Study of Unobserved Factors in Fatty Acids with Exploratory Data Analysis

Yifan Chen

Contents

1. Introduction

1.1 Background of our study

Fatty acid plays an important role in human health and fat-related diseases. Fatty acids can be obtained from various foods or biosynthesized in the body (Fig. 1.1). Food lipids are digested and absorbed at intestine, and transferred via circulation to various organs. Cells contain various lipid pools in organelle. Lipids are exchanged between organelle via intracellular lipid traffics and metabolism. Organs contain unique lipid constituents, and exchange them with plasma lipoproteins across cell membranes. Lipoproteins consist of various lipoprotein particles having unique lipid compositions, such as chylomicrons, very-low-density lipoproteins (VLDL), low-density lipoproteins (LDL), and high-density lipoproteins (HDL). Lipids are exchanged among lipoprotein particles and free fatty acids associated with albumin in circulation. Thus, plasma lipids reflect complex lipid pools and metabolism in the body.

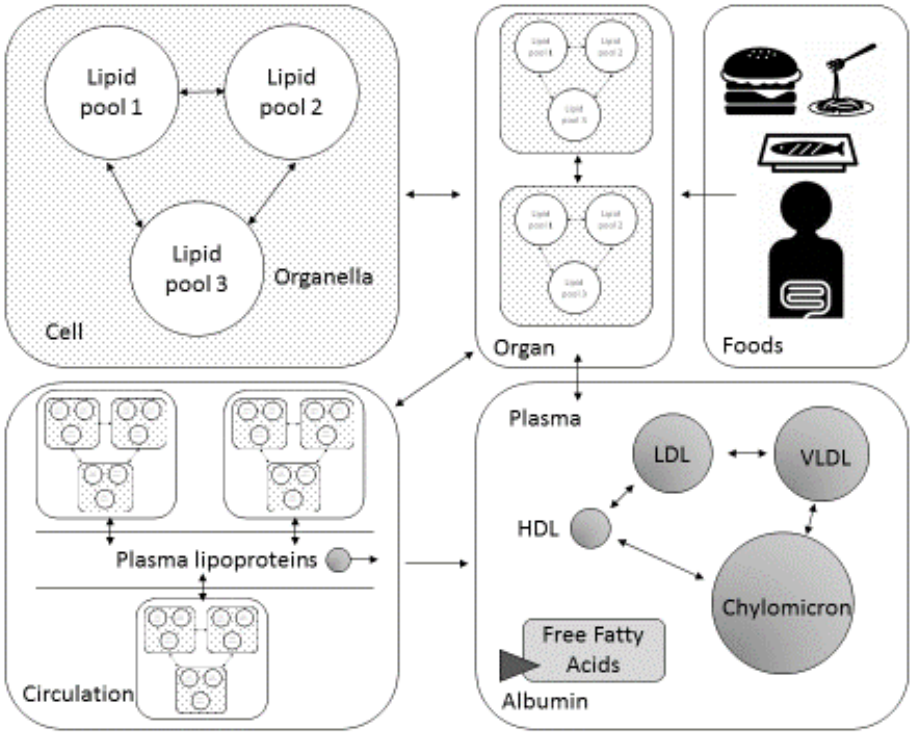

Fig. 1.1 Various lipid pools and their relationship

Currently, fatty acids are regarded as one of the most important molecules in various pathological conditions, they can provide us energy, cholesterol or other lipids, and they also have signaling function, and help cells interface. Therefore, fatty acids have an essential relationship with our body and health condition. By studying on fatty acids, we can have a good understanding on our health condition. Adjusting fatty acids intake in a scientific way is an economic and significant way to control our health condition (prevent and manage diseases), especially in remote and not well-developed areas.

Fatty acids are classified into free fatty acids and esterified fatty acids. Esterified fatty acids are further divided into cholesterol esters (CE), triglycerides, and phospholipids. According to the carbon chain-length and the number and location of double bond, CEs and triglycerides have also diversity in the structure. Moreover, fatty acids undergo various metabolic processes including dietary intake, enteric absorption, plasma lipoprotein metabolism, enzymatic modification (esterification, cleavage, elongation and desaturation of fatty acyl chain), and secretion as triglyceride in milk. Because of complexity of fatty acid, characteristics features of fatty acids in biological samples naturally generates a high-dimension dataset. Due to the variety and complexity in fatty acid dataset, it is hardly possible to extract unobserved but important factors in dataset. Therefore, it is necessary to use multivariate analysis to extract the unobserved factors from the dataset.

Exploratory data analysis is a procedure to interpret and analyze data to give information for further data study and gathering. We can find latent factors in dataset by exploratory data analysis with more ease. In exploratory data analysis, there are many classical methods, such as principal component analysis, factor analysis, and independent component analysis. It is not sufficient to induce unobserved factors only with classical methods, thus advanced methods, such as functional regression analysis and functional clustering, common principal component analysis are conducted. Exploratory data analysis is a large conception consisting of many statistical methods to deal with various dataset and give further information on more accurate data feature interpretation.

Therefore, exploratory data analysis has become important in analytical and clinical chemistry, especially for analyses of large and complicated biological or medical datasets generated from comprehensive mass spectrometry.

Previous epidemiological studies on fatty acid metabolism have not utilized exploratory data analysis, but have carried out only simple statistic approaches. For instance, fatty acid composition of CEs has been studied with special interests to 1) its correlation with other lipids or other sample characteristics such as gender effect and ageing (Chen, Z., et al., 2019), and 2) estimation of the risk for myocardial infarction (Öhrvall, M., et al., 1996) and diabetes (Vessby, B., et al., 1994), 3) search for new biomarkers of cancer (Zhang, Y., et al., 2016) and inflammation (Khorsan, R., et al., 2014), and 4) simple discriminant analysis of two or more groups of dataset (Jung, Y., et al., 2018). However, previous studies including the above literatures did not suggest inner, yet unobserved, and essential characteristics of complicated fatty acid metabolism. To attain this purpose in this study, we aim to apply more advanced exploratory data analysis methods including common principle component analysis to the datasets of fatty acids.

Furthermore, lipidomic analysis on milk fatty acid composition were reported on the basis of general multivariate data analysis (factor analysis, principal component analysis, and discriminant analysis) (Liu, H., et al., 2020; Correddua, F., et al., 2017). They led us to the understanding of the effect of geographical origins and lactation periods on the milk composition. However, these reports did not create the milk concentration changing trend as variables. Here, by generating derivatives as the changing trend of variables, we apply functional data analysis to a cow milk fatty acid dataset, and aim to reach further understanding of yet unobserved factors affecting milk fatty acid composition.

1.2 The purpose of our study

The purpose of our study is to discover the unobserved factors of fatty acid with exploratory data analysis. Unobserved factors are the latent and essential characteristics

of fatty acids cannot be seen by eyes. Fatty acids functions are based on those latent characteristics. We aim to improving medical companies profit by discovering the fatty acid characteristics. For example, median chain fatty acids have anti-inflammation characteristics, and contains less calories than long chain fatty acids. We can use exploratory data analysis to find unobserved factors to reveal the relationship between median chain fatty acids and environment factors to produce milk containing more median chain fatty acids and can be sold better.

We also aim to improve the public health including remote areas. Hokkaido is a large area but with little population, people living in Hokkaido is divided in several remote areas. People living in remote areas don't enjoy medical service as good as people in Sapporo. It also costs lot of money and time to arrive to the remote areas to study their health conditions. Therefore, we think using exploratory data analysis to discover the health conditions in remote areas by revealing the unobserved factors can reduce the cost of money and time to make the study of remote area health condition more effective and economic.

We aim to preventing disease with diet to reducing the government budget of public health expenditure. Our daily food contains plenty of fatty acids. Fatty acids are digested and transported in our body circulation. They have effect on our body health every day. If we can change the daily food intake, we can change our body health condition gradually, and we can prevent diseases. In this way we can also reduce reducing the government budget of public health expenditure because the cost of our daily food intake is lower than the cost of daily medical intake. This reduction in medical cost can relive the stress on public health expenditure of nation especially the developed countries.

There are two datasets in our study, which are serum fatty acid dataset, and milk fatty acid dataset. We used dimension reduction methods, which are principal component analysis, factor analysis, independent component analysis, and common component analysis to explore the first datasets. We used function data analysis, which are functional regression analysis and functional clustering analysis to analyze the last milk fatty acid

dataset and its corresponding environment dataset.

1.3 The structure of this study

In this study, there are six chapters. The first chapter is introduction. In the second chapter we introduced fatty acids and vitamin D knowledge including their structure, their circulation, and their functions. In the third chapter, we introduced exploratory data analysis, including the explanation about principal component analysis, factor analysis, independent component analysis, common principal component analysis, functional regression analysis and functional clustering analysis. In the fourth chapter, we conducted analysis of human serum fatty acid datasets with dimension reduction methods and the contribution of our study. In the fifth chapter we explained analysis of milk fatty acid datasets with functional data analysis and the contribution of our study. The sixth chapter is conclusion.

2. Knowledge of fatty acids

The utility of fat product for human beings from livestock, agricultural products and seafood has been in a very long history. After the industry revolution, human also started to use fat product as industry materials for daily life. Therefore, fat product or lipid is an essential part for human beings, and our study is mainly about the fatty acid composing fat product.

When majority of people hear about "fatty acid", they may think about something oil-like, or something that gives people energy and makes weight gain. However, a conception of fatty acid is actually so giant that people cannot get an impression about it with a single oral expression. It is certainly related with our daily life: we consume fatty acids, we use fatty acids, and we are actually made up of, partly, fatty acids. Thought the fat product consumed by human contains a lot of species, we only introduce two types of them for our study on fatty acids: fatty acid and triacylglycerols. They have a close relationship with each other—96 mas % of fatty acids are triacylglycerols (Leray, C., 2015).

With the large scale (various species) and complex structure, the chemical properties of fatty acids are complicated to be understood. There are mainly three parameters that affect fatty acids chemical properties: chain length, saturation, and first double bond position. These three parameters give fatty acids various characteristics.

2.1 Chain length of fatty acids

Fatty acids can be assorted into four groups by their chain length: short chain fatty acids, medium chain fatty acids, long chain fatty acids, and very long chain fatty acids (Leray, C., 2015). Chain length of fatty acids is depended on the number of carbon atoms in the acyl chain. Short chain fatty acids contain less than 6 carbon atoms, medium chain fatty acids contain more than 6 carbon atoms and less than 12 carbon atoms in their acyl chain. Long chain fatty acids contain more than 12 carbon atoms and less than 22 carbon

atoms in their acyl chain. Very long chain fatty acids contain more than 22 carbon atoms in their acyl chain.

2.2 Saturation

The fatty acids can also be assorted into two groups by their saturation: saturated group and unsaturated group (Leray, C., 2015). Saturated group contains fatty acids with no double bonds in their acyl chains, while unsaturated group contains fatty acids with double bonds in their acyl chain.

Saturated fatty acids

Fatty acids in the saturated group called saturated fatty acids. Because there is no double bond in the acyl chain, expression of saturated fatty acids is simple. There is a formula to express the general fatty acids

$$CH_3(CH_2)_zCOOH$$

If z is 6 then it is caprylic acid with 8 carbon atoms in its acyl chain, expressed as 8:0 where zero for no double bond, so caprylic acid is a medium chain and saturated fatty acid.

Because no double bond exists in their acyl chains, the physical characteristics of saturated fatty acids are majorly based on the chain length. The more carbon atoms in their acyl chains, the more oil-like saturated fatty acids are. For example, lignoceric acid, with 24 carbon atoms, are oil soluble and with 84.2 melting point, which means it is solid under the room temperature. But butyric acid, with 4 carbon atoms, are water soluble and with -5.3 melting point, which means it is liquid under the room temperature (Table 2.1).

The absorption routes of in our body fatty acids are also different because of the difference of carbon atom number in their acyl chain. For example, butyric acid is digested by intestinal bacteria in a quick speed, they barely make us obtain fat but only energy.

In nature the sources of different fatty acids are various, there is a table expressing the difference in major source of saturated fatty acids below:

Table 2.1: Fatty acids length, melting point and source

name	Carbon atoms	Melting point℃	Sources
butyric acid	4	-5.1	Animal oil plant oil
Caproic acid	6	-3.4	Goat milk
Caprylic acid	8	16.7	Coconuts milk, milk
Capric acid	10	31.3	Coconuts milk, milk
Lauric acid	12	43.5	Coconuts oil
Myristic acid	14	54.4	Coconuts oil animal oil
Palmitic acid	16	62.9	Coconuts oil animal oil, milk
Stearic acid	18	69.6	Animal oil
Arachidic acid	20	75.4	Peanut oil
Behenic acid	22	80.0	Behen (seed of Ben-oil-tree) oil
Lignoceric acid	24	84.2	Peanut oil, wood tar
Cerotic acid	26	87.7	Bees wax

Unsaturated fatty acids

Unsaturated fatty acids are much more complicated and diverse than saturated fatty acids, because of its complex structure. In order to give a better expression about unsaturated fatty acids, we need to explain the nomenclature of unsaturated fatty acids. The most widely used and well-known nomenclature of unsaturated fatty acids is defined by R.T. Holman, a physiologist, in 1964. He used a system, which is noted by p: z to define the chain length, number of double bonds. The letter p expresses the total number of carbon atoms in the acyl chain, and z expresses the total number of double bonds. Furthermore, he used omega-x to note the position of first double bonds. In this nomenclature system, omega means the position of first double bonds, the number of carbon atoms calculating from the mythel group. Different omega-x group defines different metabolic system of fatty acids. For example, erucic acid, with 22 carbon atoms,

1 double bond, and the first double bond is in the ninth carbon atom from mythel group, is expressed as fatty acid 22:1, omega-9 under this nomenclature. Unsaturated fatty acids are generally assorted by the Omega system for they share several similar properties in a same group.

Omega-9 group

One of the most important omega-9 fatty acids is oleic acid, with 18 carbon atoms and 1 double bond. It is abundant in our body, and it is the most largely existed fatty acid in our adipose tissues, and secondly large existence fatty acid in our body tissue. It is abundant in olive oil, pecan oil peanut oil, and sesame oil and so on, and largely exists in various animal fats, such as chicken and lard. The major use of oleic fatty acids is for food for its abundance on animal fats and plant oils. It is also used for soap as a fat product.

Omega-6 group

Three of the most representative fatty acids of omega-6 in our body are linoleic acid, gamma-linoleic acid and arachidonic acid. Linoleic acid is with 18 carbon atoms and 2 double bonds. It is called polyunsaturated fatty acid for its double bonds are more than one. It is rich in nuts, and can be derived from sesame seeds, corn, flax seeds, and so on. Gamma- linoleic acid is with 18 carbon atoms and 3 double bonds. It is the first derivative of linoleic acid and also a polyunsaturated fatty acid. It can be extracted from plant seeds as a medicine material. Arachidonic acid is with 20 carbon atoms and 4 double bonds. It is the final derivative of linoleic acid in human body omega-6 metabolic system and also a polyunsaturated fatty acid. It can be extracted from fish and animal fat as a medicine material. The relationship among those three fatty acids are linoleic acid desaturates to gamma- linoleic, and after series reaction, gamma- linoleic elongate and desaturates to arachidonic acid.

Omega-3 group

There are four the most representative fatty acids of omega-3 in our body, which are alpha-linoleic acid, stearidonic acid, Eicosapentaenoic acid (EPA) and Docosahexaenoic acid (DHA). Alpha-linoleic acid is with 18 carbon atoms and 3 double bonds, it is rich in nuts, and can be derived from walnuts, flax seeds, and many common vegetable oil. Stearidonic acid is with 18 carbon atoms and 4 double bonds, it is derived from alpha-linolenic acid by the enzyme. It can be derived from hemp seed oils. Eicosapentaenoic acid (EPA) is with 20 carbon atoms and 5 double bonds, it is derived from stearidonic acid. It can be extracted from fish oils. Docosahexaenoic acid (DHA) is with 22 carbon atoms and 6 double bonds, it is derived from EPA. It can also be extracted from fish oils. DHA and EPA are proved to be practical to nervous diseases, such as Alzheimer's disease (Leray, C., 2015). Besides DHA is famous for its unique metabolic system. The relationship among those three fatty acids are stearidonic acid desaturates from alpha-linoleic acid. Stearidonic acid extend its acyl chain and desaturates to become EPA. EPA elongate and desaturates to become DHA.

2.3 Triglyceride

Triglyceride is a combination of three fatty acids and one glycerol. It is the main form of fatty acids in human body. It enables the bidirectional transformation of adipose fat between liver and skin. Nature oils and fats are consisting of triglycerides generally. Here is an equation below for Triglyceride

$$HOCH_2CH(OH)CH_2OH + RCO_2H + R'CO_2H + R''CO_2H \rightarrow RCO_2CH_2CH(O_2CR')CH_2CO_2R'' + 3H_2O$$

A triglyceride molecule is generally made up of two or three kinds of fatty acids molecules. The different combination of diverse fatty acids generates the diversity of triglycerides. For example, in milk fat, there are nearly 200 kinds of triglycerides observed.

2.4 Vitamin D

Vitamin D is a group of secosteroid, which can solute in fat. It is rich in flesh fish fats and mushrooms. In humans, the most important compounds in this group are vitamin D_3 (also known as cholecalciferol) and vitamin D_2 (ergocalciferol). People can obtain vitamin D by UV radiation or oral ingestion. It is metabolized in our body as 25-hydroxycholecalciferol.

2.5 Digestion and nutrition

Triglyceride

Absorption

The triglycerides from meal will be hydrolyzed in duodenum after food intake. Then there are two different absorption metabolic routes for fatty acids in triglyceride form according to their chain length.

For short chain fatty acids, usually less than 10 carbon atoms in their acyl chain, the metabolic routes are simple. They generally absorbed directly in the intestine then to liver and give no extra storage in the adipose tissues. Such fatty acids are mainly from milk or coconut milk in our daily meal. They (no more than 6 carbon atoms) not only give us energy without extra weight gain, but also adjust the absorption of water and sodium in the intestine.

For the longer chain fatty acids, situation is more complicated. Firstly, triglycerides from meal are hydrolyzed in duodenum, then become triglycerides again or other form (cholesterol esters, phospholipids). Such products will be transported to liver, adipose tissues and other organs in the form of chylomicrons, which is a more complicated form in the lymph. In this process, the hydrolyzation is based on their saturation condition. The saturated or monounsaturated fatty acids are stored in adipose tissues generally, the rest of saturated fatty acids might be also desaturated to become unsaturated fatty acids. Omega-6 polyunsaturated fatty acid linoleic acid will be desaturated and elongated to become other fatty acids in omega-6 group, and finally the arachidonic acid which is the compound of membrane. Omega-3 polyunsaturated fatty acid EPA and DHA will be the component of membrane or oxidized to give us energy. Same as the linoleic acid in omega-6 group, alpha-linoleic acid will be desaturated and elongated to become other fatty acids in omega-3 group.

Transport

The triglycerides are transported by lipoprotein. Lipoprotein is consisting of protein and lipids. There are five principal lipoproteins in our plasma, which are chylomicrons, very large density lipoproteins (VLDLs), intermediate-density lipoproteins (IDLs), large density lipoproteins (LDLs), and high density lipoproteins (HDLs).

Chylomicrons is the first step synthesis whose density is 0.93. It contains 86% Triglyceride, 3% cholesterol esters, 2% cholesterol, 7% phospholipids, and 2% protein. It is synthesized in the intestine. VLDLs are the second step synthesis whose density is 0.95 to 1.010. It contains 55% Triglyceride, 12% cholesterol esters, 7% cholesterol, 18% phospholipids, and 8% protein. IDLs are the third step synthesis whose density is 1.008 to 1.019. It contains 23% Triglyceride, 29% cholesterol esters, 9% cholesterol, 19% phospholipids, and 19% protein. LDLs are the fourth step synthesis whose density is 1.019 to 1.060. It contains 6% Triglyceride, 42% cholesterol esters, 8% cholesterol, 22% phospholipids, and 22% protein. HDLs are the fifth step synthesis whose density is 1.125 to 1.210. It contains 3% Triglyceride, 13% cholesterol esters, 4% cholesterol, 25% phospholipids, and 55% protein. The transportation of lipoprotein in metabolic system is so complicated that we decided to divide our explanation into three parts as many experts do: the exogenous pathway, endogenous pathway, and reverse cholesterol transport.

Exogenous pathway

In this pathway, dietary lipids are hydrolyzed in intestine into chylomicrons. There is triglyceride in chylomicrons and transported in plasma. Some part of triglyceride hydrolyzed from chylomicrons are changed into adipocytes for energy storage, the others are transported to other tissues for energy use.

With the decrease of triglycerides in chylomicrons, they become VLDLs. VLDLs are transported into liver by certain receptors. After some metabolism and the decrease of proportion of triglycerides and increase of proportion of protein, VLDLs changes into HDLs, then get out of liver.

Endogenous pathway

Free fatty acids exported from VLDLs in plasma to other tissues, such as muscles, brains for energy production. In this pathway, DHA is the only fatty acid can be transported to brain for energy use, and it is a very special fatty acid in metabolic system. In endogenous pathway the rest of free fatty acids are stored in adipocytes for energy storage.

After losing free fatty acids, VLDLs become IDLs, IDLs transport to liver and macrophages by certain receptors. By releasing some free fatty acids in plasma for energy use and storage, IDLs become LDLs. LDLs are also transported to liver and tissues by corresponding receptors.

Reverse cholesterol transport

HDLs, which change from VLDLs, get out of peripheral tissues and are with cholesterols in it. HDLs uptake cholesterol and cholesterol esters from other tissues including liver, kidneys, intestines and macrophages by specific receptors. In plasma, cholesterol in one HDL will be esterified, which makes cholesterol esters exchange with triglycerides in other HDLs to make this HDL become VLDL and LDL. In liver the cholesterol will be transformed into bile and secreted by bile fluid to intestines.

Nutrition and digestion of saturated fatty acids

Fatty acid experts believe that from the point of physical view, the intake of saturated fatty acids is of no necessity, because we human body can produce the saturated fatty acids, we need, by ourselves. But they also think that a proper intake of saturated fatty acids is also of no harm especially for infants and people in a high speed of growth.

The main purpose of saturated fatty acids intake is for energy use, membrane, and protein acylation (Table 2.2). Protein acylation, acylation of protein by palmitic acid, is an essential part of nervous system. That process can only be done by saturated fatty acids.

The metabolism of saturated fatty acids with different the acyl chain length is also different. Short chain fatty acids (no more than 4 carbon atoms in the acyl chain) and median chain fatty acids (6 carbon atoms to 12 carbon atoms in the acyl chain) are different from long chain fatty acids (more than 12 carbon atoms in the acyl chain). Besides even in same chain length group, different fatty acids might even occur different effects. For example, palmitic acid (16:0) will cause inflammation, cholesterolemia, and atherosclerosis. However, myristic acid (14:0) gives hypocholesterolemia effect at the same places.

Although we can make saturated fatty acids in our body, but fatty acids experts highly recommend we intake them for the health effects of saturated fatty acids.

Table 2.2: Function of fatty acids

Function	Fatty acids			
	Saturated fatty acids	Monounsaturated Fatty acids	Polyunsaturated fatty acids omega-6	Polyunsaturated fatty acids omega-3
Energy use	○	○	○	○
membrane	○	○	○	○
Eicosanoid precursor	×	×	○	○
Protein acylation	○	×	×	×
Second messenger	×	○	○	○

Nutrition and digestion of unsaturated fatty acids

Omega-9 group

In our daily life, the most widely used omega-9 fatty acid is oleic acid (18:1, omega-9). It is mainly from animal fats, dairy products, and vegetable oils, which are common in our life.

Although, in some experiment without significant evidence, the consumption of oleic acid was proved to be related to the improvement of blood marker of cardiovascular disease, it is still considered to be a relatively healthy monounsaturated fatty acid (Table 2.3). Oleic fatty acid is generally considered to replace part consumption of saturated fatty acid, because oleic acid is a kind of acid with high stability compared with other polyunsaturated fatty acids such as linoleic acid (16:0, omega-6). According to several experiments, oleic acid was observed to have relationship with decreasing cholesterol in LDLs and HDLs. Therefore, oleic acid is considered to be an ideal fatty acid in our daily cooking. Besides, based on certain long-term experiments on oleic acid and dietary functioning food, there is a prove that oleic acid is helpful for fat losing. It is considered that oleic acid can be not only used for obesity curing, but also its related diseases, such as atherogenic disease as well as cardiovascular disease.

In the world, the 3.5%-22% energy intake is by oleic acid. In Japan, people use adjusted oleic fatty acid sunflower oil or other adjusted oil. The nutritionist of ANSES organization recommend an ideal daily intake for oleic acid, which is 42 to 55g/day. The upper limit is for the potential cardiovascular disease risk factors. The lower limit is for the replacement of saturated fatty acids to prevent obesity and its related diseases.

Table 2.3: Relationship between diseases and exchange of MFA to SFA

	ratio	P value
pathogenesis of cardiovascular disease		
Exchange MFA to SFA	1.19(1.00, 1.42)	0.32
death of cardiovascular disease		
Exchange MFA to SFA	1.01(0.73,1.14)	0.18

Omega-6 group

Linoleic acid is the root of gamma-linoleic acid and arachidonic acid. It is abundant in vegetable oils, pork fat and milk. Linoleic acid tansformed into gamma-linoleic acid. Gamma-linoleic acid elongate and desaturated into arachidonic acid, which is very important for our body tissues. Arachidonic acid is the material for thromboxanes and prostaglandins after cyclooxygenases metabolism. The prostanoids metabolized from arachidonic acid doesn't have anti-inflammation property. Linoleic acid in our body will be either stored in adipose tissues as triglycerides or used by other body tissues as cellular membrane materials in the form of phospholipids. In 2008, WHO informed that the intake of linoleic acid should be 5% to 8% of the total energy intake.

Gamma-linoleic acid, which is in some particular plat oils, is synthesized from linoleic acid by desaturation process. Since it can be synthesized form linoleic acid, no specific consumption is needed. In our body, it is used in skin for its barrier function as well as limit the epidermal hyperproliferation. Nutritionists suggest that gamma-linoleic acid might cause inflammation process such as diabetes and Alzheimer's disease.

Arachidonic acid considered as the only essential fatty acid in omega-6 series. It is rich in egg, animal fats, and can be found in fish oil as well. Since it can be synthesized form Gamma-linoleic acid, therefore no specific consumption is needed. Arachidonic acid

is also considered to cause inflammation process in our body, such as allergic reaction and eczema.

