Comparison Between Bayesian Latent Factor Model and Principal Component Analysis on Nutrition Dietary Patterns in New Zealand Senior Adults

Gasper Xin Qian

Department of Statistics

The University of Auckland

Supervisor: Beatrix Jones

Abstract

Principal component analysis (PCA) has been a popular approach in dietary patterns study. However, the arbitrariness of setting boundary in selecting food variables for dietary patterns identifications and interpretations is a drawback. Sparse latent model can be an option on covering the shortages.

The objective of this study is to compare the two methods, namely Sparse latent model and PCA, for dietary patterns identifications among New Zealand senior adults.

Dietary data was collected among 367 New Zealand senior adults aged 65 – 74 from the Researching Eating Activity and Cognitive Health (REACH) study. 109 food items are generated by using a food frequency questionnaire that obtains the frequency and portion size intakes for each item. Bayesian latent model is used for analysing both standardized data in gram scales (gram per day) and in frequency scales (intake frequency level per day) where 109 food items are combined into 57 food groups in gram scaled data, and the results are compared with PCA.

Two main dietary patterns, “healthy” and “Western” dietary patterns separately, are identified in both Sparse latent model and PCA. PCA tend to include more food groups for each dietary pattern which is represented by a principal component, more cross-loading food groups has appeared in similar dietary patterns which are categorised as the main “healthy” pattern. Sparse latent model has fewer and more representative food groups included in one dietary pattern. Sparse latent model produced more interpretable dietary patterns than PCA with frequency scaled data. However, the model recognised less pattern in gram scaled FFQ1 data, and didn’t manage to recognize any pattern with gram scaled FFQ2 data.

Sparse latent model can be used for studies of dietary pattern by reducing the arbitrariness of selecting food variables in dietary pattern identifications. However, further settings and adjustment are needed for the model to adapt various data units as well as data scales.

Contents

Chapter 1

Introduction and Literature Reviews

1.1 Introduction

1.1.1 PCA and Dietary Patterns

Dietary pattern can be defined as the amount, variety, or combination of different foods and beverage in a diet along with the habitual intake frequency. Dietary pattern analysis has been undertaken with various approaches to explore the potential combinations of different diets. The most common approach is principal component analysis (PCA) which assumes that the association between the principal component and a series of food groups can represent dietary patterns formed by combinations of food consumptions. However, the factors in PCA are linear combinations of observed variables which will include those weakly associated ones, which decreases the accuracy of the pattern that one factor is picturing (Joo et al, 2018). The situation can be moderated by setting a threshold of loading values, which brings a new problem of the proper place of the threshold.

1.1.2 Bayesian Latent Factor Model and Dietary Patterns

Bayesian factor regression model is widely used for complex high-dimensional data in gene expression patterns and other fields (Joo et al, 2018), it comes up with some characterizes that can fix the shortcoming that was faced in PCA. By forcing the loadings of weakly associated variables to zero, each factor is only represented by a subset of the significant and representative observed variables.

In this study, the cut-off for the amount of significant loadings is also clear to pick because the proportion of zero loadings, describing how strong the association is between the factor and the variables, during iterations have gotten pushed into two-sided groups of variables, ones with very high proportions of zero loadings, and the others with very low proportions. The higher the proportion of zero loadings is, the less associated the variable is to the factor. It is easy and convenient to separate the strongly associated variables from the rest by drawing such a cut-off within the gap between those two-sided groups.

There are additional covariates in this study, such as age, sex, sociocultural factors, which can influence the “dietary pattern structures”. Joo et al (2018) has studied and demonstrated a way of treating these covariates and the food group variables as in the same level, and the analysis are based on their joint distribution. We took a different look at it by treat food groups and additional variables as at different levels. In other words, we are not considering those biological data while dealing with food consumptions and dietary patterns.

1.1.3 Acquisition of Data

The data used for analysis are from REACH study which focus on exploring senior New Zealanders' dietary patterns and the relationship of the patterns with health outcomes (Mumme et al, 2019). The food intakes are measured with 10 categorical frequency variables and are converted to a frequency scales per day in 10 levels. Food frequency questionnaire (FFQ) is a questionnaire used to obtain frequency or portion size information about food and beverage consumption over a specified period. FFQ was used to give measurements on the 109 food item intakes by the participants and was administrated by the survey two times in that period of time. Most of the participants completed the survey twice in a month apart, and the data were successfully collected from 367 participants in FFQ1 and from 319 participants in FFQ2. Furthermore, the data are converted into gram scales which demonstrated the weight of food intakes in gram per day. Datasets in both scales are used in the analysis for more detailed comparisons between PCA and the Bayesian sparse latent model where 109 food items are combined into 57 food groups in gram scaled data.

1.1.4 Conclusions

Both PCA and sparse latent factor model are processed on datasets FFQ1 and FFQ2 for a detailed comparison of exploring different dietary patterns and the accuracy of the pattern that is represented by a combination of food groups. Two main dietary patterns, namely "healthy" and "western", are identified by both methods. PCA have more food groups included in each dietary pattern and more cross-loading food groups has appeared between the 2 factors within the main "healthy" pattern. Sparse latent model has fewer food groups included in one dietary pattern, with each food group being a better represented element of the pattern. Sparse latent model produced more interpretable dietary patterns than PCA with frequency scaled data. However, the model recognised less pattern in gram scaled FFQ1 data, and didn't manage to recognize any pattern with gram scaled FFQ2 data. Overall, Sparse latent model delivers a more precise and interpretable dietary pattern results by covering the shortage from PCA, but adjustments are also needed for it to adapt the data that scales are non-linearly transformed and a smaller amount of food variables.

