LATENT CLASS MODELS AND LATENT TRANSITION MODELS FOR DIETARY PATTERN ANALYSIS

Daniela Taryn Sotres-Alvarez

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics, Gillings School of Global Public Health.

Chapel Hill
2009

Approved by:

Dr. Amy H. Herring

Dr. Diane Catellier

Dr. Lloyd Edwards

Dr. Gary Koch

Dr. Anna Maria Siega-Riz

ABSTRACT

DANIELA TARYN SOTRES-ALVAREZ: Latent Class Models and Latent Transition Models for Dietary Pattern Analysis

(Under the direction of Dr. Amy H. Herring)

Dietary patterns (DP) are used to study the effects of overall diet on health outcomes as opposed to the effects of individual nutrients or foods. DP are empirically derived mostly using factor and cluster analysis. Latent class models (LCM) have been shown empirically to be more appropriate to derive DP than cluster analysis, but they have not been compared yet to those derived by factor analysis. We derive DP using LCM and factor analysis on food-items, test how well the resulting classes are characterized by the factor scores, and compare subjects' direct classification from LCM versus two a posteriori classifications from factor scores: one possible classification using tertiles and a two-step classification using LCM on previously derived factor scores.

In order to study changes in dietary patterns over time, we propose using latent transition models to study change as characterized by the movement between discrete dietary patterns. Latent transition models directly classify subjects into mutually exclusive DP at each time point and allow predictors for class membership and for probabilities of changing classes over time. There are several challenges particular to DP analysis: a large (≥80) number of food-items, non-standard mixture distributions (continuous with a mass point at zero for non-consumption), and typical assumptions (conditional independence given the class and time point, time-invariant conditional responses, and invariant transition probabilities) may not be realistic. We compare performance, capabilities and flexibility between two software packages (Mplus and a user's derived procedure in SAS) that allow fitting latent transition models.

A key decision involved when deriving DP is whether or not to collapse the primary dietary data into a smaller number of items called food groups. Advantages for collapsing include dimension reduction and decreasing the number of non-consumers to reduce the mass-point at zero. However, not collapsing helps our understanding of which combinations of specific foods are consumed. Further, food-grouping may have an impact on the association between DP and health outcomes. We explore via a Monte Carlo simulation study whether food-grouping makes a difference when deriving DP using LCM. Methods are illustrated using data from the Pregnancy, Infection and Nutrition (PIN) Study.

A Jorge, por su amor y entrega, por querer compartir y construir *Ithacas*; a Camila, por despertar en mi las emociones más hermosas; a Jimena y Lucia, por su amor incondicional; a mis papas, por enseñarme a disfrutar la vida, la música clásica y las matemáticas.

ACKNOWLEDGEMENTS

I am truly thankful to my adviser, Dr. Amy Herring, for her outstanding mentoring. As a person and as a researcher she is, and will always continue to be, an inspiration to me. As an academic adviser, she guided and encouraged me when I was (or felt) lost, she challenged me to overcome my own boundaries, and she was always available to meet, discuss and read my dissertation's drafts. Similarly, she always cared about my personal life and was a great support throughout my pregnancy and Camila's first months. It has been a wonderful experience to work with her and an honor to have her as mentor and adviser.

Thanks also to my committee members: Dr. Anna Maria Siega-Riz, Dr. Gary Koch, Dr. Diane Catellier and Dr. Loyd Edwards for their valuable comments and support. In particular, I am very grateful to Anna Maria who mentored me in the nutritional epidemiology field. I am undoubtedly in debt to Dr. Koch for all his guidance and funding support. Thanks to Dr. Lisa LaVange for providing me an office space at the CSCC to work on my dissertation. I am also very thankful to Dr. Glen Martin and the UNC Dissertation Support Group for working with me on being more confident about my work and less of a perfectionist. Thanks to Emil J. Kang, Executive Director of the Carolina Performing Arts, for articulating the marvelous performances I experienced at Memorial Hall, and to the donors for making it possible. I want also to acknowledge the Consejo Nacional para la Ciencia y Tecnología (CONACYT) from México for funding my MS and DrPH degrees. Finally, thanks to my parents and to Jorge, for their love and support, especially through the difficult moments when I struggled due to lack of confidence and motivation. ¡Gracias!

CONTENTS

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xiii

CHAPTER 1 Literature Review 1

1.1 Dietary patterns analysis 1

1.1.1 Statistical methods to derive dietary patterns 3

1.1.1.1 Empirical methods 4

Principal Components Analysis 5

Factor Analysis 7

Cluster Analysis 10

Latent class analysis 12

1.1.1.2 Reproducibility 13

1.1.1.3 Software 13

1.1.2 Dietary patterns over time 13

1.1.3 Dietary patterns during pregnancy and postpartum 16

1.2 Generalized latent variable models 18

1.2.1 Classic latent variable models 19

1.2.1.1 Continuous latent variables: Structural Equation Model (SEM) ... 19

Extension to SEM: non-normal responses 21

Software 23

1.2.1.2 Categorical latent variables: Latent Class Model (LCM) 23

Latent transition models 26

Identifiability and estimation 29

Latent class prediction 32

Number of classes..33
Software..34
1.2.2 Integration of continuous and categorical latent variables: Structural Equation Mixture Modeling (SEMM)..35
1.2.3 Model selection and goodness of fit 37

REFERENCES..49

CHAPTER 2 Estimating Dietary Patterns: Comparing Factor Analysis and Latent Class Analysis..54

2.1 Abstract 54

2.2 Introduction...55

2.3 Methodological Decisions to Derive Dietary Patterns...55
2.3.1 Factor analysis...56
2.3.2 Latent class analysis 57
2.3.3 Number of dietary patterns...57

2.4 Analysis of Dietary Patterns for Women in Pin...58
2.4.1 Study sample and dietary intake assessment 58
2.4.2 Statistical methods..60

2.5 Results 61
2.5.1 Factor analysis...61
2.5.2 Latent class analysis 63
2.5.3 Comparison between factor scores and latent classes................................64

2.6 Discussion 65

REFERENCES..76

CHAPTER 3 Latent transition models to study change in dietary patterns over time 78

3.1 Abstract 78

3.2 Introduction......79

3.3 The latent transition model (LTM) 81

3.4 Analysis of the third cohort of the Pregnancy, Infection and Nutrition (PIN) Study ...83

3.4.1 Study sample, dietary assessment and food groups 83

3.4.2 Model selection......84

3.5 Results 85

3.5.1 Identification of dietary patterns 85

3.5.2 Transition probabilities 86

3.6 Discussion 87

REFERENCES......96

CHAPTER 4 Does food-grouping makes a difference when deriving dietary patterns using Latent Class Models? A Monte Carlo Simulation Study 98

4.1 Abstract 98

4.2 Introduction......99

4.3 Latent class model considered 101

4.4 Model selection......103

4.5 Monte Carlo Simulation Study 103

4.5.1 Data generation 103

4.5.2 Data Analysis......104

4.6 An application to the Pregnancy, Infection and Nutrition (PIN) Study.....................105

4.7 Discussion...106

REFERENCES..110

CHAPTER 5 Conclusions and Future Directions 111

APPENDIX A Supplementary Tables and Figures for Chapter 2.................................114

APPENDIX B Supplementary Tables and Figures for Chapter 3.................................127

LIST OF TABLES

TABLE 1.1 Software to estimate latent class models..41

TABLE 1.2 Articles that have studied dietary patterns over time 44

TABLE 1.3 Articles that have studied dietary patterns during pregnancy and postpartum ...47

TABLE 2.1 Selected exploratory and confirmatory factor loadings for 4-factor model, in the PIN Study...73

TABLE 2.2 Median nutrient dietary intake by latent class, in the PIN Study.........................74

TABLE 2.3 Odds ratios for covariates of 3-LCM on 98 food-items, in the PIN Study 75

TABLE 3.1 Food-group median consumption among consumers at twelve months postpartum, PIN Study..92

TABLE 3.2 Model selection for 3-class latent transition model, PIN Study...........................93

TABLE 3.3 Odds ratios and 95% CI for predictors of class membership at pregnancy, 3-class latent transition model, PIN Study..94

TABLE 3.4 Class prevalences and transition probabilities from 3-class latent transition model, PIN Study 95

TABLE 4.1 Percentage of times each latent class model is chosen..................................108

TABLE 4.2 Type I error and power for the Lo-Mendell-Rubin LRT test.............................108

TABLE 4.3 Cross-classification between 3-LCA on 98 food-items and 3-LCA on 26 food-groups 109

TABLE A1. Food item's distribution in the PIN Study..114

TABLE A2. Exploratory and confirmatory factor loadings for 4-factor model in the PIN Study 117

TABLE A3. Regression coefficients for 4-factor model for the PIN Study 120

TABLE A4. Partial Spearman correlations between factors and nutrients, adjusted by energy intake, in the PIN Study...121

TABLE B1. Food-items by food group, in the PIN Study...127

LIST OF FIGURES

FIGURE 1.1 Path diagram for a two-factor model 39
FIGURE 1.2 Path diagram for a latent class model (LCM)..39
FIGURE 1.3 Path diagram for a latent transition model 40
FIGURE 1.4 Path diagram for a factor mixture model...40

FIGURE 2.1 Simplified path diagram for confirmatory factor model of dietary patterns, in the PIN Study 69
FIGURE 2.2 Probabilities of consumption for selected food-items by latent class, in the PIN Study 70
FIGURE 2.3 Factor score means by latent class from 3-LCA on 98 ordinal food-items, in the PIN Study..71
FIGURE 2.4 Factor score means by latent class from 3-LCA on 4 factor scores, in the PIN Study...72

FIGURE 3.1 Path diagram for latent transition model, PIN Study 90
FIGURE 3.2 Dietary patterns' distribution by time and transition probabilities, PIN Study 91

FIGURE A1. Probabilities of consumption by latent class from LCA on 98 food-items, PIN Study...122

FIGURE B1. Probabilities of consumption by latent class from LTM on 29 food-groups, PIN Study 130

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
B-LRT	Bootstrap likelihood ratio test
BIC	Bayesian Information Criterion
CFA	Confirmatory factor analysis
CI	Confidence Interval
CLC	Classification likelihood criterion
DIC	Deviance Information Criterion
DP	Dietary patterns
EFA	Exploratory factor analysis
EM	Expectation-Maximization algorithm
FA	Factor Analysis
FFQ	Food frequency questionnaire
GLLAMM	Generalized Linear Latent and Mixed Models
HMM	Hidden Markov Model
LCA	Latent class analysis
LCM	Latent class model
LMR-LRT	Lo-Mendell-Rubin likelihood ratio test
LRT	Likelihood ratio test
LTM	Latent transition model
MCMC	Markov chain Monte Carlo
NHANES	National Health and Nutrition Examination Survey
PIN	Pregnancy, Infection and Nutrition Study
SEM	Structural equation model

CHAPTER 1

Literature Review

1.1 Dietary patterns analysis

In the last two decades, dietary patterns (DP) have increasingly been used to study the effects of overall diet on health outcomes (Newby and Tucker, 2004; Kant, 2004) because it is well-recognized that the single nutrient approach in nutritional epidemiology has conceptual and methodological limitations (Hu 2002; Newby and Tucker, 2004; Kant, 2004; Schulze and Hoffmann, 2006). First, people do not eat isolated nutrients but rather eat meals consisting of a variety of foods with complex combinations of nutrients that are likely to interact. Second, the single nutrient approach fails to account for interactions and/or high intercorrelations between food components or foods, and is unable to detect small effects from single nutrients. By contrast, dietary patterns account for cumulative and interactive effects (Schulze and Hoffmann, 2006), for effects of physical characteristics and unknown components (van Dam, 2005), and may be particularly suitable when traditional nutrient analyses have identified few dietary associations for the disease (e.g. breast cancer), when many dietary components are relevant to the health outcome, when interest is in the effect of overall diet, or to evaluate dietary guidelines (Hu, 2002).

However, dietary patterns do have limitations. The correlated measurement error in assessing foods may distort the definition of a dietary pattern, and DP do not provide clear answers about which elements of the dietary pattern are responsible for the observed effect (Schulze and Hoffmann, 2006). Another limitation is that dietary pattern analysis does not deduce the biological mechanisms of the effect of diet on health outcomes but, from a public health perspective this is not necessary to promote healthier diets (Kant, 2004).

Nutritional epidemiologists (Hu, 2002; Newby and Tucker, 2004; Schulze and Hoffmann, 2006) recognize that dietary pattern analysis is complementary to the traditional single nutrient approach to better understand the complexity of dietary intake in relation to health outcomes. Further, some agree (Newby and Tucker, 2004; Jacobs and Steffen, 2003) that dietary patterns can even help generate or refine new diet-disease hypotheses, thought others (Kant, 2004) do not.

Dietary patterns are derived for the most part because of the interest in examining diet as a multidimensional exposure for health outcomes. The most updated literature reviews on dietary patterns (Newby and Tucker, 2004; Kant, 2004; Schulze and Hoffmann, 2006) illustrate the variety of health outcomes that have been considered, including indicators of cardiovascular or coronary heart disease, anthropometric measures, many different cancers, all-cause mortality, bone mineral density, dental caries, and birth weight, among others. However, dietary patterns have also been derived to be studied in their own interest. For example, they have been derived as a *proxy* to dietary behavior, to evaluate adherence to dietary guidelines, to assess nutritional adequacy, and to study their predictors (e.g. genetic, demographic, lifestyle, environmental). In general, dietary patterns do differ in nutrition composition, are associated with personal characteristics (e.g. sex, age, and socioeconomic status) and other health behaviors (e.g. smoking, drinking, and exercise) (Newby and Tucker, 2004; Kant, 2004; Schulze and Hoffmann, 2006).

1.1.1 Statistical methods to derive dietary patterns

Because dietary patterns cannot be measured directly, statistical methods are used to derive them. The first studies to examine dietary patterns and health in the early 80s defined food eating patterns as “foods as they are actually consumed in various characteristic combinations” (Schwerin *et al*, 1981, 1982) and were derived empirically using factor analysis. Since then, most of the literature has used two different approaches to derive dietary patterns (also referred to as food or eating patterns). The first is a theoretical (hypothesis-oriented) approach (aka *a priori* approach) that uses prevailing hypotheses and guidance from current dietary recommendations to derive the dietary patterns. One example is the Diet Quality Index-Revised DQI-R[1] (Haines *et al*, 1999), which measures diet quality relative to the Dietary Guidelines for Americans, and focuses on four major aspects of a high-quality diet: macronutrient distribution, moderation, variety and proportionality. The second approach to define dietary patterns is an empirical approach (aka *a posteriori* approach) in which DP are derived from the data. The predominant methods to derive dietary patterns empirically are factor analysis (either principal components or common factor analysis) and cluster analysis (Newby and Tucker, 2004; Kant, 2004; Schulze and Hoffmann, 2006).

Both theoretical and empirical approaches have strengths and weaknesses. One weakness of the theoretical approach is that “the index/scores focus on selected aspects of the diet and do not consider the correlation structure of food and nutrients; consequently, they do not reflect the overall effect of diet in general but only the formal sum of not-adjusted single effects” (Hoffmann *et al*, 2004). Another disadvantage of index/scores is that they

[1] The Diet Quality Index (Revised) Score is on a 100 point scale. It is created from the addition of the following ten scores: Energy from fat score, Energy from saturated fat score, cholesterol score, DQI total grains score, DQI total fruits score, DQI total vegetables and soy score, % AI calcium score, % RDA iron score, Diet Variety (weighted) score, Moderation score.

reflect the degree to which a person's diet conforms to a dietary pattern that was defined *a priori,* whereas the empirically derived dietary patterns represent real-world dietary behavior. Hence, the latter can help the conceptual understanding of human dietary practice and provide guidance for setting priorities for nutrition intervention and education (Hu, 2002; van Dam, 2005). Some weaknesses of empirically derived dietary patterns include the complete ignorance of *prior* knowledge and that by nature they are population-specific.

In 2004, reduced rank regression was proposed (Hoffmann *et al*, 2004) as one way to combine *prior* information (any continuous variables that are affected by diet and are predictive for the disease) and the data from the study. This approach is limited to studies for which knowledge about important intermediate variables exists, and when the main interest is the diet and health outcome association, but not when the main interest is the study of dietary patterns *per se*. Another option to incorporate *prior* information is to impose structure and/or adding covariates in a confirmatory factor model or structural equations model (SEM).

1.1.1.1 Empirical methods

The most common methods to derive dietary patterns empirically are principal components analysis (PCA), factor analysis (FA) and cluster analysis. PCA and FA have great appeal in nutritional epidemiology as a way to handle multicollinearity between food components and to use the principal components or the factors to define the dietary patterns. On the other hand cluster analysis provides a classification of the subjects. Regardless of the statistical method used, there are several nutritional methodological issues involved in dietary pattern analysis. For example, whether or not to collapse the primary dietary data (which ranges from 25 to 250 food items depending on the dietary assessment tool used) into a smaller number of items (called food groups), how to group the data if collapse is done, quantification of the food items (weight, frequency or percent energy contribution, etc.), the number of patterns to extract, which patterns should be reported or analyzed, and how the patterns should be named (Newby and Tucker, 2004; Schulze and Hoffman, 2006).

Principal Components Analysis

Principal component analysis (PCA) is a multivariate reduction method that reduces a set of *p* correlated variables to a smaller set of *m* uncorrelated linear combinations of the original variables, which explain a large proportion of the total variance (Mardia *et al*, 1979). One disadvantage of the principal components is that they are not scale invariant, and variables with larger variances can dominate the results. Hence, observed variables are usually standardized first. Let $\mathbf{Y}_i \quad i = 1, 2, \mathrm{K}\ , n$ be a random sample of a p-dimensional vector of continuous random variables, and let it be standardized. The $r\text{-}th$ principal component is defined as the linear combination of the *p* observed variables given by

$$Z_{ir} = \sum_{j=1}^{p} \mathrm{a}_{rj} Y_{ij} = \mathbf{a}_r^T \mathbf{Y}_i \qquad r = 1, 2, \mathrm{K}\ , p$$

where $\mathbf{a}_r$ is the $r\text{-}th$ normalized eigenvector of the sample correlation matrix $\mathbf{R}$, and their corresponding eigenvalues are ordered as $l_1 \geq l_2 \geq \mathrm{K}\ \geq l_p \geq 0$. These linear combinations are optimal in the sense that Z_1 has the largest variance over all possible linear combinations of the observed variables, Z_2 has the next largest variance under the restriction that Z_1 and Z_2 are uncorrelated, and so on. In matrix notation, the *p* principal components are given by

$$\underset{p\times 1}{\mathbf{Z}_i} = \underset{p\times p}{\mathbf{A}^T}\underset{p\times 1}{\mathbf{Y}_i}$$

and their covariance matrix is $\mathbf{L} \equiv Diag\left\{l_1, l_2, \mathrm{K}\ , l_p\right\}$ by the spectral decomposition of the sample correlation matrix of $\mathbf{Y}_i$, $\mathbf{R}{=}\mathbf{A}^T\mathbf{LA}$. The matrix $\mathbf{AL}^{1/2}$ is called the *initial loading matrix,* and its components are the correlations between the principal components and the observed variables because their covariance is $\mathrm{a}_{jr} l_r$ from

$$\mathrm{cov}(\mathbf{Z},\mathbf{Y}) = E\left[\mathbf{A}^T\mathbf{Y}\mathbf{Y}^T\right] = \mathbf{A}^T E\left[\mathbf{Y}\mathbf{Y}^T\right] = \mathbf{A}^T\mathbf{R} = \mathbf{A}^T\mathbf{A}^T\mathbf{LA} = \mathbf{LA}$$

In other words, the principal components are linear transformations of the original variables, uncorrelated with one another and with decreasing variance. Further, the sum of the variances (*total variance*) of all the principal components is equal to the total variance of the original standardized variables, which is *p*. In practice the first *m* principal components are retained in order to reduce the dimension, and the proportion of the total variance explained is $\dfrac{1}{p}\sum_{j=1}^{m} l_j$. There is no single best way to select the number of principal components to retain and hence, reduce data dimension. Some options are retaining the first *m* components that accumulate certain percent of total variation, keeping those for which their eigenvalues are above the average, which is one (known as Kaiser's rule), or based on the Cattell's Scree plot (eigenvalues vs. number of components). However, regardless of the number of principal components retained the scores for those retained are the same, which is not the case in factor analysis. The DP literature most often uses Kaiser's rule, and the percent of explained variance has ranged from 15% to 93% (Newby and Tucker, 2004).

In some applications, such as in dietary pattern analysis, there is interest in interpreting the principal components. However, usually the initial loading matrix is difficult to interpret because the components are an average of all the variables. Hence, for interpretation purposes the principal components are transformed (rotated) so that the rotated components have high loadings on a small set of variables and are close to zero for the rest. Some examples of orthogonal transformations are Varimax and Quartimax, which attempt to achieve a simple structure of the columns and rows respectively of the initial loading matrix. Examples of non-orthogonal (oblique) transformations are Promax and Quartimin. With an orthogonal transformation the components are uncorrelated whereas with an oblique rotation they are correlated. After rotation, the total variance explained by the *m* components remains the same, but the variance of each component is more evenly distributed.

Factor Analysis

In contrast to PCA, factor analysis (FA) is a multivariate method which postulates a statistical model that attempts to explain the correlations between many observed variables by few underlying but unobservable (*latent*) variables called *factors* (Mardia *et al*, 1979; Bollen, 1989). The **m-factor model** is formulated in terms of the p-dimensional vector of continuous random variables $\mathbf{Y}_i$ for subject $i = 1, 2, \mathrm{K}\ , n$ as:

$$\mathbf{Y}_i = \mathbf{\nu} + \mathbf{\Lambda}\mathbf{\eta}_i + \mathbf{\varepsilon}_i$$

$$E\left[\mathbf{\varepsilon}_i\right] = \mathbf{0} \quad \mathrm{var}\left(\mathbf{\varepsilon}_i\right) \equiv \mathbf{\Theta} = Diag\left(\mathbf{\theta}\right) \qquad E\left[\mathbf{\eta}_i\right] = \mathbf{0} \qquad \mathrm{var}\left(\mathbf{\eta}_i\right) = \mathbf{I} \quad \mathrm{cov}\left(\mathbf{\varepsilon}_i, \mathbf{\eta}_i\right) = \mathbf{0}$$

where $\mathbf{\nu}_{p\times 1}, \mathbf{\Lambda}_{p\times m}$ are parameter matrices (the elements of $\mathbf{\Lambda}$ are called factor loadings), $\mathbf{\eta}_i$ is an m-dimensional vector of latent variables (factors), and $\mathbf{\varepsilon}_i$ is a p-dimensional vector of residuals. Generally, the outcomes are centered since the interest relies on the covariance structure and hence, intercepts, $\mathbf{\nu}$, are not estimated. The m-factor model can be equivalently expressed as

$$\mathbf{\Sigma} \equiv \mathrm{var}\left(\mathbf{Y}\right) = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Theta} \qquad (1)$$

which highlights that the sample correlation matrix can be decomposed into two sources: the variance of $\mathbf{Y}_i$ that is explained by the factors and the residual variance.

There are many methods to estimate the parameters $\left(\mathbf{\Lambda} \text{ and } \mathbf{\Theta}\right)$, but in DP analysis the *principal components method* is the most widely used (Newby and Tucker, 2004) because it does not require specifying the distribution of the observed variables. This method estimates $\mathbf{\Lambda}_{p\times m}$ with the first m columns of the initial loading matrix $\mathbf{L}_{p\times m}$ from principal components and the variances of the residuals by $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T - \mathbf{R}$ from equation (1). However, it is important to highlight that PCA is not the same as FA using the principal component method of estimation. The principal components are linear combinations of the observed variables and there are no underlying unmeasurable factors, whereas the factors scores are predicted values of the unobservable (latent) factors. Unlike principal components, factor analysis is scale invariant but similarly to PCA factor loadings are not unique and indeterminacy is usually resolved by making them satisfy a constraint. Also, note that PCA alone does not allow data dimension reduction unless a few principal components are retained (*i.e. ad hoc* two step approach), whereas FA does reduce the dimension directly. The other two most used methods of estimation are common factor[2] and maximum likelihood (ML) under multivariate normality assumption.

There are two general approaches to factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). In EFA the relationship between the observed and the latent factors is not specified in advance. In confirmatory factor analysis (CFA) the model is specified *a priori* and hence, some parameters of the m-factor model are restricted to certain values, typically zero. Because latent variables are unobservable their location and scale parameters are not identifiable (i.e. unique), and hence their parameters have to

[2] Common factor analysis is an *ad hoc* two step method of estimation. First, the diagonal of the correlation matrix is replaced by the estimated residuals, and using this correlation matrix (known as reduced correlation matrix) the m-factor model is estimated (Mardia *et al*, 1979).

be restricted. A common practice is to restrict their mean to zero, and their scale is either set to the scale of one of the observed variables (aka indicators) by restricting the factor loading to one (referred as anchoring) or set to standardized units by fixing its variance to one. The scale parameterizations yield equivalent models, but anchoring is preferred because it guarantees factor invariance (Skrondal and Rabe-Hesketh, 2004). One sufficient but not necessary rule for the m-factor model to be identified requires the following three conditions: at least three indicators per factor, each indicator loading in one and only one factor, and independent residuals $\left(i.e.\ \mathbf{\Theta} = Diag\left(\mathbf{\theta}\right)\right)$ (Bollen, 1989). If there is more than one indicator per factor and/or residuals are not independent, then other parameter restrictions are needed in order to identify the model. One example of an identified two-factor model, with seven continuous outcomes (centered) is represented in a path diagram[3] in **Figure 1.1.**

Most empirical dietary patterns are derived from exploratory factor analysis using the principal components method of estimation (Newby and Tucker, 2004; Kant, 2004) to provide a unique factor score solution, and the Varimax method of orthogonal rotation to facilitate interpretability. Orthogonal rotation also simplifies future analyses, such as avoiding collinearity when using factor scores as covariates in regression models or allowing analyzing them as independent outcomes.

In general, most applications of FA involve two steps: 1) identification of the factors and 2) computation of factors scores to use them in following analyses. In such analyses, factor scores are generally then treated as known. For instance, factors are used as predictors to study association between diet and health outcomes. In nutritional epidemiology two approaches have been used to compute (predict) the factor score for each dietary pattern. The first one combines the standardized food variables with weights that are

[3] By convention, in path diagrams circles represent latent variables, squares observed variables, straight one-headed arrows ‘causal’ relationships, curved two-headed arrows correlations, and small one-headed arrows random error.

proportionate to their factor loadings (best linear unbiased predictor or BLUP, aka regression method). This approach has been used by Newby *et al* (2006) and Weismayer *et al* (2006). The second approach known as simplified dietary pattern score combines with equal weight only those standardized variables that showed high factor loadings (e.g. ≥0.25) (Schulze *et al*, 2003). Some examples are Mishra *et al* (2006) and McNaughton *et al* (2007). Using CFA on the explored DP to derive the factor scores is similar to the simplified dietary pattern score in the sense that both approaches only add foods which score highly, because the former imposes the restriction of zero loadings to foods that scored low using EFA. However, DP scores from CFA weights the foods by their factor loadings whereas the simplified dietary pattern score does not.

Cluster Analysis

Cluster analysis is another multivariate method that is appealing to derive DP empirically because it classifies individuals in groups (unknown *a priori)* such that within groups they are similar (have similar diet), and groups are different from each other. In order to interpret the DP, it is necessary to compare dietary intake (e.g. nutrient intake) between clusters. Both agglomerative hierarchical and nonhierarchical (aka optimization method) methods have been used to derive DP (Newby and Tucker, 2004). Among the advantages of hierarchical methods are: no need to specify the number of clusters in advance, allows categorical and/or continuous variables because the methods operate on a $n \times n$ matrix of pairwise distances (similarities or dissimilarities) between subjects, the tree structure (best suited for biological and zoological applications), the dendrogram, and being a non-parametric method. However, some disadvantages are: subjects (or clusters) clustered together can’t be separated, sensitive to outliers, and not practical for very large datasets because time complexity varies as the square or cube of the sample size. In particular, in nutritional epidemiology the most common agglomerative hierarchical method used is Ward's minimum variance (Ward, 1963) (aka incremental sum of squares approach) which joins the two clusters that produce the least loss of cluster homogeneity (squared Euclidian distance from the observations to the corresponding cluster mean).

In contrast to hierarchical methods, nonhierarchical methods require fixing the number of clusters from start, and clusters do not form a hierarchical structure. These methods produce a partition of the subjects by optimizing certain criterion. Because all possible enumerations of partitions of *n* observations in *K* clusters is practically impossible there are different algorithms for an optimum non exhaustive search procedure. The k-means algorithm can be derived from considering a likelihood approach or as a mixture of density functions (Magidson and Vermunt, 2004). Cluster analysis is highly sensitive to starting values because there could be several local maxima so to be more confident of reaching the global maximum several sets of starting values should be used.

In summary, the cluster and factor solutions generate patterns that differ in food composition because they are statistically different procedures (Newby and Tucker, 2004). Factor analysis groups input variables (food intake) according to the degree to which they are correlated to each other, whereas cluster analysis groups subjects into mutually exclusive categories. There are no clear advantages between factor and cluster analyses to derive dietary patterns, and it has been a matter of preference, which statistical method to use. One disadvantage of FA is the overlap in factor scores which may explain inconsistent results when comparing factors across studies (Newby and Tucker, 2004). By contrast, cluster analysis provides a direct classification of subjects, and findings are easier to interpret because subjects belong to one cluster only.

Latent class analysis

Despite that latent variable models and SEM were recognized as useful methods to reflect complex relations between diet and disease at the international workshop on dietary patterns in 2000 (Hoffmann et al, 2002) these models have been rarely been used in nutritional epidemiology. Latent class models (LCM) share the same goal as cluster analysis: to classify subjects into classes (unknown *a priori*) such that within class they are similar (have similar diet), and classes are different from each other. LCM will be discussed in chapter 1.2, but here I briefly review the only three studies that have used them to study diet. The first study (Patterson *et al*, 2002) extended the traditional LCM to complex sample survey data and used dietary data as an application. In particular, using a 2-class LCM they estimated the proportion of the population that regularly consumed vegetables (class 1) by using four repeated measurements of an indicator variable for consuming at least one vegetable in the 24hrs recall. The other two studies (Padmadas *et al*, 2006; Fahey *et al*, 2007) used LCM specifically to empirically derive DP. Padmadas *et al* (2006) used the National Family Health Survey in India (90,180 women) to fit a 5-class traditional LCM using seven food groups each with four categories (frequency of intake). The article by Fahey *et al* (2007) used a generalized latent class model to derive the DP using the 2000-2001 National Diet and Nutrition Survey for British adults (766 men and 958 women). In particular, they derived the DP by gender from 25 food groups (20 measured on a continuous scale, and 5 binary indicators of food-consumption), two covariates predicting class membership (age and energy intake), and specified different types of covariance matrix (identity, and diagonal and compound symmetric by class). They illustrated that the use of LCM to derive DP is more flexible than the traditional methods used in the past and offers the possibility of studying more complex models, which may provide interesting insights into dietary patterning.

1.1.1.2 Reproducibility

Despite the increased interest in studying dietary patterns, there are few studies to date that have examined the reproducibility and validity of the statistical methods. With respect to reproducibility of dietary patterns, it has been distinguished between reproducibility within a population (stability over time) and reproducibility between populations. Although it is reasonable to expect that dietary patterns will differ between populations, Newby and Tucker (2004) argue that some dietary patterns are more likely to be identified in several populations (reproducibility between populations). This represents a challenge because comparisons among studies are not that simple because even though patterns might be called similar they might be very different in food and nutrient composition.

1.1.1.3 Software

Cluster analysis, exploratory and confirmatory FA for continuous symmetric data can be fitted using general purpose statistical software such as SAS, STATA, and SPSS. CFA for binary or categorical outcomes requires using specialized software such as Mplus or AMOS. Some LCM can be fitted using general purpose statistical software while others require specialized software (**Table 1.1**). Reduced rank regression can be fitted in SAS.

1.1.2 Dietary patterns over time

In the last five years, there has been an increased interest among nutritional epidemiologists to study empirical dietary patterns over time. The initial motivation was to study stability (reproducibility) of DP over time to test the assumption that dietary behavior was the same throughout the follow-up period. The main challenge is measuring change in variables that are not directly observed. Also, empirical DP are population and time-specific because they are data-driven and, therefore comparisons over time are not straightforward.

Literature studying dietary behavior over time, as measured by empirically derived DP, can be classified according to two aims (**Table 1.2**). The first aim is concerned with testing stability of DP; the researcher's hypothesis is that over time the same DP can be identified (i.e. same DP structure) and that DP scores (or subject's classification) are similar. For example, in a population of interest DP might be considered stable if there are always these three DP: 'Healthy', 'Western' and 'Southern', and subjects score similarly in the three DP every time. By contrast, the second aim is interested in within-subject change in DP over time, where researchers hypothesize the structure in the DP to be the same over time though subjects may score differently. Both aims require first evaluating if the structure of the dietary patterns is the same over time (i.e. same foods and same importance) and if so, measuring within-subject's change in DP. Obviously, the study of change in DP depends if the DP are treated as continuous (as in PCA and FA) or as categorical (as in cluster analysis and LCA). All these longitudinal studies have used FA to empirically derive the DP, except for one study (Greenwood *et al*, 2003) that used cluster analysis. Regardless of the aim, investigators using continuous DP followed three steps to study them over time: identify the DP, compute DP scores at each measurement, and compare DP scores over time.

In order to identify the DP investigators derived them separately for each time point using EFA and just by visual inspection decided whether the DP were the same (i.e. same number of factors and similar factor loadings over time). Some (Newby *et al*, 2006 (Vol. 3 and 10); Weismayer *et al*, 2006) used CFA (on DP previously identified by EFA) separately by time point and hence could not compare the factor loadings over time statistically. One exception is the study by Togo *et al* (2004) who used simultaneously both time points in a mean-structure FA while keeping the loadings on the factors equal. In other words, they assumed the DP loadings were the same over time without testing if the loadings were equal or not.

There has been more variety among researchers on how to compute DP scores at each time point. Some (Newby *et al*, 2006 (Vol. 3 and 10); Weismayer *et al*, 2006; Cuco *et al*, 2006) have derived them by using the confirmed factor loadings at each time point. Others (Mishra *et al*, 2006; McNaughton *et al*, 2007) have used for all time points the same simplified dietary pattern score equation to make them less time-specific. The simplified dietary pattern score (Schulze *et al*, 2003) sums the unweighted food items which load most highly (e.g. ≥0.25) on the pattern derived using data only from one assessment. The advantage of using the simplified dietary pattern score is making them less time-specific, but it assumes that all foods with high loadings have the same contribution to the pattern, which often is not true. The study by Togo *et al* (2004) used the factor loadings that, by construction, were restricted to be equal for both time points.

The comparison of DP scores over time depends on the study aim: stability of DP or within-subject change in DP. For example, long-term (>1 yr) stability of DP using FA has been studied in Swedish (Newby *et al*, J Nutr 136 (3), 2006; Weismayer *et al*, 2006) and British (Mishra *et al*, 2006) cohorts using from two to three assessments of dietary intake over 4 to 12 years of follow-up. To evaluate if DP were stable over time they used Spearman correlations between DP scores over time (Newby *et al*, 2006 (Vol. 3); Weismayer *et al*, 2006), and agreement (using weighted Kappa statistic) between quantiles of DP scores (Mishra *et al*, 2006). One study (Cuco *et al*, 2006) used congruence coefficients to assess stability of DP. In contrast, to estimate the within-subject change in DP over time some investigators have used the difference in DP scores between time points (Togo *et al*, 2004; Newby *et al*, 2006 (Vol. 10)), and others have categorized the DP scores in quintiles and classified participants according to change in quintiles (Schulze *et al*, 2005).

On the other hand, even after finding the same DP over time and DP scores highly correlated, DP could still be internally unstable (Weismayer *et al*, 2006). For instance, within each DP there could be differences in the food-items over time (e.g. correlations, means and SD). For example, using CFA they "tested the significance of changes in the covariance matrix between baseline and follow-up" and found the alcohol pattern was internally unstable after 6 years of follow up even though the alcohol pattern scores had a Spearman correlation of 0.7 between the two measurements.

There is only one study (Greenwood *et al*, 2003) that looked at the stability of DP derived by cluster analysis. They performed cluster analysis separately at baseline and 5 years later, and then used the Kappa statistic to test agreement between clusters (DP). Overall half the women maintained the same DP, and some patterns were more stable than others (κ=0.5 suggesting moderate stability).