Besides, nutritionist suggested that Overtake of omega-6 fatty acids might give a negative influence on the intake of omega-3 fatty acids, which means that the overtake of omega-6 fatty acids might lead to the risk of pathologies relating brain and its related diseases.

Omega-3 group

Linolenic acid is the root of omega-3 group fatty acids. It desaturates and changes into stearidonic acid (18:4, omega-3). Stearidonic acid elongates into arachidonic acid, and arachidonic acid desaturates into EPA. After some specific metabolism synthesis, EPA changes into DHA. In this metabolic process, derivative action of omega-3 group (linolenic acid to EPA) and derivative action of omega-6 group (Linoleic acid to arachidonic acid) share the same enzyme, therefore omega-3 group and omega-6 group are in a competition relationship. Furthermore, linoleic acid suppresses the transform action of linolenic acid to stearidonic acid. Thus, Nutritionists inform that we should not only take care of the intake of omega-6 fatty acids but also the ratio of omega-6fatty acids/omega-3 fatty acids. The proper number of this ratio should be 1~2 (the number among Japanese is 3).

As far as we know, in human body, the transformation rate to arachidonic acid is no more than 3%. Although transformation rate of linolenic acid to EPA is no more than 10%, and as for transformation rate of linolenic DHA it is 0.5~1%, and arachidonic acid has a negative influence on the transformation from EPA to DHA. Therefore, the limitation of omega-6 intake is very important.

We should also pay attention to the derivatives of arachidonic acid, EPA, and DHA. A part of derivatives of EPA are series 3 prostaglandin and series 5 leukotriene. They have weak inflammation properties. Some derivatives of arachidonic acid are series 2 prostaglandin and series 4 leukotriene, which lead to inflammation. Meanwhile lipoxin A

from arachidonic acid, resolvin E from EPA, maresin R and protectin D from DHA have anti-inflammation properties, thus the balance of omega-6 group and omega-3 group intake is important even in the derivative point of view (Table 2.4).

Table 2.4: Function and distribution of eicosanoid

eicosanoid	function	distribution
lipoxin A	anti-inflammation	Several tissues (arachidonic acid)
resolvin E	anti-inflammation	Several tissues (EPA)
maresin R	anti-inflammation	Several tissues (DHA)
protectin D	anti-inflammation	Several tissues (DHA)
resolvin D	anti-inflammation	Several tissues (DHA)

The recommended intake of linolenic acid is 0.8 to 1.1 g/day

The recommended intake of EPA + DHA is 300 to 500 mg / day

2.6 Lipids and health

Saturated fatty acids

The intake of fatty acid is an essential part in our daily life, for it not only bring us the energy but also affect our metabolism. As for saturated fatty acids, the effect of them is actually only based on the chain length for there are no double bonds in the acyl chain. Therefore, to make it clear, we set the saturated fatty acids into three parts: long chain fatty acids, short and median chain fatty acids.

Long chain fatty acids

Cardiovascular diseases

High level of concentration of cholesterol in our plasma is proved to be the factor leading to cardiovascular diseases, therefore factor increase the concentration of cholesterol on plasma is considered to have relationship with cardiovascular diseases. Long saturated fatty acids intake was demonstrated clearly to have the property of increasing the concentration of cholesterol plasma while the short chain and unsaturated fatty acids was not. There is an equation to express the relationship between cholesterol and saturated fatty acids:

$$\Delta TC = 32 \times \Delta SFA - 6 \times \Delta MUFA - 21 \times \Delta PUFA + 31 \times \Delta TFA,$$

Where TC means total plasma cholesterol, SFA means saturated fatty acids as a percentage of total calories, MUFA means monounsaturated fatty acids, PUFA means polyunsaturated fatty acids as a percentage of total calories, TFA means trans fatty acids as a percentage of total calories. From the equation we can know that intake of long chain saturated fatty acid has a positive relationship with cardiovascular diseases. Therefore, nutritionists recommend that the substitution of intake of long chain saturated fatty acids should be done by unsaturated fatty acids.

Metabolic diseases

The relationship between long chain fatty acids and metabolic diseases is well established by scientists. In this field, the insulin resistance syndrome, which leads to type 2 diabetes, is proved to have positive relationship with long chain saturated fatty acids intake. When long chain saturated fatty acids replaced by unsaturated fatty acids, obesity, dyslipidemia, insulin resistance and other diseases in relationship with metabolic disorder decrease. Such relationship is even in gene level. Long chain fatty acids are in positive relationship with gene by affecting the transcription of genes in adipocytes. Such a relationship leads to inflammation process with long chain fatty acids overtake.

Cancer

Although, the intake of saturated fatty acids doesn't show the direct relationship with colorectal cancer, but the intake of saturated fatty acid has a direct relationship with energy intake, which has a direct positive relationship with colorectal cancer. Therefore, it is recommended by nutritionist that we should take care of the total intake of long chain saturated fatty acid.

Short chain fatty acids

Short chain fatty acids' metabolism system is not like long chain fatty acids, they are absorbed in the intestine mucosa by intestinal flora and transported in portal vein without the combination with chylomicrons. The destination of them is liver. In liver they are used as a gluconeogenesis material. SCFA not consumed by the liver has been said to be used as an energy or fat substrate in peripheral tissues. However, recent studies have shown that a small amount of SCFA in the blood affects energy metabolism through SCFA receptors (GPR41 and GPR43) expressed on adipocytes and sympathetic nerves and has anti-obesity and anti-diabetic effects.

It is well established that the risk of carcinogenesis is related to the decrease of

intestinal flora and butyrate fatty acid. Because butyrate fatty acid, which is same as other short chain fatty acids, is transported without chylomicrons, therefore it has a negative relationship with obesity.

Medium chain fatty acids

It has been well proved that medium-chain fatty acids have high solubility for water and are able to hydrolyze in water. Therefore, by the time it reaches the duodenum, most of medium chain fatty acids exist in the form of free fatty acids and there is no need to degrade pancreatic lipase. Medium-chain fatty acids have high affinity for water. They are transported directly to the liver through the portal vein, without forming bile acids and micelle. Medium chain fatty acids have been used in many diseases' treatment concerning about malabsorption syndrome for its metabolic pathways and rich in energy.

Medium-chain fatty acids oxidize faster than long-chain fatty acids, therefore diets induced temperature calculated from oxygen consumption and carbon dioxide is higher than DIT of long chain fatty acid. Medium-chain fatty acids reach the liver directly via the portal vein and are quickly metabolized. It has been established that there is no increase in blood triglycerides and chylomicrons after the intake of medium chain fatty acids. Thus, medium-chain fatty acids have the effect of increasing heat production and inhibiting triglycerides (cholesterol) after meals. It can be expected that long-term intake of medium-chain fatty acids can induce the effect of suppressing body fat accumulation compared to general edible oils.

Unsaturated fatty acid

Omega-9 group

The most representative of omega-9 group is oleic acid, which is abundant in olive oils and well seen in daily life. Therefore, we only introduce oleic acid.

Cardiovascular diseases

After the research of A. Keys done in 1980 (Leray, C., 2015)about the relationship between monosaturated fatty acids and cardiovascular diseases, it has been clearly established that the monounsaturated fatty acids have nothing to do with the increase of cardiovascular diseases incidents. Furthermore, in the aspect of nutrition, it has been well believed that a diet rich of monounsaturated fatty acids, including oleic acid, have lower probability of leading to cardiovascular disease than a diet rich of saturated fatty acids.

Cancer

Because the lack of corresponding accurate experiments, we only have a rough conclusion that the increase use of oleic acid from olive oil is linked to the decrease in the risk of colon cancer and breast cancer. Meanwhile the increase use of oleic acid from meat origins is linked to the increase in the risk of breast cancer as saturated fatty acids. From the two result, we conclude nothing but there are some factors in olive that cause the decreasing risk in some cancers. The corresponding accurate research is in need to be done in future.

Metabolic diseases

E.K. in 2011 (Leray, C., 2015) showed that the intake of oleic acid has positive relationship with the increase of TNF-a secretion of insulin, which has a relationship with type 2 diabetes. Besides, clinical trials about Mediterranean diet, rich of olive oil, showed that diet rich of oleic acid improve the sensitivity of insulin and glycemic against insulin-resistant subjects, while the diet rich of saturated fatty acids don't have this functional effect on metabolic process.

Omega-6 group

There are three most representative fatty acids in omega-6 group, which are linoleic acid, gamma-linoleic acid and arachidonic acid.

Cardiovascular diseases

It has been proved that linoleic acid has a positive effect on lower the cholesterol in plasma. Linoleic acid was reported that it also has an effect of decreasing the risk of sudden death by heart attack and stroke. Gamma-linoleic acid has been reported that its derivatives have the beneficial effect on blood vessels, platelet aggregation, and inflammation. Besides, treatment with gamma-linoleic acid have been used in the field of skin cares, such as dermatitis and eczema. On the contrary, arachidonic acid has a negative effect on decreasing the inflammation risk. Excessive accumulation of arachidonic acid will lead to inflammation diseases and cardiovascular diseases. Considering it is the derivatives of linoleic acid and gamma-linoleic acid, intake of fatty acids in omega-6 group is suggested to be in limit.

Cancers

Among 40 years studies, it is generally accepted that excessive intake of linoleic acid has an effect on carcinogenesis of various tissues. It promotes the growth of cancer cells. Although dihomo-gamma-linoleic acid has the effect of inhibiting tumor growth, but there is still no strong evidence that it has the anti-cancer effect. Therefore, the excessive intake of omega-6 fatty acid is not suggested. The total energy intake from omega-6 fatty acids should between 5% and 10% of the total daily energy intake.

Omega-3 group

There are three most representative fatty acids in omega-6 group, which are linolenic acid, EPA and DHA.

Coronary heart diseases and atherosclerosis

Although there are some conflicting results in experiments, but it is well accepted by nutritionists that omega-3 group fatty acids have anti-inflammation property. Such a property leads to the decrease of atheromatous plaque and has a positive relationship with slowing blood flow. Therefore, intake of omega-3 group fatty acids has an effect on preventing coronary heart diseases and atherosclerosis. Fish meals rich in EPA and DHA was recommended to prevent coronary heart diseases in a famous experiment. Besides, many studies showed that there is a potential possibility that the preventing of stroke has a positive relationship with diet rich in linolenic acid from vegetable oils.

Inflammation

The majority of derivatives from DHA and EPA are anti-inflammatory. Several experiments also showed that omega-3 group can slow down the inflammatory process. Therefore, diseases from inflammatory process, for example asthma, has a negative relationship with omega-3 group. In treatment of some chronic inflammatory diseases, Cronhn's disease for example, omega-3 group fatty acids is considered to be used

Metabolic

It is reported that high triglycerides and low HDL-cholesterol has relationship with omega-3 fatty acids intake. In this case omega-3 group fatty acids are considered to have different effect on metabolic function. EPA and DHA have effect on triglycerides synthesis and linolenic acid has effect on lipoprotein synthesis. In this way, n-3 fatty acids are believed to have a positive effect on reducing obesity and the related diseases, such as cardiovascular and type 2 diabetes. However, it is still difficult to demonstrate more accurate relationship between metabolic syndrome and omega-3 fatty acids, because of the diverse lifestyle of people in clinical trials.

Alzheimer's disease

For decades, it has been proved that omega-3 fatty acids especially DHA, have positive effects on the nervous process. Not only dose omega-3 group fatty acids benefit the brain but also eye nerves. Such effects show that omega-3 group fatty acids intake will benefit the cognitive development and reduce the risk of nervous aging diseases, such as Alzheimer's disease and dementia.

Vitamin D

Vitamin D is playing an important role in our health. It controls calcium absorption and metabolism in our body. By controlling the calcium metabolism, vitamin D controls the order of metabolism of our bodies.

Calcium metabolism

The main role of vitamin D is increasing the intestinal absorption of calcium and phosphate. The shortage of vitamin D would cause bone diseases and disorder of mineral metabolism. The lack of vitamin D might also cause the disorder of muscle function.

Metabolic diseases

The lack of vitamin D is considered to be related with two kinds of diabetes, type 1 diabetes and type 2 diabetes. Type 1 diabetes is insulin dependent and considered to be in inverse relation with vitamin D. vitamin D is observed as playing a protection role against inflammatory agents. Type 2 diabetes, resulted from cellular insulin sensitivity, is also considered to be in an inverse relationship with vitamin D, because vitamin D is associated with insulin resistance.

3. Knowledge of exploratory data analysis

3.1 Functional data analysis

Functional data analysis is to make data functional (Ramsay, T.O., 2005; Ramsay, T.O. & Silverman, B.V., 1982). But how to make data functional? We may ask. The process of functionalization is explained below.

3.1.1 Functionalization

Smoothing and interpolation

Measured values are observation measured in intervals.

$$y_1, \dots y_j, \dots, y_n.$$

In order to functionalize them, we need to judge whether they are errorless or not. If they are errorless, then we use the interpolation process. If they are not errorless, we need to use smoothing process.

Difference between interpolation and smoothing

Interpolation is a process assuming every data point is useful and errorless, then connect every point of observed data (Ramsay, T. O., 2005). It can be displayed below:

$$y_j = x(t_j),$$

where $x(t_j)$ is the function.

Smoothing is a process removing the effects of useless or error point to just make the points in observed data smoothing with penalty. It can be displayed below:

$$y_j = x(t_j) + \varepsilon_j,$$

where $x(t_j)$ is the function, ε_j is the error. Errors in the equation above might give influence on the functionalization.

In this equation, the variance and covariance of $x(t_j)$ is considered to be 0 matrix Σ_0, therefore the variance-covariance of y_j equals the variance covariance of ε_j. Such a relationship can be expressed as

$$\mathrm{Var}(y) = \mathrm{Var}(\varepsilon) + \sigma^2.$$

Generally speaking, in order to reduce the influence of ε_js, we use the size of variance-covariance matrix of ε_j or y_j, since they are equal.

Selection of basis function

Basis function is a statistically independent linear combination with a basis number of K. It can be expressed as below:

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t),$$

where $\phi_k(t)$ is the basis function. There are many different basis function series, such as spline function basis, Fourier basis and so on. Every basis function has its own characteristics, and data also has its own characteristics. Therefore, what we need to do is to find out the basis function with the best fitness to the observed dataset. Furthermore, since data has its own characteristics, we need to find the suitable K, which is the number of basis functions, for the particular observed dataset. Practically, it is a common sense that the smaller the K is the better the model is considered. There are four reasons below:

1. The model will be easy to calculate. Large K might not give influence that negative to some models such as Fourier basis, which is easy to calculate. However, such a K is a disaster to other models with basis function that are hard to calculate, for example, power basis.
2. The more degree of freedom we can use in our hypothesis, since the degree of freedom of hypothesis is on a negative relationship with the number of basis functions.
3. It is more possible to represent the characteristic of the observed data without overfitting.
4. It is more possible to avoid difficult calculation on the level of derivatives. Since we make data into function, therefore it has the characteristics of function. That means it also has derivatives. Thus, the difficulty in calculating derivatives should also be considered.

Fourier basis function

Fourier basis function is based on the Fourier function series. In this case, the basis function is

$$\hat{x} = f(t) = c_0 + c_1 \sin\omega t + c_2 \cos\omega t + c_3 \sin 2\omega t + c_4 \cos 2\omega t + \cdots,$$

or

$$f(t) = c_0 + \sum_{n=1} a_n \cos\left(\frac{nt\pi}{L}\right) + \sum_{n=1} b_n \sin\left(\frac{nt\pi}{L}\right),$$

where L is the half of the period of the function, $\omega = \frac{\pi}{L}, a_n = \mathrm{c}_{2n}, \text{and } b_n = c_{2n-1}$. Since Fourier basis is periodical, therefore the ω in this equation determines the period. Fourier basis function is one of the traditional basis functions in functional data analysis for its advantage in calculating and periodic characteristics, we can both obtain the coefficients c_n and $f(t_j)$ at time t_j in $O(n\mathrm{Log}n)$.

Derivatives of Fourier function basis are below:

$$Da_n \cos\left(\frac{nt\pi}{L}\right) = -a_n \frac{n\pi}{L} \sin\left(\frac{nt\pi}{L}\right)$$

$$Db_n \sin\left(\frac{nt\pi}{L}\right) = b_n \frac{n\pi}{L} \cos\left(\frac{nt\pi}{L}\right)$$

Deficits of Fourier basis function

Although Fourier basis function has advantages, there is also a disadvantage of Fourier basis function. Fourier basis function is not recommended to be used for unstable dataset where data are not in a similar order, for example dataset of nondurable goods production of the United States.

Spline basis function

As for observed data not periodic, spline basis function is the most frequently recommended basis function. In this basis function, every subinterval should be represented by a polynomial with order m_l.

$$\hat{x} = f(t, m_l) = c_0 S_0(t, m_0) + c_1 S_1(t, m_1) + c_2 S_2(t, m_2) + \cdots + c_L S_L(t, m_l),$$

or

$$\hat{x} = f(t) = \sum_{l=0}^{L} c_l S_l(t, m_l).$$

where $S_l(t, m_l)$ is a polynomial. Spline basis can be express as a basis function that has several characteristics below:

1. Spline basis in every subinterval is a spline function.
2. The combination of spline basis in every subinterval should also be a spline function.
3. Any spline basis in this linear combination can be expressed by the other spline basis in a linear combination mode.

Select interval (breakpoint) of basis function

The first step of functionalization in spline basis function is to select breakpoints. Breakpoints are the points that put in interval to divide the interval into several subintervals. In this thesis, we set the number of subintervals to be l, and breakpoints in interval to be τ_l, where l=$1, \ldots, L-1$.

Smoothness

In the smoothness part, all the subintervals should be joint with each other smoothly. Therefore, function values should be set to equal to the adjacent part of subintervals. Furthermore, the derivatives up to $m_L - 2$ order should also joint with each other smoothly, which means that values of derivatives up to $m_L - 2$ order should be set to equal to the adjacent part of subintervals.

Degree of freedom in spline basis function

The degree of freedom of spline basis function is calculated as the formula below:

$$m_L + L - 1,$$

where m is the total order of spline basis, and $L-1$ is the total number of intervals. There for we can see that a spline basis function with order m_L and $L-1$ break points is a piecewise polynomial function whose degree of freedom is $m_L + L - 1$. Besides, the piecewise and their first $m_L - 2$ derivatives must be continuous with $L-1$ distinct knots, and first $m_L - L - 1$ derivatives must be continuous with $L-1$ coincident knots.

B-spline function

B-spline basis function is based on the spline basis function. For a given spline function, it can be expressed as below

$$B_{i,n}(x) = \begin{cases} 0, & if\ x \le t_i\ \ or\ x \ge t_{i+n} \\ 1, & otherwise \end{cases}.$$

A b-spline basis function is function above with an additional constraint that is $\sum_i B_{i,n}(x) = 1,$ for all x. B-spline function act as a basis function of the spline function space, and the general structure of B-spline function is defined below,

$$B_{i,0}(x) = \begin{cases} 0, & if\ x \le t_i\ \ or\ x \ge t_{i+n} \\ 1, & otherwise \end{cases},$$

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(x).$$

The relationship between B-spline and spline function is that a linear combination of B-spline function can form spline function as

$$S_{t,n}(x) = \sum_i \alpha_i B_{i,n}(x).$$

Wavelets basis

Wavelets basis is another basis function for periodic dataset. Besides, wavelets basis function can deal with data discontinued and rapidly changing. It is transformed from a mother wavelets basis function, which is in the form of equation below:

$$\psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k),$$

Where $\psi(2^j x - k)$ is a suitable mother wavelet function, j and k are integers. $\psi_{jk}(x)$ is orthogonal with two constraints that is zero mean and one square norm. The two conditions can be expressed below:

$$\int \psi_{jk}(x)dx = 0,$$

$$\int |\psi_{jk}(x)|^2 dx = 1.$$

Wavelets basis function can be easily applied to a bonded interval for periodic observation. $\psi_{jk}(x)$ is wavering around the position $2^{-j}k$ at frequencies around $2^j c$ for some constant c with scale 2^{-j}.

Exponential and power bases

Power bases consists of several power functions,

$$t^{\lambda_1}, t^{\lambda_2}, \dots, t^{\lambda_k} \dots.$$

Exponential bases consist of several exponential functions,

$$e^{\lambda_1 t}, e^{\lambda_2 t}, \dots, e^{\lambda_k t}, \dots.$$

Polynomial basis

In functional data analysis, polynomial basis is another classic basis function, which can be expressed as

$$\psi_k(x) = (t - \omega)^k, k = 0, \dots, K,$$

where ω is considered as the center of approximated interval. There is a disadvantage

in polynomial basis function, which is the calculation of large K.

Other function basis

Although, we just introduced limited number of basis functions, but there are several basis functions we have no time to do a further introduction, such as step-function basis, constant basis.

We discuss the main part of functional data regression analysis with least squares. We will divide the part into two parts the unweighted least squares fits and the weighted least squares fits.

We first introduce the basic equation in fitting process. For a given observation

$$y_j, j = 1, \dots, n,$$

we can use the model

$$y_j = x(t_j) + \varepsilon_j$$

to express the relationship between the basis function and the observation, where $x(t_j)$ is the basis function we propose and ε_j is the error between the value from function and observation at j-th period. Besides $x(t_j)$ is the value of j-th period from the function $x(t)$, and $x(t)$ can be expressed in the form of

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \mathbf{c}^{\mathrm{T}} \boldsymbol{\phi}.$$

The last part of this equation is in the form of vector and matrix, where $\mathbf{c}$ is a vector with k scalars acting as coefficients and $\boldsymbol{\phi}$ is a matrix with k rows expressing k variables.

Unweighted least square fits is the most ordinary and simple fitting process for functional data analysis. In this process, we assume that ε_j is in a distribution independent and identical. In this distribution, ε_j is with zero mean and constant variance σ^2. We determine the coefficient vector $\boldsymbol{c}$ with lowest least squares criterion by equation below

$$\mathrm{SMSSE}(\boldsymbol{y}|\boldsymbol{c}) = \sum_{j=1}^{n} \left[y_j - \sum_{k=1}^{K} c_k \phi_k(t) \right]^2.$$

In the form of vector this equation can be expressed as below

$$\mathrm{SMSSE}(\boldsymbol{y}|\boldsymbol{c}) = (\boldsymbol{y} - \boldsymbol{\Phi c})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\Phi c}) = \|\boldsymbol{y} - \boldsymbol{\Phi c}\|^2.$$

In order to find out the minimum value of least squares criterion, we need to use the derivative of least squares criterion. Their derivatives with respect to $\boldsymbol{c}$ can be expressed as

$$\mathrm{SMSSE}'(\boldsymbol{y}|\boldsymbol{c}) = 2\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{c} - 2\boldsymbol{\Phi y}.$$

When the derivative of least squares criterion with respect to $\boldsymbol{c}$ is set to $\mathbf{0}$ as

$$\mathrm{SMSSE}'(\boldsymbol{y}|\boldsymbol{c}) = 2\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{c} - 2\boldsymbol{\Phi y} = \mathbf{0},$$

we will have the estimate coefficient vector minimizing the least squares criterion which is

$$\hat{\boldsymbol{c}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}.$$

Hence the vector $\hat{\boldsymbol{y}}$ we approximate can be expressed as

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{\Phi}\hat{\hat{\boldsymbol{c}}},$$

where $\hat{\boldsymbol{y}}$ is the vector of approximated values, and $\boldsymbol{y}$ is the vector of observations. Besides $\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}$ is always set as $\mathbf{S}$, and the relationship between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ can be written as

$$\hat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{y},$$

or more precisely

$$\hat{\boldsymbol{y}} = \hat{x}(t_j) = \sum_{l=1}^{n} S_j(t_l)y_l = \mathbf{S}\boldsymbol{y}.$$

In the equation above, $\mathbf{S}$ is a smoothing matrix to project y into $\hat{\boldsymbol{y}}$.

Weighted least square fits is slightly different from unweighted least square fits for the additional weighted matrix in the equation of least squares criterion, which can be expressed as

$$\mathrm{SMSSE}(\mathbf{y}|\mathbf{c}) = (\boldsymbol{y} - \boldsymbol{\Phi c})^{\mathrm{T}}\mathbf{W}(\boldsymbol{y} - \boldsymbol{\Phi c})$$

where W is the weighted matrix and it is symmetric and positive definite. Since we already know that the relationship between y_j and $x(t_j)$ can be expressed as

$$y_j = x(t_j) + \varepsilon_j,$$

then we can see that $\mathbf{W} = \boldsymbol{\Sigma}_{\mathrm{e}}^{-1}$. We can calculate the weighted least square estimate $\hat{\boldsymbol{c}}$ of the coefficient vector $\mathbf{c}$ as

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}\boldsymbol{y}.$$

where $\widehat{\boldsymbol{y}}$ is the vector of approximated values, and y is the vector of observations. Additionally, $\boldsymbol{\Phi}(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{W}$ is always set as $\mathbf{S}$, and the relationship between $\widehat{\boldsymbol{y}}$ and $\boldsymbol{y}$ can be written as

$$\widehat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{y},$$

or more precisely

$$\widehat{\boldsymbol{y}} = \hat{x}(t_j) = \sum_{l=1}^{n} S_j(t_l) y_l = \mathbf{S}\boldsymbol{y}.$$

In the equation above, $\mathbf{S}$ is a smoothing matrix to project y into $\widehat{\boldsymbol{y}}$.

3.1.2 Functional clustering

Cluster analysis is considered as an exploratory analysis method based on distance. Functional clustering is a combination of clustering and functional data analysis, therefore functional clustering analysis is dependent on the distances between functions.

In general clustering analysis, distance can be determined in different methods. Assuming there are a set of observed data, which are $\boldsymbol{x} = (x_1, x_2, \dots, x_p), \boldsymbol{y} = (y_1, y_2, \dots, y_p)$, the distance between two elements can be expressed in different ways, such as:

1. Euclidean distance:

$$d_{euc}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2};$$

2. Manhattan distance:

$$d_{L^1}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} |x_i - y_i|.$$

After combined with functional data analysis, functions can be inserted into distance calculation. Assuming there are two functions, f_i and g_i. The distance between two

functions is denoted as $d(f_i, g_i)$ .There are many different methods to calculate distance, such as

1. Euclidean distance (L^2 norm):

$$d_{euc}(f_i, g_i) = \sqrt{\int (f_i(t) - g_i(t))^2 dt};$$

2. Manhattan distance (L^1 norm):

$$d_{L^1}(f_i, g_i) = \int |f_i(t) - g_i(t)| dt.$$

Since we can transform data into functions, therefore we can use the characteristics of function, their derivatives, to calculate the distance. Therefore, we can obtain another two distances, which are

Euclidean distance (L^2 norm):

$$d_{euc}\left(\frac{df_i(t)}{dt}, \frac{dg_i(t)}{dt}\right) = \sqrt{\int \left(\frac{df_i(t)}{dt} - \frac{dg_i(t)}{dt}\right)^2 dt};$$

Manhattan distance (L^1 norm):

$$d_{L^1}\left(\frac{df_i(t)}{dt}, \frac{dg_i(t)}{dt}\right) = \int \left|\frac{df_i(t)}{dt} - \frac{dg_i(t)}{dt}\right| dt,$$

where $\frac{df_i(t)}{dt}$, $\frac{dg_i(t)}{dt}$ are the derivatives of $f_i(t)$ and $g_i(t)$ correspondingly.

