Application of Pattern Recognition Methods to Identify Dietary Patterns in Longitudinal Studies: A Novel approach in Nutritional Epidemiology

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfilment of the Requirements
for the degree of Master of Science
in the Collaborative Biostatistics Program of the School of Public Health
University of Saskatchewan
Saskatoon

By

Sara Serahati

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Director of School of Public Health

Health Sciences Building E-Wing, 104 Clinic Place

University of Saskatchewan

Saskatoon, Saskatchewan S7N 2Z4 Canada

OR

Dean

College of Graduate and Postdoctoral Studies

University of Saskatchewan

116 Thorvaldson Building, 110 Science Place

Saskatoon, Saskatchewan S7N 5C9 Canada

ABSTRACT

With the increasing prevalence of longitudinal nutritional data applications in medical science, there is a need for complex statistical models for the identification of dietary patterns in the longitudinal set. Advances are constantly being made in our understanding of the interpretability and application of statistical methodologies for longitudinal data. However, little guidance on these matters is available in most nutritional contexts. One of the most important features of longitudinal data is that the observations repeatedly collected over time are correlated to each other. This time-varying association among observations, which cannot be obtained solely by focusing on cross-sectional data analysis methods, is the main analytic challenge in longitudinal studies. This challenge is particularly relevant in the nutritional field, where researchers strive to identify useful and understandable dietary patterns from large-scale nutritional data. In nutritional epidemiology, dietary patterns are derived using pattern recognition (PR) methods. Generally, there are two types of PR methods; supervised and unsupervised. Although many nutritional studies applied cross-sectional PR methods to the identification of dietary patterns, however, the assumption of these methods might not be suitable for the identification of patterns in longitudinal data. Currently, extensions to both supervised and unsupervised cross-sectional PR methods for revealing patterns in longitudinal data exist in the literature. However, none of these methods have been applied to the identification of dietary patterns where nutritional data collected repeatedly over time. Recently, longitudinal principal component analysis (LPCA) and unbiased random effects expectation maximization algorithm (RE-EM) tree methods, as a substitute to principal component analysis (PCA) and regression tree analysis (RT), for revealing pattens in longitudinal studies are developed. This thesis introduces the first application of LPCA and unbiased RE-EM tree, as unsupervised and supervised PR methods, respectively, for the analysis of longitudinal nutritional data. To illustrate these methods, an analysis of dietary patterns in a representative sub-sample of the Saskatchewan Bone Mineral Study (BMAS) is presented. Results showed that the models presented in this thesis seem feasible and useful for the identification of dietary patterns and their trajectories where nutritional data is collected longitudinally. In this sense, this thesis assists the nutritional epidemiologists and researchers in understanding the importance, role, and meaning of the consideration of time-varying associations in diet. It also introduces new dietary pattern analysis methods in longitudinal nutritional studies using LPCA and unbiased RE-EM tree.

ACKNOWLEDGEMENTS

I would like to use this opportunity to express my gratitude to all the people who helped me through my M.Sc. program.

A very deep and special gratitude goes out to my supervisors Dr. Hassan Vatanparast and Dr. Punam Pahwa, for their encouragement, patience and continuous support during my M.Sc. program. Thank you for your trust, and motivations. I truly appreciate your understanding and patience, which have been a significant part of my graduate experience.

I would like to recognize and express my gratitude to my advisory committee Dr. Adam Baxter-Jones who provided me with insightful feedback and to my committee chair Dr. Cindy Feng. I would also like to appreciate Dr. Mobinul Huq for serving as my external examiner.

Most importantly, I would like to express my utmost gratitude to my entire family for their unconditional love, continuous support, and encouragement throughout my life.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

aBMD	Areal Bone Mineral Density
AHEI	Alternative Healthy Eating Index
BIA	Bioelectric Impedance Analysis
BMD	Bone Mineral Density
BMC	Bone Mineral Content
BMAD	Bone Mineral Apparent Density
BMI	Body Mass Index
CA/ CoA	Cortical Area
CoC	Cortical Circumference
CoD	Cortical Density
CSA	Cross-Sectional Area
CV	Coefficient of Variance
DASH	Dietary Approaches to Stop Hypertension
DDS	Dietary Diversity Score
DXA	Dual-energy X-ray Absorptiometry
FFMI	Fat-Free Mass Index
FFQ	Food Frequency Questionnaire
FMI	Fat Mass Index
FN	Femoral Neck
FRAX	Fracture Risk Assessment Tool
HEIC	Healthy Eating Index of Canada
LS	Lumbar Spine
MDS/MD	Mediterranean Diet Score
μMR	Micro Magnetic Resonance
MRI	Magnetic Resonance Imaging
MuA	Muscle Area
NRS	Nutritional Risk Score
OHS	Oslo Health Study
OR	Odds Ratio
PA	Physical Activity

PAQ-A	Physical Activity Questionnaire in Children
PAQ-AD	Physical Activity Questionnaire in Adolescents
PAQ-C	Physical Activity Questionnaire in Adults
PBM	Peak Bone Mass
BMAS	Pediatric Bone Mineral Accrual Study
PCA	Principal Component Analysis
PHV	Peak Height Velocity
pQCT	Peripheral Quantitative Computed Tomography
QCT	Quantitative Computed Tomography
RFS	Recommended Food Score
RRR	Reduced Ranked Regression
TB	Total Body
ToA	Total Area
ToC	Total Circumference
ToD	Total Density
TrA/TA	Trabecular Area
TrC	Trabecular Circumference
TrD	Trabecular Density
vBMD	Volumetric Bone Mineral Density
WC	Waist Circumference
WHO	World Health Organization
WHR	Waist-Hip Ratio
WHtR	Waist-Height Ratio
DP	Dietary Pattern
PCA	Principal Component Analysis
LPCA	Longitudinal Principal Component Analysis
CA	Cluster Analysis
RT	Regression Tree
RF	Random Forest
RE-EM tree	Random Effects Estimation Method
MERF	Mixed Effects Random Forest for Clustered data
MERT	Mixed effects Regression Tree
CART	Classification and Regression tree
PR	Pattern Recognition

CHAPTER 1
INTRODUCTION

1.1 Motivation

Many nutritional longitudinal studies have identified dietary patterns using cross-sectional statistical methods, including principal component analysis (PCA) and reduced rank regression (RRR). Although these methods might perform well in the identification of dietary patterns (DPs) in cross-sectional nutritional data, their assumptions may not be sufficient for longitudinal nutritional data. These methods do not consider the hypothesis of interdependence between observations collected longitudinally, which leads to erroneous conclusions (1). On the other hand, dietary intakes are dynamic and are prone to change over time. As a result, for acquiring the important information on the changes in dietary intakes over time, investigators require considering the within subject change as correlation in their analysis. In this sense, a modeling framework capable of finding DPs considering the dependence among observations affects the accuracy of the identified DPs and their trajectories.

1.2 Rational and Background Information

Longitudinal studies occupy a prominent role in biological, epidemiological, and clinical research.

These studies aim at following a set of same individuals over time, collecting repeated observations of the same variables for them. Longitudinal studies provide more sensitive detection of associations between predictors and outcome by incorporating within-subject variation. The

main goal of longitudinal design is detecting the temporal changes in the observations of the target population at both individual and population-level (2).

Longitudinal and cross-sectional studies share common characteristics of subjects, features, and values. However, longitudinal data extends beyond a single time point measurement and, as a result; they have an additional attribute time. In a longitudinal design, the occurrence of an event at a particular time point may increase or decrease the probability of the occurrence of another event in the future. In this context, repeated measurements on the same individuals are likely to be correlated, which contrasts with the cross-sectional design. Thus, analysis of longitudinal data requires the accommodation of the statistical dependence among the repeated observations within individuals (2).

Nowadays, in all disciplines such as health-related areas of research, there are highly complex, heterogeneous, multi- and high- dimensional data collected repeatedly over time. Currently, it is common to collect a large number of repeated measurements to model individual's growth pattern over time. These measurements possibly are similar to each other. The identification of dynamic patterns of such multi- or high- dimensional data has attracted increasing interest in several fields of science, including biological, clinical, brain and nutritional research (3–5). . Researchers commonly would like to answer the following question on a set of observations (e.g. dietary intakes) they are interested in: what kind of trajectory patterns do these set of observations have? For example, given the nutritional intake of individuals repeatedly over time, researchers aim at investigating both changes in dietary patterns over time and in their relationship with health-related outcomes.

Diet is an intricate combination of foods and nutrients which has been long recognized as one of the important modifiable risk factors in the maintenance and promotion of health over the

whole life span (4,6). However, many nutrients and food groups are highly correlated to each other; as a result, studying the effect of single nutrient or food groups may not capture the synergistic or interactive effects among them (7). To deal with these limitations, a complementary approach that better reflects the association of diet with health outcomes and its intricacy in everyday life, which is known as the dietary pattern has emerged (7). Dietary patterns, which are widely used in nutritional epidemiologic research, can be constructed through investigator defined (or priori method) and data-driven methods (7). Data-driven methods can be further divided into two separate methods: (1) data-driven, outcome independent or posterior approach; and, (2) data-driven, outcome dependent or the hybrid approach (7–9). Both outcome dependent (hybrid) and outcome independent (posteriori) data-driven approaches, respectively, are exactly the same methods as supervised (e.g. regression tree) and unsupervised (e.g. principal component analysis) pattern recognition (PR) methods (7).

Pattern recognition (PR) is referred to as the identification and interpretation of patterns in the data. PR is commonly categorized into supervised and unsupervised learning techniques. Pattern identification concerning the response variables (labels) is called supervised learning. However, predicting a pattern via unsupervised learning does not require labels or response variables. Supervised and unsupervised PR techniques including decision trees (DT), random forest (RF), cluster analysis (CA) and principal component analysis (PCA), etc. have been widely used for the analysis of cross-sectional nutritional data (10). However, dietary patterns identified at a single time point cannot capture the change in an individual's diet over time.

Recently, nutritional epidemiologists are interested in collecting nutritional data longitudinally to better understand the development of diet over life. Dietary intake has indeed been reported to develop from early infancy to childhood and from childhood to adulthood (11).

Dietary intake throughout life may have a cumulative effect, which means that the impacts of current diet might be confounded or influenced by previous intakes (7). Therefore, when it comes to the evaluation of diet-disease associations, particularly during adulthood, it is important to consider the effect of diet over the whole lifetime. As a result, longitudinal modeling of the change in diet as correlation is an effective way of quantifying and describing diet in nutritional epidemiological studies, especially if established during childhood.

In the longitudinal nutritional studies, different approaches have been incorporated for determining dietary patterns; however, the approaches are inconsistent. The most popular approach of identifying dietary patterns in longitudinal nutritional studies is performing separate PCA (or any method other than PCA) at each time point and selecting similar PCs (12–14). However, using PCA for the identification of dietary patterns separately at each time point will result in different number of principal components (PCs) with different interpretations. As an example, it is possible that the first PC at time one has a different meaning from the first PC at time two.

Another strategy is to create dietary patterns at one of the time points in the study (e.g. at baseline) based on, for example, the PCA method. Then, using the derived loadings of that time point, corresponding dietary patterns in other time points will be calculated (15–17). Identification of factor scores at a one-time point and using their coefficients for the calculation of corresponding factors at the other time points enables researchers to find consistent factors (or PCs) over the whole study period, which in turn allows them to track the change in diet. However, the described approach seems to be misleading for the exploration of dietary pattern trajectories. The reason is that the identified coefficients, through PCA as an example, corresponds to the dietary intakes' information at one specific time points. Broadly speaking, the PCs in PCA are computed by decomposing the covariance matrix, which might differ from a one-time point to the others due to

the difference in the linear combination of variables. This phenomenon illustrates the fact that individuals are likely to change their dietary habits as time progress.

An alternative strategy for the identification of dietary patterns in longitudinal sets is the application of the multiple-time-point method where each observation is considered as a row in the data matrix. Therefore, the same person might appear several times in the rows of the matrix. In this way, PCA is given by the linear combination of food groups which are common at different time points (18–20). However, this approach may lead to inaccuracy of results by ignoring the correlated structure of food groups across time; as a result, it might lead to a failure in detecting the pattern of change over time.

Given the above discussion, unlike cross-sectional dietary data, which lacks consideration of time, longitudinal dietary data have an extra time feature. In this scenario, traditional pattern recognition methods (e.g. PCA, RT) may not be appropriate, as observations in longitudinal data are dependent. Thus, dimension reduction or pattern recognition methods for repeated measured data are required. To the best of our knowledge, existing studies have not investigated the identification of dietary patterns considering the within-subject change in dietary intakes over time as correlation.

In this thesis, first, the application of a newly developed unsupervised PR method known as longitudinal principal component analysis (LPCA) to the identification of dietary patterns in a longitudinal design is introduced. Then, the application of unbiased regression trees (RE-EM tree) for longitudinal and clustered data as a supervised method for identifying dietary patterns is explored. In this thesis, unlike LPCA, which extracts dietary patterns based on the linear combination of food groups without considering the relationship of food groups with outcome variable, the information on total body bone mineral density will be used (TB-BMD) as a labeling

or outcome factor to derive dietary patterns applying RE-EM regression tree analysis. The introduction to the application of these methods will help non-statistician to understand the behind the scenes principles.

1.4 Significance of study

Dietary pattern analysis has become a popular method in the field of nutritional epidemiology from over a decade ago. This is an evolving field for methodologists and nutritional epidemiologist exploring and identifying the best approaches that explains the patterns of food intake in populations. However, most methods are developed for analyzing cross-sectional data, which limits the ability of researchers to use a posterior dietary pattern analyses in longitudinal studies. To the best of our knowledge no study presented the application of methodologies suitable for the analysis of longitudinal data to the identification of dietary patterns where nutritional data collected repeatedly over time. This is the first study that introduces the application of two recently developed unsupervised and supervised pattern recognition methods for longitudinal data analysis to the identification of dietary patterns. To sum, this study introduces new dietary pattern analysis methods in longitudinal studies.

1.5 Objectives

The purpose of this thesis is to examine the question of whether it is necessary to use analysis methods that address the correlated nature of data when identifying dietary patterns longitudinally. In addition, to introduce the application of some recently developed pattern recognition methods in longitudinal studies. Data from the BMAS study in which children and

adolescents aged 8-11 years at baseline (1991-1998) followed for 27 years will be used to address the following statistical objectives:

1.5.1 Objective 1

Introducing the application of longitudinal principal component analysis (LPCA) to the problem of identifying dietary patterns and their trajectories as an unsupervised pattern recognition method.

1.5.2 Objective 2

Introducing the application of unbiased regression trees for longitudinal and clustered data (RE-EM tree) as a supervised pattern recognition method to the nutritional data for identifying dietary patterns.

1.6 Overview of the Thesis

This thesis aims at introducing the application of supervised and unsupervised pattern recognition techniques to the nutritional data collected longitudinally for the identification of dietary patterns and their trajectories. This work has been conducted using a subsample of a longitudinal data on dietary intake obtained from the Saskatchewan Bone Mineral Accrual Study (BMAS). In chapter 2, a comprehensive literature review on both clinical and statistical points of views is presented to capture the current knowledge in the area of the underlying study. In chapter 3, a detailed background on clinical notation, background on the principals of longitudinal data, the details of LPCA, and unbiased RE-EM trees are provided. We proceed to chapter 4 with an introduction to the BMAS data as well as a report on the findings of this study. The last chapter, chapter 5, concluded the thesis with a short summary and directions for the future research.

CHAPTER 2
LITERATURE REVIEW

In this chapter a broad literature review has been conducted that would address the gaps in the area of nutritional epidemiology regarding dietary pattern analysis in longitudinal studies. Additionally, a comprehensive literature review is presented in this chapter to provide an in-depth background knowledge of the existing methods and their extensions over time in the literature.

2.1 Overview of Dietary Pattern

Traditional epidemiological nutritional research has mainly focused on the effect of individual nutrients or a few food groups on health outcomes (21,22). Even though these studies have produced valuable evidence, both their methodological and conceptual limitations, have necessitated researchers or nutritional epidemiologists' to instead focus on dietary patterns - which are combinations of both foods and nutrients (21). The following practical limitations can explain this shift towards dietary patterns. First, individuals typically consume nutrients and foods together, which makes it difficult to differentiate the effect of nutrients from those of foods because of the high interaction between them (23). Second, the cumulative effect of nutrients or foods in a diet-disease relationship model might be too large to be detectable (23). Third, the use of several complicated statistical tests to assess the influence of all the nutrients and foods on health outcomes may lead to false positive results (52).

Dietary patterns are now widely used in epidemiologic research and can be identified through investigator defined (or priori method) and data-driven methods. Data-driven methods can be further divided into two separate methods: a) data-driven, outcome independent or posterior approach; and, b) data-driven, outcome dependent or the hybrid approach (7–9). Both data-driven and investigator defined methods are built on some evidence, and they can have some degree of subjectivity. Supervised (e.g. regression tree) and unsupervised (e.g. principal component analysis) pattern recognition (PR) methods are examples of outcome dependent and outcome independent data-driven approaches, respectively (7). In the following sections, these patterns are described in detail.

2.1.1 Priori

Priori dietary patterns, hypothesis-oriented methods, or investigator-driven methods are specified based on the theoretical nutritional knowledge concerning what constitutes a healthy diet (7). Priori dietary patterns as a proxy for dietary diversity are predefined dietary indices or scores made up of nutrients, foods or combination of both. These indices are computed by generating specific scores based on the given dietary recommendation for the criterion food or nutrient and then summing up the overall score, ranked from minimum to maximum. These indices allow for comparability among studies and are often called measures of diet quality, because, they give answers to the question of "how near the dietary pattern in a defined population agrees with a certain set of dietary recommendation." Mediterranean diet, Oslo health study index, Dietary approaches to stop hypertension (DASH), and healthy eating index (HEI) refer to pre-existing dietary recommendations for a general or specific population to support a healthy diet for the prevention of particular diseases (24–26).

2.1.2 Posteriori

In contrast to the priori dietary patterns, posteriori dietary patterns or exploratory methods are not previously specified; they overlook previous knowledge of healthy or non-healthy foods and nutrients. Researchers have to derive these exploratory patterns from collected dietary data through questionnaires such as food frequency questionnaires (FFQ) and 24h recall, using appropriate statistical methods in a specific population (9). Statistical analysis such as cluster analysis (CA), principal component analysis (PCA), and factor analysis (FA) are used mostly for this data-driven, outcome independent dietary pattern approach. The application of each statistical method depends on the aim of the study. Posteriori dietary patterns do not always introduce dietary patterns that can be recognized as unhealthy and healthy patterns (9,26). Section 3.2 discussed the application of PCA, as an unsupervised method, to the identification of dietary patterns in longitudinal data.

2.1.3 Hybrid

Hybrid methods or data-driven, outcome dependent methods are the most recently developed methods for extracting dietary patterns which bridge the gap between priori and posteriori approaches. They are partially exploratory methods, which are mostly used to describe existing eating behaviors related to intermediate or long-term health outcomes in a specific population. Reduced rank regression (RRR), partial least squares regression (PLS) and decision tree analysis are some examples of statistical analysis techniques used with this approach (8). The derived dietary patterns through this method cannot be reproduced in other populations. However, this problem can be solved by representing the most informative foods in a dietary pattern that are an expansion of less complicated dietary pattern score in which unweighted standardized z scores of food groups that have high correlations with each other are summed (27). Section 3.3 provided

a literature review on the application of outcome-dependent pattern recognition methods to the identification of dietary patterns.

2.2 Principal Component Analysis and Dietary Pattern

Over the last few years, many researchers have devoted their attention to the identification of dietary patterns in cross-sectional studies using Principal component analysis (28–32). Principal-components analysis (PCA) is the most popular outcome-independent data-driven method for extracting dietary patterns from dietary intakes information. PCA linearly combines a set of correlated variables into a smaller set of uncorrelated variables known as principal components. The following studies present a thorough review of the current literature that have studied dietary patterns identified via PCA in longitudinal sets; the approaches used in these studies are inconsistent.

In a study by Borland et al. (18) for tracking the stability of dietary patterns over a 2-year period PCA was used to identify dietary patterns at baseline and follow-up. In a Swedish Mammography Cohort, in order to evaluate changes in the stability of dietary patterns, FFQ was administered at baseline as well as, 4-, 5-, 6- , and 7-years after baseline from 1987 to 1990. In this study, dietary patterns were extracted based on the dietary intakes at baseline and the cumulative average of dietary intakes at follow-up years using PCA (33).

The most common method of dietary patterns identification over time to date involves evaluating dietary patterns separately at each time point using PCA. These studies (12–14) considered factors with the same size of loadings at all-time points irrespective of the magnitude of their explained proportion of variance. For example, in a 21-year prospective cohort study by Mikkila et al. (34), dietary patterns were extracted separately for each study year via PCA. In this

study, for each time point, three principal components were selected. However, after a precise scrutinization the third extracted PC was excluded due to its inconsistent contribution across measurement points. Northstone et al. identified two consistent dietary patterns including "processed" and "traditional" at each time point based on the order the dietary components were extracted, together with their explained proportion of variance and the complete data available (12).