1.1.3 Dietary patterns during pregnancy and postpartum

To date there are eight published papers that have studied dietary patterns during pregnancy from which two (Cuco *et al* 2006; Northstone and Emmet, 2007) have studied them longitudinally (**Table 1.3**). For brevity, details of the populations, dietary assessment, week of gestation, DP's labels are summarized in Table 1.3 but not discussed. Except for one study (Crozier *et al*, 2008) that evaluated the impact of the dietary assessment method (FFQ vs. 4-day food diary) on the derived dietary patterns (i.e. DP's validity), the rest were interested on identifying DP and examining their association mainly with nutrient intakes (absolute intake and/or energy-adjusted) and socio-demographic and lifestyle factors, and one (Knudsen *et al*, 2007) with fetal growth. All studies collapsed the food-items into food-groups (range 21-52), used PCA on the food-groups' correlation matrix to identify the DP, and Varimax transformation to simplify their interpretation. The associations were mainly studied with continuous factor scores (DP) using Pearson's correlations and linear regressions, except for two studies (Northstone *et al*, 2007; Knudsen *et al* (2007) that categorized each DP by its quintiles. Northstone *et al* (2007) used the quintiles of each DP score separately to assess for non-linearity, whereas Knudsen *et al* (2007) classified the women into mutually exclusive classes using the two DP jointly (by cross-tabulating the DP's quintiles) to study the effect of the classes on the outcomes.

Both longitudinal studies (Cuco *et al*, 2006; Northstone and Emmett, 2007) assessed the stability and/or change of DP over time, and the former also studied their associations with predictors and other health behaviors. Cuco *et al* (2006) used 7-d dietary records on 80 women to identify the DP at each of the six timepoints (pre-pregnancy, four assessments during pregnancy, and one at 6 months postpartum) using EFA with PC method of estimation on 21 standardized food-groups (g/day). They identified two DP ('Vegetables and meat' and 'Sweetened beverages and sugars') at all timepoints except at postpartum where only the 'Sweetened beverages and sugars' was identified. They concluded that DP in their sample did not vary over time because the factorial structures of the two DP were similar and there were no mean differences over time in the defining foods of the DP. However, due to the small sample size the mean change was not significantly different from zero most likely due to lack of power. The study by Northstone and Emmett (2007) was methodologically oriented and examined the stability of DP from pregnancy to 4yr postpartum with two different forms of calculating the DP scores at postpartum: 1) using the postpartum factor loadings and 2) using the pregnancy factor loadings. To assess dietary intake they used a 44-item FFQ at pregnancy which was slightly modified into a 52-item FFQ at 47 months postpartum. At pregnancy, they derived five DP ('Health conscious', 'Traditional', 'Processed', 'Confectionery', and 'Vegetarian') using PCA with Varimax rotation for two randomly split samples (n=8,935 women in total) to assess repeatability. At postpartum, the ‘Traditional’ DP was not identified so they did not include it in any further analysis. In order to assess stability over time of each of the four common DP they did paired t-tests, Bland-Altman limits of agreement (Bland and Altman, 1986), and weighted Kappa-statistics for quintiles of DP scores. They found differences in the means and measures of agreement between the two different forms to calculate the DP scores at postpartum, and concluded that for their data it was inappropriate to apply the pregnancy factor loadings to calculate the postpartum DP scores primarily due to differences in FFQ between the two timepoints (8 additional food-items).

1.2 Generalized latent variable models

Depending on the discipline and the context, a latent variable is defined differently, but they all mean a variable that is not observed. More formally, Skrondal and Rabe-Hesketh (2004) define a latent variable as “a random variable whose realizations are hidden for us”. The presence of latent variables is commonly recognized in the social sciences because of the difficulty in measuring key variables of theoretical or substantive interest (Clogg, 1992). For example, the hypothetical construct (concept) intelligence is a latent variable because the intelligence of a person cannot be observed directly; instead an aspect of intelligence is measured in terms of a number of items using an intelligence test. Similarly, the eating behavior ‘healthy eating’ is a latent variable because it is a concept that is not directly observed but can be measured with a dietary intake instrument and empirically derived using statistical methods such as exploratory factor analysis (EFA). Other examples of latent variables in nutritional epidemiology are obesity and physical activity.

However, latent variables are also present in statistics. Some have always been recognized as such, like factors in factor analysis, but others are not usually presented as latent variables, like random effects in linear mixed models. Continuous latent variables (as the ones previously exemplified) are called factors but latent variables can also be categorical, and are referred as latent class variables. One example of a latent class variable is ‘stage of change’ used in health behavior where five classes (precontemplation, contemplation, preparation, action, and maintenance) are used to assess where a subject is located in the process of a specific behavioral change. Another example is dietary pattern where ‘high-fat’ and ‘low-fat’ classes are used to classify dietary behavior among a specific population.

1.2.1 Classic latent variable models

1.2.1.1 Continuous latent variables: Structural Equation Model (SEM)

In the social sciences, there has been a long tradition in modeling the associations between latent and observed variables using structural equations model (SEMs). In mainstream statistics these models are also known as latent variable models or covariance structure models because the fundamental hypothesis is that the covariance matrix of the observed variables is a function of a set of unknown parameters$\left(i.e.\ H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}\left(\boldsymbol{\theta}\right)\right)$, although the mean can also be modeled. SEMs allow us, through a system of linear equations, to model jointly the *measurement model* of the observed data conditional on the latent variables, and the latent variable model (aka *structural model*) which summarizes the relationships between latent variables (Bollen, 1989). Originally, SEMs assumed the population was homogeneous (i.e. a single covariance matrix could be used to summarize the associations among variables), and only allowed observed variables (aka indicators) to be continuous. However, conventional SEM was extended to multiple-groups SEM (Joreskog, 1971), and to allow the indicators to have other scales (Muthén, 2002). A recent review of SEM literature with applications to epidemiology for a statistical audience is given by Sánchez *et al* (2005). In the Gaussian linear SEM and using the notation and parameterization by Muthén (2002), the measurement model for the *i*-th subject $i = 1, 2, \ldots, n$ is defined in terms of the p-dimensional vector of continuous outcome random variables $\mathbf{Y}_i$ and is given by

$$\mathbf{Y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \mathbf{K}\mathbf{x}_i + \boldsymbol{\varepsilon}_i \qquad (2)$$

where $\boldsymbol{\nu}_{p\times 1}, \boldsymbol{\Lambda}_{p\times m}$ and $\mathbf{K}_{p\times q}$ are parameter matrices, $\boldsymbol{\eta}_i$ is an m-dimensional vector of latent variables, $\mathbf{x}_i$ is a q-dimensional vector of covariates, and $\boldsymbol{\varepsilon}_i$ is a p-dimensional vector of residuals with multivariate normal distribution with mean zero-vector and covariance matrix $\boldsymbol{\Theta}_{p\times p}$. The continuous outcome is typically a multivariate outcome but it can also be repeated measures or a combination. The structural model defines the linear relationship among latent variables and covariates

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i \qquad (3)$$

where $\boldsymbol{\alpha}_{m\times 1}, \mathbf{B}_{m\times m}$ and $\boldsymbol{\Gamma}_{m\times q}$ are parameter matrices, and $\boldsymbol{\zeta}_i$ is an m-dimensional vector of residuals with multivariate normal distribution with mean zero-vector and covariance matrix $\boldsymbol{\Psi}_{m\times m}$ that is assumed to be independent of $\boldsymbol{\varepsilon}_i$. The parameter matrix $\mathbf{B}_{m\times m}$ has zeros in the diagonal because latent variables can only be influenced by other latent variables and not themselves, and $\mathbf{I} - \mathbf{B}$ is assumed to be nonsingular. This parameterization of the model is equivalent to the well known LISREL model (Joreskog and Sorbom, 1989). Also, note that the m-factor model (section 1.1.1) is a special case of SEM.

Substituting the structural part of the model in the measurement part gives

$$\mathbf{Y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\left(\mathbf{I} - \mathbf{B}\right)^{-1}\left[\boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i\right] + \mathbf{K}\mathbf{x}_i + \boldsymbol{\varepsilon}_i$$

from which it is immediate that $\mathbf{Y}_i \,|\, \mathbf{x}_i$ has also a normal distribution because the residuals $\boldsymbol{\zeta}_i$ and $\boldsymbol{\varepsilon}_i$ are assumed normal. The full log-likelihood is given by

$$\mathrm{l}_{FULL}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{x}\right) = -\tfrac{n(p+q)}{2}\log\left(2\pi\right) - \tfrac{n}{2}\log\left|\boldsymbol{\Sigma}^{*}\right| - \tfrac{n}{2}tr\left[\boldsymbol{\Sigma}^{*-1}\left(\mathbf{S}+\left(\bar{\mathbf{V}}-\boldsymbol{\mu}^{*}\right)\left(\bar{\mathbf{V}}-\boldsymbol{\mu}^{*}\right)^{T}\right)\right]$$

where $\boldsymbol{\theta} = vech\left\{\mathbf{v},\boldsymbol{\Lambda},\mathbf{K},\boldsymbol{\alpha},\mathbf{B},\boldsymbol{\Gamma},\boldsymbol{\Theta},\boldsymbol{\Psi}\right\}$, $\mathbf{V}_i^T = (\mathbf{Y}_i^T, \mathbf{x}_i^T)$, $\boldsymbol{\mu}^{*}\left(\boldsymbol{\theta}\right) \equiv E\left[\mathbf{V}_i\right]$, $\boldsymbol{\Sigma}^{*}\left(\boldsymbol{\theta}\right) \equiv \mathrm{var}\left[\mathbf{V}_i\right]$ and $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{V}_i - \bar{\mathbf{V}})(\mathbf{V}_i - \bar{\mathbf{V}})^T$ is the sample covariance matrix. Maximizing the conditional (on the covariates) log-likelihood over $\boldsymbol{\theta}$ is equivalent to maximizing the full log-likelihood $\mathrm{l}_{FULL}\left(\boldsymbol{\theta};\mathbf{Y},\mathbf{x}\right)$ because the marginal distribution of the covariates is constant with respect to $\boldsymbol{\theta}$ and can be factorized out. In the social sciences literature and SEM software instead of maximizing the likelihood they minimize the fitting function $F_{ML}(\boldsymbol{\theta})$ which is mathematically equivalent. The fitting function is a real valued function that measures the discrepancy between the sample moment structure and the one implied by the model, and it is zero when there is perfect fit. In fact, the fitting function is the deviance.

$$F_{ML}(\boldsymbol{\theta}) = \log|\boldsymbol{\Sigma}^{*}| + tr\left[\boldsymbol{\Sigma}^{*-1}\left(\mathbf{S}+\left(\bar{\mathbf{V}}-\boldsymbol{\mu}^{*}\right)\left(\bar{\mathbf{V}}-\boldsymbol{\mu}^{*}\right)^{T}\right)\right] - \log|\mathbf{S}| - (p+q) = \frac{2}{n}\left\{\mathrm{l}_{H_1} - \mathrm{l}_{H_0}\right\}$$

Extension to SEM: non-normal responses

In SEM, non-normal responses can be modeled using a generalized linear model approach or a latent response approach. The *latent response formulation* treats the observed response variable Y as a partial observation of the continuous latent response variable Y^{*} using threshold parameters. For example, an observed three-category ordinal response is defined using two thresholds τ_1 and τ_2 as

$$Y = \begin{cases} 0 & \text{if } Y^* \leq \tau_1 \\ 1 & \text{if } \tau_1 < Y^* \leq \tau_2 \\ 2 & \text{if } Y^* > \tau_2 \end{cases}$$

and $\mathbf{Y}_i$ is replaced by $\mathbf{Y}_i^*$ in the measurement model given in (2) with the assumption of conditional normality for $\mathbf{Y}_i^*$ given $\mathbf{x}_i$. For many response types the generalized linear model and the one using the latent response formulation specify equivalent models. However, the latent response formulation allows not only responses with different scales (binary, ordinal, nominal, count, censored, semicontinuous and continuous), but to model responses with different scales simultaneously. Further, conventional generalized linear mixed effects models can be formulated in the SEM framework, and then extended. For example the linear mixed effects model for longitudinal data is the latent curve model[4] (Bollen and Curran, 2006) in the SEM framework. However, some extensions that the latent curve model allows are: to regress random coefficients on each other, to estimate factor loadings (i.e. time scores) for models that are not linear in time, and to be combined with Markov models as in the autoregressive latent trajectory (ALT) model (Bollen and Curran, 2004).

Another example of an extended model in the SEM framework is the two-part growth mixture model (Muthén, 2001) inspired from the two-part random-effects model for nonnegative continuous longitudinal data with excess of zeros (Olsen and Schafer, 2001; Berk and Lachenbruch, 2002) and a zero class in a 2-class mixture model (Carlin *et al*, 2001). The excess of zeros can be due to structural zeros (“true zeros”) yielding a *semicontinuous variable,* and/or due to left-censored values (i.e. unobserved values from the continuous distribution). Two-part models estimate, separately or jointly, two models: one for the probability of nonzero values and one conditional model for the nonzero

[4] In the SEM literature the random effects are called growth factors.

continuous values. Usually the continuous part is skewed so it is previously transformed for the distribution to be approximately normal. Both Olsen and Schafer (2001) and Berk and Lachenbruch (2002) estimated two generalized linear mixed models (logistic and conditional mean log-response models) jointly by having the random intercepts correlated, but the latter integrated the random intercepts to obtain a marginalized model. Another difference is that the Olsen-Schafer model assumed zero values were only true zeros, whereas the Berk-Lachenbruch model distinguished true zeros from censored zeros by adding a parameter for the probability of a left-censored value in the conditional mean model. The Berk-Lachenbruch model could be useful to study food intake because often the intake's distribution is a nonstandard mixture, a combination of a skewed continuous distribution and a one point mass at zero. These zeros can either be truly non-consumers or very-low-consumers who did not consume the food-item during the reference period.

Software

Gaussian linear SEM can be estimated using specialized commands in general purpose statistical software such as SAS, STATA and R. SEM extensions can be estimated using specialized SEM software like Mplus (Muthén and Muthén , 1998), LISREL (Jöreskog and Sörbom, 1989), and AMOS (Arbuckle, 2006).

1.2.1.2 Categorical latent variables: Latent Class Model (LCM)

Models involving only categorical latent variables were developed separately from SEMs, and have been used extensively and for a long time in the social sciences. These models are an application of what in statistics is known as finite mixture models (McLachlan and Peel, 2000). The mixture density $f\left(\mathbf{y}_i\right)$ can be written as:

$$f\left(\mathbf{y}_i\right) = \sum_{k=1}^{K} \pi_{ik} f_k\left(\mathbf{y}_i \mid c_{ik} = 1; \boldsymbol{\theta}_k\right) \qquad (4)$$

where $\mathbf{c}_i$ is a K-dimensional class-label vector where the *k*-th element c_{ik} is defined to be one or zero if the *i-th* subject belongs or not to the *k-th* component density, $\pi_{ik} \equiv \Pr\left(c_{ik} = 1\right)$ is the mixing proportion of the *k-th* component, and the mixing proportions add to one. The component densities are usually specified to belong to the same parametric family, and often assumed normal. Nowadays, a latent class model (LCM) is a generic term for models for which the outcome is assumed to be sampled from (4) and hence, subjects are assumed to belong to one of K mutually exclusive classes but for which class membership is unknown. The original and traditional latent class model (Lazarsfeld, 1950) uses cross-sectional data where all variables -latent and observed- are categorical, and assumes that the observed variables are conditionally independent given class-membership. The traditional LCM has been extended in cross-sectional designs to include covariates to model class probabilities and/or conditional response probabilities, allow continuous outcomes (latent profile analysis, LPA), and relax the conditional independence assumption (Skrondal and Rabe-Hesketh, 2004). **Figure 1.2** shows the path diagram for a generalized LCM for dietary pattern analysis on 25 food groups (20 measured on a continuous scale, and 5 binary indicators) and two covariates (age and energy intake) predicting class membership (Fahey *et al*, 2007).

In particular, the measurement part of the LCM with categorical outcomes and covariates, assuming conditional independence, is specified as a finite mixture of conditional response probabilities given that the *i*-th subject belongs to class $k = 1, 2, \mathrm{K}\ , K$

$$
\begin{aligned}
\Pr\left(\mathbf{U}_i = \mathbf{u}_i \middle| \mathbf{X}_i = \mathbf{x}_i\right) &= \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i) \Pr\left(\mathbf{U}_i = \mathbf{u}_i \mid c_{ik} = 1\right) \\
&= \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i) \prod_{j=1}^{p} \Pr\left[U_{ij} = u_{ij} \mid c_{ik} = 1\right] \qquad (5) \\
&= \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i) \prod_{j=1}^{p} \rho_{ij|k}
\end{aligned}
$$

where $\mathbf{U}_i$ is a p-dimensional vector of categorical random variables, $\pi_{ik} \equiv \Pr\left(c_{ik} = 1 \mid \mathbf{x}_i\right)$ is the probability to belong to class k given the covariates $\mathbf{x}_i$, $\sum_{k=1}^{K} \pi_{ik}\left(\mathbf{x}_i\right) = 1$, and $\rho_{ij|k} \equiv \Pr\left[U_{ij} = u_{ij} \mid c_{ik} = 1\right]$ is the conditional *j-th* response probability given class k. The class membership $\underset{K\times 1}{\mathbf{c}_i}$ is distributed according to a multinomial distribution, and it is modeled with a baseline-category logit model for nominal response (McCullagh and Nelder, 1989) with the particularity that $\mathbf{c}_i$ is not observed. In other words, the structural part of the LCM is given by

$$
\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\pi}_i\right)} = \underset{(K-1)\times 1}{\boldsymbol{\alpha}} + \underset{(K-1)\times q \times 1}{\boldsymbol{\Gamma}\mathbf{x}_i} \quad (6)
$$

where $\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\pi}_i\right)} \equiv \left(\log\left(\frac{\pi_{i1}}{\pi_{iK}}\right), \log\left(\frac{\pi_{i2}}{\pi_{iK}}\right), \mathrm{K}\ , \log\left(\frac{\pi_{i,K-1}}{\pi_{iK}}\right)\right)^T$, and class K is the reference class.

Note that the LCM with categorical outcomes estimates two sets of parameters: the regression coefficients predicting class membership and the conditional probabilities of the observed responses given the class. When the model does not include covariates then the latent class probabilities are estimated directly.

With respect to longitudinal data, there have been several extensions to LCM. One extension is the group-based trajectory model (Nagin, 2005) that approximates the heterogeneity of trajectories of an outcome over time assuming there are a discrete number of classes that differentiate the trajectories. In contrast to the parametric approach of multilevel (Laird and Ware, 1982) and latent curve models (growth model in the SEM framework) (Bollen and Curran, 2006) where the between-subject heterogeneity is modeled assuming a multivariate normal distribution for the random effects, the group-based trajectory model takes a semi-parametric approach using groups to approximate this distribution. This analysis is also referred as latent class growth analysis (LCGA) (Muthén, 2002).

Another extension of the LCM to longitudinal data is the latent transition model (LTM) (Collins and Wugalter, 1992) where there are multiple categorical indicators of the latent class variable repeatedly measured over T equally spaced timepoints, and the main interest is to model transition between the latent classes. Note that the LTM assumes time is a discrete process whereas the group-based trajectory model assumes time is continuous. In addition, the LTM involves several categorical latent variables (one for each time point) whereas the group-based trajectory model involves only one latent class variable, the one that classifies the trajectories. On the other hand, note that whereas in transition models (Diggle *et al*, 2005) the conditional distribution of each response is modeled explicitly as a function of the previous responses and covariates, in latent transition models the latent class is the one modeled explicitly as a function of the previous latent classes and covariates (**Figure 1.3**).

Latent transition models

Latent transition models estimate three sets of parameters: 1) regression coefficients predicting class membership, 2) conditional probabilities of the observed responses given the latent class, and 3) transition probabilities of one latent class to another. Let the vector $\underset{pT\times 1}{\mathbf{U}_i} = \left(\underset{p\times 1}{\mathbf{U}_{i1}}, \underset{p\times 1}{\mathbf{U}_{i2}}, \ldots, \underset{p\times 1}{\mathbf{U}_{iT}} \right)^T$ represent the *i-th* subject's outcomes to the *p* categorical variables for all timepoints $t = 1, 2, \ldots, T$, $\mathbf{c}_{it}$ a K-dimensional class-label vector at time *t*, and $k_t = 1, 2, \ldots, K$ the latent class at time t. In contrast to LCM, the *T* latent classes $(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T)$ are not assumed independent, but similarly to LCM the *p* responses at time point *t* $(U_{t1}, U_{t2}, \ldots, U_{tp})$ are typically assumed conditionally independent given class membership.

The measurement part of the LTM with covariates is given by

$$\Pr\left[\mathbf{U}_i = \mathbf{u}_i \mid \mathbf{X}_i = \mathbf{x}_i\right] = \sum_{t=1}^{T} \sum_{k_t=1}^{K} \overbrace{\pi_{i1k_1}}^{\substack{\text{latent class}\\ \text{probability}\\ \text{at time 1}}} \overbrace{\left(\prod_{t=2}^{T} \tau_{itk_t|k_{t-1}}\right)}^{\substack{\text{transition}\\ \text{probabilities}}} \overbrace{\left(\prod_{t=1}^{T} \prod_{j=1}^{p} \rho_{itj|k_t}\right)}^{\substack{\text{conditional}\\ \text{response probabilities}}} \qquad (7)$$

where

$$\begin{aligned}
\pi_{i1k_1} &\equiv \Pr\left[c_{i1k_1} = 1 \mid \mathbf{X}_i = \mathbf{x}_i\right] \\
\tau_{itk_t|k_{t-1}} &\equiv \Pr\left[c_{itk_t} = 1 \middle| c_{i,t-1,k_{t-1}} = 1,\ \mathbf{X}_i = \mathbf{x}_i\right] \\
\rho_{itj|k_t} &\equiv \Pr\left[U_{itj} = u_{itj} \mid c_{itk_t} = 1\right]
\end{aligned}$$

The class membership at the first time point, $\mathbf{c}_1$, is modeled with a baseline-category logit model for nominal response with the particularity that $\mathbf{c}_1$ is not observed. Similarly, transition probabilities, $\tau_{itk_t|k_{t-1}}$ $t = 2, \ldots, T$, are modeled using a baseline-category logit model for nominal response. Choosing arbitrarily class *K* as the reference, the structural part of the LTM is given by

$$\begin{aligned}
\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\pi}_{i1}\right)} &= \underset{(K-1)\times 1}{\boldsymbol{\alpha}_1} + \underset{(K-1)\times q \times 1}{\boldsymbol{\Gamma}_1 \mathbf{x}_i} \\
\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\tau}_{it|k_{t-1}}\right)} &= \underset{(K-1)\times 1}{\boldsymbol{\alpha}_t} + \underset{(K-1)\times q \times 1}{\boldsymbol{\Gamma}_t \mathbf{x}_i} \quad t = 2, \ldots, T
\end{aligned} \qquad (8)$$

where

$$\underset{K\times 1}{\boldsymbol{\pi}_{it}} \equiv \left(\pi_{it1}, \pi_{it2}, \ldots, \pi_{itK}\right)^T \quad t = 1, 2, \ldots, T$$

$$\underset{K\times 1}{\boldsymbol{\tau}_{it|k_{t-1}}} \equiv \left(\tau_{it1|k_{t-1}}, \tau_{it2|k_{t-1}}, \ldots, \tau_{itK|k_{t-1}}\right)^T \quad t = 2, \ldots, T$$

$$\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\pi}_{i1}\right)} \equiv \left(\log\left(\tfrac{\pi_{i11}}{\pi_{i1K}}\right), \log\left(\tfrac{\pi_{i12}}{\pi_{i1K}}\right), \ldots, \log\left(\tfrac{\pi_{i1,K-1}}{\pi_{i1K}}\right)\right)^T$$

$$\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\tau}_{it|k_{t-1}}\right)} \equiv \left(\log\left(\tfrac{\tau_{it1|k_{t-1}}}{\tau_{itK|k_{t-1}}}\right), \log\left(\tfrac{\tau_{it2|k_{t-1}}}{\tau_{itK|k_{t-1}}}\right), \ldots, \log\left(\tfrac{\tau_{it,K-1|k_{t-1}}}{\tau_{itK|k_{t-1}}}\right)\right)^T \quad t = 2, \ldots, T$$

A common practice is to constraint the conditional response probabilities to be time-invariant $\left(i.e.\ \rho_{itj|k_t} = \rho_{ij|k_t} \quad t = 1, 2, \ldots, T\right)$ because it simplifies interpretation of the transition probabilities (Reboussin *et al*, 1999), otherwise the characterization of the classes would change over time.

The LTM given in equation (7) without covariates is a special case of the multiple indicator hidden (latent) Markov model (HMM) (McLachlan and Peel, 2002) which is an extension of a finite mixture model to allow dependent data. Although most often LTM and HMM use discrete outcomes and first-order transition probabilities, other scales and higher-order models can be accommodated. One difference between LTM and HMM is that the former is used when there are few time points whereas HMM can handle many time points (e.g. speech recognition applications). Another difference is that LTM allows explaining individual differences in the class and transition probabilities by using time-invariant covariates. By contrast, the latent mixed Markov model (Langeheine and Van de Pol, 2000) relaxes the homogeneity assumption (i.e. same transition probabilities for all subjects) by using a latent variable that unmixes the observed distribution into S Markov chains. One example of this type of models is the mover-stayer LTM which is defined by two $\left(S = 2\right)$ hidden Markov chains: one latent mover chain and one latent stayer chain.

Identifiability and estimation

The LCM and LTM are not identified unless some constraints are imposed over the parameters. The LCM is identified when the number of distinct response patterns is larger than the number of free parameters (Reboussin *et al*, 1998). One way the LTM with three timepoints is identified is by restricting the transition probabilities to be the same over time $\left(i.e.\;\; \tau_{itk_t|k_{t-1}} = \tau_{ik_t|k_{t-1}}, \quad t = 2,3\right)$.

Maximum likelihood (ML) estimation for finite mixture models is straightforward using the **Expectation-Maximization (EM)** algorithm (Dempster *et al*, 1977). The observed random sample $\mathbf{y}_i \,|\, \mathbf{x}_i \quad i = 1, 2, \ldots, n$ from the mixture distribution in (4) is viewed as incomplete because the class-label vectors $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$ are not observed. This natural incomplete-data structure makes the EM algorithm an attractive approach for maximizing the <u>observed-data log-likelihood</u>:

$$l_{obs}\left(\boldsymbol{\Psi};\mathbf{y},\mathbf{x}\right) \equiv \log L_{obs}\left(\boldsymbol{\Psi};\mathbf{y},\mathbf{x}\right) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_{ik}\left(\boldsymbol{\alpha},\boldsymbol{\Gamma};\mathbf{x}\right) f\left(\mathbf{y}_i \,|\, c_{ik} = 1;\boldsymbol{\theta}_k\right)\right\}$$

where $\mathbf{y},\mathbf{x}$ and $\boldsymbol{\Psi}$ are stacked vectors of outcomes, covariates, and parameters respectively. The <u>complete-data log-likelihood</u> is given by

$$l_c\left(\boldsymbol{\Psi};\mathbf{y},\mathbf{x}\right) \equiv \log L_c\left(\boldsymbol{\Psi};\mathbf{y},\mathbf{x}\right) = \sum_{k=1}^{K}\sum_{i=1}^{n} c_{ik}\left\{\log\left(\pi_{ik}\left(\boldsymbol{\alpha},\boldsymbol{\Gamma};\mathbf{x}\right)\right) + \log\left(f\left(\mathbf{y}_i \,|\, c_{ik} = 1;\boldsymbol{\theta}_k\right)\right)\right\}$$

and the conditional expectation (over the latent variables) of $\log L_c$ given the observed-data at the l^{th} iteration is

$$Q\left(\boldsymbol{\Psi};\mathbf{y},\mathbf{x},\boldsymbol{\Psi}^{(l)}\right) \equiv \sum_{k=1}^{K}\sum_{i=1}^{n} \Pr\left[c_{ik} = 1 \,|\, \mathbf{y}_i;\boldsymbol{\Psi}^{(l)}\right]\left\{\log\left(\pi_{ik}\left(\boldsymbol{\alpha},\boldsymbol{\Gamma};\mathbf{x}\right)\right) + \log\left(f\left(\mathbf{y}_i \,\middle|\, c_{ik} = 1;\boldsymbol{\theta}_k\right)\right)\right\} \qquad (9)$$

For independent data, the effect of the E-step at the l^{th} iteration is to update the *posterior* probabilities of class membership, $p_{ik}{}^{(l)} \equiv \Pr\left[c_{ik} = 1 \,|\, \mathbf{y}_i;\boldsymbol{\Psi}^{(l)}\right]$ (McLachlan and

Peel, 2000). From equation (9) it can be seen that during the M-step the updated estimates for the class probabilities can be estimated independently from the updated estimates of the component densities' parameters. Specifically for the LCM defined by equations (5) and (6) where the outcomes are nominal, the M-step involves two multinomial logistic regression optimizations: one to estimate the regression coefficients predicting class membership and another to estimate the intercepts (thresholds) for the conditional response probabilities. When the outcomes are continuous it is often assumed that, conditional on class, they have a multivariate normal distribution because for normal mixtures the solution for the optimization has a closed form (McLachlan and Peel, 2000). Furthermore, such model is a model-based clustering procedure which generalizes K-means clustering because it relaxes the strict assumptions of conditional independence and same error variance for all indicators and clusters. Similarly for dependent data, as is the case for LTM and HMM, the EM algorithm can be easily implemented to maximize the observed likelihood (McLachlan and Peel, 2000). The specification of the starting parameter values for the EM algorithm is of critical importance because in finite mixture modeling it has been well documented that they could lead to quite different solutions due to the presence of several local maxima (Titterington *et al*, 1985). In addition, a poor choice of starting values could exacerbate the slowness of the EM's rate of convergence.

For finite mixture models, the standard errors of the ML estimates obtained via the EM algorithm are estimated either from information-based methods or bootstrap. However, standard errors are inaccurate if the likelihood is not smooth and quadratic, and poor or unavailable when parameters are estimated at or near the boundary (e.g. when the conditional response probability is close to one) (Chung *et al*, 2004). As an alternative to ML a data augmentation approach with a Bayesian flavor has been proposed (Lanza *et al*, 2005) to estimate parameters and standard errors for LCM and LTM. However data augmentation, being a family member of the Markov chain Monte Carlo (MCMC) algorithms, adds the complication of *label-switching,* in which class labels change across iterations because labels are arbitrary since classes are unordered. Moreover, unusual likelihood features and label-switching are exacerbated for small sample sizes. To overcome these issues in LTM Chung *et al* (2007) proposed a dynamic algorithm to pre-classify with certainty one or few individuals with high posterior probability of being in a specific class using a ‘dynamic data-dependant prior’. The advantage of this method is that it requires a simple modification of the MCMC, rather than monitoring constraints or post-processing techniques (Chung, 2004).

First and second order generalized estimating equations (GEE and GEE-2) (Liang and Zeger, 1986; Liang *et al*, 1992) have also been used to estimate latent class models in order to avoid specifying the full likelihood which can be unfeasible for complex latent models. For instance, when there are a large number of latent classes, timepoints and/or indicators. GEE-2 uses the first and second moments which are necessary for identification of the latent class model parameters (Reboussin *et al*, 2001). In particular, Reboussin *et al* (2001) estimated a latent class marginal regression model for longitudinal data where the scientific interest was on the marginal latent class prevalences and not on the associations over time. Similarly, a two-stage estimation procedure was proposed (Reboussin *et al*, 1999) to estimate the LTM defined in equations (7) and (8) with binary indicators and time-dependent covariates assuming time-invariant conditional response probabilities. Specifically, in the first stage they used a set of second order estimating equations (GEE-2), $U_{2t}\left(\underset{p\times 1}{\boldsymbol{\rho}_1},\ldots,\underset{p\times 1}{\boldsymbol{\rho}_K},\boldsymbol{\pi}_t\right)$, to estimate the conditional response probabilities $\rho_{j|k_t}$ and the class probabilities π_{tk_t} at timepoints $t=1,2,\ldots,T$. In the second stage, they used first-order estimating equations (GEE), $U_1\left(\boldsymbol{\alpha},\boldsymbol{\Gamma};\hat{\boldsymbol{\delta}}\right)$ where $\hat{\boldsymbol{\delta}}=vech\left(\underset{p\times 1}{\hat{\boldsymbol{\rho}}_1},\ldots,\underset{p\times 1}{\hat{\boldsymbol{\rho}}_K},\underset{K\times 1}{\hat{\boldsymbol{\pi}}_1},\ldots,\underset{K\times 1}{\hat{\boldsymbol{\pi}}_T}\right)$, to estimate the regression coefficients of (8) for the transition probabilities which were the parameters of interest. In contrast to the EM algorithm where standard errors are not a by product of optimization, this two-stage approach simplifies their calculation while making some assumptions about the correlation structure at each stage. For example, they did not include into the GEE-2 the cross-products of indicators more than one unit apart because of the computational burden. Their simulation study suggested that this estimation procedure has good finite sample properties and highlighted the importance of having strong indicators (in the sense of high conditional response probabilities, $0.75 \le \rho_{j|k_t} \le 1.0$). Specifically, the degree of bias was low for both parameter estimates and robust standard errors even for small sample sizes (n=400) under strong measurement precision. However, weak indicators ($0.30 \le \rho_{j|k_t} \le 0.65$) of the latent class variables increased the bias in the parameters estimates and the proportion of failure to reach convergence.

Latent class prediction

Latent class models use *posterior* probabilities of class membership to assign subjects to a specific class. Hence, each subject has a probability for belonging to each class. *Posterior* probabilities are calculated via the Bayes' theorem

$$p_{ik} \equiv \Pr\left[c_{ik} = 1 \mid \mathbf{y}_i, \mathbf{x}_i\right] = \frac{\Pr\left[c_{ik} = 1 \mid \mathbf{x}_i\right] f\left(\mathbf{y}_i \mid c_{ik} = 1\right)}{\sum_{l=1}^{K} \Pr\left[c_{il} = 1 \mid \mathbf{x}_i\right] f\left(\mathbf{y}_i \mid c_{il} = 1\right)}$$

When classification is the goal, subjects need to be assigned into a single class and, typically, they are classified into the class with the highest *posterior* probability of class membership. In practice these predicted classes are often used as a predictor variable in a second model, and frequently treated as fixed. This can bias the estimates and the efficiency of standard errors by not taking into account the error in prediction. Mixture SEM (Muthén and Shedden, 1999), which integrates continuous and categorical latent variables both in the measurement and structural models, allows estimating simultaneously this *ad hoc* two-step model.

Number of classes

So far we have assumed the number of classes K is known. However in practice this is rarely the case, and currently there is no single accepted statistical test or fit-statistic to determine the number of classes (Nylund *et al*, 2007). Hence often it is decided using a combination of information criteria and substantive theory. The usual likelihood ratio test (LRT) cannot be used to compare nested latent class models because the regularity conditions required in classical maximum likelihood theory are violated and hence, its distribution is not chi-square. In particular, to compare a model with K classes vs. one with K-1 classes the reduced model is obtained by restricting the latent class probability to zero which is a value in the boundary of the parameter space. The Lo-Mendel-Rubin likelihood ratio test LMR-LRT (Lo *et al*, 2001) compares neighboring class models using an approximation of the LRT distribution under the assumption of within-class normality conditional on covariates. Another likelihood ratio test is based on parametric bootstrap (B-LRT) to estimate its empirical distribution (McLachlan and Peel, 2000). A second option to compare models with different number of classes is to use information criteria, such as the Bayesian Information Criterion (BIC). The information criteria are based on the likelihood function so they reward models that reproduce the observed data and some, parsimony if the criteria penalizes for the number of parameters. A recent simulation study (Nylund *et al*, 2007) showed that the bootstrap LRT performed better in identifying correctly the number of classes than the naïve LRT, the LMR-LRT and the BIC for the traditional LCM with either continuous or categorical outcomes using samples sizes of 200, 500 and 1000. However, some disadvantages of the B-LRT are increase in computation time (5 to 35 times greater), and lack of robustness since misspecifications on distributional and model assumptions lead to incorrect replicated datasets and hence, incorrect p-values. A third option are classification-based information criterion (McLachlan and Peel, 2000) which reward models that produce well-separated classes, such as the normalized entropy criterion (NEC) or the integrated classification likelihood criterion (ICLC). Another option to asses whether the assumed model has the sufficient number of classes is to use Bayesian diagnostic graphical techniques (Wang *et al*, 2005; Garrett and Zeger, 2000), but these graphical methods are not available on commercial software. Deciding on the number of classes is a difficult task and, it is important to address it correctly because as cautioned by Bauer and Curran (2003) spurious latent classes can be accommodating non-normality rather than discovering subpopulations.