Agglomerative hierarchical clustering is a clustering technique that making data points, considered to be individual clusters in first level, to merge into larger clusters in upper levels as long as they are below the linkage criterion. The steps data process, can be summarized as below:

Step1. Treat every data point as one cluster

Step2. Compute the distance between clusters

Step3. Set a criterion

Step4. Merge clusters, whose distance are below the linkage criterion, into larger clusters

Step5. Repeat from Step2 to Step 4 until all cluster merged into one

---

In clustering process, choosing the criterion is important. In general, people

choose the ward's method. In ward' method, we need to minimize the cost of merging two clusters (https://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf). The cost of two clusters merging into one cluster is defined as

$$\Delta(A,B) = \sum_{i\in A\cup B} \|\vec{x_i} - \vec{m}_{A\cup B}\|^2 - \sum_{i\in A} \|\vec{x_i} - \vec{m}_A\|^2 - \sum_{i\in B} \|\vec{x_i} - \vec{m}_B\|^2,$$

which can also be defined as

$$\Delta(A,B) = \frac{n_A n_B}{n_A + n_B} \|\vec{m}_B - \vec{m}_A\|^2,$$

where $\vec{m}_i$ is the center of i-th cluster. In agglomerative hierarchical clustering, merging cost starts with zero initially, then it is increasing as the process continues. In this process, two cluster with minimized merging cost will be merged together.

3.1.3 Time series data analysis

Time series data analysis is designed to analyze a series of observation of an ordered sequence at equally spaced time intervals (Ihaka, R., Time Serires Analysis, 2005, https://www.stat.auckland.ac.nz/~ihaka/726/notes.pdf). Assuming there are observations of T times, which can be denoted as $y_1, \dots, y_T$. For every observation, there is a relationship between unobserved factors and white noises can be expressed below

$$y_t = \eta_t + \varepsilon_t, t = 1, \dots, T,$$

where η_t is the value of unobserved factors of t-th time interval and ε_t is the white noise of t-th time interval.

In time series data analysis, there are main two goals below

1. First goal is to develop a model, which generates the observation equally located in T time intervals, to obtain the feature of unobserved factors.
2. Second goal is to use the model expressing feature of observation to forecast or to monitor the observation.

Two main kinds of stationary time series

There are two main kinds of stationary time series data analysis: stationary time series and covariance stationarity times series (Brockwell. J. P. & Davis, A, R., 2002).

For a time series $\{y_{t1}, \ldots, y_{tk}: k > 0\}$, if the distribution of it is similar to the distribution of times series $\{y_{t1+\mu}, \ldots, y_{tk+\mu}: \ k > 0 \ \text{ and } \ \mu > 0\}$, it is called as stationary series. Therefore, in stationary time series, $\mu(t)$, the mean of y_T, is independent of t.

$$\mu_y(t) = \mu_y(0)$$

and

$$\gamma_y(t, t+\mu) = \mathrm{Cov}(y_t, y_{t+u}) = \gamma_y(\mu, 0).$$

For a time series $\{y_{t1}, \ldots, y_{tk}: k > 0\}$, if $E|Y_t|^2 < \infty$, mean $\mu_y(t)$ and autocovariance $\gamma_y(t, t+\mu)$ is independent of t and $t+\mu$, then it can be called covariance stationary series.

Three kinds of time series models

There are mainly three kinds of time series models we mainly use in time series analysis, which are autoregressive (AR) series, moving average (MA) series and autoregressive moving average (ARMA) series.

AR series can be defined as below

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t,$$

where $\phi_1, \ldots, \phi_p$ are constant and ε_t is white noise.

MA series can be defined as below

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\varepsilon_t, \ldots, \varepsilon_{t-q}$ are the white noises.

ARMA series can be defined as

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

It is a combination of AR series and MA series.

Time series data analysis is an analysis method similar to the function data analysis. In time series data analysis, the argument of the function is limited to the time. On the other hand, in function data analysis, various functions are used.

3.2 Dimension reduction methods

3.2.1 Principal component analysis

Principal component analysis is a dimension reduction method often used in data analysis. It is a process transforming correlated variables into scores that are linearly uncorrelated variables with orthogonal method (Mardia, K.V., et al., 1994). Therefore, each component is orthogonal to other components. Additionally, in this process, the first transformed component has the largest variance, and each component has the higher variance than the succeeding components.

For $\boldsymbol{x}_i = (x_{i1}, \dots x_{ip}), i = 1, \dots, n$, the observed matrix $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)^{\mathrm{T}}$ with it mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$ and its covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x}}$, there exist an orthogonal matrix $\boldsymbol{\Gamma}$, which can make the equation below

$$\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Gamma} = \boldsymbol{\Lambda},$$

where Λ is a diagonal matrix, whose diagonal elements have relationship below

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

There is a transformation relationship between $\boldsymbol{\Gamma}$ and $\mathbf{X}$, which can be expressed below

$$\mathbf{y} = \boldsymbol{\Gamma}^{\mathrm{T}}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu})$$

or

$$\boldsymbol{y}_i = \boldsymbol{\gamma}_i^{\mathrm{T}}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu})$$

where the $\boldsymbol{\gamma}_i$ is called the i-th principal component loading and also the i-th column of $\boldsymbol{\Gamma}$ and $\mathbf{1}$ is a column vector with n ones.

3.2.2 Factor analysis

Factor analysis is a statistical method used to transform variability among observed, correlated variables to a potentially lower number of unobserved variables which we call it common factors (Mardia, K.V., et al., 1994). In this transform process, we need to set the exact number of unobserved factors firstly, then based on the exact number of unobserved factors, we conduct the transformation to calculate the loading values of

common factors and uniqueness of each factors. Such a process can be explained by the equation below

$$\mathbf{X} - \mathbf{1}\boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon},$$

where $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)^{\mathrm{T}}$ is the observed matrix $\boldsymbol{x}_i = (x_{i1}, \dots x_{ip})$, $i = 1, \dots, n$ are the observed variable vectors with means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$. $\mathbf{1}$ is a column vector with n ones. $\mathbf{L}$ is the loading matrix and $\mathbf{F}$ is the common factor matrix $\mathbf{F} = (\boldsymbol{F}_1, \dots, \boldsymbol{F}_p)$. In the equation, $\boldsymbol{\varepsilon}$ represents the uniqueness value of each variable.

In factor analysis, there are three assumption about $\mathbf{F}$ below

1. $\mathbf{F}$ and $\boldsymbol{\varepsilon}$ are independent
2. $\mathrm{E}(\mathbf{F}) = 0$.
3. $\mathrm{Cov}(\mathbf{F}) = \mathbf{I}$.

According to the three assumption above, we know how to calculate the loading value matrix, which can be expressed as

$$\mathrm{Cov}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}) = \mathrm{Cov}(\mathbf{LF} + \boldsymbol{\varepsilon}).$$

then

$$\mathrm{Cov}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}) = \mathbf{L}\mathrm{Cov}(\mathbf{F})\mathbf{L}^{\mathrm{T}} + \mathrm{Cov}(\boldsymbol{\varepsilon}),$$

$$\mathrm{Cov}(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}) - \mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathbf{L}\mathbf{L}^{\mathrm{T}},$$

where $\mathbf{1}$ is a column vector with n ones.

Three indices in factor analysis

There are three indices we used in factor analysis on fatty acid dataset for number selection in model process.

Root Mean Square Error (RMSE)

In linear regression analysis, RMSE is the most basic criterion to adjust how well the model fit the observation (Mardia, K.V., et al, 1994). It is defined as

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{\left(z_{f_i} - z_{o_i}\right)^2}{N}},$$

where z_{f_i} is the i-th data in forecast model and z_{o_i} is the i-th observed data. Judging from RMSE, we can see the difference between the model and observation, large value means large gap between the model and observed data.

Root mean square error of approximation (RMSEA) index

RMSEA is defined as below:

$$RMSEA = \sqrt{\max\left(\frac{F\left(S, \Sigma(\hat{\theta})\right)}{df} - \frac{1}{N-1}, 0\right)}$$

where $F\left(S, \Sigma(\hat{\theta})\right)$ is the minimum if the fit function, which is of χ^2 distribution, df is its degrees of freedom, N is the sample size.

The RMSEA is bounded below by zero. Approximation is acceptable when RMSEA value is less than 0.10, The RMSEA values 0.08 and 0.10 as a mediocre fit is considered as mediocre fit between 0.05 and 0.08 is considered as an adequate fit, and values less than 0.05 is considered as a good fit (Preacher, K. J. et al., 2013).

Tucker-Lewis Index (TLI) of factoring reliability

TLI is defined as below:

$$TLI = \frac{\left(\frac{\chi_a^2}{df_a}\right) - \left(\frac{\chi_{null}^2}{df_{null}}\right)}{\left(\frac{\chi_a^2}{df_a}\right) - 1},$$

where χ_a^2 is the χ^2 distribution of the independent model, χ_{null}^2 is the χ^2 distribution of null hypothesis, df_a is the number of degrees of freedom of the independent model, df_{null} is the number of degrees of freedom of the null hypothesis. TLI values ranges from 0 to 1, and higher value means better fitness, TLI value larger than 0.95 are interpreted as acceptable fit (Hooper, D., et al, 2007).

3.2.3 Independent component analysis

Independent component analysis is a process that transform the observed dataset to independent variables under the assumption that the observed dataset coming from several independent sources (Hyvärinen, A., & Oja, E., 2000). In other word, this process is aimed to find out the independent data source of observed dataset.

In this process, there are two assumptions, which are:

1. The dataset sources are independent of each other.
2. The dataset sources are in non-Gaussian distributions.

This process is based on central limit theorem and aimed to

1. Minimize mutual information.
2. Maximize non-Gaussianity

For a given observation vector $\boldsymbol{x}_i = \left(x_{i1}, \dots, x_{ip}\right)^{\mathrm{T}}$, there exist an independent source vector $\boldsymbol{s} = (s_1, \dots, s_r)^{\mathrm{T}}$, and the two vectors has a relationship below:

$$\boldsymbol{x}_i = m_{i1}s_1 + \cdots + m_{ik}s_k + \cdots + m_{ir}s_r,$$

or

$$\boldsymbol{x}_i = \sum_{k=1}^{r} m_{ik}s_k = \boldsymbol{s}^{\mathrm{T}}\mathbf{M}$$

where $\mathbf{M}$ is the mixing matrix transforming source vector to observation vector, and $m_{i,j}$ is the scalar from mixing matrix $\mathbf{M}$. Our main object is to find an unmixing matrix to convert the former transformation and the independent source vector. Since both source vector and unmixing matrix are unknown, the gradient descent method is recommended in this calculation process.

In independent component analysis, we use kurtosis as a measure of non-Gaussianity, where $\boldsymbol{y}$ is a matrix contains $\boldsymbol{y}_i$, an independent vector, and

$$\boldsymbol{y}_i = \mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i.$$

where $\mathbf{W}$ is the unmixing matrix and. Generally, the higher the kurtosis value is, the less the Gaussianity property the distribution has.

3.2.4 Common principal component analysis

Common principal analysis is designed to reduce dimension of dataset with several groups (Flury, B. N., 1984). In order to explain the definition of common principal analysis, it is necessary to introduce several notations.

Firstly, assuming there are k groups and p variables in a dataset, which can be expressed as

$$\mathbf{X}_{k\times p} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k]^{\mathrm{T}},$$

or

$$\mathrm{X}_{k\times p} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{k1} & \cdots & x_{kp} \end{bmatrix}.$$

We can see from the matrix above that in the dataset, in every $\mathbf{X}_i$, there exist p variables. Therefore, the average vector and covariance matrix of $\mathrm{X}_{k\times p}$ can be expressed as

$$\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_k]^{\mathrm{T}}$$

and

$$\mathrm{Cov[X]} = \Phi_{kp\times kp} = \left[\Phi_{ij}\right]_{i,j=1,\dots,k} = \begin{bmatrix} \Phi_{11} & \cdots & \Phi_{1k} \\ \Phi_{21} & & \Phi_{2k} \\ \vdots & \ddots & \vdots \\ \Phi_{k1} & \cdots & \Phi_{kk} \end{bmatrix}.$$

After all the notation, we now can introduce common principal component analysis. Assuming there are $k \times p$ variance-covariance matrixes, which can be expressed below

$$\Phi_{kp\times kp} = \left[\Phi_{ij}\right]_{i=1,\dots,k,j=1,\dots,k}.$$

In order to use dimension reduction methods to all k groups, we choose to use common principal component analysis. In this method, there is a matrix $\mathbf{B}$, which makes

$$\mathbf{B}^{\mathrm{T}}\Phi_{ij}\mathbf{B} = \Lambda_{ij} = diag\left(\Lambda_{ij,1},\dots,\Lambda_{ij,p}\right),$$

where $\mathbf{B} = \left[\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_p\right]$ and is a $p \times p$ orthogonal matrix, $i = 1,\dots,k,$ and $j = 1,\dots,k.$ for all k groups.

We calculate B by the maximum likelihood function. Assuming there are k groups vectors X_i $(i = 1,\dots,k)$, whose number of variables are p. The $\mathbf{X}_i$ distributes as normal distribution $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i \in R^p$ and $\boldsymbol{\Sigma}_i$ is a definite positive matrix. Its sample variance $\mathbf{S}_i(i = 1,\dots,k)$ distributes as $W_p(n_i, \boldsymbol{\Sigma}_i)$, therefore the likelihood function of S_i is

$$L(\boldsymbol{\Sigma}_1,\dots,\boldsymbol{\Sigma}_k) = C \times \prod_{i=1}^{k} \exp\left(-\frac{n_i}{2}\boldsymbol{\Sigma}_i^{-1}\mathbf{S}_i\right)|\boldsymbol{\Sigma}_i|^{-\frac{n_i}{2}},$$

and the log likelihood function is

$$\mathrm{g}(\boldsymbol{\Sigma}_1,\dots,\boldsymbol{\Sigma}_k) = \sum_{i=1}^{k} n_i\left(\log|\boldsymbol{\Sigma}_i| + tr\boldsymbol{\Sigma}_i^{-1}\mathbf{S}_i\right),$$

where $\boldsymbol{\Sigma}_i = \sum_{j=1}^{p}\boldsymbol{\beta}_j'\Lambda_{ij}\boldsymbol{\beta}_j,\ i = 1,\dots,k.$ Thus, $\log|\boldsymbol{\Sigma}_i| = \sum_{j=1}^{p} log\,\Lambda_{ij}$ and $tr\boldsymbol{\Sigma}_i^{-1}\mathbf{S}_i = \sum_{j=1}^{p}\boldsymbol{\beta}_j'S_i\boldsymbol{\beta}_j/\Lambda_{ij}\,, i = 1,\dots,k.$

It means that $\mathrm{g}(\boldsymbol{\Sigma}_1,\dots,\boldsymbol{\Sigma}_k)$ can be expressed as

$$\mathrm{g}\left(\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_p,\Lambda_{i1},\dots,\Lambda_{ip}\right) = \sum_{i=1}^{k} n_i\left(\sum_{j=1}^{p} log\,\Lambda_{ij} + \sum_{j=1}^{p}\boldsymbol{\beta}_j'\mathbf{S}_i\boldsymbol{\beta}_j/\Lambda_{ij}\right).$$

Since B is orthogonal matrix, therefore there are two restrictions that

$$\boldsymbol{\beta}_{\mathrm{h}}^{\mathrm{T}}\boldsymbol{\beta}_j = \begin{cases} 0, & if\ h \neq j \\ 1, & if\ h = j \end{cases}.$$

Thus, we can use the Lagrange multipliers to obtain the value of B and Λ_{ij}, which makes $L(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_k)$ maximum and $\mathrm{g}(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_k)$ minimum.

The Lagrange multiplier formula can be expressed as below:

$$\mathrm{G}(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_k) = \mathrm{g}(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_k) - \sum_{h=1}^{p} \gamma_h \left(\boldsymbol{\beta}_h^{\mathrm{T}} \boldsymbol{\beta}_h - 1\right) - 2 \sum_{h<j}^{p} \gamma_{hj} \boldsymbol{\beta}_h^{\mathrm{T}} \boldsymbol{\beta}_j,$$

or

$$\mathrm{G}(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_k) = \sum_{i=1}^{k} n_i \left(\sum_{j=1}^{p} log\, \Lambda_{ij} + \sum_{j=1}^{p} \frac{\boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{S}_i \boldsymbol{\beta}_j}{\Lambda_{ij}} \right) - \sum_{h=1}^{p} \gamma_h \left(\boldsymbol{\beta}_h^{\mathrm{T}} \boldsymbol{\beta}_h - 1\right) - 2 \sum_{h<j}^{p} \gamma_{hj} \boldsymbol{\beta}_h^{\mathrm{T}} \boldsymbol{\beta}_j.$$

After we derivative the equation with respect to $\boldsymbol{\beta}_h$, Λ_{ij}, and γ_{hj}, we obtain three equation below

$$\sum_{i=1}^{k} \frac{n_i \mathbf{S}_i \boldsymbol{\beta}_j}{\Lambda_{ij}} - \sum_{h=1, h \neq j}^{p} \gamma_h \boldsymbol{\beta}_h - \gamma_j \boldsymbol{\beta}_j = 0,$$

$$\Lambda_{ij} = \boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{S}_i \boldsymbol{\beta}_j,$$

$$\gamma_{hj} = \sum_{i=1}^{k} n_i.$$

After we combine the three equations above, we can obtain the equation below:

$$\sum_{i=1}^{k} \frac{n_i \mathbf{S}_i \boldsymbol{\beta}_j}{\Lambda_{ij}} - \sum_{h=1, h \neq j}^{p} \gamma_h \boldsymbol{\beta}_h - \sum_{i=1}^{k} n_i \boldsymbol{\beta}_j = 0.$$

If we multiply the equation above by $\boldsymbol{\beta}_l$, we can obtain that

$$\sum_{i=1}^{k} \frac{n_i \boldsymbol{\beta}_l \mathbf{S}_i \boldsymbol{\beta}_j}{\Lambda_{ij}} = \gamma_{jl},$$

and if we switch l and j, we can obtain

$$\sum_{i=1}^{k} \frac{n_i \boldsymbol{\beta}_j \mathbf{S}_i \boldsymbol{\beta}_l}{\Lambda_{il}} = \gamma_{lj} = \sum_{i=1}^{k} \frac{n_i \boldsymbol{\beta}_l \mathbf{S}_i \boldsymbol{\beta}_j}{\Lambda_{ij}},$$

$$\boldsymbol{\beta}_l' \left(\sum_{i=1}^{k} \frac{n_i (\Lambda_{il} - \Lambda_{ij}) \mathbf{S}_i}{\Lambda_{il} \Lambda_{ij}} \right) \boldsymbol{\beta}_j = 0, j, l = 1, \dots, p, j \neq l.$$

The equation above can be solved by the algorithm invented by professor Flury and Gautschi (1986). After we solve the equations, we can gain the common principal eigenvectors $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p]$ and common eigenvalues $\Lambda_{ij,1}, \ldots, \Lambda_{ij,p}$.

4. Analysis on serum fatty acid dataset

4.1 Serum fatty acid dataset

Blood samples of free and total fatty acid

The present study was a cross-sectional study conducted as a work of the Dynamics of Lifestyle and Neighborhood Community on Health Study (DOSANCO Health Study). Briefly, the DOSANCO Health Study was a community-based study conducted in Suttu town, Hokkaido, Japan, during the year of 2015. A total of 2,100 participants (977 men and 1,123 women; 79.6% of all residents aged 3 years or more other than those living at nursing homes) completed a self-administered questionnaire. Of the 2,100 participants, 729 participants between the ages of 35 and 79 years were additionally asked to provide blood samples, and 545 participants (245 men and 300 women) complied (Nakamura. K. et al. 2018). The study protocol was approved by the ethic committees of the Faculty of Medicine (15-002, 16-007) and the Faculty of Health Sciences (16-10), Hokkaido University. Written informed consent was obtained from all participants.

Blood was drawn after an overnight fast. After blood coagulation at room temperature, serum was separated by centrifugation at 4°C, and stored at -80°C for no longer than 3 years before analysis. The samples were confirmed to be stable at this condition.

Datasets of free and total fatty acid

Since fatty acid measurement is commonly performed for total and free fatty acids separately in clinical medicine, we apply multivariate analysis to both total and free fatty acid datasets. Total fatty acid dataset contains the subtypes of 25-hydroxyvitamin D_3 and 16 total fatty acids 4:0, 6:0, 12:0,14:0, 16:0, 18:0, 18:1,18:2, 18:3, 20:0, 20:4, 20:5, 22:6, 22:0, 24:0, and 26:0. Free fatty acid dataset contains the subtypes of 25-

hydroxyvitamin D_3 and 12 free fatty acids 4:0, 6:0, 12:0, 14:0, 16:0, 18:0, 18:1, 18:2, 18:3, 20:4, 20:5, and 22:6 . The latter dataset lacks free fatty acids 20:0, 22:0, 24:0, and 26:0 because they were not detected.

Total fatty acids and free fatty acids are affected by diet, and therefore, they are potential confounders. However, since the information of diet was not available in this study, we didn't discuss about the confounding.

Measurement of free and total fatty acid

Serum fatty acids were determined by liquid chromatography/tandem mass spectrometry (LC/MS). Fatty acids were labeled in serum with 2-nitrophenylhydrazine hydrochloride to improve sensitivity and recovery, and then extracted in organic solvents. Serum 25-hydroxyvitamin D_3 was determined by LC/MS after extraction in organic solvents. The results of validation studies for both assays were excellent. The absolute concentration (μmol/L in serum) was determined and used for multivariate analysis (Miura, Y., et al., 2017).

Blood samples of cholesterol esters

Blood samples were obtained after an overnight fast from the persons (n = 545; 300 women and 245 men; aged 35 to 79 years old) who participated in general health examinations in the year of 2015 in Suttsu town of Hokkaido, Japan. Suttu town has a population of 3,100 approximately, and its main industry is fishery. After blood coagulation at room temperature, serum was separated by centrifugation at 4° C, and stored at -80° C for no longer than 3 years before analysis. The samples were stable at this conditions. Ethical approval was obtained by the ethic committee of the Faculty of Medicine (15-002, 16-007) and the Faculty of Health Sciences (16-10), Hokkaido University, and informed consent was obtained from each participant.

Datasets of cholesterol esters

The CE dataset contained 8 subtypes of CE molecules: cholesteryl palmitate (CE 16:0), cholesteryl stearate (CE 18:0), cholesteryl oleate (CE 18:1), cholesteryl linoleate (CE 18:2), cholesteryl linolenate (CE 18:3), cholesteryl arachidonate (CE 20:4), cholesteryl eicosapentaenoate (CE 20:5) and cholesteryl docosahexaenoate (CE 22:6).

Sample pretreatment of cholesterol esters

Serum (20 μL) was mixed with ethanol (200 μL) containing a mixture of ISs (1.2 nmol each) and added with hexane (1200 μL) and distilled water (1000 μL). After centrifugation at 1500$\times g$ for 10 min, the organic layer was collected and dried under vacuum (Tomy centrifugal concentrator, Tokyo, Japan). Then, the residue was dissolved in isopropanol (300 μL) and filtered using a centrifugal filtering device (PVDF 0.1 μm; Merck Millipore Ltd., Carrigtwohill, Ireland). Finally, the sample was then diluted 8-fold with isopropanol, and 5 μL was injected to LC-MS/MS (Miura.Y. et al. 2017).

Chemicals and reagents of cholesterol esters

Cholesterol and all LC/MS grade solvents including methanol, 2-propanol, n-hexane, and water purchased from FUJIFILM Wako Pure Chemical Corporation (Osaka, Japan). Ammonium acetate was purchased from Sigma-Aldrich (St. Louis, MO, USA). Other chemicals and reagents were purchased from Kanto Chemical Industry (Tokyo, Japan). Palmitic acid (FA16:0), stearic acid (FA18:0), oleic acid (FA18:1), FA18:2, linolenic acid (FA18:3), and 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide were purchased from Tokyo Chemical Industry Co., Ltd. (Tokyo, Japan). Arachidonic acid (FA20:4) and docosahexaenoic acid (FA22:6) were purchased from Sigma-Aldrich Co., LLC. (MO, USA). Both CE and $^{2}H_{3}$-CE (IS) were prepared as described previously, and were stored stable at −80 °C for at least 3 years.

LC/MS/MS

The detailed conditions LC/MS/MS for CE species and the results of validation studies were reported previously. Briefly, the assay was conducted using a TSQ Quantum Access MAX (Thermo Fisher Scientific, Inc., Walthum, MA, USA) linked to a Thermo Finnigan Surveyor HPLC System. LC separation was carried out on an Accucore C18 (Thermo Fisher Scientific, Inc.) at 40° C. MS detection was implemented in a selected reaction monitoring (SRM) mode under condition of atmospheric pressure chemical ionization in positive ion mode. To determine the calibration curves, the mix standards of CE solution with 0.01, 0.04, 0.2, 1, 2, 4, 10, and 20 μmol/L were prepared. Each standard solution contained 1 μmoL of the $^{2}H_{3}$-CEs. The integration of the peak area and the plotting of each calibration curve were carried out using Xcalibur 2.0.7.

Serum sample (20 μL), added with 12 μmol/L $^{2}H_{3}$-labeled CEs 16:0, 18:0, 18:1, 18:2, 18:3, 20:4, 20:5, and 22:6 as internal standards, was mixed with ethanol (200 μL), hexane (1200 μL), and deionized water (1,000 μL). Then, centrifugation (1,500×*g*, 10 min) was conducted to separate the organic layer and the aqueous layer. The collected organic layer was concentrated in a centrifugal evaporator, and redissolved in isopropanol (300 μL). The sample was then diluted 8-fold with isopropanol, and 5 μL was used for LC/MS/MS. Fine separation of all CE species and excellent results of validation studies were reported previously. CVs and recoveries ranged 1.6% - 6.6%, and 94.4% - 107.2%, respectively.

Data processing

The integration of the peak area and the plotting of each calibration curve using 1/x weighting were performed by Xcalibur 2.0.7. (Thermo Fisher Scientific, Inc., Waltham, MA, USA).

