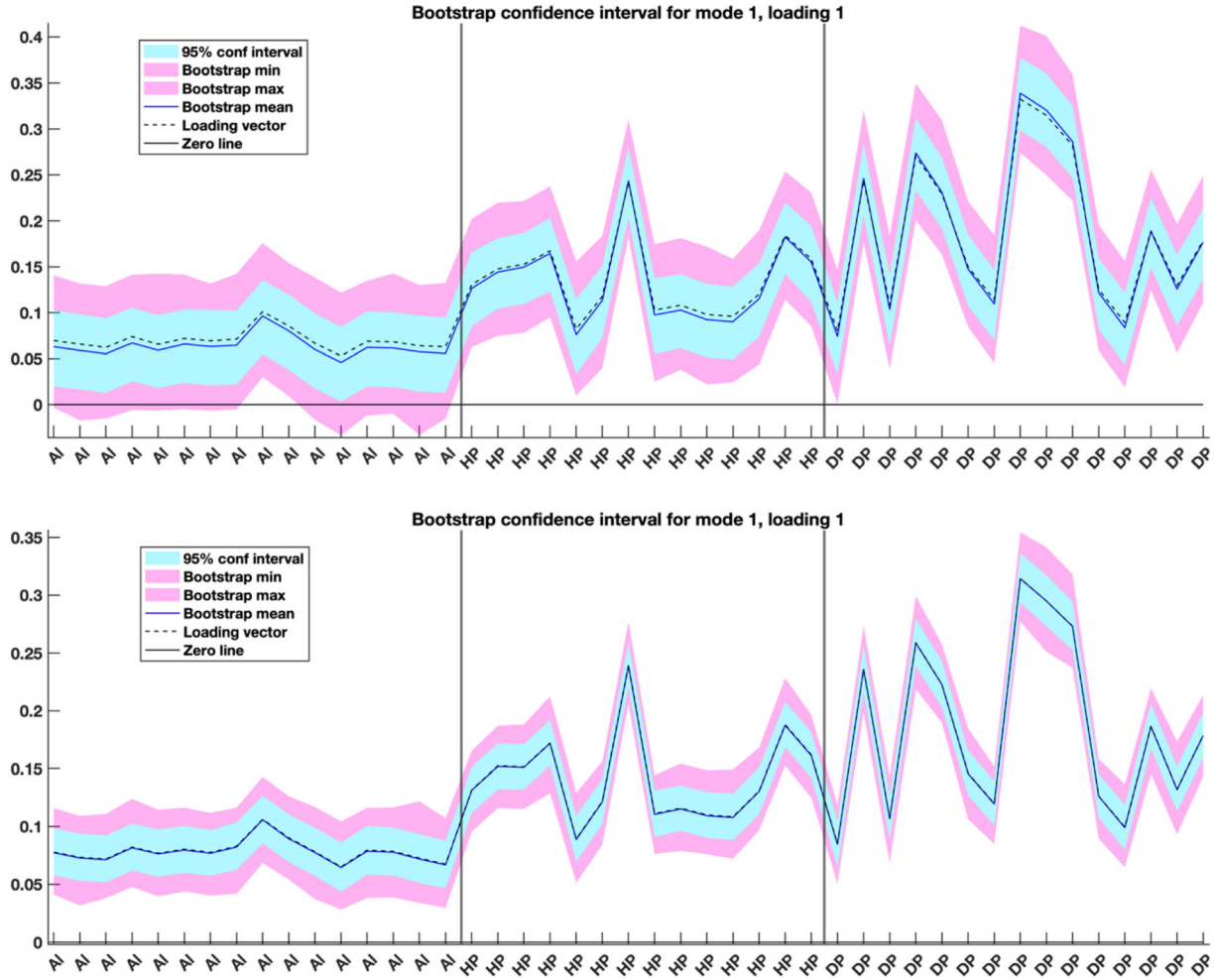


FIGURE 8 Bootstrap confidence intervals for eigenvector $\mathbf{g}_1$, top: $4 \times 5 \times 2$ Tucker3 model, bottom: $3 \times 7 \times 3$ Tucker3 model.

seem to be very large except for two of the elements, Mg and Mo. Observing $\mathbf{g}_1$ in the two models (Figure 8), again the $4 \times 5 \times 2$ bootstrap confidence intervals are much wider, and interestingly the mean bootstrap value is different than the original vector from the reference model, which indicates that the distribution of the bootstrap residuals is skewed in the $4 \times 5 \times 2$ model, whereas it is nearly symmetrical in the $3 \times 7 \times 3$ model. This suggests that the bootstrap confidence interval is approximately normally distributed in the $3 \times 7 \times 3$ model, whereas it is not in the $4 \times 5 \times 2$ model. Looking at the plots of the loadings for $\mathbf{g}_3$ (Figure 9), the pattern in the loadings for the $3 \times 7 \times 3$ model gives a much cleaner separation of Healthy Pamlico from Diseased Pamlico and Albemarle crabs, whereas it is more ambiguous for the $4 \times 5 \times 2$ model. Looking at the values of $\mathbf{e}_1$ for both models (Figure 10) shows that the values are similar in magnitude and shape, but the confidence interval is much narrower for the $3 \times 7 \times 3$ model, and the mean bootstrap value is different than the original vector from the reference model, which indicates that the distribution of the bootstrap residuals is skewed in the $4 \times 5 \times 2$ model, whereas it is nearly symmetrical in the $3 \times 7 \times 3$ model.

## 2.5 | Backward triad elimination procedure

As described above, bootstrap analysis was used to identify statistically (in)significant core values in the $3 \times 7 \times 3$ model. We next describe an ASCA+ backward elimination procedure to further reduce the complexity of Tucker3 models. In this procedure, triads are sequentially removed from the full model plus residuals, starting with the largest one first. The reduced model is then tested using ASCA+ with permutations to see if there is still detectable structure or variance because of factors A, B, or A × B in the reduced model. The result is shown in Table 5. Core values of the $3 \times 7 \times 3$ model are shown, ordered from largest variance explained along with ASCA+ $p$-values for

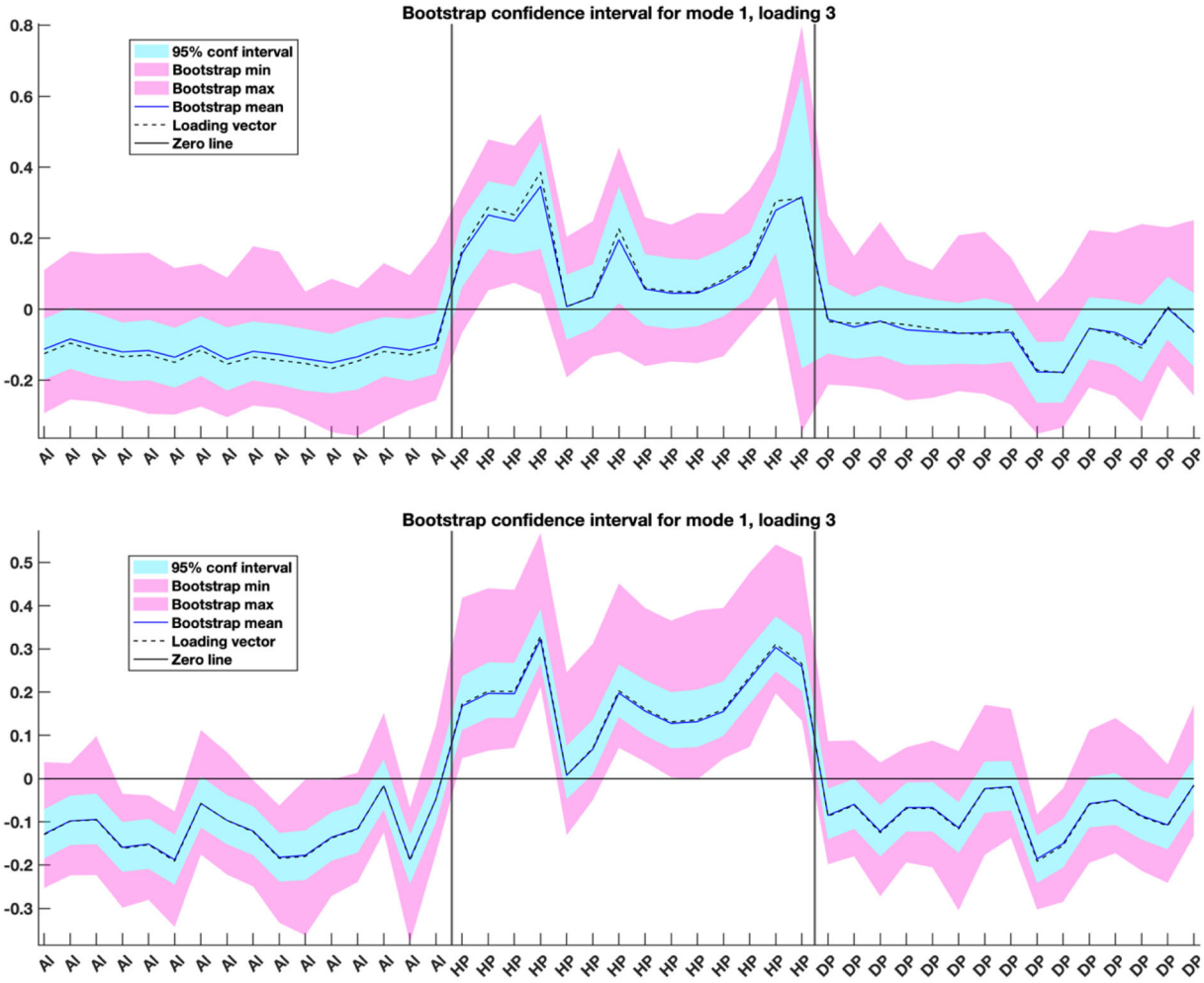


FIGURE 9 Bootstrap confidence intervals for eigenvector $\mathbf{g}_3$, top: $4 \times 5 \times 2$ Tucker3 model, bottom: $3 \times 7 \times 3$ Tucker3 model.

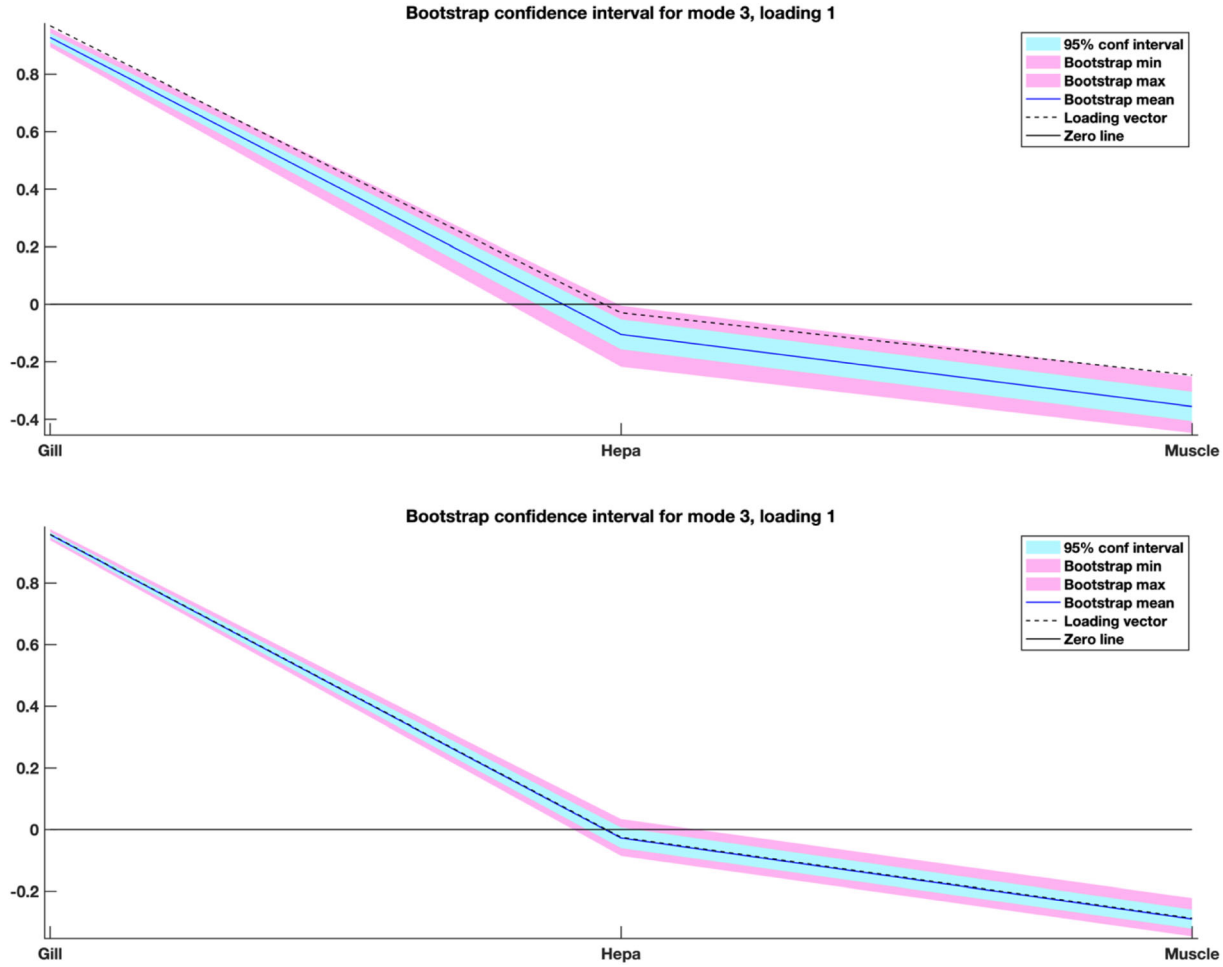


**FIGURE 10** Bootstrap confidence intervals for eigenvector $\mathbf{e}_1$, top: $4 \times 5 \times 2$ Tucker3 model, bottom: $3 \times 7 \times 3$ Tucker3 model.

the reduced models using 10,000 permutations. When the core value $c_{111}$ and its triad of eigenvectors is removed, the reduced $3 \times 7 \times 3$ model still has highly significant variance on factors A, B and A $\times$ B. Going down the list sequentially, when the 36th core value and its triad is removed, we still observe ASCA+ detectable structure, as at least one of the main factors or interactions is still significant (see Table 4). However, when we remove the 37th core element, we no longer observe ASCA+ detectable structure on factors A, B, and the interaction A $\times$ B in the reduced model. Therefore, we find that 36 core values and their associated triads are sufficient to model effects of factors A, B, and A $\times$ B. We conclude after backward elimination that the number of core values can be further reduced from 39 (after bootstrap analysis) to 36 (after backwards elimination). All removed core values account for 0.38% of the dataset variance; thus, the variance modeled by the fully reduced $3 \times 7 \times 3$ model with 36 core values retained is decreased from 78.03% to 77.65%.

## 2.6 | Interpretation of triads (factors)

Interpreting the most important triad (factor) $c_{111} \times \mathbf{g}_1 \times \mathbf{h}_1 \times \mathbf{e}_1$, (largest amount of variance explained), ASCA analysis of this unfolded tensor shows it has significant structure with respect to Factor A (disease state/region) but not Factor B (tissue). Interpretation of the individual vectors of triads is aided by this knowledge. In this triad, the vector $\mathbf{g}_1$ shows some discriminating power between Albemarle crabs which have low values, whereases the Diseased and Healthy Pamlico crabs tend to have high values (see Figure 11, bottom left panel). In vector $\mathbf{h}_1$, nine elements, Cr, V, Ti, Al, Sn, Fe, Co, Si, and Mn have high values and narrow confidence intervals, indicating they are highly significant in this triad. Dividing the value of each element in $\mathbf{h}_1$ by the bootstrap range and sorting them allows one to rank them in order of

**TABLE 5** ASCA+ backward elimination procedure.

| Order | Core element | Pct. variance of triad | Factor A (disease state/region) ASCA+ *p*-values | Factor B (tissue) ASCA+ *p*-values | Interaction (A × B) ASCA+ *p*-values |
|---|---|---|---|---|---|
| 1 | $c_{111}$ | 32.37 | 0.0001 | 0.0001 | 0.0001 |
| 2 | $c_{122}$ | 13.55 | 0.0001 | 0.0001 | 0.0001 |
| 3 | $c_{231}$ | 6.62 | 0.0001 | 0.0001 | 0.0001 |
| 4 | $c_{213}$ | 4.68 | 0.0001 | 0.0001 | 0.0001 |
| ... | ... | ... | ... | ... | ... |
| 28 | $c_{363}$ | 0.11 | 0.0490 | 0.0017 | 0.0008 |
| 29 | $c_{333}$ | 0.10 | 0.2046 | 0.0012 | 0.0012 |
| 30 | $c_{141}$ | 0.10 | 0.1673 | 0.0129 | 0.0006 |
| 31 | $c_{163}$ | 0.09 | 0.3825 | 0.0304 | 0.0016 |
| 32 | $c_{172}$ | 0.07 | 0.3730 | 0.1542 | 0.0029 |
| 33 | $c_{372}$ | 0.07 | 0.3930 | 0.1493 | 0.0068 |
| 34 | $c_{321}$ | 0.06 | 0.4606 | 0.1283 | 0.0110 |
| 35 | $c_{232}$ | 0.06 | 0.4525 | 0.4202 | 0.0063 |
| 36 | $c_{261}$ | 0.05 | 0.2744 | 0.2771 | 0.0089 |
| 37 | $c_{322}$ | 0.05 | 0.2706 | 0.2657 | 0.0742 |

*Note*: The 63 core values of the 3 × 7 × 3 model are ordered from largest variance explained to smallest with ASCA+ *p*-values shown using 10,000 permutations. The largest 37 are shown. Items in red are not significant at the 95% confidence level.

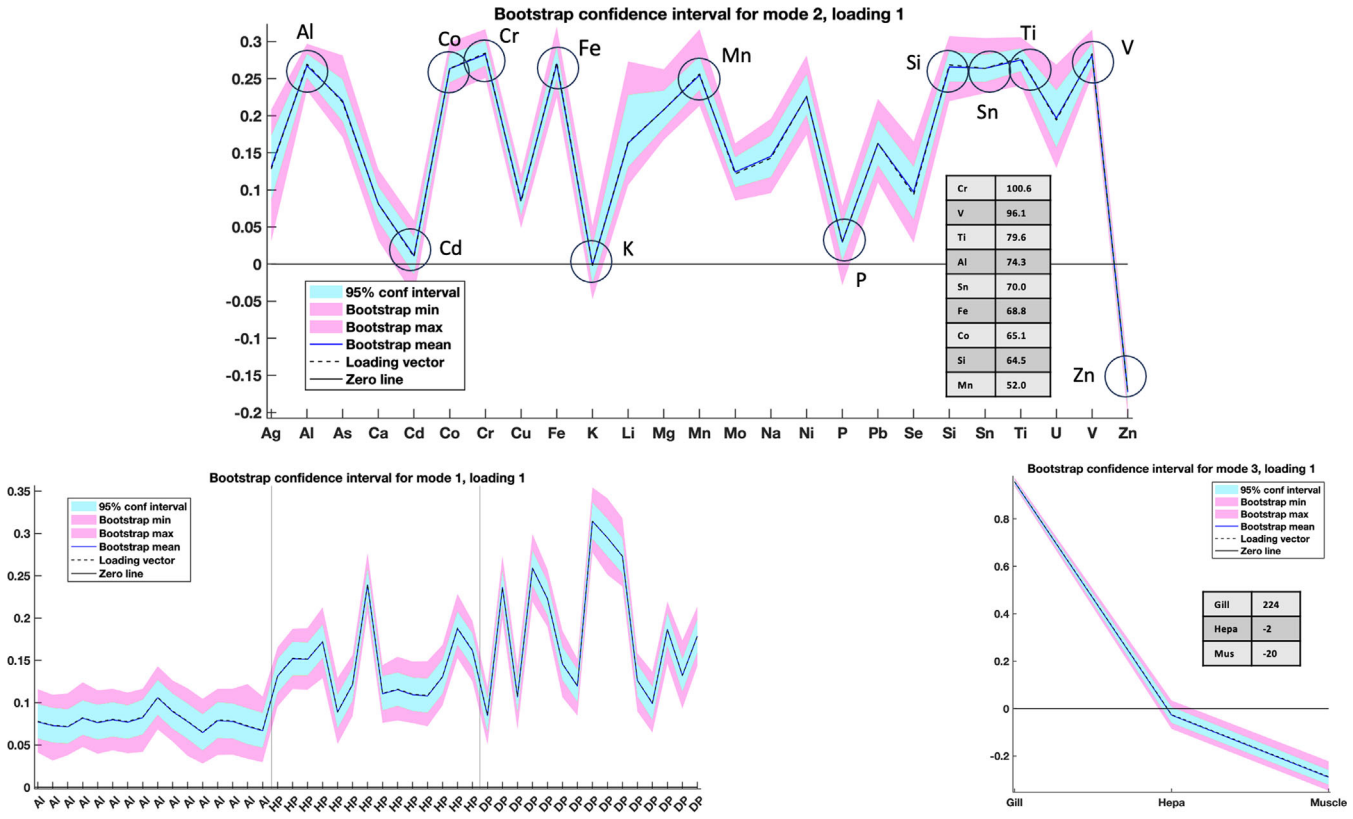


**FIGURE 11** Bootstrap confidence interval for the most important triad (factor) $c_{111} \times \mathbf{g}_1 \times \mathbf{h}_1 \times \mathbf{e}_1$.

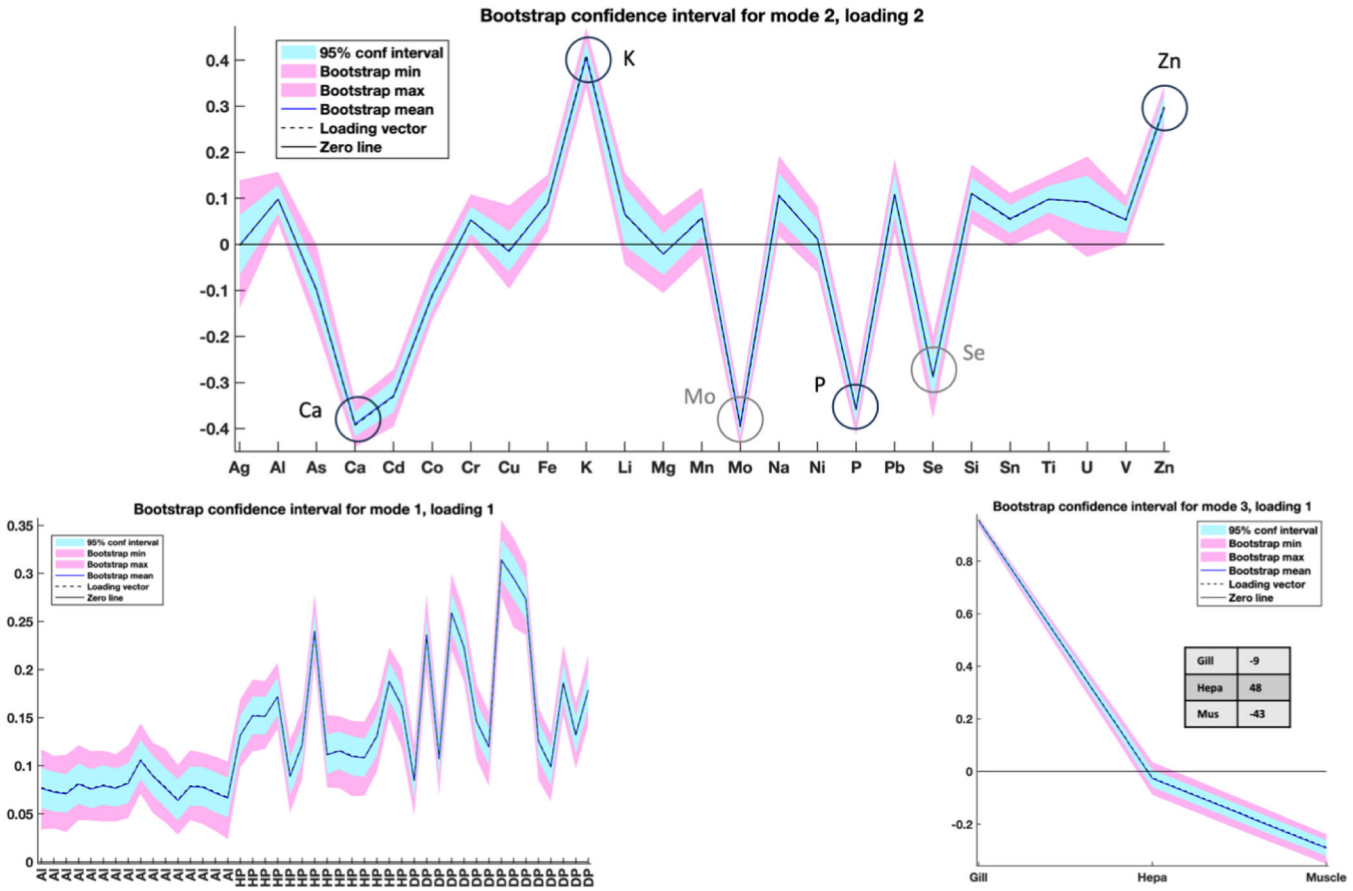


**FIGURE 12** Bootstrap confidence interval for the second important triad (factor) $c_{122} \times \mathbf{g}_1 \times \mathbf{h}_2 \times \mathbf{e}_2$.

decreasing significance (see Figure 11, inset table of top panel). Looking at the plot of $\mathbf{e}_1$ (see Figure 11, bottom right panel), it can be seen that the value for gill tissue is very large and the confidence interval is narrow, indicating it is highly significant, whereas the value for hepatopancreas is not significantly different from zero. This is consistent with ASCA results which indicates that the triad (factor) $c_{111}$ does not have statistically significant structure for explaining differences in Factor B (tissue). In summary, this triad models the response for elements listed above which are strongly correlated in gill tissue of Pamlico crabs. These elements are known to be present in the naturally occurring minerals and clay of this region and are insoluble at normal river pH; however, being "hard" metal ions, they tend to form soluble complexes with fluoride ions. The model indicates that their response is significant in gill tissue but less so in muscle and not in hepatopancreas tissue.

For the second most important triad, $c_{122}$, $\mathbf{h}_2$, potassium, K, is highly significant and negatively correlated with Ca, P, Se, and Mo; (see Figure 12, top panel) whereas Zn is positively correlated with K. The individual loadings in $\mathbf{e}_2$ for hepatopancreas and muscle are large in magnitude, negatively correlated, and the confidence intervals are narrow, indicating they are highly significant, whereas the coefficient for gill tissue is much smaller and not as significant in this triad (see Figure 12, bottom right panel). The response for these physiologically important elements is an important discriminating factor between the control group (Albemarle crabs) and Pamlico crabs (diseased and healthy). Please note that the bottom left panels in Figures 11 and 12 are the same as they come from the same loading in mode 1.

The third most important triad, $c_{231}$, is the first instance (largest amount of variance explained) that shows significant ASCA+ structure for tissue (results not shown), whereas the previous two triads (largest core values) did not. There are eight Diseased Pamlico crabs that stand out, as they have much larger coefficients when normalized by their confidence intervals. In vector $\mathbf{h}_3$, elements Cu and Na are strongly correlated, and their response is large. This is gratifying, because blue crabs achieve osmoregulation in response to varying salinity levels by adjusting copper in the hemolymph (cyanoglobin).[19] We also observe that Al is negatively correlated and Pb is slightly positively correlated, and $\mathbf{e}_1$ shows that responses for these elements are important in gill tissue. For this triad, ASCA+ shows significant structure for factors A, B, and interaction, A × B. We note that various complexes of aluminum with fluoride will increase its

overall solubility in aqueous systems at normal river pH. The fourth core value in order of decreasing size is $c_{213}$. ASCA+ shows that this is the first triad that has significant structure for tissue but not for region. Trends observed in the coefficients of $\mathbf{h}_1$ and $\mathbf{g}_2$ were previously noted above. Here, the response of these elements for the 8 Diseased Pamlico crabs are important in hepatopancreas and muscle as observed in $\mathbf{e}_2$.

This same analysis can be performed for all remaining triads as well but will not be further showcased in this study. However, it is important to note that the combination of using Tucker3 triads (which are all orthogonal with respect to each other and can therefore be analyzed in an independent way) and ASCA+ (which tells us something about the significance of the respective factors) is a powerful combination of tools that can help in the interpretation of the model and results, and to recognize important variables for the factors included in the experimental design.

## 3 | CONCLUSIONS

Analysis of Tucker3 residuals with ASCA+ allows us to identify and avoid Tucker models that do not fully model the structure in an experimental design. It is possible for ASCA to miss variation because of effects not used as factors in the design, for example, male vs. female crabs, whereas a Tucker3 model would likely capture this variation. This might cause our ASCA procedure to select an overly simplistic Tucker3 model, a potential limitation of this approach. We note, however, in ANOVA types of analyses that missed factor effects are usually confounded in the studied effects which would help guard against selecting overly simplistic Tucker3 models in our ASCA+ procedure. This is born out in the present study, where the ASCA+ residual variance and the Tucker $3 \times 7 \times 3$ residual variance were similar; 21.97% compared to 18.15%, respectively.

In conclusion, we have shown the complementary nature of Tucker3 and ANOVA simultaneous component analysis (ASCA) models for the investigation of designed multivariate experiments with multiple factors and levels. Despite the fact that Tucker3 models do not separate the variation between each factor in the way ASCA+ does, we have shown that (a) ASCA+ can be used to identify statistically sufficient Tucker3 models; (b) ASCA+ can be used to identify statistically important triads and assigning them to specific factors, making their interpretation easier; and (c) ASCA+ can be used to eliminate non-significant triads making visualization and interpretation simpler. We have also shown (d) how this approach can be combined with bootstrapping to identify statistically meaningful core values and loading values, making visualization and interpretation easier.

The power of combining these methods is clearly born out when assessing the statistical sufficiency of Tucker3 models. Compared to the original $4 \times 5 \times 2$ model,[5] which used four factors in **G,** ASCA+ analysis indicated only three factors were needed for **G**, indicating that the eigenvector matrix, **G**, was overdetermined and included an unnecessary factor. ASCA+ also showed that **H** was underdetermined in the original paper where only five factors were selected whereas seven were needed to generate a statistically sufficient model. Interpretation of the original $4 \times 5 \times 2$ was therefore incomplete, with important relationships between the different element contributions left out. The $3 \times 7 \times 3$ combination was then used throughout the paper as it was the model with the lowest complexity for which this statement was valid.

Finally, when an experimental design is known about a dataset, this strategy of using ASCA+ on model residuals is not limited to just Tucker3 analysis, but it can also be used in other decomposition methods (e.g., MCR-ALS and PARAFAC) to give a robust estimation in determining a sufficient number of model components, given that the model residuals are assumed to be normally distributed.

### CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial interest.

### PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3514.

### DATA AVAILABILITY STATEMENT

Data and code and code to reproduce the calculations in this paper are in GitHub https://github.com/FarnooshKoleini/TUSCA.

## ORCID

*Siewert Hugelier* https://orcid.org/0000-0002-6224-652X
*Hamid Abdollahi* https://orcid.org/0000-0002-5994-6365
*Paul J. Gemperline* https://orcid.org/0000-0002-3742-1170

## REFERENCES

1. Andersson CA, Bro R. *The N-way toolbox for MATLAB*. Elsevier BV; 2000. doi:10.1016/s0169-7439(00)00071-x
2. Kolda TG, Bader BW. Tensor decompositions and applications. *SIAM Rev*. 2009;51(3):455-500. doi:10.1137/07070111X
3. Kiers HAL. Bootstrap confidence intervals for three-way methods. *J Chemometr*. 2004;18(1):22-36. doi:10.1002/cem.841
4. Zwanenburg G, Hoefsloot HCJ, Westerhuis JA, Jansen JJ, Smilde AK. ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. *J Chemometr*. 2011;25(10):561-567. doi:10.1002/cem.1400
5. Gemperline PJ, Miller KH, West TL, Weinstein JE, Hamilton JC, Bray JT. Principal component analysis, trace elements, and blue crab shell disease. *Anal Chem (Washington)*. 1992;64(9):523A-532A. doi:10.1021/ac00033a719
6. Mahood K. Mapping outside the square: cultural mapping in the south-East Kimberley. *Aborig Hist*. 2011;30. doi:10.22459/AH.30.2011.02
7. Kroonenberg PM, Basford KE, Gemperline PJ. Grouping three-mode data with mixture methods: the case of the diseased blue crabs. *J Chemometr*. 2004;18(11):508-518. doi:10.1002/cem.896
8. Camacho J, Díaz C, Sánchez-Rovira P. Permutation tests for ASCA in multivariate longitudinal intervention studies. *J Chemometr*. 2022;37:e3398.
9. Madssen TS, Giskeødegård GF, Smilde AK, Westerhuis JA. Repeated measures ASCA+ for analysis of longitudinal intervention studies with multivariate outcome data. *PLoS Comput Biol*. 2021;17(11):e1009585. doi:10.1371/journal.pcbi.1009585
10. Camacho J, Pérez-Villegas A, Rodríguez-Gómez RA, Jiménez-Mañas E. Multivariate exploratory data analysis (MEDA) toolbox for Matlab. *Chemom Intel Lab Syst*. 2015;143:49-57. doi:10.1016/j.chemolab.2015.02.016
11. Timmerman ME, Kiers HAL. Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika*. 2003;68(1):105-121. doi:10.1007/BF02296656
12. McDonald JH. *Handbook of Biolological Statistics*. 2014.
13. Anderson M, Braak CT. Permutation tests for multi-factorial analysis of variance. *J Stat Comput Simul*. 2003;73(2):85-113. doi:10.1080/00949650215733
14. Liland KH, Smilde A, Marini F, Næs T. Confidence ellipsoids for ASCA models based on multivariate regression theory. *J Chemometr*. 2018;32(5):e2990. doi:10.1002/cem.2990
15. Kroonenberg PM, de Leeuw J. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*. 1980;45(1):69-97. doi:10.1007/BF02293599
16. Leardi R. Multi-way analysis with applications in the chemical sciences, Age Smilde, Rasmus Bro and Paul Geladi, Wiley, Chichester, 2004, ISBN 0-471-98691-7, 381 pp. *J Chemometr*. 2005;19(2):119-120. doi:10.1002/cem.908
17. Davison AC, Hinkley DV. *Bootstrap methods and their application*. Cambridge University Press; 1997. doi:10.1017/CBO9780511802843
18. Kruskal JB. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl*. 1977;18(2):95-138. doi:10.1016/0024-3795(77)90069-6
19. Anderson JD, Prosser CL. Osmoregulating capacity in populations occurring in different salinities. Paper presented at the. *Biol Bull*. 1953;105(2):369.

**How to cite this article:** Koleini F, Hugelier S, Lakeh MA, Abdollahi H, Camacho J, Gemperline PJ. On the complementary nature of ANOVA simultaneous component analysis (ASCA+) and Tucker3 tensor decompositions on designed multi-way datasets. *Journal of Chemometrics*. 2023;e3514. doi:10.1002/cem.3514