Software

Table 1.1 summarizes the most known procedures or software to fit latent class models. One of the most important differences is what observed outcomes' scales are handled. Both commercial software (Mplus and Latent Gold) allow observed outcomes to be nominal, ordinal, count, and continuous (censored or truncated), but they are stand-alone software and not free. Although free procedures are more restricted in the type of observed outcomes allowed, models that can be estimated, and advanced capabilities (e.g. hierarchical data, complex survey data, graphics) they are very powerful and relatively easy to implement for various useful models. For example, PROC LCA and PROC LTA are SAS procedures developed and supported by the Methodology Center at Penn State (Lanza *et al*, 2008; Collins and Lanza, 2010) to fit traditional latent class models and latent transition models on categorical outcomes. Another example is PROC TRAJ (Jones *et al*, 2001) which estimates group-based trajectory models, allowing up to a third-order polynomial in time, specifying different order polynomials across the K trajectory classes. All procedures use maximum likelihood estimation implemented by the EM algorithm and Latent Gold can also estimate parameters with posterior mode, which penalizes solutions that are too close to the boundary space. Mplus is the only software that allows modeling categorical and continuous latent variables simultaneously.

1.2.2 Integration of continuous and categorical latent variables: Structural Equation Mixture Modeling (SEMM)

Muthén and Shedden (1999) proposed a general latent variable modeling framework that integrates continuous and categorical latent variables both in the measurement and structural models. In other words, the observed variables are related to each other through *m* factors (continuous latent variables) and *K* latent classes. This general framework is a synthesis and generalization of many latent variable models; it has also been referred as 2nd generation SEM (Muthén, 2001) or mixture SEM (Muthén, 2002), and as structural equation mixture model (SEMM) (Bauer and Curran, 2004). Among many other extensions, this framework allows a multiple-groups SEM with unobserved group membership (**Figure 1.4**) which has been of much interest in applied research.

One particular model within this general framework is the heterogeneity model (Verbeke and Lesaffre, 1996) which extended the linear mixed model (Laird and Ware, 1982) by allowing the random effects to be sampled from a mixture of normal distributions and hence, incorporated an underlying latent class variable. This mixture model in addition to allow a more flexible class of distributions for fitting non-normal distributions for the random effects allows classifying subjects based on trajectory profiles (Verbeke and Melenberghs, 2000). This possibility of uncovering unobserved heterogeneity and finding substantively meaningful groups has been very appealing in both health and social sciences. For example, in the health sciences it has been used to characterize the course of low back pain (Dunn, 2006), assess patterns of physical activity (Metzger *et al*, 2008), and to classify subjects into prostate cancer risk classes (McCulloch *et al*, 2002). In the social sciences it has been popularized as the growth mixture model (GMM) (Muthén and Shedden, 1999), and has been applied to substance use and physical aggression. The group-based trajectory model (Nagin *et al*, 2005), introduced in section 1.2.1.2, is a special case where the covariance matrix of the random effects is constrained to be zero and, therefore there is no between-subject heterogeneity within classes.

However, as cautioned by Bauer and Curran (2003) latent classes and latent trajectories should not be substantively over interpreted unless there is an underlying premise that subjects belong to distinct groups (classes). Empirical evidence (Bauer and Curran, 2003) shows that multiple latent trajectory classes can be estimated and fit the data well even when the data come from a single group with a non-normal distribution. Furthermore even when latent classes truly exist, misspecification of the model and/or the distributional and linearity assumptions can lead to spurious latent classes (Bauer and Curran, 2004).

In the psychometric and social sciences, historically and currently, there is debate between using categorical or continuous latent variables to represent certain constructs (Muthén, 2006). In the nutritional epidemiology literature the debate is to a less extent, and dietary patterns are treated as continuous or discrete depending on the research question and the multivariate statistical method preferred (see section 1.1.1). In general, epidemiologists agree that 1) factor analysis is very useful to understand which foods are eaten together (from the factor loadings), reduce dimension, and examine overall diet (using DP factor scores) and disease associations, whereas 2) cluster analysis is useful to classify subjects to estimate the risk of the outcome for each exposure class compared to a reference class. However, even when DP are conceptualized and derived as continuous variables, often subjects are classified to simplify the interpretation.

1.2.3 Model selection and goodness of fit

Measures of goodness-of-ft involve the fitting function $F(\boldsymbol{\theta})$ because it measures the discrepancy between the sample moment structure and the one implied by the model. In Gaussian SEM the absolute fit for latent variable models can be evaluated using the likelihood ratio test statistic $T = (n-1)F(\hat{\boldsymbol{\theta}}) : \ \chi^2(df)$ where df is the difference between the number of non redundant elements in $\left\{\bar{\mathbf{V}}, \mathbf{S}\right\}$ and the number of free parameters. The disadvantages of this test statistic are that the larger the sample size the more likely the models do not fit well, overparameterized models have better fit, and is sensitive to violations of normality. In the social sciences literature, many goodness-of-fit indices have been proposed using different rationales to derive them to overcome the problem of sample size in the chi-square test statistic. The main problem is that their distributions are not known and, hence, statistical inference is not possible. Therefore guidelines for lack of fit are given empirically by rules of thumb or some based on simulations. In addition, currently there are more than 25 indices proposed and reported in commercial software because there is no agreement among methodologists which ones are best. Three popular comparative fit indices are the Tucker-Lewis index (TLI; Tucker and Lewis, 1973), the *incremental fit index* IFI (Bollen, 1989), and the *comparative fit index* (CFI; Bentler, 1990). Comparative fit indexes compare the fitted model to the baseline model, and are constructed to range between 0 (lack of fit) and 1 (perfect fit).

One residual-based fit index that is becoming more accepted is the root mean square error of approximation (RMSEA; Steiger, 1990). The RMSEA tests the null hypothesis of ‘close fit’ rather than exact fit (Browne and Cudeck, 1993) because the chi-square test of the null hypothesis of exact fit is an omnibus test which in practice is almost always rejected for large sample sizes. It is an estimate of the approximate fit in the population in the sense that it measures discrepancy between the true moment structure and the one implied by the approximating model. RMSEA takes values from 0 to 1 where values below 0.1 are acceptable and below 0.05 are considered a good fit (Browne and Cudeck, 1993). Confidence intervals for RMSEA can also be estimated (Browne and Cudeck, 1993).

Model selection in latent variable models includes the Bayesian Information criterion (BIC), the Akaike’s information criterion (AIC), and the Deviance Information criterion (DIC). It is not clear which ‘n’ (total sample size or total number of independent units) must be used when computing the BIC for latent variable models.

$$\underset{7\times1}{\mathbf{Y}_i} = \begin{pmatrix} 1 & 0 \\ \lambda_{12} & 0 \\ \lambda_{13} & 0 \\ \lambda_{14} & 0 \\ 0 & 1 \\ 0 & \lambda_{26} \\ 0 & \lambda_{27} \end{pmatrix} \begin{pmatrix} \eta_{1i} \\ \eta_{2i} \end{pmatrix} + \boldsymbol{\varepsilon}_i$$

$$\underset{7\times1}{\boldsymbol{\varepsilon}_i} : \; MVN\left(\mathbf{0}, Diag\left\{\boldsymbol{\theta}\right\}\right) \quad \underset{2\times1}{\boldsymbol{\eta}_i} : \; MVN\left(\mathbf{0}, \boldsymbol{\Psi}\right)$$

$$\mathrm{cov}\left(\boldsymbol{\varepsilon}_i, \boldsymbol{\eta}_i\right) = \mathbf{0} \quad \boldsymbol{\Psi} \equiv \begin{pmatrix} \psi_1 & \psi_{12} \\ \psi_{12} & \psi_2 \end{pmatrix}$$

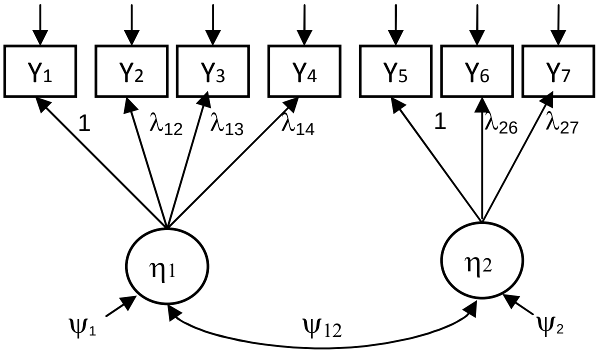


FIGURE 1.1 Path diagram for a 2-factor model

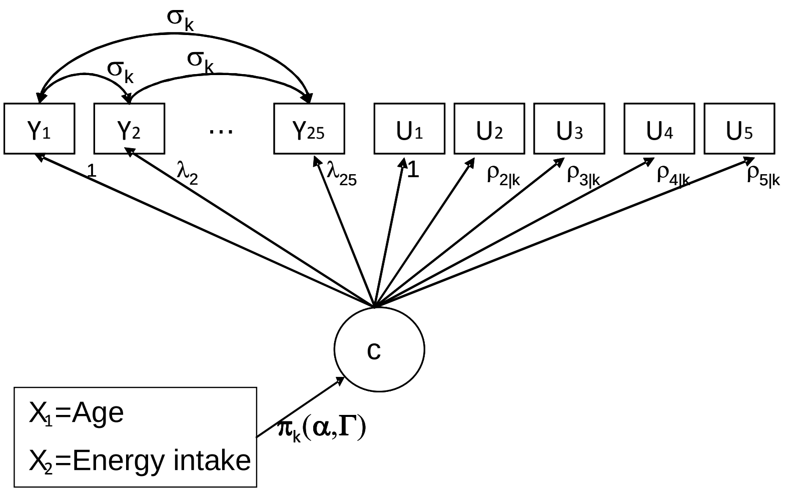


FIGURE 1.2 Path diagram for a latent class model (LCM)

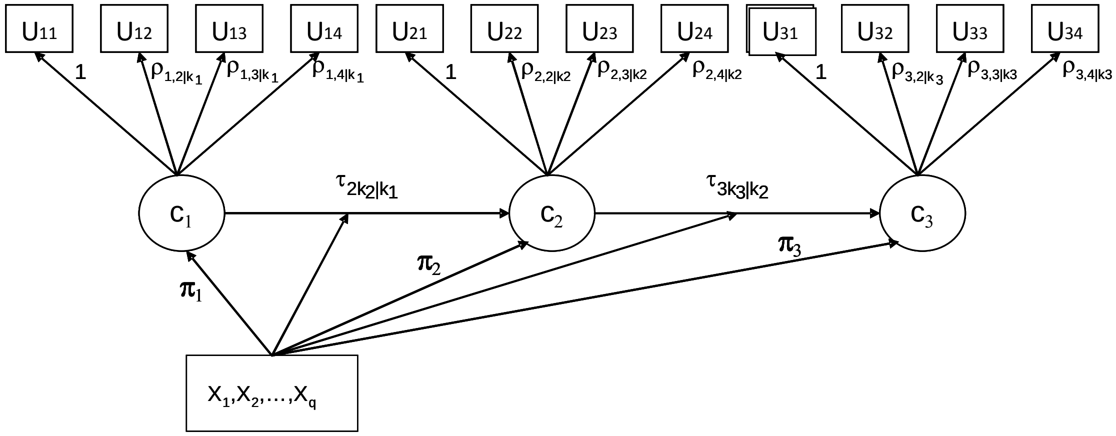


FIGURE 1.3 Path diagram for a latent transition model (LTM)

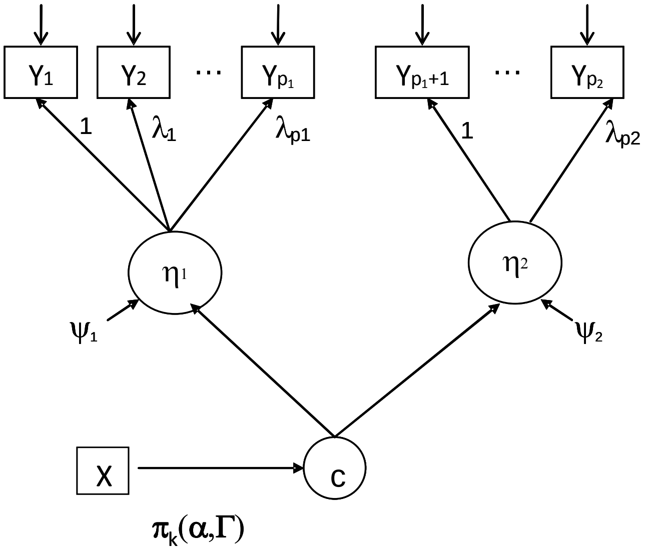


FIGURE 1.4 Path diagram for a factor mixture model

TABLE 1.1 Software to estimate latent class models

	Procedures[†] that can be accommodated on general purpose statistical software					**Stand-alone**		
	PROC LCA	**PROC LTA**	**PROC TRAJ**	**GLLAMM**	**poLCA**	**WinLTA**	**Mplus**	**Latent Gold**
Acronym	Latent class analysis	Latent transition analysis	Trajectory	Generalized Linear Latent And Mixed Models	Polytomous Variable Latent Class Analysis	Latent Class and Latent Transition Analysis		
Version	1.2.3 beta	1.2.3 beta	-	-	1.1	3.1	5.1	4.0
Date of release	Jan-10	Jan-10	Sep-05	Nov-06	Nov-07	May-02	Nov-07	
Authors	Lanza ST, Lemmon DR, Schafer JL, Collins LM	Lanza ST, Lemmon DR, Schafer JL, Collins LM	Jones B, Nagin DS, Roeder K	Rabe-Hesketh S, Skrondal A	Linzer DA, Lewis J	Lanza ST, Lemmon DR, Schafer JL, Collins LM, Flaherty BP	Muthén LK, Muthén BO	Vermunt JK, Magidson J
Institution	The Methodology Center, Penn State University	The Methodology Center, Penn State University	Carnegie Mellon University	Biostatistics, University of California at Berkeley	Political Science, Emory University and UCLA	The Methodology Center, Penn State University	Muthén and Muthén	Statistical Innovations Inc.
Free[‡]	Y	Y	Y	Y	Y	Y	N	N
Software	SAS	SAS	SAS	STATA	R	-	-	-
Outcome scale								
Nominal	Y	Y	N	Y	Y	Y	Y	Y
Ordinal	N	N	Only binary	Y	Y	N	Y	Y
Count	N	N	Poisson and zero-inflated Poisson	Y	N	N	Poisson and zero-inflated Poisson	(truncated/ overdispersed) Poisson or Binomial
Continuous	N	N	Multivariate and censored normal	Y	N	N	Multivariate and censored normal	Multivariate, censored and truncated normal
Different scale types jointly	N	N	N	Y	N	N	Y	Y

TABLE 1.1 Software to estimate latent class models (continued)

	Procedures† that can be accommodated on general purpose statistical software					Stand-alone		
	PROC LCA	**PROC LTA**	**PROC TRAJ**	**GLLAMM**	**poLCA**	**WinLTA**	**Mplus**	**Latent Gold**
Covariates	To model class membership	1) Class membership 2) Transitions over time	Y, time-stable	Y	Y, class membership	Y	Y	Y
Allows continuous and categorical latent variables simultaneously	N	N	N	N	N	N	Y	?
Parameter estimation								
Maximum likelihood using EM	Newton-Raphson	Newton-Raphson	Quasi-Newton	Newton-Raphson (uses STATA's ml with d0 method to maximize the likelihood)	Newton-Raphson	Newton-Raphson and Quasi-Newton	Newton-Raphson and Quasi-Newton	Newton-Raphson
Other	N	N	N	N	N	Posterior mode	WLS	Posterior Mode using conjugate priors
Parameter estimate standard errors	Y	Y	Observed information matrix	Observed information matrix	Empirical observed information matrix	Data Augmentation (with diffuse prior by default)	Empirical observed information, observed information or robust.	
Missing data								
Outcomes	MAR	MAR	MAR	MAR	MAR	MAR	MAR	MAR
Covariates	Listwise deletion	Listwise deletion	Listwise deletion	Listwise deletion	Listwise deletion	Listwise deletion	Listwise deletion	Listwise deletion
Hierarchical data	N	N	N	Y	N	N	Y	Y
Complex survey data	N	N	N	N	N	N	Y	Y
Multiple-group analysis	Y	Y	Y	Y	Y	Y	Y	Y

TABLE 1.1 Software to estimate latent class models (continued)

	Procedures[†] that can be accommodated on general purpose statistical software					Stand-alone		
	PROC LCA	**PROC LTA**	**PROC TRAJ**	**GLLAMM**	**poLCA**	**WinLTA**	**Mplus**	**Latent Gold**
Constraints	Y	Y	Y	Y	N	Y	Y	Y
Starting values								
By user	Y	Y	Y	Y	Y	Must	Y	Y
Multiple sets	Y	Y	N	N	Y	N	Y	Y
Goodness-of-fit[§]	N	N	AIC and BIC	Log-likelihood	AIC, BIC, Pearson and LR chi-square	AIC, BIC, Pearson and LR chi-square, RMSEA, TLI, CFI, SRMR	AIC, BIC, Pearson and LR chi-square, RMSEA, TLI, CFI, SRMR	AIC, AIC3, CAIC, BIC, DI, Pearson, LR, and Cressie-Read chi-square statistics
Generates simulated data	N	N	N	N	Y	N	Y	N
Statistical test to determine number of classes	N	N	N	N	N	N	Lo-Mendell-Rubin LRT and Bootstrap LRT	N

† These procedures are macros developed, documented and maintained by users and are not part of the software.

‡ However SAS and STATA software is not free.

§ AIC Akaike Information Criterion; AIC Akaike Information Criterion 3; CAIC Consistent Akaike Information Criterion; BIC Bayesian Information Criterion; DI Dissimilarity Index; RMSEA Root Mean Square Error of Approximation; TLI Tucker-Lewis Index; CFI Comparative Fit Index.

TABLE 1.2 Articles that have studied dietary patterns over time

Reference and Aim	Population & sample size‡	Dietary assessment	Time points	1. Identification of DP	2. Computation of DP scores	3. Analysis of DP change
Newby et al, J Nutr. (2006) 136(3). Stability	Swedish women n=33,840	semi quant FFQ (past 6 mo) servings/day	Two: 1987: (67 foods → 27 food groups) 1997: (97 foods → 32 food groups)	At each time point: 1. EFA (PCA-Varimax) 2. CFA based on EFA (loadings>0.2 and knowledge). 4 confirmed ‘food scores’: healthy, western, alcohol, sweets.	Computed for each factor, method (EFA and CFA) at each time point. DP scores=standardized intakes weighted by their factor loadings and summed.	**Stability:** Spearman correlation between timepoints, by factor (DP) and method (EFA and CFA) (Table 5).
Weismayer et al, J Nutr. (2006) 136(6). Stability	Swedish women 4 subsamples (n=1000 each)	semi quant FFQ (past 6 mo) 67foods → 25 f groups servings/day	Two: Baseline and one follow-up at either 4,5,6,7 years apart	At each time point: 1. EFA(PCA-Varimax) 2. CFA based on EFA (loadings>0.2 and knowledge). 3 confirmed ‘food scores’: healthy, western, alcohol.	Computed for each factor, method (EFA and CFA), and each time point. DP scores=standardized intakes weighted by their factor loadings and summed.	**Stability:** Spearman correlation between timepoints, by factor and method (Table 5). **Internal stability**: “Test the significance of changes in the covariance matrix between baseline and follow up”.
Newby et al, J Nutr. (2006) 136(10). Within-subject DP change	Swedish women n=33,840	semi quant FFQ (past 6 mo) servings/day	Two: 1987: (67 foods → 27 food groups) 1997: (97 foods → 32 food groups))	At each time point: 1. EFA(PCA-Varimax) 2. CFA based on EFA (loadings>0.2 and knowledge). 4 confirmed ‘food scores’: healthy, western, alcohol, sweets.	At each time point: DP scores=standardized intakes weighted by their confirmatory factor loadings and summed.	**Exposure = Change in DP scores** (1997-1987), for each DP. **Model of change in BMI:** In the linear regression models, the 4 changes in food patterns scores are used as covariates simultaneously.
Mishra et al, British J Nutr. (2006) 96. Stability	British cohort n=1,265 (M & W stratified analyses)	5-day food diary 126 binary food groups	Three: 1982 (36 y) 1989 (43 y) 1999 (53 y)	EFA at 1999 to identify number of patterns and items that loaded highly. They said that “cross-sectional analysis of DP at other two ages showed that similar DP existed at each time”. Women: 3 DP (ethnic & alcohol, F&V&dairy, meat&potato&sweets). Men: 2 DP (ethnic & alcohol, mixed).	Simplified DP score equation in 1999 (sum the unweighted food items which load most highly (≥0.25)). Exact same equation used in 1982 and 1989. Note that because binary items the score is interpreted as number of items consumed rather than quantity consumed (i.e. DP reflect variety).	**Stability:** Weighted Kappa statistic on tertiles of DP for every pair of time points by DP. **Assoc between DP and risk factors:** Linear mixed model Y_{ijk}=Simplified DP for subject i, dietary pattern j at time k NOTE: BMI is time-varying and Kappa and mixed model don’t agree.

TABLE 1.2 Articles that have studied dietary patterns over time (continued)

Reference and Aim	Population & sample size‡	Dietary assessment	Time points	1. Identification of DP	2. Computation of DP scores	3. Analysis of DP change
McNaughton et al, J Nutr. (2007) 137(1).	Same as Mishra et al, British J Nutr. (2006) 96 but using other risk factors.					Temporality? “Risk factors for this group were measured in 1999 at age 53”: waist circumference, blood pressure, blood sample red cell folate, glycated hemoglobin, total cholesterol, HDL,LDL).
Schulze el al, Obesity (2006) 14(4). Within-subject DP change	Nurses’ II n=51,670	semi quant FFQ (past year) 133 foods → 39 food groups g/day	Three: 1991 1995 1999	At each time point: PCA-Varimax eigenvalues > 1. 2 DP: Prudent and Western.	DP scores = PC scores energy-adjusted using the residuals method.	DP scores were categorized in quintiles and participants were classified according to change in category of DP score between pairs of years (low-low, high-high, low-high, high-low where low is lower two quintiles and high is upper two quintiles). Weight change model done separately by pairs of years and corresponding change in DP.
Meyerhardt et al, JAMA (2007) 298(7). Within-subject DP change	Stage III colon cancer receiving chemotherapy n=1,009	semi quant FFQ (3 mo) 131 foods → 39 food groups g/day	Two: 1. In the middle of their adjuvant chemotherapy 2. Six mo after the completion of their adjuvant chemotherapy	EFA (PCA-Varimax) on updated dietary exposure (i.e. both FFQ were combined) Eigenvalues>1.5 and interpretability. 2 DP: Prudent and Western.	DP scores = PC scores	For each outcome and each DP: Cox proportional hazards regression.
Cuco et al, Eur J Clin Nutr (2006) 60. Within-subject DP change	Spanish women n=80	7-day food diary 21 food groups g/day	Six: 1 Pre-pregnancy 4 Pregnancy {6,10,26,38 w} 1 Post-partum 6mo	At each time point: PCA (no rotation); eigenvalues >1, Scree-plot and interpretability. 2 DP: (sweetened-beverages/ sugars, meat&veg) for every time point except postpartum (sweetened beverages/sugars).	At each time point: Factor scores estimated by the regression method.	Stability: Congruence coefficients and MANOVA for the standardized foods that defined the DP

TABLE 1.2 Articles that have studied dietary patterns over time (continued)

Reference	Population & sample size[‡]	Dietary assessment	Time points	1. Identification of DP	2. Computation of DP scores	3. Analysis of DP change
Togo et al, Int. J. Obesity (2004) 28. Within-subject DP change	Danish n=2,436 (analyses stratified by gender)	FFQ (past year) 8 frequencies 26 foods groups → 21 food groups (5 omitted because of skewness)	Three: 1982 1987 1993	EFA at 1982 on a subsample of n=1,806 Women: 2 DP (green, sweet-traditional) Men: 3 DP (green, sweet, traditional) **Prospective analysis.** DP exposure is fixed at baseline and outcome is longitudinal. **Longitudinal Analysis**: DP change (exposure) and outcome is longitudinal.	**Two ways:** **1. For cross-sectional and prospective analyses:** CFA based on EFA (loadings≥0.3) using full baseline (n=3,785) Factors are correlated. **2. For longitudinal analysis:** Mean-structure FA. Loadings on the factors and thresholds were kept equal at the two time points.	**Cross-sectional analysis (baseline):** BMI = DP scores + covariates **Prospective analysis:** ΔBMI = DP scores (baseline) + covariates Two separate models: ΔBMI from 1982 to 1987 (5y) ΔBMI from 1982 to 1993 (11y) **Longitudinal Analysis**: ΔBMI= DP scores (baseline) + Δ DP scores + covariates ΔBMI from 1987 to 1993 (6 y) ΔDP scores from 1982 to 1987
Northstone and Emmett, Br J Nutr (2007) Stability	British women ALSPAC (Avon Longitudinal Study of Parents and Children) n=8,935	Pregnancy: 44-item FFQ Postpartum: 52-item FFQ	Two: 32 wk gestation 47 mo postpartum	At each time point: PCA (Varimax rotation); Scree-plot and interpretability.	**Two ways:** **1.** Component scores at each timepoint **2.** Applying the factor loadings obtained at pregnancy for both timepoints	- Pearson's correlation - Partial correlations adjusted by energy - Paired t-test - Bland-Altman limits of agreement -Weighted Kappa of quintiles of DP (pregnancy vs. postpartum)
Greenwood DC et al, Proceedings Nutr Soc of London (2003) 62. Stability	UK Women's Cohort Study n=1,938	FFQ (past year) 10 frequencies 217 foods → 74 food groups	Two: Baseline: 1995 to 1998 Follow-up: 5 years later	k-means cluster analysis (variables were not standardized)	Not applicable	1. "Participants were reclassified, using the same cluster definitions, based on their reported diet 5 years later". 2. Kappa statistic to test agreement between clusters (DP).

‡The sample size is the number of subjects used in the analysis.

TABLE 1.3 Articles that have studied dietary patterns during pregnancy and postpartum

Reference	Population & sample size[‡]	Dietary assessment	Design	Identification of DP	Computation of DP scores	Analysis
Northstone et al, Eur J Clin Nutr (2007)	British women ALSPAC (Avon Longitudinal Study of Parents and Children) (1991-1992) n=12,053	44-item FFQ Times per week	Cross-sectional (at 32 weeks of gestational age)	PCA (Varimax rotation); Scree-plot and interpretability. Repeated in two randomly selected split samples to assess repeatability. 5 DP: Health conscious; Traditional; Processed; Confectionery; Vegetarian.	Component scores	Linear regression
Northstone at al, Br J Nutr (2007)	Same as Northstone et al, Eur J Clin Nutr (2007)					Two sets of analysis: **1.** Pearson's correlations between DP and absolute nutrients intake and also partial correlations adjusting by energy intake **2.** Nutrients regressed on DP in quintiles and energy intake
Northstone and Emmett, Br J Nutr (2007)	British women ALSPAC (1991-1992) n=8,935	Pregnancy: 44-item FFQ Postpartum: 52-item FFQ Times per week	Longitudinal Two time points: 32 wk gestation 47 mo postpartum	At each time point: PCA (Varimax rotation); Scree-plot and interpretability.	**Two ways:** **3.** Component scores at each timepoint **4.** Applying the factor loadings obtained at pregnancy for both timepoints	- Pearson's correlation - Partial correlations adjusted by energy - Paired t-test - Bland-Altman limits of agreement -Weighted Kappa of quintiles of DP (pregnancy vs. postpartum)
Cuco et al, Eur J Clin Nutr (2006) 60.	Spanish Women (1992-1996) n=80	7-day food diary 21 food groups (g/day)	Longitudinal Six time points: 1 Pre-pregnancy 4 Pregnancy {6,10,26,38 w} 1 Postpartum 6mo	At each time point: PCA (no rotation); eigenvalues >1, Scree-plot and interpretability. 2 DP: (sweetened-beverages/ sugars, meat&veg) for every time point except postpartum (sweetened beverages/sugars).	At each time point: Factor scores estimated by the regression method.	**Stability:** Congruence coefficients and MANOVA for the standardized foods that defined the DP

TABLE 1.3 Articles that have studied dietary patterns during pregnancy and postpartum (continued)

Reference	Population & sample size‡	Dietary assessment	Design	Identification of DP	Computation of DP scores	Analysis
Knudsen et al, Eur J Clin Nutr (2007)	Danish women (1997-2002) n= 44,612	360-item FFQ semiquantitative (previous month) 36 food-groups (g/day)	Cross-sectional (at 32 weeks of gestational age)	EFA (Varimax rotation) eigenvalues, Scree-plot and interpretability. First they extracted 2 DP. Then they classified the women into mutually exclusive classes using the two DP jointly (by cross-tabulating the DP's quintiles)	Factor scores estimated by the regression method.	1-way ANOVA with Tukey's correction for multiple tests
Arkkola et al, Pub Health Nutr (2007)	Finish women (1997-2002) n=3,730	181-item FFQ semiquantitative Dietary intake was retrospectively assessed for the last month of pregnancy (women received FFQ at delivery and turned back at 3 months postpartum) 52 food-groups (g/day)	Cross-sectional	PCA (Varimax) 7 DP: Healthy; Fast foods; Traditional bread; traditional meat; low-fat foods; coffee; alcohol and butter	Principal components scores	- Pearson's correlation - Linear regression
Crozier et al, Br J Nutr (2008)	British women (1991-1992) n=585	100-item FFQ (3 mo) 4-day diary after FFQ -> 100 foods 49 food-groups	Cross-sectional (early pregnancy; median gestation 1`5.3 wk)	PCA on standardized variables 2 DP: Prudent and Western	Principal components scores	- Pearson's correlation - For each DP, Bland-Altman plots for agreement between FFQ and diary scores

‡The sample size is the number of subjects used in the analysis.

REFERENCES

Arkkola, T., U. Uusitalo, C. Kronberg-Kippila, S. Mannisto, M. Virtanen, M. G. Kenward, R.Veijola, M. Knip, M. L. Ovaskainen, and S. M. Virtanen. 2008. Seven distinct dietary patterns identified among pregnant finnish women--associations with nutrient intake and sociodemographic factors. *Public Health Nutrition* 11, (2) (Feb): 176-82.

Bauer, D. J., and P. J. Curran. 2003. Overextraction of latent trajectory classes: Much ado about nothing? reply to rindskopf (2003), muthen (2003), and Cudeck and henly (2003). *Psychological Methods* 8, : 384–393.

Bauer, D. J., and P. J. Curran. 2004. The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods* 9, (1) (Mar): 3-29.

———. 2003. Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods* 8, (3) (Sep): 338-63.

Berk, K. N., and P. A. Lachenbruch. 2002. Repeated measures with zeros. *Statistical Methods in Medical Research* 11, (4) (Aug): 303-16.

Bland, J. M., and D. G. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, (8476) (Feb 8): 307-10.

Bollen, K. A. 1989. *Structural equations with latent variables*. Wiley series in probability and mathematical statistics. applied probability and statistics section. New York: Wiley.

Bollen, K. A., and P. J. Curran. 2006. *Latent curve models : A structural equation perspective*. Wiley series in probability and statistics. Hoboken, N.J.: Wiley-Interscience.

Chung, H., S. T. Lanza, and E. Loken. 2007. Latent transition analysis: Inference and estimation. *Statistics in Medicine* (Dec 11).

Chung, H., E. Loken, and J. L. Schafer. 2004. Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution. *The American Statistician* 58, (2): 152-8.

Chung, H., Y. Park, and S. T. Lanza. 2005. Latent transition analysis with covariates: Pubertal timing and substance use behaviours in adolescent females. *Statistics in Medicine* 24, (18) (Sep 30): 2895-910.

Clogg, C. C. 1992. The impact of sociological methodology on statistical methodology. *Statistical Science* 7, (2): 183-96.

Collins LM and Lanza ST. 2010. Latent class and latent transition analysis: with applications in the social, behavioral, and health sciences. Wiley, Hoboken, N.J.

Collins, L. M., and S. E. Wugalter. 1992. Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research* 27, (1): 131-57.

Crozier, S. R., H. M. Inskip, K. M. Godfrey, and S. M. Robinson. 2008. Dietary patterns in pregnant women: A comparison of food-frequency questionnaires and 4 d prospective diaries. *British Journal of Nutrition* 99, (4): 869-75.

Cuco, G., J. Fernandez-Ballart, J. Sala, C. Viladrich, R. Iranzo, J. Vila, and V. Arija. 2006. Dietary patterns and associated lifestyles in preconception, pregnancy and postpartum. *European Journal of Clinical Nutrition* 60, (3) (Mar): 364-71.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37). Journal of the Royal Statistical Society.Series B, Methodological 39, : 1.

Diggle, Peter, P. J. Heagerty, K. Y. Liang, and S. L. Zeger. 2002. Analysis of longitudinal data. Oxford ;New York: Oxford University Press.

Fahey, M. T., C. W. Thane, G. D. Bramwell, and W. A. Coward. 2007. Conditional gaussian mixture modelling for dietary pattern analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170, (1): 149-66.

Garrett, E. S., and S. L. Zeger. 2000. Latent class model diagnosis. *Biometrics* 56, (4) (Dec): 1055-67.

Hoffman, K., M. B. Schulze, H. Boeing, and H. P. Altenburg. 2002. Dietary patterns: Report of an international workshop. *Public Health Nutrition* 5, (1) (Feb): 89-90.

Hoffmann, K., M. B. Schulze, A. Schienkiewitz, U. Nothlings, and H. Boeing. 2004. Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *American Journal of Epidemiology* 159, (10) (May 15): 935-44.

Hu, F. B. 2002. Dietary pattern analysis: A new direction in nutritional epidemiology. *Current Opinion in Lipidology* 13, (1) (Feb): 3-9.

Jones, B. L., D. S. Nagin, and K. Roeder. 2001. A SAS procedure based on mixture models for estimating developmental trajectories. Sociological Methods & Research 29, (3): 374.

Kant, A. K. 2004. Dietary patterns and health outcomes. *Journal of the American Dietetic Association* 104, (4) (Apr): 615-35.

Knudsen, V. K., I. M. Orozova-Bekkevold, T. B. Mikkelsen, S. Wolff, and S. F. Olsen. 2007. Major dietary patterns in pregnancy and fetal growth. *European Journal of Clinical Nutrition* (Mar 28): 1-8.

Lanza, S., and L. M. Collins. 2008. A new SAS procedure for latent transition analysis: Transitions in dating and sexual risk behavior. *Developmental Psychology* 44, (2): 446-56.

Lanza, S. T., L. M. Collins, J. L. Schafer, and B. P. Flaherty. 2005. Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychological Methods* 10, (1) (Mar): 84-100.

Lo, Y., N. R. Mendell, and D. B. Rubin. 2001. Testing the number of components in a normal mixture. *Biometrika* 88, (3): 767-78.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate analysis*. Probability and mathematical statistics. London ; New York: Academic Press.

McCulloch, C. E., H. Lin, E. H. Slate, and B. W. Turnbull. 2002. Discovering subpopulation structure with latent class mixed models. *Statistics in Medicine* 21, (3) (Feb 15): 417-29.

McLachlan, Geoffrey J., and David Peel. 2000. *Finite mixture models*. Wiley series in probability and statistics. applied probability and statistics section. New York: Wiley.

McNaughton, S. A., G. D. Mishra, A. M. Stephen, and M. E. Wadsworth. 2007. Dietary patterns throughout adult life are associated with body mass index, waist circumference, blood pressure, and red cell folate. *The Journal of Nutrition* 137, (1) (Jan): 99-105.

Metzger, J. S., D. J. Catellier, K. R. Evenson, M. S. Treuth, W. D. Rosamond, and A. M.Siega-Riz. 2008. Patterns of objectively measured physical activity in the united states. *Medicine and Science in Sports and Exercise* (Feb 29): 630-8.

Meyerhardt, J. A., D. Niedzwiecki, D. Hollis, L. B. Saltz, F. B. Hu, R. J. Mayer, H. Nelson, et al. 2007. Association of dietary patterns with cancer recurrence and survival in patients with stage III colon cancer. *JAMA : The Journal of the American Medical Association* 298, (7) (Aug 15): 754-64.

Miglioretti, D. L. 2003. Latent transition regression for mixed outcomes. *Biometrics* 59, (3) (Sep): 710-20.