4.2 Analysis on free fatty acid and total fatty acid dataset

4.2.1 Purpose

The aim of this study is to detect the latent factors of free fatty acids and total fatty acids as well as the characteristics of fatty acids, with the aid of dimension reduction methods such as principal component analysis (PCA), factor analysis, and independent component analysis (ICA) (Chen, Y., et al., 2019). To explain the reason of separating free and total fatty acids, free fatty acid is different from the rest of fatty acid species (esterified fatty acid) in molecular forms (mainly esterified vs. non-esterified for total and free, respectively), and also in localization in plasma (bound to lipoproteins vs. albumin). For this reason, separated evaluation of total and free fatty acids is common in clinical medicine. This study consists of the data of short-chain, medium-chain, long-chain, and very long-chain total and free fatty acids. The wide range of fatty acids will be a merit and novelty.

We also aim to find out possible relationship between fatty acids and 25-hydroxyvitamin D_3, an indicator of bodily storage of vitamin D. Similar to fatty acid, vitamin D is a lipophilic compound and can be synthesized de novo or taken as foods. Therefore, possible relationship between them is presumed. This report will show for the first time the relationship of 25-hydroxyvitamin D_3 with the wide range of total and free fatty acids. This study will provide the methods to extract valuable information from complex datasets of fatty acids and vitamin D. The expected results might be useful for prevention and management of fatty acid-related health disorders.

4.2.2 Result of analysis

We analyzed fatty acid dataset with three dimension reduction methods, which were factor analysis, PCA, and ICA. We applied three dimension-reduction methods to total fatty acid dataset and free fatty acid dataset respectively. Each outcome reflects basic characteristics of total fatty acids and free fatty acids.

Factor analysis

We applied factor analysis to both total and free fatty acid datasets, using the two methods, parallel analysis and BIC method, to decide the number of factors.

Factor analysis in total fatty acid dataset

In the application to total fatty acids dataset, the number of factors selected by parallel analysis was 5, and that selected by BIC values was 8. We set 5 and 8 as the factor number. Model with 8 factors gave a better fitness on the basis of RMSR index, RMSEA index, and Tucker-Lewis Index (Table 4.1).

Table 4.1: Comparison between BIC value and parallel analysis

Comparison of factor number in four indices in total fatty acids dataset		
Factor number	5	8
RMSR	0.03	0.01
RMSEA	0.136	0.068
Tucker-Lewis index	0.78	0.945

In the axis of Factor 1, the subtypes with high loading values were in the order of total fatty acids 16:0, 18:1, 18:0, 18:3, and 14:0. They belong to the group with high concentrations in the serum, and the order of their mean concentrations was quite similar with their order in size (table above). Exceptionally, total fatty acid 18:2 in this axis was with a low loading value. That is explained by the small standard deviation for total fatty acid 18:2, which is known to give a direct influence on the outcome of orthogonal rotation method. Thus, we consider Factor 1 as the size Factor. In the axis of Factor 2 (Fig.4.1), the subtypes with high loading values were total fatty acids 20:5 and 22:6, and vitamin D. Fatty acids 20:5 and 22:6 are called as eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), respectively. They are Omega-3 fatty acids enriched in fatty fish or fish oil. Vitamin D is also taken from fish among Japanese.

Thus, we consider Factor 2 as the representation of fish oil.

In the axis of Factor 3 (Fig.4.2), the subtypes with high loading values were total fatty acids 26:0, 24:0, and 22:0. They are saturated very long chain fatty acids. We consider Factor 3 as the representation of acyl chain length or fat solubility. The longer acyl chain gives the higher fat solubility (Patton. J. S., 1984). Additionally, double bond, if ever, decreases fat solubility. We observed the similar trends between the loading values and the acyl chain lengths.

In the axis of Factor 4 (Fig.4.2), the subtypes with high loading values were free fatty acids 20:0 and 22:0. They both are saturated fatty acids sharing relatively high fat solubility. Therefore, we consider Factor 4 as the representation of fat solubility in a limited range limited by yet unspecified factor(s). Intake of these fatty acids is associated with lower risk of metabolic syndrome, but its background mechanism is not known.

In the axis of Factor 5 (Fig.4.3), the subtypes with high loading values were total fatty acids 18:2 and 20:4, both belonging to omega-6 group. Fatty acid 18:2 is a major fatty acid in plant lipids, and fatty acid 20:4 is a major component of membrane phospholipids obtained from diet. Therefore, we consider Factor 5 as the representation of omega-6 group intake.

In the axis of Factor 6 (Fig.4.3), the subtypes with high loading values are total fatty acids 12:0 and 14:0, which are usually from coconut oil. Therefore, one may consider factor 6 as the representation of coconut oil. However, Japanese people do not obtain a large volume of coconuts oil. Moreover, other fatty acids enriched in coconut oil, such as fatty acids 8:0 and 10:0, did not show high loading values. Therefore, we consider Factor 6 as the representation of a unique but yet unspecified metabolic pathway shared by the two fatty acids.

In the axis of Factor 7, subtypes with high loading values are total fatty acids 4:0 and 6:0. They are short chain fatty acids derived from gut microbiota. Therefore, we consider Factor 7 as the representation of short chain fatty acids or the contribution of gut microbiota.

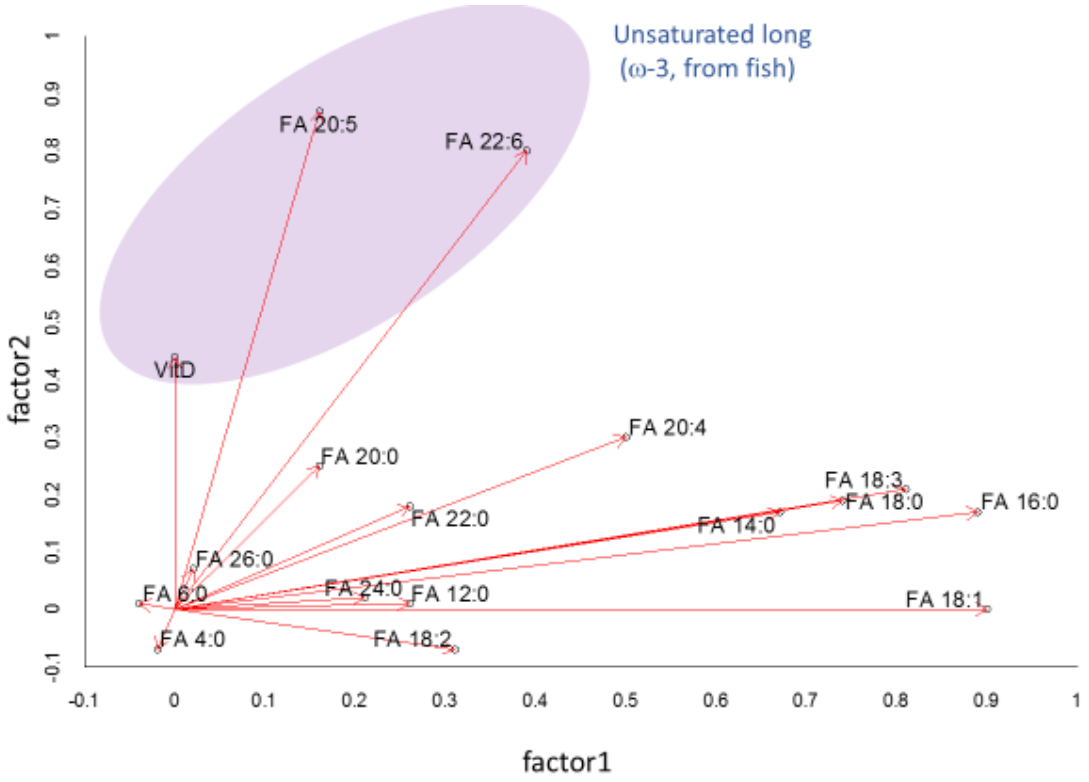

Fig.4.1 Result of total fatty acid dataset (factor analysis)

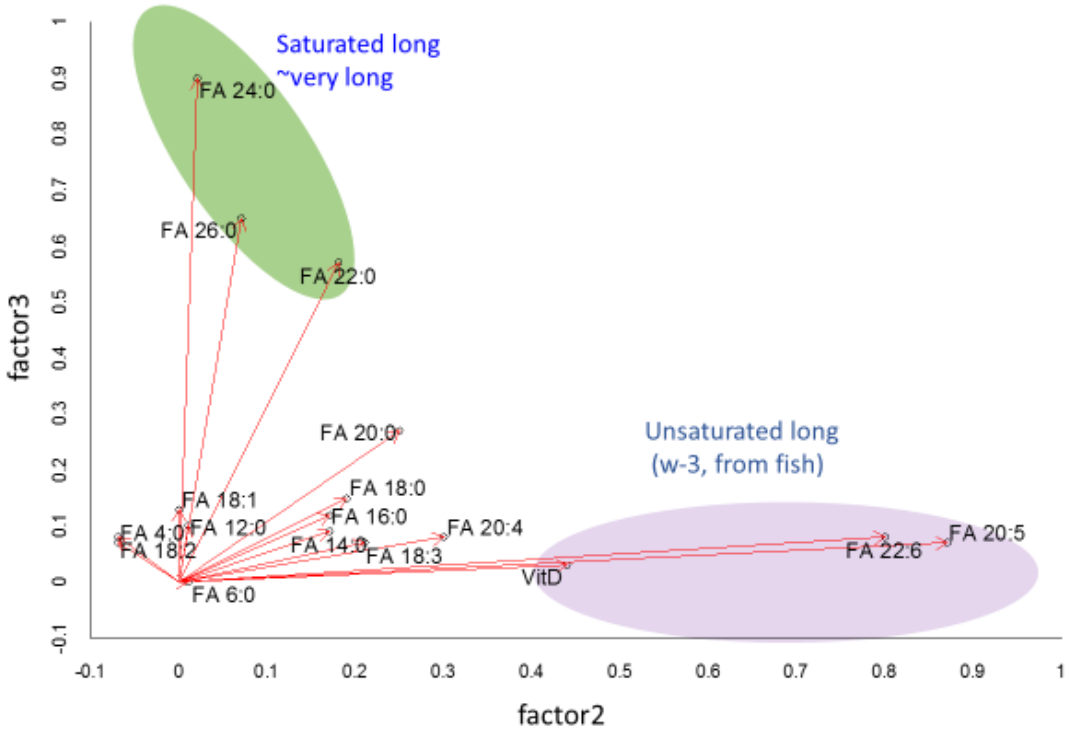

Fig.4.2 Result of total fatty acid dataset (factor analysis)

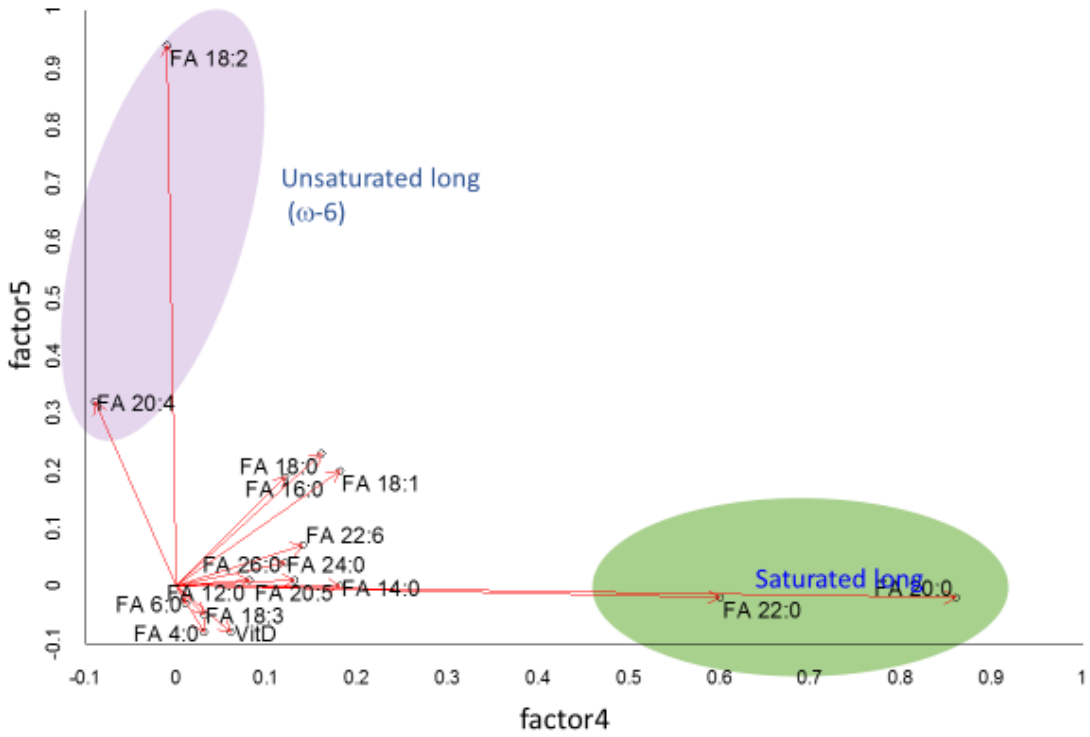

Fig.4.3 Result of total fatty acid dataset (factor analysis)

Factor analysis in free fatty acid dataset

In the application to free fatty acid dataset, the number of factors decided by parallel analysis was 2, meanwhile that decided by BIC values was 7. We tried to adopt 2 factors, but no information other than size factor was obtained. Model with 7 factors was found to be a more acceptable model (Table 4.2).

Table 4.2: Comparison between BIC value and parallel analysis

Comparison of factor number in four indices in free fatty acids dataset		
Factor number	2	7
RMSR	0.03	0
RMSEA	0.126	0.04
Tucker-Lewis index	0.872	0.987

In axis of Factor 1 (Fig.4.4), the loading values ranged in the order of free fatty acids 18:1, 18:0, 18:2, 18:3, and 16:0. This order is the same to the average concentration order, strongly indicating Factor 1 as size factor. In axis of Factor 2 (Fig.4.4), the subtypes with high loading values are free fatty acids 14:0 and 12:0, which are rich in coconut oil. Although Factor 2 might possibly be a representation of coconut oil intake, we rather consider Factor 2 as the representation of a unique but yet unspecified metabolic pathway shared by the two fatty acids with the same reason used in application to total fatty acid dataset.

In the axis of Factor 3 (Fig.4.5), free fatty acids 22:6, 20:5 and vitamin D showed high loading values. We consider Factor 3 as the representation of fish oil intake. In the axis of Factor 4 (Fig. 5), free fatty acid 4:0 showed a high loading value. We consider Factor 4 as the representation of short chain fatty acid or gut microbiota condition (Schonfeld, P. and Wojtzek, L., 2016).

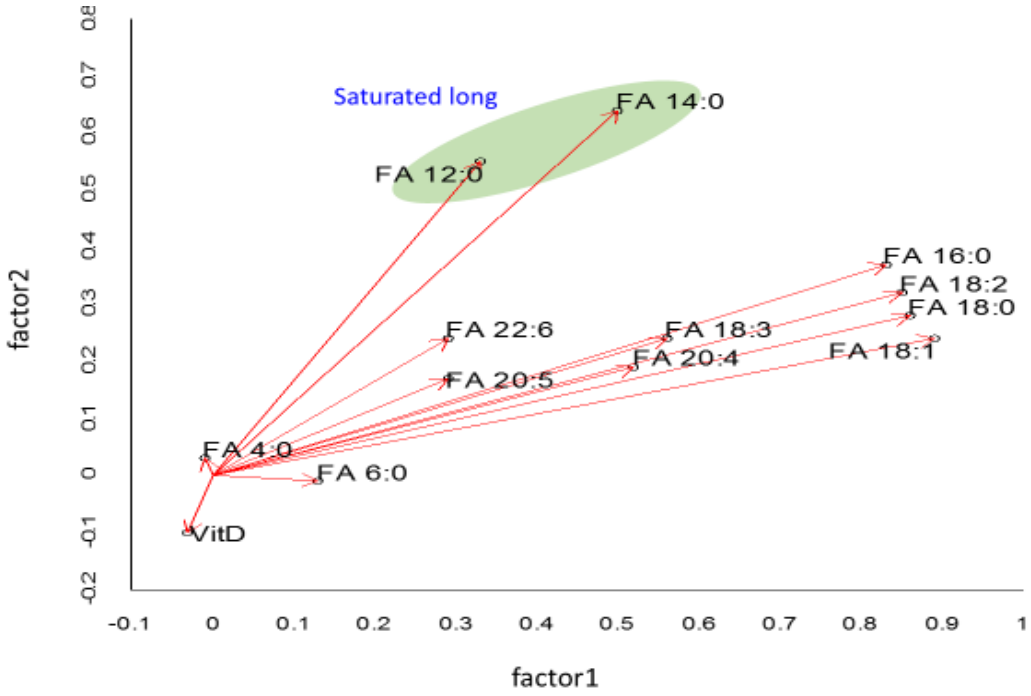

Fig.4.4 Result of free fatty acid dataset (factor analysis)

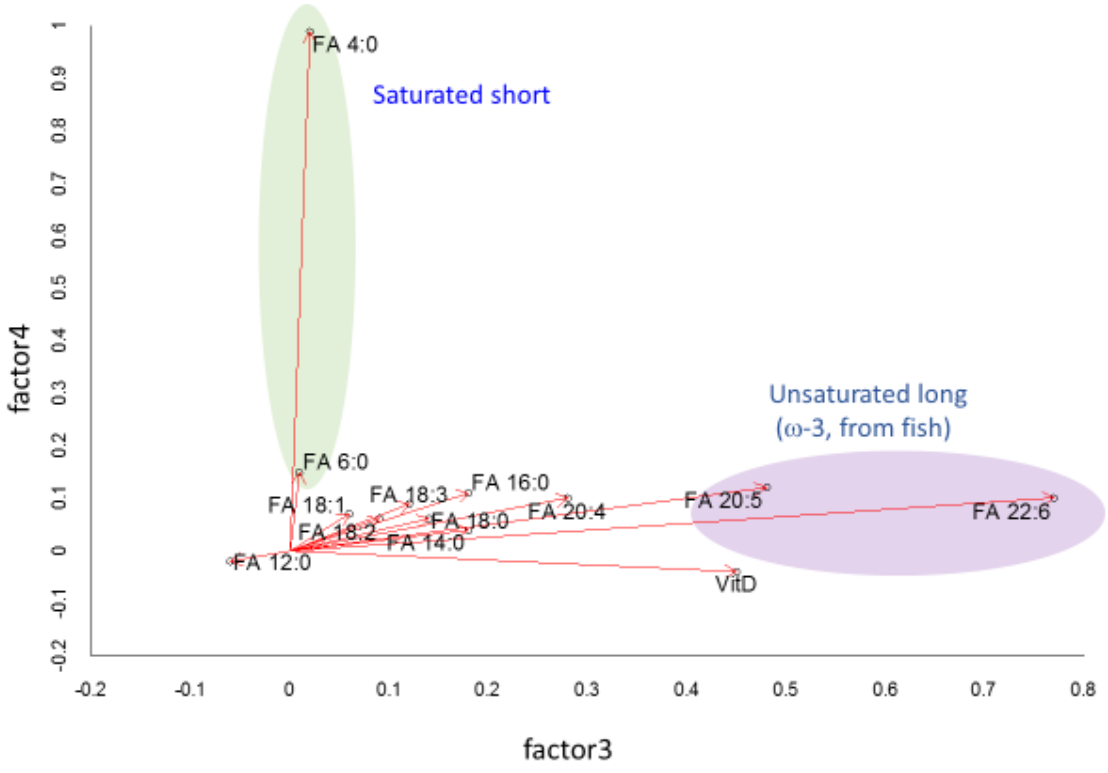

Fig.4.5 Result of free fatty acid dataset (factor analysis)

Principal component analysis

We also applied PCA to total fatty acid dataset and free fatty acid dataset. Outcomes of PCA also reflect characteristics of total fatty acids and free fatty acids.

Principal component analysis in total fatty acid dataset

We applied PCA to total fatty acid dataset firstly. In the axis of first principal component (PC1) (Fig.4.6), the subtypes with high loading values were in the order of total fatty acids 16:0, 18:1,18:0, 18:3, and 14:0, which are major components in the serum. Total fatty acid 18:2 was low because of the small standard deviation, as discussed in factor analysis application. In the axis of second principal component (PC2) (Fig. 6), the subtypes with high loading values were total fatty acid 20:0, 22:0, 24:0, and 26:0, which are saturated long (20:0) or very long fatty acids (22:0, 24:0, 26:0). Therefore, we consider PC2 as a component to represent saturated long and very long fatty acids.

In the axis of third principal component (PC3) (Fig.4.7), the subtypes with high loading values were total fatty acids 20:5, 22:6, and vitamin D. Therefore, we consider PC3 as the representation of fish oil intake. In the axis of fourth principal component (PC4) (Fig.7), the subtypes with high loading values were total fatty acids 18:2 and 20:4 with high polarity (a separation of electric charge in a molecule), and negatively 12:0, 14:0, and 8.0 with low polarity. We consider PC4 as the representation of polarity.

In the axis of fifth principal component (PC5) and axis of sixth principal component (PC6) (Fig.4.8), total fatty acid 4:0 and total fatty acid 6:0 were with high loading values in both axes, therefore PC5 and PC6 both are related to short chain fatty acids.

It is difficult to analyze principal components after PC6; therefore, we did not analyze them.

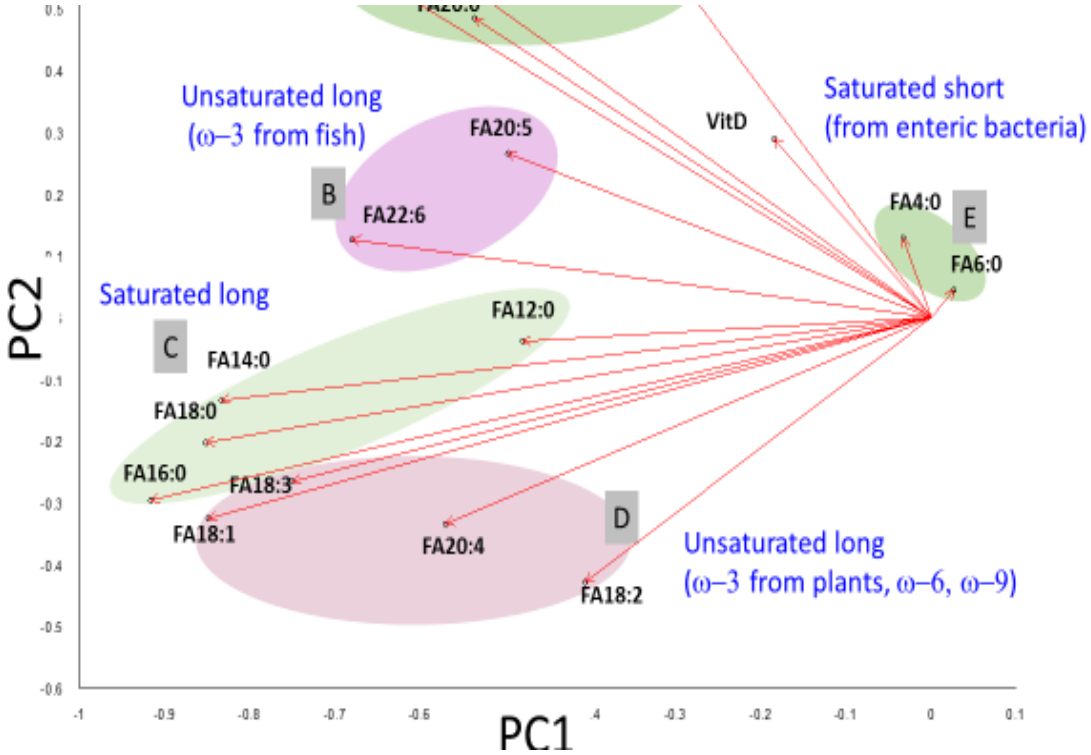

Fig.4.6 Result of total fatty acid dataset (PCA)

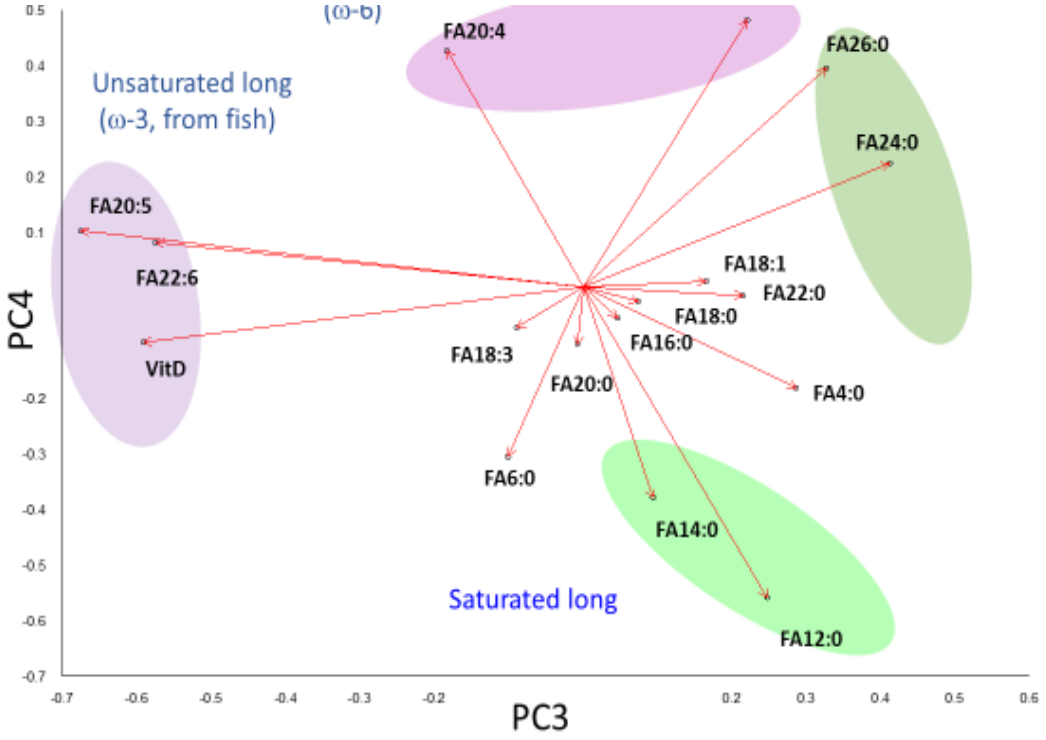

Fig.4.7 Result of total fatty acid dataset (PCA)

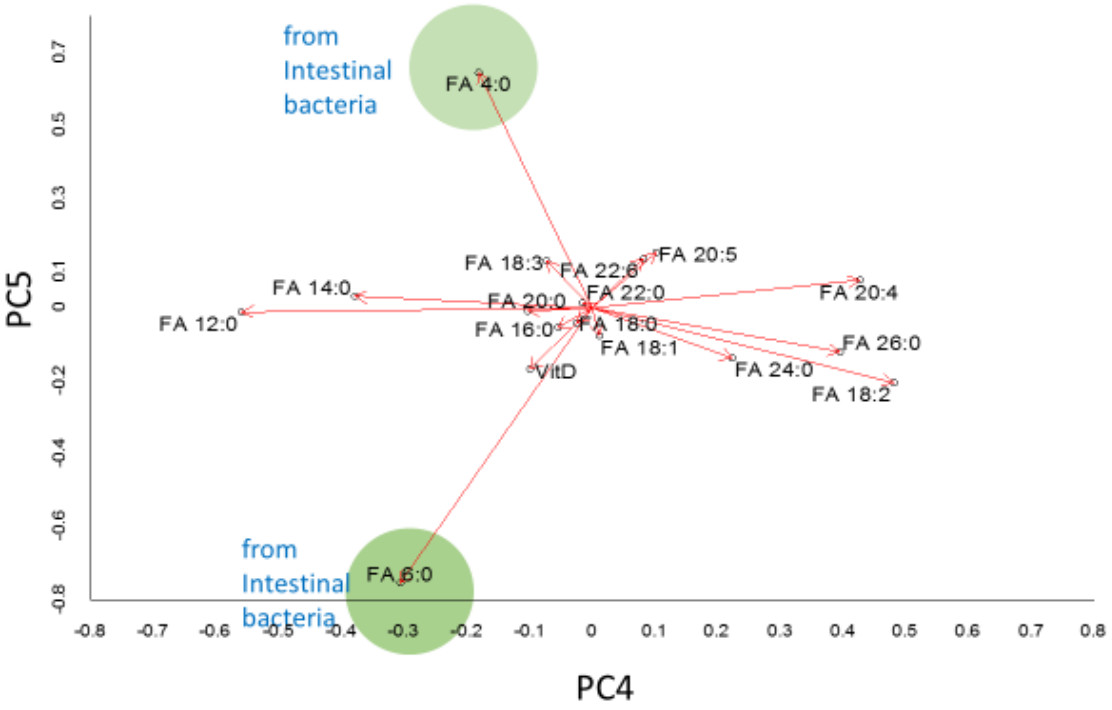

Fig.4.8 Result of total fatty acid dataset (PCA)

Principal component analysis in free fatty acid dataset

The free fatty acids with high loading values in axis of PC1 were in the order of free fatty acids 16:0, 18:1, 18:2, 18:0, 18:3, and 14:0 (Fig.4.9). We consider PC1 as the size factor. In the axis of PC2 (Fig.4.9), the subtypes with high loading values were free fatty acids 20:5, 22:6 and vitamin D, indicating PC2 as the representation of fish oil intake. In axes of PC3 and PC4 (Figs.4.10), the subtypes with high loading values were free fatty acids 4:0 and 6:0. Moreover, both subtypes showed a bidirectional distribution in PC4. Thus, we consider PC3 is the representation of short chain fatty acids, and PC4 as that of gut microbiota condition. In axis of PC5, the subtype with high loading values was free fatty acid 12:0. We suspect that PC5 represents a distinctive but yet unspecified metabolic pathway for fatty acid 12:0.