Other studies created dietary patterns at one of the time points in their follow-up (for example at baseline) based on PCA (15–17). Then, using the derived loadings of that time point, they calculated the corresponding dietary patterns in other time points. In a study by Movassagh et al. (16) to identify dietary patterns, PCA was used to compute the factor loadings based on the cumulative average of dietary intakes from the first seven years (1991 to 1997) of the study. Then, they applied the derived loadings to data at the follow-up.

Some studies used the method of multiple-time-point application to extract consistent dietary patterns across their study time points (18–20). In a study by Andersen et al. (35) PCA was used for the identification of dietary patterns. In this study, data on food groups that are collected at the age of 9, 18 and 36 months was in a long format. In the long format, each person had 25 observations on food groups (columns) in three rows for ages 9, 18, 36 months in the matrix. In this way, PCA evaluated underlying food patterns that are frequent over different phases. And, the assigned score for each observation explained the extent of a person's adherence to a particular dietary pattern at a specific phase.

Several studies, have extracted dietary patterns with the aim of evaluating diet-health outcomes associations longitudinally (36–38). In these studies, the dietary patterns have been identified in the same way as the ones described in the previous paragraphs.

As can be seen, several studies in the current literature have studied the longitudinal nature of dietary patterns derived using PCA; the approaches used in these studies are inconsistent. This inconsistency is due to the fact that PCA is not a suitable approach for the analysis of longitudinal data where diet changes over time. PCA does not consider the time-varying associations.

2.3 Hybrid Methods and Dietary Pattern

A group of cross-sectional studies has focused on the potential use of machine-learning methods in the identification of dietary patterns. The first attempt to incorporate machine-learning approaches for revealing dietary patterns was made by Hearty et al. (39). The aim of this study was to evaluate the applicability of two supervised data mining techniques including regression tree and artificial neural networks (ANNs) in predicting diet quality based on food intakes. Later, in a study by Lazarou et al. (40) dietary patterns were revealed using C4.5 algorithm and PCA. Then a comparison has been made between them. The results of this study revealed that C4.5 algorithm performs better than PCA, which supports the fact that having prior knowledge for extracting dietary patterns will improve the classification ability.

Additionally, in a cross-sectional study by Panaretos et al. (41), dietary patterns were identified using reduced ranked regression, PCA, k-nearest-neighbor's algorithm and random forest. Using cardiovascular risk factor as labeling factor for RRR and RF, the results showed that k-nearest-neighbor's algorithm and RF more accurately classified individuals compared to PCA and RRR. Biesbroek (42) applied two hybrid methods including RF and RRR and two posteriori methods including k-mean cluster analysis (KCA) and PCA in the identification and exploration of dietary patterns predictive of cardiovascular risk factors. The authors of this study concluded that RRR and RF have advantage over PCA and KCA. No study to date has applied machine-

learning methods to the analysis of longitudinal nutritional data for the identification of dietary patterns.

To sum, no study up to now applied machine-learning methods to the analysis of longitudinal nutritional data for the identification of dietary patterns and their trajectories. However, it is important to take into account that none of these used methodologies for the identification of dietary patterns in cross-sectional design are suitable for the analysis of longitudinal nutritional data.

However, over the last few years, many researchers have devoted their attention to the application of supervised or outcome-dependent pattern recognition methods for the identification of dietary patterns. However, this approach of identifying dietary patterns requires having a prior knowledge for the identification of patterns. This prior knowledge could be an outcome that diet contributed to it. Epidemiologic data suggests that diet is one of the important non-modifiable factors that have a strong effect on bone health. Therefore, modifying them may have a positive impact on the bone during early and late stages of life (43,44).

2.4 Nutritional Factors and Bone

Diet, nutrients and food group intake, throughout life, can have a potential impact on bone health, either positively or negatively. Several mechanisms through which this happens have been previously explored, they include, the rate of bone metabolism, homeostasis of calcium and other active minerals in the bone, and mainly the structure of bone (45,46). The impact of nutrients and foods have long been known to play a key role in bone mass attainment and on preventing bone loss during childhood and adulthood, and consequently a reduction in future fracture risk (47). These nutritional factors for healthy bone metabolism vary from micronutrients or minerals, such

as calcium, sodium, phosphorous and magnesium; vitamins, such as vitamin A, D, E, K, C, and certain vitamins B; to macronutrients, including fatty acids and protein (45). The role of individual nutrients including calcium and vitamin D as essential factors for promoting bone health is well established (47–50). To ensure healthy bone growth, attainment of peak bone mass and reduction in the risk of age-related bone loss, it is sufficient intake of calcium, a principal component of the bone, is essential (51). Approximately 80-90% of bone mineral content contains calcium and phosphorous (52). Sources of calcium are natural calcium-rich foods, calcium-fortified foods or calcium supplements. Also, dairy products (ex. milk, yogurt, cheese), orange, juices, bread, cereals, granola, and fruit juice, are good sources of calcium (52). Whiting et al. in their longitudinal study concluded that the need for calcium is greater during two years of peak bone accretion (53). Vitamin D acts like steroid hormones. The primary function of vitamin D is supplying calcium for bone growth and calcium homeostasis through increasing intestinal absorption of calcium (54). It also exists in different tissues such as intestine, kidney, bone, muscle, brain, skin, pancreas and immune system tissues for cell differentiation, proliferation and growth (55). Sources of vitamin D are limited, and the primary sources are fish and fish liver oils (54), however, Vitamin D supplementation and sun exposure can supply the daily requirement (55). Vitamin D as an adoptive mechanism affects bone mineral accrual during the time of peak bone mineral accrual (56). Studies on childhood and adult nutrition also focused on the impact of specific foods such as dairy products, fruits, and vegetables on bone health (57–59).

2.5 Dietary Patterns and Bone

Recent epidemiological studies are attracted to a more holistic approach to studying the effect of overall dietary patterns rather than individual nutrients or food groups on bone health

(21,22). Diet is a complex combination of foods and nutrients which has been long recognized as important determinants of bone health (60). Dietary patterns approach by studying diet as a whole, is developed to cope with the limitations that arise from individual nutrients and food studies by identifying potential synergistic, additive effects and correlations between different nutrients and foods (23). Moreover, dietary patterns are dependent on the availability of food and cultural habits consequently they differ between different populations (61). Determining nutritional patterns in various communities and age groups associated with bone health might help to find those common combinations of nutrients and food groups that are important for the improvement of bone health throughout life (61).

Studies evaluating the associations between bone health association and dietary patterns in different populations and age groups with the same measurements are limited. A scoping review by Movassagh et al. (4) reported more of the studies that have examined this association had used cross-sectional data. A comparison between adult and elderly population studies in this review showed the most reproducible dietary patterns associated with bone health are healthy, western, and traditional dietary patterns. However, the majority of these studies used data-driven dietary pattern approaches that are not comparable and reproducible across studies, partly due to different intake habits in the different populations, the subjectivity of dietary pattern approaches, and differences in bone measurements. In this scoping review, the comparison among the limited number of studies in children and adolescents that used data-driven dietary patterns led to mixed results. Besides, the authors demonstrated that most of the dietary indexes used in the included studies were developed primarily with the aim of promoting overall health and they are not bone-specific dietary indices (4). Recently, based on the information in literature a bone-specific dietary

index, the BMD Diet Score was developed (62). The number of studies that assessed the adherence to the BMD Diet Score is limited to one study (62).

Studies evaluating the effect of dietary patterns on bone health present mixed results. Several studies have reported a positive association between bone measurements and a diet high in whole grains, fish, fruit and vegetable, and low-fat dairy foods (63–66). In a study on Canadian men and women, the results showed that vegetable and fruit, and whole-grain dietary patterns lowered the risk of trauma fracture (67). Also, results from a population-based study on Japanese farm women showed that a dietary pattern characterized by higher consumption of fruits, vegetables, and fish and lower consumption of meat and processed meat has a positive association with BMD (68). Similarly, data from cross-sectional Framingham Osteoporosis Cohort Study suggest that senior men with a dietary pattern high in fruit, vegetables, cereals and low in red and processed meats, candy and soft drinks had higher BMD compared to those identified by other dietary patterns (64). Findings from Greek and Japanese women study demonstrated that higher consumption of olive oil, fish and low consumption of red meat was related to higher BMD (68,69).

Although many studies demonstrated a positive association between fruit and vegetables dietary patterns and bone factors others found no association (69–71). In a population-based cross-sectional study of Mediterranean pre- and post-menopausal women, a healthy diet, characterized as a diet rich in fruit, vegetables, and olive oil, had no significant effect on BMD or total bone body mineral content (TBBMC) (69). Similarly, two co-twin controlled studies of UK and Iran showed no significant association between fruit and vegetable dietary patterns with the BMD of postmenopausal women (70,71).

Conversely, there is abundant evidence showing that a poor diet, characterized as a diet with high consumption of processed meat, alcohol, meals low in yogurt, sugar, and confectionery; and the western dietary pattern are negatively associated with bone health (65,67,68,70). Data from a Japanese women study showed that the Western diet, defined as a diet rich in fat, oil, meats, processed meats, was associated with low BMD (68). In another study, among Australian women, investigators found that a dietary pattern, consisting of primarily soft drinks, fried potatoes, sausages, cereals, processed meats, and vegetable oils, increased the risk of TBBMC (65). A study conducted on Iranian menopausal women showed that dietary patterns rich in high-fat dairy products, red meats, processed meats, organ meats, non-refined cereals have an adverse effect on FN and LS BMD (70). Conversely, a study by Whittle et al. found that, among young adult women, a high intake of nuts, chocolates, red meat and poultry clustered as "Nuts and Meat" dietary pattern, is associated with higher FN BMD and FN BMC. Whereas, among adult men, a "refined" dietary pattern, described as a diet with high intake of puddings, crisps, chips, confectionery, chocolate, and soft drinks, has a negative effect on FN BMC (72). In a study of postmenopausal Scottish women, authors found a negative association between processed and snack food patterns and FN BMC (65). Similar findings, have also been reported in other populations, such as among Canadian men and women, and Iranian women (67,70).

To sum up, studies that showed a relationship between food groups and bone health supports the fact that diet is a modifiable risk factor for bone health. As a result, bone measurements could be used as prior knowledge for the identification of dietary patterns in a more precise way. It is important to note that, none of these studies identified dietary patterns using supervised and unsupervised PR methods suitable for longitudinal data.

Beside nutritional factors, sex differences have an impact on bone growth. Hence, when analyzing data in terms of nutritional factors and bone measurements the differences in sex should be taken into account. In the next section sex differences and bone growth will be discussed in detail.

2.6 Sex Differences and Bone Growth

Bone growth as a continuous process is highly regulated and affected by genetic, sex differences, and environmental factors. During growth, bone mass, strength, size and mineral content increases substantially, and puberty is a landmark of it (73,74). Sex differences affect bone mass and density during growth and maturation (75). The maturation time varies systematically in boys and girls of the same chronological age due to the wide variation and velocity of individual's growth parameters from one age to another (74). So, sex differences in the pattern of bone mineral accrual and skeletal growth should be based on biological age than chronological age (76,77).

Skeletal and sex maturity, puberty, and age at peak height velocity (PHV) are the three common known indicators of biological maturity (78). Age at peak height velocity (PHV) as a common maturational landmark refers to the time of maximum linear growth in height during the adolescent growth spurt (76,79). It is commonly used as an indicator of biological age for the comparison of bone mineral accrual between boys and girls (76).

Before puberty, boys and girls showed no difference in their bone mass of the axial and appendicular skeleton (56). Sex differences in bone mass first are maintained during pubertal maturation the time when the secondary sex characteristics are emerging (80). At the begging of puberty, girls showed to have more or equal radial skeletal mass (81). During mid-puberty, cross-sectional studies showed a transient thinner and less dense cortex for boys in comparison to girls

at the radius and tibia (82,83). However, after puberty, men tend to attain more bone mass in comparison to women (81,84). In a cohort study, following individuals over 2 to 3 years, the result showed that boys have more porous cortex compared to girls (85). Studies suggested that micro-architectural differences during puberty, higher rates of bone turnover, deficits at the cortex and greater peak height velocity in boys can be the reason for the higher incidence of fracture and some disadvantages during the time of rapid pubertal growth (86,87). Evidence expressed, in adolescent females, bone mass attainment reduced quickly after menarche, and no considerable attainment occurred in some bone sites after two years (56).

PBMCV, time of peak bone mineral content, is the dramatic increase in bone mineral content (BMC), which measures the amount of calcium and other minerals in bone, is a function of maturation (6report). BMC is associated with children's height and continues to increase until the adolescents' growth spurt (report). The University of Saskatchewan Pediatric Bone Mineral Accrual Study (BMAS), a mixed 7-year longitudinal study, demonstrated that PHV occurs in girls earlier than boys by a mean year of 1.5 which happened at the same time as of their timing of menarche (88). This difference is identified by boys' extended bone maturation time which increases their bone size and cortical thickness (89). Using BMAS data, Bailey et al. showed that the PBMCV occurs at after PHV with a time lag of 0.7, at the mean age of 15 years in boys and 13 years in girls (90).

2.7 Saskatchewan Bone Mineral Accrual study

Saskatchewan Pediatric Bone Mineral Accrual Study (BMAS) is a mixed longitudinal study conducted on normally active healthy Canadian children to determine growth patterns, timing, and magnitude of bone mineral accrual. BMAS provided investigators with children's

biological age that allows investigators to control for maturity differences. Biological age aligns children based on their age at peak BMC or height velocity instead of their chronological age. A detailed description of the BMAS protocol is available in chapter 4.

The longitudinal and cross-sectional studies published in scientific journals based on BMAS data have been designed to identify the relationship between physical activity and bone mineral accrual, the difference in bone mineral accrual between boys and girls, muscle-bone growth, biological maturity, and the association between nutrition and fat mass or bone. A summary of the published work on the exploration of dietary patterns and association of diet and bone based on BMAS dataset is presented below.

Until the year 2017, all the studies on BMAS data evaluated nutrients or food groups individually without considering diet as a whole both in cross-sectional and longitudinal design. As a result, none of these studies applied dietary pattern analysis approaches. A brief review on the objectives of these studies is as the following.

Iuliano et al. (91) in 1999 explored total dietary calcium intake with its main dietary sources in 226 participants aged 8 to 19 years using food recalls completed up to four times per year during the first six years of the BMAS study (1991-1996). Whiting et al. (92) assessed the association of different types of low nutrient-dense beverages intake with bone mineral content (BMC) and accrual during the two years spanning the time of peak BMC velocity age. Carter et al. in 2001 (93) analyzed the association of calcium (Ca) intake with bone mineral content (BMC) in 227 children and adolescents who had bone measurements in the fall of 1993. In 2005, Vatanparast et al. (59) investigated the effect of vegetables and fruit intake on TBBMC in 85 boys and 67 girls from childhood to late adolescence. The participants of this study were examined during the first seven years of BMAS from 1991 to 1997. Mundt et al. (94) in 2006 explored the role of sugar-

sweetened drink intake and physical activity in the development of total body fat mass (FM). Vatanparast et al. (95) in 2007 examined the effect of protein intake on bone mineral mass parameters including TBBMC, and total body bone mineral density (TBBMD) of young adults, considering calcium intake adequacy. In 2010, Vatanparast et al. (96) evaluated calcium requirements of adolescents from age 9-18 years among 85 boys and 67 girls who participated during the first seven years of BMAS based on dietary reference intake's (DRI) sex-age groups.

Later in 2017, Movassagh et al. explored the change in dietary patterns (DPs) over 20 years from childhood through adolescence and into young adulthood. In this mixed longitudinal study, 130 participants (53 females) aged 8-15 years at baseline for whom multiple 24-h recalls were gathered annually during years 1991 to 1997, 2002 to 2005, and 2010 to 2011 were included. The principal component analysis (PCA) was used to derive the dietary patterns at baseline. Then, for describing dietary patterns, the factor loadings from baseline were applied to calculate the factor scores in all other time points. Generalized estimating equations (GEEs) were used to evaluate the stability of dietary patterns over time. The results of this study demonstrated an increasing trend in adherence to a healthy dietary pattern characterized as "Vegetarian-style."

Movassagh et al. in 2018 assessed the effect of adolescents DPs on adolescents and young adult bone measurements. They also explored the change in DPs from adolescence to young adulthood. For this longitudinal study, 125 participants (53 females) within the age of peak height velocity (PHV) ± 2 years in the first measurement during the years 1991-1993 were considered as adolescents. Then 115 participants' (51 females) follow-up data collected during 2010-2011 were considered as adults. Dietary intakes from 24-h multiple recalls were summarized into 25 food groups. Then, DPs during adolescence were identified using PCA. Multivariate analysis of covariance and multiple linear regression adjusted for age, sex, age of peak height velocity,

physical activity, weight, height, and total energy intake were conducted to explore the relationship between adolescent DPs and adult's bone measurement. DPs were tracked using generalized estimating equations. This study showed that a healthy DP rich in dark green vegetables, eggs, non-refined grains, 100% fruit juice, legumes/ nuts/seeds, added fats, fruits and low-fat milk characterized as "Vegetarian-style" during adolescence positively affects bone health.

Movassagh et al. were the first investigators to identify dietary patterns based on BMAS dataset in a longitudinal set. In their studies, dietary patterns were extracted using PCA based on the cumulative average of dietary intakes during the first seven years of study (1991 to 1997). Then, to identify dietary patterns at the last follow-up, they applied the derived loadings to the data at the last follow-up.

Although Movassagh et al conducted dietary pattern analysis, however they did not identify dietary patterns considering the time-varying associations among food groups. This is mainly because for non-methodologists there is no introduction available in the literature on how to apply longitudinal PR methods to the identification of dietary patterns where nutritional data longitudinally collected. As a result, they do prefer to use the common approaches in the literature which are developed for the analysis of cross-sectional data. The next two sections provided an overview of the extension of PCA and regression tree (RT) methods to the correlated data.

2.8 Statistical Background

2.8.1 Principal Component Analysis in Longitudinal Data

Principal component analysis (PCA) is a common method for data analysis and exploration. It is a handful of a method for reducing the dimension of data. Pearson and Hotelling were the first ones who independently introduced PCA as a dimension reduction technique (97).

Over the past years, many researchers have devoted their attention to making extensions to PCA. The first attempts to adopt PCA for longitudinal data were made in the area of functional data analysis, where observations are viewed as random curves (98–101). Functional data are inherently infinite-dimensional. As a result, incorporating dimension reduction techniques is necessary to analyze them. Functional principal component analysis (FPCA) has been broadly explored for a single functional variable (98–100). Later, Greven et al. (101) devised longitudinal FPCA (LFPCA) for observations measured over more levels of time. LFPCA is similar to the classical longitudinal mixed-effects model in which random processes are used rather than random effects to incorporate longitudinal correlation. In addition, Jiang and Wang (102) proposed the covariate-adjusted FPCA where additional variable can be considered in the model. These functional PCA methods consider eigenvalues to be fixed, while eigenfunctions vary over time. Moreover, the main problem with FPCA approaches is that the covariance function is computed only for one or two variables measured at pairs of time points. Zipunnikov et al. (103) presented the longitudinal functional principal approach (LFPCA) method to deal with very high dimensional neuroimaging data. This approach creates a mixed effects model that decomposes the dynamic behavior of repeated observations into a baseline and longitudinal subject specific variability. However, it is hard to understand how this massive method works, and how well it reveals pattern in the data. Recently, a new pattern recognition method for longitudinal studies known as longitudinal principal component analysis (LPCA) has been proposed. LPCA overcomes the above-mentioned disadvantages of the prior methods. LPCA is a multivariate method, easy to understand and interpret, and its eigenvalue and eigenfunctions both vary over time (5).

2.8.2 Regression Tree in Longitudinal Data

Tree-based models are typical data mining methods that offer many advantages in comparison to parametric methods. Tree-structured models have become popular because of their potential for presenting straightforward interpretable predictive models, finding possible significant interactions between variables automatically, and offering data visualization. Since the time when Morgan and Sonquist (1963) developed the automatic interaction detection (AID) algorithm for a univariate outcome (104), a number of other tree-based algorithms such as classification and regression tree (CART) (105) and fast algorithm for classification tree (FACT) have been introduced (106). CART and FACT have had a great impact on the decision tree field because of their pruning technique and variable selection approach. CART is a non-parametric, recursive partitioning statistical technique that can be applied for prediction of both categorical and continuous (or numerical) outcomes. In comparison to other statistical analysis procedures, CART is capable of visualizing the data in the form of a tree diagram. This feature enables CART to handle complex data and illustrate the results in the form of tree structures that are straightforward to interpret. FACT and its extension, generalized unbiased interaction detection and estimation (GUIDE) (107) were developed to decrease the computational cost and bias in choosing split variables using statistical tests. These algorithms have had a significant impact on tree-based modeling methods since they not only are computationally efficient but also maintain model accuracy and interpretability.

In addition to the above-mentioned methods, a number of other extensions to the tree-based algorithms including the survival tree (108), quantile regression tree (109), regression impurity tree (110), Bayesian tree (111), classification rule with unbiased interaction selection and estimation (CRUISE) (112), logistic tree with unbiased selection (LOTUS) (113) have also been

developed. However, these methods are applicable to the cross-sectional data. The main objective of tree-based techniques is to find significant subgroups in data, which are defined by covariates and homogeneous outcomes. In longitudinal data, finding homogeneous subgroups might be pertinent to the mean and/or to the covariance structure. Several studies have extended the regression tree methodology to the longitudinal data by making some modifications to the split function.