Mishra, G. D., S. A. McNaughton, G. D. Bramwell, and M. E. Wadsworth. 2006. Longitudinal changes in dietary patterns during adult life. *The British Journal of Nutrition* 96, (4) (Oct): 735-44.

Muthen, B. 2006. Should substance use disorders be considered as categorical or dimensional? *Addiction (Abingdon, England)* 101 Suppl 1, (Sep): 6-16.

———. 2003. Statistical and substantive checking in growth mixture modeling: Comment on bauer and curran (2003). *Psychological Methods* 8, (3) (Sep): 369,77; discussion 384-93.

———. 2002. Beyond SEM: General latent variable modeling. *Behaviormetrika* 29, (1): 81-117.

———. 2001. Latent variable mixture modeling. In *New developments and techniques in structural equation modeling.*, eds. G. A. Marcoulides, R. E. Schumacker, 1-33Lawrence Erlbaum Associates.

———. 2001. Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In *New methods for the analysis of change.*, eds. L. M. Collins, A. Sayer, 291-322American Psychological Association.

Muthen, B., and L. K. Muthén (1998-2007). Mplus software, Version 5.1.

Muthen, B., and K. Shedden. 1999. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 55, (2) (Jun): 463-9.

Newby, P. K., and K. L. Tucker. 2004. Empirically derived eating patterns using factor or cluster analysis: A review. *Nutrition Reviews* 62, (5) (May): 177-203.

Newby, P. K., C. Weismayer, A. Akesson, K. L. Tucker, and A. Wolk. 2006. Longitudinal changes in food patterns predict changes in weight and body mass index and the effects are greatest in obese women. *The Journal of Nutrition* 136, (10) (Oct): 2580-7.

———. 2006. Long-term stability of food patterns identified by use of factor analysis among swedish women. *The Journal of Nutrition* 136, (3) (Mar): 626-33.

Northstone, K., P. Emmett, and I. Rogers. 2007. Dietary patterns in pregnancy and associations with socio-demographic and lifestyle factors. *European Journal of Clinical Nutrition* (Mar 21).

Northstone, K., and P. M. Emmett. 2007. A comparison of methods to assess changes in dietary patterns from pregnancy to 4 years post-partum obtained using principal components analysis. *The British Journal of Nutrition* (Oct 5): 1-8.

Northstone, K., P. M. Emmett, and I. Rogers. 2008. Dietary patterns in pregnancy and associations with nutrient intakes. *The British Journal of Nutrition* 99, (2) (Feb): 406-15.

Nylund, K. L., T. Asparouhov, and B. O. Muthén. 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *STRUCTURAL EQUATION MODELING* 14, (4): 535-69.

Padmadas, S. S., J. G. Dias, and F. J. Willekens. 2006. Disentangling women's responses on complex dietary intake patterns from an indian cross-sectional survey: A latent class analysis. *Public Health Nutrition* 9, (2) (Apr): 204-11.

Patterson, B. H., C. M. Dayton, and B. I. Graubard. 2002. Latent class analysis of complex sample survey data: Application to dietary data. *Journal of the American Statistical Association* 97, (459): 721-42.

Rabe-Hesketh, S., and A. Skrondal. 2007. Classical latent variable models for medical research. *Statistical Methods in Medical Research* 17, (1) (Sep 13): 5-32.

Reboussin, B. A., D. M. Reboussin, K. Y. Liang, and J. C. Anthony. 1998. Latent transition modeling of progression of health-risk behavior. *Multivariate Behavioral Research* 33, (4): 457-78.

Reboussin, B. A., and J. C. Anthony. 2001. Latent class marginal regression models for modelling youthful drug involvement and its suspected influences. *Statistics in Medicine* 20, (4) (Feb 28): 623-39.

Reboussin, B. A., K. Y. Liang, and D. M. Reboussin. 1999. Estimating equations for a latent transition model with multiple discrete indicators. *Biometrics* 55, (3) (Sep): 839-45.

Sánchez, B. N., E. Budtz-Jørgensen, L. M. Ryan, and H. Hu. 2005. Structural equation models: A review with applications to environmental epidemiology. *Journal of the American Statistical Association* 100, (472): 1443-55.

SAS software, Version 9.1 of the SAS System for Windows (SAS, 2002-2003). Copyright © SAS, 2002-2003. SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Schulze, M. B., T. T. Fung, J. E. Manson, W. C. Willett, and F. B. Hu. 2006. Dietary patterns and changes in body weight in women. *Obesity (Silver Spring, Md.)* 14, (8) (Aug): 1444-53.

Schulze, M. B., and K. Hoffmann. 2006. Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. *The British Journal of Nutrition* 95, (5) (May): 860-9.

Schulze, M. B., K. Hoffmann, A. Kroke, and H. Boeing. 2003. An approach to construct simplified measures of dietary patterns from exploratory factor analysis. *The British Journal of Nutrition* 89, (3) (Mar): 409-19.

Skrondal, Anders, and S. Rabe-Hesketh. 2004. *Generalized latent variable modeling :Multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC interdisciplinary statistics series. Boca Raton: Chapman & Hall/CRC.

Togo, P., M. Osler, T. I. Sorensen, and B. L. Heitmann. 2004. A longitudinal study of food intake patterns and obesity in adult danish men and women. *International Journal of Obesity and Related Metabolic Disorders : Journal of the International Association for the Study of Obesity* 28, (4) (Apr): 583-93.

Verbeke, Geert, and Geert Molenberghs. 2000. *Linear mixed models for longitudinal data*. Springer series in statistics. New York: Springer.

Wang, C. P., C. H. Brown, and K. Bandeen-Roche. 2005. Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association* 100, (471): 1054-77.

Weismayer, C., J. G. Anderson, and A. Wolk. 2006. Changes in the stability of dietary patterns in a study of middle-aged swedish women. *The Journal of Nutrition* 136, (6) (Jun): 1582-7.

CHAPTER 2

Estimating Dietary Patterns: Comparing Factor Analysis and Latent Class Analysis

2.1 Abstract

Traditional methods to derive dietary patterns (DP) include factor and cluster analysis. Latent class analysis (LCA) classifies individuals into mutually exclusive groups such that within groups diets are similar. The aim was to derive DP using LCA on 105 food-items and compare them to those derived using factor analysis. We tested how well the resulting classes were characterized by the factor scores, and compared subjects' direct classification from LCA on food-items versus two *a posteriori* classifications from factor scores: one possible classification using tertiles and a two-step classification using LCA on previously derived factor scores. Methods were illustrated using the Pregnancy, Infection and Nutrition Study, North Carolina, 2000-2005 (n=1,285 women). We found that food-items were grouped into four DP: 'Prudent', 'Prudent with alcohol and coffee', 'Western', and 'Southern'. Women were classified into three classes of approximately the same size: 'Prudent', 'Hard core Western' and 'Health conscious Western'. There was high agreement κ=0.70 (95% confidence interval 0.66, 0.73) between the direct classification from LCA on food-items and the one from the two-step LCA on factor scores. By contrast, there was poor agreement with the classification based on tertiles. The two-step classification which adds the benefit of having factor scores seems promising.

2.2 Introduction

One goal of deriving dietary patterns (DP) is to study the effects of overall diet on health outcomes as opposed to the effects of individual nutrients or foods, which maybe highly correlated (Hu, FB, 2002). DP are latent variables; while not directly observed, they can be measured with a dietary intake instrument and empirically derived using statistical methods. The predominant methods to derive them are factor and cluster analysis (Newby, P.K., 2004). Fahey MT *et al.* (2007) showed empirically that generalized latent class models (LCM) could be more appropriate to derive DP than traditional cluster analysis by allowing different outcome distributions, correlated measurement errors, and adjustment for energy intake and other covariates. Few studies (Bailey,R.L. 2006; Costacou,T. 2003; McCann S.E. 2001; Newby P.K. 2004; Velie E.M. 2005) have compared different methodologies using the same data. Only one other study (Padmadas S.S. 2006) has used latent class models to derive DP, and none have compared them to traditional methods. The aim of this paper is to compare subjects' classification into DP using factor and latent class analysis. Methods will be illustrated using data from the third cohort of the Pregnancy, Infection and Nutrition (PIN) Study.

2.3 Methodological Decisions to Derive Dietary Patterns

Nutritional epidemiologists have considered DP in both continuous and categorical scales. Principal components and exploratory factor analysis (EFA) are the predominant methods for deriving them in a continuous scale. Both methods group food-items according to the degree to which they are correlated with each other, and subjects have a score for each DP. By contrast, cluster and latent class analysis classify individuals into mutually exclusive groups (unknown *a priori)* such that within groups diets are similar. Even when DP are derived as continuous, usually investigators classify subjects based on joint classification of the factors to estimate the risk of the outcome for each group compared to a referent. In practice, factors are categorized by quantiles, and subjects are classified according to their cross-tabulation (Hu,F.B. 2002; Knudsen,V.K. 2007). Newby and Tucker (2004) review 93 studies that used principal components, factor or cluster analysis to derive dietary patterns. Here we summarize important statistical decisions relevant to FA and briefly describe LCA.

2.3.1 Factor analysis

Factor analysis postulates a statistical model to explain the correlations between many observed variables by a few underlying but unobservable (*latent*) variables called *factors* (Bollen K.A. 1989). In exploratory factor analysis the relationship between the observed and the latent factors is not specified in advance; whereas in confirmatory factor analysis (CFA) the model is specified *a priori*. Some advantages for using CFA are the ability to account for correlated errors, test if factors are uncorrelated, adjust for covariates and assess goodness-of-fit. In practice, EFA is conducted first to suggest the number and characterization of the DP, and then CFA to test hypotheses and assess goodness-of-fit.

Most empirical DP are derived using EFA with the principal components method of estimation to provide a unique factor score solution and using Varimax method of orthogonal rotation to facilitate interpretability by making factor loadings closer to 0 or ±1 rather than intermediate. Orthogonal rotation also simplifies future analyses, such as avoiding collinearity when using factor scores as covariates in regression models or allowing analysis as independent outcomes. CFA for ordinal outcomes can be estimated using the user's Stata program GLLAMM (Generalized Linear Latent and Mixed Models) (Rabe-Hesketh,S. 2004) or Mplus (Muthen L.K. and Muthen B.O. 1998-2006).

2.3.2 Latent class analysis

Latent class models (Rabe-Hesketh,S. 2007) classify subjects into unknown *a priori* classes such that within classes subjects are similar. LCM are specified as a finite mixture model (McLachlan G.J. 2000) of conditional densities given the class and are usually estimated using maximum likelihood via the Expectation-Maximization algorithm (Dempster A.P. 1977). A LCM for categorical outcomes and covariates estimates two sets of parameters: regression coefficients predicting class membership and conditional probabilities of the observed responses given the class. In contrast to cluster analysis, in LCM each subject has a predicted probability for belonging to each class. The most common way to classify subjects into a specific class is to assign them to the one with the highest probability of class membership.

2.3.3 Number of dietary patterns

There is no single best way to select the number of DP. When using EFA, the DP literature most often keeps the meaningful factors by visual inspecting the loadings in combination with eigenvalues above one (Kaiser's rule), or those before the Cattell's Scree plot (eigenvalues vs. number of factors) starts to flatten, indicating that there is no gain in explained variance by adding another factor. Similarly, in LCM there is no single accepted statistical test or fit-statistic to determine the number of latent classes. The usual likelihood ratio test (LRT) cannot be used to compare nested latent class models because the regularity conditions required in classical maximum likelihood theory are violated and hence, its distribution is not chi-square. Two approximations to the LRT are the Lo-Mendell-Rubin LRT (LMR-LRT) (Lo *et al*, 2001) and the bootstrap LRT (B-LRT) (McLachlan, GJ. 2000). Another way to compare models with different number of classes is with the Bayesian Information Criterion (BIC). A recent simulation study (Nylund, K.L. 2007) showed that the B-LRT performed better in identifying correctly the number of classes than the LMR-LRT and the BIC. However, some disadvantages of the B-LRT are requirement of large sample sizes, increase in computation time, and lack of robustness to model misspecification. Deciding on the number of classes does require care, because spurious latent classes can be accommodating non-normality rather than discovering subpopulations (Bauer DJ and Curran PJ, 2004).

2.4 Analysis of Dietary Patterns for Women in PIN

2.4.1 Study sample and dietary intake assessment

We used data from the third cohort of the Pregnancy, Infection and Nutrition (PIN) Study (December 2000 to June 2005). The study recruited pregnant women seeking services from prenatal clinics at University of North Carolina Hospitals. Study protocols were reviewed and approved by institutional review boards of the University of North Carolina. A total of 1,875 women (2,006 singleton pregnancies) were enrolled fulfilling the minimum age of 16 years and being less than 20 weeks' gestation, from which 1,352 women (1,442 pregnancies) had complete dietary data. For this analysis, only one pregnancy was randomly selected when a woman had several pregnancies with complete dietary assessments. The average age was 29.5 ± 5 years (range 16 to 47), 78.5% were married, 17.8% had ≤12 years of education, half were nulliparous, 74.4% were white and 15.9% black. Based on the categories established by the Institute of Medicine guidelines and using pregravid weight 14.3% were underweight (body mass index<19.8 kg/m^2), 52.6% normal (19.8–26.0), 10.5% overweight (>26.0-29.0) and 22.6% obese (> 29.0).

Diet intake was assessed through a self-administered semiquantitative 119 food-item Block food frequency questionnaire (FFQ) (Block G, 1992) to measure usual intake in the past three months. It was administered at 26-29 weeks' gestation to reflect diet during the second trimester. Dietsys+Plus version 5.6 with an updated food composition table based on nutrient values from the National Health and Nutrition Examination (NHANES) III and the United States Department of Agriculture (USDA) 1998 nutrient databases was used to calculate daily energy intake in kilocalories and grams/day. From the 1,352 women with complete dietary data, we excluded those with daily energy intakes below the 2.5th or above the 97.5th percentiles (1,000 and 4,765 kcal, respectively) as an attempt to exclude implausible energy intakes, leaving 1,285 women for the analysis.

The number of FFQ food-items to derive the DP was reduced from 119 to 105. Nine food-items (**Table A1**) were rarely consumed (<10% consumption). Alcoholic drinks (beer, spirits, wine) and low-fat milks (skim, 1% and 2%) were combined into two groups due to very small counts. Whole milk was excluded because the FFQ allowed selecting only one type of milk. Given that many food-items' distributions were skewed and had a lump at zero due to non-consumers, the indicators were categorized **(Table A1)**. Most were categorized into a three-level variable: non-consumers (g/day = 0) and below or above the median of consumption among consumers (g/day>0) to distinguish "low" and "high" consumption. Eleven food-items, like water and green salad, were dichotomized as below or above the median because there were too few non-consumers. Nine food-items, like meat substitute or alcohol, were dichotomized as consumed or not because there were too few consumers.

2.4.2 Statistical methods

We derived continuous and categorical dietary patterns. For continuous DP we first conducted an EFA on 105 ordinal food-items. We estimated factor loadings using weighted least squares, and factors were derived orthogonal using Varimax rotation. We decided the number of DP from a combination of the Scree-plot and the interpretation of the factors. DP names were given according to the foods with higher loadings and also based on names previously used in the literature. Second, we performed CFA on the DP derived by EFA including only food-items with loadings in absolute value ≥ 0.25, allowing food-items to load on multiple factors. We specified correlated errors between coffee and cream and iced tea and sugar/honey because the FFQ asked specifically if these condiments were usually added to these drinks. We fitted a confirmatory factor model with correlated factors to test if after constraining some of the loadings to zero, factors were still orthogonal. We adjusted for energy intake, parity, smoking status, education, age and race, and assessed goodness-of-fit with the root mean square error of approximation (Skrondal A, 2004). We studied associations between factor scores and nutrients with partial Spearman correlations adjusting for energy intake.

To determine mutually exclusive groupings, we used LCA to derive categorical DP including only ordinal food-items with EFA loadings ≥ 0.25. First, we determined the number of classes using the LMR-LRT. Next, the model was adjusted for energy intake and covariates. We interpreted and named the classes from the conditional probabilities of consumption. Finally, we compared nutrient intake between classes using the Mann-Whitney test.

We used two approaches to compare DP derived by FA and LCA. The goal for the first approach was to describe how well resulting classes could be characterized on the basis of the factor scores. To do so, we compared factor score means among latent classes.

The second approach examined whether subjects' direct classification into DP using LCA agreed with their classification using factor scores. Since FA does not classify subjects directly, we classified them *a posteriori* using a LCM on the four continuous factor scores. We assumed conditional independence given the class, and different factor means and variances by class. For this *ad hoc* two-step procedure we determined the same number of classes obtained directly from the LCM on food-items. For comparison purposes, we also classified women by cross-tabulating the four factor scores' tertiles. Because DP membership is unknown, we cannot test which classification is best but only test whether the direct classification using LCA on food-items and the two *a posteriori* classifications from factors agree or not. Agreement was assessed with the weighted Kappa statistic.

Statistical analyses were performed using SAS/STAT software, Version 9.1 of the SAS System for Windows (SAS Institute Inc 2002-2003), the procedure PROC LCA (Lanza S.T., 2007), and Mplus, Version 5.1 (Muthen L.K. and Muthen B.O., 1998-2006).

2.5 Results

2.5.1 Factor analysis

According to the Scree-plot from EFA, after the fourth factor, factors did not contribute much to explain the variance of the data (the first six eigenvalues were 10.22, 8.66, 4.36, 3.19, 2.62 and 2.59). One factor loaded high (>0.25) in many fruits and vegetables, whole grains, yogurt, vegetable soup and beans; it was called **'FA-Prudent' (Table 2.1)**. A second factor loaded high on processed meat, hamburger, French fries, soft drinks, and Southern foods like coleslaw, corn, collards, green beans, fried chicken and fish, pork, corn bread and iced tea; it was called **'FA-Southern'.** A third factor loaded on green salad and dressing, tomatoes, broccoli, spinach, fish not fried, whole grains, coffee and alcohol; it was called **'FA-Prudent with coffee & alcohol'**. The fourth factor loaded high in fast food (hamburger, French fries, pizza, cheese dish, burritos), salty snacks and sweets (doughnuts, cookies, cake); it was called **‘FA-Western’**. Most food-items loaded only on one factor, 12 loaded on two factors, and three (yogurt, whole-wheat bread, and Kool-Aid) loaded on three factors. Seven food-items (cheese, eggs, non-fortified cereal, pudding, orange juice, diet soft drinks, and butter) with EFA loadings <0.25 for all factors were excluded from CFA and LCA.

The overall test for the correlations between the four factors being zero was significant (P value < 0.0001), and this model had a slightly better fit than the one with uncorrelated factors. The highest correlation (r=0.49) was between ‘FA-Southern’ and ‘FA-Western’, and 0.38 between ‘FA-Prudent’ and ‘FA-Prudent with coffee & alcohol’. Although significant, the correlation between ‘FA-Prudent’ and ‘FA-Western’ was much smaller (r=0.17). The other correlations were not significant. The correlated errors between coffee and half & half, and iced tea and sugar/honey were significant. A simplified path diagram for the final model is presented in **Figure 2.1**. Factor loadings from EFA and CFA were similar (**Table 2.1, Table A2)** except for French fries and hamburger for ‘FA-Western’ and real fruit juice excluding orange juice for ‘FA-Southern’. The food-items that were better explained (R^2>0.4) by the factors were: green salad, fried chicken, bacon, whole wheat bread, and meat substitutes.

The 4-factor model adjusted for energy intake, nulliparous, smoker, white, education and age was significantly better (P value < 0.0001) than the one without covariates. Nulliparous women scored significantly higher in ‘FA-Prudent’ compared to multiparous and lower in ‘FA-Southern’ and ‘FA-Western’ (**Table A3**). White and more educated women scored significantly higher in ‘FA-Prudent’ and significantly lower in ‘FA-Southern’. Older women scored higher in both ‘FA-Prudent’ scores. Overweight and obese women had significantly higher scores of ‘FA-Southern’ than normal weight women. ‘FA-Prudent’ and ‘FA-Prudent with coffee & alcohol’ were positively correlated with fiber, iron, folate, calcium and vitamins, and percent of calories from protein, but ‘FA-Prudent’ was negatively correlated with fat, saturated fat, cholesterol and percent of calories from sweets (**Table A4**). ‘FA-Southern’ was positively associated with fat, saturated fat, cholesterol, and percent of sweets, and negatively correlated with fiber, iron, folate and vitamins. ‘FA-Western’ was highly associated with fat and percent of sweets.

2.5.2 Latent class analysis

We chose the latent class model with three classes (3-LCM) because it was significantly better (P value 0.0109) compared to the 2-LCM, and was not significantly different (P value 0.7475) from the 4-LCM. **Figure 2.2** shows the conditional probabilities of consumption given the class membership for selected food-items with marked differences between classes, and **Figure A1** for all 98 food-items. One class had higher probabilities of consuming more fruits and vegetables, whole grains, baked beans, nuts, fish and chicken (not fried), yogurt, water, and low-fat milk; it was called ‘**LCA-Prudent’**. Women in this class had higher consumption of fiber, folate and vitamins (**Table 2.2**). The second class had high probabilities for consuming higher amounts of fast food, salty snacks and sweets, but also for fruits and vegetables. It was called **‘LCA-Western 1’**, and had significantly higher median percent of calories from fat and sweets compared to the ‘LCA-Prudent’ class, but the micronutrient intake was similar. A third class was less likely to eat fruits, vegetables, yogurt, low-fat milk, coffee, alcohol, nuts and beans and more likely to consume fried fish and chicken, sausages, white bread, and soft drinks. It was called **‘LCA-Western 2’** and had significantly lower micronutrient intake compared to the other two classes but fat intake similar to ‘LCA-Western 1’. With respect to Southern foods, the ‘LCA-Prudent’ class had higher percentages of non-consumers, and there were no differences between the two

‘LCA-Western’ classes. There were 32.6% women in ‘LCA-Prudent’, 32.8% in ‘LCA-Western 1’, and 34.6% in ‘LCA-Western 2’.

White, older and more educated women were more likely to be in the ‘LCA-Prudent’ class than in ‘LCA-Western 2’ (**Table 2.3**). Heavier women were significantly less likely to be in ‘LCA-Prudent. Women with higher energy intake were two to three times more likely to be in ‘LCA-Western 1’ than in ‘LCA-Western 2’.

2.5.3 Comparison between factor scores and latent classes

Figure 2.3 shows the first approach for comparing DP derived from FA and LCA, in which we focused on latent classes being characterized by the factor scores. The ‘LCA-Prudent’ and ‘LCA-Western 1’ classes had significantly higher means for ‘FA-Prudent’ and ‘FA-Prudent with alcohol & coffee’ factors compared to the ‘LCA-Western 2’ class. The ‘LCA-Prudent’ class had significantly lower means for ‘FA-Southern’ and ‘FA-Western’ factor scores than the ‘LCA-Western 1’ class. The ‘LCA-Western 1’ class had a significantly higher ‘FA-Western’ mean than the ‘LCA-Western 2’ class, and the ‘FA-Southern’ means were not significantly different.

The second approach compared the direct classification into three DP using LCA on the food-items versus the a *posteriori* classification using LCA on the four factor scores. These latter classes were interpreted by comparing the means of the factors (**Figure 2.4**). We called one class **‘2-Step Prudent/Anti-Southern’** because it had means significantly higher to zero for ‘FA-Prudent’ and ‘FA-Prudent with coffee & alcohol’ factors and a negative mean for ‘FA-Southern’ factor. A second class had the highest ‘FA-Western’ mean but also had means significantly higher than zero for ‘FA-Prudent’ and ‘FA-Prudent with alcohol & coffee’; it was called **‘2-Step Western/Prudent’**. Finally, the third class had lower means for ‘FA-Prudent’, ‘FA-Prudent with alcohol & coffee’, and a higher mean for the ‘FA-Western’ factor; it was called **‘2-Step Western’.** There were 33.0% women in the ‘2-Step Prudent/Anti-Southern’ class, 29.3% in the ‘2-Step Western/Prudent’, and 37.7% in the ‘2-Step Western’. Mapping ‘2-Step Western’ to ‘LCA-Western 2’, ‘2-Step Western/Prudent’ to ‘LCA-Western 1’, and ‘2-Step Prudent/Anti-Southern’ to ‘LCA-Prudent’, the percentages of correct classification (diagonal of the contingency table between the two classifications) were 77.6%, 80.2% and 76.4% respectively, and there was high agreement κ=0.70 (95% confidence interval (CI) 0.66, 0.73) between the two classifications.

To illustrate what has been done previously in the literature to classify subjects into DP derived by FA, we categorized the four factor scores into tertiles. The total number of different combinations is 81, which is difficult to collapse into three groups without making any strong subjective decisions. We subjectively collapsed them into the same three groups obtained by the direct classification from LCA: ‘Prudent’, ‘Western 1’ and ‘Western 2’. We classified as ‘Prudent’ those with high or medium tertiles for ‘FA-Prudent’ and ‘FA-Prudent with alcohol & coffee’, and low tertiles for both ‘FA-Southern’ and ‘FA-Western’. The group ‘Western 2’ was defined as those with high or medium tertiles ‘FA-Western’, and low tertiles for both ‘FA-Prudent’ and ‘FA-Prudent with alcohol & coffee’. The remaining 72 combinations were considered ‘Western 1’. The agreement between the 3-LCM on 98 food-items and this particular classification was κ=0.29 (95% CI 0.26, 0.33). The percentages of correct classification were 91.6% (109 of 119), 40.8% (413 of 1,013) and 98.9% (86 of 87) for ‘Western 2’, ‘Western 1’, and ‘Prudent’ respectively.

2.6 Discussion

We found that food-items were grouped into three distinctive dietary patterns among pregnant women from PIN: ‘Prudent’, ‘Western’, and ‘Southern’. In addition, a fourth dietary pattern grouped alcohol and coffee with food-items also considered in a ‘Prudent’ DP.

Further, women were grouped into three classes of approximately the same size, ‘Prudent’ and two types of ‘Western’ diets: a ‘Hard core Western’ and a ‘Health conscious Western’. It seems like there is a group of women commonly in the ‘Western’ DP who because they are pregnant might be making extra efforts to eat fruits and vegetables. They have a high caloric diet with high percent of calories from fat and sweets, but they have a similar micronutrient intake compared to the ‘Prudent’ class. ‘Prudent’ and ‘Western’ patterns have been consistently derived in other populations (Newby P.K., 2004; Kant A.K., 2004), and the ‘Southern’ DP has been reported using the NHANES Survey (Tseng M., 2004). One possible reason for obtaining a ‘Prudent with alcohol and coffee’ DP among pregnant women is that other DP may have underreported alcohol and coffee consumption due to social desirability bias. Another reason is that women in the ‘Prudent’ DP were highly educated, and they could be more aware than women in ‘Southern’ and ‘Western’ DP that occasional and low consumption of alcohol and coffee has not been shown to be harmful to the fetus.

Factor and latent class analysis may derive DP differing in food composition because the former groups food-items that are correlated among each other whereas the latter groups subjects with similar food-items’ intake (Newby P.K., 2004). However, in this population, the DP derived from grouping women into latent classes were well characterized by the DP derived from factor analysis, although the correspondence was not perfect. For instance, we did not find a ‘Southern’ class even though we identified a ‘Southern’ factor. However, people rarely have an exclusive “pure” dietary pattern but rather a combination of different dietary patterns. Indeed, the ‘LCA-Prudent’ class was characterized by a low ‘FA-Southern’ mean and high ‘FA-Prudent’ mean. It was also identified as the ‘Prudent/Anti-Southern’ class from the 2-step a *posteriori* classification. The ’Hard core Western’ and ’Health conscious Western’ classes, had significantly different ‘Prudent’ and ‘Western’ means, but ‘Southern’ means were not significantly different. Similarly, Costacou *et al.*

(2003) found one cluster averaging very high on two principal components and the other very low, with no mean differences for other two principal components.

Even though factors were initially derived to be uncorrelated using EFA, in CFA we found moderate correlations between ‘Southern’ and ‘Western’, ‘Prudent’ and ‘Prudent with coffee & alcohol’, and a low one between ‘Prudent’ and ‘Western’. Factors can be correlated because many of the factor loadings in the model were restricted to zero. Testing if they are correlated is important when factors will be used in subsequent analyses to characterize DP or when they will be derived and used in the same population repeatedly over time. When the factors are to be jointly categorized to derive mutually exclusive DP before further analysis, lack of independence is less of a concern.

The main advantage for using LCM over CFA is classifying subjects into mutually exclusive DP directly as opposed from the joint classification of the factors. When there are only two factors an easy way to classify subjects is to cross-classify the factor scores’ quantiles. However, when there are more factors, LCA avoids making strong subjective decisions for collapsing all possible combinations. We found that there was high agreement between the direct classification from LCA on all 98 food-items and the *a posteriori* one from the two-step LCA on the four factors scores. By contrast, there was a poor agreement with the subjective classification due to the ‘LCA-Western 1’ group, which collapsed all the non-extreme combinations. Our experience suggests that with more than two factors, a subsequent LCM may be superior to “eyeballing” the cross-classification, which may be very time consuming and may not identify the best classification.

The advantage for using the two-step *a posteriori* procedure to classify subjects into DP instead of using LCA directly on the food-items is obtaining also the subject’s factor scores. However, the two-step *a posteriori* classification procedure uses predicted factor scores as outcomes and not fixed variables. This could bias the estimates and the efficiency of standard errors by not taking into account the error in prediction. Potentially we could fit a latent class mixture model (Muthen B. 1999, 2002, 2006) to estimate the factor scores and latent classes simultaneously. This approach also would allow within-class heterogeneity. However, this adds the challenge of determining simultaneously the number of latent classes and factors, and modeling is computational intensive.

Results from this data illustrate that using factor and latent class analysis are complementary. Finding similar dietary patterns provides more confidence that they are robust and enhance their interpretation. The advantage for using FA is finding which foods are eaten in combination, whereas for LCA is classifying the subjects. Ideally we could fit a latent class mixture model to estimate the factor scores and latent classes simultaneously, but fitting the model might not be feasible because of computing time and hence, not yet useful in practice. The proposed *ad hoc* two-step classification using a latent class model on the previously derived factor scores from FA combines both advantages and seems promising.

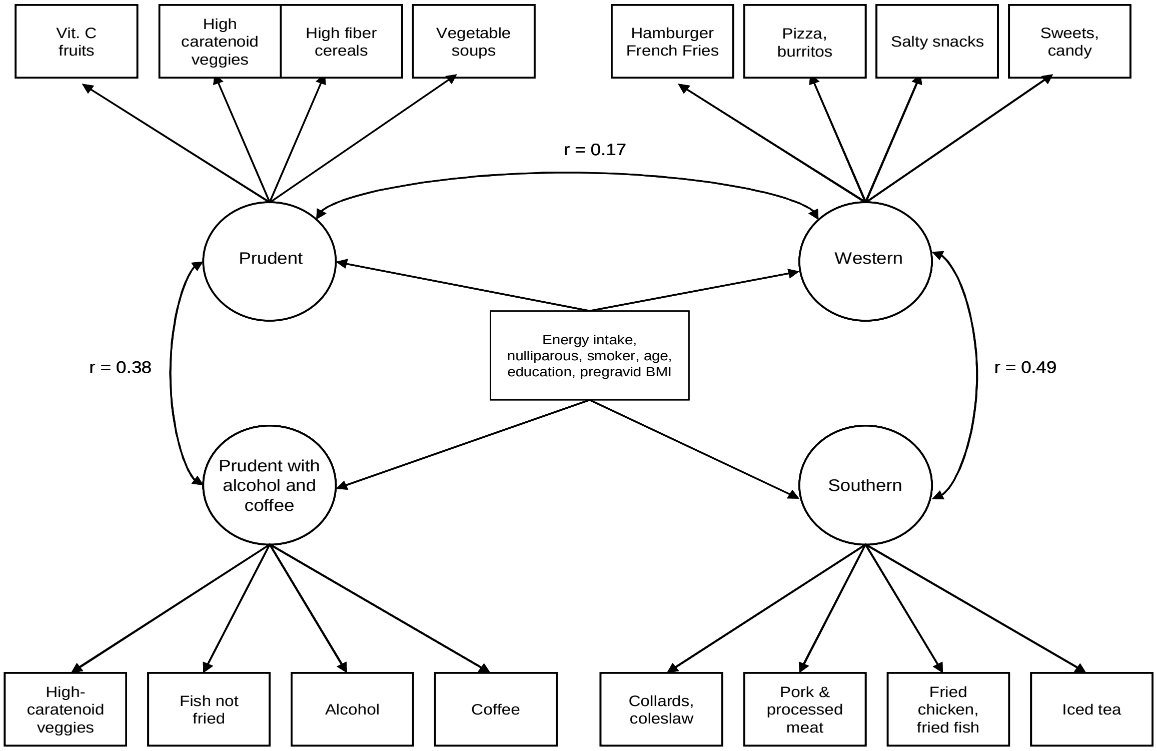


FIGURE 2.1 Simplified path diagram for confirmatory factor model for dietary patterns, PIN Study

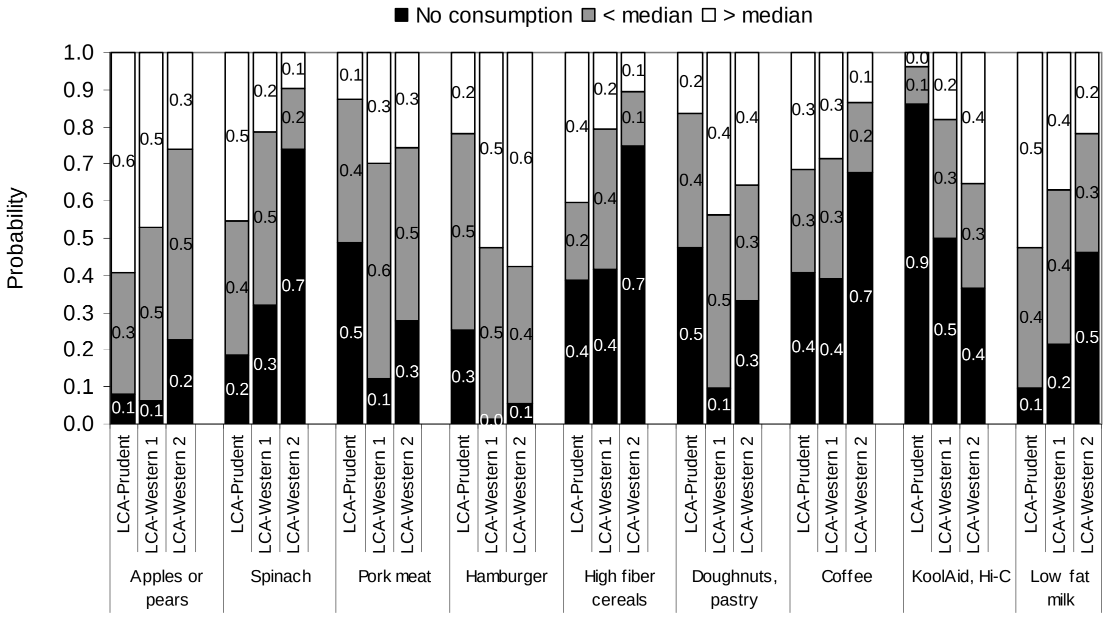


FIGURE 2.2 Probabilities of consumption for selected food-items by latent class, PIN Study

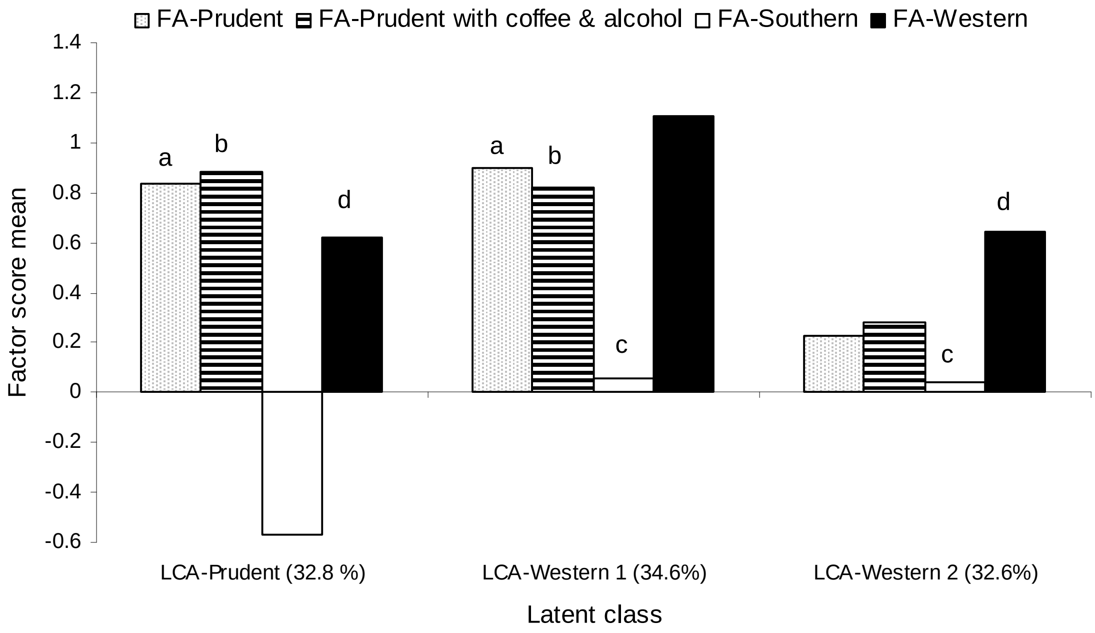


FIGURE 2.3 Factor score means by latent class from 3-LCA on 98 ordinal food-items, PIN Study