In all applications, fatty acids 20:5 (EPA) and 22:6 (DHA) showed a special relationship with vitamin D, indicating their strong relationship with fish oil intake. We think it might be the reason that vitamin D in this study is 25-hydroxyvtamin D_3, which is obtained from fish oil. Fatty acids 20:5 and fatty acid 22:6 are also mainly obtained from fish oil, therefore they are in a deep connection with each other.

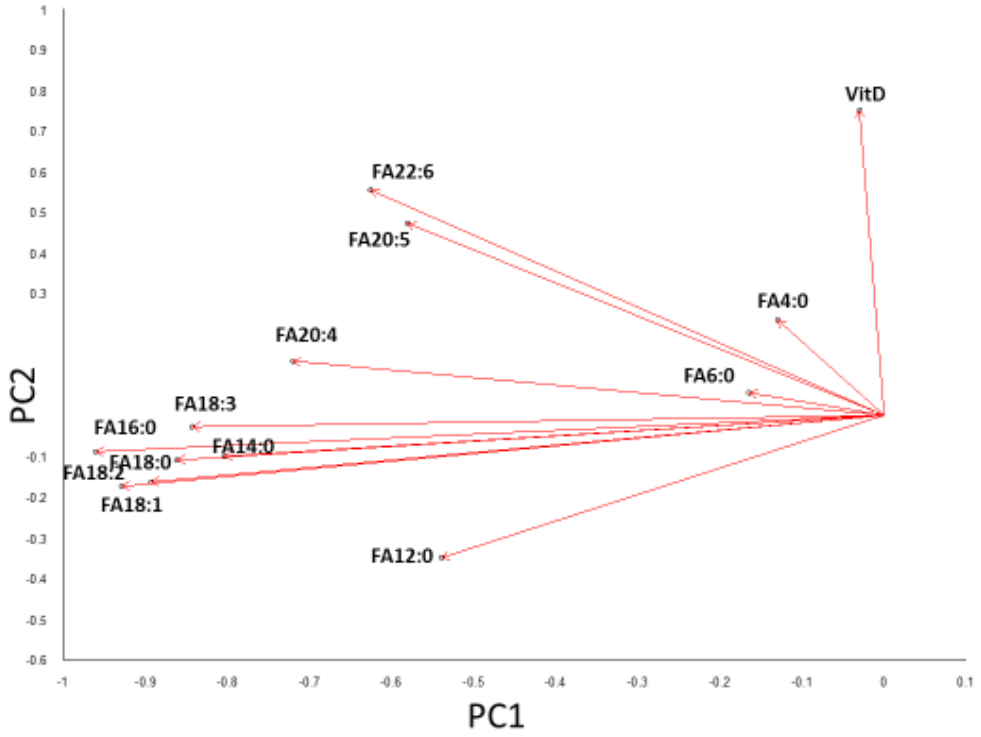

Fig.4.9 Result of free fatty acid dataset (PCA)

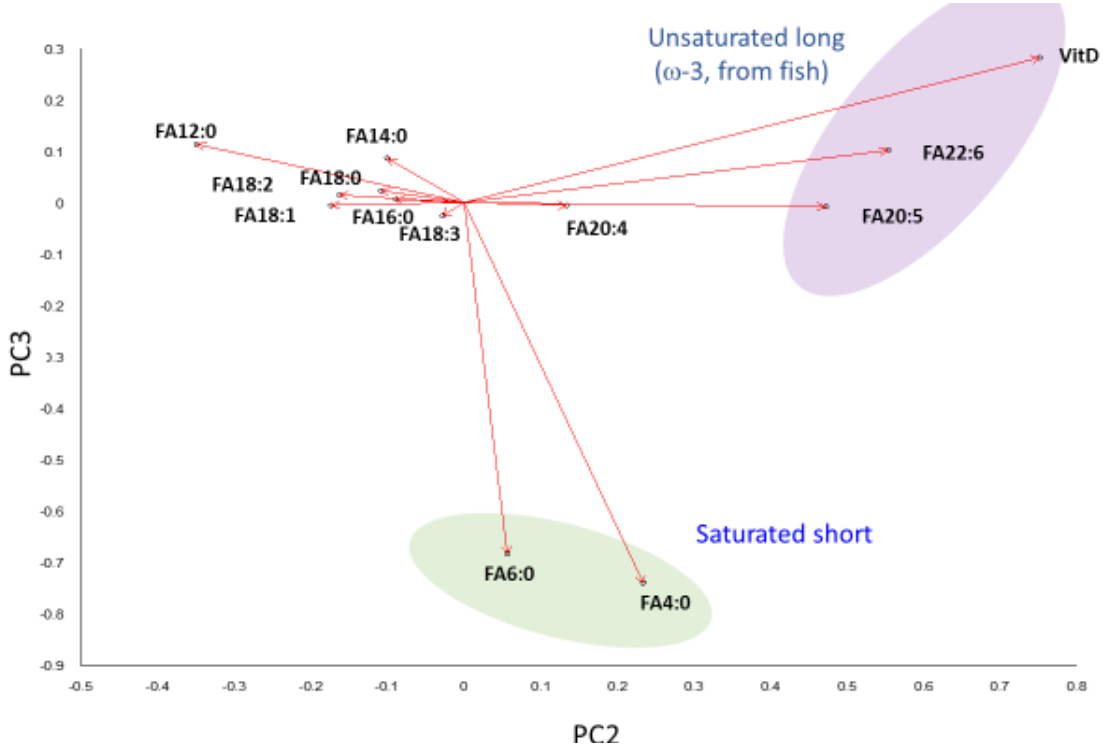

Fig.4.10 Result of free fatty acid dataset (PCA)

Independent component analysis

In the application of independent component analysis, we used three different algorithms:"ICAfast", "ICAimax", and "ICAjade". There are only three main sources of dietary fatty acids for humans: fish, animal, and plant seeds. Therefore, we set the number of components to 3, to find out the independent information source of fatty acids. Outcomes of different ICA algorithms were quite similar.

Independent component analysis in total fatty acid dataset

In the application to total fatty acid dataset, the first component was found to be size factor, because its loading values showed a similar order with that of their concentrations (Figs. 4.12). Loading value of total fatty acid 18:2 was low as seen in the factor analysis and PCA application. In the axis of the second component (Fig.4.11), we found extremely high loading values for total fatty acid 20:5 and 22:6. We consider this component related to fish intake. We cannot distinguish the species of fish with the available information from ICA application. However, it is rare to see the loading value of total fatty acid 22:6 lower than that of total fatty acid 20:5. This situation can happen in the fatty acids from Atka mackerel (Japanese name, Hokke), a most popular table fish in Hokkaido. It might be possible that the second axis reflects Atka mackerel intake more strongly than other fish species.

Independent component analysis in free fatty acid dataset

In the application to free fatty acid dataset, the first component can be considered as size factor, because order of loading values in size of the first component (Fig. 4.12) is nearly same as that of concentration in average. In the axis of the second component (Fig.4.12), subtypes with high loading values are free fatty acid 18:1 and free fatty acid 16:0. Loading value of free fatty acid 18:2 is the thirdly high. According to the proportion, we think this axis represent the animal fat intake.

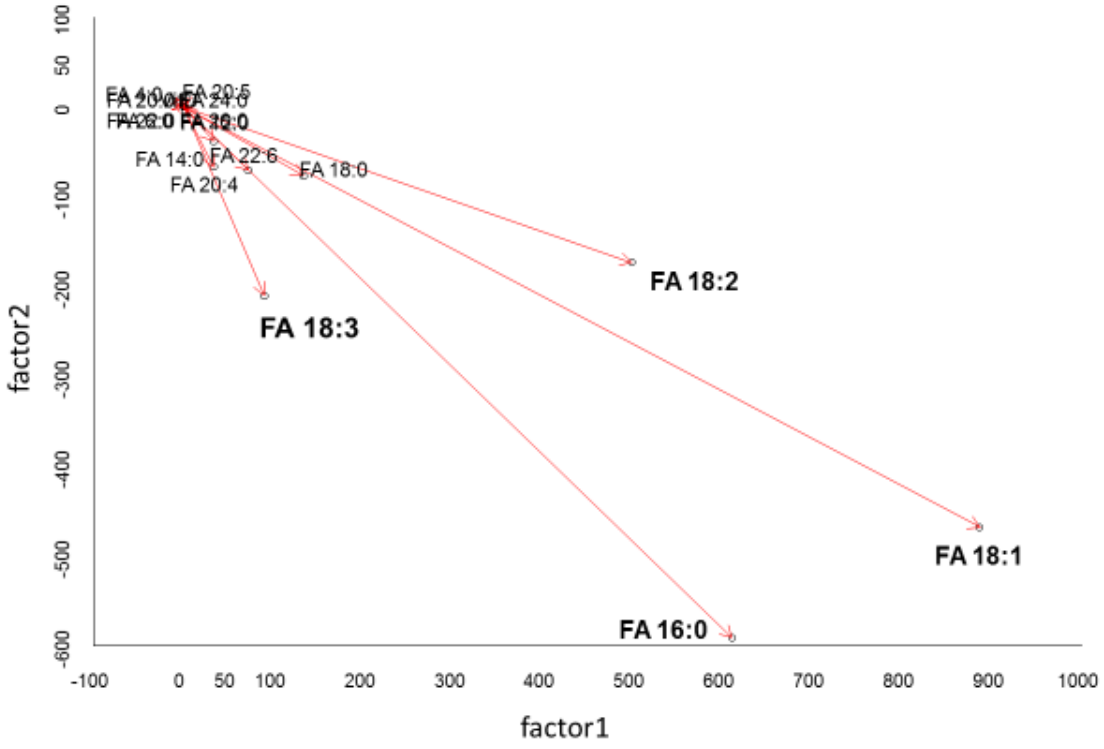

Fig.4.11 Result of total fatty acid dataset (ICA)

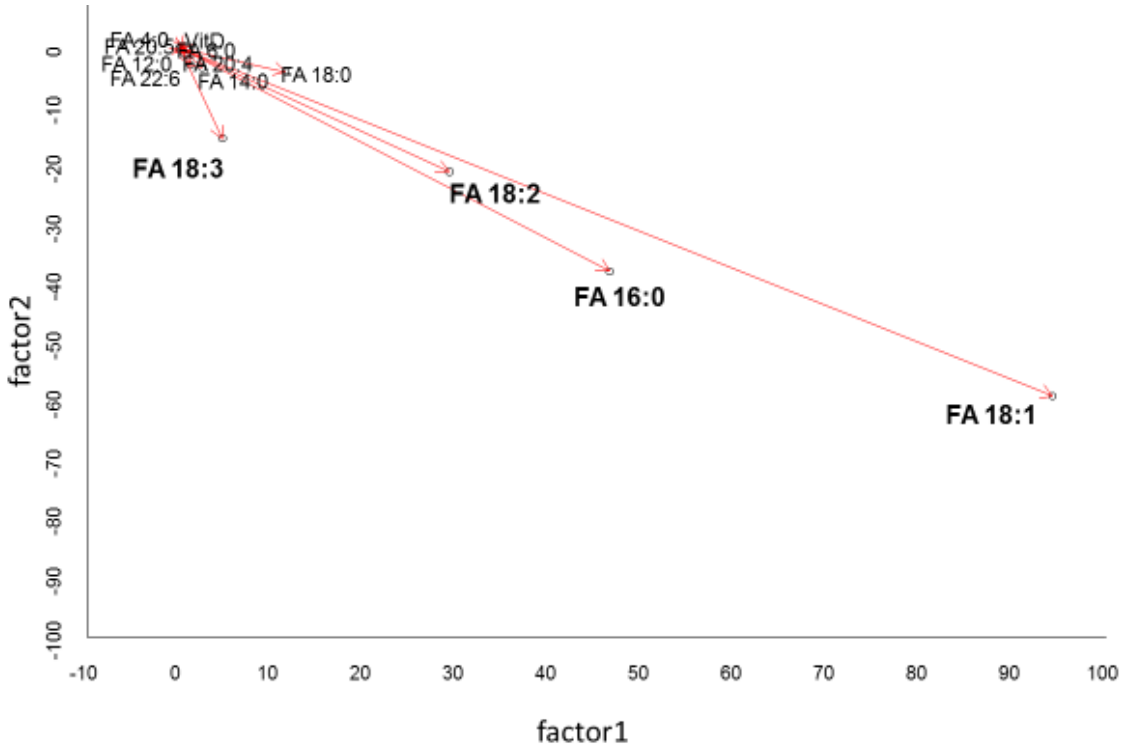

Fig.4.12 Result of tree fatty acid dataset (ICA)

4.2.3 Conclusion

In this study, we used three different dimension reduction methods to obtain different results. In the result of factor analysis in total fatty acid dataset, we found seven latent factors, which are size factor, representation of omega-3 fatty acids intake (or fish intake), representation of fat solubility under two conditions, representation of omega-6 fatty acids intake, representation of an unspecified metabolic pathway, and representation of short chain fatty acids (or the contribution of gut microbiota). In the result of factor analysis in free fatty acid dataset, we found the latent factors can be explained as size factor, representation of a unique but yet unspecified metabolic pathway, representation of fish oil intake. Besides, BIC criterion was found to be more practical than parallel analysis in the present study.

In the result of PCA in total fatty acid dataset, we found that the principal components can be explained as size factor, representation of fat solubility, representation of omega-3 fatty acids intake (or fish intake), representation of short chain fatty acids (or the contribution of gut microbiota). In the result of PCA in free fatty acid dataset, we found that principal components can be explained as size factor, representation of fish oil intake, representation of short chain fatty acids (or the contribution of gut microbiota). In the applications of factor analysis and PCA, fatty acids 20:5 (EPA) and 22:6 (DHA) showed a special connection with vitamin D, indicating their strong relationship with fish oil intake. Furthermore, in the application of factor analysis and PCA, we also found that there might be a difference between fatty acids 4:0 and 6:0, which might lead to a further research in the future.

ICA suggested non-fish obtained oil intake in free fatty acid dataset, although in total fatty acid dataset outcome was uncertain. The number of components in independent component analysis should be chosen with a more practical way in future work.

PCA and factor analysis can be used to explain the unobserved factors. PCA provided a general but rough view in the present study. In factor analysis, we obtained more sophisticated explanation based on the selected number of factors in modeling. For example, we found the latent factor representing of omega-6 fatty acids intake by

factor analysis.

Unlike the results of factor analysis and PCA, the results of ICA suggested the food sources of free fatty acids by the proportion in size. In ICA, we found the results of free fatty acid dataset, considered as the representation of food sources, are more explainable than the results of total fatty acid dataset. The reason of such situation is that concentration of free fatty acids is highly influenced by food. We chose to keep the results of ICA to make the conclusion more informative.

In conclusion, fatty acids were confirmed to be affected by factors: their acyl chain length, number of double bond, omega group number, digestion, metabolism, food sources, and gut microbiota. Furthermore, we found the strong relationship between vitamin D and fatty acids 20:5 and fatty acid 22:6, indicating the effect of fish intake with the use of factor analysis and PCA. Therefore, application of dimension reduction is considered as a tool to discover the characteristics of fatty acids and contribute to health care and disease prevention.

4.3 Analysis on serum cholesteryl ester dataset

4.3.1 Purpose

Cholesteryl ester (CE) is the ester of cholesterol and fatty acid (FA). Because of the molecular diversity of FAs, there are various kinds of CE species in human serum, where CE serves as the major core lipid of low-density lipoproteins (LDL) and high-density lipoproteins (HDL). In clinical laboratories, the concentration of each CE species is not available since CE is usually obtained as total esterified cholesterol by subtracting free cholesterol from total cholesterol both enzymatically determined. For this reason, only a limited number of literatures have reported the concentrations of CE species in a large population, and moreover, few of them contained multivariate analysis.

Warensjö, P., et al. (2005) studied the average proportions, not absolute quantities, of FA species in serum CE using gas chromatography coupled with thin-layered chromatography in the men at ages of 50 and 70 years with and without the metabolic syndrome. The authors conducted principal component analysis (PCA), and found significant relationship between the onset of metabolic syndrome and the factors of linoleic acid (FA 18:2) and omega-3 polyunsaturated FAs. Including their studies, the concentration of CE species has been commonly determined by time-consuming and laborious gas chromatography for methylated or silylated FAs, following to chromatographic separation of CE fraction from other lipid classes.

In our present study, we determined absolute concentrations of each CE species in the serum for a general Japanese population using liquid chromatography/tandem mass spectrometry (LC/MS/MS) with the use of 2H3-labeled internal standards (IS) for each CE species. To analyze the complex CE dataset, we employed an informatics approach, that is, multivariate analysis including dimension reduction methods. The aim of this report is presenting and discussing the results of our multivariate analysis for the serum CE dataset. Informatics approach might be useful to reach better understanding of the metabolisms and biological roles for each CE species (Chen, Y., et al., 2020).

4.2.2 Results of analysis

PCA of CEs

PCA was applied on the CE dataset to obtain a general view on CEs. The biplot of the first (PC1) and the second (PC2) principal components displays that larger loading values in PC1 were obtained for the group of CEs 16:0, 18:1, 20:4, 18:2, and 18:3. The smaller loading values were obtained for another group of CEs 22:6, 20:5, and 18:0 (Fig.4.13). The separation of two groups is consistent with the difference in the CE concentrations for the total, except for CE 18:3. Thus, we considered PC1 as the size factor. In PC2, CEs were separated into three groups (Fig.4.13). The first group showed positively larger loading values and contained CEs 18:2, 18:3, and 18:1. The second group showed absolutely smaller loading values and contained CEs 18:0, 16:0 and 20:4. The third group showed negatively larger loading values and contained CEs 22:6 and 20:5. Thus, we thought it was also likely to consider PC2 as the reflection of source factor: the first, second, and third groups corresponded to the plants, meats, and fish, respectively (Fig. 4.13).

The biplot of the third (PC3) and the fourth (PC4) principal components displays that positively large loading values were obtained for the group of CEs 18:0 and 16:0 in PC3 (Fig.4.14). The negatively small loading values were obtained for CEs 22:6, 18:1, 18:2, 18:3 and 20:4. Whereas, the negatively large loading values are obtained for CE 20:5. CE 18:0 and 16:0 are saturated CEs, the rest CEs are unsaturated CEs. Saturated CEs have higher fat solubility than unsaturated CEs. Therefore, PC3 was considered as representation of saturation or fat solubility. In PC4, negatively large loading value was obtained for CE 18:3, and positively large loading value was obtained for CE 20:4 and 18:2. CE 18:3 belongs to omega-3 group, which decreases inflammation. Meanwhile, CE 20:4 and 18:2 belong to omega-6 group, which increases inflammation. It is, however, unlikely that PC4 represents inflammation effect. CEs 22:6 and 20:5 showed only small loading values in PC4 in spite that they are among n-3 group and reduce inflammation more effectively than CE18:3. Thus, PC4 remains to be an obscure axis in this study, suggesting PC4 as a reflection of multiple factors. In conclusion, PCA was

found useful, but its utility was not enough for giving a precise explanation for PC4 due to complication of multiple principal components. We considered that the main solution to avoid an obscure axis was to give a specific factor number in modeling. Therefore, we applied factor analysis with a selected factor number for further understanding on the CE dataset.

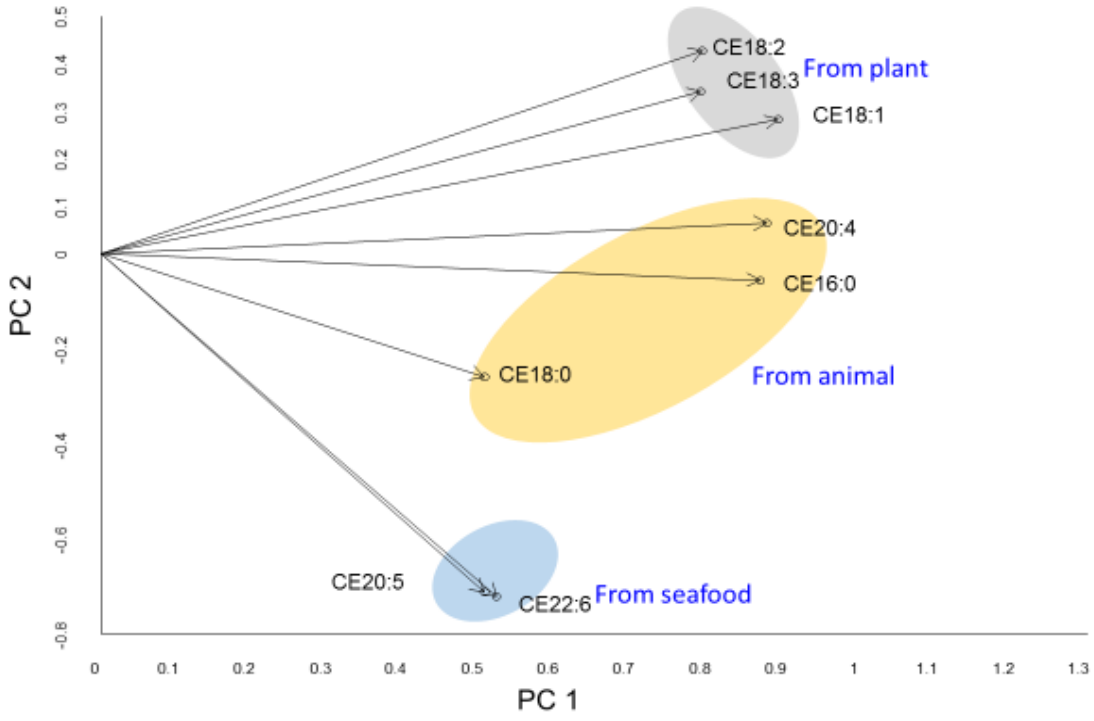

Fig.4.13 Result of CE dataset (PCA)

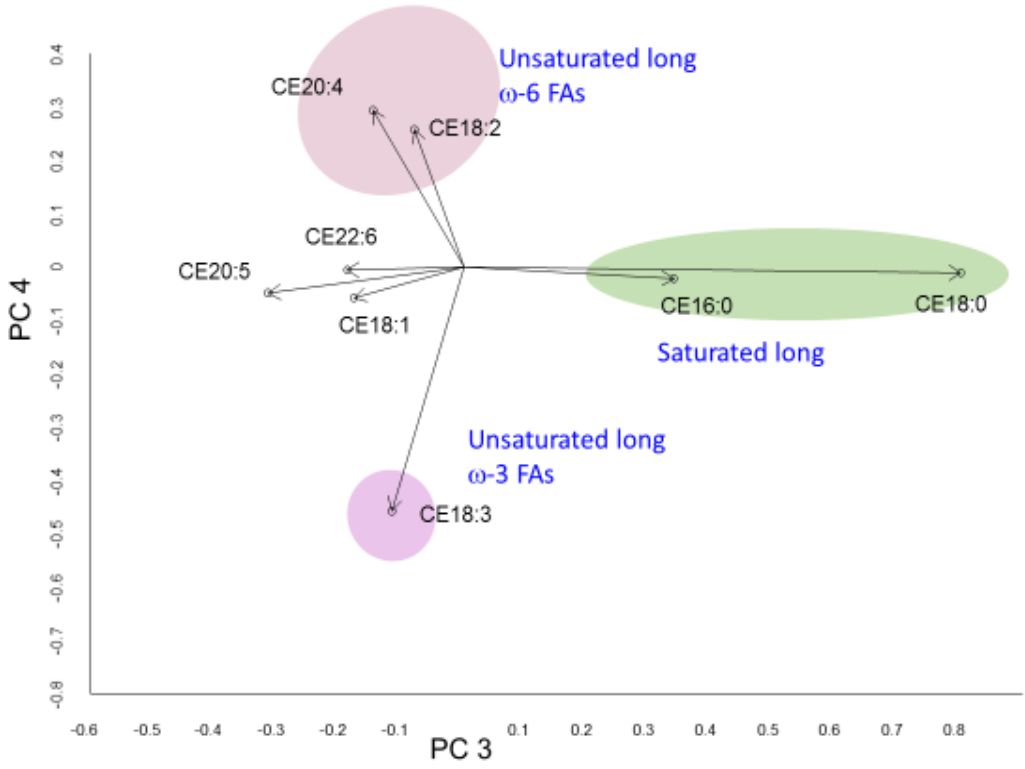

Fig.4.14 Result of CE dataset (PCA)

Factor analysis of CEs

Decision of factor number

In factor analysis, factor number is determined by a combination of two criterions: RMSEA index, and lowest BIC value. In application on the CE dataset: the factor number decided by RMSEA index, and lowest BIC value were 4 (Table 4.3), and we certainly can explain the model better than models with other number. Therefore, we chose the model with 4 factors.