In 1992, Segal published the first classification tree algorithm for continuous longitudinal data, which was inherited from CART (114). Zhang, in 1998, extended the CART algorithm to multiple binary responses data (115). In 2002, De'Ath developed the multivariate regression tree (MRT) by making changes to the CART algorithm (116). Other researchers such as Abdolell et al. (117), Hsiao and Shih (118), Eo and Cho (119) and Loh and Zheng (120) also provided extensions to the regression tree models. However, these methods are primarily based upon the multivariate repeated-measures approach. These methods also share the restriction of not allowing observation-level (i.e., time-varying) covariates. As a result, these approaches are not capable of dealing with the random or subject specific effects of observation-level covariates.

In 2011, Hajjem et al. (121) used mixed-effects approach and extended the regression tree algorithms such as CART to longitudinal and clustered data with unbalanced structure for a continuous outcome. The main idea of their mixed-effects regression tree (MERT) algorithm is to separate the fixed effects from the random effects. In 2012, Sela and Siminof (122) independently developed a similar estimation method called random effects expectation-maximization (RE-EM) trees. The main idea of both MERT and RE-EM trees is to use a tree structure for the estimation of the fixed effects after removing the random effects part of the model, update the predictions of the random effect part, and iterate until achieving convergence within the framework of the

expectation-maximization (EM) algorithm. Both methods are designed for continuous response data. Both MERT and RE-EM tree, implemented through CART algorithm, are the most common regression tree methods. However, CART introduces two critical problems including overfitting and being bias towards selection of covariates.

Although the problem of overfitting in MERT and RE-EM trees could be handled through a pruning procedure there is no solution for their bias toward variable selection which affects the interpretation of trees. To overcome the mentioned problems introduced by MERT and RE-EM trees Fu and Siminof (123) developed a revised version of RE-EM tree known as unbiased RE-EM tree that overcomes the bias problem using the conditional tree algorithm as a substitute to CART algorithm.

2.9 Conclusion

The above-mentioned research work in the area of nutritional epidemiology have entirely ignored the time varying associations among food groups to capture changes in individuals eating behavior. It is important to take into account that none of the previously employed methodologies is suitable for longitudinal nutritional data analysis. From methodological perspective, it is obvious that the development of LPCA, and RE-EM tree methodologies provided researchers with a powerful set of tools to address challenging problems raised from analyzing a multi- or high-dimensional longitudinal data. These methodologies account for the correlated nature of longitudinal data.

Until now not only no study identified dietary patterns considering the time-varying associations among food groups but also the application of LPCA and unbiased RE-EM tree have not been introduced for the analysis of longitudinal nutritional data. The contribution of this thesis

is therefore to introduce the application of LPCA and unbiased RE-EM tree for the first time to the identification of dietary patterns and their trajectories in a sub-sample of BMAS data. While introducing the application of LPCA will help researchers to identify dietary patterns based on linear relationships among them, the introduction to the unbiased RE-EM tree will help researchers to understand how to identify dietary patterns with respect to a prior knowledge or an outcome variable in longitudinal studies. As discussed above bone measurements could be considered as an outcome for discovering patterns in nutritional data so in this thesis one of the bone measurements i.e. TB-BMD will be considered as a prior knowledge.

CHAPTER 3
STATISTICAL METHODOLOGY

This chapter provides the required study background of the pattern recognition methods for longitudinal data. To address the scientific question of interest, a brief description of the background of recently developed pattern recognition approaches, including longitudinal principal component analysis (LPCA) and unbiased RE-EM trees is presented.

3.1 Longitudinal Studies

Longitudinal studies occupy a prominent role in biological, epidemiological, and clinical research. These studies aim at following a set of individuals over time, collecting repeated observations of the same variables for them. The framework of longitudinal data is described in Appendix A. Longitudinal studies provide more sensitive detection of associations between predictors and outcomes by incorporating within-subject variation. However, the main goal of longitudinal design is detecting the temporal changes in the observations of the target population at both individual and population-level (2).

Longitudinal and cross-sectional studies share common characteristics of subjects, features, and values. However, longitudinal data extend beyond a single time point measurement, and, as a result, they have an additional attribute known as time. In a longitudinal design, the incident of an event at a particular time point may increase or decrease the probability of incident

of another event in the future. In this context, repeated measurements on the same individuals are likely to be correlated, which contrasts with the cross-sectional design. Thus, analysis of longitudinal data requires the accommodation of the statistical dependence among the repeated observations within individuals, which enhances the accuracy of the estimation (2).

Currently, it is common to collect a large number of responses to model individuals' evolution patterns. The identification of dynamic patterns of such multi- or high- dimensional data has attracted increasing interest in several fields of science, including biological, clinical, brain, and nutritional research. Researchers commonly would like to answer the following question on a set of observations (e.g., dietary intakes) they are interested in: what kind of trajectory patterns do these set of observations have? However, many of these observations are highly correlated to each other; as a result, researchers need to capture information on those observations believed in making a significant contribution to total data. To discover understandable and valid patterns in a large dataset, a complementary approach known as pattern recognition (PR) is widely used by researchers (124).

3.2 Pattern Analysis

Pattern recognition (PR) is referred to as the identification and interpretation of patterns in large datasets. PR is commonly categorized into supervised and unsupervised learning techniques. Pattern identification concerning the response variables (labels) is called supervised learning. On the other hand, predicting a pattern via unsupervised learning does not require labels or response variables. Unsupervised pattern recognition methods like principal component analysis (PCA) are developed to answer the question of what the important underlying factors in accounting for the variation in responses reported by individuals are. Although unsupervised PR methods are

commonly used to investigate patterns related to health outcomes, they are not designed to extract patterns that are predictors of outcome. In other words, they do not capture patterns most related to the outcome. To overcome this problem, supervised methods such as decision trees (DT) have been developed to answer questions like what combination of variables better explains the most variation in response variable. Supervised and unsupervised PR techniques, also known as data-driven outcome dependent and data-driven outcome independent, respectively, have been extensively used for the analysis of cross-sectional data (10,124). However, it is important to note that patterns identified at a single time point could not capture the individual's evolution over time (5).

Longitudinal studies are widely conducted in medical studies as a means of studying the evolution of outcome over time, consequently, to examine how this evolution depends on subject-specific differences. Currently, it is common to collect a large number of variables to model individual's evolution patterns. These variables possibly are very similar to each other. Considering an increase interest in collecting multi- or high- dimensional longitudinal data the need for the development of pattern recognition (PR) approaches emerged. Currently, existing PR methods analyze longitudinal data in a cross-sectional manner; as a result, they are not directly applicable to the analysis of longitudinal data, as they do not consider the longitudinal information (3,5,123).

Recently, several PR methods are developed to handle multi- or high- dimensional longitudinal data especially in the field of neuroimaging (3,101,125). However, little is known in terms of application of such methods in different areas of research. While PR methods are not straightforwardly applicable to the longitudinal data, the classical methods in statistics is to analyze longitudinal data based on mixed effects model. Importantly, these models are able to capture

individual differences in longitudinal data with univariate outcomes based on linear mixed models (126). The next section discussed a brief introduction to the background of newly developed Longitudinal Principal Component Analysis (LPCA) and unbiased RE-EM tree as an extensions of PCA and regression tree (RT) , respectively, for the analysis of high dimensional longitudinal data (5,123).

3.3 Longitudinal Principal Component Analysis

Longitudinal principal component analysis (LPCA) is an extension to the principal component analysis (PCA) with the aim of data exploration and analysis of longitudinal data. LPCA solves the within-cluster correlation problem in longitudinal studies by estimating eigenvalues and eigenvectors as time-varying functions. In this chapter, the Principal component analysis (PCA) and its extension, longitudinal Principal Component Analysis (LPCA) formulation is presented, followed by the application of the LPCA method on the BMAS data.

3.3.1 Principal component analysis

Principal component analysis (PCA) is a widely used technique in multivariate data analysis for data exploration, pattern analysis, and dimension reduction (97). PCA is usually conducted prior to the application of any statistical procedure to avoid the curse of dimensionality. PCA reduces the dimensionality by combining the observations linearly. These linear combinations are highlighting the greatest possible variations in the data.

The framework of PCA is provided elsewhere (97). Briefly, in the multivariate situation when data for N individuals is collected for j random variables, suppose X be a j-component random vector of j random variables and the covariance matrix $\sum$ is defined as $E(XX')$. Let e_k be

a vector of weights with length j i.e. $k = 1,2, \dots, j$ then the principal components or the new variables could be defined as:

$$\varepsilon_k = \sum_{k=1}^{j} e_k' X \qquad for\ i = 1 \dots N$$

ε_k is a linear combination of j original variables and e_k' is a group of orthogonal weights which maximizes the variation in the ε_k. The variance of this principal component is defined as $\lambda_k = E(\varepsilon\varepsilon')$ which equals to $\lambda_k = E(e_k' X X' e_k)$ or $\lambda_k = e_k' \sum e_k$. The maximization problem associated with PCA is as follow:

$$max_{e_k} \frac{e_k^T \sum e_k}{e_k^T e_k} \ , \qquad k = 1, \dots \dots, j \qquad (1)$$

$$s.t. \qquad e_k^T e_k = 1 \ and \ e_i^T e_k = 0 \ , for \ i = 1, \dots, k-1$$

The solution to the objective function (1) could be found through the identification of the eigen-decomposition of covariance matrix of the original data. The sample variance-covariance matrix is given by:

$$\Sigma = \sum_{k=1}^{j} \lambda_k e_k e_k' \qquad (2)$$

The equation (2) is known as eigen-equation, which is satisfied by a sequence of pairs of eigenvalue-eigenvectors ($\lambda_k e_k$) where e_k is orthogonal.

In the longitudinal setting, it is common to conduct discretized principal component analysis (DPCA) that is the decomposing of the eigen-equation (variance-covariance matrix) at each time point, separately. However, DPCA may not always results in improved and clear interpretation of the eigenvectors due to lack of consideration of the variability of observations over time. To tackle the problem of variability of observation over time, recently a new methodology known as longitudinal principal component analysis (LPCA) is proposed which considers this within subject

variation as correlation (5). In the next section, a quick review of the inference and estimation procedure of LPCA will be given.

3.3.2 Longitudinal principal component analysis

Longitudinal principal component analysis (LPCA) reduces to the algorithm of the Estimated Eigen-analysis with Random Effects (EERE) model. The EERE algorithm is as following:

1. Start with initializing the random effects $\gamma_i^{(0)}$, the covariance matrix $D^{(0)}$, the overall mean $\mu_t^{(0)}$, and $j \times j$ residual covariance $R^{(0)}$;

2. Iterate the steps below until the convergence, or $\left\| \gamma_i^{(l)} - \gamma_i^{(l-1)} \right\| < \epsilon_\gamma$, where ϵ_γ is a pre-defined tolerance level.

a. Estimate the random effects $\gamma_i^{(l)}$ through the Estimation-Substitution (ES) algorithm bellow:

$\hat{\gamma} = Z\,((\,1_t 1_t^T) \otimes D + R)^{-1}(1_t^T \otimes D\,)^T$

Where $R = diag(R_1, R_2, \dots, R_t)$ is a $jt \times jt$ block diagonal matrix, $Z = (y_1^{obs} - (1_N \otimes \mu_1^T), y_2^{obs} - (1_N \otimes \mu_2^T), \dots, y_t^{obs} - (1_N \otimes \mu_t^T)$ is an $N \times jt$ matrix, y_t^{obs} is a $N \times j$ matrix of observations at each measurement point, and 1_t is a t-dimensional vector of ones for $t = 1, \dots, t$.

b. At the *l*-th step, estimate the parameters D^l, μ^l and R^l where:

$\widehat{D}^{(l)} = \widehat{D}^{(l-1)} - (1_t^T \otimes \widehat{D}^{(l-1)})((1_t 1_t^T) \otimes \widehat{D}^{(l-1)} + \hat{R}^{(l-1)})^{-1}(1_t^T \otimes \widehat{D}^{(l-1)})^T$;

$\hat{\mu}_t^l = \frac{1}{N}\sum_{i=1}^{N}(y_{it}^{obs} - \hat{\gamma}_i^{(l-1)})$; And,

$\hat{R}_t^{(l)} = \frac{1}{N}\sum_{i=1}^{N}(y_{it}^{obs} - \hat{\gamma}_i^{(l-1)} - \hat{\mu}_t^{(l-1)})(y_{it}^{obs} - \hat{\gamma}_i^{(l-1)} - \hat{\mu}_t^{(l-1)})^T$.

3. Initialize the eigenvector values as the sample eigenvectors: $e_k(t)^{(0)} = \tilde{e}_k(t)$

4. Based on the current eigenvectors $e_k(t)^{(m-1)}$ update the following parameters:

a. Update the eigenvalues $\alpha_k(t)^m$ by minimizing the objective function (3):

$$\sum_{t=1}^{t}\left(\sum_{i=1}^{N}\frac{h_{it}^{*T}h_{it}^{*}}{N}+\phi\sum_{i\neq j}\|e_i(t)^T e_k(t)\|_2^2\right) \quad (3)$$

b. Given $\alpha_k^{(m)}$ update $e_k(t)^{(m)}$ through the Newton-Raphson algorithm

5. Repeat step 4 until the convergence criteria a or b meet:

a. $\left\|\alpha_k(t)^{(m)}-\alpha_k(t)^{(m-1)}\right\|>\epsilon_\alpha$, and ϵ_α is a pre-defined tolerance level, or

b. $\left\|(e_k(t)^{(m)})(e_k(t)^{(m)})^T-(e_k(t)^{(m-1)})(e_k(t)^{(m-1)})^T\right\|>\epsilon_e$, and $\epsilon_e$is a pre-defined tolerance level.

3.3.3 Choosing the number of principal components

A consensus has yet to be reached on the identification of the number of principal components (PCs) in the PCA. However, one of the common methods for identifying the number of PCs in cross-sectional studies is creating a scree plot by which the number of components could be chosen at the elbow of the graph (97). In longitudinal studies, graphing the scree plot at each time point will not result in efficient number of PCs; since each time point leads to different number of PCs. It is suggested that, first the overall variation in data be calculated over the entire study period then the scree plot be analyzed (5). There are other ways of identifying the PCs which are explained more elsewhere (5).

3.4 Classification and regression tree

Classification and regression trees (CART) are non-parametric supervised learning methods which recursively partition the variable space into R_j separate regions, $j=1,2,\dots,J$. The partitioning can be illustrated by a decision tree. A decision tree is a tree-like structure where the

paths from root to leaf illustrates the partitioning rules. The partitions are constructed by binary split of the variables' values; $X \leq c$ if X is an ordered variable with m different values, and $X \in S$ if X is categorical. Then, the mean or mode of the outcome of the training observations in each region is used to make estimation or prediction for observations belonging to the same region. Tree models where the outcome variable is discrete are called classification trees. Decision trees where the outcome variable is continuous (typically real numbers) are called regression trees (105). Below diagram (127) illustrate the basic CART:

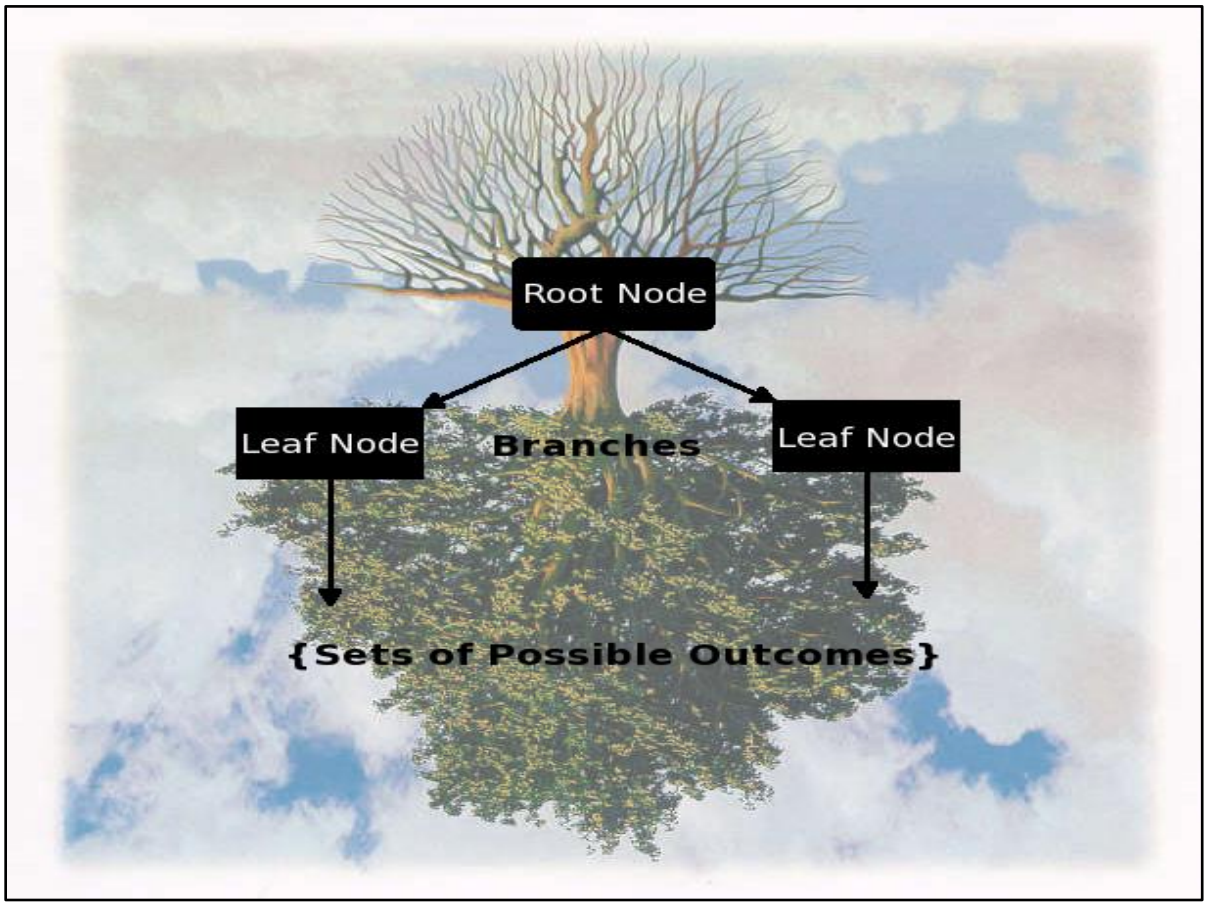

Figure 3-1 A classical classification and regression tree framework, adopted from Leo Pekelis (http://statweb.stanford.edu/~lpekelis/13_datafest_cart/13_datafest_cart_talk.html#(6)) *(127)*

In this section, a brief background on Regression trees (RTs) is provided. This section continued by presenting the details on unbiased RE-EM regression tree as modifications to the RT for analyzing longitudinal data.

3.4.1 Unbiased RE-EM regression tree

The detail of unbiased RE-EM tree is described elsewhere (123). Briefly, longitudinal data includes a panel of individuals $i = 1, \dots, N \ at \ times \ t = 1, \dots, T_i$ where (i, t) is a single observation for an individual i at time t. For each subject there is a vector of independent variables, $x_{it} = (x_{it1}, \dots, x_{itp})'$ where $p = 1, \dots, j$. Let y_{it} be a response variable for subject i at time t, the formula below is a basic formula of longitudinal data:

$$y_{it} = Z_{it}\alpha_i + \beta_1 x_{it1} + \cdots + \beta_p x_{itp} + \varepsilon_{it}$$

Where β_k is an unknown vector of population parameters, x_{itp} is the p[th] covariate for individual i at time t, ε_{it} is an unknown vector of errors for individual i at time t with the assumption bellow:

$$\begin{pmatrix} \varepsilon_{it} \\ \vdots \\ \varepsilon_{it} \end{pmatrix} \sim N\ (0, R_i)$$

And, Z_{it} is a known design matrix which accounts for the differences between individuals across time points, and α_i is the subject-specific intercept for individual i, the assumptions for α_i is as following:

- $\alpha_i \sim N(0, D)$
- $cov\ (\alpha_i, \varepsilon_{it}) = 0$

- $E(y_{it}|\alpha_i) = \alpha_i + x_{it}\beta$ where Z_i is a matrix of one that is just the intercept varies across individuals. Then the model is called mixed model or random-intercept model.

This linear mixed effect model presented above is based on the theory of best linear unbiased prediction (BLUP). And, it assumes a parametric form for the f where $f = X\beta$ that it might be a very restrictive assumption. To overcome this assumption, regression tree tries to estimate f based on the following algorithm:

1. Start with an initial value for the estimated random effects, $\hat{b}$, and set $\hat{b}_i$ to zero.
2. Given the target variable $y_{it} - Z_{it}\hat{b}_i$ and variables $x_{it} = (x_{it1}, \dots, x_{itk}), for\ i = 1, \dots, I\ and\ t = 1, \dots, T_i$, approximate f and estimate a regression tree. Build a set of indicator variables $I(x_{it} \in g_p)$ using this regression tree, where g_p ranges over the whole terminal nodes in the tree. This can be obtained through any tree algorithm. Here, the conditional inference tree (ctree) (horthorn et al. 2006) is used.
3. Estimate the linear mixed effects model, $y_{it} = Z_{it}b_i + \sum_p I(x_{it} \in g_p)\mu_p + \varepsilon_{it}$ using maximum likelihood (ML) or restricted maximum likelihood (REML). Then, extract $\hat{b}$ from the fitted linear mixed model.
4. Repeat step 2 and 3 until the estimated $\hat{b}_i$ converge given the change in the likelihood or restricted likelihood chosen to e less than some tolerance value.
5. Update the predicted response at each terminal node of the tree using the estimated population level predicted response $\hat{\mu}_p$ gained from the estimated linear mixed effects model in 3.

CHAPTER 4
APPLICATION to BMAS DATA

This chapter covers the application of the methodologies, discussed in chapter 3, to the nutritional data collected repeatedly over time with the aim of identifying dietary patterns and their trajectories. To illustrate the application of these methods, an analysis of dietary patterns and their trajectories in a subsample of Saskatchewan Pediatric Bone Mineral Accrual Study (BMAS) data is presented. Besides, the problems with the common approaches already exist in the literature for the identification of dietary patterns in longitudinal studies is addressed. This chapter starts with an introduction to the Saskatchewan Pediatric Bone Mineral Accrual Study (BMAS) data; followed by application of the models presented in chapter 3 to the BMAS data.

4.1 Study Population

In this thesis, the real data used to illustrate the application of longitudinal principal component analysis (LPCA) and unbiased RE-EM tree methods is a subsample of Saskatchewan Bone Mineral Accrual Study (BMAS) dataset. BMAS includes 2109 observations aged ≥8 years measured over 20 years. In the original BMAS study from 1991 to 1993 with annual measurements, 251 students were included, 23 of which were excluded from the study because of incomplete data at baseline. Then, 238 participants were followed in the years 1994 (N=197), 1995 (N=101), 1996 (N=165),

1997 (N=120), 2002 (N=148), 2003 (N=137), 2004 (N=134), 2005 (N=139), 2010 (N=108), 2011 (N=76) and 2018 (N=54).

For the current study, 104 individuals out of 228 BMAS data within the age of peak height velocity (PHV) ± 2 years who had dietary information for 25 food group intakes at least in five measurements were considered. After this inclusion, the number of participants at each measurement changed to 1991 (N=104), 1992 (N=83), 1993 (N=97), 1994 (N=94), 1995 (N=91), 1996 (N=81), 1997 (N=63), 2002 (N=66), 2003 (N=73), 2004 (N=69), 2005 (N=35), 2010 (N=55), 2011 (N=46) and 2018 (N=24); information from the year 2018 is excluded for the current study. The outcome variable, which used for the clustering of 25 food groups, is the Total body bone mineral density (TB-BMD). Both food groups and TB-BMD are continuous variables. A full description of the original BMAS dataset is provided in the next section. In the next section the original BMAS data is described I more details.

4.1.1 BMAS Data

The participants of Pediatric Bone Mineral Accrual Study (BMAS), which is a prospective, population-based mixed longitudinal study, were recruited from two elementary schools in the city of Saskatoon, Saskatchewan. Approximately all the participants were Caucasian, living in a middle-class area of Saskatoon. In 1991, 228 students (113 boys and 115 girls) aged 8-14 years old were entered the study. In years of 1992 and 1993 a total of 23 new cases were included in the study. Therefore, in the original study from 1991 to 1997 a total of 251 students were included, from which 230 students (109 boys) had measurements on two or more follow-ups. In 2001, 197 participants who had not withdrawn from the study were contacted to participate in the follow-up measurements in the BMAS study. In the first follow up study, from 197 participants, 173

participated, however, 169 (94 males) participants aged 18.2 to 27.5 years were measured on at least one occasion. In the second follow up study conducted between 2009 and 2012, 121 participants (48 males), aged 24 to 34 years were examined. Moreover, in the third follow-up study 54 participants (26 males) aged 31 to 40 years were measured. In the original study, children who had a history of chronic disease or chronic medication use, medical conditions, allergies or medications usage is known to affect bone metabolism, or calcium balance was excluded. Participants themselves or their parents signed a written consent form. Table 4-1 demonstrates the frequency of subjects stratified by their age and sex, who had bone scans at each test.

Table 4-1 Age and gender distribution of BMAS from 1991 to 2018

	years																	
Age	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**2002**	**2003**	**2004**	**2005**	**2010**	**2011**	**2017**	**2018**	**Male**	**Total**
Sequence	**(2)**	**(4)**	**(6)**	**(8)**	**(10)**	**(12)**	**(14)**	**(16)**	**(24)**	**(26)**	**(28)**	**(30)**	**(40)**	**(42)**	**(52)**	**(54)**	**(Female)**	
8	3 (7)	5 (10)	0 (2)														8 (19)	27
9	12 (16)	3 (10)	5 (9)	0 (2)													20 (37)	57
10	17 (14)	9 (17)	3 (11)	5 (10)	0 (2)												34 (54)	88
11	17 (13)	19 (15)	11 (18)	3 (9)	5 (10)	0 (2)											55 (67)	122
12	19 (18)	17 (12)	19 (14)	13 (14)	3 (10)	5 (10)	0 (2)										76 (80)	156
13	20 (23)	21 (17)	16 (12)	15 (15)	12 (14)	3 (10)	5 (7)										92 (98)	190
14	19 (15)	14 (19)	19 (15)	12 (11)	13 (15)	10 (12)	3 (9)	0 (2)									90 (98)	188
15	5 (15)	19 (15)	14 (20)	14 (15)	12 (11)	10 (16)	7 (8)	0 (5)									81 (105)	186
16		3 (8)	18 (14)	13 (13)	13 (14)	10 (7)	9 (11)										66 (67)	133
17			3 (7)	13 (11)	14 (12)	12 (13)	9 (5)		0 (2)								51 (50)	101
18				3 (6)	12 (8)	13 (12)	8 (10)		3 (3)	0 (2)							39 (41)	80
19					3 (5)	7 (6)	7 (8)		4 (13)	3 (2)	0 (2)						24 (36)	60
20						3 (4)	5 (3)		6 (13)	2 (12)	5 (4)						21 (36)	57
21							2 (2)		6 (9)	5 (7)	2 (9)	0 (2)					15 (29)	44
22									12 (9)	6 (11)	6 (10)	4 (7)					28 (37)	65
23									8 (10)	10 (8)	7 (9)	3 (7)					28 (34)	62
24									14 (13)	9 (12)	10 (7)	7 (11)	0 (2)				40 (45)	85
25									8 (6)	11 (13)	8 (14)	5 (12)	1 (1)				33 (46)	79
26									3 (5)	9 (6)	10 (11)	7 (8)	1 (8)	1 (1)			31 (39)	70
27									1 (0)	4 (5)	9 (5)	12 (11)	2 (4)	1 (3)			29 (28)	57
28											3 (3)	11 (14)	8 (8)	3 (8)			25 (33)	58
29												9 (5)	7 (10)	4 (3)			20 (18)	38
30												2 (2)	4 (5)	6 (4)			12 (11)	23
31													9 (11)	2 (4)	0 (1)		11 (16)	27
32													4 (8)	12 (7)	3(2)	0 (1)	19 (16)	37
33													5 (7)	5 (7)	0 (2)	0 (0)	10 (16)	26
34													1 (0)	5 (3)	2 (0)	2 (2)	10 (5)	15
35														1 (1)	3 (2)	2 (1)	6 (4)	10
36															2 (1)	1 (1)	3 (2)	5
37															1 (4)	2 (1)	3 (5)	8
38															0 (4)	3 (1)	3 (5)	8
39															1 (2)	3 (2)	4 (4)	8
40															0 (1)	1 (0)	1 (1)	2
Male	112	110	108	91	87	73	55	0	65	59	60	60	42	40	12	14	902 (1148)	
(Female)	(121)	(123)	(122)	(106)	(101)	(92)	(65)	(7)	(83)	(78)	(74)	(79)	(64)	(36)	(19)	(9)		
Total	223	233	230	197	188	165	120	7	148	137	134	139	108	76	31	23		2109

BMAS, Bone Mineral Accrual Study

4.1.2 Measurements of BMAS Study

In BMAS study many variables including bone measurements, nutritional factors, physical activity, body composition and anthropometrics, biomarkers and socio-demographic factors is collected. However, among them for the objectives of this thesis total bone mineral content (TBBMD), nutritional and socio-demographic factors are considered. Table 4-2 illustrate the variables of interest of the current study measured in BMAS.

Table 4-2 Description of BMAS measurements

	Study	Follow-up	Tool/measurement method	Variable
Bone measurements	**BMAS**	1991-1997: 113 boys and 115 girls	DXA	DXA measures: TB BMD
		2002-2005; 94 boys and 75 girls		
		2009-2011; 48 boys and 73 girls	DXA	
Nutrition	**BMAS**	1991-1997: 113 boys and 115 girls	24-hour recall	All data on nutrients (macro and micro) and food groups and food items
		2002-2005; 94 boys and 75 girls		
		2009-2011; 48 boys and 73 girls		
Socio-demographics	**BMAS**	1991-1997: 113 boys and 115 girls	Questionnaire	Sex, age
		2002-2005; 94 boys and 75 girls		
		2009-2011; 48 boys and 73 girls		

4.1.3 Bone Measurements

Throughout the first seven years of the study, DXA using the Hologic 2000 QDR (Hologic, Inc., Waltham, MA, U.S.A.) was used to estimate bone measurements such as total body bone mineral content (TB BMC), poster-anterior lumbar spine (LS (L1-L4) BMC), and the femoral neck bone mineral content (FN BMC), annually. For minimizing the measurement related variability,

the same operator conducted DXA scans, and the same person analyzed all the scans during the original (7-years) and follow-up studies. For all bone scans, array mode with using enhanced global software version 7.10 for analysis was used. Software version 5.67 A was used to analyze the total body scan. In vivo, the coefficient of variation of TB, LS and FN scans (0.60, 0.61 and 0.91, respectively) were comparable to the values from other studies that employed the QDR 2000 in the array mode. During the first follow up study years 2002-2005 the same measurements with DXA were measured.

4.1.4 Dietary Intake

In the original study, whenever bone scans were done at hospitals, at both participation schools and hospital setting, dietary intake information was collected using the administering 24-hour dietary recall at least three times per year during the first three years of the study and at least two times per year after that. For data collection on 24h dietary recall all days of the week, except Friday and Saturday, were included. A trained interviewer recorded the verbally provided information from children of grades 2 and 3, but for other participants the 24 h recall was self-administered. For this purpose, a training session on food portion sizes was administered at the beginning of the study for children before filling the 24 h recall questionnaire. Additionally, display boards of life-size pictures of portion sizes and foods were showed at each administration of 24-hour recall to help subjects make accurate estimates of their dietary intake. The coded dietary information derived from 24-hour recalls for nutritional content was analyzed using NUTS, nutritional assessment software (version 3.7 Quilchena Consulting Ltd., Victoria, BC), which used the 1988 Canadian Nutrient File information. This software stratified each food according to servings from the 1982 Canada's Food Guide, which, except for the names and recommended

servings of the food groups and graphics to illustrate the Guide, was similar to the current version of Food Guide. For foods labeled as "other foods," two separate groups were used: fats and oils, and sweets and desserts. Nutrient supplement use was included in nutrient intake data when supplement use was considered consistent, that is at least two-thirds of the time. The same individual checked and coded all the forms and analyzed dietary intake data. The intake of food and nutrients from consecutive 24-hour recalls were averaged for each year of study to obtain the usual intake of subjects.

For the first follow-up study from 2002 to 2005, two 24-hour recalls at the time of bone scans and one through phone interview were obtained and they averaged to represent a usual intake of BMAS young adults. Food intake was analyzed by Food Processor (version 8.0, ESHA Research Inc, Salem OR) that contained food from the 1997 Canadian Nutrient File. Dietary intake of BMAS adults was assessed using two 24-hour recalls conducted at the time of pQCT scanning during the second (2010-2011) and last (2017-2018) follow-up study and data were analyzed using Food Processor (Version 10.0, ESHA Research) that included foods from the 2007 Canadian Nutrient File. To include in dietary pattern analysis, in the first step quantities of all consumed foods and beverages were converted into grams per day then all items were categorized into 25 predefined non overlapping food groups manually regarding similar nutrient or culinary consumption of them (Table 4-3).

4.1.5 Anthropometrics

Trained staff following standard protocols collected weight and height examinations. Before weight measurement participants were asked to wear light clothes and remove their shoes and

Table 4-3 Food groups utilized for identifying dietary patterns using principal component analysis at each time points of the study

Food Groups	Food Items
Dark green vegetables	Asparagus, green beans, broccoli, lettuce, green pepper, seaweed, spinach, mixed greens, snow peas
Eggs	Eggs
Non-refined grains	Whole grains and partially whole grains (60%) mostly cereals, mixed granola/ grain bar, cracker, oat flakes, wheat germ, whole wheat bread, puffed wheat, brown and wild rice, popcorn, barley
Fruit juice 100%	Apple juice, limeade, apple, lemon, orange juice canned or bottled, unsweetened cranberry
Legumes, nuts, and seeds	Beans (black, kidney, lima, navy, small white, soy), chickpeas, hummus, tofu, brazil nuts, coconut, almond, hazelnuts, cashew, peanuts, mixed nuts, pecans, peanut butter, sunflower seeds
Added fats	Butter, margarine, vegetable oil, cooking oil, mayonnaise (salad dressing, miracle whip), coconut milk, dream whip, olive oil, pesto, meatless bacon bits
Fruits	All fresh and dried fruits, canned fruits (not sweetened), avocado, olives
Low-fat milk	1%, skim, rice beverage, soy beverage
Fruit drinks	Fruit juice (sweetened), fruit drinks, iced tea
Refined grains	Refined cereals, white bread, white rice, refined pasta, noodles, pop corns, ice cream cone, pie crust, pizza pop, pizza pocket
Cream	Sour cream, cream (10%, whipped or low fat)
Poultry	Chicken and Turkey
Processed meats	Burger patties (beef, ham, chicken, etc.), sausages, bacon, canned meat (beef, ham, pork, chicken, turkey), dry ribs, fried chicken, nugget
High-fat milk	2%, whole, or almond milk
Tomato	Tomato (fresh, raw, cooked, canned), ketchup, clamato juice, tomato juice, pasta sauce, pizza sauce, salsa, spaghetti sauce, tomato sauce
Red meat	Beef, ham, pork, bison (ground, loin, rib, steak, stew, fried, pot roast, balls, loaf, chop)
Cheese	All kind of cheese
Yogurt	Yogurt (plain, vanilla, or fruit)
Desserts and sweets	Sweet baked products, milk desserts, jelly, chocolate, sugar, jam, syrups, honey, and candies
Fish and seafood	Fish, Shrimp, lobster, mussels, pickerel, prawns, scallops
Dressing, sauces, gravy	Gravy, dressing, Caesar, French, ranch, Italian, 1000 island, Alfredo, blue cheese, chip chip, Greek, honey garlic, white sauce, sandwich spread, tartar, teen, sundried tomato
Vegetables, others	Carrots, snap beans, cabbage, cauliflower, celery, cucumber, garlic, mushroom, pepper, squash, bean sprouts, beets, onion, eggplant, radish, zucchini, potato, green peas, corn, sweet potato, and soups
Chips and fries	Potato chips, fries, corn chips, nacho, hash brown
Soft drinks	Soft drinks (sugar-sweetened or diet)
Others	Salt, spices, seasoning, additive, pickles (dill, beet), low fat sauces (mustard, hot, soy, teriyaki), vinegar

jewelry, then it was measured to the nearest 0.01 kg on a SECA electronic scale at original and first follow-up study, and using a calibrated mechanical scale at the second follow-up study.

And, height in standing condition was measured to the nearest 0.1 cm by a wall-mounted stadiometer. For height measurement, the participants were asked breath in and to stand barefoot, keep their hands by their sides and their heels touching the wall. To eliminate human error, a second-time measurement was taken then the average of the first and second measurements was recorded.

4.1.6 Ethical Approval

For the original BMAS and consequent follow-up studies (Bio # 88-102) (128), ethical approvals were granted through the University of Saskatchewan and Royal University Hospital Advisory Committee on Ethics in Human Experimentation.

4.2 Statistical analysis

All the analysis was conducted using SPSS version 26 (SPSS, Chicago, IL, USA) and R version 6.1 for windows (2019 The R foundation for statistical computing). The data management and descriptive analysis were conducted using SPSS software. All continuous variables are illustrated as means and standard deviations, and the differences were evaluated using t-test. Categorical variables are illustrated as frequency and percentages and were compared using Chi-square test. For comparing the differences P-values lower than 0.05 were considered as statistically significant threshold. The trend in the mean of four food groups is illustrated and a test for trend is computed based on generalized estimating equation (GEE) method. And, the trend in individual's consumption is illustrated using spaghetti plot. Principal component analysis was

conducted separately at each measurement and first two identified principal components (PC) at each measurement were selected. Then the correlation of first and second PC among measurements were computed, separately. Also, the correlation among food groups measured at each measurement were computed. Correlograms[1] were used to illustrate the Pearson correlations (129).

For the analysis of LPCA, we had to modify the primary codes which was available through request from the author (5). The primary R codes were appropriate for the time when the longitudinal data is balanced. So, the codes were modified in a way that it could handle unbalanced longitudinal data. Number of principal components to be created were identified using scree plot. Heatmaps were created to illustrate the changes over time among different variables. The number of iterations were considered 350 and the threshold were considered as 0.0001. Dendrogram[2] was used to classify the food groups score over time.

For conducting the analysis of unbiased RE-EM tree, the R codes were available from a link introduced by authors (http://people.stern.nyu.edu/jsimonof/unbiasedREEM/). Additionally, REEMtree package was used (https://cran.r-project.org/web/packages/REEMtree/index.html). The leave-one-out cross-validation was incorporated into the R codes. To identify dietary patterns, the decision tree was graphed, and the most important food groups were selected for each sex separately. However, not all food groups that were selected with unbiased RE-EM tree contributed to the classification of dietary patterns because one of the drawbacks of all regression trees is that they might overfit the data. So, for the definition of dietary patterns, first the data set were classified based on categorizes that unbiased RE-EM tree selected. Then, the mean of food groups in each category were computed and dietary patterns were defined based on the food groups contributed to the highest mean in the categories. The categories with almost same values were merged.

[1] An image of correlation statistics
[2] A diagram that shows the hierarchical clustering

4.3 Descriptive Results

In this study, 104 participants (58 males) with a mean age of 12.05 ± 1.62 and age range of 8-16 years at baseline were followed for 20 years; participants represented approximately 46% of the total sample and had the food group intakes information in at least 5 of 13 study time points. Totally 957 observations were considered for the analysis. Table 4.4 demonstrates the distribution of age and sex among different time points, for those who were included in this study and had their dietary information at least at five time points. Because the number of individuals who had dietary information in the last time point, years 2017 and 2018, was small (N= 27), the information from this time point is excluded for this study. Since this study aims at describing why and how to apply pattern recognition methods in the longitudinal nutritional studies rather than discussing the clinical results, removing the last time point will not have any effect on it. However, in order to show that this sub-sample is representative, a comparison between the sub-sample BMAS data used in this thesis with the excluded BMAS data in terms of their baseline characteristics is made.

Table 4-4 Age and sex distribution among different time points, Saskatchewan Bone Mineral Accrual Study, 1991-2011

	N	Age Mean (SD)	Age Min-Max	Boys N (%)
1991	104	12.01 (1.62)	8-16	58 (55.8)
1992	83	12.93 (1.72)	9-16	44 (53.0)
1993	97	13.56 (1.89)	9-17	54 (55.7)
1994	94	14.58 (1.90)	10-18	51 (54.3)
1995	91	15.59 (1.92)	11-19	47 (51.6)
1996	81	16.48 (1.94)	12-20	41 (50.6)
1997	63	17.25 (1.99)	13-21	33 (52.4)
2002	66	22.87 (1.92)	19-27	41 (62.1)
2003	73	23.88 (1.92)	20-27	40 (54.8)
2004	69	24.94 (1.87)	21-28	37 (53.6)
2005	35	26.32 (1.82)	22-29	21 (60.0)
2010	55	29.85 (2.05)	26-34	29 (52.7)
2011	46	31.36 (1.95)	28-34	26 (56.5)
Total	957	12.05 (1.62)	8-34	522 (54.5)

SD, Standard Deviation; N, number of individuals

Table 4.5 illustrated the baseline characteristics of those who were included in the study to those who were not included. There was no significant difference between these two groups regarding age, sex, height, weight and TBBMD.

Table 4-5 Comparison of differences in the baseline characteristics of those who were included to those who were not, Saskatchewan Bone Mineral Accrual Study, 1991

	Included N=104	**Excluded N=124**	**P-value**
Age (y)	11.6 ± 1.8	11.5 ± 3.8	0.869
Height (m^2)	151.8 ± 13.3	152.4 ± 13.1	0.693
Weight (Kg)	43.0 ± 12.4	45.0 ± 14.6	0.264
TBBMD	0.9 ± 0.1	0.9 ± 0.1	0.770
Male	58 (55.8)	54 (43.5)	0.066

TBBMD, Total body bone mineral density.
[1]Values are expressed as mean ± SD or as N (%)
[2] P<0.05 is considered statistically significant

The baseline characteristics of those who participated in the last measurement (year 2011) compared to those who did not are illustrated in Table 4-6; there were no clinically significant differences between these two groups regarding age, sex, biological age, height, weight and TBBMD.

Table 4-6 Comparison of baseline characteristic of study participants who completed vs. not completed the last measurement (2011), Saskatchewan Bone Mineral Accrual Study, 1991

	Complete N=56	**Loss to follow N=46**	**P-value**
Age (y)	11.8 ± 1.6	12.3 ± 1.6	0.120
Height (m^2)	158.4 ± 11.7	161.3 ± 11.6	0.221
Weight (Kg)	43.2 ± 11.5	46.9 ± 12.2	0.108
TBBMD	0.9 ± 0.1	0.9 ± 0.1	0.157
Male	32 (55.2)	26 (56.5)	0.891

TBBMD, Total body bone mineral density.
[1]Values are expressed as mean ± SD or as N (%)
[2] P<0.05 is considered statistically significant

Results from table 4.5 and table 4.6 showed that this study sample could be a representative of the whole sample. Figure 4-1 shows the average food consumed by individuals for four food groups, Dark green vegetables, Eggs, Red meat, and Desserts and sweets. We noticed that individuals tend to have an irregular trend in Desserts and sweets (P=0.562) which is not significant, while they tend to have a relatively increasing trend in Dark green vegetables (P<0.001), and Eggs (P<0.001). The trend in red meat (0.019) was almost stable at first but after the year 2003 it started to increase.

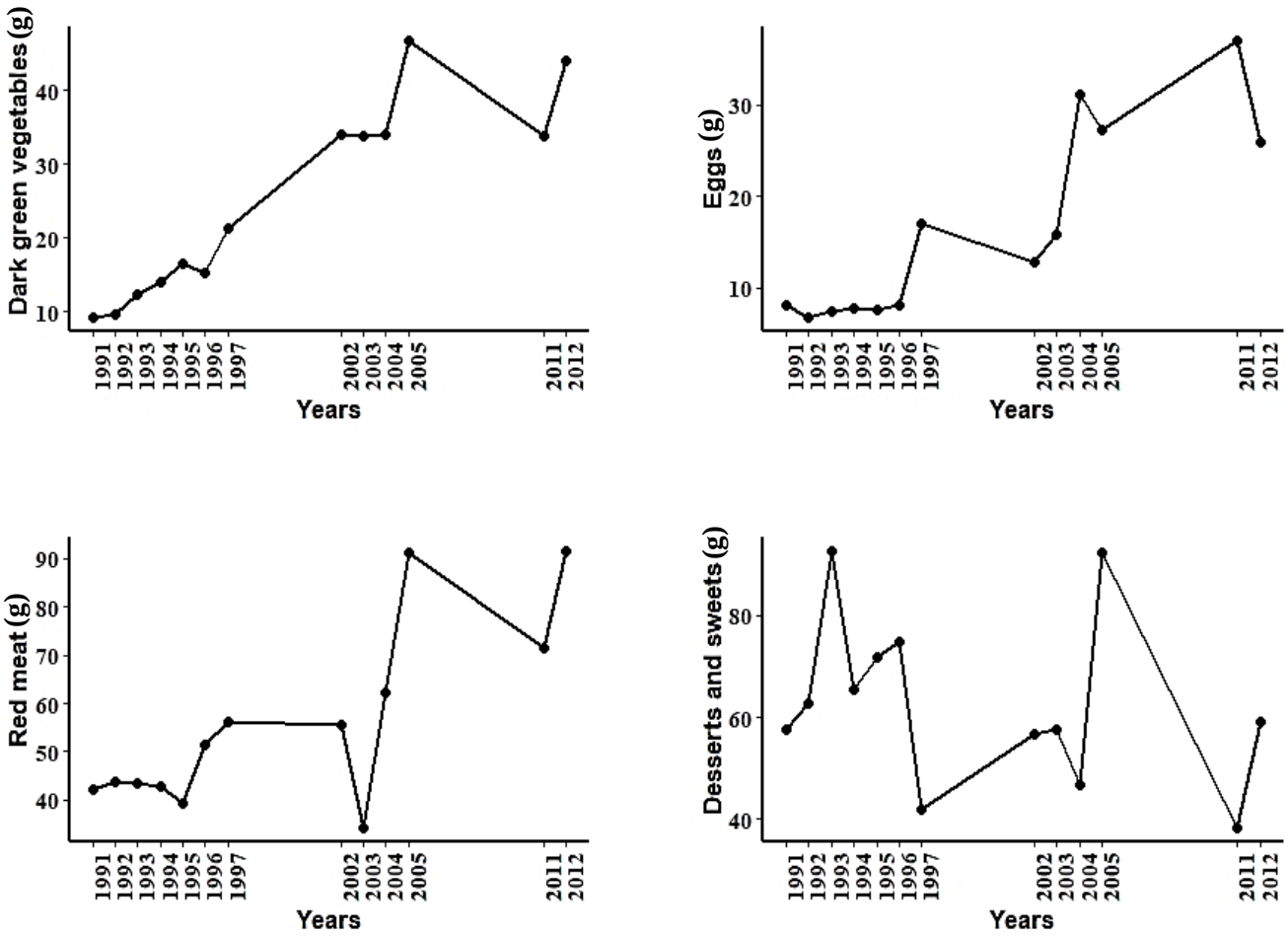

Figure 4-1 Mean levels of Dark green vegetables, Eggs, Red meat, and Desserts and sweets, across different time points of BMAS; The values of food groups are in grams

Figure 4-2 illustrates that selected food groups consumed by individuals present different patterns over time.

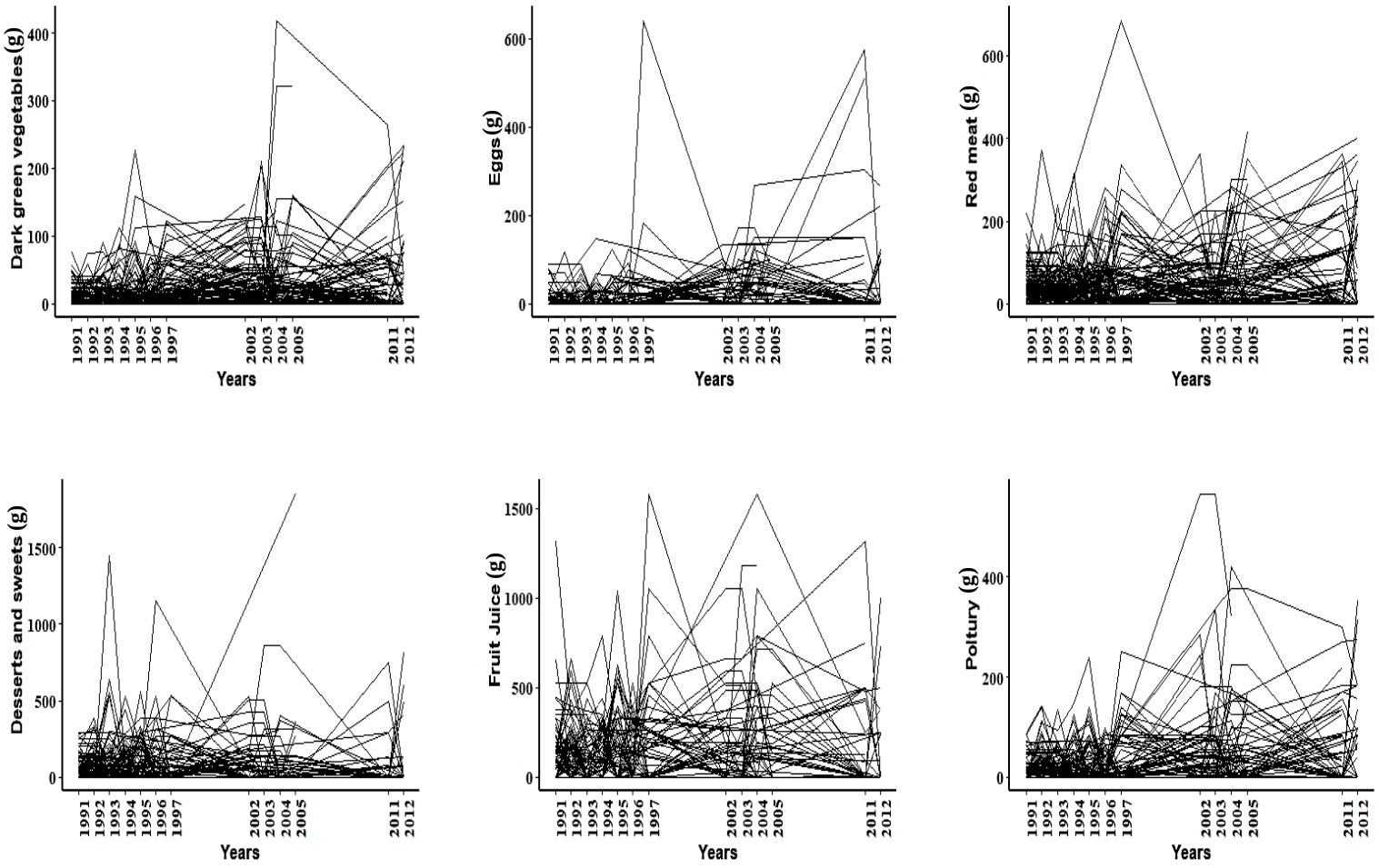

Figure 4-2 The spaghetti plot of Dark green vegetables, Eggs, Red meat, Desserts and sweets, Fruit Juice, and Poultry's change overtime by subject in BMAS; The values of food groups are in grams

This longitudinal nutritional data exhibits interesting features. However, results from the graphs 4-1 and 4-2 support the fact that analyzing this data is challenging due to the variability, number of food groups, and time-varying nature of the data. The variation in the food groups consumption can be due to the several intrinsic factors, such as changing diet due to the illness, immigration and transforming to the healthier diet, etc. In fact, capturing the transition in diet is essential in understanding the disease-diet associations. In the next section the current methods for the identification of dietary patterns and their trajectories will be discussed.

4.4 Problem with Common Methods

It is essential to investigate the problems with the common methods for the analysis of longitudinal nutritional data in the literature. This aids nutritional epidemiologist and researchers to better understand why they need to use longitudinal statistical methods for the analysis of nutritional data that is collected repeatedly over time. The most popular method of identifying dietary patterns in longitudinal studies is to perform separate PCA at each time point and selecting similar PCs. Figure 4-3 shows the correlograms for the correlation matrices of first two PCs between different study time points; PCs, derived from performing separate PCA at each time point.

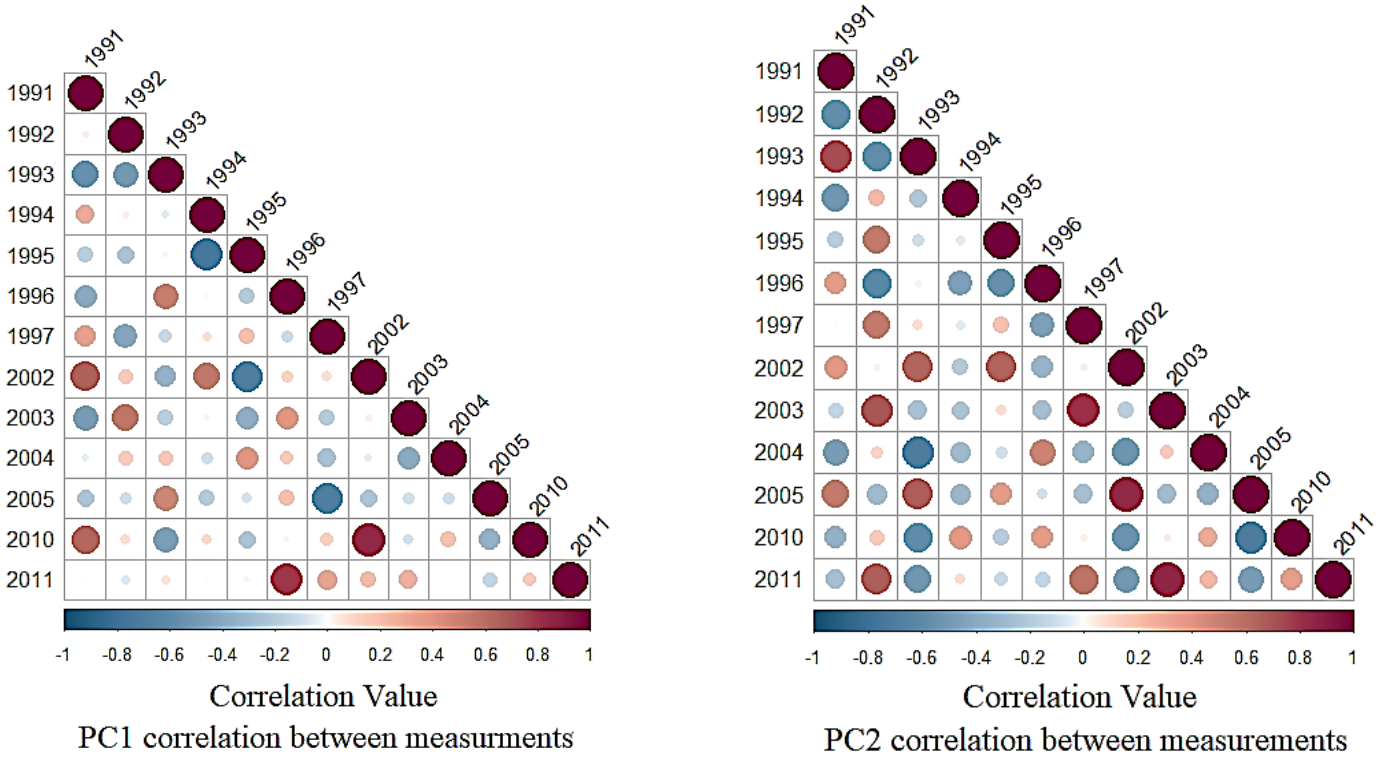

Figure 4-3 Pearson correlation of first two PCs between measurements; PCs, derived from performing separate PCA at each time point; the color and size of the figures represents the correlation value, colors change by the values and size changes by the magnitude

Figure 4-4 shows the correlograms for the correlation matrices among several selected food groups.

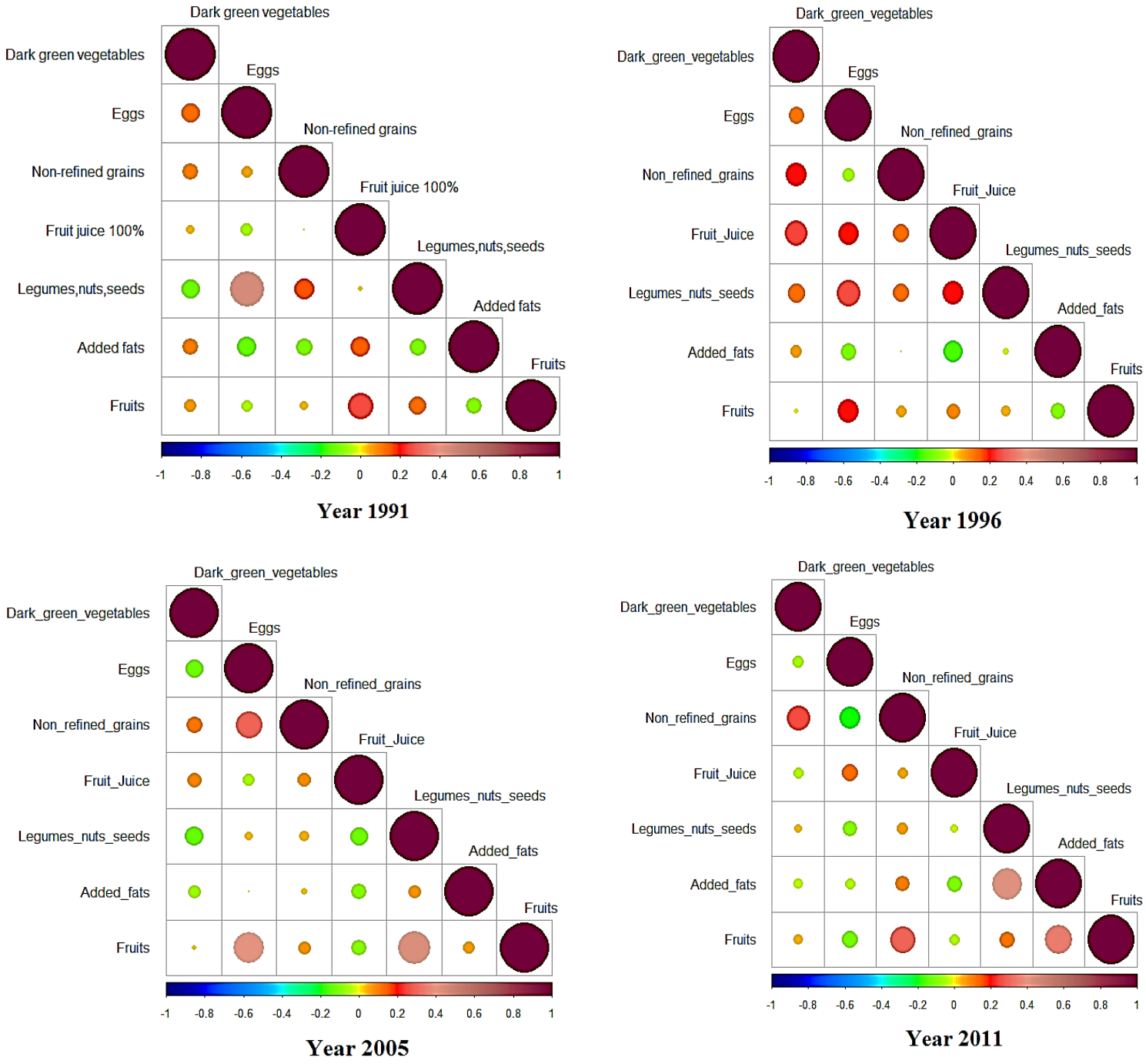

Figure 4-4 Pearson correlation between food groups at the years 1991, 1996 and 2011 ; Capture association among food groups over time; The values of food groups are in grams the color and size of the figures represents the correlation value, colors change by the values and size changes by the magnitude

The most popular method of identifying dietary patterns in longitudinal studies is performing separate PCA at each time point and selecting similar PCs. In this dataset, performing PCA at different time points resulted in different number of PCs (data not shown). Additionally, results from figure 4-3 showed that the magnitude among the pairwise correlations for PCs changes as time progresses. Also, loadings from the most important two PCs, PC1 and PC2, at each time point are not correlated highly to their peer at another time points or they are not correlated to each other among different time point with the same value. For example, the correlation between PC1 at time 1 and PC1 at time 5 is small and it is almost -0.2. Additionally, the correlation between PC1 at time 1 and PC1 at time 5 (r=-0.2) is different from correlation from PC1 at time 1 and PC1 at time 12 (r=0.6). Therefore, performing separate PCA at each time point will result in different PCs with different interpretations, which will prevent researchers from tracking the patterns of diet over time.

Another common method for the identification of dietary patterns in longitudinal studies is to perform PCA at one of the time points of the study and applying the loadings from that time point to the data at another one. Results from figure 4-4 describes the fact that individuals are likely to change their dietary habits over time. As a result, applying loadings derived by PCA at one time point to the data at other time points is an inefficient method. Which will prevent researchers from identifying trajectories over time in a reliable way. To address these problems, in the following sections the application of LPCA, and unbiased RE-EM tree methods for the identification of dietary patterns and their trajectories in longitudinal studies is presented.

4.5 Application of LPCA

In this section the application of longitudinal principal component analysis (LPCA) as an extension to PCA for the analysis of longitudinal nutritional data is illustrated. Similar to PCA, LPCA could be considered as a data-driven outcome-independent approach. The point on the scree plot (Fig 4-5) where the slope of the curve leveling off shows the number of principal components that should be created which is two.

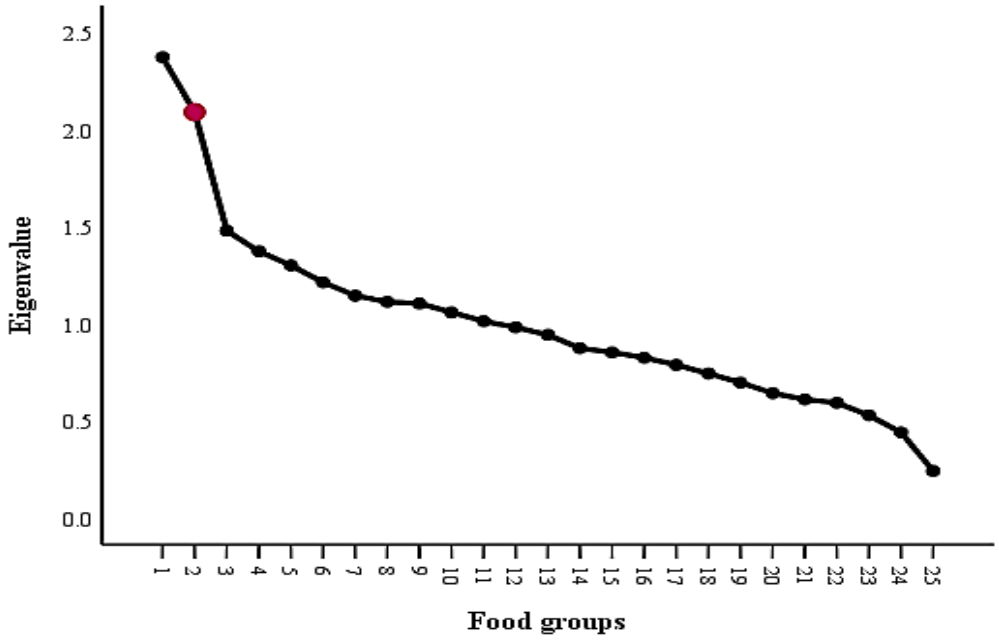

Figure 4-5 Scree plot obtained using principal component analysis of 25 food groups to derive dietary patterns of BMAS participants over 20 years, the red dot presents the number of principal components

Results from implementing the Estimated Eigenanalysis with Random Effects (EERE) or LPCA approach indicated that the two derived PCs are very similar to each other in terms of categorizing the variables (Fig 4-6 & 4-7). However, as it is apparent the scores differ. Therefore, here the results from first EERE is interpreted.

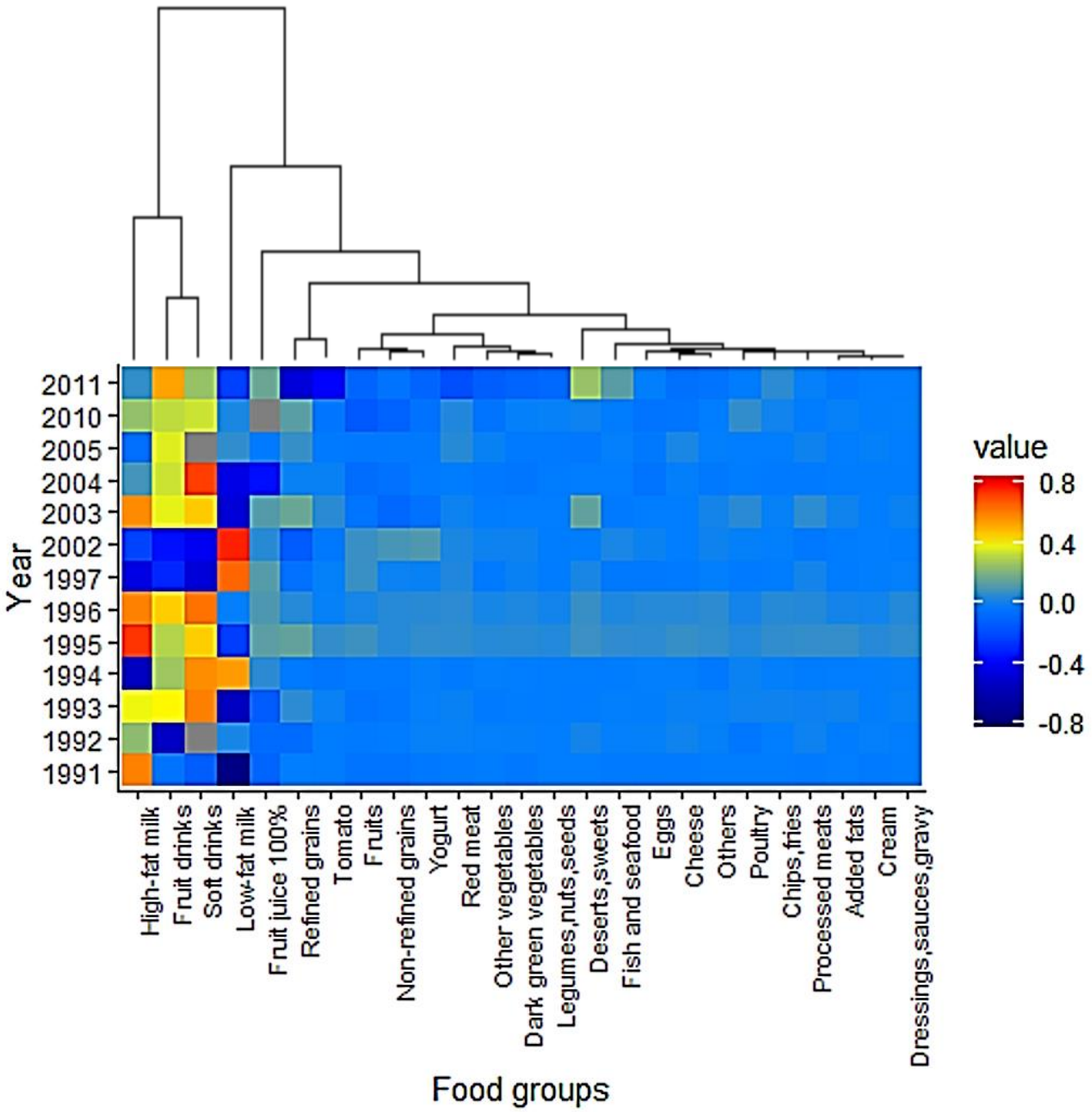

Figure 4-6 First Eigenvector obtained from Estimated Eigenanalysis with random effects EERE analysis; colors indicates the value of the eigenvector's scores

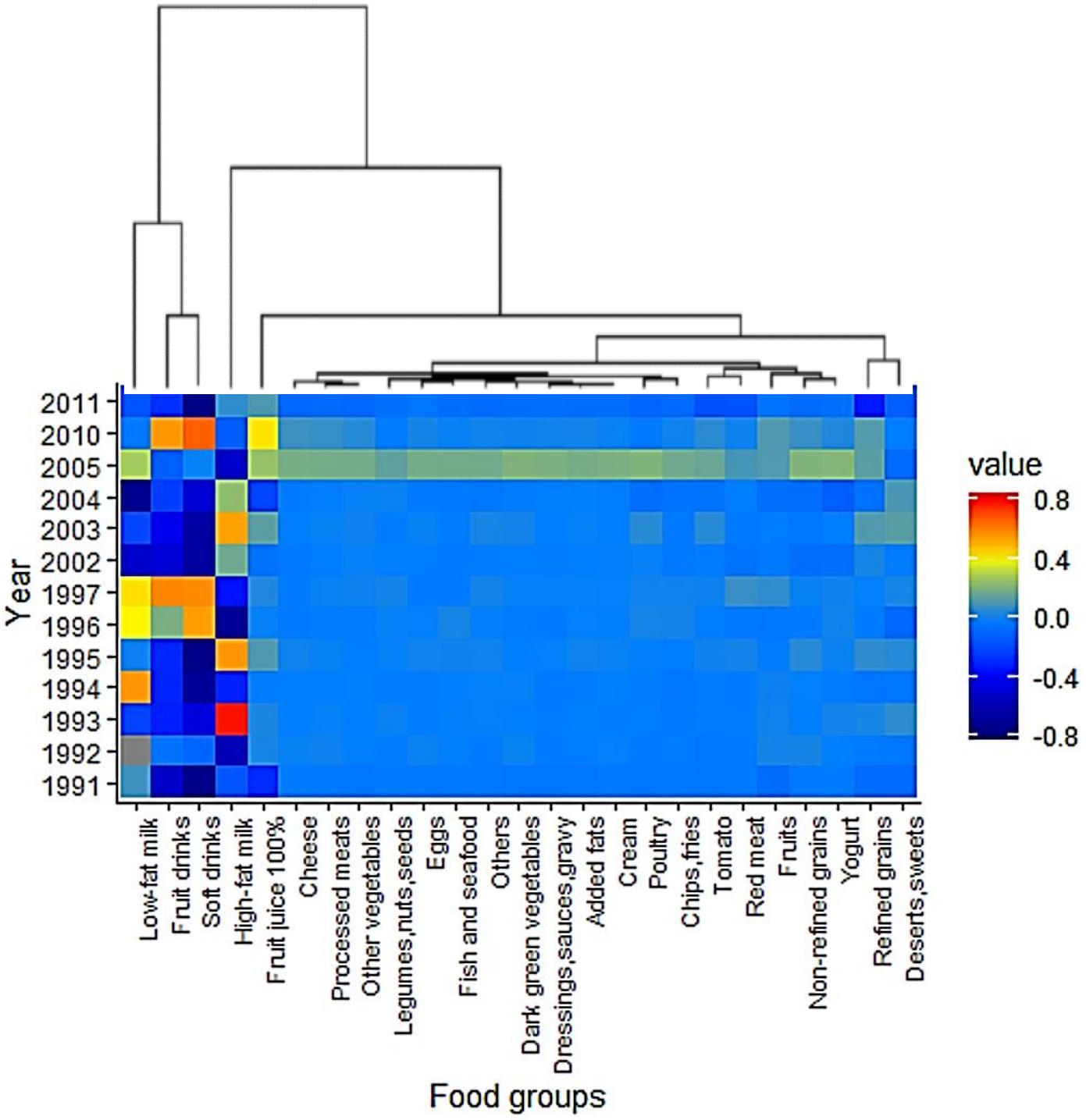

Figure 4-7 Second Eigenvector obtained from Estimated Eigenanalysis with random effects EERE analysis; colors indicates the value of the eigenvector's scores

Based on the first eigenvector (EERE) we can see the most change in high-fat milk, fruit drinks, soft drinks and fruit juice 100%. For example, the change in the scores value as it is apparent in the Figure 4-6 suggest that intake of high-fat milk, fruit drinks and soft drinks shifts from positive values (1991-1996) to the negative values (1997-2007) then again, they shifts toward positive values. Changes in low fat milk is inconsistent. However, after the year 2002 it shifted toward negative values. Fruit juice shifted relatively from negative values to positive values. Other food groups seem to have a similar change that is not noticeable. Based on these colors the change in food groups is noticeable, however, for a precise identification of food groups with the same trend over time dendrograms were used.

Dendrogram was used to cluster the food groups with the same evolution pattern over time in the same categories. Based on the dendrogram the clusters could be considered as high-fat milk, high sugar (Fruit drinks and soft drinks), low-fat milk, Fruit juice 100% and mixed dietary patterns. Additionally, results indicate that individuals are attached mostly to their diet developed during childhood. The loadings of this method also could be used for the prediction purposes which is not the objective of the current study. For extracting PCs obtained from LPCA in dataset, like cross-sectional PCA approach, one can first standardized the food groups then multiply the standardized values by the eigenvector scores.

4.6 Application of Unbiased RE-EM tree

In this section the application of unbiased RE-EM tree is illustrated as an extension to the regression tree (RT) for the identification of dietary patterns where nutritional data collected repeatedly over time. Similar to RT, unbiased RE-EM tree is a data-driven outcome-dependent approach for dietary pattern analysis. Unbiased RE-EM tree considers ….. Based on unbiased RE-

EM tree illustrated in figure 4-8, the following food groups selected as important contributors to the TB-BMD development: Dark green vegetables, creams, and poultry for girls, and, Dark green vegetables, red meat, soft drinks, Eggs, and tomato for boys. This unbiased Re-EM tree resulted in 11 nodes based on the combination of these food groups. For example, participants in the 3rd group are girls who are classified based on low consumption of dark green vegetables (≤27.03 gr), and high consumption of creams (>10.57 gr). And, the combination which resulted in the 6th node includes males with dark green vegetables consumption more than 12.4 grams, tomato consumption more than 18.4 grams and again tomato consumption more than 131.5 grams. As it is apparent in the figure 4-8 most of nodes almost have similar values. Additionally, not all 25 food groups that are considered as important variables contributed to the definition of dietary pattern categories with unbiased RE-EM tree. Totally, because regression trees might result in overfitting so based on the identified food groups through unbiased RE-EM tree the primary dataset were classified into 11 categories (or nodes). Then, these nodes were used to classify the original dataset and the mean intake of 25 food groups were computed in each of the 11 clusters separately (Fig 4-9). Finally, nodes were combined based on their similarity in mean intake of food groups to extract the dietary patterns. The emerged dietary patterns were donated names subjectively regarding food groups with higher mean intake.

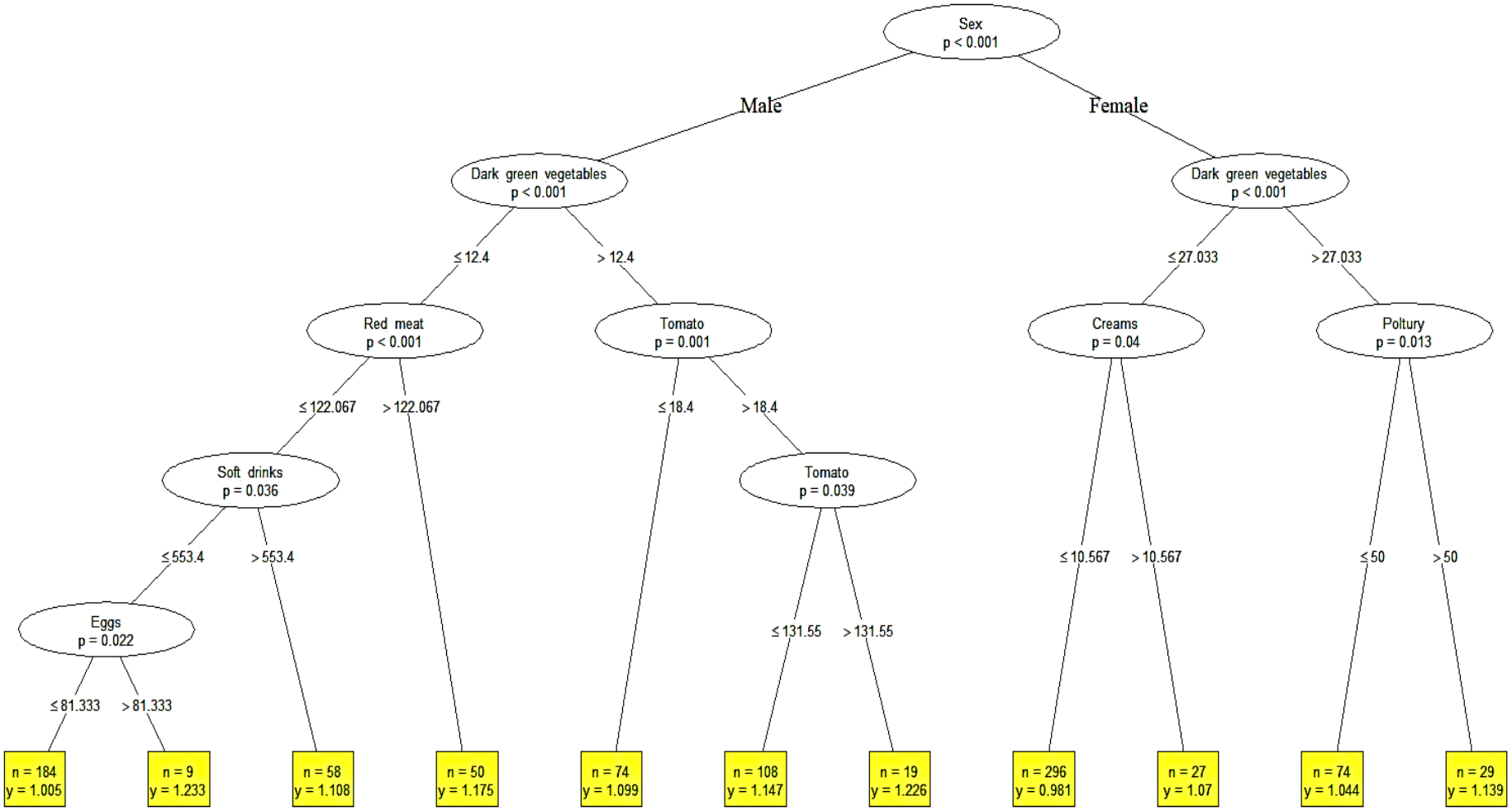

Figure 4-8 Tree structure estimated by the unbiased RE-EM tree method for the BMAS data with 25 food groups as variables, and TBBMD as response variable; all the numbers used for splitting the tree are in grams

Based on the mean scores of food groups in each node, figure 4-9 presents that node 1 and 10 could be combined and considered as “low-fat milk”, nodes 2,4,6,7, and 8 as “Mixed” and nodes 3,5, and 9 as “ high-sugar”, and node 11 as “High-fat milk” dietary patterns. Table A.1 in Appendix A summarizes the values of food groups in each node. Interestingly, results illustrate that individuals are attached to healthy diet more than other food groups as the change in healthy food groups is not noticeable, this is exactly the results from regression tree which was not helpful in the definition of dietary patterns and identification of change in diet.

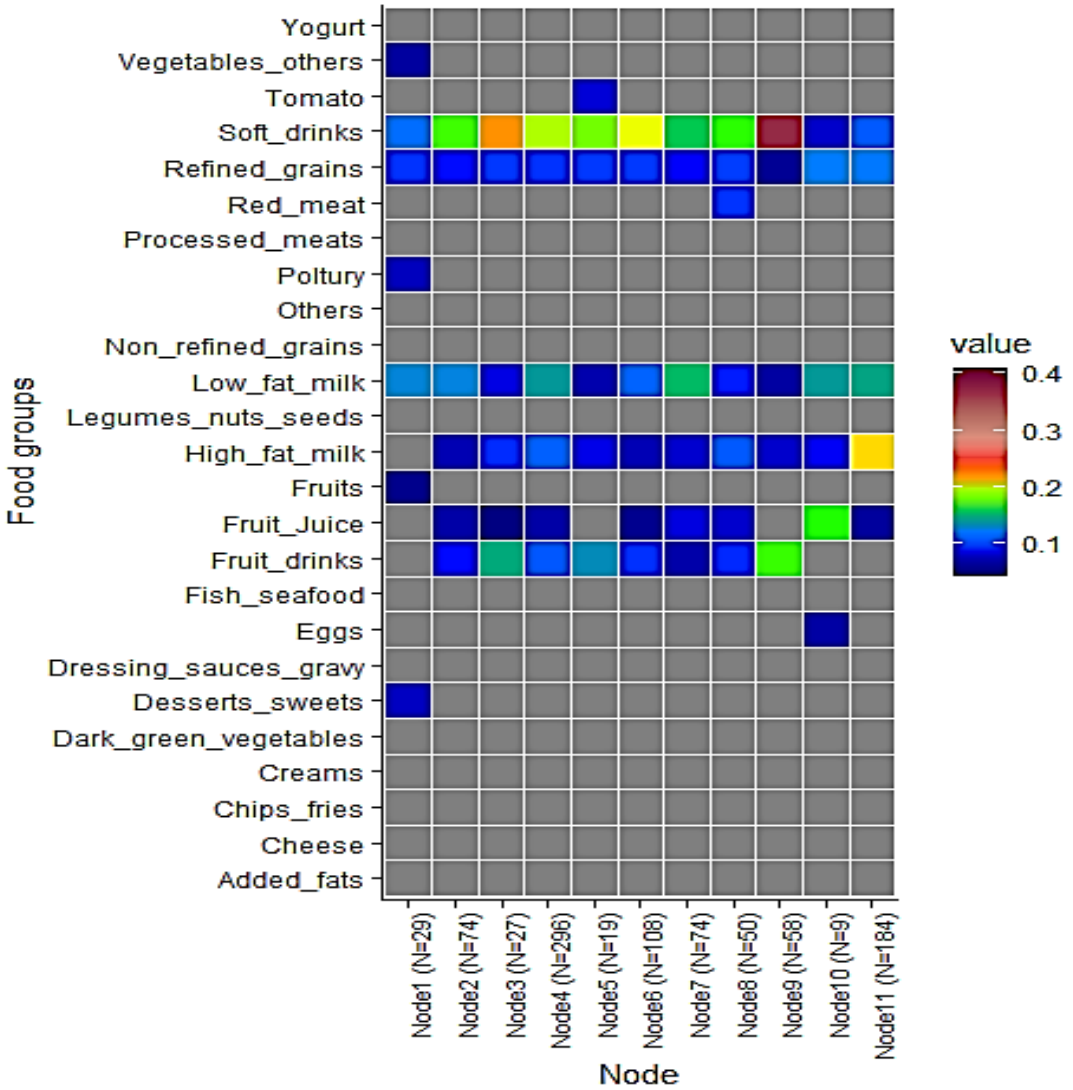

Figure 4-9 Mean score of food groups in each node separately, means are rescaled by dividing each mean value in a node to the total mean value in the node

In order to identify the stability of the dietary patterns over time a sequence plot was graphed. Examining the cluster membership over time using sequence plot showed that, while children do change their diet, they are more likely to continue following the same dietary pattern as they did at an earlier age; about half of the children continued to follow the same pattern at a later age. This helps to quantify the extent to which dietary patterns are formed in childhood and continued into adulthood (Fig 4-10). The next important step in supervised method for the identification of dietary patterns is to identify the prediction ability of these derived patterns for bone development over time which is not described in here as the aim of this thesis is showing how to identify patterns more accurately based on a labeling factor.

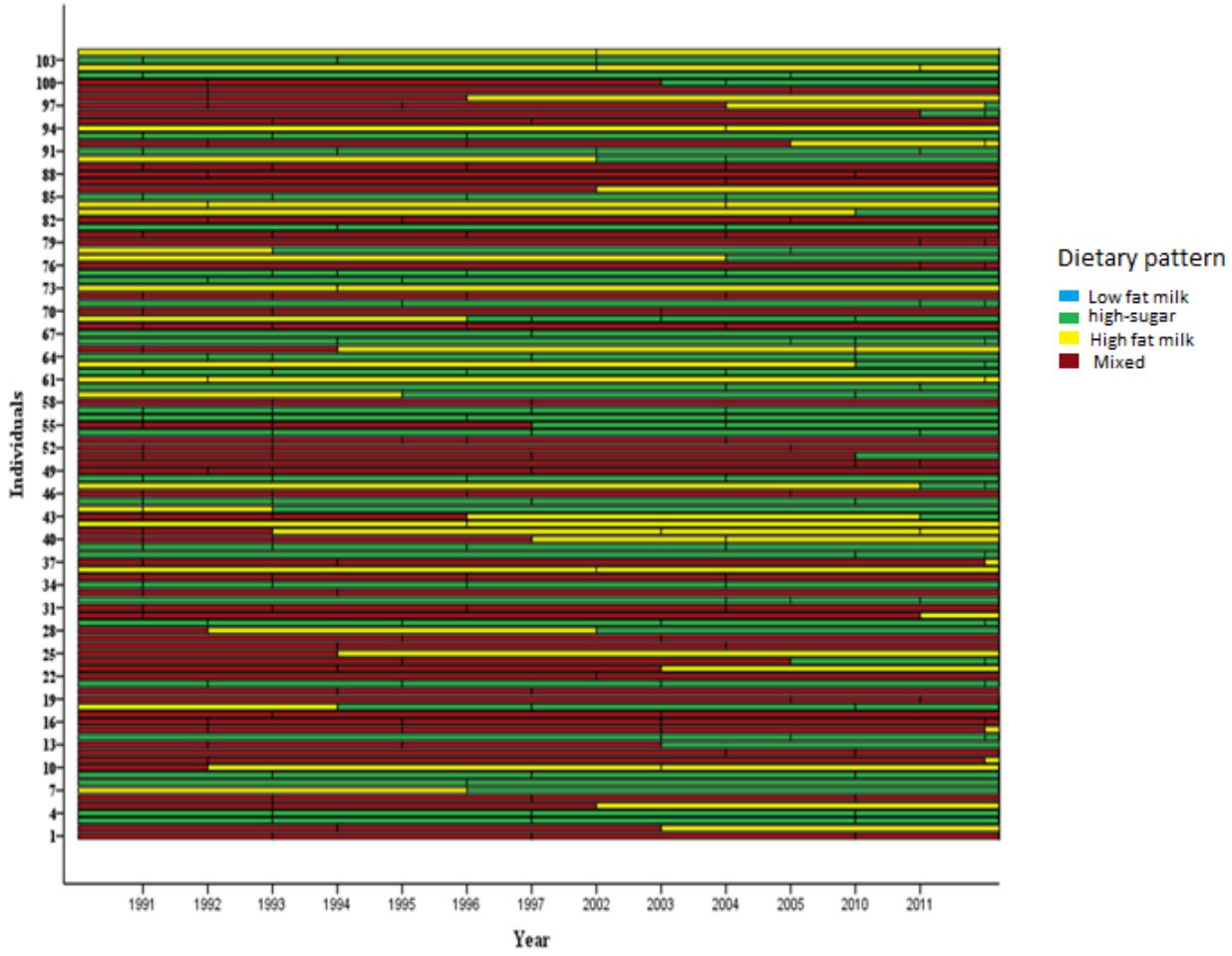

Figure 4-10 Sequence index plot illustrating changes in cluster membership over time for each individual

CHAPTER 5
DISCUSSION

An essential aspect of data analysis is to realize when a model for the analysis of correlated data is sufficient, and when an independent model is appropriate for data analysis. Many nutritional longitudinal studies have identified dietary patterns using cross-sectional statistical methods, including principal component analysis (PCA) and classification and regression trees (CART). Although these methods perform well in the identification of dietary patterns (DPs) in cross-sectional nutritional data, their assumptions may not be reasonable for the identification of dietary patterns in longitudinal nutritional data. This thesis motivated by the analytical challenges encountered in identifying dietary patterns in longitudinal data, introduced the application of LPCA and unbiased RE-EM tree methods to a representative sub-sample of Saskatchewan Bone Mineral Study (BMAS) for the identification of dietary patterns.

Although, cross-sectional pattern recognition (PR) methods such as PCA are widely applied to the identification of dietary patterns in longitudinal nutritional studies, the application of PR methods suitable for the analysis of longitudinal data remained scarce. For longitudinal nutritional data, the application of traditional cross-sectional PR methods is not adequate; in particular, the change in diet over time necessitates the consideration of statistical dependence

arising from the same individuals measured repeatedly over time. Results from several cross-sectional PR approaches to dietary pattern analysis in longitudinal studies are not considered in longitudinal studies are not considered the time-varying association among food groups (12,15,16), as a results studies applied the same cross-sectional PR method (e.g. PCA) in different ways to their data. To overcome the instability of the analysis approach of longitudinal nutritional data, which restricts the comparisons among studies, methods suitable for the analysis of longitudinal nutritional data should be considered. The LPCA and unbiased RE-EM tree represent extensions of PCA and regression tree (RT) to multi- or high-dimensional correlated data.

The example to describe the importance of considering time-varying associations among individuals, centered on scrutinizing the common approaches for the identification of dietary patterns in our longitudinal study of BMAS sample. Conducting separate PCA, as one of the popular cross-sectional PR methods, at each time point of BMAS data yielded in different number of principal components (PCs) with different interpretations. Additionally, results from the correlation analysis among different time points with respect to the first two PCs showed that the correlation between loadings not only is not high among most of the time points but also is not the same for all time points. For example, the correlation of first PC between years 1991 and 1993 is different from the correlation of first PC between years 1991 and 1995 (Fig 4-3). It is important to note that the correlation of first PC between years 1991 and 1993 (~ -0.75) is almost high but the correlation of first PC between years 1991 and 1995 (~ -0.2) is not high. Therefore, there is a chance that the first PC at year 1991 has a different meaning from the first PC at year 1995. In this sense, it is possible that studies arbitrarily select PCs with almost similar loadings at all-time points ignoring the proportion of variance explained by PCs. This arbitrarily selection of PCs leads to

losing important information by removing the most important PCs that explain the highest percentage of total variability at a specific time point in data.

Another common approach for identifying dietary patterns and their trajectories in the literature is to perform PCA (or other methods) at one of the time points of the study and applying the loadings from that time point to the data at another one (15,16). However, results from the correlation analysis between food groups at each time point separately showed that the correlation between food groups changes from one time point to the other. As an example, the correlation between Added fats and fruits is around -0.2, -0.1, 0.1, 0.6 at years 1991, 1996, 2005, and 2011 which shows different amount of food consumption over time. As a result, applying the loadings derived from PCA at one time point to the data at other time points will not discover the true contribution of food groups to the PCs because PCA creates PCs based on the linear combination of food groups. In this thesis, problems of applying cross-sectional PR methods to the identification of dietary patterns in longitudinal studies were assessed using PCA. However, it is expected that the application of regression tree to the longitudinal data for the identification of dietary patterns lead to the same problems.

In sum, these approaches did not considered the time-varying associations among individual's food consumption to handle large-dimensional nutritional data. Other approaches including cluster analysis (CA) and reduced rank regression (RRR) are also applied to the longitudinal data for the identification of dietary patterns in the literature, however, these approaches share the same problems defined by illustrating the application of PCA to the BMAS data.

In this thesis, the simplicity in the computation and interpretation of results from the application of the LPCA and RE-EM to the BMAS sample showed that these methods could

provide a better solution than their comparable cross-sectional approaches available in the literature. This superiority resides in the fundamental differences in the concepts behind these methods that is LPCA and RE-EM tree are considering the time-varying associations among food groups. Beside the incorporation of longitudinal information, the main benefit of these PR methods for the analysis of longitudinal data is the accurate estimation and simple interpretation of the emerged patterns. The results of the application of LPCA and unbiased RE-EM tree irrespective of their methodological differences were almost similar. The results showed that while BMAS participants do change their diet, they are more likely to continue following the same dietary patterns as they did at an earlier age. More interestingly, both methods identifies relatively the same dietary patterns including mixed, high-sugar, low-fat milk and high-fat milk dietary patterns. Moreover, based on both methods the variation in healthy food groups were not noticeable. A main explanation for the similarity in the identified dietary patterns based on the two different approach may be associated with the correlation among food groups. However, future work is required to compare these two methods in terms of their prediction ability.

In concluding, the findings of this study suggest that application of LPCA and RE-EM tree to the longitudinal data for the identification of dietary patterns and their trajectories represents progress. Additionally, it is important to emphasize that identifying patterns in longitudinal nutritional data using these methods is not computationally as expensive as identifying patterns using cross-sectional PR methods. By introducing the application of these methods as an alternative to the cross-sectional approaches, this thesis is expected to assist the nutritional epidemiologists and researchers to become more familiar with the application of appropriate methods to the identification of dietary patterns and their trajectories in longitudinal design.

This thesis marked the first application of the LPCA and unbiased RE-EM tree methods, as a substitute to the PCA and RT, to the identification of dietary patterns and their meaningful dynamic changes in longitudinal studies. Additionally, this thesis for the first time discussed the problems associated with the application of cross-sectional PR methods to the identification of dietary patterns in longitudinal studies. However, this study had some limitation. The main limitation of this study was the small sample size because of which a separate analysis for females and males could not be conducted based on LPCA method. Additionally, due to the small sample size a comparison could not be made between this study and other studies in terms of the identified dietary patterns because the results could not be generalized. Only one 24-hour recall was collected each year for BMAS participants at study years after 1997 which might misrepresent the usual intake of the BMAS participant and affects the final results. Finally, BMAS participants were a grab sample of Caucasian people which means that the participants were drawn from the available population. Therefore, the results of this study could not be generalized to the whole population.

Even though this study showed that using LPCA and unbiased RE-EM tree is computationally cheap, this is not to say that these methods are the only existing methods in the literature for extracting patterns in longitudinal studies. Future work requires the application of other pattern recognition methods suitable for the analysis of longitudinal nutritional data such as mixed effect random forest for correlated data (121). Methodologists also should pay more attention to the area of pattern recognition in longitudinal studies in terms of both developing efficient methods and introducing their application.

Finally yet importantly, this study showed the importance of considering time-varying associations among food groups in longitudinal studies and the application of two PR methods for the analysis of longitudinal data, but no comparison is made between the methods in terms of their

prediction ability. Future studies should be made to compare these methods to each other and to the other methods appropriate for this kind of data. In future studies, it is also suggested to apply these methods to data with larger sample size and make clinical comparisons among studies in terms of the derived dietary patterns using methods discussed in this thesis. Unfortunately, the task of analyzing data with longitudinal PR methods has not been greatly facilitated because software's are not widely available.

BIBLIOGRAPHY

1. Liang K-Y, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika [Internet]. 1986;73(1):13–22. Available from: http://www.biostat.jhsph.edu/~fdominic/teaching/bio655/references/extra/liang.bka.1986.pdf
2. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. Analysis of Longitudinal Data. 2002;379.
3. Aksman LM, Lythgoe DJ, Williams SCR, Jokisch M, Mönninghoff C, Streffer J, et al. Making use of longitudinal information in pattern recognition. Hum Brain Mapp. 2016;37(12):4385–404.
4. Movassagh EZ, Vatanparast H. Current Evidence on the Association of Dietary Patterns and Bone Health: A Scoping Review. Adv Nutr An Int Rev J. 2017 Jan 20;8(1):1.2-16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28096123
5. Kinson CL, Annie Q, Culpepper S, Marden J, Simpson D. Longitudinal principal component analysis for binary and continuous data. 2017. Available from: http://hdl.handle.net/2142/98374
6. Nadeau KJ, Maahs DM, Daniels SR, Eckel RH. Childhood obesity and cardiovascular disease: links and prevention strategies. Nat Rev Cardiol. 2011 Jun 14;8(9):513–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21670745
7. Kant AK. Dietary patterns and health outcomes. J Am Diet Assoc. 2004 Apr;104(4):615–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15054348
8. Hoffmann K, Schulze MB, Schienkiewitz A, Nöthlings U, Boeing H. Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. Am J Epidemiol. 2004;159:935–44. Available from: http://dx.doi.org/10.1093/aje/kwh134
9. Newby PK, Tucker KL. Empirically derived eating patterns using factor or cluster analysis: a review. Nutr Rev. 2004;62:177-203.
10. Jordan M, Kleinberg J, Scho B. Pattern Recognition and Machine Learning. Second. Springer

(India) Private Limited;

11. Birch LL. Development of Food Preferences. Annu Rev Nutr. 1999;19(1):41–62.
12. Northstone K, Emmett PM. Are dietary patterns stable throughout early and mid-childhood? A birth cohort study. Br J Nutr. 2008 Nov 1;100(5):1069–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18377690
13. Mikkilä V, Räsänen L, Raitakari OT, Pietinen P, Viikari J. Consistent dietary patterns identified from childhood to adulthood: The Cardiovascular Risk in Young Finns Study. Br J Nutr. 2005;93(6):923–31.
14. Northstone K, Emmett P. Multivariate analysis of diet in children at four and seven years of age and associations with socio-demographic characteristics. Eur J Clin Nutr. 2005;59(6):751–60.
15. Asghari G, Rezazadeh A, Hosseini-Esfahani F, Mehrabi Y, Mirmiran P, Azizi F. Reliability, comparative validity and stability of dietary patterns derived from an FFQ in the Tehran Lipid and Glucose Study. Br J Nutr. 2012;108(6):1109–17.
16. Movassagh E, Baxter-Jones A, Kontulainen S, Whiting S, Vatanparast H. Tracking Dietary Patterns over 20 Years from Childhood through Adolescence into Young Adulthood: The Saskatchewan Pediatric Bone Mineral Accrual Study. Nutrients. 2017 Sep 8;9(9):990. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28885565
17. Crozier SR, Robinson SM, Godfrey KM, Cooper C, Inskip HM. Dietary patterns change little from before to during pregnancy. J Nutr. 2009;139(10):1956–63.
18. Borland SE, Robinson SM, Crozier SR, Inskip HM. Stability of dietary patterns in young women over a 2-year period. Eur J Clin Nutr. 2008;62(1):119–26.
19. Lioret S, Betoko A, Forhan A, Charles M-A, Heude B, de Lauzon-Guillain B. Dietary Patterns Track from Infancy to Preschool Age: Cross-Sectional and Longitudinal Perspectives. J Nutr. 2015;145(4):775–82.
20. Brazionis L, Golley RK, Mittinty MN, Smithers LG, Emmett P, Northstone K, et al. Characterization of transition diets spanning infancy and toddlerhood: A novel, multiple-time-point

application of principal components analysis. Am J Clin Nutr. 2012;95(5):1200–8.

21. Shao A, Drewnowski A, Willcox DC, Kramer L, Lausted C, Eggersdorfer M, et al. Optimal nutrition and the ever-changing dietary landscape: a conference report. Eur J Nutr [Internet]. 2017;56:1-21. doi: 10.1007/s00394-017-1460-9. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5442251/

22. Tapsell LC, Neale EP, Satija A, Hu FB. Foods, Nutrients, and Dietary Patterns: Interconnections and Implications for Dietary Guidelines. Adv Nutr. 2016;7(3):445–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27184272

23. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol. 2002;13:3–9.

24. Kant AK. Indexes of overall diet quality: a review. J Am Diet Assoc. 1996;96:785-91. doi: 10.1016/S0002-8223(96)00217-9.

25. Waijers PM, Feskens EJ, Ocke MC. A critical review of predefined diet quality scores. Br J Nutr. 2007;97:219-31. doi: 10.1017/S0007114507250421.

26. Hodge A, Bassett J. What can we learn from dietary pattern analysis? Public Heal Nutr. 2016;19:191-4. doi: 10.1017/S1368980015003730.

27. Schulze MB, Hoffmann K, Kroke A, Boeing H. An approach to construct simplified measures of dietary patterns from exploratory factor analysis. Br J Nutr. 2003;89:409-19. doi: 10.1079/BJN2002778.

28. S Eilat-Adar, M Mete, A Fretts, RR Fabsitz, V Handeland, ET Lee, C Loria, J Xu, J Yeh BH. Dietary Patterns and Their Association with Cardiovascular Risk Factors in a Population Undergoing Lifestyle Changes: The Strong Heart Study. Nutr Metab Cardiovasc Dis. 2013;23(6):528–35.

29. Smith ADAC, Emmett PM, Newby PK, Northstone K. Dietary patterns obtained through principal components analysis: the effect of input variable quantification. Br J Nutr. 2013 May 28;109(10):1881–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22950853

30. Schwedhelm C, Iqbal K, Knüppel S, Schwingshackl L, Boeing H. Contribution to the understanding

of how principal component analysis–derived dietary patterns emerge from habitual data on food consumption. Am J Clin Nutr. 2018;107(2):227–35.

31. Thorpe MG, Milte CM, Crawford D, McNaughton SA. A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. Int J Behav Nutr Phys Act. 2016;13(1):1–14. Available from: http://dx.doi.org/10.1186/s12966-016-0353-2

32. Garcia-Larsen V, Morton V, Norat T, Moreira A, Potts JF, Reeves T, et al. Dietary patterns derived from principal component analysis (PCA) and risk of colorectal cancer: a systematic review and meta-analysis. Eur J Clin Nutr. 2019;73(3):366–86. Available from: http://dx.doi.org/10.1038/s41430-018-0234-7

33. Weismayer C, Anderson JG, Wolk A. Changes in the Stability of Dietary Patterns in a Study of Middle-Aged Swedish Women. J Nutr. 2006;136(6):1582–7.

34. Mikkilä V, Räsänen L, Raitakari OT, Pietinen P, Viikari J. Consistent dietary patterns identified from childhood to adulthood: the cardiovascular risk in Young Finns Study. Br J Nutr. 2005 Jun;93(6):923–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16022763

35. Andersen LB, Mølgaard C, Ejlerskov KT, Trolle E, Michaelsen KF, Bro R, et al. Development of Dietary Patterns Spanning Infancy and Toddlerhood: Relation to Body Size, Composition and Metabolic Risk Markers at Three Years. AIMS public Heal. 2015;2(3):332–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29546114

36. Northstone K, Joinson C, Emmett P, Ness A, Paus T. Are dietary patterns in childhood associated with IQ at 8 years of age? A population-based cohort study. J Epidemiol Community Health. 2012;66(7):624–8.

37. G.L. A, P.M. E, K. N, S.A. J. Tracking a dietary pattern associated with increased adiposity in childhood and adolescence. Obesity. 2014;22(2):458–65. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed12&NEWS=N&AN=2014092602

38. Mertens E, Markey O, Geleijnse JM, Givens DI, Lovegrove JA. Dietary patterns in relation to

cardiovascular disease incidence and risk markers in a middle-aged british male population: Data from the caerphilly prospective study. Nutrients. 2017;9(1):1–16.

39. Hearty ÁP, Gibney MJ. Analysis of meal patterns with the use of supervised data mining techniques - Artificial neural networks and decision trees. Am J Clin Nutr. 2008;88(6):1632–42.

40. Lazarou C, Karaolis M, Matalas AL, Panagiotakos DB. Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. Comput Methods Programs Biomed. 2012;108(2):706–14. Available from: http://dx.doi.org/10.1016/j.cmpb.2011.12.011

41. Panaretos D, Koloverou E, Dimopoulos AC, Kouli GM, Vamvakari M, Tzavelas G, et al. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002-2012): The ATTICA study. Br J Nutr. 2018;120(3):326–34.

42. Biesbroek S, Van Der A. DL, Brosens MCC, Beulens JWJ, Verschuren WMM, Van Der Schouw YT, et al. Identifying cardiovascular risk factor-related dietary patterns with reduced rank regression and random forest in the EPIC-NL cohort. Am J Clin Nutr. 2015;102(1):146–54.

43. Kontulainen SA, Kawalilak CE, Johnston JD, Bailey DA. Prevention of Osteoporosis and Bone Fragility:A Pediatric Concern. Am J Lifestyle Med. 2013 May 23;7:405–17. Available from: https://doi.org/10.1177/1559827613487664

44. Alghadir AH, Gabr SA, Al-Eisa E. Physical activity and lifestyle effects on bone mineral density among young adults: sociodemographic and biochemical analysis. J Phys Ther Sci. 2015 Jul [cited 2018 Nov 10];27(7):2261–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26311965

45. Cashman KD. Diet and control of osteoporosis. Funct foods, ageing Degener Dis Cambridge, UK Woodhead Publ Ltd. 2004;83–114.

46. Cashman KD. Diet, nutrition, and bone health. J Nutr. 2007;137:2507S-2512S. doi: 10.1093/jn/137.11.2507S.

47. Heaney RP, Abrams S, Dawson-Hughes B, Looker A, Marcus R, Matkovic V, et al. Peak bone mass. Osteoporos Int. 2000;11:985-1009. doi: 10.1007/s001980070020.

48. Bonjour JP, Gueguen L, Palacios C, Shearer MJ, Weaver CM. Minerals and vitamins in bone health: the potential value of dietary enhancement. Br J Nutr. 2009;101:1581-96. doi: 10.1017/S0007114509311721. Epub 2009.

49. Huncharek M, Muscat J, Kupelnick B. Impact of dairy products and dietary calcium on bone-mineral content in children: results of a meta-analysis. Bone. 2008;43:312-21. doi: 10.1016/j.bone.2008.02.022. Epub 2008.

50. Mouratidou T, Vicente-Rodriguez G, Gracia-Marco L, Huybrechts I, Sioen I, Widhalm K, et al. Associations of dietary calcium, vitamin D, milk intakes, and 25-hydroxyvitamin D with bone mass in Spanish adolescents: the HELENA study. J Clin Densitom. 2013;16:110-7. doi: 10.1016/j.jocd.2012.07.008. Epub 2012.

51. Optimal calcium intake. NIH Consens Statement. 1994;12:1–31.

52. Staff I of M, Vedral JL. Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride. National Academies Press; 1900. 448 p.

53. Whiting SJ, Vatanparast H, Baxter-Jones A, Faulkner RA, Mirwald R, Bailey DA. Factors that Affect Bone Mineral Accrual in the Adolescent Growth Spurt. J Nutr. 2004 Mar 1;134(3):696S-700S. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14988470

54. Gropper SAS, Smith JL. Advanced nutrition and human metabolism. Wadsworth/Cengage Learning; 2013. 586 p.

55. Holick MF. Sunlight and vitamin D for bone health and prevention of autoimmune diseases, cancers, and cardiovascular disease. Am J Clin Nutr. 2004 Dec 1;80(6):1678S-1688S. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15585788

56. Bonjour J-P, Ammann P, Chevalley T, Ferrari S, Rizzoli R, Ferrari S, et al. Nutritional aspects of bone growth; and overview. In: Nutritional Aspects of Bone Health. Cambridge: Royal Society of Chemistry; 2003. p. 111–27. Available from: http://ebook.rsc.org/?DOI=10.1039/9781847551559-00111

57. Huang RY, Huang CC, Hu FB, Chavarro JE. Vegetarian Diets and Weight Reduction: a Meta-

Analysis of Randomized Controlled Trials. J Gen Intern Med. 2016;31:109–16.

58. McGartland CP, Robson PJ, Murray LJ, Cran GW, Savage MJ, Watkins DC, et al. Fruit and vegetable consumption and bone mineral density: the Northern Ireland Young Hearts Project. Am J Clin Nutr. 2004;80:1019-23. doi: 10.1093/ajcn/80.4.1019.

59. Vatanparast H, Baxter-Jones A, Faulkner RA, Bailey DA, Whiting SJ. Positive effects of vegetable and fruit consumption and calcium intake on bone mineral accrual in boys during growth from childhood to adolescence: the University of Saskatchewan Pediatric Bone Mineral Accrual Study. Am J Clin Nutr. 2005;82:700-6. doi: 10.1093/ajcn.82.3.700.

60. New SA. Exercise, bone and nutrition. Proc Nutr Soc. 2001 May;60(2):265–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11681642

61. de Jonge EAL, Rivadeneira F, Erler NS, Hofman A, Uitterlinden AG, Franco OH, et al. Dietary patterns in an elderly population and their relation with bone mineral density: the Rotterdam Study. Eur J Nutr. 2018 Feb 24;57(1):61–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27557817

62. de Jonge EALL, Kiefte-de Jong JC, de Groot LCPGMPGM, Voortman T, Schoufour JD, Zillikens MC, et al. Development of a Food Group-Based Diet Score and Its Association with Bone Mineral Density in the Elderly: The Rotterdam Study. Nutrients. 2015 Aug 18;7(8). Available from: http://www.mdpi.com/2072-6643/7/8/5317

63. McNaughton SA, Wattanapenpaiboon N, Wark JD, Nowson CA. An Energy-Dense, Nutrient-Poor Dietary Pattern Is Inversely Associated with Bone Health in Women. J Nutr. 2011 Aug 1;141(8):1516–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21653576

64. Tucker KL, Chen H, Hannan MT, Cupples LA, Wilson PW, Felson D, et al. Bone mineral density and dietary patterns in older adults: the Framingham Osteoporosis Study. Am J Clin Nutr. 2002 Jul 1;76(1):245–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12081842

65. Hardcastle AC, Aucott L, Fraser WD, Reid DM, Macdonald HM. Dietary patterns, bone resorption and bone mineral density in early post-menopausal Scottish women. Eur J Clin Nutr. 2011 Mar

22;65(3):378–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21179049

66. Ward KA, Prentice A, Kuh DL, Adams JE, Ambrosini GL. Life Course Dietary Patterns and Bone Health in Later Life in a British Birth Cohort Study. J Bone Miner Res. 2016;31(6):1167–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26817442

67. Langsetmo L, Hanley DA, Prior JC, Barr SI, Anastassiades T, Towheed T, et al. Dietary patterns and incident low-trauma fractures in postmenopausal women and men aged ≥50 y: a population-based cohort study. Am J Clin Nutr. 2011 Jan 1;93(1):192–9. Available from: https://academic.oup.com/ajcn/article/93/1/192/4597637

68. Okubo H, Sasaki S, Horiguchi H, Oguma E, Miyamoto K, Hosoi Y, et al. Dietary patterns associated with bone mineral density in premenopausal Japanese farmwomen. Am J Clin Nutr. 2006 May 1;83(5):1185–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16685064

69. Kontogianni MD, Melistas L, Yannakoulia M, Malagaris I, Panagiotakos DB, Yiannakouris N. Association between dietary patterns and indices of bone mass in a sample of Mediterranean women. Nutrition. 2009 Feb;25(2):165–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18849146

70. Karamati M, Jessri M, Shariati-Bafghi S-E, Rashidkhani B. Dietary Patterns in Relation to Bone Mineral Density Among Menopausal Iranian Women. Calcif Tissue Int. 2012 Jul 27;91(1):40–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22644320

71. Fairweather-Tait SJ, Skinner J, Guile GR, Cassidy A, Spector TD, MacGregor AJ. Diet and bone mineral density study in postmenopausal women from the TwinsUK registry shows a negative association with a traditional English dietary pattern and a positive association with wine. Am J Clin Nutr. 2011 Nov 1;94(5):1371–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21940596

72. Whittle CR, Woodside J V., Cardwell CR, McCourt HJ, Young IS, Murray LJ, et al. Dietary patterns and bone mineral status in young adults: the Northern Ireland Young Hearts Project. Br J Nutr. 2012 Oct 4;108(08):1494–504. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22214826

73. Anderson JJB, Garner SC, Klemmer PJ. Diet, Nutrients, and Bone Health. CRC Press; 2011.

74. Malina RM, Bouchard C, Bar-Or O. Growth, maturation, and physical activity. Human kinetics; 2004.

75. Karlberg J. Secular Trends in Pubertal Development. Horm Res Paediatr. 2002;57(2):19–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12065922

76. Baxter-Jones AD, Mirwald RL, McKay HA, Bailey DA. A longitudinal analysis of sex differences in bone mineral accrual in healthy 8-19-year-old boys and girls. Ann Hum Biol. 2003;30:160–75.

77. Molgaard C, Thomsen BL, Michaelsen KF. Whole body bone mineral accretion in healthy children and adolescents. Arch Dis Child. 1999;81:10–5.

78. Beunen GP, Rogol AD, Malina RM. Indicators of biological maturation and secular changes in biological maturation. Food Nutr Bull. 2006;27.

79. Gabel L, Macdonald HM, McKay HA. Sex differences and growth-related adaptations in bone microarchitecture, geometry, density and strength from childhood to early adulthood: a mixed longitudinal HR-pQCT study. J Bone Miner Res. 2017;32:250–63. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5233447/

80. Papadimitriou A. Sex differences in the secular changes in pubertal maturation. Pediatrics. 2001;108.

81. Chevalley T, Rizzoli R, Nydegger V, Slosman D, Rapin CH, Michel JP, et al. Effects of calcium supplements on femoral bone mineral density and vertebral fracture rate in vitamin-D-replete elderly patients. Osteoporos Int. 1994;4:245–52.

82. Wang Q, Wang XF, Iuliano-Burns S, Ghasem-Zadeh A, Zebaze R, Seeman E. Rapid growth produces transient cortical weakness: a risk factor for metaphyseal fractures during puberty. J Bone Min Res. 2010;25:1521–6.

83. Kirmani S, Christen D, van Lenthe GH, Fischer PR, Bouxsein ML, McCready LK, et al. Bone Structure at the Distal Radius During Adolescent Growth. J Bone Miner Res. 2009;24:1033–42. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2683647/

84. Maynard LM, Guo SS, Chumlea WC, Roche AF, Wisemandle WA, Zeller CM, et al. Total-body

and regional bone mineral content and areal bone mineral density in children aged 8-18 y: the Fels Longitudinal Study. Am J Clin Nutr . 1998;68:1111–7. Available from: http://dx.doi.org/10.1093/ajcn/68.5.1111

85. Nishiyama KK, Macdonald HM, Moore SA, Fung T, Boyd SK, McKay HA. Cortical porosity is higher in boys compared with girls at the distal radius and distal tibia during pubertal growth: an HR-pQCT study. J Bone Min Res. 2012;27:273–82.

86. Bailey DA, Wedge JH, McCulloch RG, Martin AD, Bernhardson SC. Epidemiology of fractures of the distal end of the radius in children as associated with growth. J Bone Jt Surg Am. 1989;71:1225–31.

87. Parfitt AM. The two faces of growth: benefits and risks to bone integrity. Osteoporos Int. 1994;4:382–98.

88. McKay HA, Petit MA, Bailey DA, Wallace WM, Schutz RW, Khan KM. Analysis of proximal femur DXA scans in growing children: comparisons of different protocols for cross-sectional 8-month and 7-year longitudinal data. J Bone Min Res. 2000;15:1181–8.

89. Bonjour JP, Ammann P, Chevalley T, Rizzoli R. Protein intake and bone growth. Can J Appl Physiol. 2001;26:S153-66.

90. Bailey DA, Martin AD, McKay HA, Whiting S, Mirwald R. Calcium accretion in girls and boys during puberty: a longitudinal analysis. J Bone Min Res. 2000;15:2245–50.

91. Iuliano-Burns S, Whiting SJ, Faulkner RA, Bailey DA. Levels, sources, and seasonality of dietary calcium intake in children and adolescents enrolled in the University of Saskatchewan pediatric bone mineral accrual study. Nutr Res. 1999;

92. Whiting SJ, Healey A, Psiuk S, Mirwald R, Kowalski K, Bailey DA. Relationship between carbonated and other low nutrient dense beverages and bone mineral content of adolescents. Nutr Res. 2001;

93. Carter LM, Whiting SJ, Drinkwater DT, Zello GA, Drinkwater DT, Faulkner RA, et al. Self-Reported Calcium Intake and Bone Mineral Content in Children and Adolescents. J Am Coll Nutr.

2001;20(5):502–9.

94. Mundt CA, Baxter-Jones ADG, Whiting SJ, Bailey DA, Faulkner RA, Mirwald RL. Relationships of activity and sugar drink intake on fat mass development in youths. Med Sci Sports Exerc. 2006;38(7):1245–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16826021

95. Vatanparast H, Bailey DA, Baxter-Jones ADG, Whiting SJ. The Effects of Dietary Protein on Bone Mineral Mass in Young Adults May Be Modulated by Adolescent Calcium Intake. J Nutr. 2007 Dec 1;137(12):2674–9.

96. Vatanparast H, Bailey DA, Baxter-Jones ADG, Whiting SJ. Calcium requirements for bone growth in Canadian boys and girls during adolescence. Br J Nutr. 2010 Feb;103(4):575–80.

97. Jolliffe I. Principal Compopnent Analysis. 2nd ed. Wiley online library. Wiley online library; 2002.

98. Ramsay JO, Silverman BW. Functional Data Analysis. In: Springer New York. 2nd ed. Springer New York; 2005. p. 514–8. Available from: http://www.bookmetrix.com/detail/book/231b74d4-0a74-4145-952b-e1b4b2a22fc3#citations

99. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. J Am Stat Assoc. 2005;100(470):577–90.

100. Hall P, Hosseini-nasab M. On Properties of Functional Principal Components Analysis Author (s): Peter Hall and Mohammad Hosseini-Nasab Source : Journal of the Royal Statistical Society . Series B (Statistical Methodology), Vol . 68 , Published by : Wiley for the Royal Statisti. Jrss. 2006;68(1):109–26.

101. Greven S, Crainiceanu C, Caffo B, Reich D. Longitudinal functional principal component analysis. Electron J Stat. 2010;4(2):1022–54.

102. Jiang CR, Wang JL. Covariate adjusted functional principal components analysis for longitudinal data. Ann Stat. 2010;38(2):1194–226.

103. Zipunnikov V, Greven S, Shou H, Caffo BS, Reich DS, Crainiceanu CM. Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis. Ann Appl Stat. 2014;8(4):2175–202.

104. Morgan JN, Sonquist JA. Problems in the Analysis of Survey Data , and a Proposal. Am Stat Assoc. 1963;58(302):415–34.

105. Brieman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. 2nd ed. Chapman & Hall; 1984.

106. Loh W-Y, Vanichsetakul N. Tree-Structured Classification Via Generalized Discriminant Analysis. Am Stat Assoc. 1988;83(403):715–25.

107. Loh W. Regression Trees with Unbiased Variable Selection and interaction detection. Stat Sin. 2002;12:361–86.

108. Ahn H, Loh W-Y. Tree-Structured Proportional Hazards Regression Modeling. Int Biometric Soc. 1994;50(2):471–85.

109. Chaudhuri P, Huang M-C, Loh W-Y, Yao R. Piecewise-polynomial regression trees. Stat Sin. 1994;4(1):143–67.

110. Alexander WP, Grimshaw SD. Treed Regression. J Comput Graph Stat. 1996;5(2):156–75.

111. Chipman HA, George EI, Mcculloch RE, Chipman HA, George E, Mcculloch RE. Bayesian CART Model Search. Am Stat Assoc. 1998;93(443):935–48.

112. Kim H, Loh W-Y. Classification Trees with Unbiased Multiway Splits. Am Stat Assoc. 2001;96(454):589–604.

113. Chan KY, Loh WY. Lotus: An algorithm for building accurate and comprehensible logistic regression trees. J Comput Graph Stat. 2004;13(4):826–52.

114. Segal MR. Tree-structured methods for longitudinal data. J Am Stat Assoc. 1992;87(418):407–18.

115. Zhang H. Classification trees for multiple binary responses. J Am Stat Assoc. 1998;93(441):180–93.

116. De'ath G. Multivariate regression trees: A new technique for modeling species-environment relationships. Ecology. 2002;83(4):1105–17.

117. Abdolell M, LeBlanc M, Stephens D, Harrison R V. Binary partitioning for continuous longitudinal data: Categorizing a prognostic variable. Stat Med. 2002;21(22):3395–409.

118. Hsiao WC, Shih YS. Splitting variable selection for multivariate regression trees. Stat Probab Lett. 2007;77(3):265–71.

119. Eo SH, Cho HJ. Tree-Structured Mixed-Effects Regression Modeling for Longitudinal Data. J Comput Graph Stat. 2014;23(3):740–60.

120. Loh WY, Zheng W. Regression trees for longitudinal and multiresponse data. Ann Appl Stat. 2013;7(1):495–522.

121. Hajjem A, Bellavance F, Larocque D. Mixed effects regression trees for clustered data. Stat Probab Lett. 2011;81(4):451–9. Available from: http://dx.doi.org/10.1016/j.spl.2010.12.003

122. Sela RJ, Simonoff JS. RE-EM trees: A data mining approach for longitudinal and clustered data. Mach Learn. 2012;86(2):169–207.

123. Fu W, Simonoff JS. Unbiased regression trees for longitudinal and clustered data. Comput Stat Data Anal. 2015;88:53–74. Available from: http://dx.doi.org/10.1016/j.csda.2015.02.004

124. Sergios Theodoridis, Koutroumbas K, Koutroumbas ST& K. Pattern Recognition. 3rd ed. Academic Press; 2006.

125. Schoenau E. Bone mass increase in puberty: what makes it happen? Horm Res. 2006;2:2–10.

126. Laird N, JH W. Random-effects models for longitudinal data. Biometrics. 1982;38(4):963–74.

127. Pekelis L. Classification and Regression Trees: A Practical Guide for Describing a Dataset. 2013. http://statweb.stanford.edu/~lpekelis/13_datafest_cart/13_datafest_cart_talk.html#(6)

128. Bailey D. The Saskatchewan Pediatric Bone Mineral Accrual Study: Bone Mineral Acquisition During the Growing Years. Int J Sports Med. 1997 Jul 9 ;18(S 3):S191–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9272847

129. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. Radiology. 2003;227(3).

APPENDIX. A

Table A.1 Mean score of food groups in each node separately, means are rescaled by dividing each mean value in a node to the total value in the node

	Node1 (N=29)	Node2 (N=74)	Node3 (N=27)	Node4 (N=296)	Node5 (N=19)	Node6 (N=108)	Node7 (N=74)	Node8 (N=50)	Node9 (N=58)	Node10 (N=9)	Node11 (N=184)
High fat milk	0.0398	0.0746	0.0950	0.1115	0.0864	0.0752	0.0821	0.1085	0.0813	0.0881	0.2082
Low fat milk	0.1303	0.1297	0.0857	0.1406	0.0725	0.1129	0.1545	0.0924	0.0680	0.1411	0.1465
Refined grains	0.0963	0.0902	0.0975	0.0963	0.0977	0.0977	0.0888	0.0991	0.0614	0.1261	0.1240
Soft drinks	0.1177	0.1719	0.2203	0.1977	0.1824	0.2047	0.1588	0.1696	0.3682	0.0812	0.1087
Fruit Juice	0.0246	0.0702	0.0511	0.0696	0.0342	0.0588	0.0852	0.0815	0.0426	0.1685	0.0656
Fruit drinks	0.0427	0.0897	0.1487	0.1078	0.1341	0.0960	0.0711	0.0946	0.1711	0.0218	0.0474
Desserts and sweets	0.0789	0.0238	0.0165	0.0380	0.0324	0.0276	0.0366	0.0244	0.0339	0.0472	0.0454
Nonrefined grains	0.0373	0.0341	0.0323	0.0261	0.0358	0.0410	0.0386	0.0232	0.0205	0.0216	0.0407
Fruits	0.0569	0.0470	0.0335	0.0428	0.0327	0.0350	0.0461	0.0407	0.0205	0.0279	0.0322
Cheese	0.0253	0.0292	0.0237	0.0230	0.0193	0.0216	0.0211	0.0203	0.0147	0.0395	0.0251
Red meat	0.0101	0.0269	0.0402	0.0222	0.0291	0.0291	0.0384	0.0967	0.0129	0.0242	0.0219
Tomato	0.0433	0.0267	0.0270	0.0232	0.0836	0.0294	0.0021	0.0234	0.0176	0.0301	0.0200
Processed meats	0.0086	0.0138	0.0101	0.0085	0.0148	0.0151	0.0224	0.0178	0.0113	0.0256	0.0183
Chips and fries	0.0128	0.0247	0.0089	0.0181	0.0148	0.0147	0.0139	0.0186	0.0260	0.0114	0.0165
Vegetables and others	0.0668	0.0301	0.0200	0.0194	0.0164	0.0259	0.0301	0.0129	0.0109	0.0135	0.0158
Poltury	0.0782	0.0060	0.0166	0.0135	0.0094	0.0279	0.0209	0.0105	0.0118	0.0329	0.0134
Others	0.0067	0.0042	0.0094	0.0040	0.0062	0.0095	0.0093	0.0061	0.0039	0.0033	0.0085
Yogurt	0.0188	0.0201	0.0162	0.0082	0.0112	0.0114	0.0091	0.0017	0.0046	0.0000	0.0081
Legumes, nuts and seeds	0.0097	0.0049	0.0086	0.0055	0.0120	0.0069	0.0093	0.0200	0.0044	0.0054	0.0077
Dressing, sauces and gravy	0.0108	0.0140	0.0014	0.0045	0.0127	0.0067	0.0103	0.0076	0.0042	0.0000	0.0072
Fish and seafood	0.0120	0.0100	0.0000	0.0044	0.0116	0.0057	0.0103	0.0085	0.0018	0.0150	0.0069
Added fats	0.0069	0.0077	0.0083	0.0050	0.0076	0.0065	0.0050	0.0048	0.0025	0.0038	0.0054
Eggs	0.0147	0.0118	0.0041	0.0062	0.0127	0.0106	0.0073	0.0136	0.0038	0.0690	0.0036
Creams	0.0041	0.0018	0.0203	0.0003	0.0011	0.0028	0.0009	0.0025	0.0010	0.0026	0.0015
Dark green vegetables	0.0468	0.0371	0.0047	0.0033	0.0292	0.0276	0.0277	0.0010	0.0009	0.0002	0.0014

All the values are the ration of the mean of each food intake to the total mean in a specific node; all values are in grams