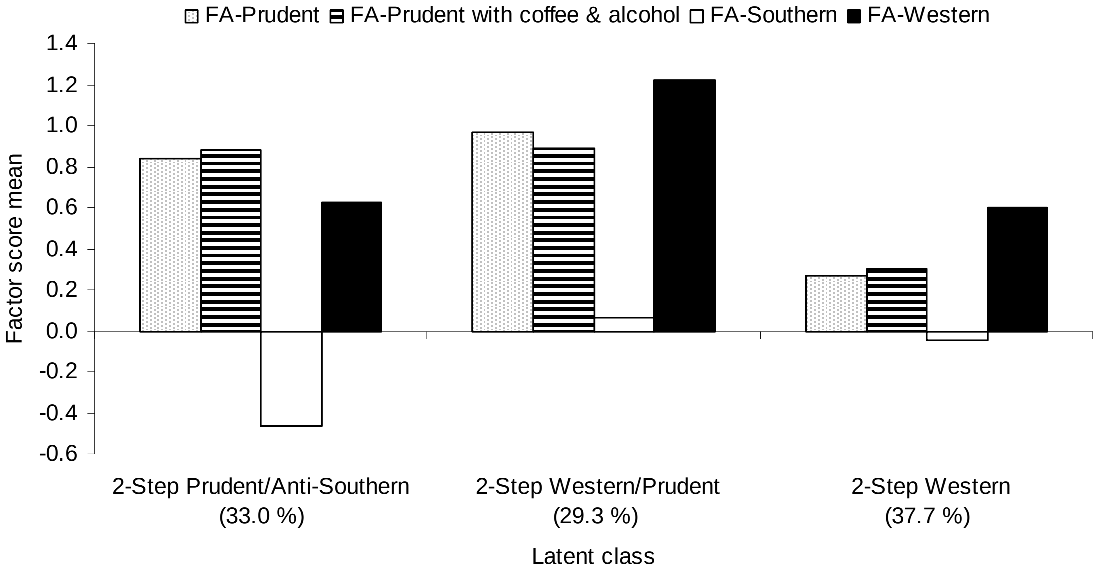


FIGURE 2.4 Factor score means by latent class from 3-LCA on 4 factor scores, PIN Study

TABLE 2.1 Selected exploratory and confirmatory factor loadings for 4-factor model, PIN Study

Food item[a]	**FA-Prudent**		**FA-Southern**		**FA-Western**		**FA-Prudent coffee & alcohol**		**R^2**
	EFA	**CFA**[b]	**EFA**	**CFA**[b]	**EFA**	**CFA**[b]	**EFA**	**CFA**[b]	
Oranges, tangerines	0.47	0.49							0.23
Apples or pears	0.45	0.51							0.24
Coleslaw, cabbage			0.50	0.46					0.20
Greens like collards			0.51	0.41	-0.31	-0.19			0.14
Raw tomatoes							0.54	0.65	0.37
Spinach (cooked)	0.36	0.25					0.50	0.47	0.35
Carrots	0.41	0.40					0.37	0.26	0.30
Green salad							0.66	0.70	0.42
Salad dressing							0.53	0.52	0.25
Yogurt	0.36	0.36	-0.31	-0.25			0.38	0.29	0.36
Low fat milk			-0.41	-0.27					0.07
Baked beans, blackeye p, pintos	0.40	0.43							0.18
Vegetable stew	0.50	0.51							0.24
Beef (roast, steak, sandwiches)			0.46	0.61					0.33
Pork chops, roasts, dinner ham			0.50	0.63					0.34
Ribs, spareribs			0.58	0.62					0.33
Fried chicken			0.63	0.73					0.44
Fried fish			0.48	0.46					0.20
Chicken not fried							0.36	0.29	0.08
Fish not fried							0.56	0.67	0.39
Hot dogs or dinner sausage			0.55	0.66					0.37
Bacon			0.53	0.69					0.40
Breakfast sausage			0.56	0.65					0.36
Meat substitutes (not just soy)	0.40	0.56	-0.52	-0.54					0.53
White bread, French, Ital.,etc			0.40	0.48					0.21
Bagels, Eng. muffins, buns					0.44	0.35			0.12
Dark bread, whole wheat, rye	0.33	0.35	-0.42	-0.36			0.43	0.29	0.40
High fiber cereals	0.30	0.40							0.15
Salty snacks (chips, popcorn)					0.44	0.35			0.12
Ice cream					0.41	0.37			0.13
Doughnuts, pastry					0.43	0.60			0.31
Cake – regular					0.39	0.54			0.26
Coffee							0.36	0.30	0.09
Alcohol							0.32	0.20	0.04
KoolAid, Hi-C,Vit.C-rich drinks			0.53	0.40			-0.38	-0.19	0.19
Drinks w. some juice, Sunny D			0.47	0.46					0.20
French fries, fried potatoes			0.40	0.40	0.34	0.12			0.20
Hamburger, cheeseburger			0.50	0.67	0.34	0.03			0.39
Pizza					0.48	0.38			0.14
Cheese dish like macaroni/cheese					0.41	0.44			0.18
Tacos or burritos					0.49	0.52			0.25

Abbreviations: CFA, confirmatory factor analysis; EFA, exploratory factor analysis; FA, factor analysis.

a The full table with all 105 food items included in EFA is available as a supplementary table in appendix A.

b The confirmatory 4-factor model was adjusted for energy intake, nulliparous, smoker, white, education and age. It included correlated errors between coffee and half & half, and iced tea and sugar/honey. Some factors were correlated; r=0.49 between 'FA-Southern' and 'FA-Western', 0.38 between 'FA-Prudent' and 'FA-Prudent with coffee & alcohol' and r=0.17 between 'FA-Prudent' and 'FA-Western'.

c Sample size was 1,285 women for EFA and 1,219 women for CFA due to missing values in some covariates.

TABLE 2.2 Median nutrient dietary intake by latent class, PIN Study

	LCA-Prudent	**LCA-Western 1**	**LCA-Western 2**	**Overall**
Frequency	400	422	397	1,219
Prevalence	**32.8%**	**34.6%**	**32.6%**	**100.0%**
Total energy, *kcal*	1,865.2^{A}	2,186.7	2,011.5^{A}	2,014.80
Fat, *g*	65.4	79.6	72.1	72.3
Saturated fat, *g*	21.8	26.3^{A}	24.5^{A}	24.0
Cholesterol, *mg*	177.8	230.4^{A}	220.9^{A}	210.1
Omega-3 fatty acids	1.8^{A}	2.0	1.7^{A}	1.8
Fiber, *g*	19.6	18.2	13.5	17.3
Iron, *mg*	14.9^{A}	15.4^{A}	13.4	14.6
Folate, *mcg*	424.5^{A}	406.3^{A}	342.6	393.1
Calcium, *mg*	1,060.4^{A}	984.2^{A}	848.8	970
Vitamin D, *mg*	199.4^{A}	189.4^{A,B}	155.9^{B}	180.8
Vitamin A, *IU*	10,763.7	9,401.8	5,929.2	8,620.1
Vitamin E	9.9^{A}	10.3^{A}	7.6	9.3
Zinc, *mg*	11.0^{A}	11.7^{A}	9.1	10.6
Alpha-carotene	660.7^{A}	595.1^{A}	322.5	538.4
Beta carotene	3,681.1^{A}	3,341.9^{A}	1,917.2	3,045.8
% calories from fat	31.6	33.6^{A}	33.3^{A}	32.9
% calories from protein	15.2	14.1	13.0	14.2
% calories from carbohydrates	55.8^{A}	54.3^{B}	55.2^{A,B}	55.1
% calories from sweets	8.5	11.3^{A}	11.3^{A}	10.2
Number of foods consumed	68	81	64	72

Abbreviation: LCA, latent class analysis.
a The latent class model was adjusted for energy intake, nulliparous, smoker, white, education and age. It included correlated errors between coffee and half & half, and iced tea and sugar/honey. Sample size was 1,219 women due to missing values in some covariates. Classes sharing same upper case letter are not significantly different at 0.0025 level (Bonferroni correction for 20 multiple comparisons within class).

TABLE 2.3 Odds ratios for covariates of 3-LCM on 98 food-items, PIN Study

Covariate	**LCA-Prudent**		**LCA-Western 1**	
	Odds Ratio	**P value**	**Odds Ratio**	**P value**
Nulliparous	1.7	0.011	1.2	0.287
Smoker	0.4	0.053	0.8	0.356
White	3.3	< 0.0001	2.4	0.001
Age, *years*				
25-29	2.7	0.016	2.3	0.005
30-34	8.7	< 0.0001	4.7	< 0.0001
35-47	9.1	< 0.0001	5.7	< 0.0001
Education				
Grades 13-16	3.9	0.003	1.8	0.027
>= Grade 17	11.6	< 0.0001	3.2	0.002
Pregravid BMI				
Low weight	2.1	0.013	1.6	0.117
Over weight	0.4	0.011	0.9	0.706
Obese	0.2	< 0.0001	0.7	0.198
Energy intake				
2nd quartile	1.1	0.687	2.1	0.013
3rd quartile	1.2	0.500	3.4	< 0.0001
4th quartile	0.7	0.386	3.4	0.003

Abbreviations: BMI, body mass index; LCA, latent class analysis.

The reference class is 'LCA-Western 2'.

REFERENCES

Bailey RL, Gutschall MD, Mitchell DC, et al. Comparative strategies for using cluster analysis to assess dietary patterns. *J Am Diet Assoc*. 2006;106(8):1194-1200.

Bauer DJ, Curran PJ. The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychol Methods*. 2004;9(1):3-29.

Block G, Thompson FE, Hartman AM, et al. Comparison of two dietary questionnaires validated against multiple dietary records collected during a 1-year period. *J Am Diet Assoc,* 1992;92(6):686-693.

Bollen KA. Structural equations with latent variables. New York: Wiley, 1989.

Costacou T, Bamia C, Ferrari P, et al. Tracing the Mediterranean diet through principal components and cluster analyses in the Greek population. *Eur J Clin Nutr*. 2003;57(11):1378-1385.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37). *Journal of the Royal Statistical Society.Series B, Methodological.* 1977;39(1):1:38.

Fahey MT, Thane CW, Bramwell GD, et al. Conditional Gaussian mixture modelling for dietary pattern analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2007;170(1):149-166.

Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol.* 2002;13(1):3-9.

Kant AK. Dietary patterns and health outcomes. *J Am Diet Assoc*. 2004;104(4):615-635.

Knudsen VK, Orozova-Bekkevold IM, Mikkelsen TB, et al. Major dietary patterns in pregnancy and fetal growth. *Eur J Clin Nutr.* 2007:1-8.

Lanza ST, Collins LM, Lemmon DR, et al. PROC LCA: A SAS procedure for latent class analysis. *Struct Equ Modeling*. 2007;14(4):671-694.

Lo Y, Mendell NR, Rubin DB. Testing the number of components in a normal mixture. *Biometrika*. 2001;88(3):767-778.

McCann SE, Marshall JR, Brasure JR, et al. Analysis of patterns of food intake in nutritional epidemiology: food classification in principal components analysis and the subsequent impact on estimates for endometrial cancer. *Public Health Nutr.* 2001;4(5):989-997.

McLachlan GJ, Peel D. Finite mixture models. New York: Wiley, 2000.

Muthen B. Beyond SEM: General latent variable modeling. *Behaviormetrika*. 2002;29(1):81-117.

Muthen L.K., Muthen B. *Mplus User's Guide*. Los Angeles, CA: Muthen & Muthen, 1998-2006.

Muthen B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*. 1999;55(2):463-469.

Muthen B. Should substance use disorders be considered as categorical or dimensional? *Addiction*. 2006;101 Suppl 1:6-16.

Newby PK, Tucker KL. Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev.* 2004;62(5):177-203.

Newby PK, Muller D, Tucker KL. Associations of empirically derived eating patterns with plasma lipid biomarkers: a comparison of factor and cluster analysis methods. *Am J Clin. Nutr*. 2004;80(3):759-767.

Nylund KL, Asparouhov T, Muthén BO. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Struct Equ Modeling*. 2007;14(4):535-569.

Padmadas SS, Dias JG, Willekens FJ. Disentangling women's responses on complex dietary intake patterns from an Indian cross-sectional survey: a latent class analysis. *Public Health Nutr*. 2006;9(2):204-211.

Rabe-Hesketh S, Skrondal A. Classical latent variable models for medical research. *Stat Methods Med Res*. 2007;17(1):5-32.

SAS software, Version 9.1 of the SAS System for Windows (SAS, 2002-2003). Copyright © SAS, 2002-2003. SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC.

Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. Boca Raton: Chapman & Hall/CRC, 2004.

Stata programs for estimating, predicting, simulating generalized linear latent and mixed models (GLLAMM) CA, U.C. Berkeley Division of Biostatistics Working Paper Series, 2004. (http://www.gllamm.org/). (Accessed September 25, 2009).

Tseng M, Breslow RA, DeVellis RF, et al. Dietary patterns and prostate cancer risk in the national health and nutrition examination survey epidemiological follow-up study cohort. *Cancer Epidemiol Biomarkers Prev*. 2004;13(1):71-77.

Velie EM, Schairer C, Flood A, et al. Empirically derived dietary patterns and risk of postmenopausal breast cancer in a large prospective cohort study. *Am J Clin Nutr*. 2005;82(6):1308-1319.

CHAPTER 3

Latent transition models to study change in dietary patterns over time

3.1 Abstract

Latent class models (LCM) have been shown empirically to be more appropriate to derive dietary patterns (DP) than cluster analysis since they allow different outcome distributions, correlated measurement errors, and adjustment for energy intake and other covariates. The latent transition model, which is an extension of the LCM to longitudinal data, might be useful to study change as characterized by the movement between discrete DP. Our goal was to identify DP during pregnancy and postpartum, estimate their prevalences, and model the transition probabilities between DP as a function of covariates. Women in the Pregnancy, Infection and Nutrition Study were followed for one year postpartum and diet was assessed in the 2nd trimester, 3 and 12 months postpartum using a 119-item semiquantitative food frequency questionnaire (n=519, 484 and 374, respectively). Foods were aggregated into 29 food-groups (g/day) and categorized into three levels (zero or very low consumption, < median or ≥median). Adjusting for parity and energy intake, three dietary patterns were identified at pregnancy and postpartum: ‘Prudent’, ‘Health Conscious Western’ and ‘Hard Core Western’. Prevalences depend on parity, smoking status, being white and education. First time moms were 4.5 and 2.8 more likely to be in

‘Prudent’ than in ‘Health Conscious Western’ or ‘Hard Core Western’ respectively, and smokers were 2.9 times more likely to be in ‘Hard Core Western’ than in ‘Prudent’. Transition probabilities from one dietary pattern to another were not significantly different between primiparous and not primiparous. Neither the transition probabilities from pregnancy to three months post partum and from three to twelve months post partum were different. The three dietary patterns were very stable; the probability of staying on the same dietary pattern at 3 and 12 months post partum was 0.89, 0.85 and 0.96, for ‘Prudent’, ‘Health Conscious Western’ and ‘Hard Core Western’ respectively. Understanding the different DP during pregnancy and postpartum, and what determines transition among DP, may be important in developing interventions to improve health outcomes.

3.2 Introduction

Literature studying dietary behavior over time, as measured by empirically derived dietary patterns (DP), can be classified according to two goals. The first one is concerned with testing stability of DP because some studies require long periods of follow-up. One hypothesis is that over time the same DP structure can be identified and that DP scores (or subject’s classification) are similar. By contrast, the second goal is quantifying subjects’ changes in DP over time. Some of the challenges when studying DP over time are measuring change in variables that are not directly observed, determining the number of DP and their characterization, and emerging/disappearing DP.

Few studies have studied changes in DP (Newby et al, 2006; Schulze et al, 2006; Meyerrhardt et al, 2007; Togo et al, 2004; Cuco et al, 2006). Only one study (Schulze et al, 2006) studied change in DP in a categorical scale, though they derived DP in a continuous scale and later categorized them using quintiles. Schulze et al (2006) estimated within-subject change in DP by classifying participants according to the cross-tabulation of quintiles of scores at each time point. An alternative approach would be to use a latent transition model to directly classify subjects into mutually exclusive DP at each time point and to estimate the probabilities of changing classes over time.

The latent transition model (LTM) (Collins & Wugalter; 1992, Chung *et al*, 2007) is one extension of latent class models (LCM) (Rabe-Hesketh et al, 2007; Fahey et al, 2007) to longitudinal data in which multiple categorical indicators of the latent class variable (e.g. dietary pattern) are repeatedly measured over T equally spaced time points, and the main interest is to model transition probabilities between latent classes. Traditional LTM involves categorical outcomes and typical assumptions are: 1) responses at each time point are conditionally independent given the class, 2) conditional response probabilities are time-invariant so the characterization of the classes do not change over time, and 3) transition probabilities might need to be time-invariant for model identification. These assumptions may not be realistic when studying dietary patterns from pregnancy to postpartum. For some food-items, the conditional response probabilities (e.g. probability of consuming raw fish given that the women belongs to the 'Health conscious' DP) are not time-invariant because during pregnancy certain foods are craved (desserts and beverages), encouraged (those rich in calcium, iron and folate) or restricted (raw fish, alcohol and caffeine). Second, the probability of switching DP from pregnancy to 3 months postpartum might not be the same that from 3 to 12 month postpartum but may depend on factors such as parity, breastfeeding practices, and weight gain. Lastly, the conditional independence assumption might not be true because correlated errors are expected among foods due to the nature of the food frequency questionnaire (e.g. multi-pass probing and foods organized in sections of food-groups and meal patterns) and due to self-report bias for groups of foods that are perceived as healthy (Kipnis *et al*, 1999). To date, change as characterized by the movement between discrete dietary patterns over time has not been studied using latent transition models. In this paper we review the LTM and illustrate a model selection strategy using data from women in the Pregnancy, Infection and Nutrition (PIN) Study who were followed after delivery.

3.3 The latent transition model (LTM)

The latent transition model (LTM) (Collins & Wugalter, 1992; Chung *et al*, 2007) is one extension of finite mixture models to longitudinal data in which multiple categorical indicators of the latent class variable are repeatedly measured over T equally spaced time points, and the main interest is in modeling transition probabilities between latent classes. LTMs assume time is a discrete process and involve one categorical latent variable for each time point. Whereas in transition models (Diggle et al, 2005) the conditional distribution of each response is modeled explicitly as a function of the previous responses and covariates, in latent transition models the latent class is the one modeled explicitly as a function of the previous latent classes and covariates. **Figure 3.1** shows a simplified path diagram[5] of a latent transition model for the PIN Study.

Let $\mathbf{U}_i$ be a $pT-$dimensional vector of the *i-th* subject's *p* categorical outcomes for all *T* time points, $\mathbf{c}_{it}$ a K-dimensional class-label vector at time $t=1,2,\ldots,T$, and $k_t=1,2,\ldots,K$ the latent class at time t. In contrast to LCM, the *T* latent classes $\left(\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_T\right)$ are not assumed independent, but similarly to LCM the *p* responses at time point *t* are typically assumed conditionally independent given class membership. The measurement part of the LTM with covariates is given by

$$\Pr\left[\mathbf{U}_i=\mathbf{u}_i \mid \mathbf{X}_i=\mathbf{x}_i\right]=\sum_{t=1}^{T}\sum_{k_t=1}^{K}\overbrace{\pi_{i1k_1}}^{\substack{\text{latent class}\\ \text{probability}\\ \text{at time 1}}}\overbrace{\left(\prod_{t=2}^{T}\tau_{itk_t|k_{t-1}}\right)}^{\substack{\text{transition}\\ \text{probabilities}}}\overbrace{\left(\prod_{t=1}^{T}\prod_{j=1}^{p}\rho_{itj|k_t}\right)}^{\substack{\text{conditional}\\ \text{response probabilities}}}$$

[5] By convention, in path diagrams circles represent latent variables, squares observed variables, straight one-headed arrows 'causal' relationships, curved two-headed arrows correlations, and small one-headed arrows random error.

where

$$\pi_{i1k_1} \equiv \Pr\left[c_{i1k_1} = 1 \mid \mathbf{X}_i = \mathbf{x}_i\right]$$
$$\tau_{itk_t|k_{t-1}} \equiv \Pr\left[c_{itk_t} = 1 \middle| c_{i,t-1,k_{t-1}} = 1,\ \mathbf{X}_i = \mathbf{x}_i\right]$$
$$\rho_{itj|k_t} \equiv \Pr\left[U_{itj} = u_{itj} \mid c_{itk_t} = 1\right]$$

The class membership at the first time point, $\mathbf{c}_1$, is modeled with a baseline-category logit model for nominal response with the particularity that $\mathbf{c}_1$ is not observed. Similarly, transition probabilities, $\tau_{itk_t|k_{t-1}} \quad t = 2,\ldots,T$, are modeled using a baseline-category logit model for nominal response. Choosing arbitrarily class *K* as the reference, the structural part of the LTM is given by

$$\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\pi}_{i1}\right)} = \underset{(K-1)\times 1}{\boldsymbol{\alpha}_1} + \underset{(K-1)\times q\times q\times 1}{\mathbf{B}\mathbf{x}_i}$$
$$\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\tau}_{it|k_{t-1}}\right)} = \underset{(K-1)\times 1}{\boldsymbol{\alpha}_t} + \underset{(K-1)\times q\times q\times 1}{\boldsymbol{\Gamma}_t\mathbf{x}_i} \qquad t = 2,\ldots,T$$

where

$$\underset{K\times 1}{\boldsymbol{\tau}_{it|k_{t-1}}} \equiv \left(\tau_{it1|k_{t-1}}, \tau_{it2|k_{t-1}}, \ldots, \tau_{itK|k_{t-1}}\right)^T \qquad t = 2,\ldots,T$$
$$\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\tau}_{it|k_{t-1}}\right)} \equiv \left(\log\left(\tfrac{\tau_{it1|k_{t-1}}}{\tau_{itK|k_{t-1}}}\right), \log\left(\tfrac{\tau_{it2|k_{t-1}}}{\tau_{itK|k_{t-1}}}\right), \ldots, \log\left(\tfrac{\tau_{it,K-1|k_{t-1}}}{\tau_{itK|k_{t-1}}}\right)\right)^T \qquad t = 2,\ldots,T, \quad k_{t-1} = 1,2,\ldots,K$$

In summary, latent transition models estimate three sets of parameters: 1) regression coefficients predicting class membership at the first time point, 2) conditional probabilities of the observed responses given the latent class, and 3) regression coefficients predicting transition probabilities of one latent class to another. Model selection in latent transition models requires determining the number of latent classes and their characterization (i.e. testing whether item-response probabilities are time-invariant), adjusting the latent class membership and transition probabilities for covariates as needed, and testing measurement invariance of transition probabilities. The order of decisions might have an impact on the results, even if the number of classes is predetermined. In addition, identification might be an issue.

3.4 Analysis of the third cohort of the Pregnancy, Infection and Nutrition (PIN) Study

3.4.1 Study sample, dietary assessment and food groups

The data analyzed are from women from the third cohort of the Pregnancy, Infection and Nutrition (PIN) Study who were followed after delivery. Between 2000 and 2005, pregnant women seeking services from prenatal clinics at University of North Carolina Hospitals were recruited for enrolment. From a total of 2,006 pregnancies (1875 women) enrolled for PIN, 1169 women were eligible for the postpartum recruitment, 938 women were asked to participate, and 688 (73.3%) agreed to participate and completed a three-month home interview. There were no significant differences (P value < 0.05) between the women who completed the 3-month interview and those that were excluded or refused, as well as pregravid BMI, parity, bed rest, general health, and total physical activity. There were 571, 545 and 424 pregnancies with complete dietary assessment at pregnancy, three and twelve months postpartum, respectively.

Dietary intake was assessed through a self-administered semiquantitative 119 food-item Block Food Frequency Questionnaire (FFQ) (Block, 1992) to measure usual intake in the past three months. The same dietary instrument was administered at 26-29 weeks' gestation, and at three and twelve months postpartum. Dietsys+Plus version 5.6 with an updated food composition table based on nutrient values from NHANES III and USDA's 1998 nutrient databases was used to calculate daily energy intake in kilocalories, nutrients and grams per day. Pregnancies with daily energy intakes below the 2.5th or above the 97.5th percentiles were dropped as an attempt to exclude implausible energy intakes. For this analysis only one pregnancy per woman was selected based on keeping the pregnancy with the greatest number of completed FFQs. The final sample size was 519, 484 and 374 women at pregnancy, three and twelve months postpartum respectively.

Because the sample size was small relative to the number of outcomes being studied, 104 foods-items were aggregated *a priori* into 29 food-groups according to nutrient content and culinary usage (**Table B1**). From the 119 food-items assessed in the FFQ, fifteen were excluded due to very low consumption (**Table A1**). Given that many food-groups' distributions (g/day) were skewed and had a lump at zero due to non-consumers, the indicators were categorized into three-level variables: non-consumers (g/day below the 5th percentile) and below or above the median of consumption among consumers (g/day>0) to distinguish "low" and "high" consumption. We used the same percentiles for all three time points to make the levels comparable across time, selecting dietary intake at twelve months postpartum (**Table 3.1**) to estimate these cut points because that time point more likely to represent 'typical diet'.

Maternal weight, height, age, education level, race, smoking behavior during pregnancy, and parity were assessed at enrollment through a self-reported questionnaire.

3.4.2 Model selection

In this paper we do not focus on determining the number of latent classes. Instead, we considered the number of classes as 'known' because the sample size was too small (n=519) relative to the number of outcomes being studied (p=29) for comparing latent transition models with more than three classes. We chose three classes based on the results on latent class models to derive dietary patterns using the complete sample of pregnant women in PIN (n=1,285) (chapter 2).

We followed three general steps in order to select the “best” latent transition model. First, we tested measurement invariance across time for certain food-groups, adjusting by energy intake. This tests whether the conditional response probabilities for these groups changed over time. We guided our selection of groups to test by choosing food-groups that were significantly different over time using the correlation statistic test on the categorical variables which accounts for both outcome and time being ordinal, and adjusting for multiple comparisons by Bonferroni’s method. Next**,** we adjusted the model for class membership at pregnancy by energy intake and covariates (maternal age, education level, race, and smoking behavior during pregnancy). Third, we added covariates to model latent transition probabilities and tested if they were time-invariant. Before adding covariates to the transition model, we constrained to zero some transition probabilities that were negligible (< 0.05) (Lanza et al, 2007).

3.5 Results

3.5.1 Identification of dietary patterns

The distributions of the following eight food groups changed significantly (P value < 0.002) from pregnancy to postpartum: vitamin C fruits, other fruits, 100% juice, coffee, alcohol, diet soft drinks, water and fish not fried. During pregnancy, women consumed greater amounts of fruits, 100% juice and water, and less alcohol, coffee, diet soft drinks and non-fried fish. Hence, we first assessed measurement invariance over time. **Table 3.2** presents the models that were compared for model selection. The reduced model with only alcohol and coffee changing over time was preferred by the BIC over the model with eight food-groups changing. However, the omnibus test was highly significant (P value = 0.001), despite the women’s classification being almost identical. In light of these conflicting criteria, we chose the most parsimonious model. We examined visually the conditional response probabilities to interpret the dietary patterns and found that the three classes were meaningful **(Figure B1).** One class had higher probabilities of consuming more fruits and vegetables, whole grains, beans, nuts, fish and chicken (not fried), water, and low-fat dairy; it was called ‘Prudent’. A second class had high probabilities for consuming higher amounts of fast food, salty snacks and sweets, but also for fruits and vegetables. It was called ‘Health Conscious Western’. A third class was less likely to eat fruits, vegetables, and more likely to consume fried fish and chicken, and soft drinks. It was called ‘Hard Core Western’.

First time moms were 4.5 and 2.8 more likely to be in ‘Prudent’ than in ‘Health Conscious Western’ or ‘Hard Core Western’ respectively (**Table 3.3**). Smokers were 2.9 times more likely to be in ‘Hard Core Western’ than in ‘Prudent’. White and more educated women were more likely to be in ‘Prudent’ than in ‘Hard Core Western’. Women with higher energy intake were more likely to be in ‘Health Conscious Western’ than ‘Prudent’. Overall, the percent of women in each of the three dietary patterns was approximately one third (**Figure 3.2**). However, the prevalence depends on parity, smoking status, being white and education.

3.5.2 Transition probabilities

Transition probabilities were not significantly different between primiparous and multiparous women (P value = 0.1690). Furthermore, transition probabilities from pregnancy to three months post partum and from three to twelve months post partum were not significantly different (P value = 0.7253) (**Table 3.4**). The three dietary patterns were generally very stable, with probabilities less than 0.1 of switching dietary patterns from one period to the next (**Figure 3.2**). The ‘Hard Core Western’ was the most stable pattern, with a probability of women staying on the same dietary pattern of 0.96 at any given time. The least stable dietary pattern was ‘Health Conscious Western’ with a probability of 0.85 of remaining in that pattern.

3.6 Discussion

In this paper we studied how latent transition models might be useful to study movement between discrete dietary patterns. In particular, using latent transition models, women from the PIN study were classified into three dietary patterns: ‘Prudent’, ‘Health Conscious Western’ and ‘Hard Core Western’ at pregnancy, 3 and 12 months postpartum. Prevalences depend on parity, smoking status, being white and education. Transition probabilities from one dietary pattern to another were not significantly different between primiparous and not primiparous. Neither the transition probabilities from pregnancy to three months post partum and from three to twelve months post partum were different. The three dietary patterns were very stable; the probability of staying on the same dietary pattern at 3 and 12 months post partum was 0.89, 0.85 and 0.96, for ‘Prudent’, ‘Health Conscious Western’ and ‘Hard Core Western’ respectively. In contrast, in the only study (Greenwood et al, 2003) that has directly classified subjects to study stability of DP, half the women maintained the same DP, and some patterns were more stable than others (κ=0.5 suggesting moderate stability). However, they performed cluster analysis separately at baseline and 5 years later.

Most of the studies concerned on studying changes in dietary patterns over time have used DP in a continuous scale derived using principal components or factor analysis, and have followed three steps: identify the DP, compute DP scores at each time point, and compare DP scores over time. DP are often identified separately at each time point and verified by visual inspection if they have the same number of factors and similar factor loadings over time. Next, DP scores are calculated at each time point either using the time-specific factor loadings (Newby *et al*, 2006 (Vol. 3 and 10); Weismayer *et al*, 2006; Cuco *et al*, 2006) or the same factor loadings for all time points (Togo *et al*, 2004; Mishra *et al*, 2006; McNaughton *et al*, 2007; Northstone and Emmett, 2007). When the goal is to study DP stability, associations between factors over time using time-specific loadings is appropriate. However, when the goal is to study a subject's changes in DP, using time-specific loadings is not correct because the dietary patterns would not be equally measured over time. Similarly, when dietary patterns are considered in a categorical scale they have to be also derived (measured) time-invariant.

Two studies (Cuco *et al*, 2006; Northstone and Emmett, 2007) have studied dietary patterns longitudinally during pregnancy and postpartum to assess stability of DP and to study their associations with predictors and other health behaviors. Both studies derived the DP separately at each time point using principal component analysis, decided by visual inspection that for some dietary patterns the structure was similar, and found one dietary pattern fewer at postpartum. Specifically, Cuco *et al* (2006) identified two DP: 'Sweetened beverages and sugars' and 'Vegetables and meat' at preconception and pregnancy (four time points), but only 'Sweetened beverages and sugars' at six months postpartum. Northstone and Emmett (2007) found both at pregnancy and at four years postpartum the following four DP: 'Health conscious', 'Processed', Confectionery' and 'Vegetarian', but the 'Traditional' DP only at pregnancy. In contrast to our approach, in which women were directly classified into mutually exclusive dietary patterns, these two studies did not investigate the subject's overall dietary pattern as a combination of the derived factors.

There are several advantages of using LTM to study dietary patterns over time. First, LTMs provide a direct classification of the subjects into mutually exclusive DP and by constraining for measurement invariance it guarantees that the same dietary patterns are measured over time. Second, it allows not only estimation of the transition probabilities but also testing what factors determine the transitions over time. One limitation for studying dietary patterns using latent class models is a relatively large sample size required because for each class there are a large number of parameters being estimated for the item response probabilities (food-items). For example, in this application with 27 time-invariant three-level ordinal outcomes and two non-invariant over three time points there were 66 parameters per class (two thresholds for 27 outcomes plus six thresholds for the two non-invariant outcomes). The large number of parameters being estimated could contribute to poor power for testing whether the transition probabilities were different by primiparous (model 6 vs. model 5) and whether the transition probabilities were the same over time (model 7 vs. model 5).

DP can change over time for different reasons such as nutritional advice, changes in food supply or major life events like pregnancy and motherhood. Understanding the different dietary patterns during pregnancy and the first year postpartum, and what determines transition among dietary patterns, could help create more effective interventions during pregnancy, which could be an excellent period to modify or improve health behaviors that should be maintained over time.