Table 4.3: Comparison between different factor number

Comparison in indexes of different factor numbers for free fatty acids				
Factor number	2	3	4	5
RMSEA	0.246	0.126	0.092	NA
BIC value	355.6	22.71	-1.47	NA

Results of factor analysis

In the result of factor analysis, all loading values were positive, and therefore, the factor pattern was simple and easy to interpret. The subtypes with the loading values no less than 0.60 were generally considered to be significant. The biplot of factors 1 and 2 displays that the CE subtypes with the loading values ≥ 0.6 were CEs 18:1, 18:3, 18:2, 20:4 and 16:0 in factor 1 (Fig.4.15). Factor 1 was thought to be the size factor with the same reason mentioned in PC1 analysis. In factor 2, CEs 22:6 and 20:5 showed the large loading values ≥0.6. We considered factor 2 as a reflection of the fish intake because CEs 22:6 and 20:5 are usually obtained from fish oil. On the other hand, we also found that subtypes were also divided into three groups. CEs 20:4, 16:0, 18:0, and 18:1 were in group 1, which are mainly obtained from animal oil. CEs 18:2 and 18:3 were in group 2, which are mainly obtained from plants oil. CEs 22:6 and 20:5 were in

group 3, which are mainly obtained from fish oil. Therefore, factor 2 was also likely to reflect food source of FAs.

In the biplot of factor 3 and 4, the subtypes with the large loading values ≥ 0.5 in factor 3 were CEs 16:0 and 18:0, both of which were with saturated FAs (Fig.4.16). On the other hand, the rest of CEs showing the small loading values were those with unsaturated FAs. Therefore, we considered factor 3 to reflect fat solubility with the same reason in PC3 analysis. In factor 4, only CE 20:4 showed a large loading value (Fig.4.16). FA 20:4 (arachidonic acid) is metabolized to eicosanoids (prostacyclin, thromboxanes, leukotrienes, and prostaglandins) that are signaling molecules and play an important role in mediating inflammatory responses, exerting a wide spectrum of biologic actions in various body systems (Ricciotti, E. and FitzGerald, G. A., 2011). In contrast, CEs 20:5 and 22:6 showed the smaller loading values than CE 20:4. FAs 20:5 (EPA) and 22:6 (DHA) are metabolized to proresolving mediators, respectively, resolvin and protectin (Serhan, C. N., 2014). Therefore, we consider that factor 4 might reflect inflammatory aspect of CEs.

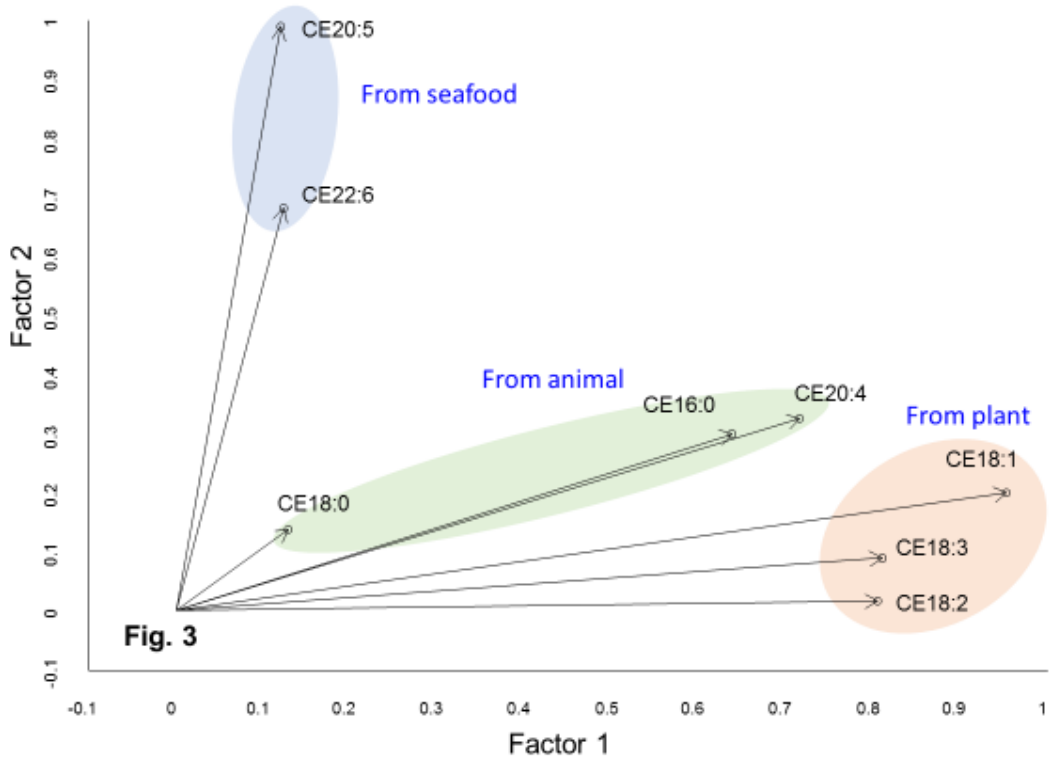

Fig.4.15 Result of CE dataset (factor analysis)

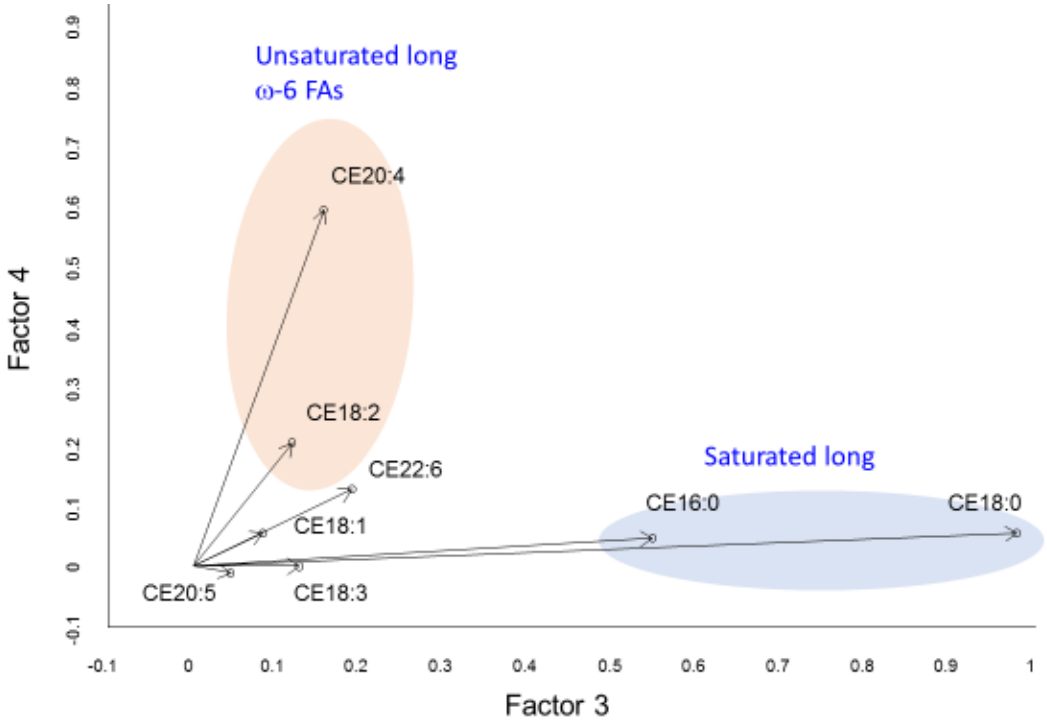

Fig.4.16 Result of CE dataset (factor analysis)

Summary and general consideration

First, the order of the CE species in PC1 and factor 1 were parallel to the order of the concentrations of CE species, suggesting PC1 and factor 1 serve as the size factors. The PC1, PC2, and PC3 in PCA, and also the factor 1, factor 2, and factor 3 in factor analysis well reflected the size, the food source (or fish intake), and the fat solubility, respectively (Figs. 4.13-16). In comparison between PC4 and factor 4, however, only factor 4 reflected the inflammation and signaling, whereas PC4 failed to do that. It means that factor analysis performed better than PCA in this study.

In factor analysis of the CE dataset, the uniqueness value of CE 22:6 was remarkably high among those of CEs. In contrast, the uniqueness of free FA 22:6 and total FA 22:6 were not so high. Furthermore, another n-3 FA-containing CE, namely, CE 20:5 was the lowest in the uniqueness value among all CEs. Uniqueness contains two parts: specific variance and error variance. Specific variance is specific to a particular item. Error variance is from errors of measurement. In the present case, the accuracy and precision of the measurement of CE 22:6 was confirmed in the validation studies. Therefore, the observed high uniqueness of CE 22:6 should largely reflect the specific variance. These findings might indicate the possibility of unique metabolic pathway for CE 22:6. FA 22:6 is enriched in the brain, representing > 40% of total brain PUFA. Since the synthesis of FA 22:6 from the n-3 precursor, α-linolenic acid, is low in the brain, a continuous supply of FA 22:6 across the blood-brain barrier is required. In plasma, FA 22:6 can be in several fractions, such as phospholipids, lysophospholipids, CEs, TGs, and free FAs. It is of our interest that lyosophosphatidylcholine 22:6 is transported from plasma to the brain and the eye by a specific transporter, namely Mfsd2a. Also, it is reported that FA binding protein 5 (FABP5) mediates the transport of free FA 22:6 across the blood-brain barrier. The high uniqueness value for CE22:6 might suggest a possibility of a unique transport system for CE 22:6 from plasma to the brain, in spite that CE 22:6 is a minor part (< 1% by weight) of total CE in plasma.

One limitation of this study is that the present informatics approach provides only suggestions, but not conclusions. Based on the suggestions, however, we can plan

future experiments to test the suggestions. For example, the suggested uniqueness of CE 22:6 might raise a possibility of a unique metabolic pathway, which might lead us to identify possible receptors or transporters specific to CE 22:6 (Nguyen, Long. N. et al., 2014; Wong, B. H. et al., 2016). We conclude that informatics approach, especially factor analysis, could be a useful tool to extract important information from chaotic lipid datasets.

4.3.4 Conclusion

Use of two basic dimension reduction methods (PCA and factor analysis), had advantages in the analysis of CE molecules in the epidemiological study. PCA and factor analysis seemed to suggest the quantity, the food source, fat solubility, and the biological function of CE molecules. Factor analysis following to a factor selection process using BIC and RMSEA was found more practical than PCA in this study. Informatics approach will be useful for analysis of a large and complex lipid dataset generated by mass spectrometry. This study uncovered the uniqueness of CE 22:6, which remains to be elucidated in the future.

4.4 Analysis on common component analysis

4.4.1 Purpose

In this study, we analyzed the cholesterol ester dataset, free fatty acid dataset and total fatty acid dataset in one time to discover the common characteristics in the three datasets. In the combined dataset, there are 24 variables, which are CE16:0, CE18:0, CE18:1, CE18:2, CE18:3, CE20:4, CE20:5, CE22:6, FFA16:0, FFA18:0, FFA18:1, FFA18:2, FFA18:3, FFA20:4, FFA20:5, FFA22:6, TFA16:0, TFA18:0, TFA18:1, TFA18:2, TFA18:3, TFA20:4, TFA20:5, and TFA22:6. In order to explain the result clearly, we named the common variables as FA16:0, FA18:0, FA18:1, FA18:2, FA18:3, FA20:4, FA20:5, and FA22:6 in the result of the common principal component analysis.

4.4.2 Result of analysis

In the result of combined dataset (Fig.4.17), we found there are four explainable components of this combined dataset. In the first axis, components with high loading values are FA18:1, FA16:0, FA18:2, which are generally high in concentration in serum, therefore the first component is considered as the size factor. In the second axis (Fig.4.18), the component with relatively high loading values are FA18:2, FA16:0 and FA18:1. The combination and proportion of loading values seems to be similar to the proportion in bean oil and corn oil. Therefore, we considered the second axis indicating fatty acids extracted from plants, especially bean oil and corn oil, which are generally used in japan.

In the third axis (Fig.4.18), the component with high loading values are FA16:0, FA18:1, which are generally from animal fat, thus, the third axis is considered to represent fatty acids from animal fat. In the fourth axis (Fig.4.18), the component with high loading values are FA20:5, FA20:4, and FA22:6. FA20:5 and FA22:6 are generally from fish oil, and FA20:4 are partly from fish oil, therefore we considered the forth component representing the fish oil intake.

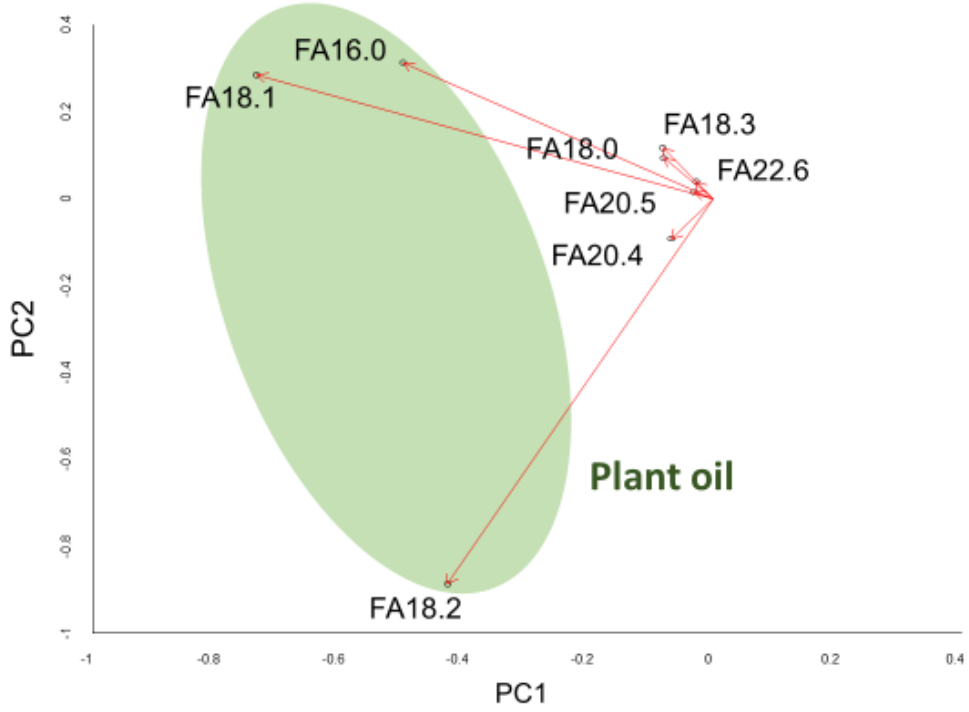

Fig.4.17 Result of fatty acid dataset (CPCA)

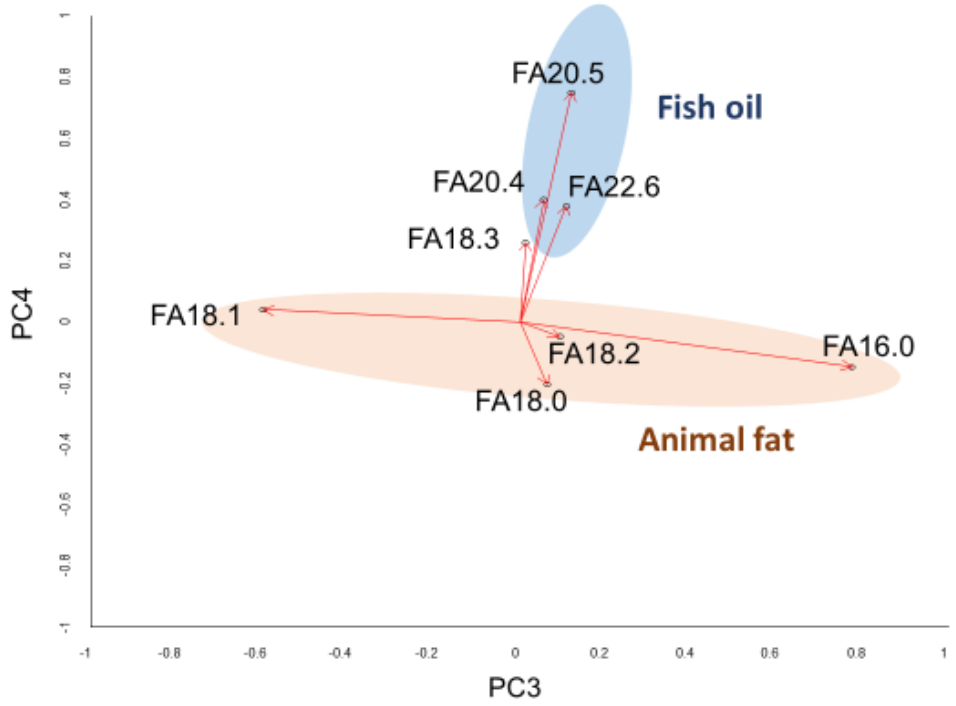

Fig.4.18 Result of fatty acid dataset (CPCA)

4.4.3 Conclusion

Comparing with other outcomes, the representation of each component indicating the food sources is more practical and can be wildly used in the clinical health test as food intake background information. In the clinical test, common principal component analysis can be more time saving than the handwriting questionnaire. Therefore, common principal component analysis is practical in the field of clinic.

Compared with other three dimension reduction methods (PCA, factor analysis, and independent component analysis), common principal component analysis has its advantages as well.

Independent component analysis showed the food sources of fatty acids, but the result of independent component analysis is less accurate than the results of common principal component analysis. In the results of independent component analysis, we can only understand the outcome of animal fat intake source in free fatty acid dataset and a suspectable guess of fish oil intake source in total fatty acids separately. But in the results of common principal component analysis, we can understand the three food sources, plant oil, animal fat and fish oil, at once.

Components in the result of principal component analysis are explainable, but the results are rough. Components in the result of factor analysis are explainable and accurate, but the results are not as useful as the results of common principal component analysis. Therefore, the outcome from common principal component analysis seems to be more competitive than the other three analysis methods in the field of utility and accuracy.

5. Analysis on milk fatty acid dataset

Concentration variation happens everywhere, such as medicine aspect and dietary aspect. It has a great meaning in our daily life, therefore revelation of the affecting factors of concentration variation is important and necessary (Chen. Y., et al. 2017).The purpose of this research is to solve real problems like "what are affecting factors to concentration variation".

In this research, we propose a method to analyze relationship between concentration and affecting factors. Meanwhile we also use functional data analysis to explain the relationship between the changing trends of concentration and those of affecting factors. Finally, to get a better representation, we show some examples of execution of the proposed method.

5.1 Milk fatty acid dataset

The milk fatty acid dataset contained only total fatty acid 8:0 concentrations. Five sets of milk samples obtained at each of 18 locations ($n = 18$) in Hokkaido area over May, July, September, and November in 2014 and January in 2015 were provided from commercial companies. We also obtained three additional datasets as variables ($p = 3$) for the corresponding locations and periods, that is, every day speed of wind, every day sum of rainfall, and every day average temperature. The weather data was obtained at the homepage of the Japan Meteorological Agency (https://www.jma.go.jp/jma/indexe.html).

Milk total fatty acid 8:0 concentrations were determined by high-performance liquid chromatography (HPLC), as previously reported (Shrestha, R. et al., 2018). Briefly, every milk sample was saponified with potassium hydroxide and derivatized with 2-nitrophenylhydrazine (NPH), extracted in diethyl ether, and then analyzed in HPLC. Standard fatty acids were commercially obtained. Two internal standards, that is, 1,2,3-triundecanoylglycerol and 1,2,3-trinonadecanoylglycerol, were chemically synthesized as previously reported (Hui, S.P., et al., 2018). Total fatty acid 8:0 concentrations were determined on the basis of the peak height ratios of the analyte and the internal standards.

Accuracy, precision, and recovery of this assay were studied 6 times and confirmed to range within acceptable ranges.

5.2 Method of analysis

We may see in the traditional data analysis as well, including functional clustering and functional multidimensional scaling (MDS).

Before the introduction of our method, some notations should be claimed.

We assume there are time dependent multivariate observations below:

$$y_{ij}, x_{ij}^{(1)}, x_{ij}^{(2)}, \dots, x_{ij}^{(p)} \quad i = 1, \dots, n;\ j = 1, \dots, m \qquad (1)$$

where p is the number of variables, n is the number of observations, m is the number of period, and $x_{ij}^{(r)}$ refers to the r-th variable for the i-th object and j-th period.

In this research, we assume y_{ij} as an object of concentration, and $x_{ij}^{(1)}, x_{ij}^{(2)}, \dots, x_{ij}^{(p)}$ as the affecting factors at the i-th object and j-th period.

We describe the proposed method as follows:

STEP 1: Functionalization

STEP 2: Functional hierarchical clustering

STEP 3: Measure of similarity between dendrograms

STEP 4: Visualization of dissimilarity

Functionalization step is a step of smoothing. In this step, basis functions are chosen to make data into functions, which we can calculate by any argument t, with basis function system.

It is a linear combination of basis functions $\Phi_k = (\phi_1(t), \dots, \phi_k(t))^{\mathrm{T}}$ and coefficients that are independent from each other. There are several options, such as Legendre polynomials and Fourier basis system.

Legendre polynomials systems is a linear combination of Legendre polynomials functions, which are the solutions to Legendre's differential equation with order n:

$$P_{\mathrm{n}}(t) = \frac{1}{2^n n!}\frac{d^n}{dt^n}(t^2 - 1)^n, \qquad (2)$$

therefore, the basis function can be written as

$$\Phi(t) = \big(P_0(t), P_1(t), \dots, P_k(t)\big)^{\mathrm{T}}.$$

Fourier basis system is a linear combination as well, but this time it is with the Fourier series, which is:

$$c_0 + c_1 \sin(wt) + c_2 \cos(wt) + c_3 \sin(2wt) + c_4 \cos(2wt) + \cdots, \quad (3)$$

and its corresponding basis function can be written as

$$\Phi(t) = \left(\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\sin(wt), \frac{1}{\sqrt{\pi}}\cos(wt), \frac{1}{\sqrt{\pi}}\sin(2wt), \frac{1}{\sqrt{\pi}}\cos(2wt), \cdots\right)^{\mathrm{T}}$$

Since basis functions are independent, basis function system can be an orthonormal system. Thus, we assume basis functions are orthonormal,

$$\int \phi_k(t)^{\mathrm{T}} \phi_s(t) = \begin{cases} 1 & if\ s = k \\ 0 & if\ s \neq k \end{cases}$$

We can choose the kind of basis function system according to the characteristic of dataset. In this research, the corresponding coefficient $c_{ki}^{(r)}$ and c_{ki} are calculated by minimizing the least square criterion.

$$\mathrm{SMSSE}(c_{ki}^{(r)} | x_{ij}^{(r)}, \ldots, x_{im}^{(r)}) = \sum_{j=1}^{m} \left(x_{ij}^{(r)} - \sum_{k=1}^{K} c_{ki}^{(r)} \phi_k(t_j)\right)^2, \quad (5)$$

$$\mathrm{SMSSE}(c_i | y_{ij}, \ldots, y_{im}) = \sum_{j=1}^{m} \left(y_{ij} - \sum_{k=1}^{K} c_{ki} \phi_k(t_j)\right)^2, \quad (6)$$

where $x_{ij}^{(r)}$ and y_{ij} are the observed values, $\sum_{k=1}^{K} c_{ki}^{(r)} \phi_k(t_j)$and $\sum_{k=1}^{K} c_{ki} \phi_k(t_j)$ are the corresponding functions.

Then, for each i and r, $\{x_{ij}^{(r)}, j = 1, \ldots, m\}$, can be approximated well arbitrarily with a linear combination of K numbers.

The formula for the functional data with basis function system can be:

$$x_{ij}^{(r)} = \sum_{k=1}^{K} c_{ki}^{(r)} \phi_k(t_j) = \mathbf{c}_i^{(r)\mathrm{T}} \Phi(t)$$

$$y_{ij} = \sum_{k=1}^{K} c_{ki} \phi_k(t_j) = \mathbf{c}_i^{\mathrm{T}} \Phi(t)$$

where $\mathbf{c}_i^{(r)} = \left(c_{1i}^{(r)}, c_{2i}^{(r)}, \ldots, c_{Ki}^{(r)}\right)^{\mathrm{T}}$ and $\boldsymbol{c}_i = (c_{1i}, c_{2i}, \ldots, c_{Ki})^{\mathrm{T}}$ correspondingly.

With the functionalization, data have the characteristics of function, which means that we can take derivatives and find the integration of those dataset. More precisely, we can explain changing trends with the derivative of the functions.

To find the effects of changing trends of data, we focus on the derivatives of basis functions. So we calculated their derivatives of functions as

$\frac{\mathrm{d}}{\mathrm{dt}} y_i(t), \frac{d}{dt} x_i(t)^{(1)}, \ldots, \frac{d}{dt} x_i(t)^{(p)}$from $y_i(t), x_i(t)^{(1)}, \ldots, x_i(t)^{(p)}$ correspondingly.

Functional hierarchical clustering step is to get an interpretation of the relationship between observed objects. Functional clustering is not only used in the functions, but also the derivatives of the functions, because both original functions and their derivative functions may be useful for analysis of concentration changes.