1.2 Literature Reviews

This review is focused on sparse latent models with Bayesian specifications, more specifically to answer questions on what structure of the model is, how the model is used in Bayesian approach, how to deal with the nonnormality of the data, and how MCMC analysis is combined with the model.

The use of generalised shrinkage priors and high-dimensional predictors is useful especially when the amount of predictors is large while the amount of data available is relatively small (West, 2003). Sparse latent models are having a design assuming linear relationships between the latent factors and variables in measurement as well as shrinking the insignificant loadings to zero while keeping the significant loadings as non-zero make itself useful in multivariate analysis.

Latent factor models have a basic structure that can be applied into dietary pattern study:
Defining the product of $p * k$ loading matrix as A, A_{ij} therefore is one loading in A over i^{th} food item and j^{th} latent factor, the food intakes x_i, the normally distributed sum of random noise v_i, and k-vector latent factor scores λ_i (p is the number of variables and k is the number of factors).

$$x_i = A * \lambda_i + v_i \qquad (1)$$

Dietary pattern analysis requires a precise description of the pattern using a combination of representative food variables while excluding the rest unassociated variables, which leads to many zeros on each row of the loading matrix. It is intuitive and reasonable to use priors that can induce the sparsity in the loading matrix (West, 2003). Therefore, the Bayesian approach is preferred.

We assume that λ_i has its priors with T distribution and the variance for each vector inside of λ_i is in Gamma distribution with mean and variance both are equal to r/2, where r is the degree of freedom parameter. The reason for assuming independent T distributions is that it allows shrinkages with various degrees for different factor dimension (West, 2003). Bayesian approach uses prior on each A_{ij} by assign a probability π_j to induce sparsity (zeros) with a high probability:

$$A_{ij} \sim \pi_j \, \sigma_0(A_{ij}) + (1 - \pi_j) N(A_{ij} | 0, \tau_j) \qquad (2)$$

Where $\sigma_0(.)$ is the Dirac delta function at 0, and τ_j is the variance of a normal prior from which the nonzero loadings on factor j are drawn. Having π_j's prior heavily favours 1 draws non-zero loadings with a very low probability (Carvalho et al, 2008).

Based on such a heavy prior concentration near 1, important loadings are able to escape the shrinkage and to obtain a non-zero values whereas other loadings are getting zeros by the shrinkage, and our loading matrix will be formed with lots of zeros and sparse non-zero salient loadings. Therefore, posterior distribution can be calculated over prior and the likelihood based on our data, which defines the dietary pattern formed by significant food variables.

Taking the arbitrary non-normal data structure into consideration, which is likely to be encounter in our study when there is a strong positive skewness for food variables, Dirichlet process (DP) framework is used for modelling the latent factor.

Defining a k-variate distribution function $F(\lambda_i)$ a latent factor λ_i, where $F \sim (\alpha\ F_0)$ being a DP prior with base measure αF_0, $\alpha > 0$, and prior expectation $F_0(\lambda) = N(\lambda|0, I)$. Let λ_{-i} be the set of latent factors excluding λ_i, we have the DP model for λ_i marginalized over F as:

$$(\lambda_i|\lambda_{-i}) \sim a_{n-1}N(\lambda_i|0, I) + (1 - a_{n-1})\sum_{r=1,r\neq i}^{n}\sigma_{\lambda_r}(\lambda_i) \tag{3}$$

Where $\sigma_\lambda(.)$ is the Dirac delta function at λ_i, and $a_{n-1} = \frac{\alpha}{\alpha+n-1}$. This model indicates that λ_i is assigned over a normal prior with probability a_{n-1}, it takes the value from one existing λ_r in equal probabilities within these n-1 values. The model has a feature of reducing a sample of n factors into k values with the k-variate distribution and configure the samples across the k clusters in factor space (Carvalho et al, 2008). The larger the α is, the more clusters it will be. Dirichlet process mixture model can mix this many normal distributions and that the model can capture any continuous distribution if α is set to approach infinity. It is used in our study by setting α to a finite number that can adapt the non-normal structure of the data.

With the interest of exploring what the linear combination of food variables over each factor is, generating posterior samples of the combination is helpful for the estimation of the posterior means. Let β be the p-vector of parameters representing a vector of all food items that we have, we want to find a way of generating sequences of posterior samples for $A\beta$, which is a factor length vector of food items for the factor variables. Markov Chain Monte Carlo method is used to help generating the posterior samples that we want to explore.

Markov Chain Monte Carlo method (MCMC) is an algorithm that can generate samples through posterior distributions for inferencing on posterior means over the parameters. Despite of that it can be time-consuming for the chain to converge to its stationary distribution without determine the steps needed, it works well especially in high dimensional space (Mossel et al, 2006). Extending Bayesian latent model with MCMC method, which can not only utilise the priors on the loading matrix A but also sequence through the columns of A (each factor) to sample the set of A_{ij} for fixed j.

MCMC is constructed for estimations on parameters by building a Markov Chain with a state space S starting from some initial value $\boldsymbol{\theta}_0$. The Markov Chain is run for a lot of iterations to generate an approximate sample $\boldsymbol{\theta}_n$ at time n. One way of going from $\boldsymbol{\theta}_n$ to $\boldsymbol{\theta}_{n+1}$ from time n to n + 1 is by using a proposal in which we sample a candidate $\boldsymbol{\theta}'_{n+1}$ by taking a step from $\boldsymbol{\theta}_n$ under a distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$. We accept the proposal of moving to $\boldsymbol{\theta}'_{n+1}$with a probability α where:

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}'_{n+1})p(x|\boldsymbol{\theta}'_{n+1})}{p(\boldsymbol{\theta}_n)p(x|\boldsymbol{\theta}_n)}\right\} \tag{4}$$

A Markov chain is considered as mixed well when it reaches the convergence, a stationary distribution in other words. The estimations won't get close to stationary results without giving a sufficient number of iterations. One convergence-monitoring strategy that can be applied to decide the time of reaching equilibrium is to look at the trace plots of the chain. Regarding the output as a time series, the equilibrium is not reached when the sample is not showing a random move around one value, and the first-order autocorrelation is not close to 0. We can monitor the convergence of loadings through iterations in our study by defining the sparsity:

$$Sparsity = \frac{num\ of\ zeros - mnvariables}{num\ of\ factors \times mnvariables} \tag{5}$$

where "mnvariables" is the number of variables in our dataset, "num of zeros" is the number of zeros is the number of zero loadings in our loading matrix, and "num of factors" is the number of factors that is set for representing different dietary patterns.

Chapter 2

Bayesian Sparse Latent Factor Model and Dietary Patterns

2.1 Study population

The REACH study is a cohort study in New Zealand with a primary goal of exploring the dietary patterns of 65-74 aged adults and their association with non-food factors including cognition, metabolic syndrome, as well as body composition (Mumme et al, 2019). There are over 360 participants in the study with their in-depth dietary data collected The food intakes are measured with 10 categorical variables, ranging from "never eat this food", "1 to 3 times a month", "2 to 3 times per day" to "6 plus per day", along with a convention formula for consumption in different period units (1 month = 4.34 weeks = 30.42 days). With the knowledge of average intake quantity for each food item, the frequency scale can also get transformed into gram scales, in which the food intakes are measured in weight (gram per day). Both gram scaled FFQ1 and FFQ2 data are the dataset used in our study. The 109 food items are combined into 57 different food groups. Overall, the two datasets, namely FFQ1 and FFQ2, with 367, 319 participants 57 food groups included in the dataset, are available for the analysis.

The dataset FFQ1 was completed by 367 participants along with 108 different food, which are combined and categorized into 57 different food categories. There were 614 values missing, spread over 242 participants and 99 foods. The most common missed items were tea (7%), tomatoes (6%), wholemeal bread (5%) and brown rice (5%). The dataset FFQ2 was completed by 318 participants along with 632 values missing (1.8%), spreading over 175 participants and 108 foods.

The values for each participant are the food intakes per day regarding to each food group. They are measured by the intake weights (in gram). When inspecting the data, the range of intakes varies by different food intakes, and missing values are assigned to extreme values to avoid influential points.

Data standardization is applied to both dataset with the reason that absolute intake weights can vary for different food groups, and it can cause confusions. For example, the 100 grams intake of olive oil is the same as 100 grams intake of bananas in weight, but it might be the largest intake in terms of olive oil in the dataset whereas it might be below the average banana intake in the dataset. Some food groups may appear more often in our factors just because they have greater intake weights in general. Standardization calculates the relative weights and put them back onto the same scale.

However, extreme values will end up appearing as outliers and will affect the outputs because of their extremeness. We replaced the missing values as the average values of those food groups as more realistic estimations instead.

2.2 Methodology

Bayesian latent factor model is carried out in processing both datasets, which produces factors with only a small number of variables involved by forcing the little associations between factors and variables to 0 association so that it is easier for us to reveal a set of food that shows significant association with specific dietary patterns.

The algorithm is adapted from Sparse Bayesian factor regression models built by Mike West team from Duke University, and the outputs are set of posterior means that are based on Monte Carlo approximations. The sparsity of our model is calculated through MCMC iterations, and it is considered as to judge the most efficient MCMC iterations amount which gives reliable results and not consuming extra iterations (Carvalho et al, 2008). The loadings are calculated for each dietary consumption against different factors, which represent the relation between our variables and the factors. The "non-Gaussian structure" parameter is selected for our data to adapt the skewness.

After trying out 10000 iterations with 1000 burn-in, it was found that the sparsity tended to be more stable after 7000 iterations. According to the trace plot of the relationship between the number of iterations and the sparsity for our model, the sparsity stays around 0.11 for 3 and 7 factors; it stays around 0.13 for 5 factors' model. We want our model to have the right sparsity so that it can identify different dietary patterns without including irrelevant food groups.

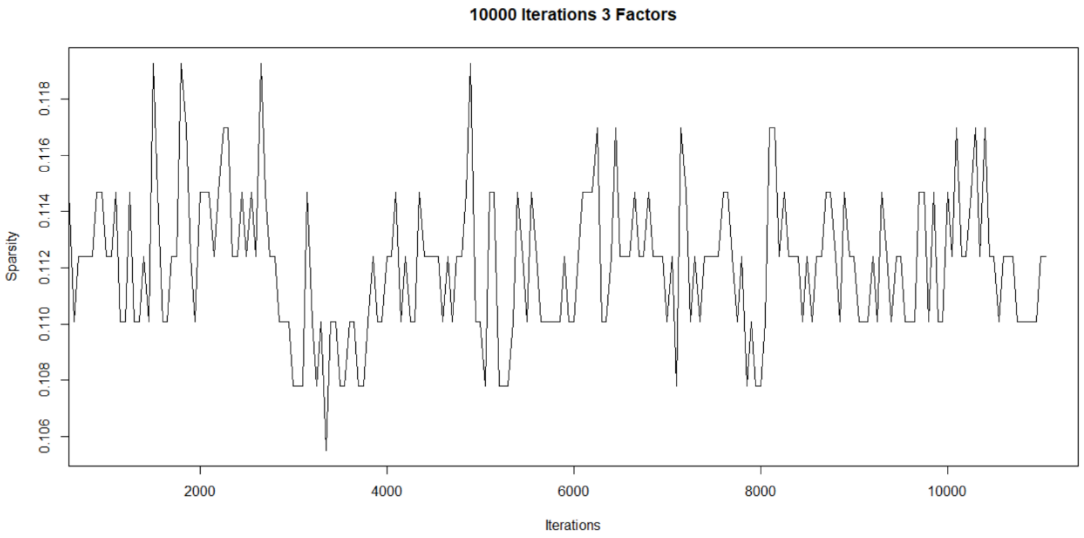
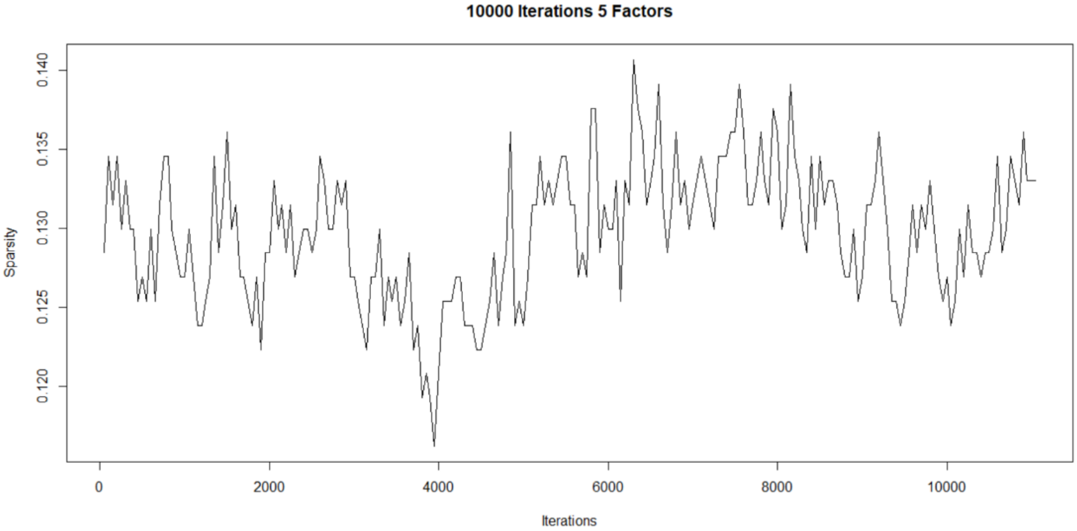


Figure 2.1.1 *Sparsity against 10000 iterations for models with 3, 5, 7 factors*



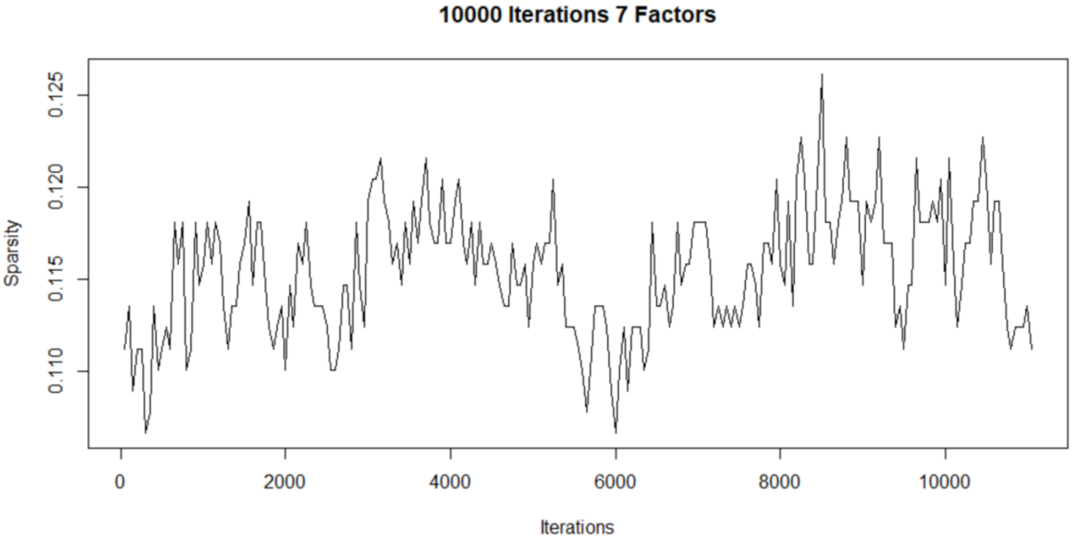


Figure 2.1.2 *Sparsity against 10000 iterations for models with 3, 5, 7 factors*

As the cut-offs for significant loadings can be ambiguous and hard to define when there is no obvious separation between loadings, the proportion of zeros for each dietary intake against different factors are calculated to give a better solution for cut-offs. The higher the proportion of zero is, the more zero loadings there are in one food group across 4000 iterations, and the less correlated the food group is to that factor. After plotting the proportion histograms for each factor, it is easy to see that most of the food groups are clustered at either very low proportions (<20%) or very high proportions (>80%). Since the food groups are separated to two sides, it is easy to pick a cut-off in between and select the significant dietary intakes which are with less 0 loadings.

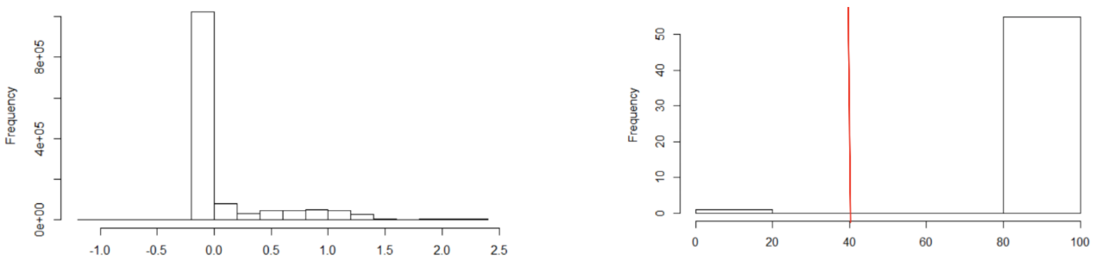


Figure 2.2 *The frequency histogram on the left contains values of loadings from the loading matrix output by Bayesian latent model. The histogram on the right contains posterior probability of zeros calculated based on the loading outputs, from which a stronger contrast is observed for identifying food patterns from less associated ones. Although the elbow from the left plot can be identified, it is much more obvious to set a cut-off on the right plot as the values tend to skew on two sides.*

2.3 Results

After applying the Bayesian latent model on FFQ1 data and separate out the food groups that are more associated and represented by each factor by setting the cut-offs based on the proportion of zero loadings calculation, the results are demonstrated in Table 2.1.

All the loadings of the food groups in the table are having the same loading sign for each factor, which means the food groups are positively associated. Each factor is represented by at least one food group, but not all factors are indicating general dietary patterns. For instance, factor 3 is associated with banana only and it will be a misinterpretation to consider banana can be a generalised dietary pattern. The threshold on the dietary pattern representations can be arbitrary but a factor with at least 3 food groups associated is considered as more likely to represent a dietary pattern.

The dietary patterns recognised from these 5 factors are included in Table 2.1, at which the first factor is recognised as a healthy dietary pattern and the second factor as a western healthy dietary.

The outputs with data FFQ2, included in Table 2.2, are not similar to the ones I got from FFQ1 with the same model settings. Each factor is represented by once one food group, which is very hard to be generalised into a dietary pattern.

With the goal of precision as well as reducing misrepresentations, the model with 3 and 7 factors are also processed on FFQ1 data. Comparing to the outputs from the model with 5 factors, the food groups represented by 3 factors, demonstrated in Table 2.3, are all appeared in the dietary patterns from the 5 factors' model (Table 2.1), which means setting the latent factor number to 5 not only includes all the information that the model with 3 factors has, but giving more information and dietary patterns that are better represented by more food items. On the contrary, more dietary patterns are identified by the model with 7 factors according to the outputs in Table 2.4. For instance, Dietary pattern 3 in Table 2.4 represents an unhealthy western food pattern. However, the other three recognised patterns are similar and with several food groups in all of them. They are also included in the outputs of 5 factors' model (Table 2.1), which made us suspect of that too many factors can deliver duplicated information and redundant patterns.

Bayesian Latent Model with 5 Factors over FFQ1 Data	Food Groups
Factor 1: Dietary Pattern 1 (*healthy*)	Wholegrain; Alliums; Alternate; Fresh.frozen.Legumes; Refined grain; Root.starchy.vegetable
Factor 2: Dietary Pattern 2 (*western)*	Sauces, Chutneys; Processed meats; Dressings; Biscuits, Cakes and Pastries; Diet rinks; Savoury; Tomatoes; Confectionery; Chocolate; Stone fruit; Processed fish; Cheese and Creamy diary; Beer
Factor 3	Banana
Factor 4	All other fruit
Factor 5	Alternate

Table 2.1 *The significant food groups separated from the Bayesian latent model with 5 factors using the gram scaled and standardized FFQ1 data. Factors with more than three factors represented are selected as introducing dietary patterns to make sure that the food patterns are well identified and indicated by the factors having sufficient numbers of food groups represented.*

Bayesian Latent Model with 5 Factors over FFQ2 Data	Food Groups
Factor 1	Alliums
Factor 2	Apple.Pear
Factor 3	Banana
Factor 4	All other fruit
Factor 5	Alternate

Table 2.2 *Dietary patterns identified from the Bayesian latent model with 5 factors using the gram scaled and standardized FFQ2 data*

Bayesian Latent Model with 3 Factors over FFQ1 Data	Food Groups
Factor 1: Dietary Pattern 1 (*healthy*)	Alliums; Oily.fish; Olives and Avocados; White fish
Factor 2: Dietary Pattern 2 (*healthy*)	Alternate; Carrots; Green.leafy.cruciferous; Water
Factor 3: Dietary Pattern 3 (*healthy*)	All other fruit; Dried.Legumes; Fresh.frozen.Legumes; Other.vegetables; Root.starchy

Table 2.3 *Dietary patterns identified from the Bayesian latent model with 3 factors using the gram scaled and standardized FFQ1 data*

Bayesian Latent Model with 7 Factors over FFQ1 Data	Food Groups
Factor 1: Dietary Pattern 1 (*healthy*)	Wholegrain; Alliums; Alternate; Green.leafy.cruciferous; Dried.Legumes; Fresh.frozen.Legumes; Refined.grain; Root.starchy.vegetables; Spices; Carrots; Cruciferous
Factor 2: Dietary Pattern 2 (*healthy)*	Apple.Pear; Carrots; Nuts and Seeds; Other vegetables; Eggs
Factor 3	Banana
Factor 4	All other fruit
Factor 5	Alliums; Milk
Factor 6 Dietary Pattern 3 (*western)*	Beer; Processed meats; Red wine; Savoury
Factor 7: Dietary Pattern 4 (*healthy)*	Berry fruits; Egg; Nuts and Seeds; Olives and Avocados; Salad vegetables; Stone fruits; Tea and Coffee; Tomatoes

Table 2.4 *Dietary patterns identified from the Bayesian latent model with 7 factors using the gram scaled and standardized FFQ1 data. Factors with more than three factors represented are selected as introducing dietary patterns to make sure that the food patterns are well identified and indicated by the factors having sufficient numbers of food groups represented.*

2.4 Discussion

The results and outputs with FFQ1 data turn out to be closer to our goal of dietary pattern identifications comparing to the ones from FFQ2 data. We are able to recognise more than one dietary pattern which is represented by a suitable amount of food groups with FFQ1 data. However, there are still misrepresentations such that some factors that are only represented by one food group. There are also cross-loading food groups between two or three dietary patterns. In terms of the reason behind getting fewer dietary patterns than the number of factors in the model, one plausible explanation is that the factors represented by a small number of food groups are involved in mixing the MCMC chain to help reaching stationary distributions. The results don't seem to be replicable with FFQ2 data as each factor is only associated with one food intake that we could hardly identify any dietary pattern.

Chapter 3

Comparison between Bayesian Latent Factor Model and PCA in Dietary Patterns

3.1 PCA Methodology

Principal component analysis (PCA) is used for dietary patterns identification by assuming that food intake items are represented by a small amount of principal components which are also recognized as dietary patterns. With the fact that each principal component is the linear combination of multiple food groups, one of the drawback of using PCA is that we tend to remove the food intakes which are less associated with a given dietary pattern while defining the cut-off for the loading values can be ambiguous and rarely a proper rule to follow. Moreover, it may cause more information loss if there is no obvious elbow in the scree plot, which may limit the amount of recognized patterns.

Principal component analysis (PCA) was processed on the same standardized data, and the results have shown some difference from the sparse latent factor approach. PCA worked fine with both FFQ1 and FFQ2 dataset.

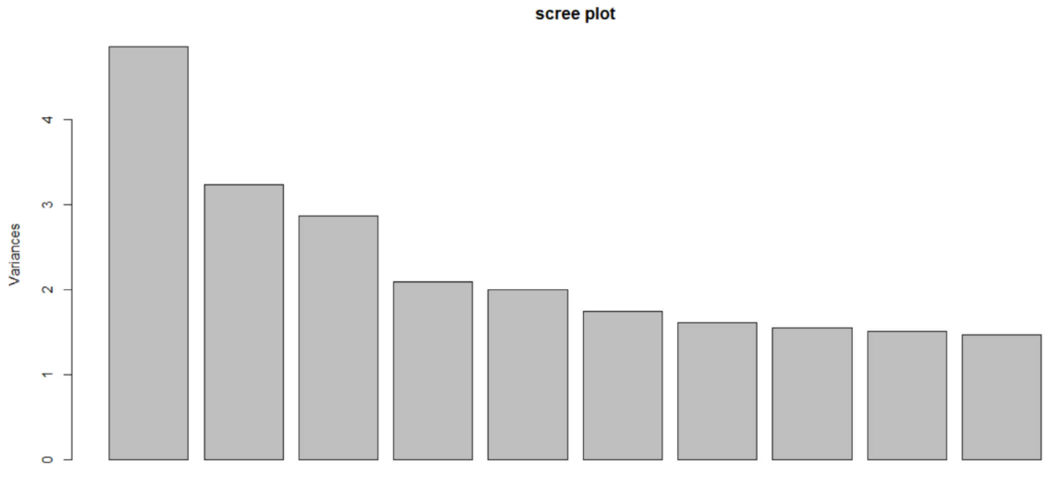


Figure 3.1 *The scree plot of the PCA on our dataset. There is no obvious elbow in the plot, but it can still be recognised that the first two or three principal components represent a good amount of information then the rest of the principal components. The first two components explain 15% of the variance, and the first three components explain 20% of the variance.*

The models with two and three rotated principal components are processed on our data, and food groups with loadings greater than 0.3 are kept in the result. The cut-off is set to be 0.3 following "a factor loading > 0.298 is significant in a sample size of 300" by Stevens (1992). According to the outputs, PCA worked moderately on our datasets with off-diagonal values, the covariances between variables and the extend of distortion in the data, being 0.41 and 0.46 for 3 factors and 2 factors separately. According to the outputs from Table 3.1 to 3.4, each dietary pattern includes at least ten food groups, and two main dietary patterns, healthy pattern and western pattern, are recognized. The existence of crossing-loading food groups was found with frankly a certain amount between the factors representing similar dietary patterns, which indicates that the pattern identification is limited and partially duplicated.

3.2 Comparison

Compare the sparse latent model to PCA from the outputs, Each factor in PCA is associated with more food intake groups, which demonstrates a suspicion of an increased redundancy in a sense that some food group with relatively weak associations are included too, or that each food group become less representative when several similar food groups are included in the principal component. However, PCA has given a better output than the sparse latent model when it comes to FFQ2 data. The two main dietary patterns are identified with PCA whereas the outputs from sparse latent model with FFQ2 failed to show any significant dietary pattern.

Two Rotated Components with FFQ1 Data	Food Groups
Dietary Pattern 1 (*healthy*)	salad veg; other veg; green leafy cruciferous; alliums; cruciferous; carrots; dried legumes; nuts and seeds; water; berry fruits; all other fruit; olives and avocados; alternate; root, starchy veg; fresh or frozen legumes; oily fish; wholegrain; white fish; spices, tomatoes, stone fruit
Dietary Pattern 2 (*western)*	sauces, chutneys; processed meats; dressings; biscuits, cakes and pastries; diet drinks; savoury; tomatoes; confectionery; chocolate; stone fruit; processed fish; cheese and creamy diary; beer

Table 3.1 *PCA outputs with two principal components on FFQ1 data*

Three Rotated Components with FFQ1 Data	Food Groups
Dietary Pattern 1 (*healthy*)	salad vegetables; alliums; other vegetables; tomatoes; olives and avocados; sauces and chutneys; berry fruits; green leafy cruciferous; stone fruit; dressings; all other fruit; water; oily fish; eggs; shellfish
Dietary Pattern 2 (*healthy*)	dried, fresh or frozen legumes; alternate; root, starchy veg; carrots; wholegrain; cruciferous; poultry; green leafy cruciferous
Dietary Pattern 3 (*western)*	processed meats; biscuits, cakes and pastries; savoury; suces, chutneys; confectionery; diet drinks; processed fish; chocolate; breakfast cereals; uSFA; beer; dressings; and *negative loading* nuts and seeds

Table 3.2 *PCA outputs with three principal components on FFQ1 data*

Two Rotated Components with FFQ2 Data	Food Groups
Dietary Pattern 1 (*healthy*)	cruciferous; carrots; alliums; green leafy cruciferous; fresh or frozen legumes; citrus; uSFA; other veg; root, starchy veg
Dietary Pattern 2 (*healthy*)	nuts and seeds; berry fruits; olives and avocados; oily fish; salad veg; spices; tomato; apple and pear; water; dried legumes; white fish; other vegetable; all other fruit and *negative loadings:* savoury; processed fish; processed meats; confectionary; red meat; biscuits, cakes and pastries; breakfast cereals

Table 3.3 *PCA outputs with two principal components on FFQ2 data*

Three Rotated Components with FFQ2 Data	Food Groups
Dietary Pattern 1 (*healthy*)	cruciferous; carrots; alliums; green leafy cruciferous; fresh or frozen legumes; citrus; uSFA; other veg; root, starchy veg
Dietary Pattern 2 (*healthy*)	tomatoes; salad veg; berry fruits; nuts and seeds; olives and avocados; dressings; oily fish; water; other vegetables; bran cereal; dried legumes; yoghurt; apple and pear; white fish; spices; banana
Dietary Pattern 3 (*western)*	savoury; processed fish; biscuits, cakes and pastries; processed meats; beer; chocolate; sauces chutneys; confectionery; dressings; red meat; breakfast cereals

Table 3.4 *PCA outputs with three principal components on FFQ2 data*

Chapter 4

Inconsistency from Non-Linear Data Transformation

4.1 Comparison

Sparse latent model and PCA are used on the frequency scaled data. The difference from the ones with gram scaled data is that sparse latent model is managed to identify dietary patterns with FFQ2 data. The outputs from PCA are similar to the expended outputs that are included in Table 3.1 to 3.4. The outputs from the sparse latent model on frequency scaled FFQ1 data included two main dietary patterns too. There are cross loading food items which are included in multiple factors representing the same dietary pattern, and the recognised dietary patterns along with the associated food items generated are in Table 4.1. However, sparse latent model with 5 factors has successfully delivered two dietary patterns with opposite loading signs from one factor despite that other factors are still associated with only one or two food items. According to the outputs from Table 4.2, the two main dietary patterns are represented by at least eight food items and are not conflicting the results from previous identifications. Overall, sparse latent model is supportive in dietary pattern identifications with frequency scaled data in our study, but the model needs change to adapt our data which are gram scaled.

Dietary Patterns Identified with FFQ1 Data	Food Items
Dietary Pattern 1 (*healthy*)	Berries; Dried.fruit; Apples.pears; Olives; Pickles; Coconut.oil; Vegetable.oils Marmite.vegemite; Brown rice; Non.milk.based.puddings; Carrots; Grain; Herbal.tea; Other.root.vegetable; Salad.vegetables; Tomatoes; Onions.leeks.garlic; Sausages; Seeds; Kumara.taro
Dietary Pattern 2 (*western)*	Hot.potato.chips.French.fries.wedges; Biscuits; Cakes; Port.sherry.liquors; Margarine; Coconut.cream; Creamy dressings; White.sauce; Sweets; Soft.fizzy.drinks

Table 4.1 *Sparse latent 5 factors model outputs with two dietary patterns identified on FFQ1 frequency scaled data*

Bayesian Latent Model with 5 Factors over FFQ2 Frequency Scaled Data	Food Items
Factor 1 with Positive Loading Signs: Dietary Pattern 1 (*healthy*)	Berries; All other fruits; Root vegetables; Salad vegetables; Leafy vegetables; Tomatoes; Beans; Onion.leaks.garlic; Kumura.taro; Dried fruits; Broccoli.cauliflower.brussel.sprouts.cabbage; White fish; Light dressings; All other vegetables; Nuts; Seeds; Peas
Factor 1 with Negative Loading Signs: Dietary Pattern 2 (*western)*	Avocado; Processed meat; Low.calorie.cordials; Chocolate; Sparkling water; Yogurt; Couscous; Cereal
Factor 2	Banana; Hot.potato.chips.French.fries.wedges
Factor 3	Apples.pears
Factor 4	Citrus fruits
Factor 5	Stone fruits

Table 4.2 *Sparse latent 5 factors model outputs with two dietary patterns identified on FFQ2 frequency scaled data*

4.2 Discussion

The inconsistency of the results is high likely caused by the non-linear transformation of our data while converting from one scale to another. With the fact that each portion of food item weights differently than the other and the intake portion for each food also various, calculating the intake weight for food items based on the frequency reforms the structure of the dataset non-linearly, which induces skewness and outliers that is not fitting our model as the frequency scale did despite the construction to accommodate non-normality. Joo et al (2018) have managed to derive distinct and interpretable dietary patterns by producing factors composed by a subset of food items with trivial associations using FFQ scaled data collected from a teenager group. Compare to the inconsistency of our results, it gives us suspicion that FFQ scale can be a more tractable form of non-normality with limited range than grams scale.

However, this is more complicated than just standardise the data as the outputs we had in Table 2.1 to 2.4 because we use sparse latent model with an intuition that the default model works with great motivation when the amount of predictors is large and the number of data is relatively moderate or small. Reducing the amount of predictors by combining food items may reduce the accuracy of the interpretation if not adjusting the settings of the model. With the consistent outputs from PCA, we can conclude that further adjustments on both the data side and the model side are needed if our data exhibits skewness or outliers.

Chapter 5

Conclusions

Both Bayesian sparse latent model and PCA are useful models for dietary pattern study with data collected from REACH study. They are managed to derive two main dietary patterns, including healthy and western patterns, by introducing a combination of food intake variables for each pattern. In contrast to PCA which includes too many food variables as well as cross loading variables within two factors, Bayesian sparse latent model processes fewer food variables in each dietary pattern to reduce the information redundancy. What is more, the arbitrariness of food items selection with PCA can be covered through inducing the sparsity into the loading matrix with Bayesian sparse latent model.

Sparse latent model produced more interpretable dietary patterns than PCA with frequency scaled data which also has more food variables, but the model delivers a less interpretable output when analysing gram scaled dataset with fewer food variables included. Overall, Sparse latent model delivers a more precise and interpretable dietary pattern results by covering the shortage from PCA, but adjustments on both model settings and data scale units are needed especially when there are fewer regression parameters in the dataset or the data is scaled by non-linear transformation.

Reference

Carvalho C.M., Chang J., Lucas J.E., Nevins J. R., Wang Q., West M. (2008) High-dimensional sparse factor modelling: applications in gene expression genomics *J Am Stat Assoc;* 103: 1438 – 56.

Joo, J., Vazquez, A. I., Fernandez, J. R., Bray, M. S., and Williamson, S. A. (2018) Advanced Dietary Patterns Analysis Using Sparse Latent Factor Models in Young Adults. *The Journal of Nutrition*, 12, 1984--1992.

Mossel E., Vigoda E. (2006) Limitations of Markov Chain Monte Carlo Algorithms for Bayesian Inference of Phylogeny *The Annals of Applied Probability* Vol. 16, No. 4, 2215-2234

Mumme, K.D., Conlon, C.; Hurst, P., Jones, M.B., Haskell-Ramsay, C., Stonehouse, W., Heath, A.-L., Coad, J., Seymour, J., Beck, K. (2019) Dietary Patterns and Associations with Socio-Demographic Factors in Older New Zealand Adults: The REACH Study. Proceedings 2019, 37, 40.

Mumme, K.D., von Hurst P.R., Conlon C.A., et al. (2019) Study Protocol: Associations between Dietary Patterns, Cognitive Function and Metabolic Syndrome in Older Adults - a Cross-Sectional Study. *BMC Public Health*. 2019;19(1):535. Published 2019 May 10. doi:10.1186/s12889-019-6900-4

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). *Lawrence Erlbaum Associates, Inc*.

West (2003) Bayesian Factor Regression Models in the “Large p, Small n” Paradigm
, ISDS, Duke University.