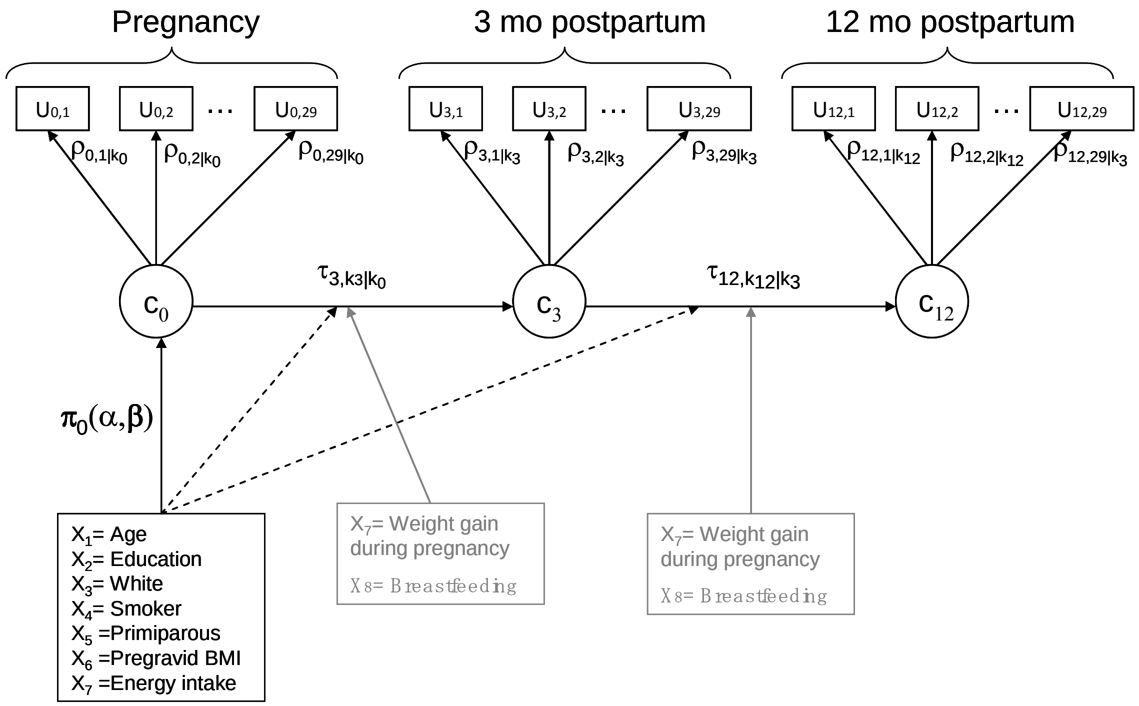


FIGURE 3.1 Path diagram for latent transition model, PIN Study

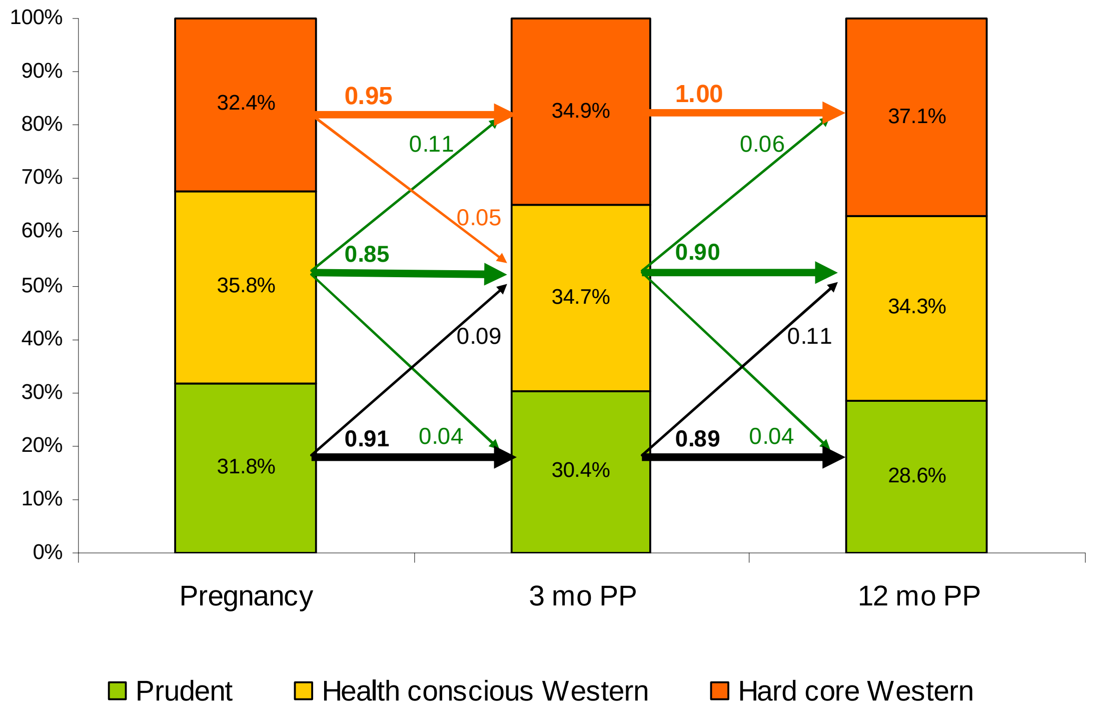


FIGURE 3.2 Dietary patterns' distribution by time and transition probabilities, PIN Study

TABLE 3.1 Food-group median consumption among consumers

at twelve months postpartum, PIN Study

Food group	Freq	Median g/day
Vitamin C fruits	247	10.1
Other fruits	373	97.6
Vegetables	373	82.5
High caratenoid vegetables	371	50.4
High fat dairy	366	18.2
Low fat dairy	350	232.2
Nuts	329	9.7
Beans	336	13.7
Mixed dish w/meat	372	93.3
Eggs	351	7.7
Beef	272	6.5
Pork	278	8.2
Fried chicken or fried fish	222	8.5
Chicken NOT fried	338	10.8
Fish NOT fried	334	11.9
Processed meat	348	14.9
Refined grains	374	83.6
Whole grains	322	24.3
Salty snacks	360	9.0
Sweets	372	41.9
Real 100% fruit juice	339	240.5
Coffee	236	300.0
Alcohol	249	148.0
Soft drinks	365	495.4
Diet soft drinks	186	284.0
Fast food	373	59.0
Condiments with fat	359	5.4
Salad dressing	350	9.1
Water	365	720.0

TABLE 3.2 Model selection for 3-class latent transition model, PIN Study

Model description	Class membership model	Transition probabilities model	# Parame-ters	BIC	log likelihood	Models compared	P value
I. Food-groups changing over time							
M0 None	KCAL€	TIME	194	73,534.5	-36,160.8		
M1 Eight£	KCAL	TIME	290	73,378.4	-35,782.7	M1 vs M0	<0.0001
M2 Coffee and alcohol	KCAL	TIME	218	73,158.4	-35,897.8	M2 vs M1	0.0010
II. Predictors of class membership							
M3 All	KCAL + COVARIATES	TIME	238	70,962.2	-34,740.8	M3 vs M2	0.0000
M4 Only significant	KCAL + COVARIATES	TIME	226	71,028.0	-34,811.1	M4 vs M2	0.0000
III. Predictors of transition probabilities							
M5 Restrict 4 negligible transitions ($\tau = 0$)	KCAL + COVARIATES	TIME&	222	71,014.1	-34,816.5	M5 vs M4	0.2415
M6 Different transitions over time	KCAL + COVARIATES	TIME& | NULLIPAR	230	71,040.6	-34,804.9	M6 vs M5	0.1690
M7 Same transitions over time	KCAL + COVARIATES	Intercept	216	70,984.0	-34,820.2	M7 vs M5	0.7253

£ Vit C fruits, other fruits, fish not fried, 100% juice, coffee, alcohol, diet soft drinks, and water.

N = 519, 484, 374 at pregnancy, 3 and 12 months postpartum.

€ Energy (Kcal/day) was coded as three dummy variables for the three upper quartiles of energy intake.

Covariates were dummy variables for BMI categories, groups of age, education level, white, smoking behavior during pregnancy, and primiparous.

ł BIC from these models cannot be compared to the rest because observations are a subsample due to missing values in covariates.

TABLE 3.3 Odds ratios and 95% CI for predictors of class membership at pregnancy, 3-class latent transition model, PIN Study

Covariate	Prudent	Health conscious Western	Hard core Western (soft drinks & low F&V)	Change in 2 log likelihood	df	P value
		OR	OR			
Nulliparous	1	0.22	0.36	27.7	2	< 0.0001
Smoker	1	1.4	2.9	7.8	2	0.0207
White	1	0.7	0.2	25.5	2	< 0.0001
≥ Grade 17	1	1.0	0.2	33.4	2	0.0496
2nd quartile Kcal at pregnancy	1	3.6	1.0	9.9	2	0.0070
3rd quartile Kcal at pregnancy	1	11.6	1.0	57.0	2	< 0.0001
4th quartile Kcal at pregnancy	1	31.7	1.3	98.7	2	< 0.0001

€ Model 6: Alcohol and coffee distribution allowed to change over time, different transition probabilities over time and some constrained to zero.

TABLE 3.4 Class prevalences and transition probabilities from 3-class latent transition model, PIN Study

	Class prevalence			Transition probabilities					
Dietary pattern		Postpartum		Pregnancy to 3 mo postpartum			3 to 12 mo postpartum		
	Pregnancy	3 mo	12 mo	Prudent	Health conscious Western	Hard core Western[ƚ]	Prudent	Health conscious Western	Hard core Western[ƚ]
M6 Different transitions over time									
Prudent	31.8	30.4	28.6	**0.913**	0.087	0[&]	**0.893**	0.107	0[&]
Health conscious Western	35.8	34.7	34.3	0.039	**0.847**	0.114	0.042	**0.897**	0.062
Hard core Western (soft drinks & low F&V)	32.4	34.9	37.1	0[&]	0.048	**0.952**	0[&]	0.000	**1.000**
M7 Same transitions over time									
Prudent	32.1	30.0	28.1	**0.890**	0.110	0[&]	**0.890**	0.110	0[&]
Health conscious Western	36.0	35.4	34.7	0.040	**0.845**	0.115	0.040	**0.845**	0.115
Hard core Western (soft drinks & low F&V)	31.9	34.7	37.2	0[&]	0.044	**0.955**	0[&]	0.044	**0.955**

€ Alcohol and coffee distribution allowed to change over time, some transition probabilities constrained to zero.

ƚ Hard core Western (high soft drinks and low fruits and vegetables).

& Constrained to zero.

REFERENCES

Chung, H., S. T. Lanza, and E. Loken. 2007. Latent transition analysis: Inference and estimation. *Statistics in Medicine* (Dec 11).

Chung, H., E. Loken, and J. L. Schafer. 2004. Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution. *The American Statistician* 58, (2): 152-8.

Collins, L. M., and S. E. Wugalter. 1992. Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research* 27, (1): 131-57.

Cuco, G., J. Fernandez-Ballart, J. Sala, C. Viladrich, R. Iranzo, J. Vila, and V. Arija. 2006. Dietary patterns and associated lifestyles in preconception, pregnancy and postpartum. *European Journal of Clinical Nutrition* 60, (3) (Mar): 364-71.

Diggle, Peter, P. J. Heagerty, K. Y. Liang, and S. L. Zeger. 2002. Analysis of longitudinal data. Oxford ;New York: Oxford University Press.

Fahey, M. T., C. W. Thane, G. D. Bramwell, and W. A. Coward. 2007. Conditional gaussian mixture modelling for dietary pattern analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170, (1): 149-66.

Kipnis, V., R. J. Carroll, L. S. Freedman, and L. Li. 1999. Implications of a new dietary measurement error model for estimation of relative risk: Application to four calibration studies. *American Journal of Epidemiology* 150, (6): 642.

Lanza, S., and L. M. Collins. 2008. A new SAS procedure for latent transition analysis: Transitions in dating and sexual risk behavior. *Developmental Psychology* 44, (2): 446-56.

McNaughton, S. A., G. D. Mishra, A. M. Stephen, and M. E. Wadsworth. 2007. Dietary patterns throughout adult life are associated with body mass index, waist circumference, blood pressure, and red cell folate. *The Journal of Nutrition* 137, (1) (Jan): 99-105.

Mishra, G. D., S. A. McNaughton, G. D. Bramwell, and M. E. Wadsworth. 2006. Longitudinal changes in dietary patterns during adult life. *The British Journal of Nutrition* 96, (4) (Oct): 735-44.

Muthén, B., and L. K. Muthén (1998-2007). Mplus software, Version 5.1.

Newby, P. K., C. Weismayer, A. Akesson, K. L. Tucker, and A. Wolk. 2006. Longitudinal changes in food patterns predict changes in weight and body mass index and the effects are greatest in obese women. *The Journal of Nutrition* 136, (10) (Oct): 2580-7.

———. 2006. Long-term stability of food patterns identified by use of factor analysis among swedish women. *The Journal of Nutrition* 136, (3) (Mar): 626-33.

Northstone, K., and P. M. Emmett. 2007. A comparison of methods to assess changes in dietary patterns from pregnancy to 4 years post-partum obtained using principal components analysis. *The British Journal of Nutrition* (Oct 5): 1-8.

Rabe-Hesketh, S., and A. Skrondal. 2007. Classical latent variable models for medical research. *Statistical Methods in Medical Research* 17, (1) (Sep 13): 5-32.

SAS software, Version 9.1 of the SAS System for Windows (SAS, 2002-2003). Copyright © SAS, 2002-2003. SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Schulze, M. B., T. T. Fung, J. E. Manson, W. C. Willett, and F. B. Hu. 2006. Dietary patterns and changes in body weight in women. *Obesity (Silver Spring, Md.)* 14, (8) (Aug): 1444-53.

Togo, P., M. Osler, T. I. Sorensen, and B. L. Heitmann. 2004. A longitudinal study of food intake patterns and obesity in adult danish men and women. *International Journal of Obesity and Related Metabolic Disorders : Journal of the International Association for the Study of Obesity* 28, (4) (Apr): 583-93.

Weismayer, C., J. G. Anderson, and A. Wolk. 2006. Changes in the stability of dietary patterns in a study of middle-aged swedish women. The Journal of Nutrition 136, (6) (Jun): 1582-7.

CHAPTER 4

Does food-grouping makes a difference when deriving dietary patterns using Latent Class Models? A Monte Carlo Simulation Study

4.1 Abstract

A key decision involved when deriving dietary patterns is whether or not to collapse the primary dietary, which ranges from 25 to 250 food items depending on the dietary assessment tool, into a smaller number of items (called food groups). One advantage for collapsing is dimension reduction. Another is decreasing the number of non-consumers, as non-consumption leads to a semicontinuous distribution of quantity consumed due to the mass-point at zero for non-consumers. However, one advantage of not collapsing the food items *a priori* is the ability to derive dietary patterns as *a proxy* to dietary behavior in order to better understand which combinations of specific foods are consumed. Also, there is some evidence that food grouping can have an impact on the association between DP and health outcomes showing that greater dietary detail maybe be important. The goal of this paper is to explore via a Monte Carlo simulation study whether food-grouping makes a difference when deriving dietary patterns using Latent Class Models.

4.2 Introduction

Dietary patterns (DP) are commonly derived to study the effects of overall diet on health outcomes as opposed to the effects of individual nutrients or foods (Hu, F.B. 2002). To date, principal components and exploratory factor analysis are the predominant methods for deriving them in a continuous scale. By contrast, latent class models (LCM) have been rarely used (Padmadas *et al*, 2006; Fahey *et al*, 2007) despite that they were recognized as useful methods to reflect complex relations between diet and disease at the international workshop on dietary patterns in 2000 (Hoffmann et al, 2002). Similar to cluster analysis, latent class models classify subjects into classes such that within class they have similar diet, and classes are different from each other.

Regardless of the statistical method used, there are several nutritional methodological issues involved in dietary pattern analysis. For example, whether or not to collapse the primary dietary data, which are measured with 25 to 250 food items depending on the dietary assessment tool, into a smaller number of items (called food groups), how to group the data if collapse is done, and deciding on the number of dietary patterns (Newby and Tucker, 2004).

The main advantage for collapsing is dimension reduction. Another is decreasing the number of non-consumers, as non-consumption leads to a semicontinuous distribution of quantity consumed due to the mass-point at zero for non-consumers. The proportion of habitual non-consumers is naturally greater when working with food-items than with food-groups because after collapsing there is usually a much lower percentage of subjects not consuming any food from the food group. However, one advantage of not collapsing the food items *a priori* is the ability to derive dietary patterns as *a proxy* to dietary behavior in order to better understand which combinations of specific foods are consumed. Also, there is some evidence (McCann *et al*, 2001) that food grouping can have an impact on the association between DP and health outcomes, showing that greater dietary detail maybe be important. Unfortunately, there are not many studies that have compared the performance and effect of deriving dietary patterns with food items and food groups.

Deciding on the number of latent classes is a difficult task. The usual likelihood ratio test (LRT) cannot be used to compare nested latent class models because the regularity conditions required in classical maximum likelihood theory are violated and hence, its distribution is not chi-square. In particular, to compare a model with K classes vs. one with K-1 classes, the reduced model is obtained by restricting the latent class probability to zero, which is a value in the boundary of the parameter space. The Lo-Mendel-Rubin likelihood ratio test (LMR-LRT) (Lo *et al*, 2001) compares neighboring class models using an approximation of the LRT distribution under the assumption of within-class normality conditional on covariates. Another likelihood ratio test is based on the parametric bootstrap to estimate its empirical distribution (McLachlan and Peel, 2000) but is very computationally intensive. A second option to compare models with different number of classes is to use information criteria, such as the Bayesian Information Criterion (BIC). Information criteria are based on the likelihood function, so they reward models that reproduce the observed data, and they usually incorporate a penalty for the number of parameters. The BIC is widely used due to its relatively strong penalty on the number of parameters relative to that of the Akaike Information Criterion (AIC). A recent simulation study (Nylund *et al*, 2007) showed that the bootstrap LRT performed better in identifying correctly the number of classes than the naïve LRT, the LMR-LRT and the BIC for the traditional LCM with few outcomes (8 and 15) with either continuous or categorical outcomes using samples sizes of 200, 500 and 1000. A third option is classification-based information criterion (McLachlan and Peel, 2000) which reward models that produce well-separated classes. For example, the classification likelihood criterion (CLC) uses the estimated entropy to penalize the model for its complexity. The entropy is the directed divergence between the multinomial distribution with posterior probabilities of component membership and the one with equal probabilities (1/k). The estimated entropy cannot be used directly to select the number of classes because it is an increasing function of the number of classes.

The goal of this paper is to explore via Monte Carlo simulation study whether food-grouping makes a difference when deriving dietary patterns using latent class models. In particular, we will compare between deriving dietary patterns using a large number of outcomes (from 80 to 120 food-items) and a medium size number (from 25 to 60) of food-groups in terms of:

1) number of classes (dietary patterns) chosen. We will use two information criteria (BIC and AIC), a statistical test (Lo-Mendell-Rubin likelihood ratio test) and a summary measure of classification (CLC).
2) characterization of the classes (dietary patterns).
3) performance of the estimates. We will assess the parameter bias, standard error bias and parameter coverage.
4) true classification of the subjects. We will predict latent class and compare it to the true class using the weighted Kappa statistic.

4.3 Latent class model considered

The LCM with categorical outcomes and covariates is specified as a finite mixture (McLachlan and Peel, 2000) of conditional response probabilities given that the *i*-th subject belongs to class $k = 1, 2, \ldots, K$. The contribution to the likelihood by subject i is given by

$$\begin{aligned}\Pr\left(\mathbf{U}_i = \mathbf{u}_i \middle| \mathbf{X}_i = \mathbf{x}_i\right) &= \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i) \Pr\left(\mathbf{U}_i = \mathbf{u}_i \mid c_{ik} = 1\right) \\ &= \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i) \prod_{j=1}^{p} \rho_{ij|k}\end{aligned}$$

where $\mathbf{U}_i$ is a p-dimensional vector of categorical random variables (food-items), $\pi_{ik}\left(\mathbf{x}_i\right) \equiv \Pr\left(c_{ik} = 1 \mid \mathbf{x}_i\right)$ is the probability to belong to class k given the covariates $\mathbf{x}_i$, and they add to one, $\mathbf{c}_i$ is a K-dimensional class-label vector where the *k*-th element c_{ik} is defined to be one or zero if the *i-th* subject belongs or not to the *k-th* class, and $\rho_{ij|k} \equiv \Pr\left[U_{ij} = u_{ij} \mid c_{ik} = 1\right]$ is the conditional *j-th* response probability given class k. The number of levels for each outcome is r_j. The class membership $\underset{K\times 1}{\mathbf{c}_i}$ is distributed according to a multinomial distribution, and it is modeled with a baseline-category logit model for nominal response (Agresti, 2002) with the particularity that $\mathbf{c}_i$ is not observed. Mathematically,

$$\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\pi}_i\right)} = \underset{(K-1)\times 1}{\boldsymbol{\alpha}} + \underset{(K-1)\times q\times 1}{\boldsymbol{\Gamma}\mathbf{x}_i}$$

where $\underset{(K-1)\times 1}{\operatorname{logit}\left(\boldsymbol{\pi}_i\right)} \equiv \left(\log\left(\frac{\pi_{i1}}{\pi_{iK}}\right), \log\left(\frac{\pi_{i2}}{\pi_{iK}}\right), \mathrm{K}\ , \log\left(\frac{\pi_{i,K-1}}{\pi_{iK}}\right)\right)^T$ with class K as the reference class. In summary, two sets of parameters are estimated: regression coefficients predicting class membership and conditional probabilities of the observed responses given the class. We estimated the LCM with Mplus (Muthen L.K., Muthen B, 2006), which uses maximum likelihood via the Expectation-Maximization (EM) algorithm (Dempster et al, 1977). During the E-step the *posterior* probabilities of class membership are updated. The M-step involves two multinomial logistic regression optimizations: one to estimate the regression coefficients predicting class membership and another to estimate the intercepts (thresholds) for the conditional response probabilities. *Posterior* probabilities of class membership were calculated via the Bayes' theorem, and subjects were classified into the class with the highest *posterior* probability.

4.4 Model selection

To determine the number of classes we used the Lo-Mendel-Rubin likelihood ratio test LMR-LRT (Lo *et al*, 2001), the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), a sample-size adjusted BIC (Sclove, 1987) and the classification likelihood criterion (CLC). They are defined as:

$$AIC = -2\log L + 2r$$
$$BIC = -2\log L + r\log n$$
$$CLC = -2\log L + 2Entropy \quad \text{where } Entropy = 1 - \frac{\sum_i \sum_k \left(-\hat{p}_{ik} \ln \hat{p}_{ik}\right)}{n \ln K}$$

where *r* is the number of free parameters, *n* the number of independent observations and $\hat{p}_{ik}$ the posterior probability of subject *i* in class *k*. The sample-size adjusted BIC replaces *n* by $n^* = (n+2)/24$. Models with smaller values for the information criteria are preferred. Entropy ranges from zero to one with values closer to one indicating better classification.

4.5 Monte Carlo Simulation Study

4.5.1 Data generation

We generated 1,000 replications for each of four total sample sizes: n=1000, 2000, 5000 and 10000. To generate the data we used as parameters the sample moments for 98 food-items (g/day) from the Pregnancy, Infection and Nutrition (PIN) Study. Because latent class models rely on the assumption of conditional independence given the latent class we generated the data independently given the class. We assumed the true number of latent classes was K=3 and we randomly divided each sample into three classes with equal proportion of belonging to each class (i.e. 0.33). Then given a class, for each subject and food-item we first generated a pseudo-random number from a uniform distribution over the interval 0 to 1 to determine whether the subject consumed the food-item or not. If the food-item was consumed we then generated a pseudo-random number from a Normal distribution with mean and variance of log-transformed g/day, and then transformed back to a log-normal in order to simulate skewed distributions. Hence, every subject belonged to a class and had for each food-item zero grams if not consumed and g/day if consumed. Total amount consumed (g/day) was created by adding all food-items.

Once the food-items were simulated they were aggregated by adding the g/day into food-groups according to nutrient content and culinary usage. We considered different ways of grouping the food-items in order to vary the amount of detail. For example, we could aggregate or not high-caratenoid vegetables with the rest of the vegetables.

Given that many food-items' distributions were intentionally skewed and had a lump at zero due to non-consumers the indicators were categorized, as it is often done in practice. Food-items were categorized into a three-level variable: non-consumers (g/day = 0) and below or above the median of consumption among consumers (g/day>0) to distinguish from "low" and "high" consumption. Food-groups were also categorized into three-level variables, but the level for non-consumers also included those with very low consumption (below the 5th percentile) because for most food-groups there was a very low percent of not consuming the food group after collapsing. Data generation was performed using SAS/STAT software, Version 9.1 of the SAS System for Windows (SAS, 2002-2003).

4.5.2 Data Analysis

For each of the four sample sizes, replication and food-grouping we fitted latent class models with K = 2, 3, and 4 classes using the Monte Carlo facility in Mplus. For all models, we included total amount consumed (g/day) as predictor of class membership using three dummy variables for the 4-level categorical variable created from the quartiles of total g/day. Given that latent class models are sensitive to starting values because there could be several local maxima, we used ten sets of starting values in order to be more confident of reaching the global maximum. Furthermore, in Mplus maximum likelihood optimization was first done over ten random sets of starting values, and then two optimizations were done using as starting values the ending values with the highest log likelihoods from the first optimizations.

4.6 An application to the Pregnancy, Infection and Nutrition (PIN) Study

To compare the dietary patterns derived using all the food-items individually to those derived using food-groups instead we used data from women of the third cohort of the Pregnancy, Infection and Nutrition (PIN) Study. Diet intake was assessed through a self-administered semiquantitative 119 food-item Block food frequency questionnaire (FFQ) (Block et al, 1992) to measure usual intake in the past three months. For this analysis we used 1,285 women and 98 food-items; details of the study sample and the rationale for selecting these particular food-items has been described previously (chapter 2). In that paper, we derived three dietary patterns using a latent class model on 98 categorical food-items. The ‘Prudent’ class had higher probabilities of consuming more fruits and vegetables, whole grains, beans, nuts, fish and chicken (not fried), water, and low-fat dairy. The ‘Health Conscious Western’ class had high probabilities for consuming higher amounts of fast food, salty snacks and sweets, but also for fruits and vegetables. It was called. The ‘Hard Core Western’ class was less likely to eat fruits, vegetables, and more likely to consume fried fish and chicken, and soft drinks.

Food-items (g/day) were aggregated *a priori* into 26 food-groups and the latter categorized into three-level variables: non-consumers (g/day below the 5th percentile) and below or above the median of consumption among consumers (g/day>0) to distinguish “low” and “high” consumption. We fitted a latent class model with 3 classes on the 26 categorical food-groups. By comparing the conditional response probabilities, we found that the classes could be interpreted similarly to those derived using all food-items individually. However, there was only moderate agreement κ=0.48 (95% CI: 0.44, 0.53) between classifying women using 3-LCA on 98 food-items and 3-LCA on 26 food-groups. The percentages of agreement on classification (diagonal of the contingency table between the two classifications) were 63.1%, 77.5% and 62.8% for Prudent, Health Conscious Western and Hard Core Western respectively (**Table 4.3**).

4.7 Discussion

In this paper we explored via Monte Carlo simulation whether food-grouping makes a difference when deriving dietary patterns using latent class models. In particular, for a scenario of 98 food-items aggregated into 26 food-groups, we compared the effect of sample size for selecting the number of classes using two information criteria (AIC and BIC) and a modified likelihood ratio test (Lo-Mendell-Rubin LRT). Either using food-items or food-groups, for all criteria the percentage of times each selected the true number of classes decreased as sample size increased (**Table 4.1**). In all cases, this decrement was because the criteria selected models with at least one extra class. In general, BIC performed better than AIC and adjusted BIC, but slightly worst than the LMR-LRT. The only exception was when using food-groups for smaller sample sizes (n= 1,000 or 2,000), when BIC performed better than the LMR-LRT. Because the data were simulated at the food-item level, the percentages of times the true number of classes is chosen were obviously much larger than those at the food-group level. The power of the LMR-LRT (i.e. the probability of rejecting a LCM with K-1 classes when the true model has K classes) was very high both when using food-items and food-groups (**Table 4.2**). By contrast, the type I error for the LMR-LRT (i.e. probability of rejecting a K-LCM given that the true number of classes is K) was outrageously large when using food-groups, indicating the preference of LMR-LRT for models with at least one extra class.

Results from this simulation, particularly that the percentage of times each criteria selected the true number of classes decreased as sample size increased, are surprising and go on opposite direction as those found by a recent simulation study (Nylund *et al*, 2007). Both studies simulated the data conditionally independent of the class and considered a complex structure for the outcomes, where none of the outcomes had particularly high or low probabilities for a specific class. However, Nylund *et al* considered a small number of outcomes (ten binary outcomes) and sample sizes of n=200, 500 and 1000 and hence, the ratio of subjects per parameter were much higher than the ones we had.

TABLE 4.1 Percentage of times each latent class model is chosen

n	Successful replications for K=4 / Requested	AIC			Adj BIC			BIC			LMR-LRT€			
		Classes (K)			Classes (K)			Classes (K)			Classes (K)			
		2	**3**	4	2	**3**	4	2	**3**	4	1	2	**3**	> 4
98 food-items														
1,000ŧ	1 / 1,000	0.0	**100.0**	NC	0.0	**100.0**	NC	0.0	**100.0**	NC	0.0	0.0	**100.0**	NC
2,000ŧ	0 / 100	0.0	**100.0**	NC	0.0	**100.0**	NC	0.0	**100.0**	NC	0.0	0.0	**100.0**	NC
5,000	71 / 100	0.0	**67.6**	32.4	0.0	**87.3**	12.7	0.0	**97.2**	2.8	0.0	0.0	**98.6**	1.7
10,000	70 / 100	0.0	**64.3**	35.7	0.0	**77.1**	22.9	0.0	**84.3**	15.7	0.0	0.0	**100.0**	0.0
26 food-groups														
1,000	994 / 1,000	0.0	**25.9**	74.1	0.0	**50.5**	49.5	0.0	**87.0**	13.0	4.1	0.1	**44.0**	51.8
2,000	721 / 1,000	0.0	**14.2**	85.9	0.0	**33.6**	66.4	0.0	**62.1**	37.9	1.7	0.0	**40.9**	57.4
5,000	995 / 1,000	0.0	**3.5**	96.5	0.0	**10.1**	90.0	0.0	**20.3**	79.7	0.0	0.0	**33.5**	66.5
10,000	100 / 100	0.0	**0.0**	100.0	0.0	**0.0**	100.0	0.0	**0.0**	100.0	0.0	0.0	**25.0**	75.0

Columns in bold highlight the true number of classes for the latent class model.

ŧ These percentages were calculated including ONLY models with 2 and 3 classes since none of the replications were completed for models with 4 classes.

NC The replications were not completed because "THE MODEL ESTIMATION DID NOT TERMINATE NORMALLY DUE TO AN ILL-CONDITIONED FISHER INFORMATION MATRIX."

€ The LMR-LRT compares a model with K classes vs. a model with one class less. Thus the range of models is from 1 class to 4 classes. We selected the model based on the occurrence of the first non significant p value from the LMR-LRT.

TABLE 4.2 Type I error and power for the Lo-Mendell-Rubin LRT test

N	Type I error	Power
	Ho: 3-LCA (true) vs. H_1:4-LCA	Ho: 2-LCA vs. H_1: 3-LCA (true)
98 food-items		
1,000	NC	1.00
2,000	NC	1.00
5,000	0.01	1.00
10,000	0.00	1.00
# parameters	796	596
26 food-groups		
1,000	0.54	99.9
2,000	0.59	1.00
5,000	0.67	1.00
10,000	0.75	1.00
# parameters	220	164

TABLE 4.3 Cross-classification between 3-LCA on 98 food-items and 3-LCA on 26 food-groups, PIN Study

	3-LCA from 26 food-groups			
3-LCA from 98 food-groups	Prudent	Health conscious Western	Hard core Western	Total
Prudent	**294**	116	56	466
Health conscious Western	77	**303**	11	391
Hard core Western	113	47	**268**	428
Total	484	466	335	1285

REFERENCES

Agresti, A. 2002. Categorical data analysis. New York: Wiley-Interscience

Block G, Thompson FE, Hartman AM, et al. Comparison of two dietary questionnaires validated against multiple dietary records collected during a 1-year period. *J Am Diet Assoc,* 1992;92(6):686-693.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37). Journal of the Royal Statistical Society.Series B, Methodological 39, : 1.

Hoffman, K., M. B. Schulze, H. Boeing, and H. P. Altenburg. 2002. Dietary patterns: Report of an international workshop. *Public Health Nutrition* 5, (1) (Feb): 89-90.

Hu, F. B. 2002. Dietary pattern analysis: A new direction in nutritional epidemiology. *Current Opinion in Lipidology* 13, (1) (Feb): 3-9.

Lo Y, Mendell NR, Rubin DB. Testing the number of components in a normal mixture. *Biometrika*. 2001;88(3):767-778.

McCann SE, Marshall JR, Brasure JR, et al. Analysis of patterns of food intake in nutritional epidemiology: food classification in principal components analysis and the subsequent impact on estimates for endometrial cancer. *Public Health Nutr.* 2001;4(5):989-997.

McLachlan GJ, Peel D. Finite mixture models. New York: Wiley, 2000.
Muthen L.K., Muthen B. *Mplus User's Guide*. Los Angeles, CA: Muthen & Muthen, 1998-2006.

Newby, P. K., and K. L. Tucker. 2004. Empirically derived eating patterns using factor or cluster analysis: A review. *Nutrition Reviews* 62, (5) (May): 177-203.

Nylund KL, Asparouhov T, Muthén BO. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Struct Equ Modeling*. 2007;14(4):535-569.

Padmadas, S. S., J. G. Dias, and F. J. Willekens. 2006. Disentangling women's responses on complex dietary intake patterns from an indian cross-sectional survey: A latent class analysis. *Public Health Nutrition* 9, (2) (Apr): 204-11.

SAS software, Version 9.1 of the SAS System for Windows (SAS, 2002-2003). Copyright © SAS, 2002-2003. SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC.

Sclove, S. L. 1987. Application of model-selection criteria to some problems in multivariate analysis. Psychometrika 52, (3): 333-43.

CHAPTER 5

Conclusions and Future Directions

It has been a matter of preference among epidemiologists whether empirical dietary patterns (DP) are conceptualized, and derived, as continuous (as in principal components or common factor analysis) or as categorical (as in cluster or latent class analysis). The scale of the underlying latent variable (dietary pattern) makes them differ in food composition because they are derived using two different statistical methods. Factor analysis is useful to understand which foods are eaten together, and reduce dimension whereas cluster and latent class analysis are useful to classify subjects in order to estimate the risk of the outcome for each exposure class compared to a reference class. Even when dietary patterns are derived as continuous, usually investigators want to classify the subjects based on the joint classification of the factors. When there are only two factors an easy way to classify subjects is to cross-classify the factor scores' quantiles, as is often done in practice. However, when there are more than two factors, the number of different combinations increases substantially, making it difficult to collapse into K groups without taking any strong subjective decisions. Instead, in paper 1 we propose a two-step method in which we first derive the dietary patterns using factor analysis and then use latent class models (LCM) on the previously derived factor scores to classify the subjects. We found that there was high agreement between the direct classification using LCM on all food-items and the *a posteriori* one from the two-step LCM on the previously derived factor scores. By contrast, there was a poor agreement with the subjective classification due to collapsing all the non-extreme combinations. Hence, we recommend that with more than two factors, a subsequent LCM may be superior to “eyeballing” the cross-classification, which may be very time consuming and may not identify the best classification. The advantage for using the two-step *a posteriori* procedure to classify subjects into dietary patterns instead of using LCM directly on the food-items is obtaining also the subject’s factor scores.

In paper 2, we propose using latent transition models (LTM) to study change as characterized by the movement between discrete dietary patterns. There are several advantages of using LTM to study dietary patterns over time. First, LTMs provide a direct classification of the subjects into mutually exclusive dietary patterns and by constraining for measurement invariance it guarantees that the same dietary patterns are measured over time. Second, it allows not only estimation of the transition probabilities but also testing of what factors determine the transitions over time. However, one limitation for studying dietary patterns using latent class models is the relatively large sample size required because for each class there are a large number of parameters being estimated for the item response probabilities (food-items). Dietary patterns can change over time for different reasons such as nutritional advice, changes in food supply, or major life events like pregnancy and motherhood. Understanding the different dietary patterns during pregnancy and the first year postpartum, and what determines transitions among dietary patterns, could help create more effective interventions during pregnancy, which could be an excellent period to modify or improve health behaviors that should be maintained over time.

Because a key decision involved when deriving dietary patterns is whether or not to collapse the primary dietary data, in paper 3 we explore via a Monte Carlo simulation study whether food-grouping makes a difference when deriving dietary patterns using latent class models. We compare the effect of sample size for selecting the number of classes using information criteria (AIC, BIC and CLC) and a modified likelihood ratio test (Lo-Mendell-Rubin LRT).

Future research for paper 1 includes fitting a latent class mixture model to estimate the factor scores and latent classes simultaneously. This approach also would allow within-class heterogeneity. However, this adds the challenge of determining simultaneously the number of latent classes and factors, and modeling is computational intensive. Future research for paper 2 includes writing a tutorial for a biostatistics audience on latent transition models with a simple example using ordinal outcomes comparing MPLUS and PROC LTM. In addition, we will perform a Monte Carlo simulation on LTM with small sample sizes to study convergence, stability and power for testing group effect (e.g. parity) and/or predictors for transition probabilities. Future research for paper 3 includes considering different ways of grouping the food-items in order to vary the amount of detail. Also, we will compare the effect of the food-groupings in the characterization of the classes (dietary patterns), the performance of the estimates (parameter bias, standard error bias and parameter coverage) and in their classification of the subjects.

APPENDIX A

Supplementary Tables and Figures for Chapter 2

TABLE A1. Food item's distribution, PIN Study

Food item	No consumption	≤ Median	> Median
Vitamin C fruits			
Oranges, tangerines	27.2	44.4	28.5
Grapefruit[a]	73.5		
Other fruits			
Apples or pears	12.8	44.2	43.0
Bananas	13.2	55.6	31.2
Peaches, apricots, fresh	42.9	36.6	20.5
Cantaloupe (year round)	35.9	32.1	32.0
Watermelon	53.4	26.5	20.2
Canned fruit, applesauce, etc.	28.4	46.4	25.2
Strawberries	20.5	46.0	33.5
Other fruits	11.1	48.1	40.8
Vegetables			
Green beans or peas	8.9	56.5	34.6
Corn, fresh, frozen or canned	9.2	47.3	43.5
Coleslaw, cabbage	41.2	33.0	25.8
Green salad[b]		57.7	42.3
White potatoes baked, mashed[b]		61.6	38.4
Other vegetables	26.4	45.1	28.5
High-caratenoid vegetables			
Raw tomatoes	20.9	40.6	38.5
Broccoli	14.2	43.9	41.9
Spinach (cooked)	41.6	33.2	25.2
Greens like collards	65.9	17.5	16.6
Carrots	14.6	43.3	42.0
Sweet potatoes	57.7	21.9	20.4
Dairy			
Cheese and cheese spreads[b]		54.2	45.8
Yogurt	29.6	42.6	27.8
Frozen yogurt/regular	65.8	19.0	15.3
Low fat milk	26.5	36.8	36.7
Soy milk[c]	0.98		
Rice milk[c]	1.00		
Nuts and beans			
Peanut butter	22.4	43.7	33.9

Peanuts, other nuts & seeds	27.2	40.5	32.3
Baked beans, blackeye p, pintos	29.1	46.5	24.4
Chili with beans	53.5	29.3	17.2
Refried beans, bean burritos	43.0	29.2	27.8
Mixed dish w/meat			
Vegetable stew	63.5	19.8	16.7
Spaghetti w/tomato sauce and meat[b]		63.8	36.2
Vegetable soup	31.0	36.1	32.9
Other soups like chicken noodle	30.3	36.1	33.6
Mixed dishes with beef or pork	51.5	24.4	24.0
Pasta salad, other pasta dish	6.8	47.0	46.1
Chicken stew, pot pie	19.6	40.9	39.5
Eggs and meat			
Eggs or egg biscuits[b]		64.4	35.6
Beef (roast, steak, sandwiches)	24.9	41.2	33.9
Liver, liverwurst[c]	0.92		
Pork chops, roasts, dinner ham	28.6	47.9	23.5
Ribs, spareribs	67.7	18.8	13.5
Gizzard, neckbones, chitlins[c]	0.94		
Fried chicken	40.9	30.3	28.9
Chicken not fried	10.5	45.4	44.0
Fried fish	61.6	19.4	19.0
Fish not fried	56.3	26.1	17.5
Tuna casserole, tuna sandwich	47.9	26.5	25.5
Shellfish (shrimp, crab, etc.)	32.1	38.4	29.4
Oysters[c]	0.92		
Processed meat			
Hot dogs or dinner sausage	30.7	38.7	30.6
Ham, boloney, lunch meats	18.8	49.6	31.5
Bacon	24.4	40.4	35.2
Breakfast sausage	46.4	27.2	26.5
Rice or dishes with rice	8.0	50.0	41.9
Refined grains			
White bread, French, Ital.,etc	14.8	43.2	42.0
Cornbread or hush puppies	58.0	23.8	18.2
Cereal excl. fiber or fortified	21.7	46.2	32.1
Cooked cereal or grits	36.7	38.4	25.0
Bagels, Eng.muffins, buns[b]		54.2	45.8
Biscuits, muffins	11.7	53.9	34.4
Pancakes, waffles, Pop Tarts	16.7	47.7	35.6
Tortillas - Corn or flour	31.7	34.8	33.5
Whole grains			
Dark bread, whole wheat, rye	29.8	40.2	30.0
High fiber cereals	51.8	24.2	24.0
Product 19, Total, Just Right[a]	85.9		
Salty snacks and sweets			
Salty snacks (chips, popcorn)[b]		58.3	41.7
Crackers	14.2	45.0	40.8
Ice cream[b]		59.6	40.4

Doughnuts, pastry	30.5	37.6	31.9
Pumpkin pie, sweet potato pie[a]	81.9		
Pie or cobbler	53.5	33.2	13.4
Chocolate candy, candy bars	13.5	50.1	36.3
Candy (not chocolate)	31.8	35.9	32.4
Pudding	62.4	20.2	17.4
Cookies, regular	10.5	51.2	38.3
Cake - regular	22.5	40.9	36.7
Jelly, jam, syrup	16.4	42.0	41.6
Beverages			
Orange juice, grapefruit juice	8.3	46.8	44.8
Tomato juice, V-8 juice[a]	81.1		
Real frt juice excl orange,grft	18.8	40.8	40.4
Coffee	49.9	25.8	24.3
Alcohol[a]	90.5		
Light beer[c]	0.99		
Non-alcoholic beer[c]	0.93		
Drinks w. some juice, Sunny D[a]	73.5		
KoolAid, Hi-C,Vit.C-rich drinks	56.2	23.6	20.2
Soft drinks or Snapple not diet	24.4	40.2	35.4
Tea or iced tea (not herb tea)	33.3	34.9	31.8
Cranberry juice cocktail	54.6	23.0	22.3
Diet soft drinks[a]	69.1		
Breakfast or diet shakes[c]	0.91		
Water[b]		57.4	42.6
Fast food			
French fries, fried potatoes[b]		58.3	41.7
Hamburger, cheeseburger	10.4	45.2	44.4
Pizza[b]		62.4	37.6
Cheese dish like macaroni/cheese	15.9	43.7	40.5
Chinese dishes	22.6	38.8	38.5
Tacos or burritos	21.1	44.3	34.6
Condiments and other food items			
Butter	34.6	39.5	25.9
Margarine	39.5	33.3	27.2
Gravy	53.9	25.4	20.8
Mayonnaise, sandwich spreads	24.2	48.1	27.7
Salad dressing	7.9	56.2	36.0
Salsa, ketchup, taco sauce	15.2	47.2	37.6
Mustard, BBQ sauce, other sauce	13.7	44.0	42.3
Cream or half & half[a]	80.3		
Non dairy creamer[c]	0.94		
Sugar or honey in coffee/tea	43.2	28.7	28.1
Breakfast bars, Power bars	46.3	28.5	25.2
Meat substitutes (not just soy)[a]	76.7		

CFA, confirmatory factor analysis; EFA, exploratory factor analysis; LCA, latent class analysis.

a Food-items were dichotomized as consumed or not because there were too few consumers; only percent of non consumers shown.

b Food-items were dichotomized as below or above the median because there were too few non consumers.

c Food-items were rarely consumed (<10% consumption) and were not included in EFA, CFA and LCA because they did not add any useful information.

TABLE A2. Exploratory and confirmatory factor loadings for 4-factor model, PIN Study

Food item	FA-Prudent		FA-Southern		FA-Western		FA-Prudent coffee & alcohol		
	EFA	CFA[a]	EFA	CFA[a]	EFA	CFA[a]	EFA	CFA[a]	R^2
Oranges, tangerines	0.47	0.49							0.23
Grapefruit	0.42	0.52							0.25
Apples or pears	0.45	0.51							0.24
Bananas	0.29	0.36							0.13
Peaches, apricots, fresh							0.29	0.47	0.20
Cantaloupe (year round)							0.32	0.44	0.19
Watermelon							0.26	0.29	0.08
Canned fruit, applesauce, etc.	0.43	0.39							0.15
Strawberries							0.34	0.41	0.16
Other fruits	0.33	0.24					0.32	0.33	0.23
Green beans or peas	0.27	0.28	0.27	0.24					0.12
Corn, fresh, frozen or canned			0.35	0.38					0.14
Coleslaw, cabbage			0.50	0.46					0.20
Green salad							0.66	0.70	0.42
White potatoes baked, mashed			0.32	0.27	0.31	0.23			0.16
Other vegetables	0.34	0.27					0.41	0.33	0.25
Raw tomatoes							0.54	0.65	0.37
Broccoli	0.35	0.30					0.40	0.35	0.29
Spinach (cooked)	0.36	0.25					0.50	0.47	0.35
Greens like collards			0.51	0.41	-0.31	-0.19			0.14
Carrots	0.41	0.40					0.37	0.26	0.30
Sweet potatoes	0.51	0.57							0.30
Cheese and cheese spreads[c]	0.13		-0.10		0.19		0.11		
Yogurt	0.36	0.36	-0.31	-0.25			0.38	0.29	0.36
Frozen yogurt/regular	0.25	0.39							0.15
Low fat milk			-0.41	-0.27					0.07
Peanut butter					0.32	0.27			0.07
Peanuts, other nuts & seeds	0.36	0.50							0.24
Baked beans, blackeye p, pintos	0.40	0.43							0.18
Chili with beans	0.41	0.46							0.20
Refried beans, bean burritos	0.31	0.46							0.20
Vegetable stew	0.50	0.51							0.24
Spaghetti w/Tom. sauce + meat					0.30	0.23			0.05
Vegetable soup	0.49	0.54							0.27
Other soups like chicken noodle	0.41	0.43							0.18
Mixed dishes with beef or pork			0.35	0.52					0.24
Pasta salad, other pasta dish					0.45	0.39			0.14
Chicken stew, pot pie					0.38	0.49			0.22
Eggs or egg biscuits[c]	0.09		0.22		-0.03		0.12		
Beef (roast, steak, sandwiches)			0.46	0.61					0.33
Pork chops, roasts, dinner ham			0.50	0.63					0.34

Ribs, spareribs			0.58	0.62					0.33
Fried chicken			0.63	0.73					0.44
Fried fish			0.48	0.46					0.20
Chicken not fried							0.36	0.29	0.08
Tuna casserole, tuna sandwich							0.33	0.44	0.18
Shellfish (shrimp, crab, etc.)							0.39	0.39	0.14
Fish not fried							0.56	0.67	0.39
Hot dogs or dinner sausage			0.55	0.66					0.37
Ham, boloney, lunch meats			0.28	0.36					0.12
Bacon			0.53	0.69					0.40
Breakfast sausage			0.56	0.65					0.36
Rice or dishes with rice							0.27	0.34	0.11
White bread, French, Ital.,etc			0.40	0.48					0.21
Cornbread or hush puppies	0.37	0.32	0.32	0.45					0.26
Cereal excl. fiber or fortified[c]	0.06		0.05		0.23		-0.10		
Cooked cereal or grits	0.36	0.37							0.13
Bagels, Eng.muffins, buns					0.44	0.35			0.12
Biscuits, muffins					0.36	0.49			0.21
Pancakes, waffles, Pop Tarts					0.31	0.41			0.16
Tortillas - Corn or flour					0.41	0.35			0.11
Dark bread, whole wheat, rye	0.33	0.35	-0.42	-0.36			0.43	0.29	0.40
High fiber cereals	0.30	0.40							0.15
Product 19, Total, Just Right	0.30	0.40							0.15
Salty snacks (chips, popcorn)					0.44	0.35			0.12
Crackers					0.38	0.34			0.11
Ice cream					0.41	0.37			0.13
Doughnuts, pastry					0.43	0.60			0.31
Pumpkin pie, sweet potato pie	0.42	0.33							0.11
Pie or cobbler					0.29	0.55			0.27
Chocolate candy, candy bars					0.48	0.45			0.19
Candy (not chocolate)					0.40	0.40			0.15
Pudding[c]	0.23		0.16		0.24		-0.11		
Cookies, regular					0.45	0.39			0.14
Cake - regular					0.39	0.54			0.26
Jelly, jam, syrup					0.33	0.37			0.13
Orange juice, grapefruit juice[c]	0.13		0.05		0.00		0.01		
Tomato juice, V-8 juice							0.25	0.38	0.14
Real frt juice excl orange,grft			0.25	0.11					0.01
Coffee							0.36	0.30	0.09
Alcohol							0.32	0.20	0.04
Drinks w. some juice, Sunny D			0.47	0.46					0.20
KoolAid, Hi-C,Vit.C-rich drinks			0.53	0.40			-0.38	-0.19	0.19
Soft drinks or Snapple not diet			0.26	0.25					0.06
Tea or iced tea (not herb tea)			0.38	0.23					0.05
Cranberry juice cocktail	0.26	0.27							0.07
Diet soft drinks[c]	-0.08		-0.11		0.11		0.20		
French fries, fried potatoes			0.40	0.40	0.34	0.12			0.20
Hamburger, cheeseburger			0.50	0.67	0.34	0.03			0.39
Pizza					0.48	0.38			0.14
Cheese dish like macaroni/cheese					0.41	0.44			0.18

Chinese dishes							0.27	0.33	0.10
Tacos or burritos					0.49	0.52			0.25
Butter[c]	0.10		0.11		0.08		0.13		
Margarine			0.26	0.26	0.26	0.16			0.12
Gravy			0.54	0.68					0.39
Mayonnaise, sandwich spreads			0.34	0.28					0.08
Salad dressing							0.53	0.52	0.25
Water							0.27	0.36	0.13
Salsa, ketchup, taco sauce					0.42	0.52			0.24
Mustard, BBQ sauce, other sauce					0.38	0.47			0.20
Cream or half & half							0.32	0.20	0.04
Sugar or honey in coffee/tea			0.38	0.20					0.04
Breakfast bars, Power bars	0.27	0.33							0.11
Meat substitutes (not just soy)	0.40	0.56	-0.52	-0.54					0.53

Abbreviations: CFA, confirmatory factor analysis; EFA, exploratory factor analysis; FA, factor analysis; LCA, latent class analysis.

a The 4-factor model was adjusted for energy intake, nulliparous, smoker, white, education and age. It included correlated errors between coffee and half & half, and iced tea and sugar/honey. Some factors were correlated; r=0.49 between 'FA-Southern' and 'FA-Western', 0.38 between 'FA-Prudent' and 'FA-Prudent with coffee & alcohol' and r=0.17 between 'FA-Prudent' and 'FA-Western'.

b Sample size was 1,285 women for EFA and 1,219 women for CFA due to missing values in some covariates.

c Food-items with EFA loadings < 0.25 for all factors were excluded from CFA and LCA.

TABLE A3. Regression coefficients for 4-factor model, PIN Study

	FA-Prudent		**FA-Southern**		**FA-Western**		**FA-Prudent with alcohol & coffee**	
Covariate	**Estimate**	**P value**	**Estimate**	**P value**	**Estimate**	**P value**	**Estimate**	**P value**
Nulliparous	0.13	< 0.0001	-0.12	< 0.0001	-0.07	0.011	0.03	0.311
Smoker	-0.17	0.004	0.19	< 0.0001	-0.13	0.007	-0.07	0.135
White	0.15	< 0.0001	-0.23	< 0.0001	0.37	< 0.0001	0.00	0.890
Age, *years*								
25-29	0.14	0.006	0.01	0.806	0.06	0.158	0.24	< 0.0001
30-34	0.21	< 0.0001	-0.06	0.116	0.04	0.356	0.37	< 0.0001
35-47	0.18	0.002	-0.03	0.492	-0.06	0.255	0.41	< 0.0001
Education								
Grades 13-16	0.12	0.020	-0.13	0.001	0.21	< 0.0001	0.20	< 0.0001
>= Grade 17	0.22	< 0.0001	-0.25	< 0.0001	0.16	0.002	0.36	< 0.0001
Pregravid BMI								
Low weight	0.07	0.105	-0.09	0.013	-0.07	0.067	-0.06	0.139
Over weight	-0.13	0.011	0.11	0.015	0.02	0.649	-0.01	0.795
Obese	-0.06	0.143	0.14	< 0.0001	-0.02	0.645	-0.13	< 0.0001
Energy intake								
Kcal 2nd quartile	0.18	< 0.0001	0.13	< 0.0001	0.30	< 0.0001	0.15	< 0.0001
Kcal 3rd quartile	0.31	< 0.0001	0.28	< 0.0001	0.51	< 0.0001	0.25	< 0.0001
Kcal 4th quartile	0.42	< 0.0001	0.44	< 0.0001	0.79	< 0.0001	0.33	< 0.0001

Abbreviations: BMI, body mass index; FA, factor analysis.

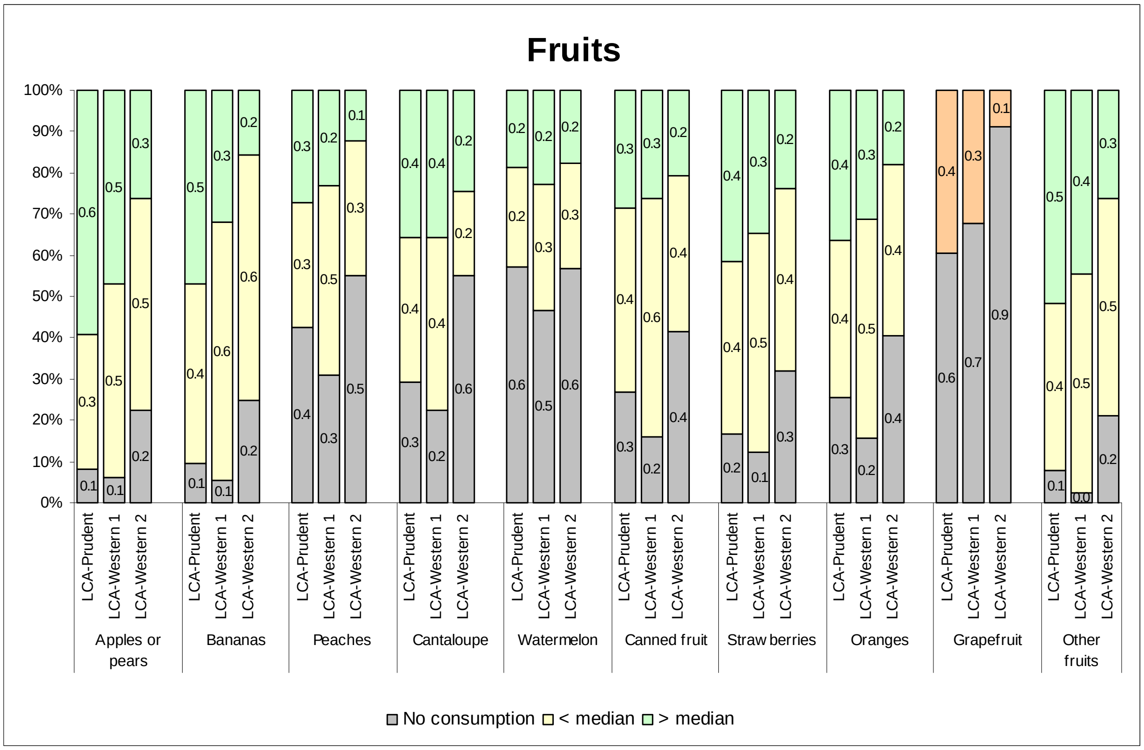
TABLE A4. Partial Spearman correlations between factors and nutrients, adjusted by energy intake, PIN Study

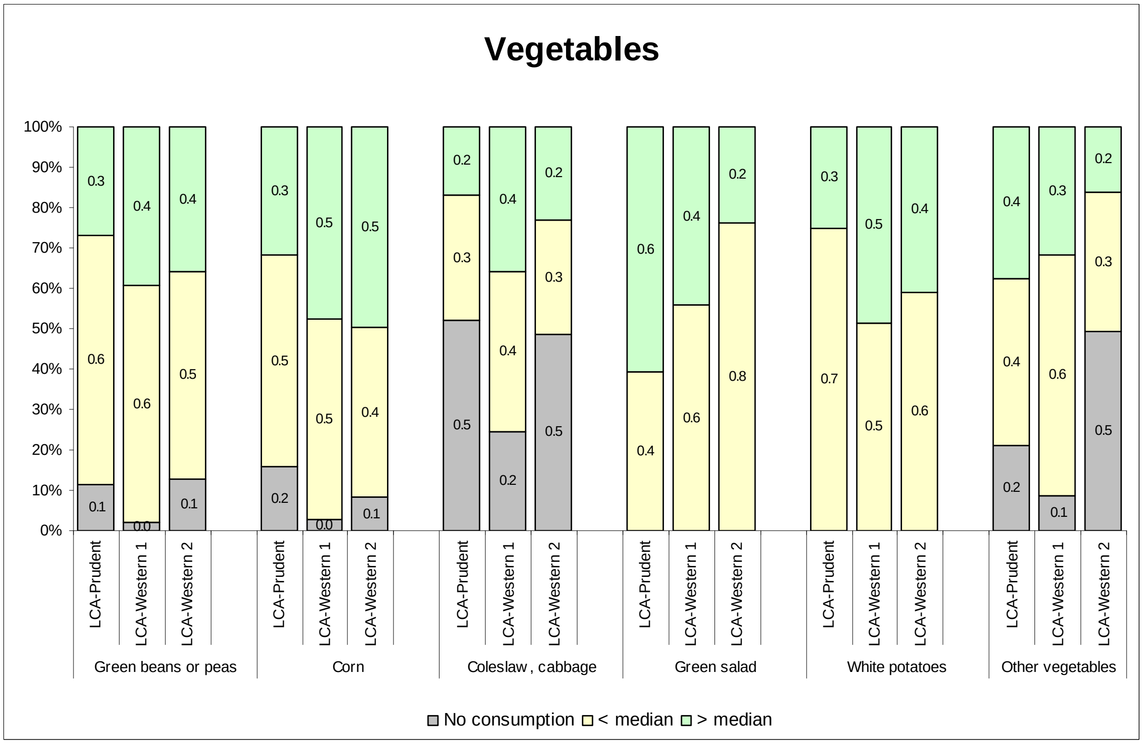
Nutrient	**FA-Prudent**	**FA-Southern**	**FA-Prudent with alcohol & coffee**	**FA-Western**
Fat, *g*	-0.11	0.17	0.04	0.21
Saturated fat, *g*	-0.21	0.21	-0.09	0.11
Cholesterol, *mg*	-0.12	0.29	0.03	-0.04
Omega-3 fatty acids	0.02	0.03	0.21	0.14
Fiber, *g*	0.64	-0.46	0.52	0.01
Iron, *mg*	0.27	-0.25	0.23	0.06
Folate, *mcg*	0.47	-0.40	0.41	-0.01
Calcium, *mg*	0.26	-0.39	0.21	-0.08
Vitamin D, *mg*	0.09	-0.15	0.08	-0.12
Vitamin A, *IU*	0.57	-0.30	0.54	-0.02
Vitamin E,	0.39	-0.31	0.47	0.14
Zinc, *mg*	0.35	-0.23	0.35	0.04
Alpha-carotene	0.53	-0.23	0.40	0.03
Beta carotene	0.55	-0.21	0.53	-0.03
% calories from fat	-0.13	0.18	0.03	0.20
% calories from protein	0.26	-0.22	0.38	0.03
% calories from carbohydrates	0.09	-0.12	-0.11	-0.17
% calories from sweets	-0.17	0.15	-0.20	0.20
% calories from alcohol	0.04	-0.06	0.17	0.06
Number of foods consumed	0.61	0.31	0.45	0.54

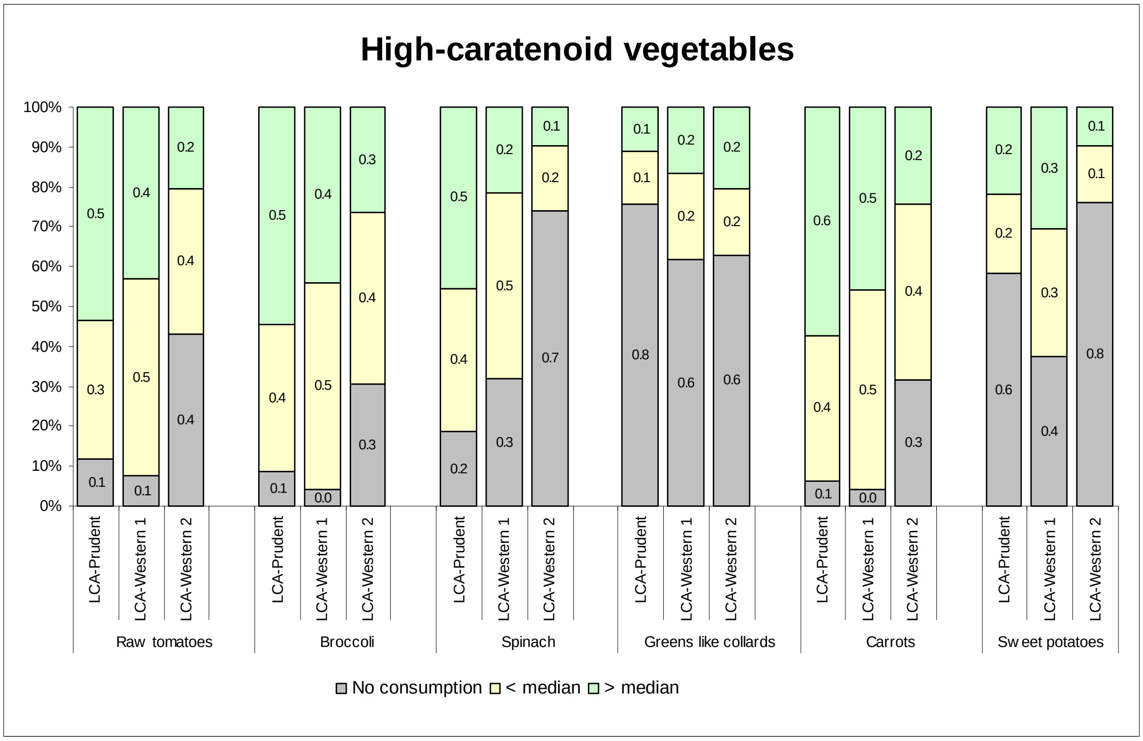
Abbreviation: FA, factor analysis.

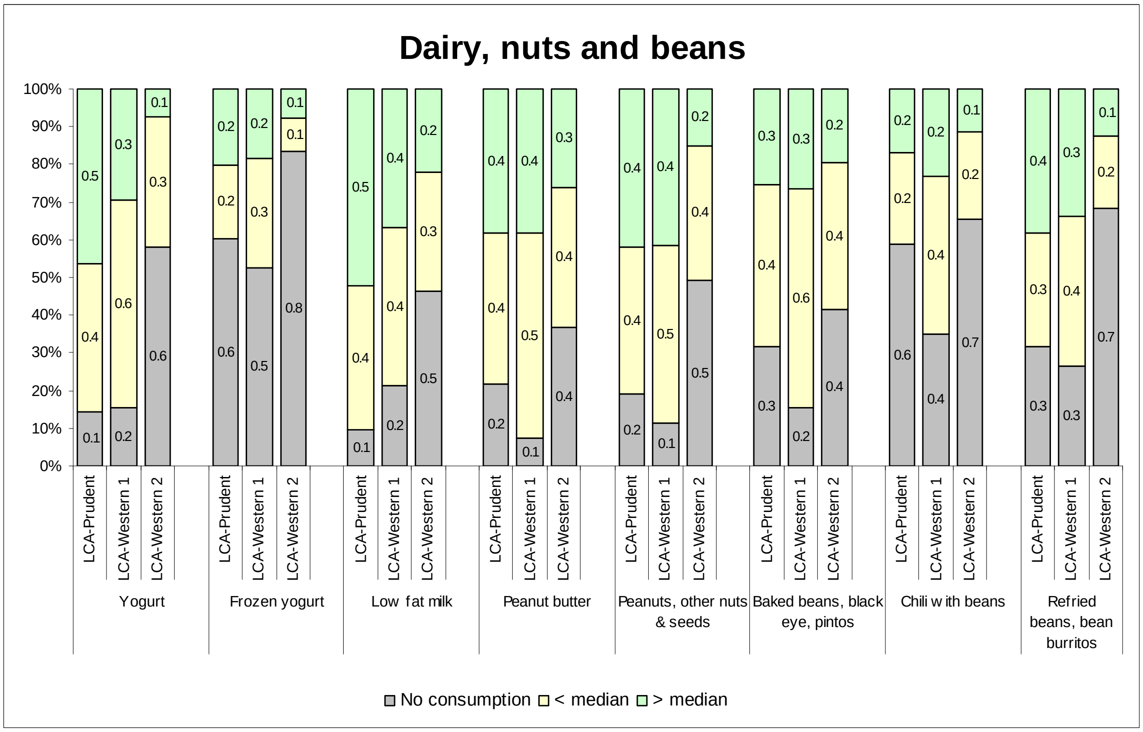
FIGURE A1. Probabilities of consumption by latent class

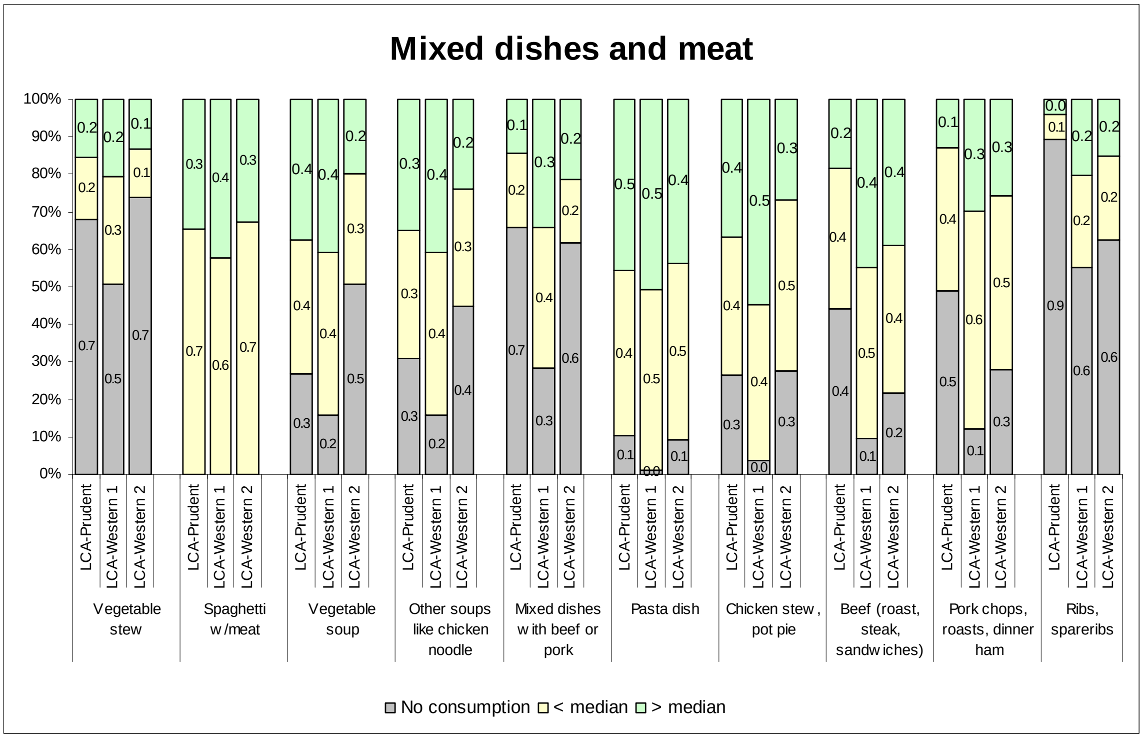
from LCA on 98 food-items, PIN Study

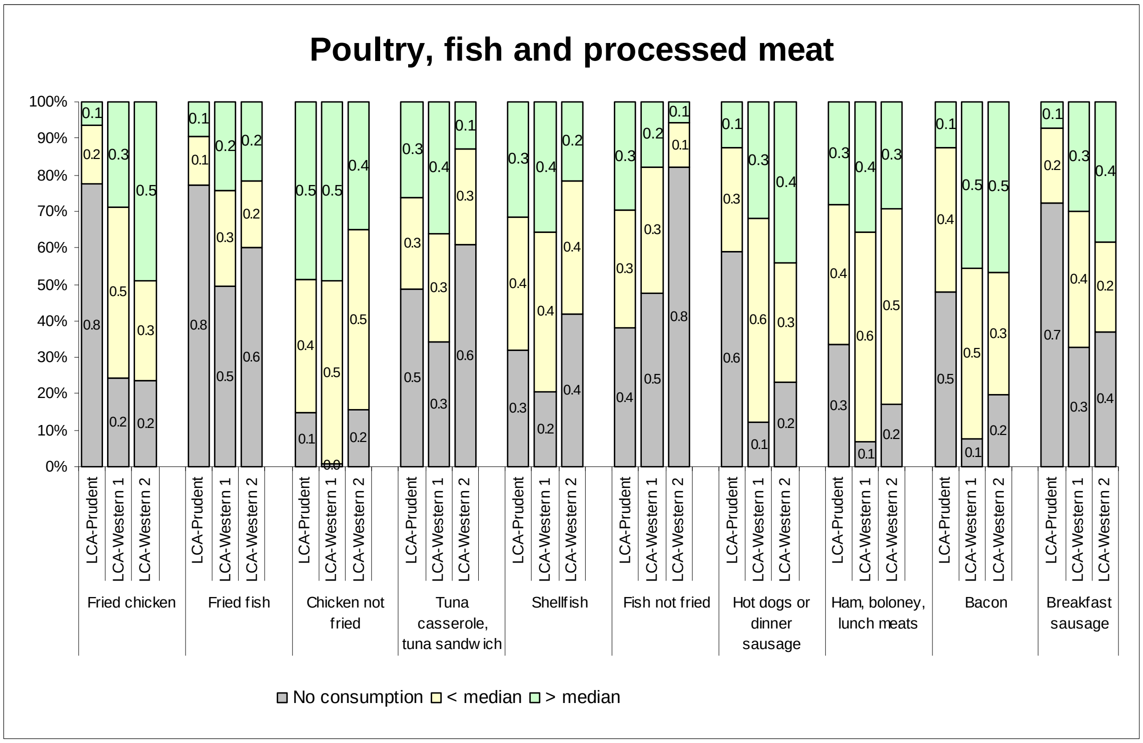


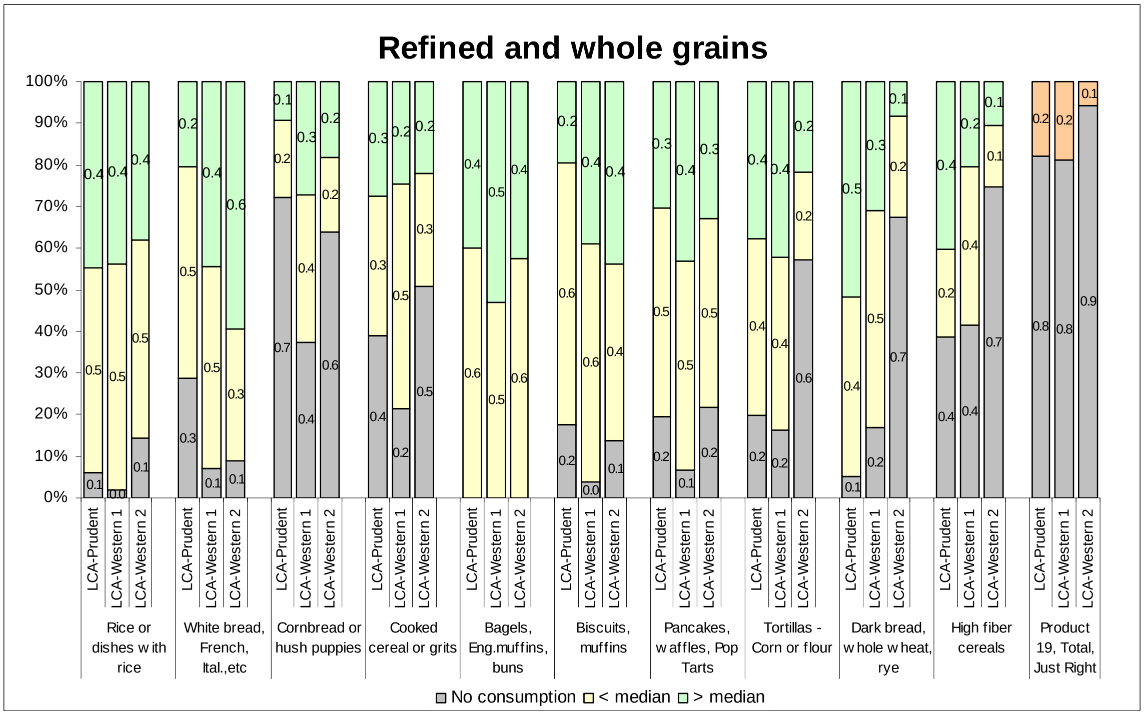


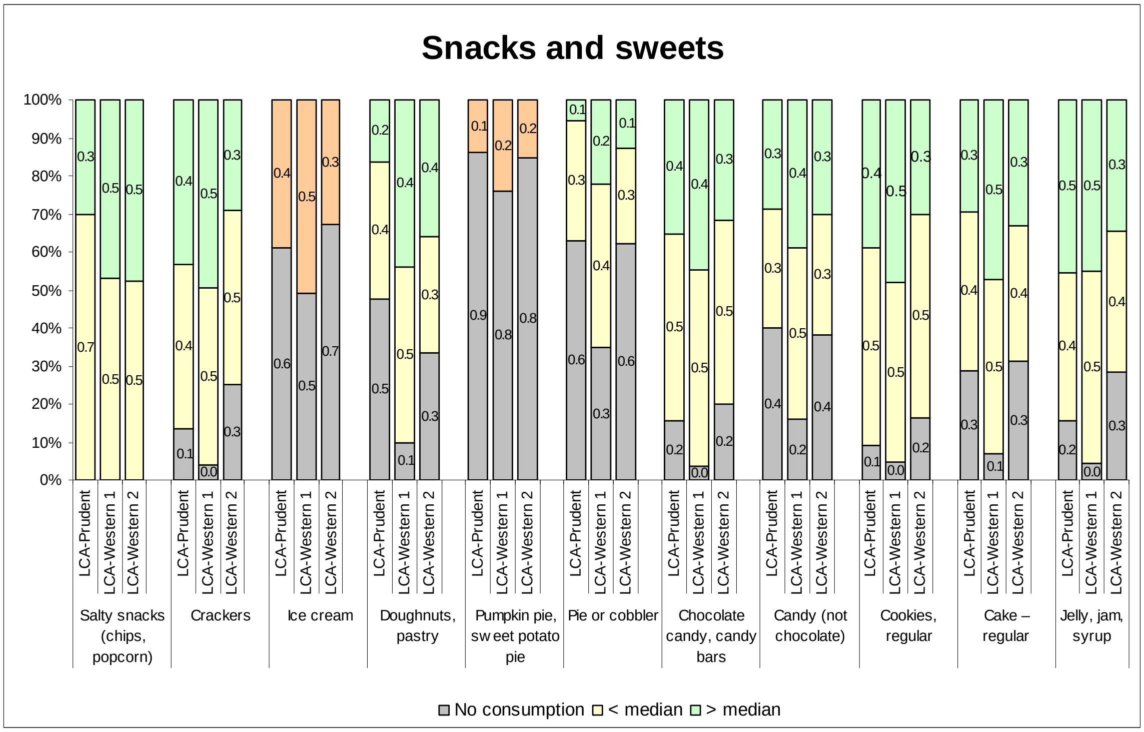

High-caratenoid vegetables
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
LCA-Prudent
LCA-Western 1
LCA-Western 2
Raw tomatoes
0.1
0.3
0.5
0.1
0.5
0.4
0.4
0.4
0.2
Broccoli
0.1
0.4
0.5
0.0
0.5
0.4
0.3
0.4
0.3
Spinach
0.2
0.4
0.5
0.3
0.5
0.2
0.7
0.2
0.1
Greens like collards
0.8
0.1
0.1
0.6
0.2
0.2
0.6
0.2
0.2
Carrots
0.1
0.4
0.6
0.0
0.5
0.5
0.3
0.4
0.2
Sw eet potatoes
0.6
0.2
0.2
0.4
0.3
0.3
0.8
0.1
0.1
No consumption
< median
> median

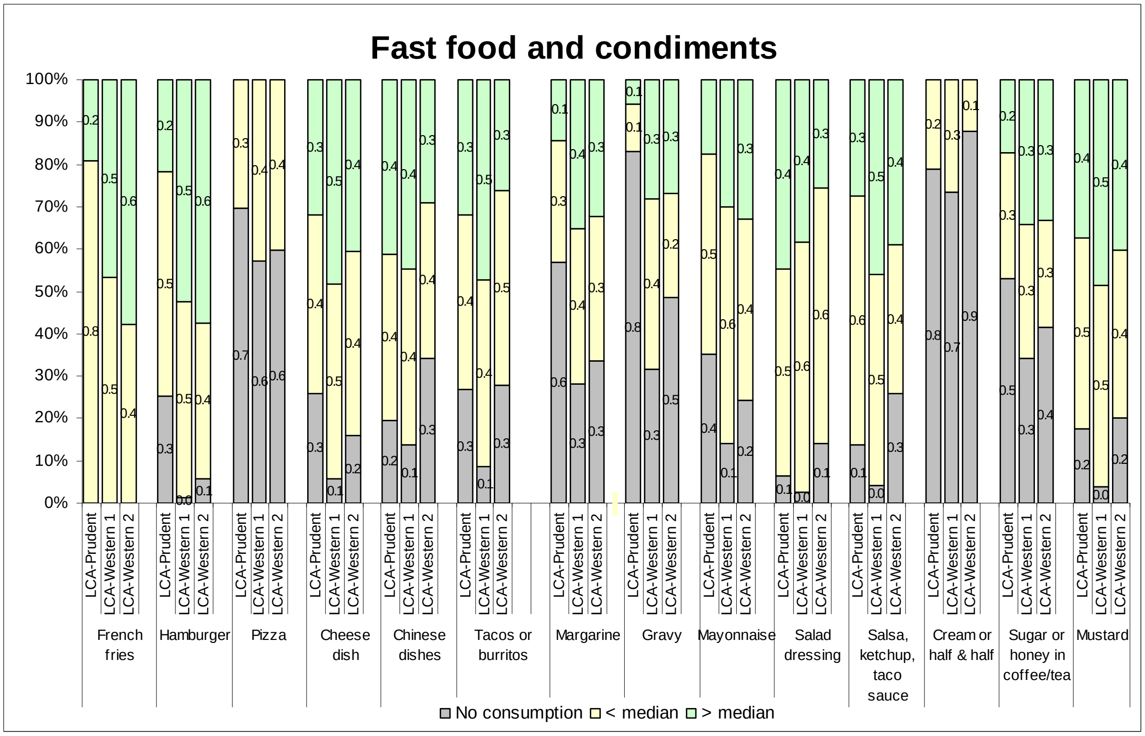

Dairy, nuts and beans
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
LCA-Prudent
LCA-Western 1
LCA-Western 2
Yogurt
0.1
0.4
0.5
0.2
0.6
0.3
0.6
0.3
0.1
Frozen yogurt
0.6
0.2
0.2
0.5
0.3
0.2
0.8
0.1
0.1
Low fat milk
0.1
0.4
0.5
0.2
0.4
0.4
0.5
0.3
0.2
Peanut butter
0.2
0.4
0.4
0.1
0.5
0.4
0.4
0.4
0.3
Peanuts, other nuts & seeds
0.2
0.4
0.4
0.1
0.5
0.4
0.5
0.4
0.2
Baked beans, black eye, pintos
0.3
0.4
0.3
0.2
0.6
0.3
0.4
0.4
0.2
Chili w ith beans
0.6
0.2
0.2
0.4
0.4
0.2
0.7
0.2
0.1
Refried beans, bean burritos
0.3
0.3
0.4
0.3
0.4
0.3
0.7
0.2
0.1
No consumption
< median
> median

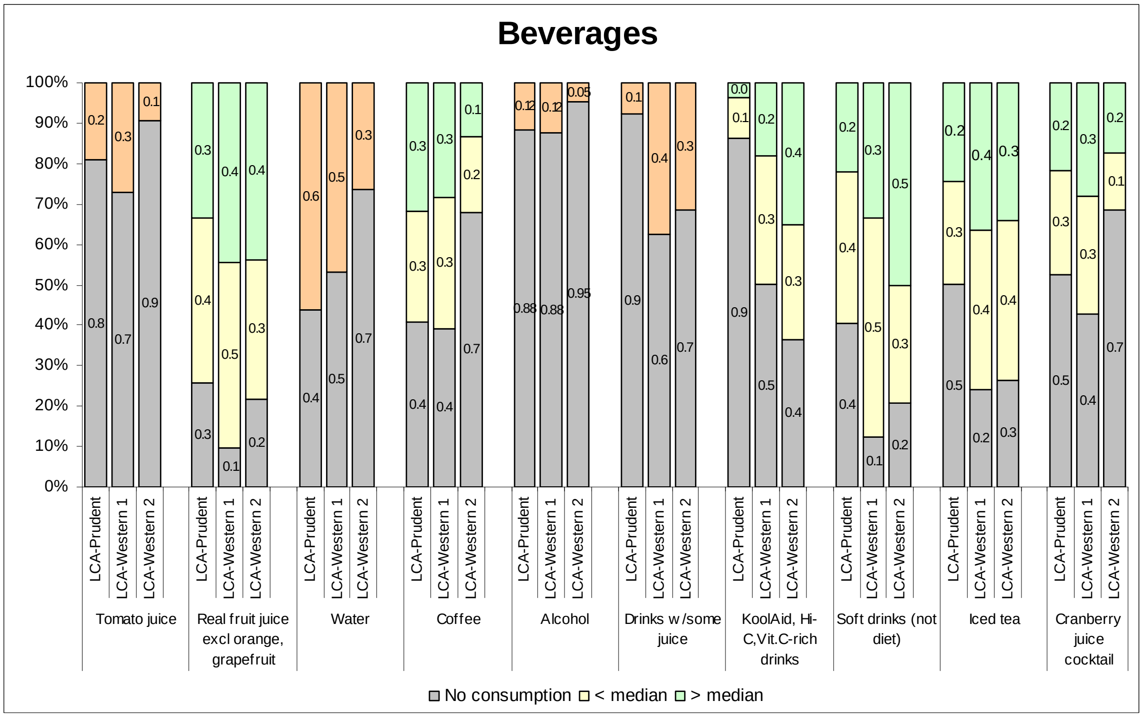

Mixed dishes and meat
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
LCA-Prudent
LCA-Western 1
LCA-Western 2
Vegetable stew
0.7 0.2 0.2
0.5 0.3 0.2
0.7 0.1 0.1
Spaghetti w/meat
0.7 0.3
0.6 0.4
0.7 0.3
Vegetable soup
0.3 0.4 0.4
0.2 0.4 0.4
0.5 0.3 0.2
Other soups like chicken noodle
0.3 0.3 0.3
0.2 0.4 0.4
0.4 0.3 0.2
Mixed dishes with beef or pork
0.7 0.2 0.1
0.3 0.4 0.3
0.6 0.2 0.2
Pasta dish
0.1 0.4 0.5
0.0 0.5 0.5
0.1 0.5 0.4
Chicken stew, pot pie
0.3 0.4 0.4
0.0 0.4 0.5
0.3 0.5 0.3
Beef (roast, steak, sandwiches)
0.4 0.4 0.2
0.1 0.5 0.4
0.2 0.4 0.4
Pork chops, roasts, dinner ham
0.5 0.4 0.1
0.1 0.6 0.3
0.3 0.5 0.3
Ribs, spareribs
0.9 0.1 0.0
0.6 0.2 0.2
0.6 0.2 0.2
No consumption
< median
> median


Poultry, fish and processed meat
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
LCA-Prudent
LCA-Western 1
LCA-Western 2
Fried chicken
0.8 0.2 0.1
0.2 0.5 0.3
0.2 0.3 0.5
Fried fish
0.8 0.1 0.1
0.5 0.3 0.2
0.6 0.2 0.2
Chicken not fried
0.1 0.4 0.5
0.0 0.5 0.5
0.2 0.5 0.4
Tuna casserole, tuna sandwich
0.5 0.3 0.3
0.3 0.3 0.4
0.6 0.3 0.1
Shellfish
0.3 0.4 0.3
0.2 0.4 0.4
0.4 0.4 0.2
Fish not fried
0.4 0.3 0.3
0.5 0.3 0.2
0.8 0.1 0.1
Hot dogs or dinner sausage
0.6 0.3 0.1
0.1 0.6 0.3
0.2 0.3 0.4
Ham, boloney, lunch meats
0.3 0.4 0.3
0.1 0.6 0.4
0.2 0.5 0.3
Bacon
0.5 0.4 0.1
0.1 0.5 0.5
0.2 0.3 0.5
Breakfast sausage
0.7 0.2 0.1
0.3 0.4 0.3
0.4 0.2 0.4
No consumption
< median
> median


Refined and whole grains
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0.1
0.5
0.4
0.0
0.5
0.4
0.1
0.5
0.4
0.3
0.5
0.2
0.1
0.5
0.4
0.1
0.3
0.6
0.7
0.2
0.1
0.4
0.4
0.3
0.6
0.2
0.2
0.4
0.3
0.3
0.2
0.5
0.2
0.5
0.3
0.2
0.6
0.4
0.5
0.5
0.6
0.4
0.2
0.6
0.2
0.0
0.6
0.4
0.1
0.4
0.4
0.2
0.5
0.3
0.1
0.5
0.4
0.2
0.5
0.3
0.2
0.4
0.4
0.2
0.4
0.4
0.6
0.2
0.2
0.1
0.4
0.5
0.2
0.5
0.3
0.7
0.2
0.1
0.4
0.2
0.4
0.4
0.4
0.2
0.7
0.1
0.1
0.8
0.2
0.8
0.2
0.9
0.1
LCA-Prudent
LCA-Western 1
LCA-Western 2
Rice or dishes with rice
White bread, French, Ital.,etc
Cornbread or hush puppies
Cooked cereal or grits
Bagels, Eng.muffins, buns
Biscuits, muffins
Pancakes, waffles, Pop Tarts
Tortillas - Corn or flour
Dark bread, whole wheat, rye
High fiber cereals
Product 19, Total, Just Right
No consumption
< median
> median


Snacks and sweets
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0.7
0.3
0.5
0.5
0.5
0.5
0.1
0.4
0.4
0.0
0.5
0.5
0.3
0.5
0.3
0.6
0.4
0.5
0.5
0.7
0.3
0.5
0.4
0.2
0.1
0.5
0.4
0.3
0.3
0.4
0.9
0.1
0.8
0.2
0.8
0.2
0.6
0.3
0.1
0.3
0.4
0.2
0.6
0.3
0.1
0.2
0.5
0.4
0.0
0.5
0.4
0.2
0.5
0.3
0.4
0.3
0.3
0.2
0.5
0.4
0.4
0.3
0.3
0.1
0.5
0.4
0.0
0.5
0.5
0.2
0.5
0.3
0.3
0.4
0.3
0.1
0.5
0.5
0.3
0.4
0.3
0.2
0.4
0.5
0.0
0.5
0.5
0.3
0.4
0.3
LCA-Prudent
LCA-Western 1
LCA-Western 2
Salty snacks (chips, popcorn)
Crackers
Ice cream
Doughnuts, pastry
Pumpkin pie, sweet potato pie
Pie or cobbler
Chocolate candy, candy bars
Candy (not chocolate)
Cookies, regular
Cake – regular
Jelly, jam, syrup
No consumption
< median
> median


Fast food and condiments
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
LCA-Prudent
LCA-Western 1
LCA-Western 2
French fries
0.8
0.2
0.5
0.5
0.4
0.6
Hamburger
0.3
0.5
0.2
0.0
0.5
0.5
0.1
0.4
0.6
Pizza
0.7
0.3
0.6
0.4
0.6
0.4
Cheese dish
0.3
0.4
0.3
0.1
0.5
0.5
0.2
0.4
0.4
Chinese dishes
0.2
0.4
0.4
0.1
0.4
0.4
0.3
0.4
0.3
Tacos or burritos
0.3
0.4
0.3
0.1
0.4
0.5
0.3
0.5
0.3
Margarine
0.6
0.3
0.1
0.3
0.4
0.4
0.3
0.3
0.3
Gravy
0.8
0.1
0.1
0.3
0.4
0.3
0.5
0.2
0.3
Mayonnaise
0.4
0.5
0.1
0.6
0.2
0.4
0.3
Salad dressing
0.1
0.5
0.4
0.0
0.6
0.4
0.1
0.6
0.3
Salsa, ketchup, taco sauce
0.1
0.6
0.3
0.0
0.5
0.5
0.3
0.4
0.4
Cream or half & half
0.8
0.2
0.7
0.3
0.9
0.1
Sugar or honey in coffee/tea
0.5
0.3
0.2
0.3
0.3
0.3
0.4
0.3
0.3
Mustard
0.2
0.5
0.4
0.0
0.5
0.5
0.2
0.4
0.4
▨ No consumption □ < median □ > median


Beverages
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
LCA-Prudent
LCA-Western 1
LCA-Western 2
Tomato juice
0.8
0.2
0.7
0.3
0.9
0.1
Real fruit juice excl orange, grapefruit
0.3
0.4
0.3
0.1
0.5
0.4
0.2
0.3
0.4
Water
0.4
0.6
0.5
0.5
0.7
0.3
Coffee
0.4
0.3
0.3
0.4
0.3
0.3
0.7
0.2
0.1
Alcohol
0.88
0.12
0.88
0.12
0.95
0.05
Drinks w/some juice
0.9
0.1
0.6
0.4
0.7
0.3
KoolAid, Hi-C,Vit.C-rich drinks
0.9
0.1
0.0
0.5
0.3
0.2
0.4
0.3
0.4
Soft drinks (not diet)
0.4
0.4
0.2
0.1
0.5
0.3
0.2
0.3
0.5
Iced tea
0.5
0.3
0.2
0.2
0.4
0.4
0.3
0.4
0.3
Cranberry juice cocktail
0.5
0.3
0.2
0.4
0.3
0.3
0.7
0.1
0.2
▨ No consumption □ < median □ > median

APPENDIX B

Supplementary Tables and Figures for Chapter 3

TABLE B1. Food-items by food group, PIN Study

Food group		Food-item	
Code	**Name**	**Code**	**Name**
1	Vitamin C fruits	10	Oranges, tangerines
1	Vitamin C fruits	12	Grapefruit
2	Other fruits	1	Apples or pears
2	Other fruits	2	Bananas
2	Other fruits	4	Peaches, apricots, fresh
2	Other fruits	6	Cantaloupe (year round)
2	Other fruits	7	Watermelon
2	Other fruits	8	Canned fruit, applesauce, etc.
2	Other fruits	9	Strawberries
2	Other fruits	15	Other fruits
3	Vegetables	16	Green beans or peas
3	Vegetables	19	Corn, fresh, frozen or canned
3	Vegetables	29	Coleslaw, cabbage
3	Vegetables	31	Green salad
3	Vegetables	34	White potatoes baked, mashed
3	Vegetables	36	Other vegetables
4	High-caratenoid vegetables	22	Raw tomatoes
4	High-caratenoid vegetables	24	Broccoli
4	High-caratenoid vegetables	27	Spinach (cooked)
4	High-caratenoid vegetables	28	Greens like collards
4	High-caratenoid vegetables	30	Carrots
4	High-caratenoid vegetables	33	Sweet potatoes
5	Cheese & whole milk	81	Cheese and cheese spreads
5	Cheese & whole milk	83	Whole milk
6	Low fat dairy	82	Yogurt
6	Low fat dairy	84	Reduced fat 2 % Milk
6	Low fat dairy	85	Nonfat milk
6	Low fat dairy	125	Frozen yogurt/regular
6	Low fat dairy	156	Low-fat 1% milk
7	Nuts	61	Peanut butter
7	Nuts	170	Peanuts, other nuts & seeds
8	Beans	18	Baked beans, black eye, pintos
8	Beans	102	Chili with beans
8	Beans	109	Refried beans, bean burritos
9	Mixed dish w/meat	40	Vegetable stew
9	Mixed dish w/meat	49	Spaghetti w/Tom. sauce + meat

TABLE B1. Food-items by food group, PIN Study (continued)

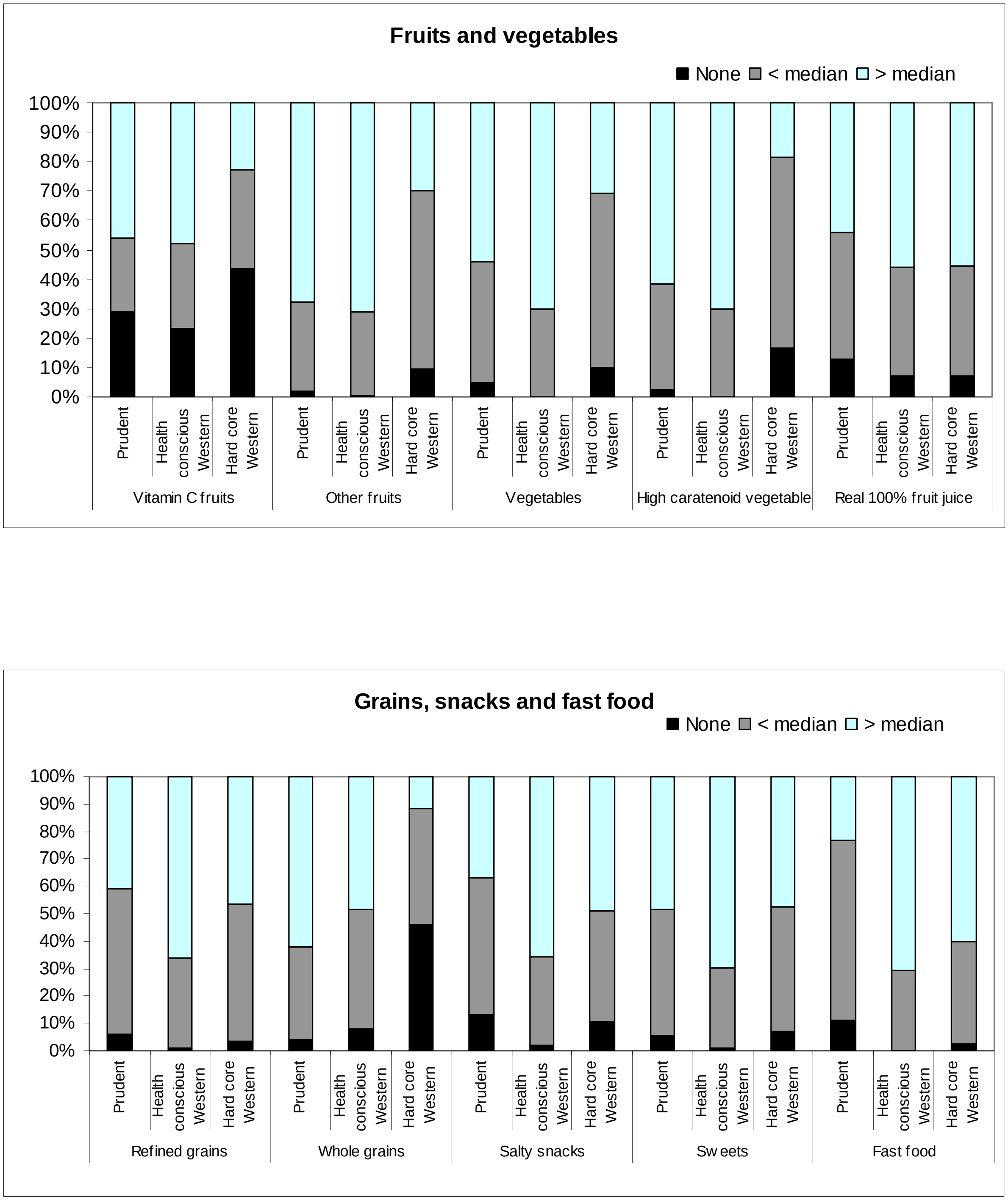
Food group		Food-item	
Code	**Name**	**Code**	**Name**
9	Mixed dish w/meat	55	Vegetable soup
9	Mixed dish w/meat	56	Other soups like chicken noodle
9	Mixed dish w/meat	104	Mixed dishes with beef or pork
9	Mixed dish w/meat	180	Pasta salad, other pasta dish
9	Mixed dish w/meat	187	Chicken stew, pot pie
10	Eggs	71	Eggs or egg biscuits
11	Beef	39	Beef (roast, steak, sandwiches)
12	Pork	42	Pork chops, roasts, dinner ham
12	Pork	143	Ribs, spareribs
13	Fried chicken or fish	43	Fried chicken
13	Fried chicken or fish	45	Fried fish
14	Chicken not fried	44	Chicken not fried
15	Fish not fried	46	Tuna casserole, tuna sandwich
15	Fish not fried	47	Shellfish
15	Fish not fried	48	Fish not fried
16	Processed meat	53	Hot dogs or dinner sausage
16	Processed meat	54	Ham, boloney, lunch meats
16	Processed meat	72	Bacon
16	Processed meat	73	Breakfast sausage
17	Refined grains	35	Rice or dishes with rice
17	Refined grains	57	White bread, French, Ital.,etc
17	Refined grains	59	Cornbread or hush puppies
17	Refined grains	68	Cereal excl. fiber or fortified
17	Refined grains	69	Cooked cereal or grits
17	Refined grains	119	Bagels, Eng.muffins, buns
17	Refined grains	120	Biscuits, muffins
17	Refined grains	121	Pancakes, waffles, Pop Tarts
17	Refined grains	181	Tortillas - Corn or flour
18	Whole grains	58	Dark bread, whole wheat, rye
18	Whole grains	66	High fiber cereals
18	Whole grains	67	Product 19, Total, Just Right
19	Salty snacks	60	Salty snacks (chips, popcorn)
19	Salty snacks	112	Crackers
20	Sweets	74	Ice cream
20	Sweets	75	Doughnuts, pastry
20	Sweets	76	Pumpkin pie, sweet potato pie
20	Sweets	77	Pie or cobbler
20	Sweets	78	Chocolate candy, candy bars
20	Sweets	79	Candy (not chocolate)
20	Sweets	123	Pudding
20	Sweets	133	Cookies, regular
20	Sweets	178	Cake – regular
20	Sweets	179	Jelly, jam, syrup
21	Real 100% juice	11	Orange juice, grapefruit juice

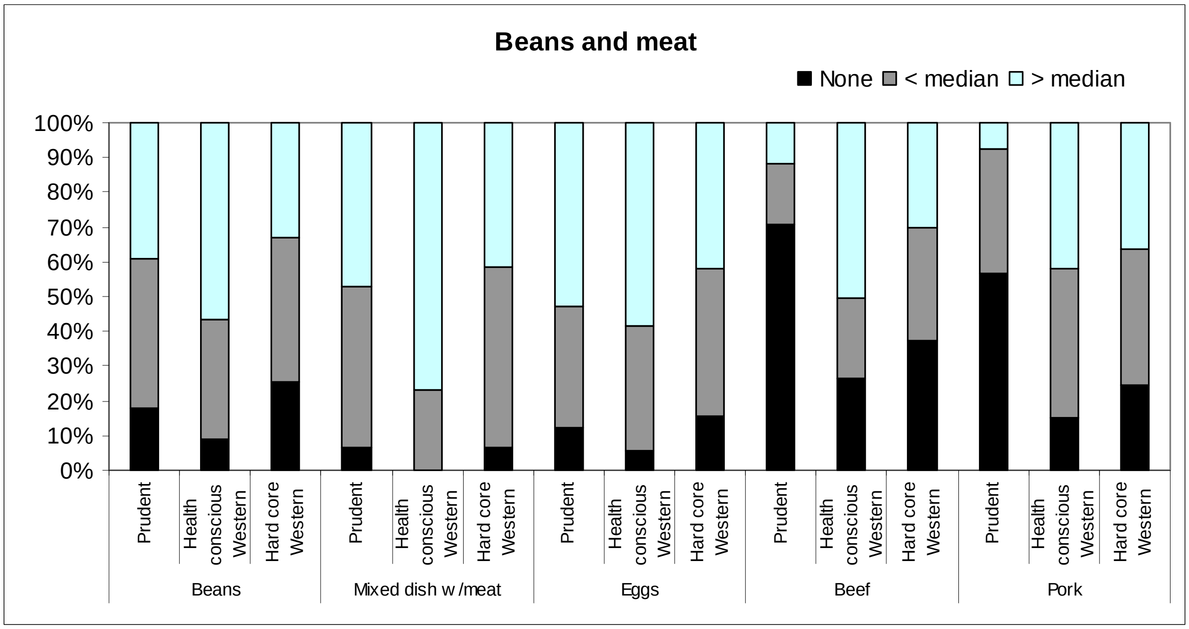
TABLE B1. Food-items by food group, PIN Study (continued)

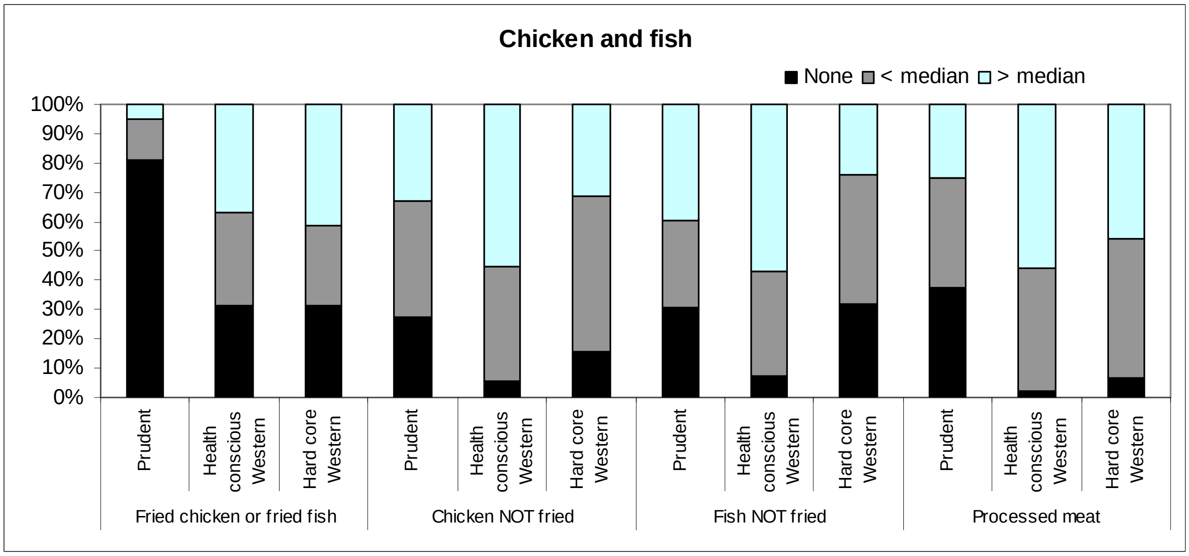
Food group		Food-item	
Code	**Name**	**Code**	**Name**
21	Real 100% juice	152	Tomato juice, V-8 juice
21	Real 100% juice	160	Real fruit juice excl orange, grapefruit
22	Coffee	92	Coffee
23	Alcohol	88	Beer (regular)
23	Alcohol	89	Wine or wine coolers
23	Alcohol	90	Liquor or mixed drinks
24	Soft drinks	13	Drinks w. some juice, Sunny D
24	Soft drinks	14	KoolAid, Hi-C,Vit.C-rich drinks
24	Soft drinks	86	Soft drinks or Snapple not diet
24	Soft drinks	93	Tea or iced tea (not herb tea)
24	Soft drinks	164	Cranberry juice cocktail
25	Diet soft drinks	87	Diet soft drinks
26	Fast food	32	French fries, fried potatoes
26	Fast food	38	Hamburger, cheeseburger
26	Fast food	50	Pizza
26	Fast food	51	Cheese dish like macaroni/cheese
26	Fast food	106	Chinese dishes
26	Fast food	185	Tacos or burritos
27	Mayo, gravy, butter or margarine	62	Butter
27	Mayo, gravy, butter or margarine	63	Margarine
27	Mayo, gravy, butter or margarine	65	Gravy
27	Mayo, gravy, butter or margarine	198	Mayonnaise, sandwich spreads
28	Salad dressing	64	Salad dressing
29	Water	99	Water

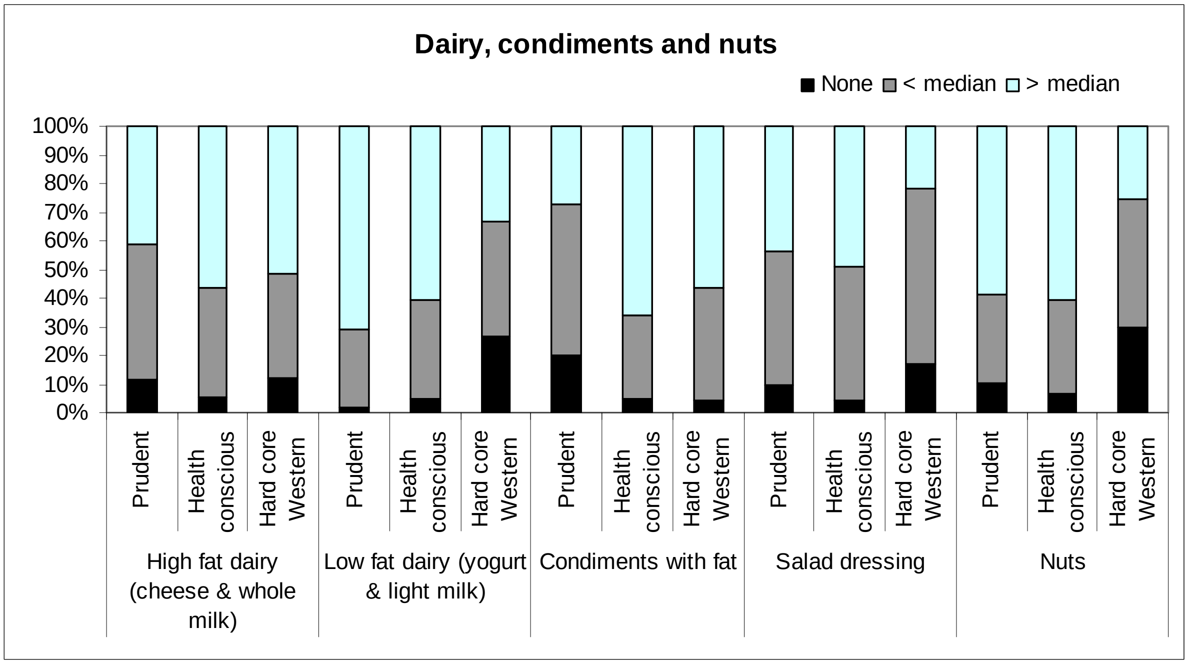
FIGURE B1. Probabilities of consumption by latent class

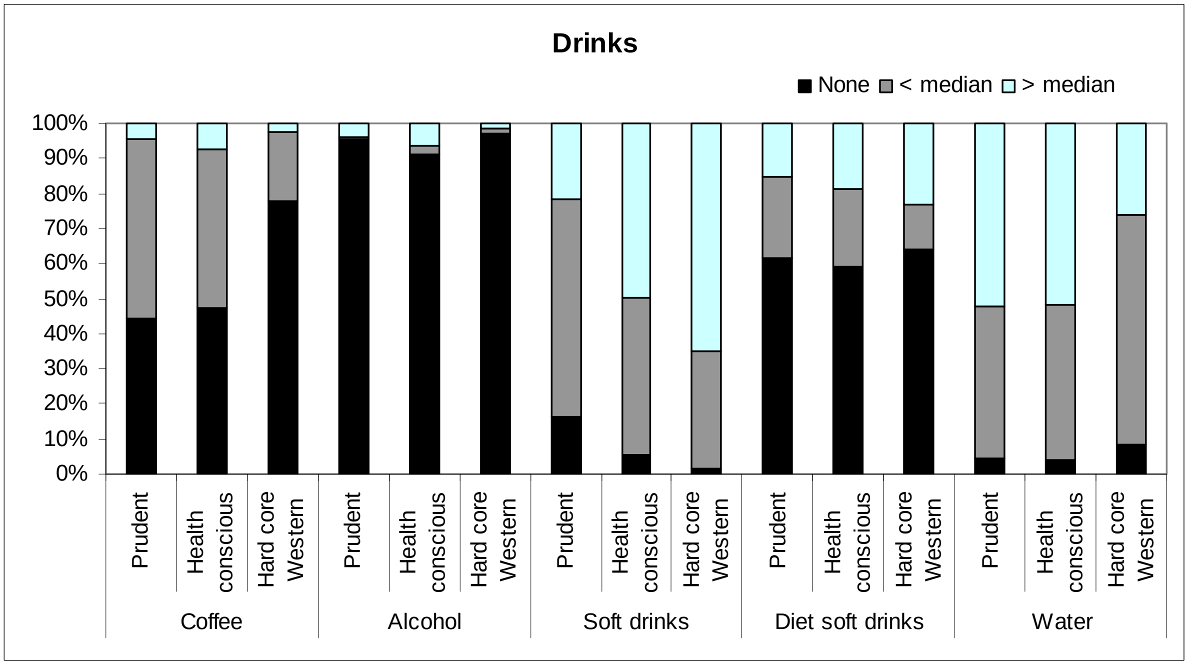
from LTM on 29 food-groups, PIN Study




Beans and meat
■ None ▨ < median □ > median
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Prudent
Health conscious Western
Hard core Western
Beans
Mixed dish w /meat
Eggs
Beef
Pork


Chicken and fish
■ None ▨ < median □ > median
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Prudent
Health conscious Western
Hard core Western
Fried chicken or fried fish
Chicken NOT fried
Fish NOT fried
Processed meat


Dairy, condiments and nuts
■ None ▨ < median □ > median
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Prudent
Health conscious
Hard core Western
High fat dairy (cheese & whole milk)
Low fat dairy (yogurt & light milk)
Condiments with fat
Salad dressing
Nuts


Drinks
■ None ▨ < median □ > median
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Prudent
Health conscious
Hard core Western
Coffee
Alcohol
Soft drinks
Diet soft drinks
Water