Firstly, dissimilarity should be calculated. Dissimilarity between two functions $f(t)$, $g(t)$ is defined by

$$|f - g|_1 = \int |f(t) - g(t)| dt$$

or

$$|f - g|_2 = \int \big(f(t) - g(t)\big)^2 dt.$$

Then, we carried out functional hierarchical clustering to functions $\{y_i(t); i = 1, \ldots, n\}, \left\{\frac{\mathrm{d}}{\mathrm{dt}} y_i(t); i = 1, \ldots, n\right\}, \left\{x_i^{(r)}(t); i = 1, \ldots, n\right\}, \left\{\frac{\mathrm{d}}{\mathrm{dt}} x_i^{(r)}(t); i = 1, \ldots, n\right\}$. We can get $2p$+2 dendrograms.

In measure of similarity between dendrograms step, a criterion should be used to evaluate similarities. Cophenetic correlation coefficient is used in this research. It is the height between two nodes in the dendrograms and usually considered as a measure to judge how well dendrograms can fit the original distances.

More precisely speaking, cophenetic distance between two observations that have been clustered represents the intergroup dissimilarity at which the two observations are first combined into a single cluster. That means cophenetic distance is the height of point where two elements first intersect as a union. Cophenetic correlation coefficient varies from 0 to 1, when two kinds of distance fit in 100%, it becomes 1, when two kinds of distance don't fit each other at all, it moves to 0 (Saraçli, S., Doğan, N. and Doğan, İ., 2013).

In this research, we use it as a measure of similarity of two classifications or two dendrograms. We denote i_a and i_b as the i-th elements of the $\mathrm{dendrogram_a}$ and $\mathrm{dendrogram_b}$ respectively. Also $\mathrm{D_a}(i,j)$, $\mathrm{D_b}(i,j)$ are denoted as the cophenetic distance between objects i and j in $\mathrm{dendrogram_a}$ and i and j in $\mathrm{dendrogram_b}$

respectively, and D_a, D_b as the average cophenetic distance in $\text{dendrogram}_\text{a}$ and $\text{dendrogram}_\text{b}$ respectively.

Cophenetic correlation coefficient is defined as below:

$$\mathrm{C} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(D_a(i,j) - D_a)(D_b(i,j) - D_b)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(D_a(i,j) - D_a)}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(D_b(i,j) - D_b)}}.$$

If the outcome is close to 1, the correlation between two dendrogram is strong. If it is moving to 0, the correlation between two dendrograms goes weak.

In the final step, Multidimensional scaling is applied as a visualization method measuring the level of similarity in individual cases of a dataset with a dissimilarity matrix. Multidimensional scaling is a way to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissimilarities (distances) between the investigated objects.

The visualization in 2-dimension is a plot showing the dissimilarity between points corresponding to those input objects, the closer those points are in the plot, the much similar the input objects are, and vice versa. Such a visualized plot is created by reproduced dissimilarity d_{ij} in Y-axis, and the originally input dissimilarity in X-axis which makes the stress function in the lowest value,

$$\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(f(x_{ij}) - d_{ij}(t)\right)^2},$$

Where d_{ij} represents the reproduced dissimilarity, $f(x_{ij})$ represents the transformation of the observed input data x_{ij} (In metric scaling, $(x_{ij}) = x_{ij}$, the element of i-th row and the j-th column of dissimilarity matrix).

Multidimensional Scaling (MDS) is used, from a proximity matrix (similarity or dissimilarity) between a series of n objects, to produce the coordinates of these same objects in a low dimensional space (2-dimension in this research) so as to reproduce the observed distances. As a result, we can illustrate the distances in terms of

underlying dimensions.

5.3 Result of analysis

The proposed method is applied to an actual data set below:

$$\left\{\left(y_{ij}, x_{ij}^{(1)}, x_{ij}^{(2)}, x_{ij}^{(3)}\right); i = 1, \ldots, n, j = 1, \ldots, m\right\},$$

where y_{ij} is the concentration value, $x_{ij}^{(1)}$is the sum of rainfall every day, $x_{ij}^{(2)}$ is the average wind speed every day, and $x_{ij}^{(3)}$ is the average temperature every day for the i-th place and the j-th period. The number of data from different places is eighteen (n=18). All milk fatty acid data are observed every two months and continued in one year (m=6).

In this analysis, the concentration varies from places to places in different time. The main characteristics of corresponding places are also obtained as the observed dataset in sum of rain, average speed of wind and average temperature in one month. Therefore, the goal is to find out whether the factors are effective to the changing by using the proposed method above.

In functionalization of this experiment, Fourier basis system resulting from Fourier series is used, for the dataset is periodic and Fourier series is suitable.

Fourier basis function system below is chosen with five basis functions, here is the outcome of functionalization on concentration:

$$x_i^{(r)}(t) = a_{i0}^{(r)} + a_{i1}^{(r)} \sin(wt) + a_{i2}^{(r)} \cos(wt) + a_{i3}^{(r)} \sin(2wt) + a_{i4}^{(r)} \cos(2wt),$$

and the derivatives as follow:

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i^{(r)}(t) = a_{i1}^{(r)} \cos(wt) - a_{i2}^{(r)} \sin(wt) + 2a_{i3}^{(r)} \cos(2wt) - 2a_{i4}^{(r)} \sin(2wt)$$

Three more variables are obtained as $\frac{\mathrm{d}}{\mathrm{d}t} x_i^{(1)}(t)$, $\frac{\mathrm{d}}{\mathrm{d}t} x_i^{(2)}(t)$, $\frac{\mathrm{d}}{\mathrm{d}t} x_i^{(3)}(t)$, which are representations of changing trends of sum of rainfall every day, changing trends of average wind speed every day, changing trends of average temperature every day correspondingly.

In this case, eight dendrograms can be obtained by L_2-norm. We can see four dendrograms of the outcomes below Figs. 5.1,3-5.

There are four dendrograms of clustering outcomes from the differential functions below Figs.5.2, 6-8. In the figures, the number denote eighteen different locations, which means same number corresponds to same place.

Then cophenetic correlation coefficient is applied to measure the similarity between dendrograms of concentration and the variables. In this case, "Kendall" method is used instead of "Pearson" method, because Coefficient of Rank Correlation is considered to be a more suitable measure of rank than others in hierarchical clustering.

We summarize the similarities in Table, where "concentration_d", "rain_d", "wind_d", "temperature_d" represent derivatives of concentration, derivatives of sum of rainfall every day, derivatives of average wind speed every day, derivatives of average temperature every day correspondingly.

We can see from the table that average speed of wind and average temperature have effect on milk concentration, while the derivatives of sum of rain and the derivatives of average speed of wind in one month play important roles on derivatives of concentration. With the functionalization, factors affecting the changing trends of concentration is shown.

In the final step, multidimensional scaling is used to visualize the dissimilarity between variables.

Table 5.1: Cophenetic similarity distance between variables

Cophenetic similarity						
	Wind everyday	Temp everyday	Rain everyday	Wind_d everyday	Temp_d everyday	Rain_d everyday
Concentration	0.55	0.46	0.30	0.22	0.23	0.34
Concentration_d	0.52	0.44	0.27	0.23	0.29	0.63

Multidimensional scaling (Torgerson method) is used to visualize the dissimilarity between dendrograms. We can see from Fig.5.9 that it reflects directly the outcome of cophenetic correlation coefficients. It also shows that the dendrogram of the speed of wind has the closest relationship with the dendrogram of concentration, while the dendrograms of derivatives of sum of rain and the derivatives of average speed of wind in one month are in close relationship with the dendrogram of derivatives of concentration. With the functionalization, factors affecting the changing trends of

concentration is shown.

5.4 Conclusion

Functional data analysis, especially functional clustering, gives us more choices to analyze dataset. Cophenetic correlation coefficient can be used as a measure of degree of fitness of a classification to a set of data to reveal relationship between two dendrograms. Multidimensional scaling is a good visualization to show the dissimilarity between variables.

