Cancer Subtype Detection Using Biomarker Discovery in Multi-Omics Tensor Datasets

By

Farnoosh Koleini

July, 2023

Directors of Thesis: Dr. Paul J. Gemperline, and Dr. Nasseh Tabrizi

Major Departments: Computer Science, and Chemistry

ABSTRACT

This thesis begins with a thorough review of research trends from 2015 to 2022, examining the challenges and issues related to biomarker discovery in multi-omics datasets. The review covers areas of application, proposed methodologies, and evaluation criteria used to assess performance, as well as limitations and drawbacks that require further investigation and improvement. This comprehensive overview serves to provide a deeper understanding of the current state of research in this field and the opportunities for future research. It will be particularly useful for those who are interested in this area of study and seeking to expand their knowledge. In the second part of this thesis, a novel methodology is proposed for the identification of significant biomarkers in a multi-omics colon cancer dataset. The integration of clinical features with biomarker discovery has the potential to facilitate the early identification of mortality risk and the development of personalized therapies for a range of diseases, including cancer and stroke. Despite extensive efforts towards discovering disease-associated biomolecules by analyzing data from various “omics” experiments, such as genomics, transcriptomics, and metabolomics, the poor integration of diverse forms of 'omics' data has

made the integrative analysis of multi-omics data a daunting task. Our research includes ANOVA simultaneous component analysis (ASCA) and Tucker3 modeling to analyze a multivariate dataset with an underlying experimental design. By comparing the spaces spanned by different model components we showed how the two methods can be used for confirmatory analysis and provide complementary information. We demonstrated the novel use of ASCA to analyze the residuals of Tucker3 models to find the optimum one. Increasing the model complexity to more factors removed the last remaining ASCA detectable structure in the residuals. Bootstrap analysis of the core matrix values of the Tucker3 models was used to check that additional triads of eigenvectors were needed to describe the remaining structure in the residuals. Also, we developed a new simple, novel strategy for aligning Tucker3 bootstrap models with the Tucker3 model of the original data so that eigenvectors of the three modes, the order of the values in the core matrix, and their algebraic signs match the original Tucker3 model without the need for complicated bookkeeping strategies or performing rotational transformations. Additionally, to avoid getting an overparameterized Tucker3 model, we used the bootstrap method to determine 95% confidence intervals of the loadings and core values. Also, important variables for classification were identified by inspection of loading confidence intervals. The experimental results obtained using the colon cancer dataset demonstrate that our proposed methodology is effective in improving the performance of biomarker discovery in a multi-omics cancer dataset. Overall, our study highlights the potential of integrating multi-omics data with machine learning methods to gain deeper insights into the complex biological mechanisms underlying cancer and other diseases.

Cancer Subtype Detection Using Biomarker Discovery in Multi-Omics Tensor Datasets

A Thesis

Presented to the Faculty of the Computer Science and Chemistry Departments

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Data Science and Chemistry

By

Farnoosh Koleini

July, 2023

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Biomarker discovery has emerged as a crucial field in the battle against diseases like cancer. By exploring various biological indicators present in patients, researchers can identify specific molecules, genes, proteins, or characteristics that can serve as reliable biomarkers for early detection, prognosis, and personalized treatment strategies. These biomarkers not only aid in the early identification of cancer, enabling timely intervention and potentially improved outcomes, but they also contribute to a deeper understanding of the underlying mechanisms of the disease [1]. Through advanced technologies such as genomics, proteomics, and metabolomics, scientists can analyze large datasets and compare them to healthy controls, unveiling patterns and signatures that are unique to cancer. This invaluable information paves the way for the development of innovative diagnostic tests, targeted therapies, and more accurate patient monitoring. Biomarker discovery represents a transformative approach that holds tremendous promise for revolutionizing cancer care by enhancing detection sensitivity, reducing healthcare costs, and ultimately improving patient survival rates [2].

Recent advancements in integrated analysis have shown promising results in enhancing knowledge discovery by utilizing tensor decompositions on data sourced from multiple outlets. In the field of metabolomics, for instance, diverse analytical techniques are employed to study biological fluids like blood or urine, aiming to identify disease or metabolites. To address the challenge of data fusion, a joint factorization approach has been devised. This method allows

data from various sources to be represented as multiple matrices, which are subsequently evaluated collectively through tensor decomposition techniques [3].

In the past, researchers used single-omics investigations to uncover disease causes and aid in treatment selection or design. However, many diseases involve intricate molecular pathways where different biological layers interact. As a result, there is now a growing need for biomarker discovery in multi-omics investigations, which can integrate multiple layers of biological information and offer a more comprehensive understanding of a specific phenotype. This approach allows for a more holistic and detailed perspective, enabling researchers to gain a fuller picture of complex diseases and potentially improve diagnostic accuracy and treatment outcomes [28].

Integrating multi-omics datasets faces several challenges. Some obstacles, like missing values and class imbalance, are common in machine learning analysis. Class imbalance occurs when there is an unequal distribution of classes in the learning data, often seen in rare events. A classification dataset with skewed class proportions is called imbalanced. Classes that make up a large proportion of the dataset are called majority classes. Those that make up a smaller proportion are minority classes. Strategies such as sampling and cost-sensitive learning can address this issue [35, 36]. 'Omics' datasets, being biological in nature, are inherently noisy and complex, making it challenging to identify relevant patterns across multiple datasets. Limited availability of substantial biomedical data due to factors like financial constraints and rarity of the desired phenotype can result in high-dimensional datasets with more variables than samples, leading to overfitting and reduced generalizability [36]. Moreover, heterogeneity among 'omics' methodologies and varying dataset sizes pose integration and learning imbalances [37].

Scalability is another technical concern when working with large and heterogeneous multi-omics datasets, requiring efficient methods for data processing and analysis. Scalability refers to the ability of a system, application, or infrastructure to handle and accommodate an increasing amount of work, data, or users while maintaining or improving its performance and efficiency. Overcoming these challenges is crucial for effective biomarker discovery in multi-omics datasets, with a need for comprehensive investigation into data heterogeneity [38-40].

As we will discuss later in Chapter 2, several studies try to deal with these challenges to detect biomarkers in complex diseases, however, none of them have used Tucker3 as a tensor decomposition method and ANOVA Simultaneous Component Analysis as a confirmatory analysis and provide complementary information for biomarker discovery in colon cancer multi-omics datasets.

In this thesis, we report a comprehensive review of biomarker discovery in multi-omics datasets using tensor decompositions as the basis idea of our proposed model. We then perform a novel application of ASCA to analyze the residuals of Tucker3 models to find the optimum one. We use bootstrap analysis of the core matrix values of the Tucker3 models to indicate whether there is a need for additional sets of eigenvectors to describe the remaining structures in the residuals or not. Furthermore, we introduce a simple and innovative strategy to align the bootstrap models with the original Tucker3 model, ensuring the eigenvectors, the order of values in the core matrix, and their signs are matched without requiring complex bookkeeping or rotational transformations. To prevent an overly complex Tucker3 model, we employ the bootstrap method to determine 95% confidence intervals for the loadings and core values. Additionally, we identify

significant multi-omics features, biomarkers, for classification by examining the confidence intervals of the loadings.

1.1 Research Contribution

In this research work, we first report a comprehensive literature survey on biomarker discovery in multi-omics datasets using tensor decompositions as the basis of the proposed approach. In this literature review, we comprehensively review the trend of research conducted from 2015 to 2022 in terms of challenges and problems regarding biomarker discovery in multi-omics datasets, areas of application, proposed methodologies, evaluation criteria used to assess the performance, limitations, and drawbacks that require investigation and improvements.

The second and primary contribution of this thesis is developing a methodology that uses ASCA to analyze the residuals of Tucker3 models to find an optimum model. We then utilizing bootstrap analysis of the core matrix values of Tucker3 models to determine if additional sets of eigenvectors are necessary to explain the remaining structure in the residuals. To prevent an overly complex Tucker3 model, we apply the bootstrap method to establish 95% confidence intervals for the loadings and core values. Additionally, we identify significant multi-omics features or biomarkers for classification by examining the confidence intervals of the loadings.

1.2 Thesis structure

The thesis is structured as follows: Chapter 2 presents a comprehensive literature survey on biomarker discovery in multi-omics datasets using tensor decompositions. In this chapter, existing studies, methodologies, and findings related to biomarker discovery are discussed, with

a specific focus on the application of tensor decompositions in multi-omics datasets. Chapter 3 provides the related Literature on multi-dimensional data analysis in cancer studies.

Chapter 4 describes the proposed methodology in detail, explaining the rationale behind using tensor decompositions for biomarker discovery and providing a step-by-step approach for implementing the methodology. The methodology is then applied to a multi-dimensional dataset, the Blue Crab dataset, serving as a case study. Moving forward, Chapter 5 provides the performance analysis of the proposed method using one of the most extensive public datasets from the National Institutes of Health (NIH), specifically the colon cancer dataset. The obtained results are thoroughly analyzed, highlighting the significance of the identified biomarkers. Finally, in Chapter 6, the current research is concluded, summarizing the key findings of the thesis. Additionally, future research directions in biomarker discovery using tensor decompositions are discussed, outlining potential areas of exploration.

Chapter 2

Comprehensive Literature Survey

2.1 Introduction

Biomarkers are biological molecules that are indicative of normal or abnormal processes, such as disease states or responses to treatments. These biological molecules may be found in any type of organism by sampling tissue and body fluids followed by biochemical analysis. The development of high throughput methods has facilitated an explosion of research in this field. When combined with clinical data, the resulting information can be used for earlier detection of diseases and the development of personalized therapies. Moreover, new developments in ‘‘omics’’ technology provide researchers the chance to look for disease biomarkers at the system level [1]. A Tremendous amount of work has gone into discovering disease-associated biomolecules by analyzing data obtained from different ‘‘omics’’ experiments (genomics, transcriptomics, metabolomics). However, due to the complexity of biological systems and the poor integration of various forms of ‘‘omics’’ data, integrative analysis of multi-omics data is a difficult undertaking. Various feature selection procedures have been shown to provide different sets of biomarkers [2]. A classic approach to biomarker selection comprises statistical approaches such as the Student’s t-test and ANOVA, which find and choose biomolecules with a significant change in expression level between separate biological groups (normal vs. disease; untreated vs. treated). One clear disadvantage of these methods is that they ignore the fact that biomolecules

in a biological system are densely interconnected and interact with one another. Integrated analysis using tensor decompositions of data from many sources has recently demonstrated the ability to improve knowledge discovery. In metabolomics, for example, biological fluids such as blood or urine are examined using various analytical techniques to find molecules associated with specific diseases or diets [3]. A joint factorization problem has been developed for the topic of data fusion [3]. Data from many sources can be represented as several matrices, which can then be evaluated jointly using tensor decomposition methods. The tensor factorization has also been found to be effective in other domains, including social network analysis [4-8], signal processing [9,10], and bioinformatics [11-13]. Also, coupled tensor decomposition methods have been developed and employed in chemometrics [14], bioinformatics [11,12], signal processing [9,15,16], and data mining [17,18]. With the introduction of high throughput technology capable of extensive analysis of genes, transcripts, proteins, and other significant biological molecules, the identification of molecular markers of disease processes has become a reality on a scale never before seen. It has, however, made it more difficult to extract relevant molecular markers of biological processes from these complex datasets. The process of biomarker discovery and characterization allows for more sophisticated approaches to integrating purely statistical and expert knowledge-based approaches, and tensor decompositions provide a great opportunity to aid in the interpretation of such interactions and the identification of reliable biomarkers [19]. There are several review papers on biomarker discovery using tensor decompositions published in the last few years [20], [21].

This paper reviews research in this area from 2015 to 2022 to provide useful insights into the recent advances in biomarker discovery using tensor decompositions and suggests future

research directions. The challenges, drawbacks, and new opportunities that have arisen due to the availability of more multi-omics data and information have called for studies on developing tensor decomposition methods to detect biomarkers in recent years. Figure 2.1 shows the number of publications that use tensor decompositions for biomarker detection or deal with biomarker discovery challenges published between 2015 and early 2022.

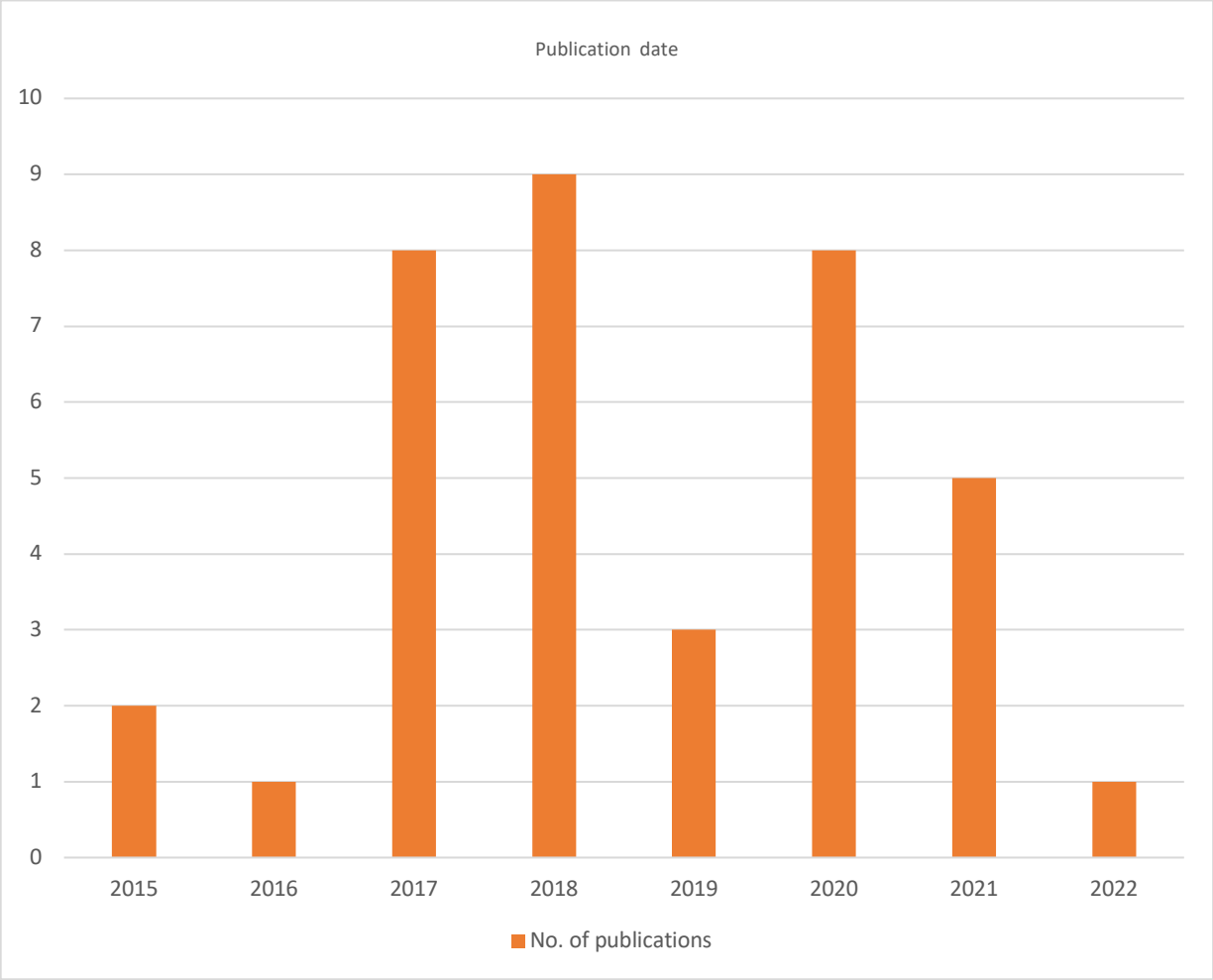


Fig.2.1: Number of papers from 2015 to 2022

2.3 Systematic Review

The first step in this systematic review was to define the goals of the survey. These goals are described as follows.

- Identifying the problems and challenges regarding biomarker discovery in multi-omics datasets by tensor decompositions.
- Identifying algorithms and methodologies employed to solve these problems and their challenges.
- Identifying areas of application for biomarker discovery in multi-omics datasets.
- Identifying evaluation criteria used to evaluate developed TD-based methods.

In this systematic review, we first searched the literature for publications using scientific search engines and collected databases of publications. The search query used was (“biomarker” AND “discovery”) AND (“multi-omics”) AND (“tensor” AND “decompositions”). This search query was used on several databases including IEEE Xplore, ACM Digital Library, Lynda.com, ScienceDirect, and SpringerLink. Then, the selected publications were studied, and the information was used to answer the main questions of this systematic review.

2.4 Integrating Multi-Omics Datasets: Opportunities and Challenges

Developing computational models to discover potential biomarker-disease connections in multi-omics data, which could provide insight into disease pathophysiology and improve illness diagnostic and prognostic accuracy, is gaining popularity. The recent introduction of effective and low-cost screening technologies has resulted in massive amounts of biological data, paving the

door for a new era of treatments and customized medicine [22, 23]. Clinical information and ‘‘omics’’ data can be acquired directly from databases or collected through screening technologies for disease [24], class prediction [25], biomarker identification [26], disease subtyping [24], better system biology understanding [27], drug repurposing, and other applications. Each ‘‘omics’’ data type is specific to a single "layer" of biological information, such as genomics, epigenomics, transcriptomics, proteomics, or metabolomics, and provides a complementary medical perspective of a biological system or an individual [22].

1) *Integration analysis of multi-omics datasets:*

Single-omics investigations were previously conducted to discover the causes of diseases to help design or pick a suitable treatment. Most diseases, on the other hand, involve complicated molecular pathways in which distinct biological layers interact with one another. Therefore, there is a greater demand for biomarker discovery in multi-omics investigations that can incorporate several layers and provide a fuller picture of a particular phenotype [28]. Faint patterns in gene expression data can be enhanced by several ‘‘omics’’ methods [29]. For example, complementary information can be exploited to better explain classification results [30], improve prediction performance [31, 32], or comprehend complex molecular pathways [33]. Multi-omics studies, on the other hand, comprise data of varying types, scales, and distributions, with thousands of variables and only a few samples. Furthermore, biological datasets are complicated and noisy, with the possibility of errors due to measurement errors or unique biological variances. Finding relevant information and incorporating ‘‘omics’’ data into a useful model is difficult, and several methods and tactics have been developed in recent years to address this difficulty [24, 34]. As a result, researchers are seeking approaches that, by adding additional ‘‘omics’’ data, result in an

increase in performance rather than simply increasing the complexity and processing time of the task.

2) *Challenges of multi-omics datasets:*

When integrating multi-omics datasets, several obstacles occur. Some of these, such as the existence of missing values or class imbalance, are general to machine learning analysis. When working on rare events, such as an uncommon attribute in a population, class imbalance occurs when the distribution of classes in the learning data is biased. This problem can be solved using a variety of strategies, including sampling and cost-sensitive learning. Sampling tries to balance the dataset before the integration process, where either the majority class is randomly under-sampled, or the minority class is oversampled by creating new artificial observations, or a combination of both methods. Cost-sensitive learning is directly integrated into the algorithm and balances the learning process by giving more weight to misclassified minority observations [35, 36]. Some are more specific and include the noisiness and complexity of ‘‘omics’’ datasets, which naturally occur in biological data. Relevant patterns can occasionally be obscure and involve a large number of molecules from various ‘‘omics’’ layers. Therefore, identifying those patterns across numerous datasets is a challenging endeavor. Furthermore, due to financial constraints, the rarity of the desired phenotype, or a lack of willing volunteers, etc., the collection of substantial volumes of biomedical data is frequently only possible on a small sample of patients when conducting ‘‘omics’’ or multi-omics investigations. This results in datasets with variables greatly exceeding the number of samples. Machine learning algorithms have a propensity to overfit these high-dimensional datasets, which reduces their generalizability to new data. This problem is known as the ‘‘curse of dimensionality’’ [36]. Another difficulty is their

heterogeneity, which must be handled properly because various ‘‘omics’’ methodologies may provide data with varying distributions of types (e.g., numerical, categorical, continuous, discrete, etc.). Furthermore, ‘‘omics’’ datasets can vary greatly in size (number of features), with a typical gene expression dataset having tens of thousands of variables and a metabolomics dataset having only a few thousand. Disparities between ‘‘omics’’ datasets might impede integration and create an imbalance in the learning process [37]. Scalability is an additional technical issue regarding multi-omics datasets. The scope of genomics research has been broadened from a narrow single-layer examination to a comprehensive multi-dimensional interpretation of biological data as a result of the accessibility of these massive multidimensional and heterogeneous datasets. To create rich, multi-scale characterizations of biological systems, the emphasis is on combining various forms of omics data from many layers of biological regulation. However, it necessitates systems that can scale across heterogeneous datasets while also centralizing data processing analysis, and interpretation inside a unified inference framework [38-40]. Therefore, developing a quick and effective method that can compute tensor decompositions of larger quantities of data would lead to more effective biomarker discovery in multi-omics datasets. Figure 2.2 shows the number of papers that deal with specific problems in biomarker discovery in multi-omics datasets: the curse of dimensionality, scalability, and noisiness problems. Typically, these papers describe the development of procedures that perform better. While these issues are still being researched to improve biomarker identification in multi-omics datasets, data heterogeneity necessitates a more thorough investigation.

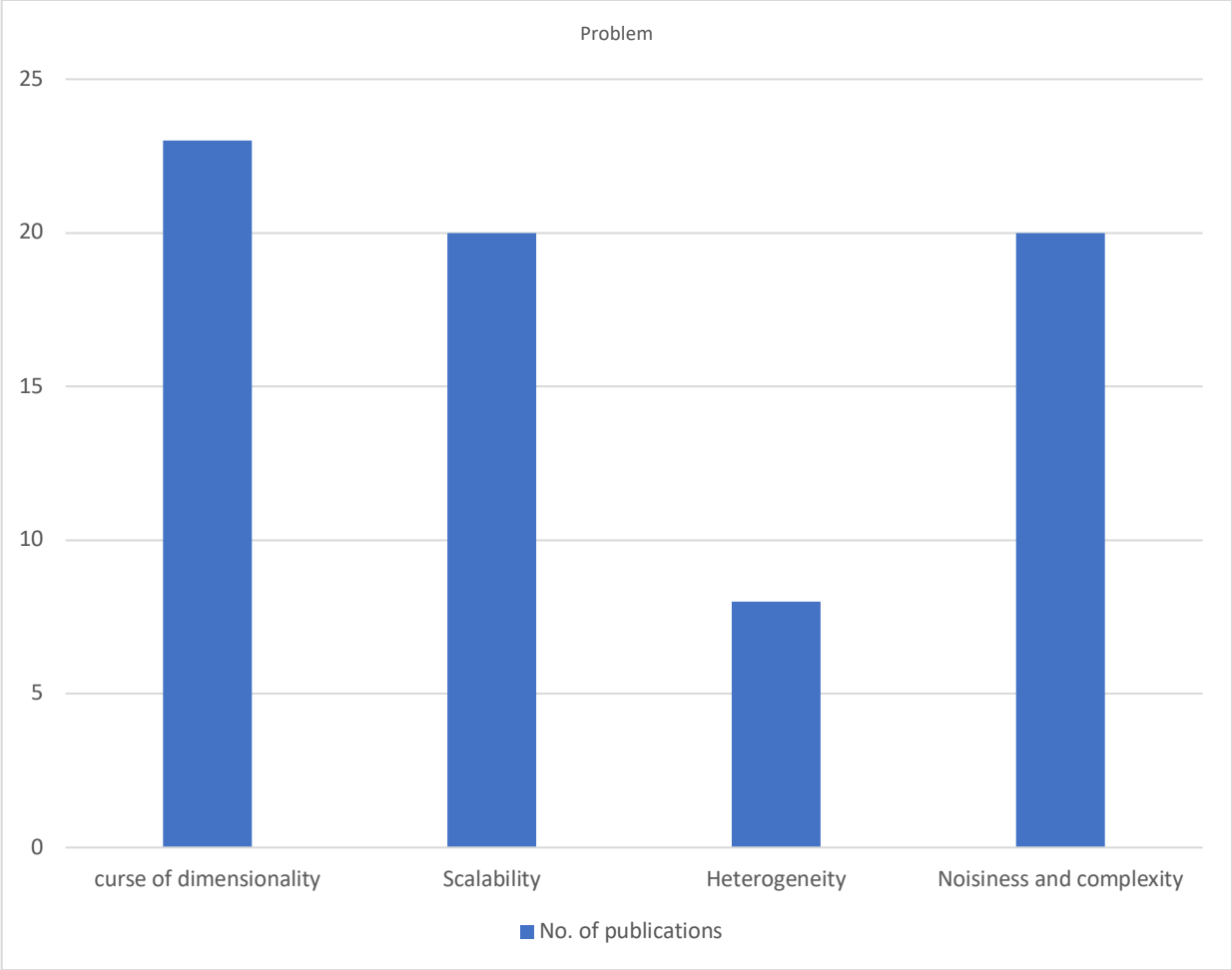


Fig.2.2: Number of studies dealing with a specific problem in biomarker discovery in multi-omics dataset

2.5 Methodologies

In this section, we review the approaches and methodologies that are described in the literature on biomarker discovery in multi-omics datasets by tensor decompositions.

2.5.1 Parallel Factor Analysis

Parallel Factor Analysis (PARAFAC) is a popular tensor decomposition method that is widely used in biomarker discovery in multi-omics datasets. It is a method for decomposing multidimensional arrays to focus on the aspects of interest and provide a clear illustration of the results. PARAFAC

is based on a mathematical model that depicts the interactions of the dimensions to be evaluated in the input data. The analysis dimensions must be defined before performing PARAFAC. Each input value can then be related to an index for each of the dimensions. Assuming *N*=3 dimensions, for example, x_{ijk}, identifies the measured value for index i in the first dimension, j in the second dimension, and k in the third dimension. Equation 2.1 represents the PARAFAC model, where *F* denotes the number of so-called components and defined so-called loading matrices **A**, **B**, and **C** of dimensions $I \times F$, $J \times F$, and $K \times F$ and with elements a_{if}, b_{jf}, and c_{kf}, respectively, and the model error, ε_{ijk}.

$$x_{ijk} = \sum a_{if}\, b_{jf} c_{kf} + \varepsilon_{ijk} \qquad (2.1)$$

Reference [41] provides the generic model that PARAFAC uses to represent the input data. A graphical illustration of this model is given in Figure 2.3. The data is decomposed into triads or trilinear components, where each component comprises one score vector and two loading vectors rather than one score vector and one loading vector as in bilinear PCA. It is the standard three-way procedure to consider scores and loadings numerically similarly, without making any distinction between the two. A well-established advantage of the PARAFAC model is the mathematical uniqueness of the solution. Unique solutions can be expected if the loading vectors are linearly independent in two of the modes and the third mode, and if no two loading vectors are linearly dependent in the third mode.

PARAFAC applications:

Zhang et al defined ‘‘a temporal and spatial feature similarity measure to calculate the rate of change and velocity of each biomarker in MRI to form a vector that represents the morphological

change of the biomarker, then calculating the similarity of the changing trend between two biomarkers to encode the data in a third-order tensor to extract interpretable biomarker latent factors from the original data using PARAFAC decomposition.'' [42].

Jung et al proposed ''a multi-omics analysis method called MONTI (Multi-omics Non-negative Tensor decomposition for Integrative analysis), that selects multi-omics features that can represent trait-specific characteristics.'' They describe the usefulness of multi-omics integrated analysis for cancer subtyping. The multi-omics data were first merged in a biologically meaningful way to generate a three-dimensional tensor, which was then decomposed using the PARAFAC method. MONTI was then utilized to identify highly informative subtype-specific multi-omics features [43].

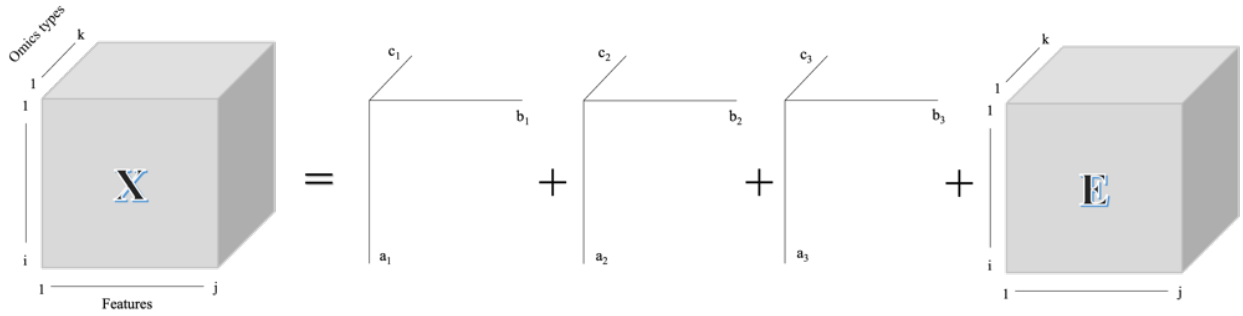


Figure2.3: A graphical illustration of the PARAFAC model.

2.5.2 Tucker3

Tucker3 is another tensor decomposition method that could be used in multi-omics datasets to detect biomarkers. The Tucker3 model name is taken from psychometrician Ledyard R. Tucker who in 1966 proposed the model. He also presented a method for calculating the model's parameters, and several changes to the algorithmic approach have subsequently been suggested.

The model has remained a powerful tool for analyzing three-way (and higher way) data arrays. The Tucker3 model is frequently used for decomposition, compression, and interpretation in many applications because of its generality and the way it treats the PARAFAC model as a particular instance. The Tucker3 model can be seen as an extension of the PARAFAC-CANDECOMP model along the line of outer products. Kroonenberg provided ‘‘a full mathematical description of this model as well as advanced topics such as data preparation/scaling and core rotation. Different numbers of factors in each of the modes can be extracted using the Tucker3 model [44].’’ Figure 2.4 is used to provide a simple explanation of the model.

Tucker3 applications:

Taguchi has focused on post-traumatic stress disorder (PTSD), a mental condition that can cause symptoms that do not appear to be immediately related to the central nervous system, which is thought to be directly affected by PTSD. PTSD-mediated heart disease is one such secondary disorder [45]. The spatial separation between the heart and the brain hindered researchers from clarifying the mechanisms that link the two disorders, despite the strong associations between PTSD and heart diseases. Their goal was to discover the genes that link cardiac problems with PTSD. To execute gene selection, they employed Tucker3 factorization as the tensor decomposition method to examine the gene expression profiles in diverse tissues, such as the heart and brain. The gene expression profiles were regarded as tensors. Gene expression profiles in diverse tissues were studied under various conditions such as stressful or unstressful, with varying periods of stress and rest time following the application of a stressor. Approximately 400 potential genes were identified that may mediate heart problems related to PTSD based on the

obtained features. Additionally, before being applied to gene expression profiles, Tucker3 was applied to a synthetic data set to illustrate the utility of their technique [45,46].

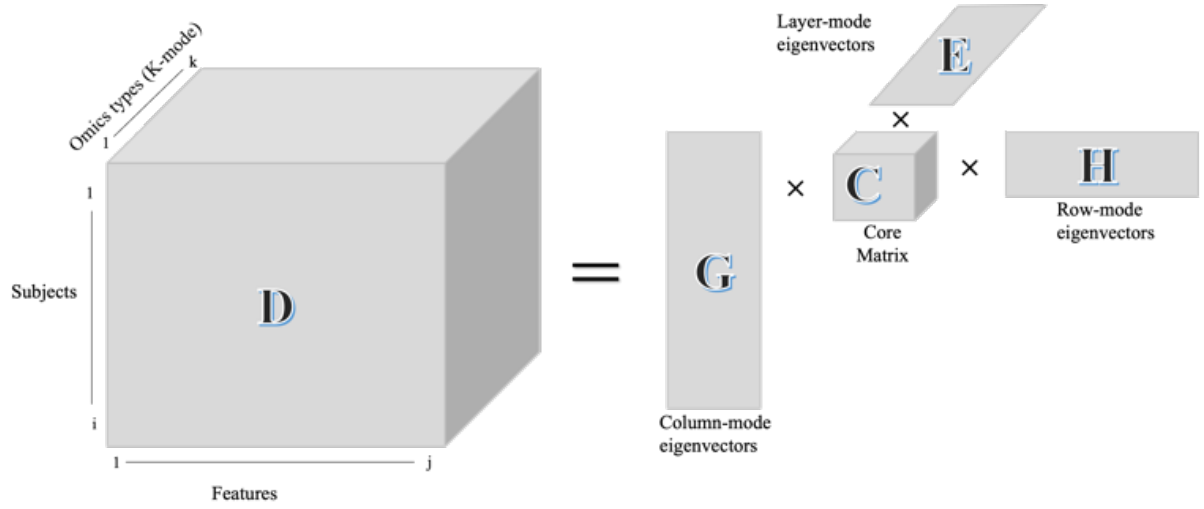


Fig.2.4: A graphical illustration of the Tucker3 model.

2.5.3 Hybrid and other techniques

Feature extraction methods are a class of techniques that try to turn a set of input biomarkers into another set of variables that are linear or non-linear combinations of the original biomarkers. The goal is to extract features in such a way that the resulting new variables retain useful information while being less noisy and less redundant. Learning from a smaller set of features or biomarkers reduces complexity while increasing computational efficiency. The interpretability of a model may be compromised by feature extraction methods since the derived features are no longer biological measurements. Feature extraction methods are frequently employed experimentally to visualize data and uncover significant features.

Principal Component Analysis (PCA) is the most extensively used feature extraction approach. [47] PCA creates new variables called principal components, which are uncorrelated linear

combinations of the original features and optimize the description of variance in the dataset; however, PCA is sensitive to outliers and is unable to describe non-linear trends in the data. To address these issues, several extensions have been developed, such as Kernel PCA [48] and Bayesian PCA [49]. Other similar methods such as Principal Coordinates Analysis (PCoA) [50], Correspondence Analysis (CA) [51], and Independent Component Analysis (ICA) [52] may improve PCA in certain ways. The majority of feature extraction techniques have also been developed with sparsity constraints. Sparse feature extraction methods can be used for feature selection with methods such as Sparse PCA (sPCA) [53], Sparse Canonical Correlation Analysis (CCA) [54], Sparse Non-negative Matrix Factorization (Sparse NMF) [55], and Sparse CA [56]. These approaches, however, fail to examine multi-omics datasets since applying them to concatenated ‘‘omics’’ typically yields unsatisfactory results. As a result, feature extraction methods are frequently used on each ‘‘omics’’ dataset for either block scaling or after concatenation of the extracted features or clustering, or other downstream analysis [38].

2.5.4 AI for biomarker discovery in multi-omics datasets

Gene regulatory networks, which are critical for understanding complicated disease mechanisms, have become one of the most popular topics for biomarker identification in multi-omics datasets. Several large-scale projects have been done and significant amounts of ‘‘omics’’ data have been released to identify heterogeneous genetic networks that underlie complex human diseases. The gene networks scale is increasing, and methodologies for analyzing large-scale gene networks have been proposed. Park et al. proposed a novel AI technique for analyzing gene regulation networks in depth. The multilayer networks were decomposed using an AI technique based on deep learning to identify all-encompassing gene regulatory systems distinguished by patient

clinical features. They extracted global and unique mechanisms of gene regulatory systems from the vast multiple networks using an AI technique based on tensor decomposition. They developed a novel technique to do integrative analysis of multilayer gene networks, which is an essential tool for precision medicine. In their method, gene regulatory networks were built under varied sample conditions, and the multilayer networks were thoroughly examined using an AI algorithm. To construct a low-dimensional subspace of the multiway interaction between genes, a deep learning algorithm for tensor decomposition was applied to the gene network for a target sample. They were able to understand the constructed large-scale gene networks since prediction and interpretation were carried out on the constructed low-dimensional subspace. Their technique is divided into two stages: building sample-specific gene regulatory networks and globally analyzing large-scale multiple gene networks using AI technology [57,58].

2.6 Applications

Discovering biomarkers has various uses in the healthcare system, such as early disease detection, disease prevention, identifying an individual's risk, monitoring disease, and drug development in the pharmaceutical sector. Therefore, biomarker discovery, specifically in multi-omics datasets by tensor decompositions could help a lot to develop biomarker applications. In this section, we will cover some of the important applications of biomarkers in the literature.

2.6.1 Early disease detection, prevention, and monitoring

Measures for the early detection of various diseases such as different cancers and stroke offer the opportunity to help control rising healthcare costs. We can already see that alternative disease prevention strategies will be used in the future because these strategies can and should be tailored to each patient based on their unique risk profiles. Fortunately, biomarkers make it

possible to detect diseases such as Alzheimer's and certain cancers at a disease stage even when the patient shows no symptoms. The recent failures of potential medications that are tailored for various conditions may be an indication that the clinical trial participants are too far along to benefit clinically. Therefore, the development of new therapeutics will be greatly influenced by validated biomarkers for the early detection and precise diagnosis of diseases in their preclinical phases. When biomarkers are used synthetically, they may someday be able to identify patients in the initial stages of the disease, when therapeutic modification is most likely possible. Because whether medicine is likely to work can frequently be a genetic issue, biomarkers are also important in the development of individualized treatment. As a result, determining or excluding specific genetic variations can make a significant contribution to therapeutic management, not only reducing costs and side effects but also improving treatment quality. Biomarkers can also be used to track treatment response [42,59,60].

2.6.2 Risk assessment

Biomarkers can be classified into susceptibility, effect, and exposure indicators. It is commonly expected that current developments in genomics, proteomics, and metabolomics will eventually translate into a constellation of advantages for human health. However, only a few biomarkers have been reported in the past ten years for risk assessment using "omics" technologies; however, there is a wide range of potential applications for "omics" technology. The lack of integrated bioinformatics techniques, statistical analysis, and predictive models frequently severely restricts the use of biomarker-based monitoring systems as a tool for environmental risk assessment. Therefore, identifying pertinent and reliable biomarkers that contribute to the assessment of environmental and health risks may be necessary [61,62].

2.6.3 Drug discovery and development

Biomarkers that are robust and verified are required to improve diagnosis, monitor drug activity, therapeutic response, and lead the development of safer and more tailored therapeutics for a variety of diseases. The development of specialized biomarkers for complicated chronic diseases can now be discovered and developed more quickly thanks to recent developments in multi-omics techniques, bioinformatics, and biostatistics. Even though there are still many obstacles to overcome, current initiatives for the discovery and development of disease-related biomarkers will help with the best decision-making during the medication development process and further our comprehension of the disease processes. To the benefit of patients, healthcare professionals, and the biopharmaceutical industry, good preclinical biomarker translation into the clinic will pave the path for the effective execution of personalized therapies across a range of complex disease areas [63,64]. Figure 2.5 illustrates the distribution of studies focusing on each area of application.

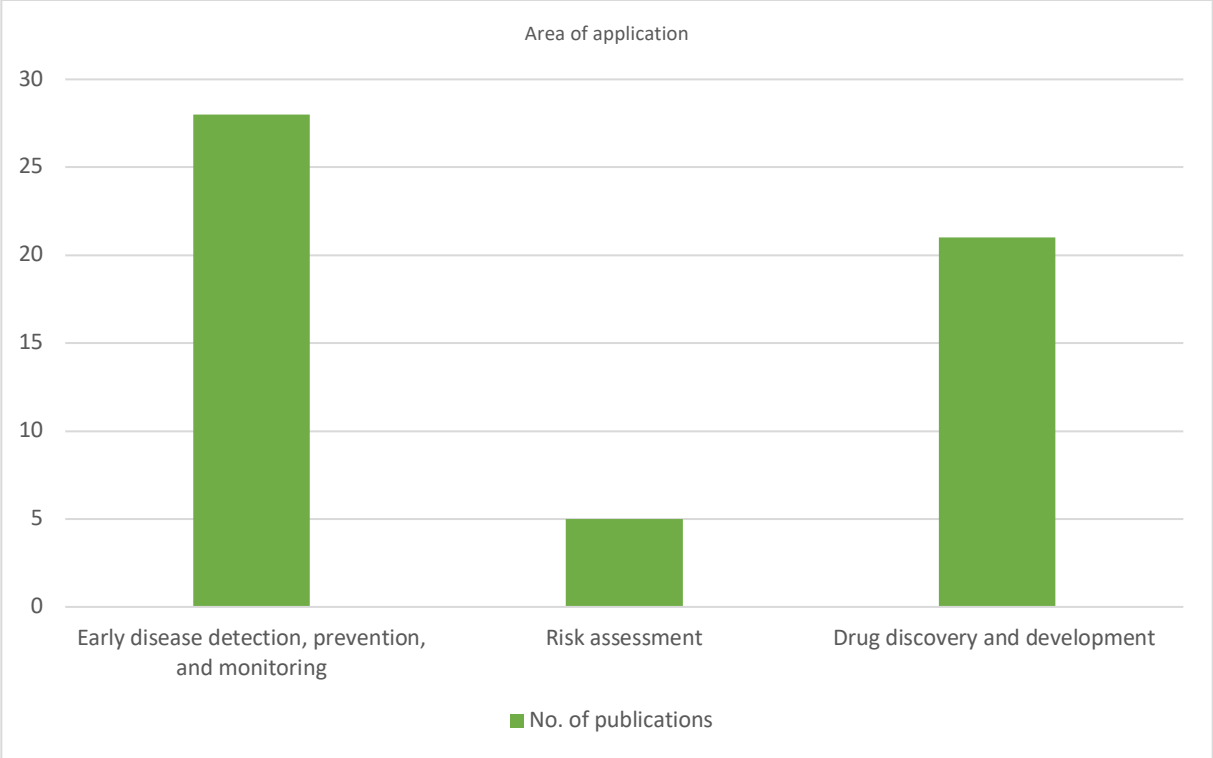


Fig.2.5: Areas of application for biomarker discovery in multi-omics datasets

2.7 Evaluation Criteria

In terms of evaluation, there are several metrics available. These metrics include and are not limited to the residual sum of squares, precision, and *f*-scores. The residual sum of squares is the sum of the squares of residuals (deviation of predicted from actual empirical values of data). It serves as a gauge for the disparity between data and the estimated model. One measure of precision is the proportion of correctly selected biomarkers over the whole set of biomarkers. The recall rate is calculated as the ratio of the number of correctly selected biomarkers to the total number of test biomarkers. The *f*-score is obtained using a harmonic mean between recall and precision. Other metrics, which are introduced based on the nature of the problem and the proposed model, can be established and used to analyze the success of biomarker identification

approaches employing tensor decompositions in multi-omics datasets. We describe the criteria used in the surveyed papers as follows.

- Root Mean Square Error (RMSE) can be formulated as shown in Equation (2.1), where t_i is the test rating value and p_i is the predicted rating value [38,65].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - p_i)^2} \quad (2.1)$$

- Residuals Sum of Squares (RSS) is the measure of the discrepancy between the data and the estimated model. The important point in tensor decompositions is that the trilinear model is found to minimize the RSS. Equation 2.2 shows this metric, where y_i is the *i*th value of the biomarker to be predicted, and $f(x_i)$ is the predicted value of y_i [43].

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2 \quad (2.2)$$

- Precision (p) is the proportion of the relevant features among all retrieved feature sets and assesses the predictive power of a method. Precision can be formulated as follows where $t\mathrm{p}$ is the true positive, and $f\mathrm{p}$ is the false-positive selected cases [43,66].

$$p = \frac{t\mathrm{p}}{t\mathrm{p}+f\mathrm{p}} \quad (2.3)$$

- Recall rate (r) calculates the proportion of the selected features as part of the optimal feature set relative to all features and assesses the effectiveness of an algorithm in identifying the true positive features. Recall can be formulated as follows where tp is the true positive, and fn is the false negative in selected cases [66].

$$r = \frac{t\mathrm{p}}{t\mathrm{p}+f\mathrm{n}} \tag{2.4}$$

- *The f*-score which utilizes precision (p) and recall (r) can be defined as follows. Recall and precision are balanced in the *f*-score when the β constant parameter is set to 1 and is in favor of precision when β > 1 [43,66].

$$f = \frac{(\beta^2+1)pr}{(\beta^2 p)+r} \tag{2.5}$$

- *P*-values are a commonly used criterion used for ranking biomarker candidates and determining the top set of markers considered for further development and validation. Thus, statistical *P*-values can play a fundamental role in the evaluation of biomarker discovery studies. In the case of control studies, the *P*-value associated with a statistic is defined as follows: [63,67]

$$P - value \;=\; Probability\,(statistic \;\geq\; observed\ data\ statistic \mid cases\ same\ as\ controls) \tag{2.6}$$

- Sensitivity and Specificity are two other measures that evaluate the diagnostic performance of a biomarker. Sensitivity is the ability to detect a disease in patients in whom the disease is truly present (i.e., a true positive), and specificity is the ability to rule out the diseases in patients in whom the disease is truly absent (i.e., a true negative) [66,68,69].
- Computation time and cost for biomarker detection in multi-omics datasets is an important evaluation criterion, especially where the problem requires a real-time application or there is a large amount of data for the computation [70].

Figure 2.6 shows the distribution of evaluation criteria used in the reviewed papers. The top 2 criteria are RSS and Precision. However, the majority of the papers used a combination of criteria to enhance their performance evaluation.

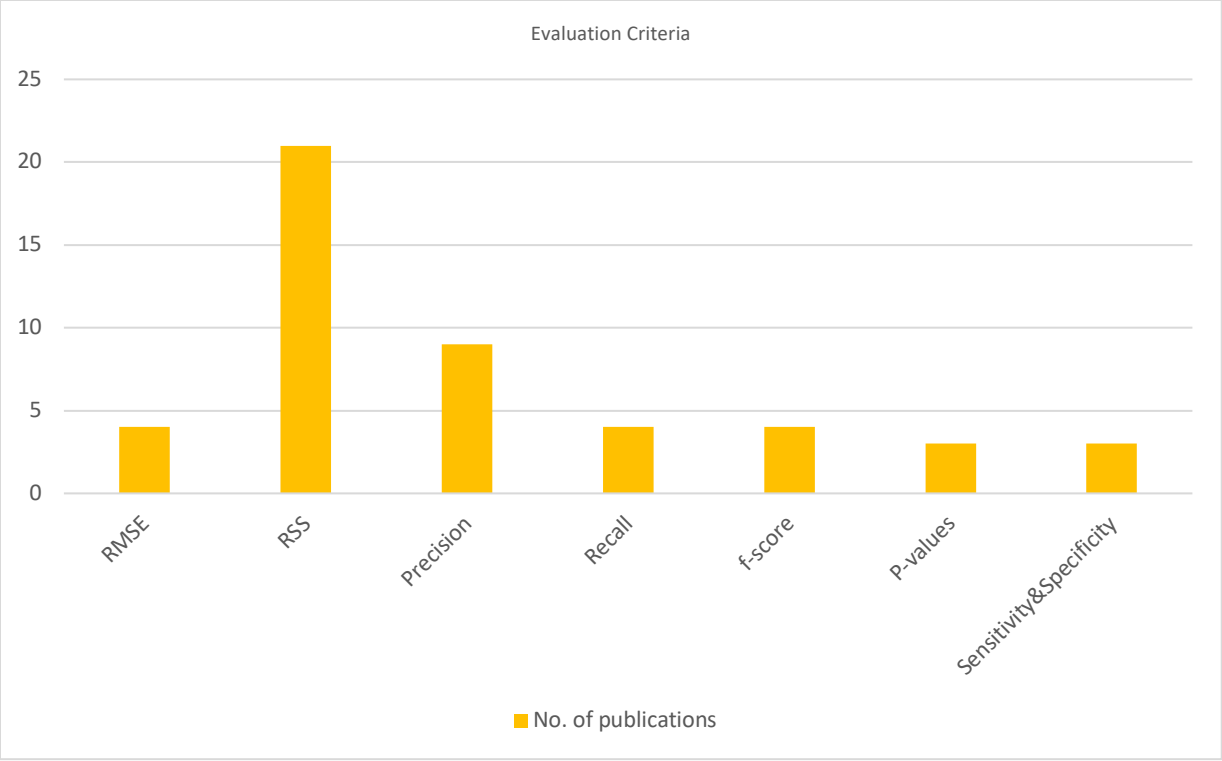


Fig.2.6: Evaluation Criteria

2.8 Conclusion and Future Research Directions

There are many directions for future research. Dealing with long computation times and the associated costs is one of the most significant issues. Depending on the application, a large quantity of data may need to be evaluated in order to design a Tensor Decomposition (TD)-based strategy for biomarker discovery. The majority of research in the literature focuses on the development of solutions for difficulties such as interpretability and scalability; however, they rarely focus on the efficiency of the models. Proposing and developing TD-based strategies for dealing with massive amounts of data from various ‘‘omics’’ types is a research area that has not been extensively investigated. Another issue is that the recommended solutions are often developed for a specific application area. Future research directions might find it interesting to

offer a framework that encompasses several application areas. Recently, some researchers have used neural networks and machine learning to find biomarkers in multi-omics datasets. On this subject, several machine learning and deep learning models may be developed, and their performance can be compared to that of conventional approaches [57,58]. Additionally, different machine learning and deep learning models in terms of chemometrics can be developed in biomarker discovery in multi-omics datasets by tensor decompositions.

2.9 Chapter Conclusions

Recent advancements in "omics" technology have opened up new avenues for researchers to explore disease biomarkers at a systemic level. Extensive efforts have been dedicated to uncovering disease-associated biomolecules by analyzing data from various "omics" experiments, such as genomics, transcriptomics, and metabolomics. In light of these developments since 2015, it was essential to conduct a comprehensive review of the existing research in this field. To achieve this, we established clear goals and research questions to guide our review of biomarker discovery in multi-omics datasets using tensor decompositions. Following a defined protocol, we systematically retrieved and refined relevant articles, carefully reviewing, and analyzing the information extracted from these papers. Our review began by highlighting the challenges and issues that motivate researchers to develop tensor decomposition-based models for biomarker discovery in multi-omics datasets. We delve into the methodologies and models employed by researchers to address these challenges, exploring their strengths and limitations. Additionally, we cover the diverse application fields in which biomarker discovery has been developed. By examining these areas, we gain valuable insights into the

current state of research and the potential impact of biomarker discovery in various domains. Furthermore, we critically discuss the limitations within the field of biomarker discovery in multi-omics datasets using tensor decompositions. This assessment helps to identify areas that require further investigation and improvement. In light of these limitations, we outline future research directions and highlight opportunities that researchers should focus on. By elucidating these future directions, we aim to inspire and guide the scientific community in advancing the field of biomarker discovery, ultimately contributing to improved diagnostic and therapeutic approaches.

Chapter 3

Related Literature on multi-dimensional data analysis in cancer studies

Multi-dimensional data refers to data sets that contain multiple attributes or features. Each attribute represents a different dimension or variable, and the combination of these dimensions creates a multi-dimensional space. This type of data is commonly encountered in various fields, including statistics, machine learning, data mining, and computer graphics [72]. Multi-dimensional data presents challenges such as the curse of dimensionality, visualization difficulties, storage and computational complexity, dimensionality reduction needs, correlation and redundancy issues, noise and outliers, and interpretability concerns. Tensor decomposition is an approach that enhances the capability of multi-dimensional data analysis. It allows for the decomposition of a three-way tensor into lower-dimensional components, enabling dimensionality reduction, feature extraction, and pattern discovery [44]. By leveraging tensor decomposition, meaningful insights can be obtained from complex multi-dimensional data. There are several studies that aim to incorporate tensor decomposition into various applications and domains. These studies utilize Tensor decomposition to analyze and extract insights from multi-dimensional data in fields such as chemometrics [14], and bioinformatics [12].

Tensor decomposition has been increasingly utilized in cancer research to analyze multi-dimensional data. Several studies have employed tensor decomposition to integrate genomics, transcriptomics, proteomics, and other omics data types. By applying tensor decomposition, these studies aim to uncover underlying molecular mechanisms, identify biomarkers, and

enhance our understanding of cancer progression and treatment response. This approach enables the integration of diverse data sources and facilitates comprehensive analysis, leading to improved insights into cancer biology and potential advancements in personalized medicine.

With the purpose of cancer study, research on the integration analysis of multi-dimensional datasets using tensor decompositions can be broadly classified into two major groups. The first group encompasses methods based on canonical polyadic decomposition (or PARAFAC). The second category comprises algorithms based on Tucker3 decomposition.

Deng et al. [90] proposed a novel semi-symmetric PARAFAC decomposition method and introduced the concept of a correlation tensor, which captures spatial correlation. Their aim was to do a breast cancer study that uses images taken of different regions with varying photon wavelengths. Tumor-associated microvesicles (TMVs) are a strong indication of invasive tumors. Unlike other imaging studies, TMVs in breast cancer can appear randomly. To efficiently identify TMVs, the article proposes incorporating pixel correlation in multimodality image analysis. This method efficiently recovers correlation structures among pixels, allowing for the extraction of important features for disease diagnosis, even with limited modalities.

In another paper, the authors of [91] focus on the challenges of analyzing multimodality breast cancer imaging data, specifically the random distribution and heterogeneous patterns of tumor-associated micro-vesicles. To address these challenges, the researchers propose a novel multilayer tensor learning method that incorporates heterogeneity into a higher-order canonical polyadic decomposition. By utilizing subject-wise imaging features and multimodality

information, they develop an approach that efficiently captures the heterogeneous spatial features of signals and integrates multimodality information simultaneously.

Diaz et.al [92] proposed CLIGEN, a computational pipeline for unsupervised subtyping of complex diseases using non-negative PARAFAC decomposition of a binary tensor that combines clinical and somatic mutation patient data. The evaluation of breast cancer subtypes discovered by CLIGEN shows promising results in refining known subtypes and revealing new characteristics, particularly for triple-negative breast cancer. CLIGEN demonstrates the potential for high-throughput subtyping of complex diseases in precision medicine. Additionally, it is found that patient membership proportions in CLIGEN-discovered subtypes are better predictors of survival time compared to data-driven molecular and clinical phenotypes.

In addition, Taguchi et al [45] proposed a tucker3-decomposition based unsupervised feature extraction approach for prostate cancer multi-omics datasets with a large number of features and a small number of samples. Their method outperforms other supervised and unsupervised feature selection methods when applied to synthetic and multi-omics datasets. The genes selected by tucker3-based unsupervised feature extraction are enriched with known tissue-related genes and transcription factors.

In the context of analyzing multi-dimensional image datasets, Lu et al. [93] introduced a new spatial-spectral classification framework based on Tucker3 modeling for hyperspectral imaging in the application of head and neck cancer detection. This method incorporates both spatial and spectral information of the hypercube and performs dimensionality reduction. With the

proposed classification framework, they were able to distinguish between tumor and normal tissue in animal experiments with different tumor sizes.

The authors of [43] proposed a novel multi-omics analysis approach named MONTI (Multi-Omics Non-Negative Tensor decomposition for Integrative analysis). The primary objective of MONTI is to identify multi-omics features that accurately represent specific traits. Their research demonstrates the effectiveness of integrated analysis using multiple omics datasets, particularly in the context of cancer subtyping. To achieve this, they began by integrating diverse multi-omics data in a biologically meaningful manner, resulting in a three-dimensional tensor. This tensor is then subjected to a non-negative PARAFAC decomposition method. Through this process, MONTI effectively identifies subtype-specific multi-omics features that carry significant information. They applied the MONTI method to three case studies involving 597 breast cancer, 314 colon cancer, and 305 stomach cancer cohorts. Remarkably, in all these cases, the classification accuracy of cancer subtypes significantly improved when leveraging the entirety of available multi-omics data.

The studies mentioned above did not utilize a combination of Tucker3 decompositions and ASCA+ [78] with bootstrapping to identify biomarkers in multi-omics datasets specifically for cancer subtyping. To the author's knowledge, this thesis describes the first study to extract multi-omics features, or biomarkers, from the colon cancer dataset to develop a model capable of accurately classifying the four primary subtypes of this cancer. In the next chapter, we introduce our novel approach, outlining how it addresses this research gap and presents a comprehensive methodology for the identification and classification of colon cancer subtypes.

Chapter 4

Methodology- Tested on a published dataset (Blue Crab data)

The complementary nature of 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. ASCA+ is used to determine the significance of experimental factors and their interactions and detect three outliers that were previously undiagnosed using the Mahalanobis Distance method of ASCA. ASCA+ with permutations was used to analyze Tucker3 residuals. A statistically sufficient Tucker 3 model was created by adding factors to the Tucker3 model in a step-wise manner until no ASCA detectable structure was observed in the residuals. The original Tucker3 model with 4×5×2 factors, was insufficient to account for all of the systematic variation in the dataset, whereas the new 3×7×3 Tucker3 model adequately explains the experimental factors A (disease state/region), B (tissue type), and their interaction, A×B. A 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

3×7×3 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 for visualization and interpretation of the model. An ASCA backward elimination strategy is reported to further simplify the Tucker3 3×7×3 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. Figure 4.1 shows the graphical representation of a Tucker3 model of the blue crab dataset.

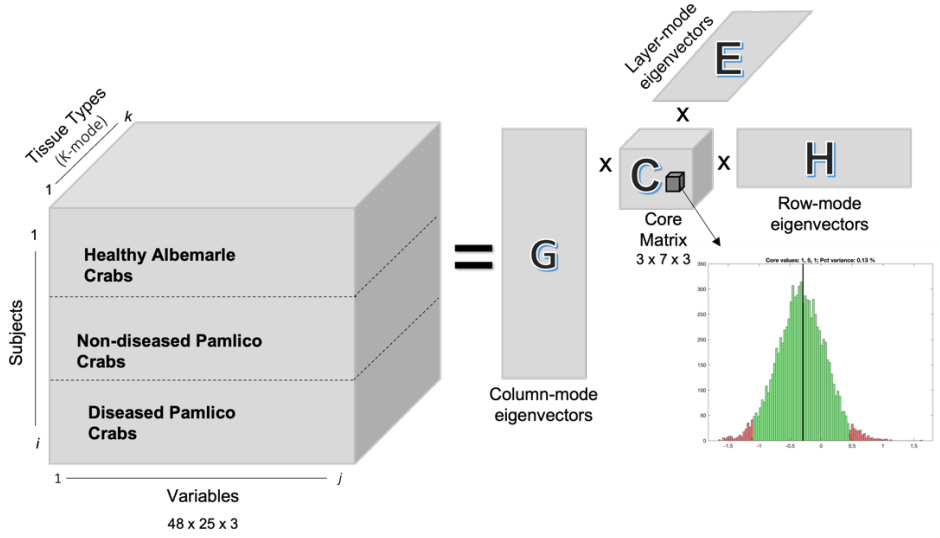


Figure 4.1 A graphical representation of a Tucker3 model of the blue crab dataset.

4.1 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 [71,72]

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 [72]. 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 [73].

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 [74].

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 [75] and also 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.

4.2 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 [75]. A similar issue was discovered in the Saint Johns River near a phosphate strip mine in Florida [76]. 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 [76]. In a study by Gemperline *et al.*, it was hypothesized that environmental stress due to 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 [75].

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) [75]. The dataset was arranged into a three-way array of 48 individuals×25 elements×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 PCA analysis was used to construct a Tucker3 model of rank 4×5×2 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 [77] and confirmed the existence of the clusters that were only 'visually' observed in the original report [75].

4.2.1 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+ method called ParGLM [78] (https://github.com/josecamachop/MEDA-Toolbox) that performs permutation analysis of unbalanced nested designs to determine the statistical significance of experimental factors and their interactions [79]. ASCA is a multivariate extension of the analysis of variance (ANOVA). It 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 [80]. 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 [81].

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 that measures the variance over different disease state/regions.
2. Factor B: Tissue type, three levels (gill, hepatopancreas, and muscle); a fixed factor 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 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 **X** of measurements can be decomposed as follows using ASCA:

$$\mathbf{X} = \mathbf{1}\mathbf{m}^{\top} + \mathbf{A} + \mathbf{B} + \mathbf{C(A)} + \mathbf{AB} + \mathbf{R} \qquad (4.1)$$

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+[79] as implemented in ParGLM[78] to account for the study's unbalanceness. In ASCA+, the original ASCA methodology is extended to unbalanced designs by using general linear models (GLM) to estimate the effect matrices, instead of the classical ANOVA estimators based on differences in means [78,79]

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:

$$\mathbf{X}_i = \mathbf{T}_i \mathbf{P}_i^T + \mathbf{R}_i \quad (4.2)$$

where $\mathbf{T}_i$ and $\mathbf{P}_i^T$ are the scores and loadings for the i^{th} partition, respectively, where a partition, *I*, represents an experimental factor or interaction, and $\mathbf{R}_i$ is the corresponding residual matrix. ASCA is a supervised method where external knowledge about the experimental design is used. Factor A and Factor B in this study have three levels each, so the dimensionality of the PCA visualizations of the effect matrices **A** and **B** is constrained to rank two. Rank four PCA models were used to visualize the interaction matrix, **AB**. Unconstrained permutations on the residuals of reduced ANOVA models was used, the most promising approximate test following Anderson and Braak [82] and implemented for nested designs in ParGLM[78].

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 significant have ASCA detectable structure at the 95% confidence level. It is important to note that permutation tests are only exact for main effects, but approximate tests for interactions have nonetheless been proved to be useful [83,84].

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 original data were calculated for each experimental factor in ASCA, according to [84]. Objects that lie outside of the 95% confidence interval using Hoteling's T^2 distribution were considered to be outliers. Details are provided in the Results and Discussion section.

4.2.2 ASCA+ 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 [85], where three matrices of eigenvectors are computed (orthonormal loading vectors), one for each dimension in the original data table (see Figure 4.2). Equation 4.3 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 48×25×3); *P*, *Q*, and *R* represent the number of factors selected for eigenvectors **G**, **H** and **E** (in this case 3×7×3) and c_{pqr} is an element of the core matrix, **C**, a 3×7×3 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 [85, 86].

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} c_{pqr}\,(g_{ip}h_{jq}e_{kr}) + \varepsilon_{ijk} \tag{4.3}$$

The total variance can be partitioned into two parts according to Equation 4.4.

$$SS_{total} = SS_{fit} + SS_{residual} \quad (4.4)$$

where SS_{fit} is the sum of squares explained by the three-mode model and $SS_{residual}$ is the residual sum of squares. In the previous work [77], 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 about 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 4×5×2 model residuals, indicating that an insufficient number of factors were selected. Details are discussed in the Results and Discussion section.

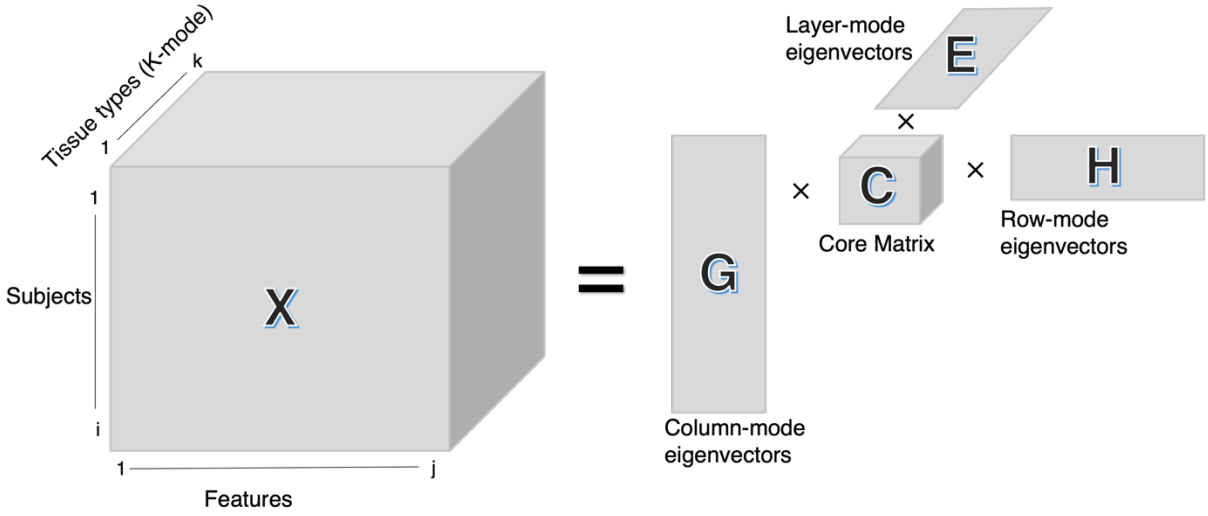


Figure 4.2 Diagram of the Tucker3 model of the dataset.

4.2.3 Bootstrap analysis

There are three major strategies for performing bootstrap analysis: the parametric bootstrap, resampling of residuals, and resampling cases or whole data points [87]. 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 [73] which assumes that the entities in the first mode are a random sample from a population of such entities. Kiers uses resampling of cases (rows) from $\mathbb{X}$ with replacement to produce 'pseudo populations', $\mathbb{X}_b$. In the case of the blue crab data, resampling rows of $\mathbb{X}$ would disrupt the experimental design, e.g., 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.

4.3 Software

ASCA analysis was done 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.

4.4 Results and discussion

4.4.1 Outlier detection

The Mahalanobis distance method was used to detect outlier samples based on ellipsoids at the 95% confidence interval on the ASCA score plots for each main factor and the interactions [84]. The score plots of ASCA on the auto-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, data rows 17, 36, and 128 appear to be outlier objects in both factor A and factor B based on visual inspection (see Figure 4.3). These results are summarized in Table 4.1, showing 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 4.1).

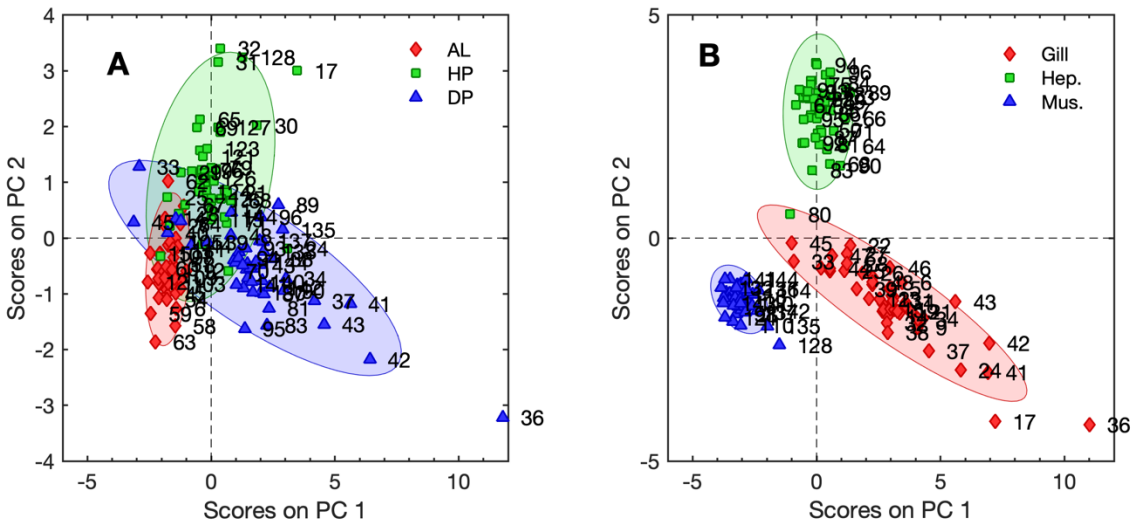


Figure 4.3 Score plots of ASCA on the auto-scaled data. (a) Score plot for factor A (disease state/region), $\mathbf{X}_a$, (b) Score plot for factor B (tissue), $\mathbf{X}_b$.

Table 4.1 Outliers detected using the Mahalanobis distance and probability density (only samples with a probability density < 0.05 are shown) on ASCA plots of factor A (disease state/region), and factor B (tissue).

Factor A			Factor B		
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			

4.4.2 ANOVA-Simultaneous Component Analysis

ASCA+ as described in the Methods section was used to assess the significance of the underlying factors and their interactions. Although the underlying experimental design was unbalanced due to outlier removal, ParGLM was able to accommodate the unbalanced design using general linear models (GLM) to estimate the effect matrices [78,79]. 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%.

Table 4.2 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	1461	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		

Inspection of the table shows that tissue (factor B) is the largest effect accounting for more than 40 of the modeled variation, whereas the disease state/region (factor A) is a much smaller effect (about 11 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 4.2 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 4.2, both the main effects and their interaction are statistically significant ($p < 0.05$), indicating that the interaction is statistically significant, and the concentration of the trace elements in the various tissues is dependent on the population measured.

4.4.3 ASCA analysis of the Tucker3 residuals

Tucker3 models were constructed using the TuckerALS method as noted in the method section. The core matrix value associated with each triad of eigenvectors represents the total variance

explained by the corresponding triad (see Figure 4.1). For Tucker3 models, the total variance can be partitioned into two parts: the sum of squares explained by the three-mode model and the residual sum of squares. In the original paper, a 4×5×2 Tucker3 model was used to explain 70% 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, i.e., the residuals still contained structure that could be associated with the main factors and their interactions indicating that the 4×5×2 model does not sufficiently explain the main factors and interactions. Figure 4.4 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, a result that was not expected.

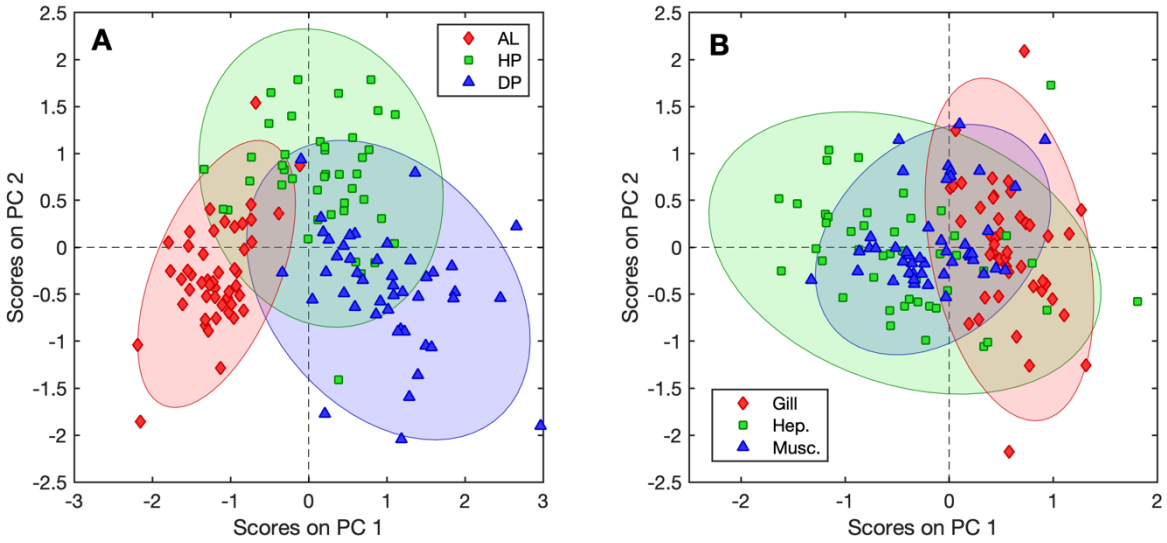


Figure 4.4 Score plots of ASCA on the 4×5×2 Tucker3 model residuals. (a) Score plot for factor A (disease state/region), $\mathbf{X}_a$, (b) Score plot for factor B (tissue), $\mathbf{X}_b$.

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 [88] 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 dataset was the 3×7×3 Tucker3 model with residual variance of 21.97%, compared to ASCA+, 18.15%. The results obtained by ASCA analysis of its residuals show that the *p*-value for factors A, B and A×B were larger than 0.01 ($p = 0.883$, $p = 1$, and $p = 0.953$, respectively), indicating there was not any ASCA detectable structure in the residuals. Figure 4.5 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.

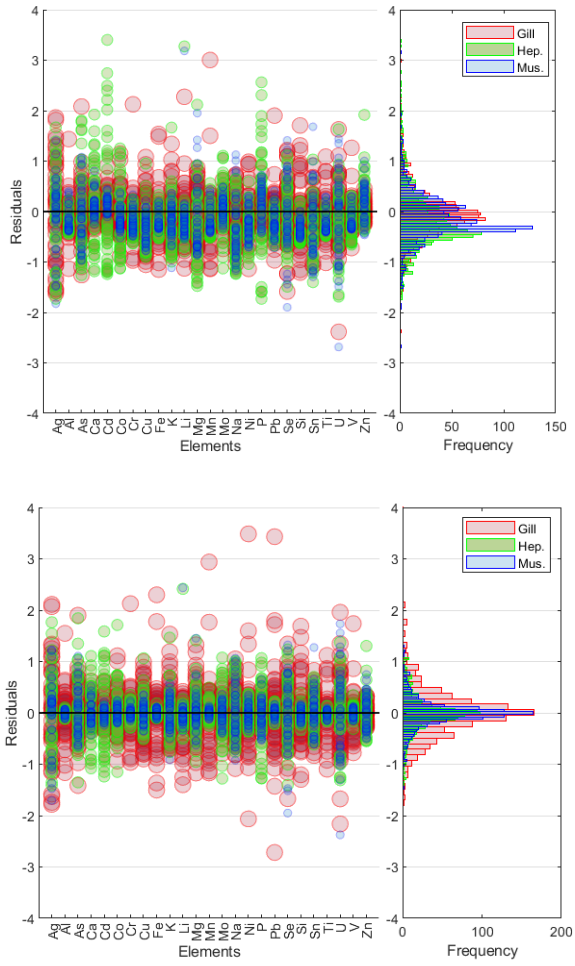


Figure 4.5 Distribution of the residuals for each variable in all three tissue types, top: 4×5×2 Tucker3 model, bottom: 3×7×3 Tucker3 model.

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 so small, that it is unlikely that there is any variation remaining in the ASCA residual matrix or the 3×7×3 model residuals that could be relevant for interpretation.

4.4.4 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 exists in the core values and in corresponding triads of eigenvectors or loadings in Tucker3 models [71]. We also observed this ambiguity in the bootstrap 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, $\mathbb{X}_b$, without having to implement an extensive bookkeeping strategy. As an example, Figure 4.6 shows the bootstrap distribution for core value c_{212} before and after correction.

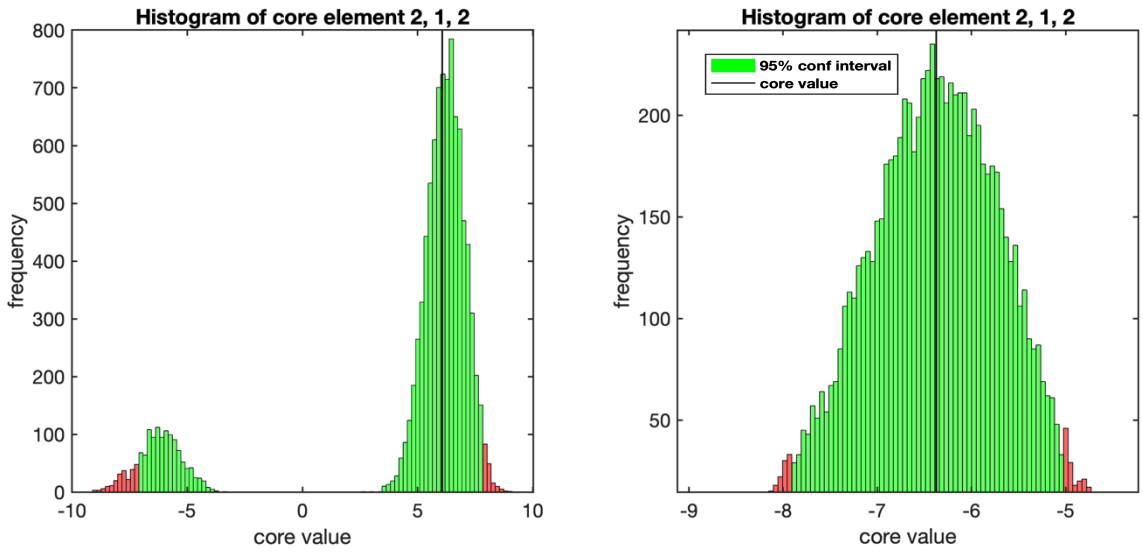


Figure 4.6 The bootstrap distribution for core value c_{212} before (left panel) and after (right panel) sign flipping correction. Green areas indicate core element values within the 95% confidence interval, red areas indicate core element values outside the 95% confidence interval, and the solid line indicates the value of the core element of the original reference model (before bootstrapping).

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 4.6 shows the distribution of the core values of the first three factors. The green region of the histogram lies 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 4.7 represent c_{111} (left panel) and c_{211} (middle panel), reveal 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 done for all

core values. Table 4.3 shows that 21 of the 63 core values are not statistically significant in the 3×7×3 Tucker3 model.

Table 4.3 Statistically insignificant core values determined by bootstrap analysis. H_0: $c_{ijk} = 0$. The 39 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			

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 2-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 3×7×3 model. These removed core values account for only 0.13% of the dataset variance, thus the variance modeled by the 3x7x3 model is decreased from 78.03% to 77.90%.

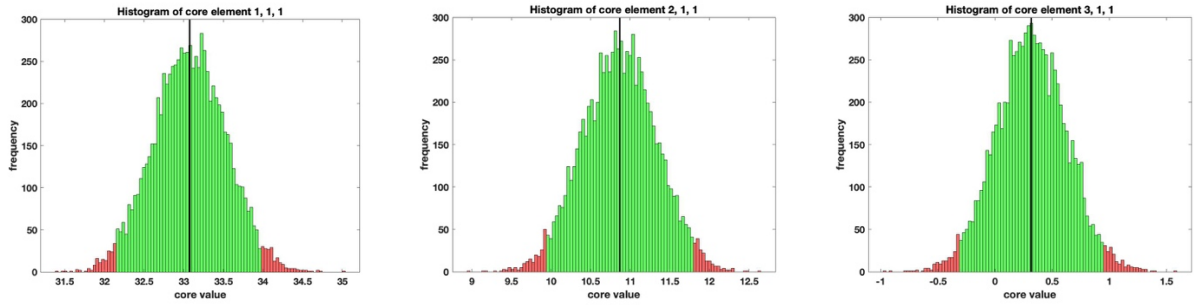
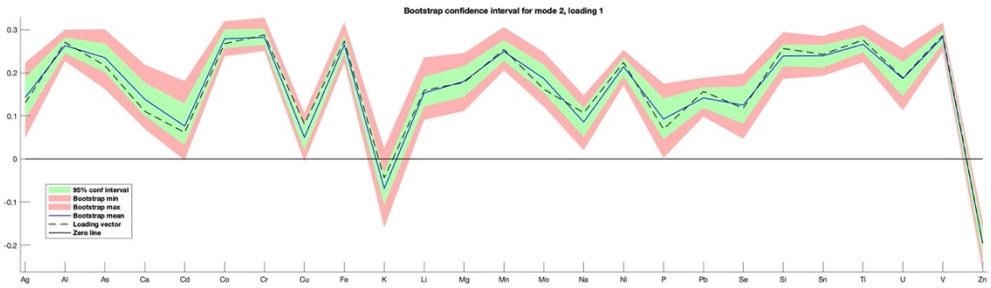


Figure 4.7 Distribution histograms (frequency vs core element value) for the null hypothesis obtained by bootstrap analysis of selected core values. The green region is inside the 95% confidence interval, the red region is outside the 95% confidence interval. The solid line shows the core value of the reference model.

4.4.5 Interpretation of the model loadings

When comparing the 3×7×3 Tucker3 model with 4×5×2 model, the bootstrap confidence intervals for $\mathbf{h}_1$ are narrower with the 3×7×3 model. This is because the residuals for the 3×7×3 model are smaller with no ASCA detectable structure is left in them, and thus it is to be expected that the 4×5×2 residuals give larger confidence intervals compared to the 3×7×3 residuals (Figure 4.8). Eigenvectors associated with small core values are computed with greater uncertainty in the 4×5×2 model, as can be seen in the confidence intervals. The shape of the first loading vector for

the 4×5×2 model compared to the 3×7×3 model is slightly different, although the differences do not seem to be very large except for two of the elements, Mg and Mo. Observing $\mathbf{g}_1$ in the two models (Figure 4.9), again the 4×5×2 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 4×5×2 model, whereas it is nearly symmetrical in the 3×7×3 model. This suggests that the bootstrap confidence interval is approximately normally distributed in the 3×7×3 model whereas it is not in the 4×5×2 model. Looking at the plots of the loadings for $\mathbf{g}_3$ (Figure 4.10), the pattern in the loadings for the 3×7×3 model gives a much cleaner separation of Healthy Pamlico from Diseased Pamlico and Albemarle crabs whereas it is more ambiguous for the 4×5×2 model. Looking at the values of $\mathbf{e}_1$ for both models (Figure 4.11) shows that the values are similar in magnitude and shape, but the confidence interval is much narrower for the 3×7×3 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 4×5×2 model, whereas it is nearly symmetrical in the 3×7×3 model.



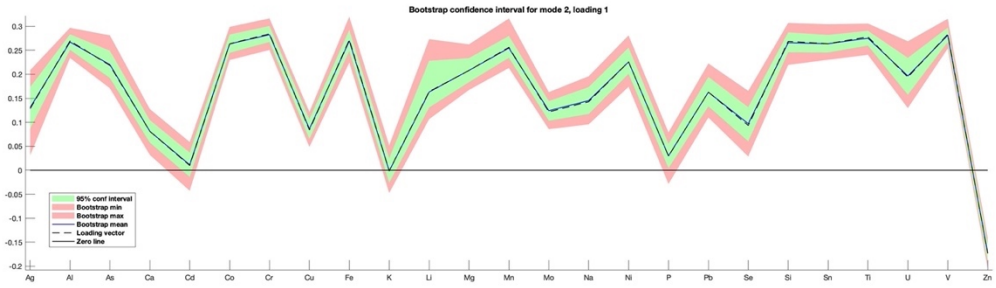


Figure 4.8 Bootstrap confidence intervals for eigenvector $\mathbf{h}_1$, left: 4×5×2 Tucker3 model, bottom: 3×7×3 Tucker3 model

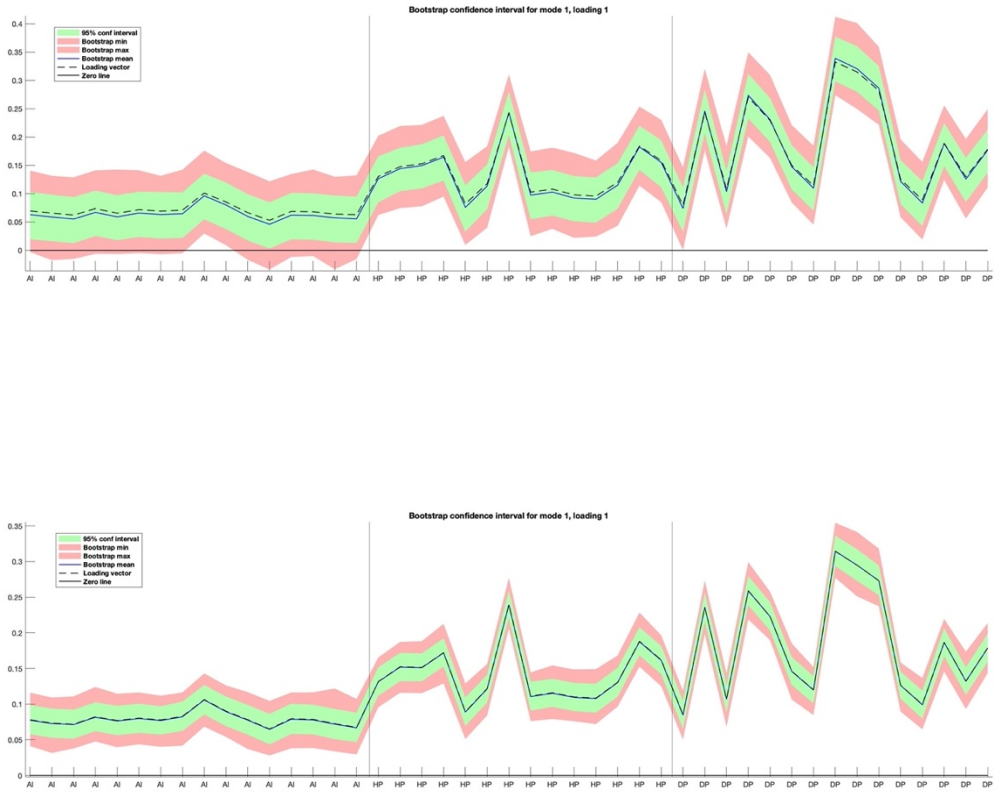


Figure 4.9 Bootstrap confidence intervals for eigenvector $\mathbf{g}_1$, top: 4×5×2 Tucker3 model, bottom: 3×7×3 Tucker3 model

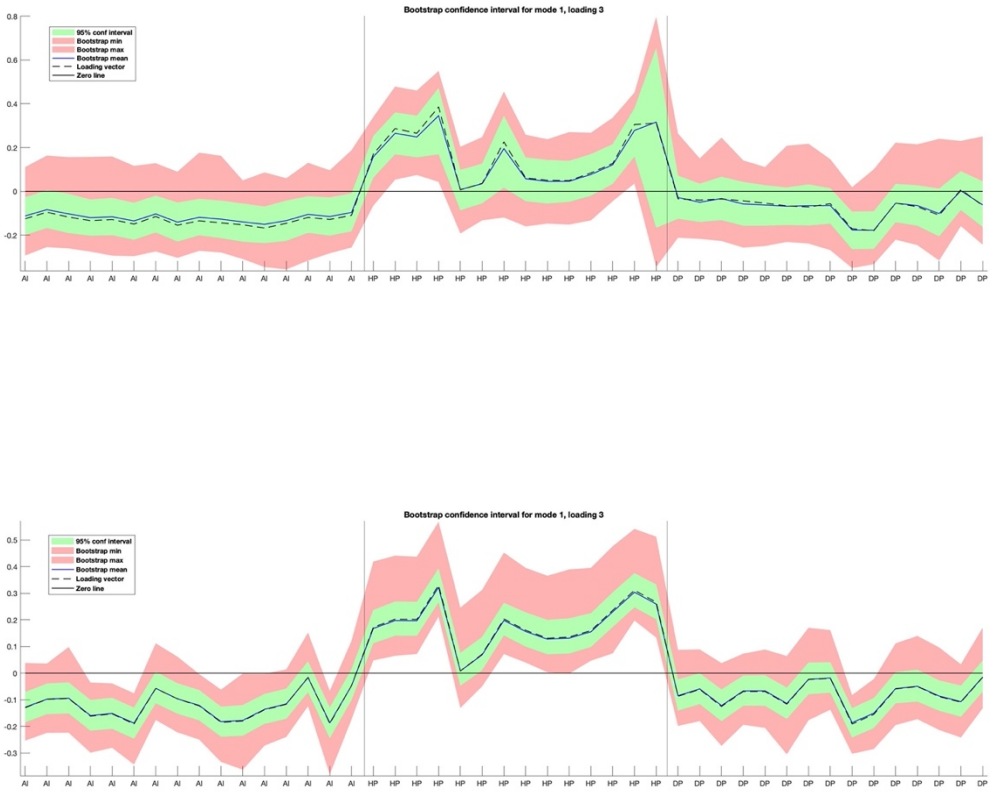
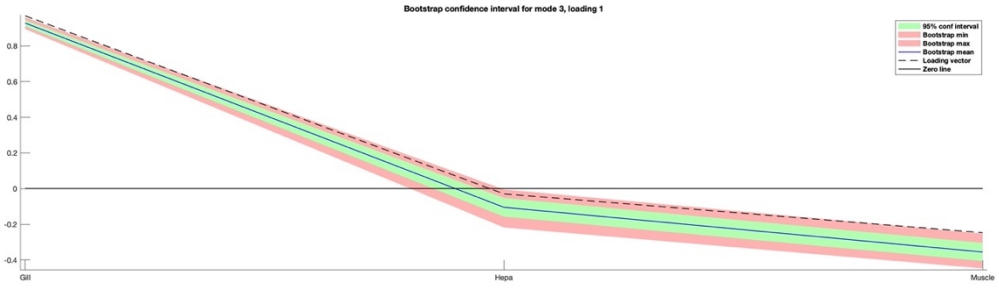


Figure 4.10 Bootstrap confidence intervals for eigenvector $\mathbf{g}_3$, top: 4×5×2 Tucker3 model, bottom: 3×7×3 Tucker3 model



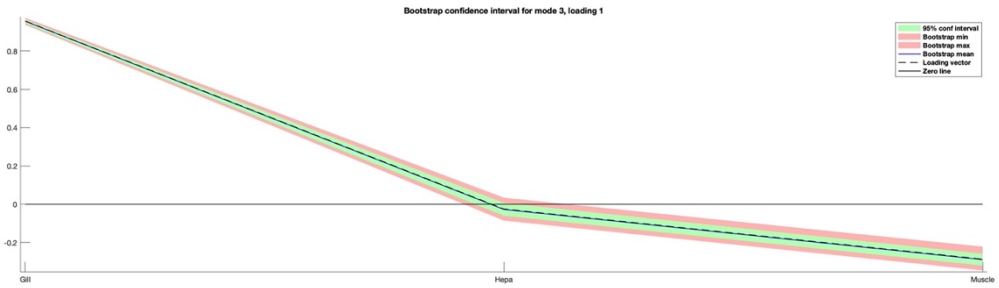


Figure 4.11 Bootstrap confidence intervals for eigenvector $\mathbf{e}_1$, top: 4×5×2 Tucker3 model, bottom: 3×7×3 Tucker3 model

4.4.6 Backwards triad elimination procedure

As described above, bootstrap analysis was used to identify statistically (in)significant core values in the 3×7×3 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 due to factors A, B or A×B in the reduced model. The result is shown in Table 4.4 Core values of the 3×7×3 model are shown, ordered from largest variance explained along with ASCA *p*-values for the reduced models using 10,000 permutations. When the core value c_{111} and its triad of eigenvectors is removed, the reduced 3×7×3 model still has highly significant variance on factors A, B and A×B. Going down the list sequentially, when the 36th core value and its triad is removed we still observe ASCA detectable structure (see Table 4.4). When we remove the 37th core element we no longer observe ASCA detectable structure on factors A, B and the interaction A×B in the reduced model. Continuing in this fashion we find that 36 core values and their associated triads are sufficient to

model effects of factors A, B and A×B. We conclude after backwards elimination that the number of core values can be further reduced from 39 to 36. All removed core values account for 0.38% of the dataset variance, thus the variance modeled by the fully reduced 3x7x3 model with 36 core values retained is decreased from 78.03% to 77.65%.

Table 4.4 ASCA backward elimination procedure. 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.

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

4.4.7 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 are 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 4.12 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 decreasing significance (see Figure 4.12, inset table of middle panel). Looking at the plot of $\mathbf{e}_1$ (see Figure 4.12, 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 loadings for hepatopancreas is not significant. This is consistent with ASCA results which indicate that the triad (factor) c_{111} does not have statistically significant structure for explaining differences in tissues, 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 their response is significant in gill tissue, but less so in muscle and not in hepatopancreas tissue.

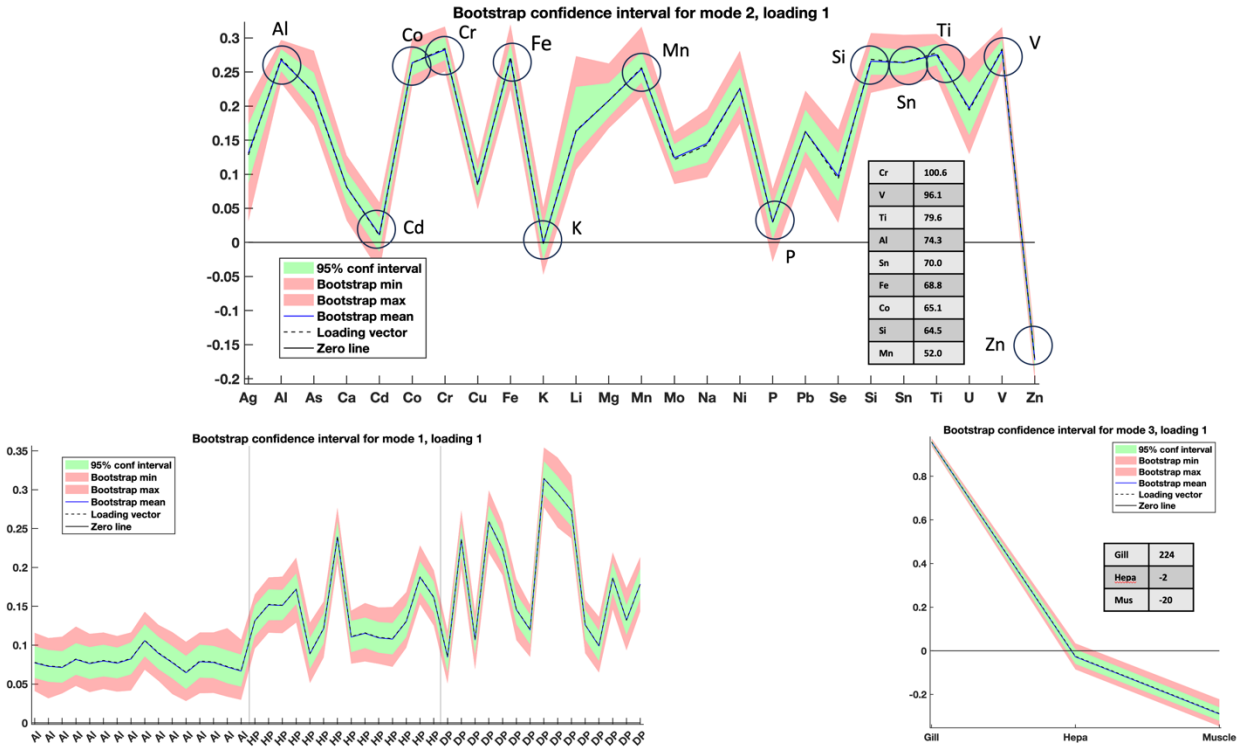


Figure 4.12:Bootstrap confidence interval for the most important triad (factor) $c_{111\times}\mathbf{g}_{1\times}\mathbf{h}_{1\times}\mathbf{e}_{1,}$

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 4.13 middle 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. The response for these physiologically important elements is an important discriminating factor between the control group (Albemarle crabs) and Pamlico crabs (diseased and healthy).

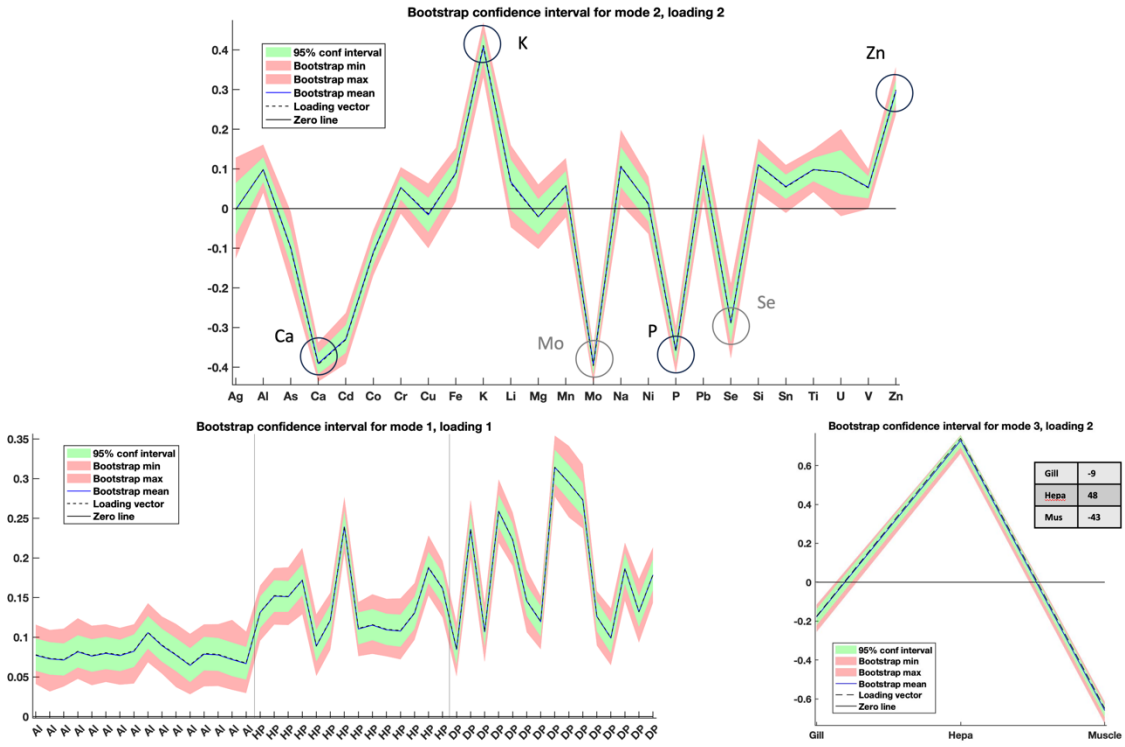


Figure 4.13 Bootstrap confidence interval for the second important triad (factor) $c_{122}\times\mathbf{g}_{1}\times\mathbf{h}_{2}\times\mathbf{e}_{2}$

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) [79]. We also observe that Al is negatively correlated and Pb is slightly positively correlated, and $\mathbf{e}_1$ shows 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.

4.5 Chapter conclusion

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 due to 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, 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 3×7×3 residual variance were similar; 21.97% compared to 18.15%, respectively.

In conclusion, we showed 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 showed 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 4×5×2 model [5], which used 4 factors in **G,** ASCA analysis indicated only 3 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 5 factors were selected whereas 7 were needed to generate a statistically sufficient model. Interpretation of the original 4×5×2 was therefore incomplete, with important relationships between the different element contributions left out. The 3×7×3 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, PARAFAC, etc.) to give a robust estimation in determining a sufficient number of model components, given that the model residuals are assumed to be normally distributed.

Chapter 5

Experimental Results

5.1 Dataset

In this research, we use the Cancer Genome Atlas (TCGA) [https://www.cancer.gov/tcga] which is a comprehensive and publicly available resource that aims to understand the molecular basis of cancer through the analysis of various genomic, epigenomic, transcriptomic, and proteomic data across multiple cancer types. TCGA provides a wealth of information that has greatly contributed to our understanding of cancer biology. Regarding colon cancer, TCGA has generated multi-omics data for colon adenocarcinoma, which is the most common type of colon cancer. This data encompasses several molecular levels, including genomics, transcriptomics, DNA methylation, and proteomics. Here is an overview of the different types of data available in TCGA for colon cancer:

1. Genomics: TCGA has performed whole-exome sequencing (WES) on colon cancer samples, providing information about the coding regions of genes and detecting mutations. This data allows researchers to identify genetic alterations, such as somatic mutations, copy number variations (CNVs), and structural variations, which may contribute to colon cancer development and progression.
2. Transcriptomics: TCGA has generated RNA sequencing (RNA-seq) data, which provides information about gene expression levels in colon cancer samples. This data helps identify differentially expressed genes between cancerous and normal tissue, as well as molecular

subtypes of colon cancer. It can also reveal potential therapeutic targets and molecular pathways involved in the disease.

3. DNA Methylation: TCGA has also performed DNA methylation profiling using array-based technologies, such as the Illumina Infinium platform. DNA methylation is an epigenetic modification that can regulate gene expression. Methylation data from TCGA allows researchers to identify differentially methylated regions associated with colon cancer, providing insights into the epigenetic alterations involved in the disease.
4. Proteomics: TCGA has generated proteomic data using mass spectrometry techniques to measure protein expression levels in colon cancer samples. This data complements the transcriptomic data and can provide additional insights into the functional consequences of gene expression changes.

In this study, we conducted a comprehensive analysis of three distinct omics datasets obtained from 315 patients diagnosed with colon cancer. The datasets utilized include gene expression, miRNA expression, and DNA methylation profiles. Each dataset contains patient IDs for easy identification and integration of the data. Among the colon cancer patients in our study, we identified four main cancer subtypes, each exhibiting distinct molecular characteristics. The first subtype, CMS1, consisted of 43 patients, while the second subtype, CMS2, comprised 125 patients. The third subtype, CMS3, included 48 patients, and the fourth subtype, CMS4, comprised 99 patients [43].

Based on the previous study [43], only primary tumor samples that had complete matching omics datasets were chosen for the analysis. The gene and miRNA expression values from TCGA, which were already pre-quantified, were directly utilized. Subtype information was obtained from the original studies [43]. Our main objective was to uncover gene regulatory multi-omics features, so each omics dataset was individually processed to create a gene-centric two-dimensional sample (patient)-gene matrix. For this, values within each omics matrix were computed and associated with their respective genes. To ensure a uniform tensor structure, all slices needed to have the same size. Therefore, although each omics matrix underwent separate processing, they shared the same set of genes and samples [43]. The gene expression values were prepared for analysis using TCGA level 3 gene expression data, and a log2 quantile normalization process was applied across the samples. As for miRNA data, they were organized into bundles based on target genes, aligning the number of bundles with the number of genes. The geometric mean of miRNA expression within each bundle was then assigned to its corresponding gene. The miRNA expression values were also log2 quantile normalized. Regarding methylation data, probes located within the transcription start site and 2 Kb upstream of gene promoter regions were grouped together per gene. The average methylation level per gene was subsequently quantile normalized. Since tensor decomposition requires consistency in the range of omics values, each matrix was scaled within a common range. This step was essential to prevent an imbalance where omics matrices with significantly larger values, like gene expression, could overpower other matrices with comparatively lower values. To achieve this, all normalized matrices were further scaled to a range of 0 to 1. In the final step, the processed omics matrices were combined along an orthogonal axis, creating a three-dimensional tensor structure [43].

Figure 5.1 displays a surface plot representing a subset of the entire dataset containing 500 randomly selected genes (variables) after undergoing preprocessing.

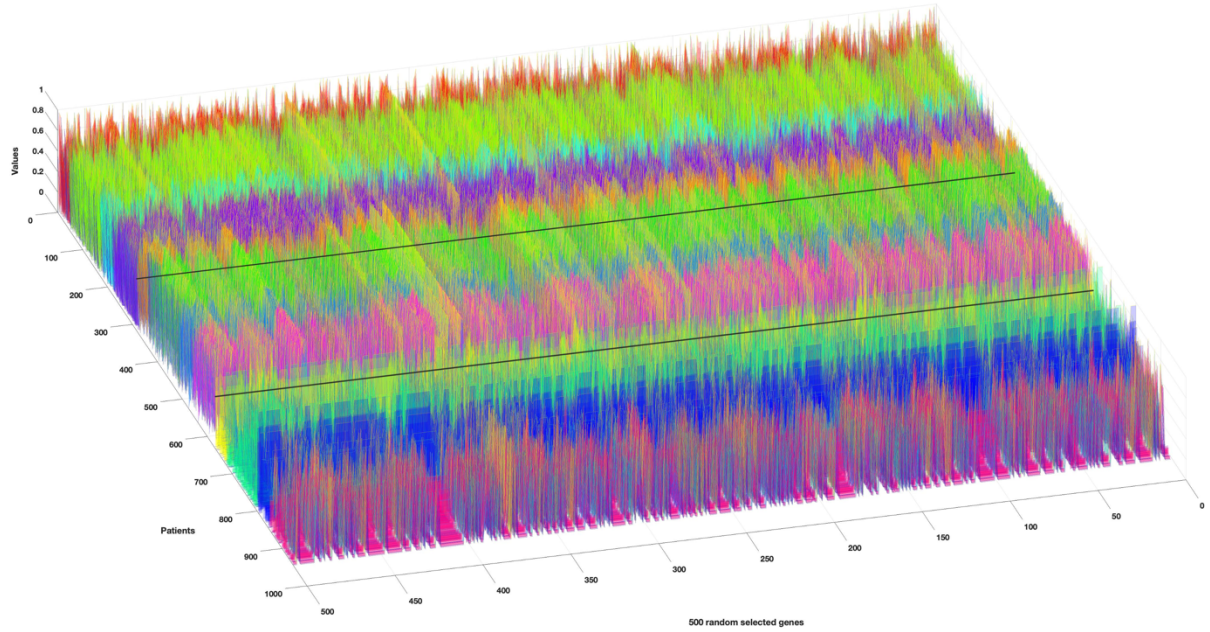


Figure 5.1 A surface plot of a subset of 500 randomly selected genes from the whole dataset.

5.2 TUSCA (Tucker3+ASCA), and Bootstrap analysis results

ASCA+ was employed to evaluate the significance of the underlying factors and their interactions. Despite the experimental design being unbalanced due to varying patient numbers in each subtype, ParGLM managed to handle the unbalanced nature using general linear models (GLM) for effect matrix estimation [78,79]. Table 5.1 presents the sum of squares and *p*-values for the various factors in the blue crab dataset.

Table 5.1 Sum of Squares and their *p*-values (10,000 permutations) by the different effects in ASCA.

Source	SumSq	*p*-value
Mean	82123	-
A: Disease state/subtype	434.04	9.999e-05
B: Omics	22002	9.999e-05
Interaction A×B	820.41	9.999e-05
Residuals	7925.2	-

Inspection of the table shows that omics (factor B) is the largest effect, whereas the disease state/subtype (factor A) is a much smaller effect. In general, the ASCA results reported in Table 5.1 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 5.1, both the main effects and their interaction are statistically significant ($p < 0.05$), indicating that the interaction is statistically significant, and the expression level of each gene in the various omics is dependent on the population measured.

The TUSCA (Tucker3+ASCA) method (described in chapter 4) was implemented to find the optimum 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 15 factors for *P*, 1 to 25 factors for *Q*, and 1 to 3 factors for *R*. 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 dataset was the 12×20×3 Tucker3 model with explained variance of 52.19%. Bootstrap analysis (10,000 randomizations) was performed to determine the significance of the Tucker3 core values.

In our study, the 12×20×3 Tucker3 model has a total of 720 core values, out of which 367 core values were found to be statistically significant at a significance level of $p < 0.05$. Table 5.2 presents the ten most important triads (factors) ranked by the amount of variance they explain. In the upcoming section, we focus on the interpretation of the two largest factors;, the triad $c_{111\times}\mathbf{g}_{1\times}\mathbf{h}_{1\times}\mathbf{e}_1$, which accounts for the largest amount of variance explained and the second significant factor, $c_{322\times}\mathbf{g}_{3\times}\mathbf{h}_{2\times}\mathbf{e}_2$.

Table 5.2 Statistically insignificant core values determined by bootstrap analysis (99% confidence level).

Core value (c_{ijk})	Explained variance (%)	p-value
1,1,1	22.778	0
3,2,2	3.470	0
4,3,3	2.082	0
2,4,2	0.837	0
2,5,3	0.754	0
7,6,3	0.708	0
2,2,2	0.668	0
4,4,3	0.531	0
3,1,2	0.482	0
2,3,3	0.454	0

5.3 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/subtype), but not factor B (omics). Interpretation of the individual vectors of triads are aided by this knowledge. This triad is of particular interest due to the vector $\mathbf{g}_1$, which exhibits some discriminative properties among different subtypes. Notably, CMS3 patients demonstrate high values in this triad, as illustrated in Figure 5.2. However, it should be noted that there is a large offset in the whole dataset because of the q-norm preprocessing method which is captured in this tria, thus the interpretation of this loading is not helpful.

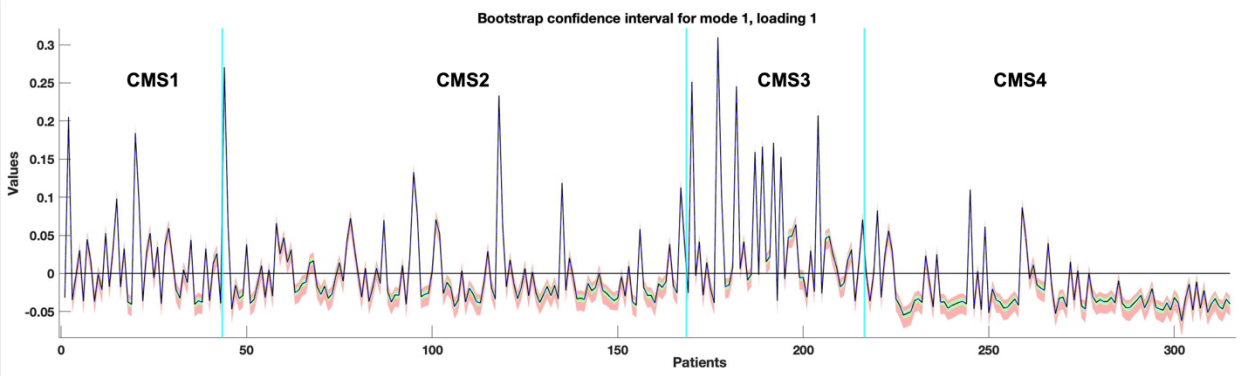


Figure 5.2 Bootstrap confidence intervals for eigenvector $\mathbf{g}_1$

In vector $\mathbf{h}_1$, we see (Figure 5.3) a large offset which is necessary to account for offsets introduced by q-norm scaling whereas autoscaling as seen in the blue crab data does not introduce offsets. For this reason, it is difficult to interpret the $\mathbf{h}_1$ loading.

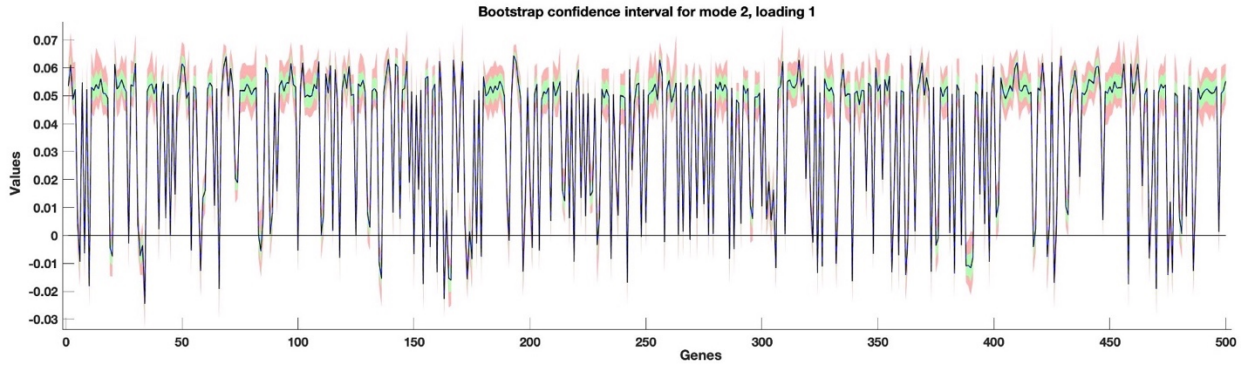


Figure 5.3 Bootstrap confidence intervals for eigenvector $\mathbf{h}_1$

Looking at the plot of $\mathbf{e}_1$ (see Figure 5.4) it can be seen that the value for miRNA omic is very large and the confidence interval is narrow, indicating it is highly significant, whereas the value of the loading for methylation is not significant. This is consistent with the ASCA results which indicates that the triad (factor) c_{111} does not have statistically significant structure for explaining differences in omics, Factor B.

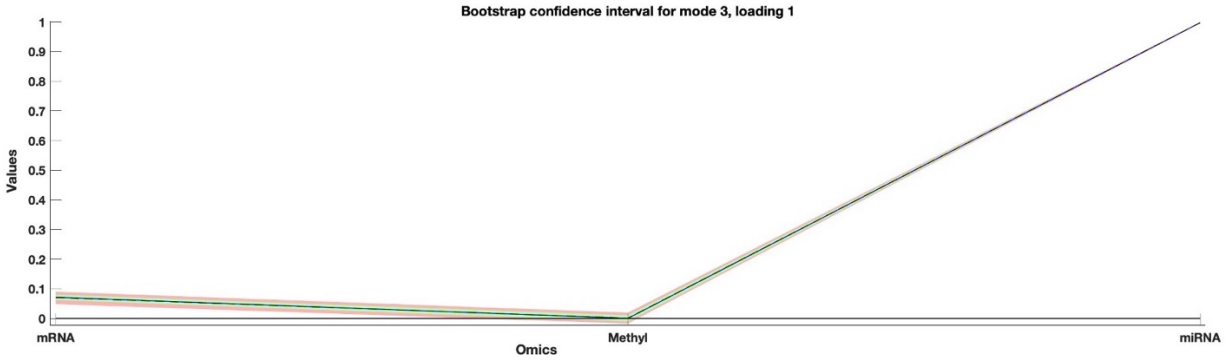
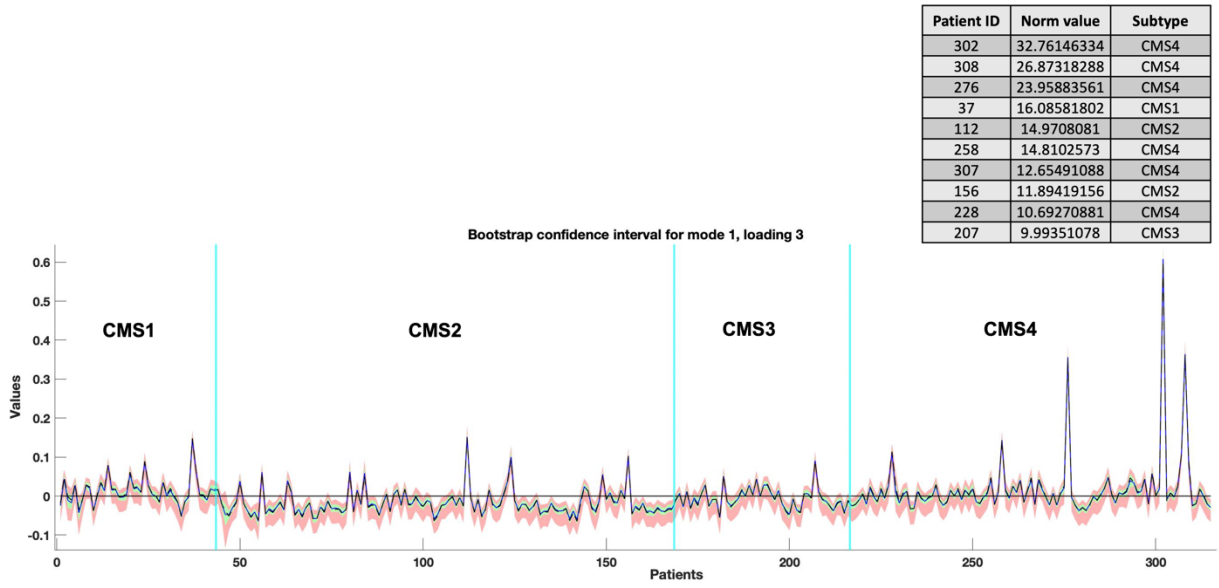


Figure 5.4 Bootstrap confidence intervals for eigenvector $\mathbf{e}_1$

For the second most important triad, c_{322}, $\mathbf{g}_3$ shows some CMS4 patients demonstrate high values in this triad, as illustrated in Figure 5.5. In this vector, the top 10 patients with high values and narrow confidence intervals are identified, indicating they are highly significant in this triad. Dividing the value of each patient in $\mathbf{g}_3$ by the range of the 95% confidence level of the bootstrap range (range = $x_{upper} - x_{lower}$) and sorting them allows one to rank them in order of decreasing significance (see Figure 5.5, inset table).



Patient ID	Norm value	Subtype
302	32.76146334	CMS4
308	26.87318288	CMS4
276	23.95883561	CMS4
37	16.08581802	CMS1
112	14.9708081	CMS2
258	14.8102573	CMS4
307	12.65491088	CMS4
156	11.89419156	CMS2
228	10.69270881	CMS4
207	9.99351078	CMS3

Figure 5.5 Bootstrap confidence intervals for eigenvector $\mathbf{g}_3$

In vector $\mathbf{h}_2$, we have identified 10 highly significant genes, as highlighted in Figure 5.6. These findings underscore the potential importance of vector $\mathbf{h}_2$ in our analysis. In the literature, gene AC156455.1 has been identified as a potential biomarker for diagnosis and prognosis in colorectal cancer and was upregulated in colon cancer tissues compared with adjacent normal tissues [94]. Also, another point is that the other detected significant genes in this triad have low cancer

specificity which means these biomarkers are less exclusive to cancer cells and may also be present in non-cancerous conditions or healthy tissues [https://www.cancer.gov/tcga]

Gene ID	Norm value	Label
193	23.12049399	SPINT2
364	23.12049399	CFL1P8
421	17.68779763	EID1
429	16.39259407	AC156455.1
316	16.39080488	AC138409.2
194	16.33990237	CABP2
69	15.96478384	KDM4C
171	15.9603988	ACOX1
32	15.86251818	ICAM3
102	15.86250877	SLC12A4

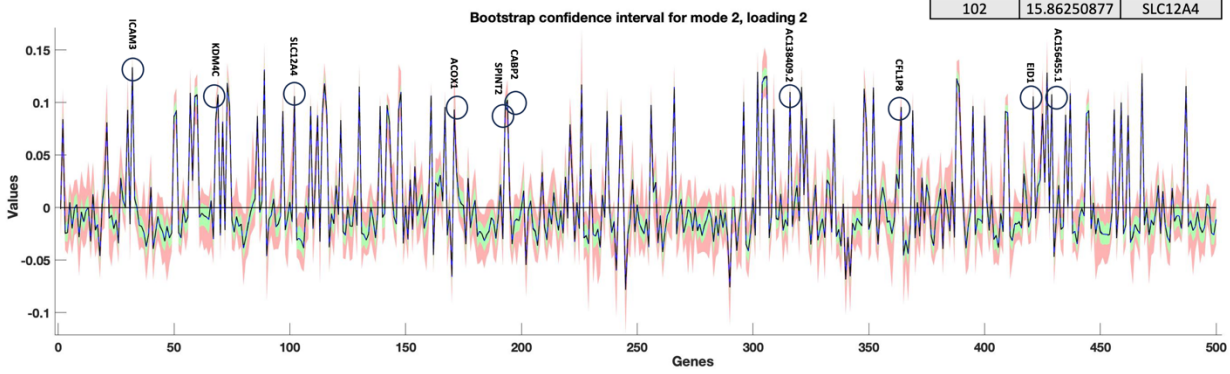


Figure 5.6 Bootstrap confidence intervals for eigenvector $\mathbf{h}_2$

Looking at the plot of $\mathbf{e}_2$ (see Figure 5.7) it can be seen that the value for methylation is very large and the confidence interval is narrow, indicating it is highly significant, whereas the values for mRNA and miRNA are not significant. Again, this is consistent with the ASCA results, which indicates that the triad (factor) c_{322} does not have statistically significant structure for explaining differences in omics, Factor B.

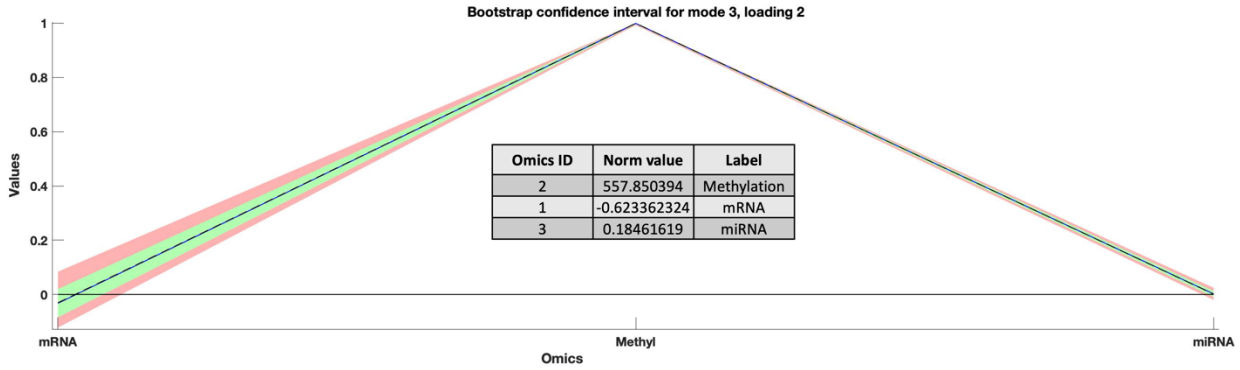


Figure 5.7 Bootstrap confidence intervals for eigenvector $\mathbf{e}_2$

In summary, this triad models the response for genes listed above, which exhibit strong correlations in methylation sites among the top 10 patients and additional patients with large absolute value scores and narrow bootstrap range not listed here.

For the third most important triad, c_{433}, $\mathbf{g}_4$ shows some CMS1 and CMS4 patients demonstrate high values in this triad, as illustrated in Figure 5.8. In this vector, the top 10 patients with high values and narrow confidence intervals are listed, indicating they are highly significant in this triad (see Figure 5.7, inset table). None of the top 10 patients in $\mathbf{g}_4$ are common to the top 10 in $\mathbf{g}_1$, $\mathbf{g}_2$, or $\mathbf{g}_3$.

Patient ID	Norm value	Subtype
6	28.2791857	CMS1
46	25.57033697	CMS2
55	23.05463365	CMS2
229	20.88724481	CMS4
3	16.61304072	CMS1
4	15.8321915	CMS1
1	13.32069975	CMS1
221	10.0405666	CMS4
249	-8.117555645	CMS4
259	-8.010229333	CMS4

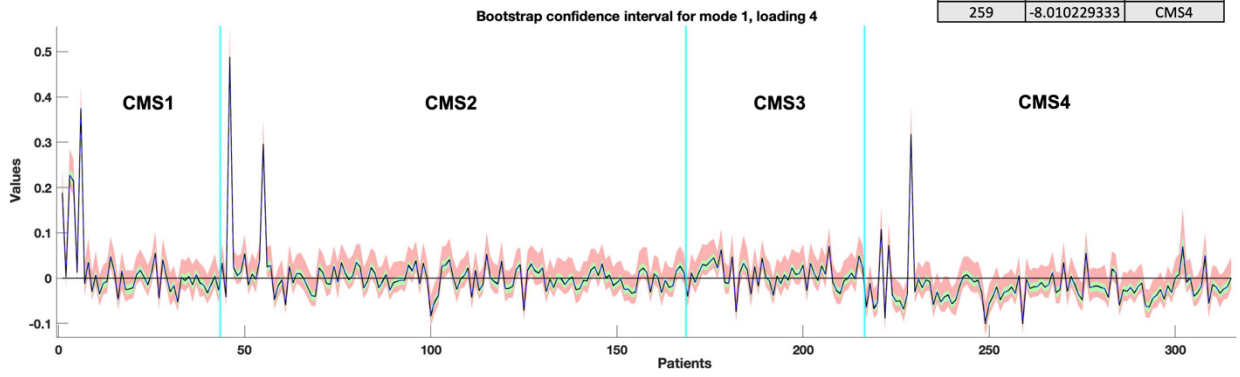


Figure 5.8 Bootstrap confidence intervals for eigenvector $\mathbf{g}_4$

In vector $\mathbf{h}_3$, we again have identified 10 highly significant genes, as highlighted in Figure 5.9.

Gene ID	Norm value	Label
371	21.11346293	AC018442.1
87	15.61019068	ERBIN
403	15.37340259	AC106772.2
20	13.85559491	HDAC7
120	12.94814747	ZDHHC4
180	12.66602905	FAM208A
156	12.41590226	VPS26B
273	11.92484975	RF00019
155	11.65056047	KCTD14
111	11.41443858	NES

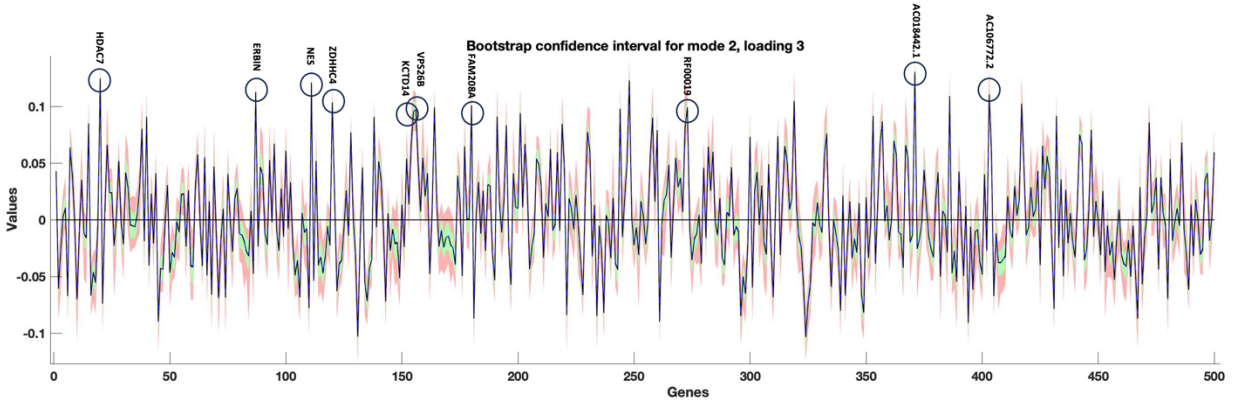


Figure 5.9 Bootstrap confidence intervals for eigenvector $\mathbf{h}_3$

Looking at the plot of $\mathbf{e}_3$ (see Figure 5.10) it can be seen that the value for mRNA is very large and the confidence interval is narrow (normed value 498.95), indicating it is highly significant, whereas the normed value for methylation on this loading is not significant at the 95% CL. Although miRNA value is negatively correlated with mRNA (-8.0 normed value), it is much less significant, thus the top 10 genes identified in $\mathbf{h}_3$ (see below) have significant correlated responses with miRNA but not gene methylation and only slight correlation with gene expression (mRNA).

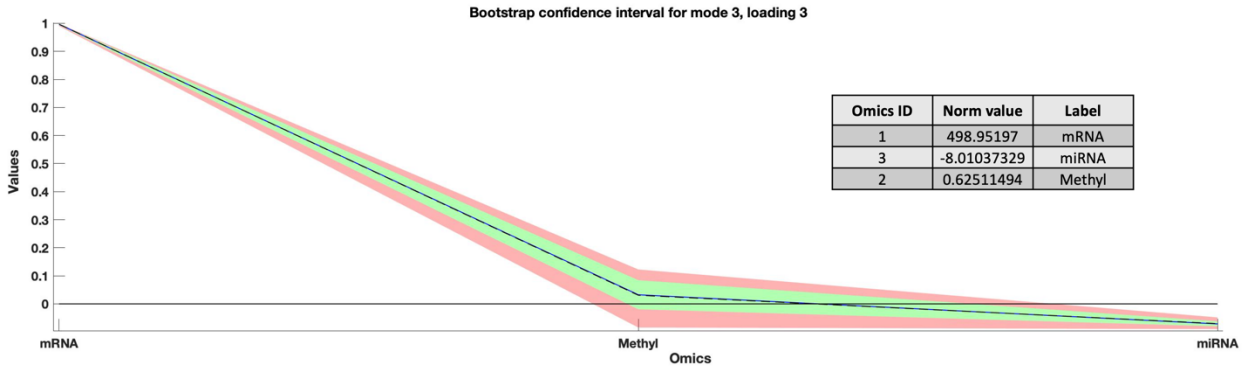
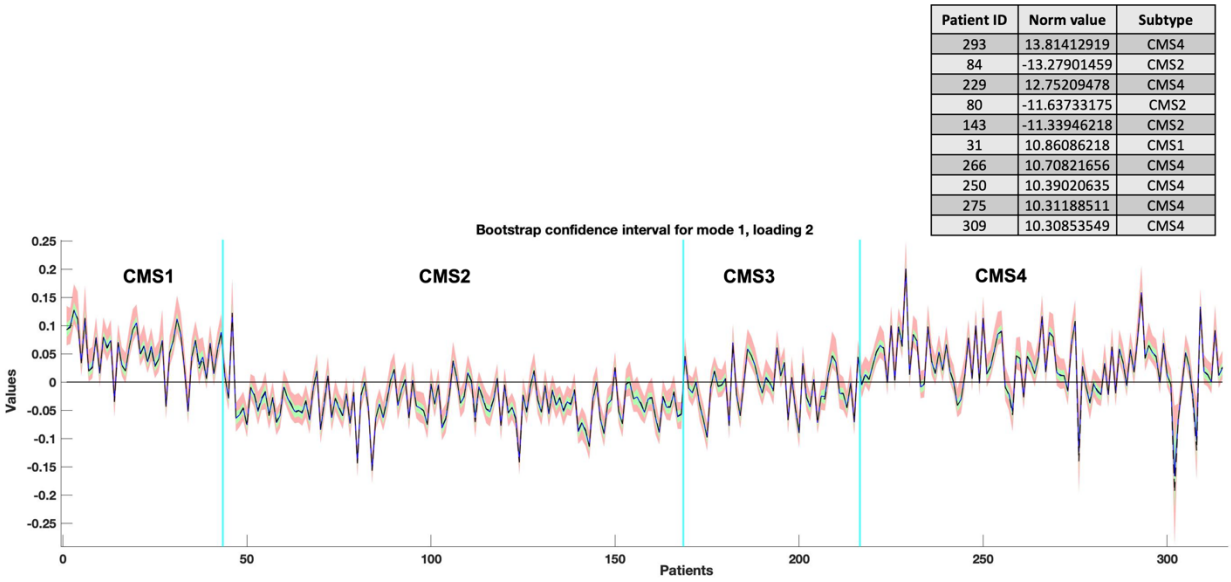


Figure 5.10 Bootstrap confidence intervals for eigenvector $\mathbf{e}_3$

It is interesting to look at the fourth most important triad as well, c_{242}, where $\mathbf{g}_2$ shows a different pattern for each subtype (see Figure 5.11). Also, in $\mathbf{g}_2$, patients in subtypes CMS2 and CMS4 demonstrate high values in this triad, as illustrated Figure 5.7, inset table.



Patient ID	Norm value	Subtype
293	13.81412919	CMS4
84	-13.27901459	CMS2
229	12.75209478	CMS4
80	-11.63733175	CMS2
143	-11.33946218	CMS2
31	10.86086218	CMS1
266	10.70821656	CMS4
250	10.39020635	CMS4
275	10.31188511	CMS4
309	10.30853549	CMS4

Figure 5.11 Bootstrap confidence intervals for eigenvector $\mathbf{g}_2$

In vector $\mathbf{h}_4$, we also have identified 10 highly significant genes, as highlighted in Figure 5.12. Interestingly, gene AC024560.2 has been identified as a potential biomarker and may provide broader perspective for combating cervical cancer metastasis [95].

Gene ID	Norm value	Label
496	-13.28730231	AC024560.3
288	12.25557349	RNU6-468P
416	10.44491262	AC100810.2
422	-9.919266526	AP000867.3
483	9.511623033	SIGLEC27P
443	9.244110088	CES1P2
153	8.875457807	OR5M9
96	8.404918264	MLH3
406	8.09101187	PRODH2
238	8.072892623	MUC6

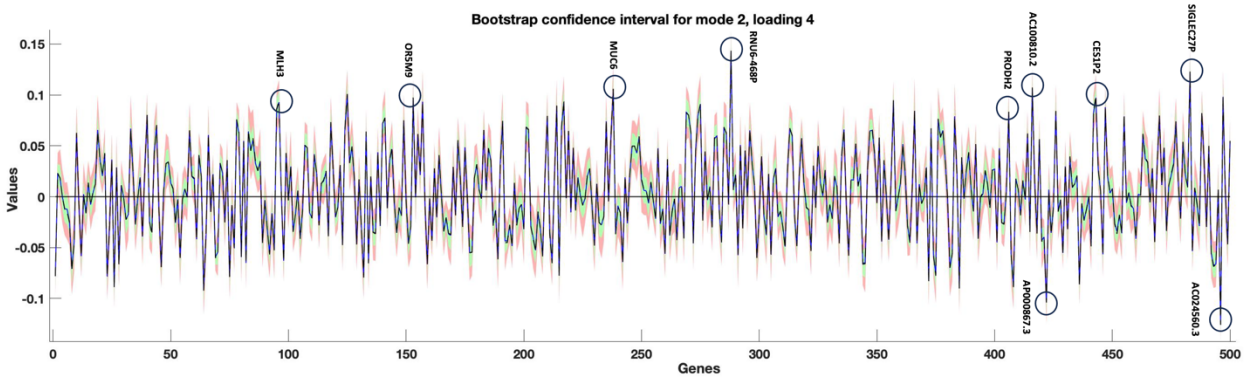


Figure 5.12 Bootstrap confidence intervals for eigenvector $\mathbf{h}_4$

Again, in this triad, by looking at the plot of $\mathbf{e}_2$ (see Figure 5.7) it can be seen that the value for methylation is very large and the confidence interval is narrow, indicating it is highly significant, whereas the normed values for mRNA and miRNA on this loading are not significant.

The analysis can be extended to all the remaining triads, but those results are not presented in this study. Nevertheless, it is worth emphasizing the strength of using Tucker3 triads, which are orthogonal and can be independently analyzed in conjunction with ASCA. This combination of tools is powerful and aids in interpreting the model and results. It helps identify crucial multi-omics features (biomarkers) associated with the factors considered in the experimental design. Also, our analysis shows that the three omics types in this data set, mRNA, methylation, and miRNA have a natural tendency to be mutually orthogonal. Considering that this analysis was a random sample of about 3% of more than 14,000 genes, analysis of the full data set may reveal

some correlations between omics types that were not represented in the 500 genes selected for this study.

Chapter 6

Future Work and Conclusion

Biomarker discovery has become pivotal in the fight against illnesses such as cancer. Scientists can examine different biological cues in patients to pinpoint distinct molecules, genes, proteins, or attributes that function as dependable biomarkers for detecting diseases early, forecasting outcomes, and tailoring personalized treatment approaches. In this thesis, we first systematically defined goals and questions for reviewing the research studies conducted regarding biomarker discovery in multi-omics datasets using tensor decompositions. We followed a defined protocol to retrieve and refine articles, and reviewed and analyzed the information extracted from the papers. We also introduced challenges and problems that motivate researchers to develop tensor decomposition-based methods for biomarker discovery in multi-omics data. Additionally, we presented the methodologies and models they employ to tackle these challenges. The application fields for biomarker discovery have been covered in this thesis. Last but not least, we discussed the limitations in the field of biomarker discovery using tensor decompositions and describe future research directions and opportunities that require researchers' focus.

Additionally, we proposed a novel approach to demonstrate the complementary nature of ASCA and Tucker3 tensor decompositions on design datasets. We introduced ASCA+ to achieve three main objectives: (a) identifying statistically sufficient Tucker3 models, (b) pinpointing important triads for easier interpretation, and (c) eliminating non-significant triads to facilitate visualization and interpretation. To perform this analysis, we applied ASCA+ to the unfolded matrix in datasets with at least two factors, while employing Tucker3 modeling on the folded tensor. We introduced

innovative strategies for evaluating the statistical significance of Tucker3 models, utilizing a published dataset. To further enhance our analysis, we conducted a bootstrap analysis of the Tucker3 model residuals, enabling us to determine confidence intervals for the loadings and individual elements of the core matrix. Finally, we applied the entire procedure to a large cancer dataset, successfully identifying multi-omics features (biomarkers) with significant potential for cancer research. In conclusion, our proposed method showcases the synergy between ASCA and Tucker3 tensor decompositions, and its successful application on a substantial cancer dataset holds promise for future biomarker discovery endeavors.

As we move towards future research, our focus will be on applying the proposed strategy to the complete dataset. The analysis conducted so far was based on a random sample of approximately 3% of over 14,000 genes. Analyzing the full dataset may unveil correlations between different omics types that were not represented in the 500 genes selected for this study. Furthermore, the computational time for this small subset was significant, taking approximately 30 hours. To address this challenge and expedite our modeling process, we aim to optimize our code using a "parfor" loop, a parallel for loop construct designed for parallel computing in MATLAB. Leveraging the "parfor" loop offers the potential to significantly reduce execution time, particularly when dealing with large datasets or repetitive computations. This optimization will allow us to achieve faster and more efficient results in our analysis.

Bibliography

[1] Z. Fan, Y. Zhou, and H. W. Ressom, "MOTA: Multi-omic integrative analysis for biomarker discovery," in Jul 2019, Available: https://ieeexplore.ieee.org/document/8857049. DOI: 10.1109/EMBC.2019.8857049.

[2] E. G. Armitage and C. Barbas, "Metabolomics in cancer biomarker discovery: current trends and future perspectives," *J. Pharm. Biomed. Anal.,* vol. 87, pp. 1-11, 2014. Available: http://europepmc.org/abstract/MED/24091079 https://doi.org/10.1016/j.jpba.2013.08.041. DOI: 10.1016/j.jpba.2013.08.041.

[3] E. Acar, R. Br,o and A. K. Smilde, "Data Fusion in Metabolomics Using Coupled Matrix and Tensor Factorizations," *Jproc,* vol. 103, *(9),* pp. 1602-1620, 2015. Available: https://ieeexplore.ieee.org/document/7202834. DOI: 10.1109/JPROC.2015.2438719.

[4] W. Ma *et al*, "Local probabilistic matrix factorization for a personal recommendation," in Dec 2017, Available: https://ieeexplore.ieee.org/document/8288451. DOI: 10.1109/CIS.2017.00029.

[5] Y. Lin *et al*, "Community Discovery via Metagraph Factorization," *ACM Transactions on Knowledge Discovery from Data,* vol. 5, *(3),* pp. 1-44, 2011. Available: http://dl.acm.org/citation.cfm?id=1993081. DOI: 10.1145/1993077.1993081.

[6] H. Shuang *et al*, "Like like alike -Joint Friendship and Interest Propagation in Social Networks Hongyuan Zha".

[7] H. Mohammadi and V. Marojevic, “Artificial neuronal networks for empowering radio transceivers: Opportunities and challenges,” in 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall). IEEE, 2021, pp. 1–5.

[8] B. Ermis, A. T. Cemgil, and E. Acar, "Generalized coupled symmetric tensor factorization for link prediction," in Apr 2013, Available: https://ieeexplore.ieee.org/document/6531411. DOI: 10.1109/SIU.2013.6531411.

[9] Jiho Yoo *et al*, "Nonnegative matrix partial co-factorization for drum source separation," in Mar 2010, Available: https://ieeexplore.ieee.org/document/5495305. DOI: 10.1109/ICASSP.2010.5495305.

[10] B. Ermiş, E. Acar and A. T. Cemgil, "Link prediction in heterogeneous data via generalized coupled tensor factorization," *Data Min Knowl Disc,* vol. 29, *(1),* pp. 203-236, 2013.

Available: https://link.springer.com/article/10.1007/s10618-013-0341-y. DOI: 10.1007/s10618-013-0341-y.

[11] O. Alter, P. O. Brown and D. Botstein, "Generalized Singular Value Decomposition for Comparative Analysis of Genome-Scale Expression Data Sets of Two Different Organisms," *Proceedings of the National Academy of Sciences - PNAS,* vol. 100, *(6),* pp. 3351-3356, 2003. Available: https://www.jstor.org/stable/3139363. DOI: 10.1073/pnas.0530258100.

[12] L. Badea, "Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing,* pp. 267, 2008. Available: https://www.ncbi.nlm.nih.gov/pubmed/18229692.

[13] E. Acar, G. E. Plopper and B. Yener, "Coupled analysis of in vitro and histology tissue samples to quantify structure-function relationship," *PLoS ONE,* vol. 7, *(3),* pp. e32227, 2012. Available: https://www.ncbi.nlm.nih.gov/pubmed/22479315. DOI: 10.1371/journal.pone.0032227.

[14] A. K. Smilde, J. A. Westerhuis, and S. de Jong, "A framework for sequential multiblock component methods," *Journal of Chemometrics,* vol. 17, *(6),* pp. 323-337, 2003. Available: https://api.istex.fr/ark:/67375/WNG-VKWV2690-G/fulltext.pdf. DOI: 10.1002/cem.811.

[15] A. Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," *Tsp,* vol. 50, *(7),* pp. 1545-1553, 2002. Available: https://ieeexplore.ieee.org/document/1011195. DOI: 10.1109/TSP.2002.1011195.

[16] A. Ziehe *et al*, "A Fast Algorithm for Joint Diagonalization with Non-orthogonal Transformations and its Application to Blind Source Separation Klaus-Robert Müller," *Journal of Machine Learning Research,* vol. 5, pp. 777, 2004.

[17] A. P. Singh and G. J. Gordon, "Relational Learning via Collective Matrix Factorization," *Kdd'08.* DOI: 10.21236/ada486804.

[18] B. Long *et al*, "Spectral clustering for multi-type relational data," on Jun 25, 2006, Available: http://dl.acm.org/citation.cfm?id=1143918. DOI: 10.1145/1143844.1143918.

[19] J. E. McDermott *et al*, "Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data," *Expert Opinion on Medical Diagnostics, 7(1):37-51,* vol. 7, *(1),* pp. 37-51, 2013.
Available: https://www.tandfonline.com/doi/abs/10.1517/17530059.2012.718329. DOI: 10.1517/17530059.2012.718329.
[20] C. M. Ghantous *et al*, "Advances in Cardiovascular Biomarker Discovery," *Biomedicines,* vol. 8, *(12),* 2020. . DOI: 10.3390/biomedicines8120552.
[21] D. Ledesma, S. Symes, and S. Richards, "Advancements within Modern Machine Learning Methodology: Impacts and Prospects in Biomarker Discovery," *Curr. Med. Chem.,* vol. 28, 2021. . DOI: 10.2174/0929867328666210208111821.
[22] B. B. Misra *et al*, "Integrated omics: tools, advances, and future approaches," *Journal of Molecular Endocrinology,* vol. 62, *(1),* pp. R21, 2019. . DOI: 10.1530/jme-18-0055.
[23] Z. Ahmed, "Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis," *Hum Genomics,* vol. 14, *(1),* 2020. . DOI: 10.1186/s40246-020-00287-z.
[24] O. Menyhárt and B. Győrffy, "Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis," *Computational and Structural Biotechnology Journal,* vol. 19, pp. 949-960, 2021. Available: https://dx.doi.org/10.1016/j.csbj.2021.01.009. DOI: 10.1016/j.csbj.2021.01.009.
[25] Y. Hasin, M. Seldin and A. Lusis, "Multi-omics approaches to disease," *Genome Biol,* vol. 18, *(1),* 2017. . DOI: 10.1186/s13059-017-1215-1.
[26] Y. V. Sun and Y. Hu, "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases," *Advances in Genetics,* vol. 93, pp. 147-190, 2016. Available: https://www.ncbi.nlm.nih.gov/pubmed/26915271. DOI: 10.1016/bs.adgen.2015.11.004.
[27] S. Dahal *et al*, "Synthesizing Systems Biology Knowledge from Omics Using Genome-Scale Models," *Proteomics,* vol. 20, *(17-18),* 2020. . DOI: 10.1002/pmic.201900282.
[28] J. Yan *et al*, "Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data," *Briefings in Bioinformatics,* vol. 19, *(6),* pp. 1370-

1381, 2018. Available: https://www.ncbi.nlm.nih.gov/pubmed/28679163. DOI: 10.1093/bib/bbx066.

[29] M. Kang *et al*, "Integration of multi-omics data for integrative gene regulatory network inference," *International Journal of Data Mining and Bioinformatics,* vol. 18, *(3),* pp. 223-239, 2017. Available: https://search.proquest.com/docview/1989915459. DOI: 10.1504/IJDMB.2017.10008266.

[30] N. Rappoport *et al*, "MONET: Multi-omic module discovery by omic selection," *PLoS Computational Biology,* vol. 16, *(9),* pp. e1008182, 2020. Available: https://search.proquest.com/docview/2451546964. DOI: 10.1371/journal.pcbi.1008182.

[31] H. Sharifi-Noghabi *et al*, "MOLI: multi-omics late integration with deep neural networks for drug response prediction," *Bioinformatics,* vol. 35, *(14),* pp. i501-i509, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/31510700. DOI: 10.1093/bioinformatics/btz318.

[32] G. Tini *et al*, "Multi-omics integration-a comparison of unsupervised clustering methodologies," *Briefings in Bioinformatics,* vol. 20, *(4),* pp. 1269-1279, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/29272335. DOI: 10.1093/bib/bbx167.

[33] M. Akhmedov *et al*, "OmicsNet: Integration of Multi-Omics Data using Path Analysis in Multilayer Networks," 2017. Available: https://explore.openaire.eu/search/publication?articleId=sharebioRxiv::43bbef39d5491455c65759257b27a0a7. DOI: 10.1101/238766.

[34] S. Canzler *et al*, "Prospects and challenges of multi-omics data integration in toxicology," *Arch Toxicol,* vol. 94, *(2),* pp. 371-388, 2020. Available: https://link.springer.com/article/10.1007/s00204-020-02656-y. DOI: 10.1007/s00204-020-02656-y.

[35] M. Song *et al*, "A Review of Integrative Imputation for Multi-Omics Datasets," *Frontiers in Genetics,* vol. 11, pp. 570255, 2020. Available: https://search.proquest.com/docview/2461001785. DOI: 10.3389/fgene.2020.570255.

[36] B. Mirza *et al*, "Machine Learning and Integrative Analysis of Biomedical Big Data," *Genes,* vol. 10, *(2),* 2019. . DOI: 10.3390/genes10020087.

[37] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM,* pp. 78-87, 2012. Available: http://dl.acm.org/citation.cfm?id=2347755. DOI: 10.1145/2347736.2347755.

[38] M. Picard *et al*, "Integration strategies of multi-omics data for machine learning analysis," *Computational and Structural Biotechnology Journal,* vol. 19, pp. 3735-3746, 2021. Available: https://dx.doi.org/10.1016/j.csbj.2021.06.030. DOI: 10.1016/j.csbj.2021.06.030.

[39] D. Choi, J. Jang and U. Kang, "S3CMTF: Fast, accurate, and scalable method for incomplete coupled matrix-tensor factorization," *PLoS ONE,* vol. 14, *(6),* pp. e0217316, 2019. Available: https://search.proquest.com/docview/2249031337. DOI: 10.1371/journal.pone.0217316.

[40] Kijung Shin, Lee Sael and U. Kang, "Fully Scalable Methods for Distributed Tensor Factorization," *Tkde,* vol. 29, *(1),* pp. 100-113, 2017. Available: https://ieeexplore.ieee.org/document/7569093. DOI: 10.1109/TKDE.2016.2610420.

[41] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems,* vol. 38, *(2),* pp. 149, 1997. . DOI: 10.1016/s0169-7439(97)00032-4.

[42] Y. Zhang, P. Yang and V. Lanfranchi, "Tensor multi-task learning for predicting alzheimer's disease progression using MRI data with spatio-temporal similarity measurement," in Jul 21, 2021, Available: https://ieeexplore.ieee.org/document/9557584. DOI: 10.1109/INDIN45523.2021.9557584.

[43] I. Jung *et al*, "MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis," *Frontiers in Genetics,* vol. 12, *(1),* pp. 682841, 2021. Available: https://search.proquest.com/docview/2576911840. DOI: 10.3389/fgene.2021.682841.

[44] P. M. Kroonenberg, *Three-Mode Principal Component Analysis.* 1983Available: http://www.econis.eu/PPNSET?PPN=014065479.

[45] Y. Taguchi, "Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases," *BMC*

Medical Genomics, vol. 10, *(Suppl 4),* pp. 67, 2017. Available: https://www.ncbi.nlm.nih.gov/pubmed/29322921. DOI: 10.1186/s12920-017-0302-1.

[46] Y. Taguchi, "Tensor decomposition-based unsupervised feature extraction applied to matrix products for multi-view data processing," *PLoS ONE,* vol. 12, *(8),* pp. e0183933, 2017. Available: https://www.ncbi.nlm.nih.gov/pubmed/28841719. DOI: 10.1371/journal.pone.0183933.

[47] M. RINGNER, "What is principal component analysis?" *Nature Biotechnology,* vol. 26, *(3),* pp. 303-304, 2008. Available: http://dx.doi.org/10.1038/nbt0308-303. DOI: 10.1038/nbt0308-303.

[48] B. Schölkopf, A. Smola and K. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation,* vol. 10, *(5),* pp. 1299-1319, 1998. Available: https://direct.mit.edu/neco/article/doi/10.1162/089976698300017467. DOI: 10.1162/089976698300017467.

[49] M. N. Nounou *et al*, "Bayesian principal component analysis," *Journal of Chemometrics,* vol. 16, *(11),* pp. 576-595, 2002. Available: https://api.istex.fr/ark:/67375/WNG-9XMC3V01-W/fulltext.pdf. DOI: 10.1002/cem.759.

[50] Y. Xie *et al*, "Robust principal component analysis by projection pursuit," *Journal of Chemometrics,* vol. 7, *(6),* pp. 527-541, 1993. Available: https://api.istex.fr/ark:/67375/WNG-6CX42F08-2/fulltext.pdf. DOI: 10.1002/cem.1180070606.

[51] E. J. Beh, "Simple Correspondence Analysis: A Bibliographic Review," *International Statistical Review,* vol. 72, *(2),* pp. 257-284, 2004. Available: https://api.istex.fr/ark:/67375/WNG-6GWHW1JD-G/fulltext.pdf. DOI: 10.1111/j.1751-5823.2004.tb00236.x.

[52] N. Sompairac *et al*, "Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets," *Ijms,* vol. 20, *(18),* 2019. . DOI: 10.3390/ijms20184414.

[53] H. Zou, T. Hastie and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics,* vol. 15, *(2),* pp. 265-286, 2006. Available: https://www.tandfonline.com/doi/abs/10.1198/106186006X113430. DOI: 10.1198/106186006X113430.

[54] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Mach Learn,* vol. 83, *(3),* pp. 331-353, 2010. Available: https://link.springer.com/article/10.1007/s10994-010-5222-7. DOI: 10.1007/s10994-010-5222-7.

[55] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with ℓ0-constraints," *Neurocomputing,* vol. 80, *(1),* pp. 38-46, 2012. Available: https://dx.doi.org/10.1016/j.neucom.2011.09.024. DOI: 10.1016/j.neucom.2011.09.024.

[56] L. Liu and V. W. Berger, *Two by Two Contingency Tables.* 2014Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06160. DOI: 10.1002/9781118445112.stat06160.

[57] H. Park *et al*, "Global gene network exploration based on explainable artificial intelligence approach," *PLoS ONE,* vol. 15, *(11),* 2020. . DOI: 10.1371/journal.pone.0241508.

[58] X. Chen *et al*, "S.I. : HEALTHCARE ANALYTICS Global research on artificial intelligence-enhanced human electroencephalogram analysis," 2009. . DOI: 10.1007/s00521-020-05588-x(.

[59] R. de Kock *et al*, "Circulating biomarkers for monitoring therapy response and detection of disease progression in lung cancer patients," *Cancer Treatment and Research Communications,* vol. 28, pp. 100410, 2021. Available: https://dx.doi.org/10.1016/j.ctarc.2021.100410. DOI: 10.1016/j.ctarc.2021.100410.

[60] S. E. Counts *et al*, "Biomarkers for the Early Detection and Progression of Alzheimer's Disease," *Neurotherapeutics,* vol. 14, *(1),* pp. 35-53, 2016. Available: https://link.springer.com/article/10.1007/s13311-016-0481-z. DOI: 10.1007/s13311-016-0481-z.

[61] H. Fang *et al*, "Chapter 11 - omics biomarkers in risk assessment: A bioinformatics perspective," in *Computational Toxicology*, B. A. Fowler, Ed. 2013, Available: https://www.sciencedirect.com/science/article/pii/B9780123964618000130. DOI: https://doi.org/10.1016/B978-0-12-396461-8.00013-0.

[62] H. O. World and International Programme on, Chemical Safety, "Biomarkers in risk assessment: validity and validation," 2001. Available: https://apps.who.int/iris/handle/10665/42363.

[63] Y. Taguchi, "Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data," *BMC Bioinformatics,* vol. 19, *(Suppl 13),* pp. 388, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/30717646. DOI: 10.1186/s12859-018-2395-8.

[64] M. W. Yeung *et al*, "Machine learning in cardiovascular genomics, proteomics, and drug discovery," *Machine Learning in Cardiovascular Medicine,* pp. 325, 2021. . DOI: 10.1016/b978-0-12-820273-9.00014-2.

[65] J. Lee, S. Oh and L. Sael, "GIFT: Guided and Interpretable Factorization for Tensors with an application to large-scale multi-platform cancer analysis," *Bioinformatics,* vol. 34, *(24),* pp. 4151-4158, 2018. Available: https://www.ncbi.nlm.nih.gov/pubmed/29931238. DOI: 10.1093/bioinformatics/bty490.

[66] C. Christin *et al*, "A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics," *Molecular & Cellular Proteomics,* vol. 12, *(1),* pp. 263-276, 2013. Available: https://dx.doi.org/10.1074/mcp.M112.022566. DOI: 10.1074/mcp.M112.022566.

[67] M. F. Buas *et al*, "Recommendation to use exact P-values in biomarker discovery research in place of approximate P-values," *Cancer Epidemiology,* vol. 56, pp. 83-89, 2018. Available: https://dx.doi.org/10.1016/j.canep.2018.07.014. DOI: 10.1016/j.canep.2018.07.014.

[68] B. Cao, X. Kong and P. S. Yu, "A review of heterogeneous data mining for brain disorder identification," *Brain Inf,* vol. 2, *(4),* pp. 253-264, 2015.

Available: https://link.springer.com/article/10.1007/s40708-015-0021-3. DOI: 10.1007/s40708-015-0021-3.

[69] D. S. Warner *et al*, "Statistical Evaluation of a Biomarker," *Anesthesiology,* vol. 112, pp. 1023, 2010.

[70] K. Ng and Y. Taguchi, "Identification of miRNA signatures for kidney renal clear cell carcinoma using the tensor-decomposition method," *Sci Rep,* vol. 10, *(1),* 123456789. . DOI: 10.1038/s41598-020-71997-6.

[71] Andersson, Claus A., and Rasmus Bro. "The N-way toolbox for MATLAB." *Chemometrics and intelligent laboratory systems* 52.1 (2000): 1-4.

[72] KOLDA, T. G., & BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Review, 51*(3), 455-500. doi:10.1137/07070111X

[73] Kiers, H. A. L. (2004). Bootstrap confidence intervals for three-way methods. *Journal of Chemometrics, 18*(1), 22-36. doi:10.1002/cem.841

[74] Zwanenburg, G., Hoefsloot, H. C. J., Westerhuis, J. A., Jansen, J. J., & Smilde, A. K. (2011). ANOVA-principal component analysis and ANOVA-simultaneous component analysis: A comparison. *Journal of Chemometrics, 25*(10), 561-567. doi:10.1002/cem.1400

[75] Gemperline, P. J., Miller, K. H., West, T. L., Weinstein, J. E., Hamilton, J. C., & Bray, J. T. (1992). Principal component analysis, trace elements, and blue crab shell disease. *Analytical Chemistry (Washington), 64*(9), 523A-532A. doi:10.1021/ac00033a719

[76] Mahood, K. (2011). Mapping outside the square: Cultural mapping in the south-east kimberley. *Aboriginal History, 30* doi:10.22459/AH.30.2011.02

[77] Kroonenberg, P. M., Basford, K. E., & Gemperline, P. J. (2004). Grouping three-mode data with mixture methods: The case of the diseased blue crabs. *Journal of Chemometrics, 18*(11), 508-518. doi:10.1002/cem.896

[78] Camacho, J., Díaz, C., & Sánchez-Rovira, P. (2022). Permutation tests for ASCA in multivariate longitudinal intervention studies. *Journal of Chemometrics*, e3398.

[79] Madssen, T. S., Giskeødegård, G. F., Smilde, A. K., & Westerhuis, J. A. (2021). Repeated measures ASCA+ for analysis of longitudinal intervention studies with multivariate outcome data. *PLoS Computational Biology*, *17*(11), e1009585.

[80] Timmerman, M. E., & Kiers, H. A. L. (2003). Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika, 68*(1), 105-121. doi:10.1007/BF02296656

[81] McDonald, J. H. (2014). *Handbook of biolological statistics*. New York•.

[82] Anderson, M., & Braak, C. T. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation, 73*(2), 85-113. doi:10.1080/00949650215733

[83] Vis, D. J., Westerhuis, J. A., Smilde, A. K., & Van Der Greef, J. (2007). *Statistical validation of megavariate effects in ASCA* Springer Science and Business Media LLC. doi:10.1186/1471-2105-8-322

[84] Liland, K. H., Smilde, A., Marini, F., & Næs, T. (2018). Confidence ellipsoids for ASCA models based on multivariate regression theory. *Journal of Chemometrics, 32*(5), e2990. doi:https://doi.org/10.1002/cem.2990

[85] Kroonenberg, P. M., & de Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika, 45*(1), 69-97. doi:10.1007/BF02293599

[86] Leardi, R. (2005). 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. *Journal of Chemometrics, 19*(2), 119-120. doi:https://doi.org/10.1002/cem.908

[87] Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511802843

[88] Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, *18*(2), 95-138.

[89] Anderson, J. D., & Prosser, C. L. (1953). Osmoregulating capacity in populations occurring in different salinities. Paper presented at the *Biological Bulletin, , 105.* (2) pp. 369.

[90] Deng, Y., Tang, X., & Qu, A. (2021). Correlation tensor decomposition and its application in spatial imaging data. *Journal of the American Statistical Association*, 1-17.

[91] Tang, X., Bi, X., & Qu, A. (2020). Individualized multilayer tensor learning with an application in imaging analysis. *Journal of the American Statistical Association*, *115*(530), 836-851.

[92] Diaz, D., Bollig-Fischer, A., & Kotov, A. (2019, November). Tensor decomposition for sub-typing of complex diseases based on clinical and genomic data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 647-651). IEEE.

[93] Lu, G., Halig, L., Wang, D., Chen, Z. G., & Fei, B. (2014, March). Spectral-spatial classification using tensor modeling for cancer detection with hyperspectral imaging. In *Medical Imaging 2014: Image Processing* (Vol. 9034, pp. 267-277). SPIE.

[94] Kuang, Fei, et al. "lncRNAs AC156455. 1 and AC104532. 2 as Biomarkers for Diagnosis and Prognosis in Colorectal Cancer." *Disease Markers* 2022 (2022).

[95] Shang, Chunliang, et al. "Characterization of long non-coding RNA expression profiles in lymph node metastasis of early-stage cervical cancer." *Oncology Reports* 35.6 (2016): 3185-3197.