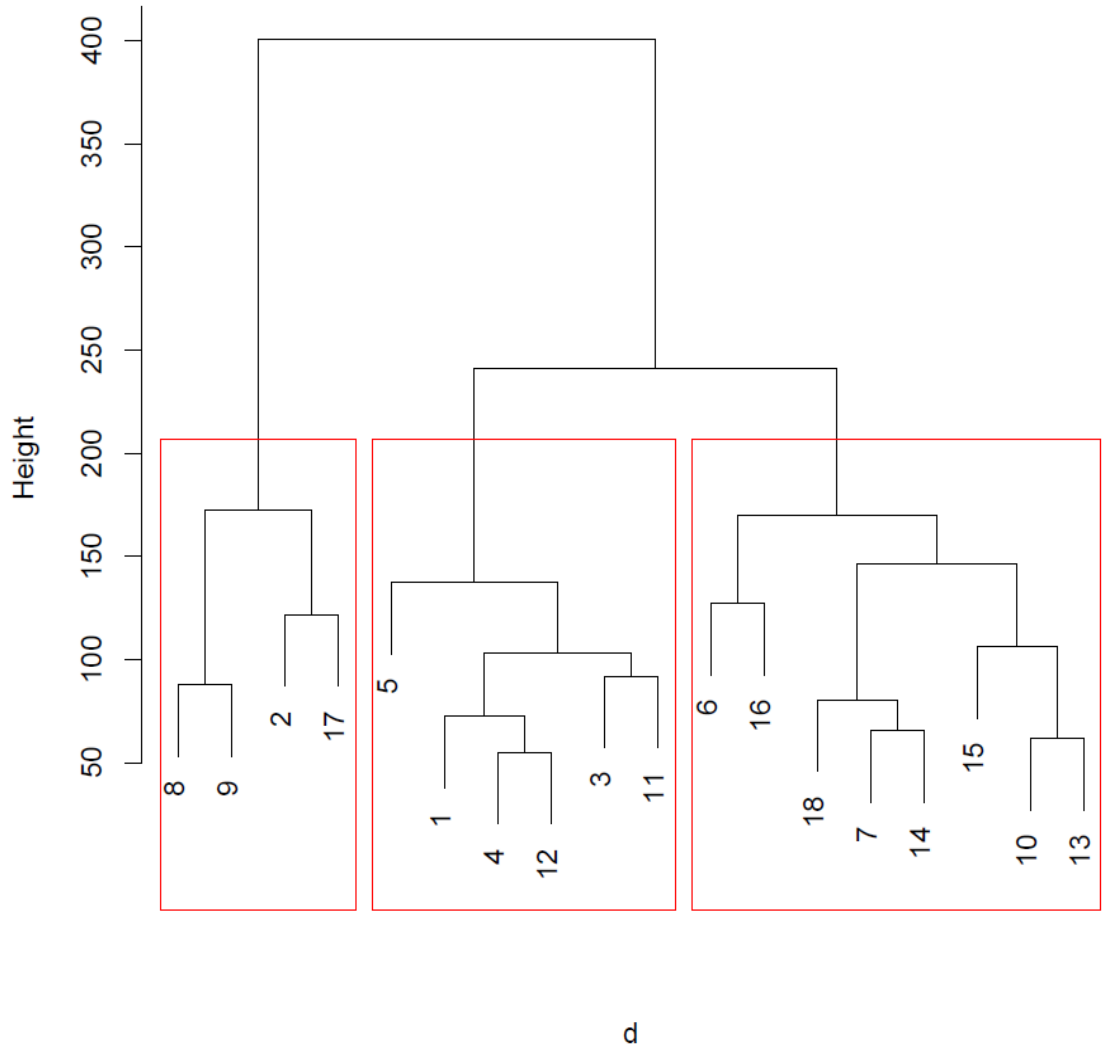

Fig.5.1 Clustering for milk concentration FA 8:0

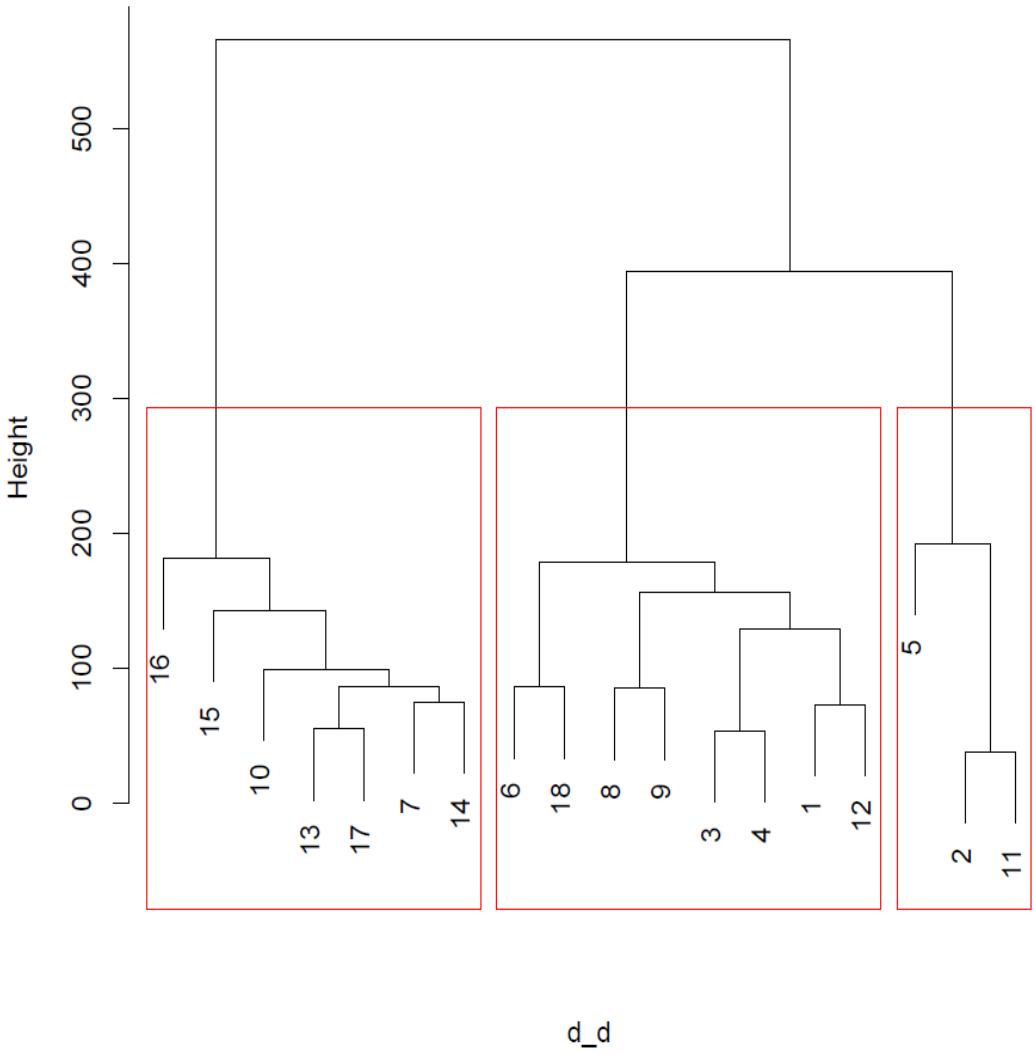

Fig.5.2 Clustering for milk concentration FA 8:0 derivatives

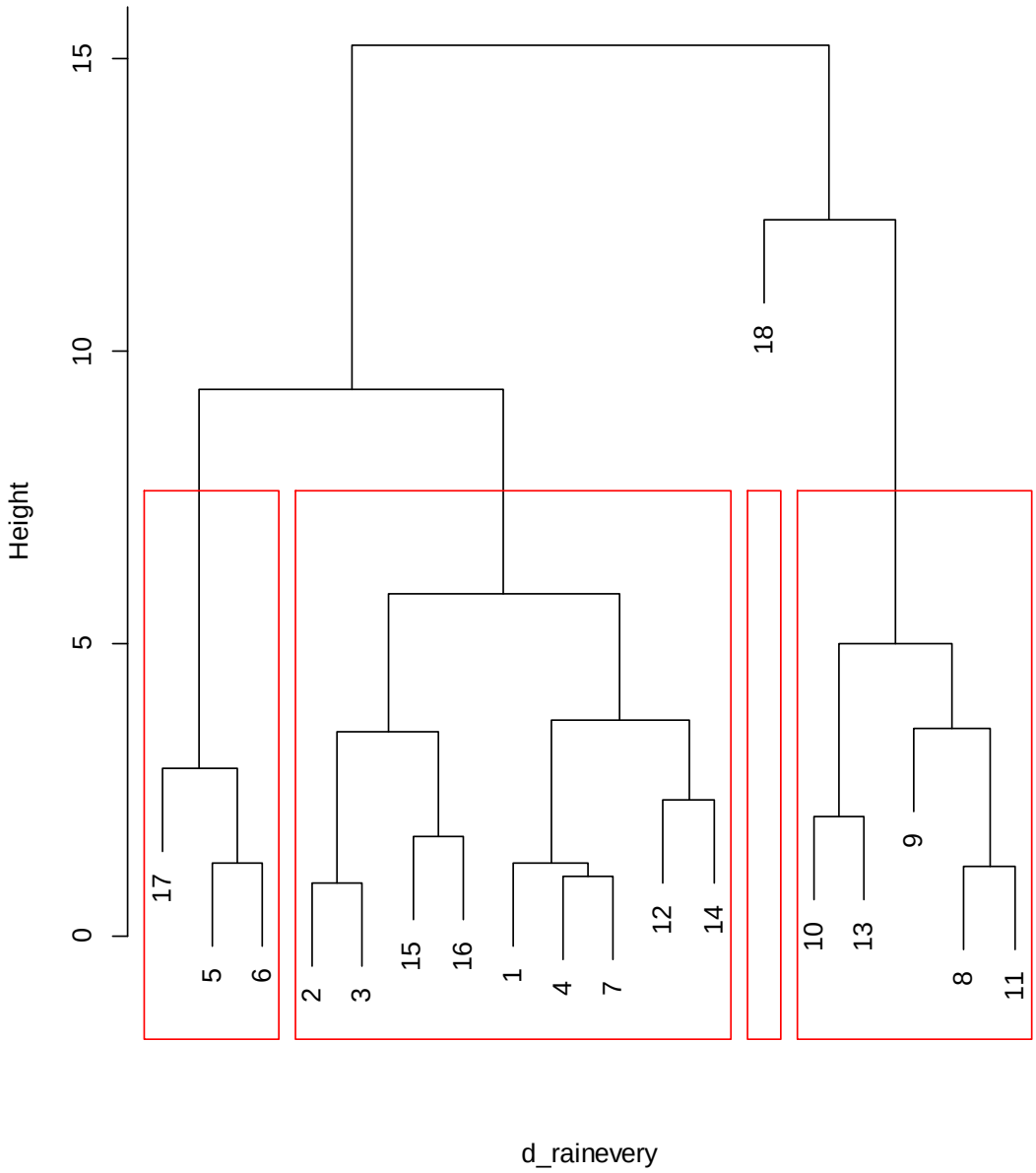

Fig.5.3 Clustering for sum of rain everyday

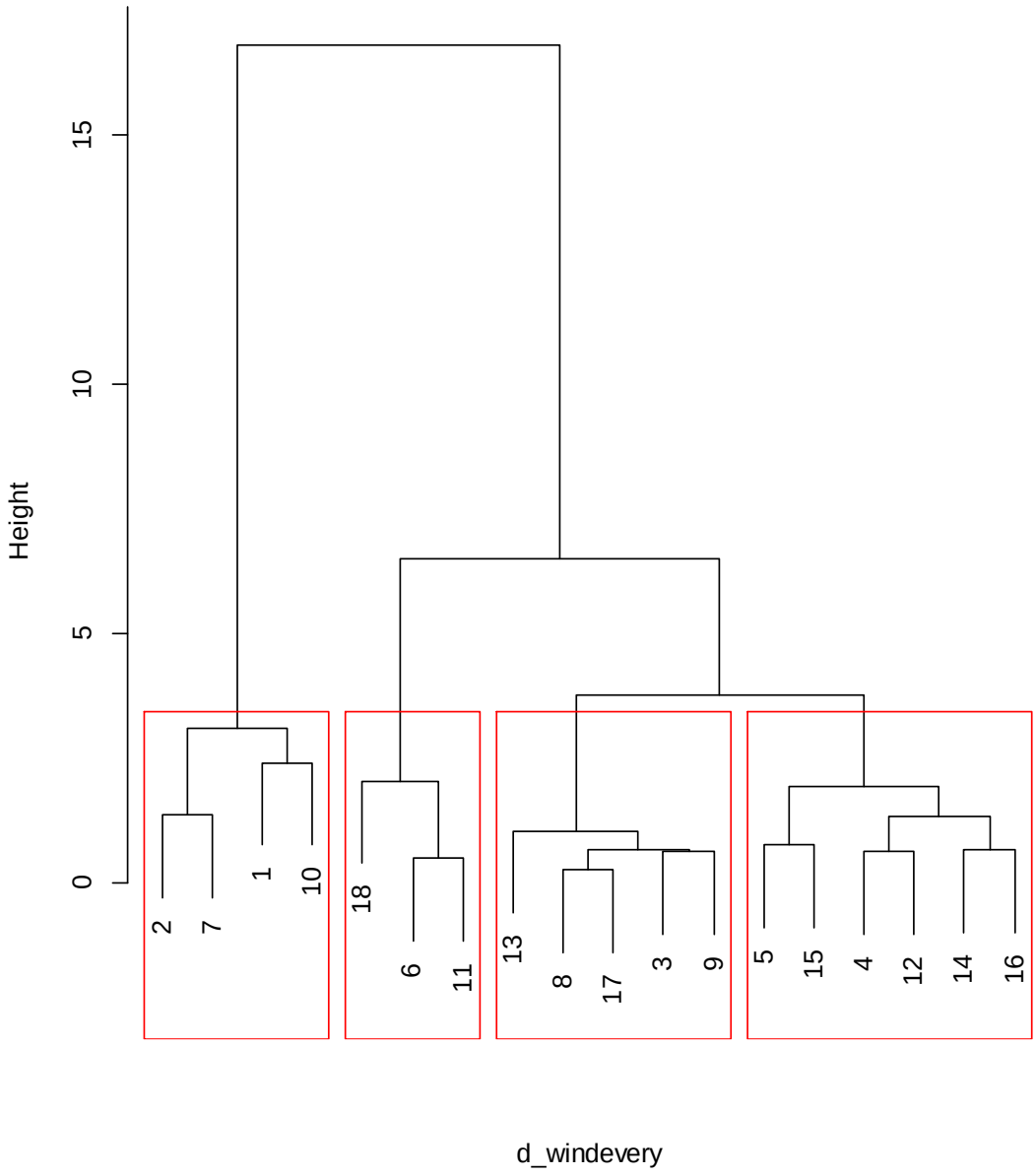

Fig.5.4 Clustering for wind speed everyday

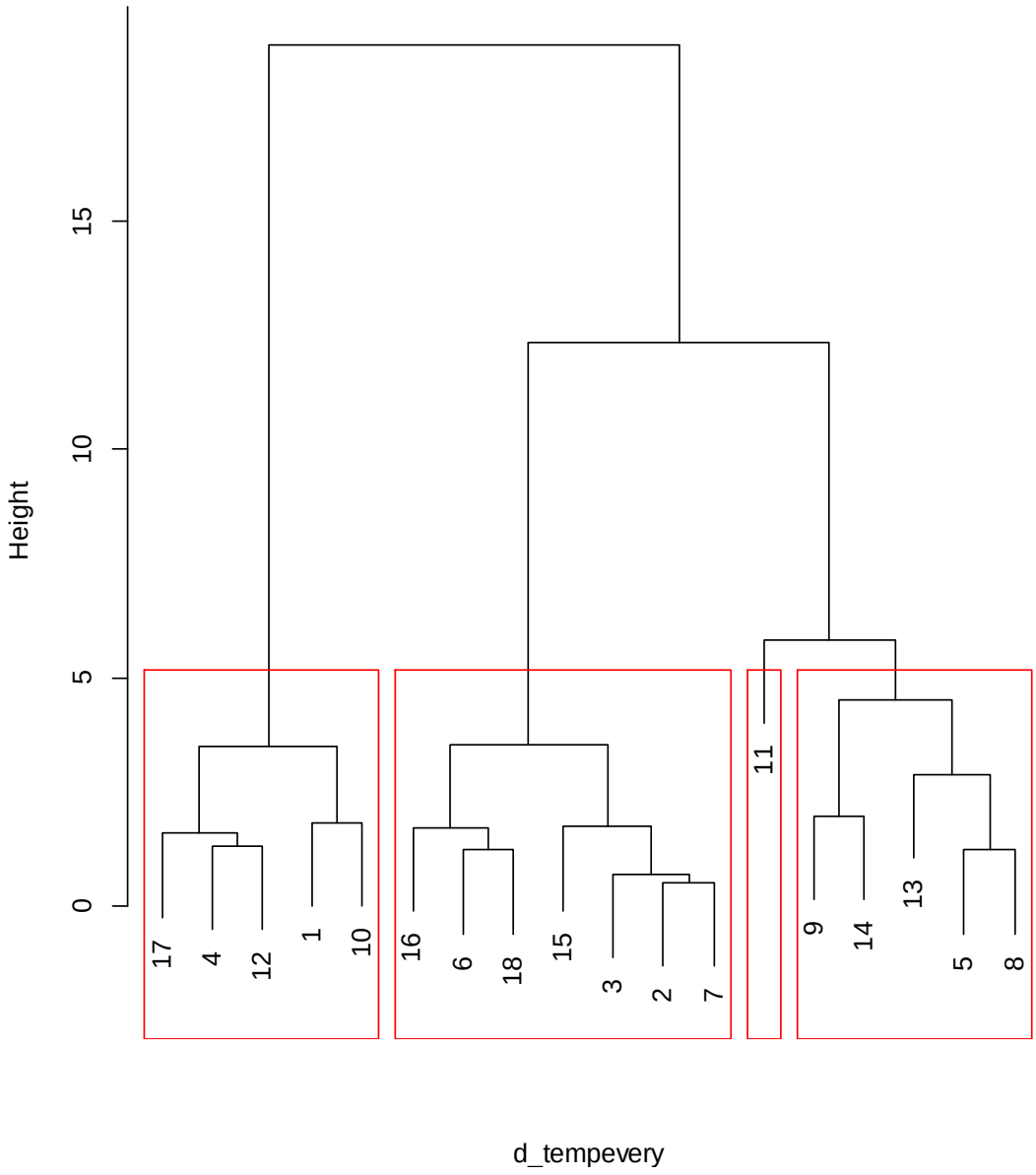

Fig.5.5 Clustering for average temperature everyday

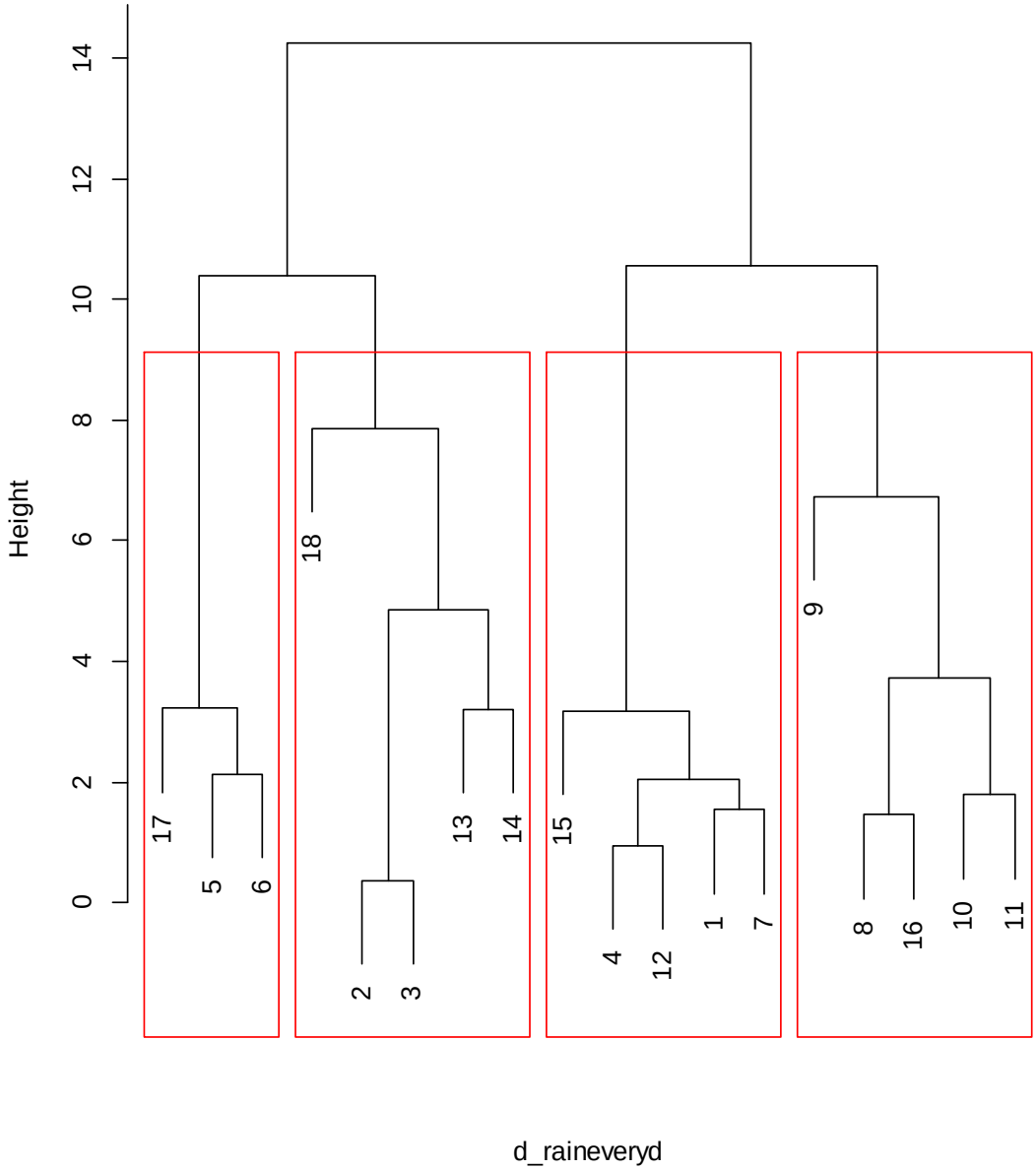

Fig.5.6 Clustering for sum of rain everyday derivatives

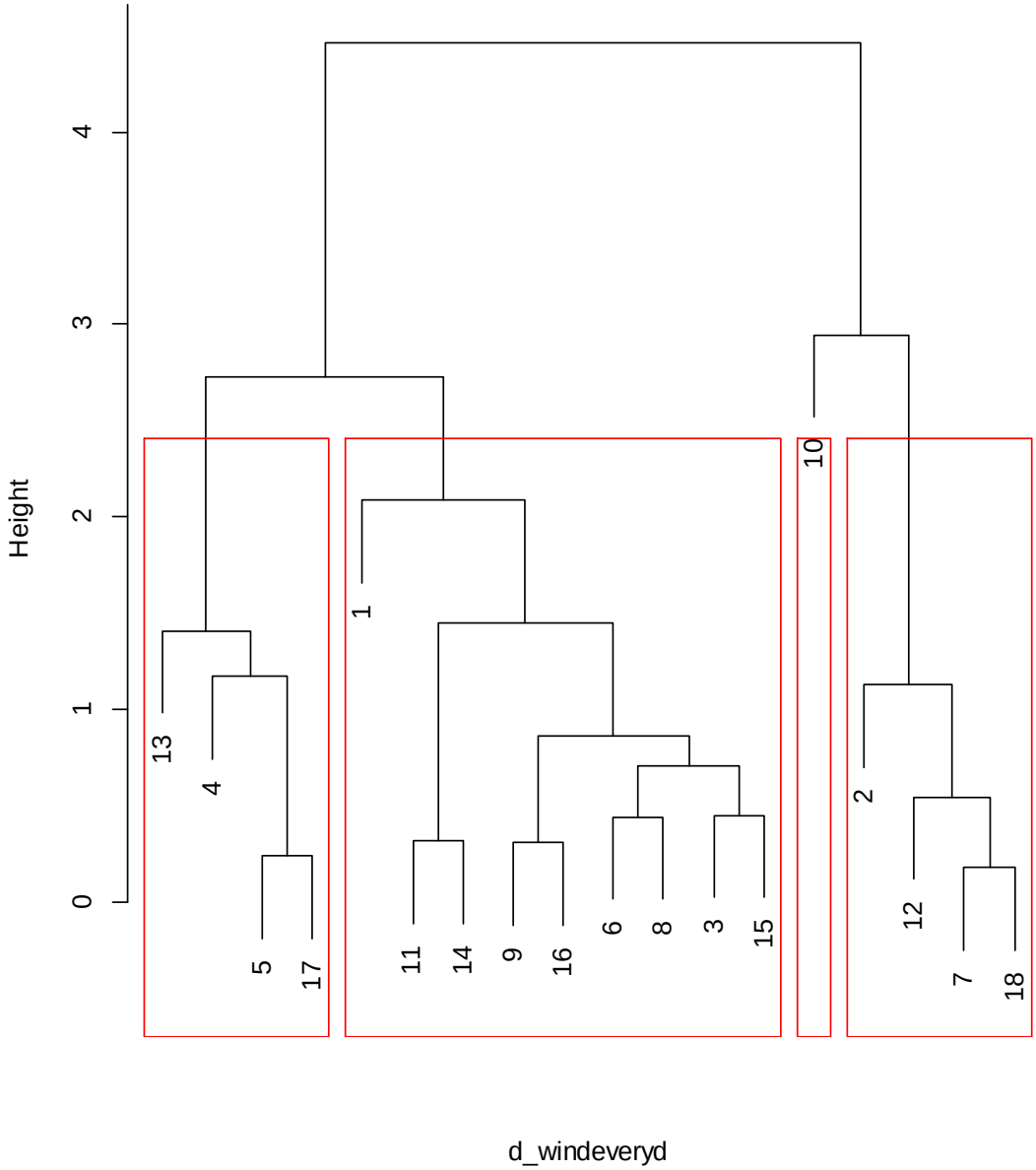

Fig.5.7 Clustering for wind speed everyday derivatives

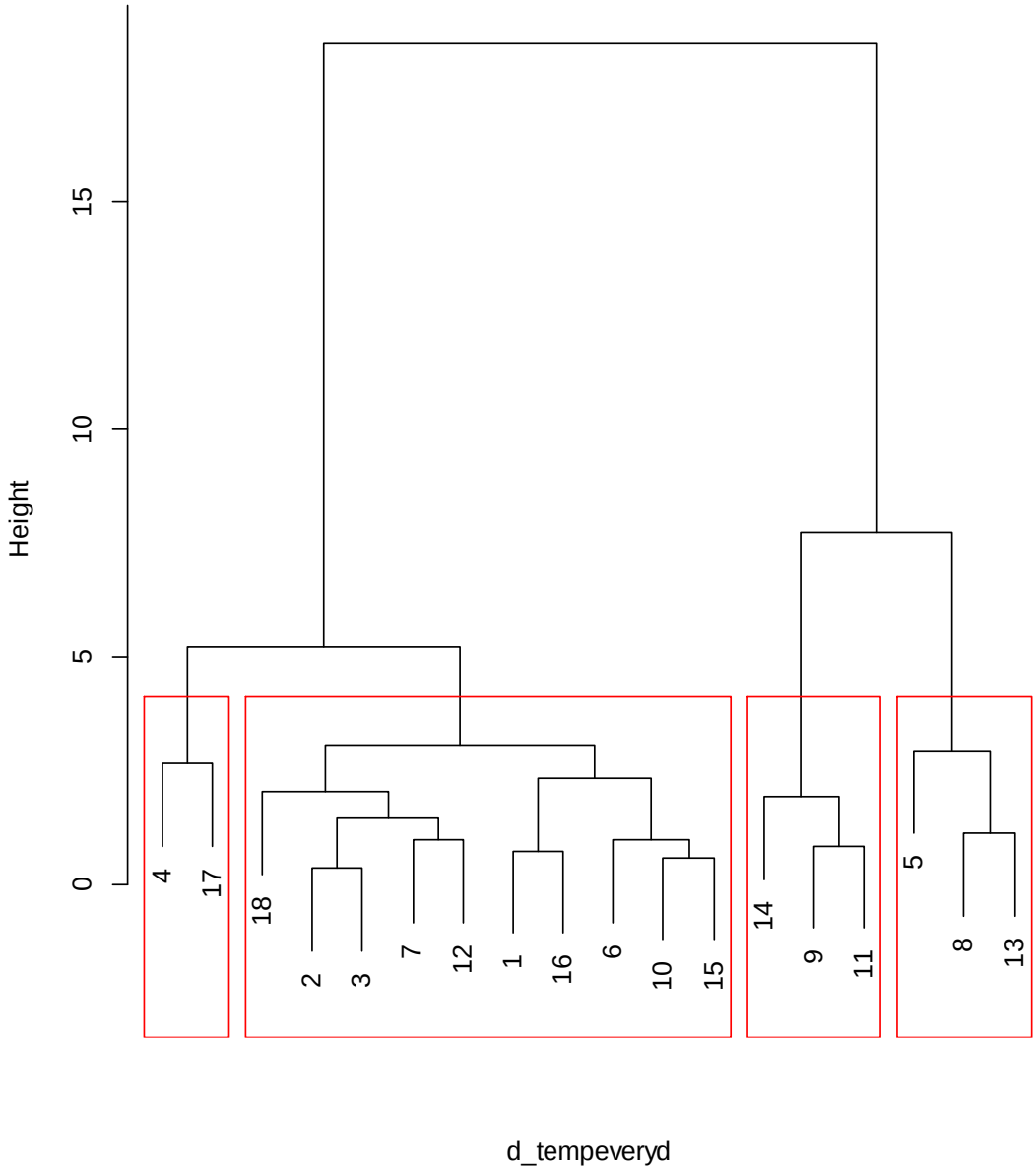

Fig.5.8 Clustering for average temperature everyday derivatives

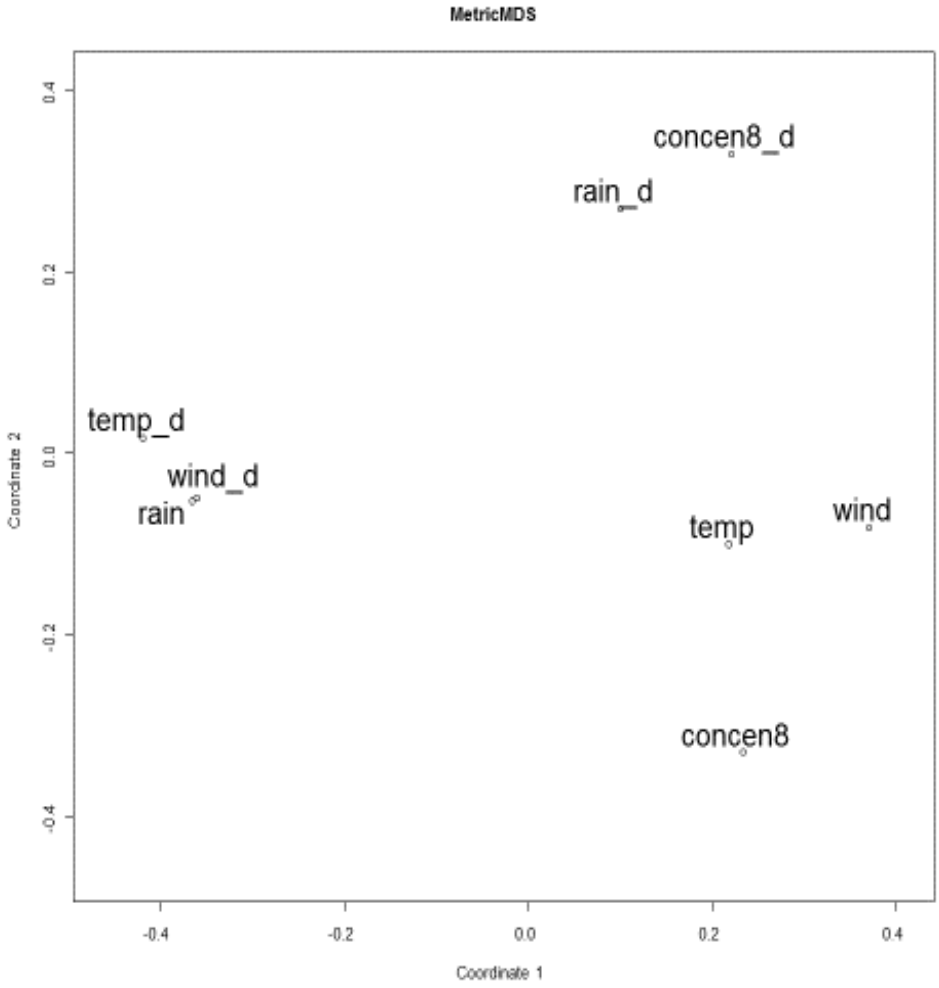

Fig.5.9 Cophenetic similarity distance between variables

6. Conclusion

In our study, we extracted successfully unobserved factors of the fatty acid datasets, reflecting their physicochemical natures, metabolisms, and food sources by three basic dimension reduction methods (factor analysis, principal component analysis, and independent component analysis) as exploratory data analysis. Factor analysis and principal component demonstrated the association of omega-3 fatty acids (20:5 and 22:6) with vitamin D, suggesting fish oil as the common source of vitamin D and we were told that it is the first time to find out vitamin D and omega-3 fatty acids (20:5 and 22:6) are in a positive relation in the aspect of clinic. Secondly, two basic dimension reduction methods (principal component analysis and factor analysis) were conducted to variables related to CE. They both reflected size (concentration), food source, fat solubility, and biological aspect of CE species. However, factor analysis was found more suggestive for a biological aspect of omega group in this study. Cholesteryl docosahexaenoate (DHA) was found unique by factor analysis, possibly relevant to the unique accumulation of DHA in the brain. We analyzed the dataset by common principal component analysis (CPCA) as another method of exploratory data analysis to find out successively that there were four components explainable in the common serum fatty acid dataset: first one was considered as size factor, and more importantly the rest three were considered as plant intake, animal fat intake and fish oil intake correspondingly.

Functional data analysis (functional regression analysis and functional clustering analysis) were conducted as exploratory data analysis to find the unobserved factors of milk fatty acid concentration and the corresponding environment datasets (daily sum of rainfall, daily average wind speed, and daily average temperature). With the advantage of functional data analysis (combine the time characteristics with dataset), we analyzed not only the periodic data themselves but also the periodic changing trend from other environment datasets. We obtained eight dendrograms in functional clustering analysis to reveal the unobserved factors in their relationship. Cophenetic correlation coefficient was used to measure the similarities among the dendrograms. Multidimensional scaling was used to clearly visualize the dissimilarities of the dendrograms. As a result, we found that milk fatty acid concentration has relationships with the average wind speed and the average

temperature, and also that the trend of milk concentration has relation with the trends of daily sum of rainfall and average wind speed. This result can benefit farmers and even milk company to find a proper location in Hokkaido for economical milk production.

We concluded that dimension reductions and functional data analysis can serve as exploratory data analysis to extract valuable information from complex datasets. This combination of dimension reduction and functional data analysis can be considered useful in the field of fatty acid dataset analysis.

In the present study, we conducted three dimension reduction method (principal component analysis, factor analysis, independent component analysis) on the total fatty acid, free fatty acid, cholesteryl ester dataset separately at first. The results revealed several unobserved factors of fatty acid characteristics, such as chain length and saturation. However, those unobserved factors are not as practical as we think in the field of clinical medicine. Therefore, we conducted common principal component analysis on a combined dataset of total fatty acids, free fatty acids and cholesteryl esters. The result suggested possible speculation of food intake sources of fatty acids. This finding might be helpful for reducing the risk of collecting uncertain information based on people's memories typically associated with food frequency questionnaire (FFQ). This approach with common principal component analysis will help, at least, verification of FFQ and possible reduction of research expense and time consumption.

Although, there are many multivariate analysis on milk fatty acid dataset, the study of milk fatty acid dataset with functional data analysis provides a new view point of the milk fatty acid composition. With functional data analysis, we can create new variables of changing trend by first derivatives, furthermore, we can also create new variables of changing acceleration by using second derivatives and forecast the maximum concentration by function characteristics. Additionally, we can combine different characteristics of a dataset with function by the use of basis function. Therefore, functional data analysis was considered to be a powerful tool in exploratory data analysis to capture the unobserved characteristics of dataset concerned concentration.

In our study, we combined cophenetic correlation coefficient with functional data analysis as a new tool to calculate the similarity between dendrograms generated from functional clustering. This combination not only revealed the similarities between

variables, but also revealed the similarities between changing trend of variables. We considered this combination can be applied in the field of concentration analysis.

Therefore, we conclude that functional data analysis is a flexible tool in the study of exploratory data analysis for milk fatty acid concentrations. Functional data analysis with the combination of other useful exploratory data analysis such as common principal component analysis can be useful in future studies. So far we have not found any reports of fatty acid study combined with common principal component analysis or functional data analysis to our best knowledge.

Reference

1. Brockwell. J. P. and Davis, A, R., Introduction to Time Series and Forcasting (second edition),Springer, 2002
2. Chen, Y., Shrestha, R., Chen, Z., Chiba, H., Hui, SP., Okada, E., Ukawa, S., Nakagawa, T., Nakamura, K., Tamakoshi, A., Minami, M., Mizuta, M., Multivariate statistical analysis of serum cholesteryl esters determined by liquid chromatography/mass spectrometry, Analytical Science 36: 373-378, 2020
3. Chen, Y., Shrestha, R., Hirano, K., Hui, SP., Chiba, H., Komiya, Y., and Mizuta, M., Functional clustering approach for fatty acid concentration changes. IFCS2017, Conference Program and book of abstracts August, p.69, 2017
4. Chen, Y., Shrestha, R., Chen, Z., Chiba, H., Hui, SP., Okada, E., Ukawa, S., Nakagawa, T., Nakamura, K., Tamakoshi, A., Minami, H., Mizuta, M., Analyses of Serum Fatty Acids and Vitamin D with Dimension Reduction Methods, ECDA2019, ARCHIVES O F DATA SCIENCE, SERIES A, (accept in 2019)
5. Chen, Z., WU, Y., Shrestha, R., Gao, Z., Zhao, Y., Miura, Y., Tamakoshi, A., Chiba, H., and Hui, SP., Determination of total, free and esterified short-chain fatty acid in human serum by liquid chromatography-mass spectrometry. Annals of Clinical Biochemistry , Vol. 56(2) 190-197, 2019
6. Correddua, F., Serdinoa, J., Mancaa, M., Cosenzab, G., Pauciulloc, A., Ramunnob, L., and Macciottaa, N., Use of multivariate factor analysis to characterize the fatty acid profile of buffalo milk, Journal of Food Composition and Analysis 60, 25–31,2017
7. Flury, B. N., Common principal component analysis in k groups. Journal of the American Statistical Association. 79:388,892-898, 1984
8. Flury, B. N, and Gauschi, W., An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symmetric Matrices to Nearly Diagonal Form, SIAM Journal on Scientific and Statistical Computing, · January, 1986
9. Hooper, D., Coughlan, J., and Mullen, M. R., Structure equation modelling guideline for determining model fit, Electronic Journal on Business Research Methods 6(1) · November, 2007
10. Hui, SP., Murai, T., Yoshimura, T., Chiba, H., and Kurosawa, T., Simple Chemical syntheses of TAG monohydroperoxides, Lipids, Vol. 38, no. 12, 2003

11. Hyvärinen, A. and Oja, E., Independent Component Analysis: Algorithms and Applications, Neural Networks, 13(4-5):411-430, 2000
12. Ihaka, R., Time Serires Analysis, 2005, https://www.stat.auckland.ac.nz/~ihaka/726/notes.pdf
13. Leray, C., Lipids: Nutrition and Health (first edition), CRC Press, 2015
14. Liu, H., Guo, X., Zhao, Q., Qin, Y., and Zhang, J., Lipidomics analysis for identifying the geographical origin and lactation stage of goat milk, Food chemistry 309 125765, 2020
15. Jung, Y., Cho, Y., Kim, N., Oh, Il-Y., Kang, S., Choi, E., Hwang, G., Lipidomic profiling reveals free fatty acid alterations in plasma from patients with atrial fibrillation, PLoS ONE 13(5): e0196709. 2018
16. Khorsan, R., Crawford,C., Ives, J. A., Walter, Avi. R., Jonas,W. B., The Effect of Omega-3 Fatty Acids on Biomarkers of Inflammation: A Rapid Evidence Assessment of the Literature, MILITARY MEDICINE, 179, 11:2, 2014
17. Mardia, K.V., Kent, J.T., and Bibby, J. M., *"Multivariate analysis",* ACADEMIC PRESS, San Diego, CA, U. S. A., 1994.
18. Miura, Y., Furukawa, T., kobayashi, M., Shrestha, R., Takahashi, R., Shimizu, C., Chiba, H., and Hui SP., Absolute quantification of cholesteryl esters using liquid chromatographytandem mass spectrometry uncovers novel diagnostic potential of urinary, *Steroids*, 123, 43, **2017**
19. Nakamura, K., Hui. SP., Ukawa, S., Okada, E., Nakagawa, T., Okabe, H., Chen, Z., Miura, Y., Chiba, H., and Tamakoshi, A., Serum 25-hydroxyvtamin D_3 levels and poor sleep quality in a Japanese population: the DOSANCO Health Study. Sleep Medicine 57 135-140, 2019
20. Nguyen, Long. N., Shui, D. Ma, G., Wong, P., Cazenave-Gassiot, A., Zhang, X., Wenk, M. R., Goh, E. L. K., and Silver, D. L., Mfsd2a is a transporter for the essential omega-3 fatty acid docosahexaenoic acid, Nature, 509, 502, 2014
21. Öhrvall, M., Berglund, L., Salminen, I., lithell, H., Aro, A., and Vessby, B., The serum cholesterol ester fatty acid composition but not the serum concentration of alpha tocopherol predicts the development of myocardial infarction in 50-year-old mean: 19 years follow up, Atherosclerosis, 27, 65,1996
22. Patton, J.S., Stone, B., Papa, C., Abramowitz, R., and Yalkowsky, S. H., Solubility of

fatty acids and other hydrophobic molecules in liquid trioleoylglycerol, J. Lipid Res., 25, 189,1984

23. Preacher, K. J., Zhang, G., Kim, C., and Mels, G., Choosing the Optimal Number of Factors in Exploratory Factor Analysis: A Model Selection Perspective, *Multivariate Behav. Res.,* 48, 28, 2013
24. Ramsay, T. O., Functional data analysis (2nd ed.), Springer, 2005.
25. Ramsay, T. O., and Silverman. B.W.: When the data are functions. Psychometrika December , Volume 47, Issue 4, pp 379-396, 1982
26. Ricciotti, E. and FitzGerald, G.A., Prostaglandins and Inflammation, Arterioscler. Thromb. Vasc. Biol., 31, 986, 2011
27. Saraçli, S., Doğan, N. and Doğan, İ., Comparison of hierarchical cluster analysis methods by cophenetic correlation coefficient. Journal of Inequalities and Applications December, 2013:203, 2013
28. Schonfeld, P. and Wojtczak, L. (2016) Short- and median-chain fatty acids in energy metabolism: the cellular perspective. Journal of Lipid Research Volume 57, DOI 10.1194/jlr. R067629, 2016
29. Serhan, C.N., Pro-resolving lipid mediators are leads for resolution physiology, Nature, 510, 92, 2014
30. Shrestha, R., Miura, Y., Hirano, K., Chen, Z., Okabe, H., Chiba, H., and Hui, SP., Microwave-assisted Derivatization of Fatty Acids for Its Measurement in Milk Using High-Performance Liquid Chromatography, NALYTICAL SCIENCES, MAY, VOL. 34, 2018
31. Vessby, B., Aro, A., Skarfors, E., Berglund, L., Salminen, I., and Lithell, H., The Risk to Develop NIDDM Is Related to the Fatty Acid Composition of the Serum Cholesterol Esters, Diabetes, 43, 1353, 1994
32. Warensjö, E., Risérus, U., and Vessby, B., Fatty Acid Composition of Serum Lipids Predicts the Development of the Metabolic Syndrome in Men, Diabetologia. Oct;48(10):1999-2005, 2005
33. Wong, B.H., Chan, J. P., Cazenave-Gassiot, A., Poh, R. W., Foo, J. C., Galam, D. L. A., Ghosh, S., Nguyen, L. N., Barathi, V. A., Yeo, S. W., Luu, Chi. D., and Wenk, M. R., Mfsd2a is a transporter for the essential ω–3 fatty acid docosahexaenoic acid

(DHA) in eye and is important for photoreceptor cell development, J Biol Chem, 291, 10501, 2016

34. Zhang, Y., Liu, Y., Li, L., Wei, J., Xiong, S., Zhao, Z., High resolution mass spectrometry coupled with multivariate data analysis revealing plasma lipidomic alteration in ovarian cancer in Asian women, Talanta, 150, 88–96, 2016
35. Distances between Clustering, Hierarchical Clustering, https://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf