SEM Analysis of Epigenetic Data

By Azadeh Chizarifard

Department of Statistics
Uppsala University

Supervisors: Åsa Johansson, Rolf Larsson

2014
Abstract

DNA methylation as well as glucosylceramide has been suggested as important factors in cancer development. However, influence of glucosylceramide and DNA methylation on each other, especially direction of these influences, or affection of unknown phenomena on both is unknown. In the present study, multivariate regression and multiple regression analysis are employed to inquire the influences of the glucosylceramide level on the DNA methylation and vice versa. Nine DNA methylation sites were selected based to their pairwise association with glucosylceramide. We investigated the causal relations between the methylation sites and glucosylceramide level by structural equation modeling (SEM) analysis. The multiple regression models suggested that only a subset of the DNA methylation sites were associated with glucosylceramide levels. Even though the DNA methylation sites were selected to be independent, we detected collinearity between them, using multiple regression analyses. Three different models were suggested when SEM analysis was performed with only observed variables. However, the collinearity between DNA methylation sites might suggest the existence of latent variables. When including a latent variable in the SEM analyzes, the model with eliminating sex influence was the best fit to the data. In the other models, "omitted variables bias" problem happened when both sex and age are considered in the model.

Keywords. DNA methylation, Glucosylceramide, Mitochondria, Multivariate regression, Multiple regression, Multicollinearity, Structural Equation Modeling (SEM), Latent variable, Omitted variables bias.
# Contents

1 Introduction 3

2 Background 5
   2.1 Genetics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
      2.1.1 What is DNA? . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
      2.1.2 DNA’s code . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
   2.2 DNA methylation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
   2.3 Glucosylceramide . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
   2.4 Mitochondrion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

3 Statistical Methods 8
   3.1 Regression Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
   3.2 Multicollinearity . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
   3.3 Structural Equation Modeling (SEM) . . . . . . . . . . . . . . . . . . . . . . 11

4 Material 17
   4.1 Study group . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
   4.2 Pre-study knowledge . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
      4.2.1 Determination of DNA methylation status . . . . . . . . . . . . . . . . 17
      4.2.2 Determination of glucosylceramide level . . . . . . . . . . . . . . . . 18

5 Statistical Analysis 18
   5.1 Data screening . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
   5.2 Regression Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
      5.2.1 Multivariate Regression . . . . . . . . . . . . . . . . . . . . . . . . . . 20
      5.2.2 Multiple Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
   5.3 SEM Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
      5.3.1 SEM analysis with observed variables . . . . . . . . . . . . . . . . . . 22
      5.3.2 SEM analysis with observed and latent variable . . . . . . . . . . . . 27

6 Conclusion 36

A Plots of raw dataset 40

B Estimated parameters of section 5.3.1 43
1 Introduction

In recent years, DNA methylation and its role in carcinogenesis has been an important topic. DNA methylation is an important regulator of gene expression. Alterations in DNA methylation are common in a variety of tumors as well as in development (Partha and Rakesh [1]). Glucosylceramide is a membrane lipid that belong the glycosphingolipid family. It has an important and ambiguous role in mammalian cells. High levels of glucosylceramide is associated with cancer and cardiovascular disease. Even though both DNA methylation and glucosylceramide have been associated with cancer development, the influence of DNA methylation on glucosylceramide levels, or vice versa, is unknown. So, discovering whether there is a relationship between DNA methylation and glucosylceramide level is an important research question. The aim of this project is to find out whether or not methylation influences glucosylceramide, if glucosylceramide influences methylation or if unknown variables influence both methylation and glucosylceramide. The study dataset consisted of measurements of the level of DNA methylation at nine sites in the genome, glucosylceramide levels, sex and age of each individual. The nine methylation sites, have been selected out of 480,000 sites through the genome, due to their strong association with glucosylceramide levels. The research question was divided into four parts:

1. If glucosylceramide has influence on methylation sites
2. If methylation has influence on glucosylceramide level
3. If there is a causality between glucosylceramide level and methylation sites
4. If there are some unknown factors behind the relationship of methylation sites and glucosylceramide level

To answer these questions, employed the following statistical analysis techniques: regression analysis and structural equation modeling (SEM). The regression part contained multivariate regression and multiple regression. In the multivariate regression analysis, the methylation sites were considered as response variables. Here, the level of glucosylceramide was defined as an explanatory variable. In other words, the question was detection of linear regressions between the methylation sites and glucosylceramide level individually for each of the nine methylation sites. In the multiple regression analysis, the glucosylceramide level was taken into account as a response variable and the nine methylation sites were considered as explanatory variables; then the impact of the nine methylation sites on glucosylceramide level was investigated. The purpose of this part was to find out which linear combination of methylation sites can explain the variation of glucosylceramide level. In the SEM analysis, at first normality of the dataset was tested since the default estimation method of LISREL, the program used in this study, is Maximum Likelihood. In the present study, the causality investigation was examined at first among observed variables and then among observed variables and a latent variable. In the case of observed variables, the nine methylation sites and glucosylceramide level were taken into
account as the endogenous (or dependent) variables, and sex and age as the exogenous (or independent) variables. The selection criteria of the nine methylation sites have been designed so that they were not supposed to be correlated. Because of finding correlation between them, we tried to find some possible explanations. We saw that four of these methylation sites matched to multiple locations in the genome. Interestingly, all these four methylation sites also mapped to the mitochondria. Therefore, we further designed a SEM model with these four methylation sites, sex and age as the observed variables and the mitochondria as a latent variable.

The structure of the remaining part of the paper is as follows: Section 2 presents some biological background which is useful to understand DNA methylation and glucosylceramide definitions. Section 3 introduces the employed statistical methods in this study. The study materials such as study group are expressed in section 4 and then the statistical results are discussed in section 5. The conclusion of the study is presented in section 6.
2 Background

2.1 Genetics

The material on the presented section comes from [2, 3].

2.1.1 What is DNA?

Our body consists of 100,000,000,000,000 cells which are the basic units of living things. Each cell includes a special set of instructions to make our cells and their components, this set of instruments is called a genome. The human genome is similar among all people and that is the reason of human beings. The position of DNA in a cell is displayed in Figure 1. We get two genomes at the moment of fertilization: one copy of our genome comes from our mother and one copy from our father. In other words, a sperm cell (from the father) and an egg cell (from the mother) give us only one copy of our genome. At the moment of fertilization, a sperm cell and an egg cell join together to make a fertilized egg cell which contains two genomes to make a new person.

Our genome is made from a chemical called deoxyribonucleic acid (henceforth DNA). DNA is a molecule that encodes our genetic instructions and has a specific shape. The shape of DNA looks like a twisted ladder. Imagine that you have a rubber ladder, then hold the bottom of this ladder and twist it from the top. So, you will get a ladder similar to DNA shape. The DNA helix and base pairs shape are shown in Figure 2. Researchers call this shape a "double helix". DNA has rungs which are called base pairs and these rungs are very important in DNA instructions. These pairs can break and allow the sides of the helix to unravel. By this property, DNA can copy itself.

All living organisms have their own package of DNA which is stored in their cells. In plants and animals, the DNA is wrapped around a scaffold of protein. Chromosome is a strand of DNA in a cell. So, we can say that chromosomes are the genetic package which are stored in the nucleus of cells. The situation of chromosome is displayed in Figure 3. The number of chromosomes varies between different species of living organisms. Humans have 46 chromosomes (23 pairs), carp fish have 104 chromosomes (52 pairs) while broad beans have 12 chromosomes (6 pairs).
2.1.2 DNA’s code

We use alphabets as codes in our daily communications. In other words, each word can be translated by the alphabet’s code. For example, "koala" is a word which refers to a special animal that lives in Australia and eats eucalyptus leaves. One way to understand the meaning of "koala" is in terms of the particular order of the letters 'k', 'o', 'a', 'l' and 'a' so, if we change this order we will not get the intended meaning. DNA’s code is considered as an alphabet with only four letters, called A, C, T and G. The meaning of the DNA code depends on the sequence of these four letters. Our cells read the DNA sequence to make the chemicals that we need to survive. A gene is a set of DNA that carries the information to make proteins and usually stored in DNA sequences. In our cells, protein has a very special duty; they break down our food to release energy. They are responsible for every activity in our cells. Genes are the unit of heredity. The gene combinations of living organisms specify the characteristics of that organism such as ‘eye color’, ‘hair color’, ‘blood type’ or ‘smell of a plant’. To show the gene characteristics, it should be translated to the protein.
2.2 DNA methylation

As mentioned above, DNA is an important nucleic acid that stores the genetic information for any given organism. It is made up of four different molecules known as nucleotides; these are referred to as adenine, cytosine, guanine, and thymine. The structure of DNA nucleotide is shown in Figure 4. DNA methylation involves the addition of a methyl group to the fifth position of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring (cytosine and adenine are two of the four bases of DNA). This can be seen in Figure 5. DNA methylation is the biological process by which a methyl group, which is an organic functional group with the formula $CH_3$, is added to the DNA nucleotide. The addition of a methyl group to these nucleotides can serve many important biological purposes, such as preventing potentially damaging viral genetic information that is present in the human genome. [7, 8, 9]

![Figure 4: Structure of DNA nucleotide](image)

![Figure 5: Structures of cytosine and 5-methylcytosine](image)

2.3 Glucosylceramide

According to Messner MC [10], "Glucosylceramide has a unique and often ambiguous role in mammalian cells. Alterations in the level of glucosylceramide are noted in cells and tissues in response to cardiovascular disease, diabetes, skin disorders and cancer. Overall, up-regulation of glucosylceramide offers cellular protection and primes certain cells for proliferation". In other words, this is a molecule that is measured in blood and high levels of this marker is associated with cancer and cardiovascular disease.
2.4 Mitochondrion

Mitochondria is the plural form of mitochondrion. The mitochondria are of great importance for maintaining the function of our body by having the role of as the cell’s powerhouse. They convert energy from food to the form that cells can use. As mentioned most of the DNA is packed in chromosomes within the nucleus of cells and mitochondria have their own DNA. The number of the mitochondria in each cell depends on what the purpose of the cell is; for instance the cell that transmits nerve impulses has fewer mitochondria than in a muscle cell that needs to load of energy. In addition to energy production, mitochondria also play a role in the aging process [19].

3 Statistical Methods

3.1 Regression Analysis

The employed linear regression analyses in this study are multivariate regression and multiple regression.

- **Multivariate linear regression**: There are more than one response (dependent) and predictors (independent) variables and the basic assumptions of multivariate regression are:

  1. multivariate normality of the residuals
  2. homogeneous variances of residuals conditional on predictors
  3. common covariance structure across observations
  4. independent observations

The multivariate regression with \( n \) dependent and \( r \) independent variables can be shown as:

\[
\begin{bmatrix}
  Y_1 \\
  Y_2 \\
  \vdots \\
  Y_n
\end{bmatrix}
= \begin{bmatrix}
  1 & x_{11} & x_{12} & \ldots & x_{1r} \\
  1 & x_{21} & x_{22} & \ldots & x_{2r} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  1 & x_{n1} & x_{n2} & \ldots & x_{nr}
\end{bmatrix}
\begin{bmatrix}
  \beta_0 \\
  \beta_1 \\
  \vdots \\
  \beta_r
\end{bmatrix}
+ \begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \vdots \\
  \epsilon_n
\end{bmatrix}
\]

- **Multiple linear regression**: This type of regression involves more than one regressor (independent) and the most important assumptions concerning regression analysis are (Montgomery [12]):

  1. The relationship between the response and the explanatory variables should be approximately linear.
  2. The error term has zero mean.
  3. The error term has constant variance.
4. The error terms are uncorrelated.
5. The error terms are normally distributed.

The general form of the multiple linear regression with $k$ regressors is presented as:

$$ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon $$

### 3.2 Multicollinearity

The aim of multiple regression analysis is to find the estimates of individual regression coefficients. In order to do multiple regression, if there is no linear relation between the regressors they are called orthogonal. However, in most cases the orthogonality of the regressors is violated. In Montgomery, Peck, Vining [12], four sources of multicollinearity are mentioned:

1. The data collection method used
2. Constraints on the model or in the population
3. Model specification
4. An overdefined model

There are some tests to investigate multicollinearity which are listed in the following process.

- **Correlation Matrix**: By inspection of the off-diagonal of correlation matrix, there is high correlation between some explanatory variables.

- **VIF (variance inflation factor)**: VIF is an index which measures how much variance of an estimated regression coefficient is increased because of multicollinearity and it is represented as

$$ VIF_j = \frac{1}{1 - R_j^2} $$

for $j = 1, \ldots, p - 1$ where $R_j^2$ is the coefficient of multiple determination from the regression of the $j^{th}$ regressor on the remaining regressors.

The values of VIF are interpreted as below:

1. $VIF_j = 1$ when $R_j^2 = 0$, i.e. when the $j^{th}$ variable is not linearly related to the other predictor variables.
2. $VIF_j \rightarrow \infty$ when $R_j^2 \rightarrow 1$, i.e. when the $j^{th}$ variable is linearly related to the other predictor variables.

**Rule of Thumb**: If any of the VIF values exceeds 5 or 10, it implies that the associated regression coefficients are poorly estimated because of multicollinearity. Also, if any of the square roots of VIF are greater than 2 it’s a sign of multicollinearity. (Montgomery et al[12])
• Eigen-system analysis of correlation matrix: If multicollinearity is present in the predictor variables, one or more of the eigenvalues will be small (near to zero).

Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of the correlation matrix. The condition number of correlation matrix is defined as $\kappa = \sqrt{\frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}}$ and $\kappa_j = \sqrt{\frac{\lambda_{\text{max}}}{\lambda_j}}, j=1, \ldots, p$.

**Rule of Thumb:** If one or more of the eigenvalues are small (close to zero) and the corresponding condition number is large, then it indicates multicollinearity. (Montgomery [12])

The presence of multicollinearity has a number of serious effects on the least-square estimations: (Montgomery [12])

1. Existence of multicollinearity causes large variances and covariances for the least-square estimators of the regression coefficients. Although, it is not the only reason of large variances and covariances of regression coefficients. Consider the diagonal elements of the $C=(X'X)^{-1}$ matrix as:

$$C_{jj} = \frac{1}{1-R_j^2}, j=1, \ldots, p$$

where $R_j^2$ is the coefficient of multiple determination from the regression of the $j^{th}$ regressor on the remaining regressors. So, if there is strong multicollinearity between the $j^{th}$ regressor and any subset of regressors then the value of $R_j^2$ tends to unity. On the other hand, if the variance of $\hat{\beta}_j$ is $\text{Var}(\hat{\beta}_j) = C_{jj} \sigma^2$, strong multicollinearity leads to high value of the variance of the least-square estimators of the regression coefficients. Moreover, the covariance of $\hat{\beta}_i$ and $\hat{\beta}_j$ will be large if there is multicollinearity.

2. Multicollinearity leads to too large estimates of individual regression coefficients in absolute value. Define the squared distance from $\hat{\beta}$ to the true vector $\beta$ as:

$$L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta)$$

The expected value of squared distance is defined as

$$E(L_1^2) = E((\hat{\beta} - \beta)'(\hat{\beta} - \beta)) = \sigma^2 Tr(X'X)^{-1},$$

where $Tr$ defies trace of a matrix as we know the trace of a matrix equals to the sum of its eigenvalues (or sum of the main diagonal elements of $X'X$ matrix). So, the expected value is:

$$E(L_1^2) = \sum_{j=1}^{p} \frac{1}{\lambda_j} \text{ where } \lambda_j > 0, j=1, \ldots, p, \text{ are the eigenvalues of } X'X.$$ 

As is explained above, multicollinearity causes some of the eigenvalues of $X'X$ to get small. Then, the distance of the least-squares estimate $\hat{\beta}$ to the true vector $\beta$ will be large. In other words, the vector of $\hat{\beta}$ will get longer than the vector $\beta$. This shows that the least-square method produces poor estimators in the presence of multicollinearity.
3.3 Structural Equation Modeling (SEM)

1. Basic SEM Concepts

The term *structural equation modeling* refers to sets of simultaneous linear equations. These multi equation models can contain observed variables or observed and latent variables. Other terms such as *covariance structure analysis*, *covariance structure modeling* or *causal modeling* might be used in other literature. Causality refers to causal processes which produce observations on multiple variables (Jöreskog[15]). Structural equation modeling contains two important features:

(a) causality is presented by regression equations.

(b) these relations can be modeled such that to reveal the theory under study.

As mentioned before, SEM can examine the relation between the measured and unmeasured (latent) variables. Sometimes the researchers are interested to study some theoretical concepts that are not observable or measurable. These unobservable variables are called *latent variables* or *factors*. So, the researchers should define the latent variable in terms of their scientific belief or knowledge. The measured variables are called observed variables or *indicators* of the assumed underlying constructs.

Factor analysis is the well known statistical method to investigate the relationships among observed variables in terms of a lower number of latent variables. In other words, factor analysis is employed for dimension reduction of variables. The main purpose of factor analysis is to describe the covariation among observed variables in terms of possible underlying latent variables. There are two types of factor analysis: explanatory factor analysis (EFA) and confirmatory factor analysis (CFA).

In EFA, the relation between observed and latent variables is unknown. So, it is used to detect and assess the sources of covariation in observed variables via latent variables. However in CFA, a hypothesized model exists in advance based on the researcher’s knowledge. So, this pre-specified model will be assessed.

There is another model under the SEM concept which is called the *full latent variable model*. In contrast to the factor analysis models, the full latent variable model (LV) can explain the relationships using latent variables. This model is called the full or complete model because it involves a *measurement model* and a *structural model*. The measurement model defines the relation between the observed and latent variables; the structural model defines relations among the latent variables.

The general purpose of SEM is to test how well the dataset fits to the hypothesized model as well as the other statistical modeling. So, the investigators test the goodness of fit between the sample and the SEM model. However, there does not exist a perfect fit between the sample and hypothesized model in reality. Therefore, the existence of discrepancy between estimates and population values is obvious and called *residual*. Then, the model fitting process can be defined as:
Data = Specified Model + Residual

The basic steps in the SEM method are listed as follows: (Kline [14])

(a) Specify the model
This is the most important and difficult step because it is assumed that the model is correct. Then, the results based on the specified model will be analyzed. The causal assumptions derive from prior studies, research design, scientific judgment, or other justifying sources (Morgan [13]).

(b) Evaluate model identification
Most of the time, the researcher cannot analyze the data without monitoring the identification. A model is identified if all of its unknown parameters are identified. The unknown parameter is identified if there is a unique estimated parameter value. In other words, there must be more known parameters than unknown parameters to be estimated. So, models that are not identified should be respecified or we should return to step (a). The different types of identifications can be listed as: Under Identification, Just Identification, Over Identification.

(c) Estimate the model:
The default estimation method of the program (LISREL) is Maximum Likelihood. That means one more assumption needs to be made about the observed variables, the assumption that they follow the multinormal distribution. In practice, the assumption of a multivariate normal distribution of the dataset often does not hold. Violating of the assumption of normality causes the chi-square to be too large so too many models are rejected. Moreover, standard errors will be too small so significance tests will cause the Type I error. There are some approaches to deal with the nonnormality problem such as using GLM (Generalized Least Squares), WLS (Weighted Least Squares), DWLS (Diagonally Weighted Least Squares) or RML (Robust ML) instead of OLS (Ordinary Least Squares) to estimate the coefficients (Jöreskog [16]). Also, Satorra and Bentler [9] found a method to scale chi-square and "robust" standard errors which is a good general approach to dealing with non-normality.

There are two important problems about estimating the model:

i. Evaluate the model fitness: concerning overall fitness of the model
ii. Interpret parameter estimates: concerning validity and reliability of the indicators

- Validity: Validity is defined as a measure of whether an indicator is measuring what it is intended to measure. If a loading is significant on a 5% significance level, it can conclude that it can be regarded as a valid indicator of the concept; which means that the indicator is measuring what it is intended to measure.
Reliability: the squared multiple correlation $R^2$ for each relationship is interpreted as the reliability. This is a measure of the strength of the linear relationship between the dependent variable and the independent variables. The range of $R^2$ is between zero and one; a value close to one indicates that the model is effective.

2. The LISREL and Statistical Model and Notations

There are several SEM-specific packages in R, SAS and LISREL. In the present study, LISREL is used. Understanding the output of LISREL requires some knowledge about its Greek notations which will be explained briefly in the following lines.

Concerning measurement and structural models, we need to define exogenous and endogenous variables. Exogenous variable refers to the independent variable and change in the values of exogenous variables cannot be explained by the model. Endogenous variables point out the dependent variable which is effected by exogenous variables in the model directly or indirectly.

According to Jöreskog’s instructions [16], matrices and their elements are represented by upper and lower case Greek letters respectively; the elements of matrices represent the parameters of the model. The observed variables are shown with Roman letters such that X-variable and Y-variable indicate the exogenous and endogenous variables respectively.

Measurement Model for the X-variables:

$$x = \Lambda_x \xi + \delta$$ (1)

Measurement Model for the Y-variables:

$$y = \Lambda_y \eta + \epsilon$$ (2)

Structural Equation Model:

$$\eta = B\eta + \Gamma \xi + \zeta$$ (3)

The above used notations are defined as follows:

- $x$ := $q \times 1$ vector of observed exogenous variables
- $y$ := $p \times 1$ vector of observed endogenous variables
- $\xi$ := $n \times 1$ vector of latent exogenous variables
- $\eta$ := $m \times 1$ vector of latent endogenous variables
- $\delta$ := $q \times 1$ vector of measurement errors in $x$
- $\epsilon$ := $p \times 1$ vector of measurement errors in $y$
- $\Lambda_x$ := $q \times n$ coefficient matrix relating $x$ to $\xi$
- $\Lambda_y$ := $p \times m$ coefficient matrix relating $y$ to $\eta$
- $\Gamma$ := $m \times n$ coefficient matrix for latent exogenous variable
- $B$ := $m \times m$ coefficient matrix for latent endogenous variable
- $\zeta$ := $m \times 1$ vector of latent errors in equation

Furthermore, we have the following assumptions:
\( E(\eta) = 0, E(\xi) = 0, E(\delta) = 0, E(\epsilon) = 0, E(\zeta) = 0, \)
\( \epsilon \) uncorrelated with \( \eta, \xi \) and \( \delta \)
\( \delta \) uncorrelated with \( \eta, \xi \) and \( \epsilon \)
\( \zeta \) uncorrelated with \( \xi \), \( (I - B) \) is a nonsingular matrix

Then, the general covariance structure can be explained as:
\[ \Phi = Cov(\xi) \]
\[ \Theta = \begin{pmatrix} \Theta_\epsilon & \Theta'_\delta \\ \Theta'_\delta & \Theta_\delta \end{pmatrix} = Cov \begin{pmatrix} \epsilon \\ \delta \end{pmatrix}, \]
then the covariance matrix \( \Sigma \) of \( z = (y', x')' \) is defined as
\[ \Sigma = \begin{pmatrix} \Lambda_Y A (\Gamma \Phi' + \Psi) A' \Lambda'_Y + \Theta_\epsilon & \Lambda_Y A \Gamma \Phi \Lambda_X' + \Theta'_{\delta \epsilon} \\ \Lambda_X \Phi' A' \Lambda'_Y + \Theta_{\delta \epsilon} & \Lambda_X \Phi \Lambda_X' + \Theta_{\delta \delta} \end{pmatrix}, \]
where \( A = (I - B)^{-1} \).

Obviously seen, the elements of \( \Sigma \) are functions of the elements of \( \Lambda_Y, \Lambda_X, B, \Gamma, \Phi, \Psi, \Theta_\epsilon, \Theta_{\delta \epsilon}, \Theta_{\delta \delta} \) which can be represented as three kinds:

- fixed parameters that have specific values
- constrained parameters that are not known but they can be explained in relation to the other parameters
- free parameters that are not known and constrained

3. Fitting and testing a covariance structure

Before explaining the SEM hypothesis, it is necessary to define three situations after specifying the SEM models. According to Jöreskog [16]:

- **SC** or strictly confirmatory: This defines a situation that one single model is formulated by the researcher and the empirical data is gathered to test the model. The presented model is accepted or rejected.

- **AM** or alternative models: Several competing models have been defined. Test them based on a single dataset and select one of them.

- **MG** or model generating: There is a tentative initial model. After testing the initial model, if it does not fit to the certain dataset the model would be modified and tested again. Several models might be tested during the process and finally we get the best model which fits well to the dataset. So, this situation is model generating rather than model testing.

In the **SC** situation, the hypothesis of overall fit of the model to the data is defined as:

\[ H_0 : \Sigma = \Sigma(\theta) \]
\[ H_1 : \Sigma \text{ unconstrained} \]

where, \( \Sigma = \text{population covariance matrix} \)
\( \theta = (\theta_1, \theta_2, ..., \theta_t) \)
\( \Sigma(\theta) \) = model implied covariance matrix

The hypothesis tests the discrepancy between the sample covariance matrix and the covariance matrix implied by the model with the parameter estimates. If it is assumed that the empirical data is a random sample of size \( N \) then the sample covariance matrix \( S \) is obtained. So, the best fit model is achieved where \( \theta \) is estimated in such a way that the covariance matrix \( \Sigma(\theta) \) gets a value equal to \( S \) or by minimizing a fit function \( F[S, \Sigma(\theta)] \).

There are several different estimation methods which can be used in LISREL such e.g. ULS, GLS, ML, DWLS and WLS. To find out the precise definition of them, you can refer to the statistical books. (Jöreskog and Sörbom[20])

It is supposed that \( S \) converges to \( \Sigma_0 \) in probability when sample size grows and, \( \theta_0 \) is the value that minimizes \( F[\Sigma_0, \Sigma(\theta)] \). Let \( \hat{\theta} \) be the estimated value of \( \theta \) which minimizes \( F[S, \Sigma(\theta)] \) and \( n = N - 1 \). So, to test the model calculate \( c = nF[S, \Sigma(\hat{\theta})] \) such that for large sample size \( c \) is approximately distributed as \( \chi^2 \) with \( d = s - t \) degrees of freedom, where \( s = \frac{k(k+1)}{2} \), \( t = \# \) of independent parameters estimated and \( k = \# \) of observed variables. (Jöreskog[16])

4. Selection criteria of the specified models

In the AM situation, the selection process involves three fundamental criteria called the AIC measure of Akaike, the CAIC by Bozdogan, and the single sample cross-validation index ECVI by Browne & Cudeck[16]. They are represented as:

\[
\text{AIC} = c + 2t, \quad \text{CAIC} = c + (1 + lnN)t, \quad \text{ECVI} = (\frac{c}{n}) + 2(\frac{t}{n})
\]

where \( c, t \) and, \( n \) are defined in the past lines.

The definitions of AIC and CAIC are based on information theory, while according to Jöreskog[16]: "ECVI indicates a measure of the discrepancy between the fitted covariance matrix in the analyzed sample and the expected covariance matrix that would be obtained in another sample of the same size". The derived decision based on these three criteria is illustrated such that we choose the model associated with the smallest value.

In the MG case, the researchers do not only test a single model for accepting or rejecting, because they are not in the AM situation to select a model among several specified models. The investigators have a desire to improve an initial model which does not fit so well to the given sample. Here, the main purpose of investigation is to find a model which either fits well to the data or has the property of every parameter being interpretable. The output of the SEM program gives a set of useful information to evaluate and assess the model.

- Examine the solution e.g. look at the \( R^2 \)
- Examine overall fit
- Assess the fitness in details: look at the residuals and standardized residuals
To improve the specified model, we should look at the modification indices which are reported by the program. A modification index is computed for each fixed and constrained parameter in the model. Each modification index indicates how much the chi-square value will be decreased if the associated parameter is set free and the model is reestimated. These indices are applied in the process of model evaluation such that, if the chi-square value is large relative to the degrees of freedom, the modification indices must be examined in order to relax the parameter with the largest modification index value (Jöreskog [16]).

To evaluate the overall fitness of the model, the program gives a long list of fit indices which can assess the model in different perspectives. The indices that are most useful in large sample size and number of variables are introduced here.

- **Chi-Square statistic:** It is the traditional measure for testing the model against the alternative hypothesis. The chi-Square value is a magnitude of discrepancy between sample and fitted covariance matrices. This value is sensitive to violation of the normality test and sample size. Due to large sample size, the $\chi^2$ statistic nearly always rejects the model.

- **Root mean square error of approximation (RMSEA):** It is a measure of discrepancy per degree of freedom. The RMSEA is defined as

$$\epsilon = \sqrt{\frac{\hat{F}_0}{d}}$$

where $\hat{F}_0 = \text{Max}\{\hat{F} - \left(\frac{d}{n}\right), 0\}$, $\hat{F}$ = minimum value of the fit function, $n = N - 1$, $d$ = degree of freedom and $N$ = # sample size

RMSEA is sensitive to increase the number of estimated parameters due to decreased values of $\hat{F}_0$. Brown & Cudeck suggested that a value $0.05 < \epsilon < 0.08$ and $\epsilon > 0.08$ indicates a close fit, mediocre fit and poor fit respectively (Jöreskog [16]).

- **Standardized root mean square residual (SRMR):** Standardized RMR presents the square root of the difference between the residuals of the sample covariance matrix and the hypothesized covariance matrix. The range of its value is between 0 and 1, well fitness of the model is indicated with values less than 0.05. Values higher than 0.08 show poor fitness. SRMR values reduce when there is a high number of parameters and large sample size.

- **Comparative fit index (CFI):** This is a revised version of normal-fit index (NFI) that considers sample size affection. CFI assumes that latent variables are uncorrelated like NFI and constructs the comparison of the sample covariance matrix with the hypothesized covariance matrix based on this assumption. The CFI range is between 0 and 1 with closer values to 1 showing good fit. A threshold of $\text{CFI} \geq 0.95$ is indicating of good fit. CFI is one of the goodness of fit indices least effected by sample size (Hu and Bentler[17]).

- **Parsimony fit indices:** This class of fit indices includes two kinds of parsimony fit indices; the Parsimony Goodness-of-Fit Index (PGFI) and the Parsimonious Normal Fit Index (PNFI). Both of them take into account penalization for model complexity. Although no
threshold levels have been suggested for these indices, obtaining parsimony fit indices inside the 0.5 region while other goodness of fit indices attain over 0.9 is acceptable (Mulaik, et al [18]).

4 Material

4.1 Study group

The study group has been described previously (Besingi and Johansson [11]),"This study is based on the dataset of the Northern Sweden population Health Study (NSPHS) that was initiated in 2006 to provide health survey of the population in the parishes of Karesuando and Soppero, county of Norrbotten and to study the medical consequences of lifestyle and genetics”.

They invited all 3000 inhabitants of which 1069 met the study eligibility criteria, such that they should be greater than 15 years old, and volunteered to participate in the study. The participants’ blood were stored at $-70^\circ$C and genomic DNA for methylation analysis was extracted from the previously frozen peripheral blood using a phenol:chloroform protocol.

4.2 Pre-study knowledge

4.2.1 Determination of DNA methylation status

As described previously (Besingi and Johansson [11]), "Genomic DNA from 432 samples was bisulfite converted using the EZ-DNA methylation kit (ZYMO research) according to the manufacturer’s recommendations. Genome-wide DNA methylation status of 476 366 CpG sites was assessed using the HumanMethylation450k BeadChip (Illumina, San Diego, USA) according to the standard protocol." As a second phase, another set of 310 samples were processed in the same way

Preliminary analysis has been performed, similar as described previously (Besingi and Johansson [11]) such that 476 366 tests, one for each methylation site, were performed to test for association between the methylation level and glucosylceramide level, using sex, age and smoking status as covariates. Two selection criteria were used to select the methylation sites for this project:

1. The sites should be located on different chromosomes.

2. The p-value for the association should be $< 10e-8$.

Based on these two criteria, 9 methylation sites were included in this study.

Out of the 1069 participants, methylation was measured in 734 unique individuals. Table 1 includes the names and the positions of the methylation sites.
### Table 1: Table of methylation sites

<table>
<thead>
<tr>
<th>ID</th>
<th>chromosome</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met1</td>
<td>cg04041942</td>
<td>8</td>
</tr>
<tr>
<td>Met2</td>
<td>cg13771517</td>
<td>9</td>
</tr>
<tr>
<td>Met3</td>
<td>cg22256960</td>
<td>15</td>
</tr>
<tr>
<td>Met4</td>
<td>cg01070250</td>
<td>1 This site maps also to the mitochondria</td>
</tr>
<tr>
<td>Met5</td>
<td>cg26563414</td>
<td>2 This site maps also to the mitochondria</td>
</tr>
<tr>
<td>Met6</td>
<td>cg15890734</td>
<td>5 This site maps also to the mitochondria</td>
</tr>
<tr>
<td>Met7</td>
<td>cg05740793</td>
<td>11 This site maps also to the mitochondria</td>
</tr>
<tr>
<td>Met8</td>
<td>cg08715914</td>
<td>12</td>
</tr>
<tr>
<td>Met9</td>
<td>cg01485645</td>
<td>17</td>
</tr>
</tbody>
</table>

Note: Met 1, ... Met 9 ≡ methylation 1, ..., methylation 9

#### 4.2.2 Determination of glucosylceramide level

The lipid species were quantified by electrospray ionization tandem mass spectrometry (ES-IMS/MS) using methods validated and described previously. Here, the number of sample size of measured glucosylceramide was 700 out of 1069 (Demirkan [21]).

### 5 Statistical Analysis

#### 5.1 Data screening

A total of 681 individuals with measurements of DNA methylation and glucosylceramide were included in this study. Beside the measure of methylation and glucosylceramide level, sex and age are included when investigating association between methylation sites and glucosylceramide level.

The histogram of glucosylceramide level and it's transformation is displayed in Figure 12 which can be found in Appendix A. Obviously, the histogram of glucosylceramide level is positively skewed. However, the histogram of the logarithmic transformation is almost symmetric and looks like a normal distribution. Also, the histogram of methylation sites, which has been represented in Figure 13, shows that most of the histograms are skewer than a normal distribution. Furthermore, normality of each variable and multi normality of all variables was tested; the results can be found in Table 3 and 4 respectively. Although Table 3 shows that the logarithm of glucosylceramide and methylation 9 follow the univariate normal distribution; multivariate normality is not confirmed for the whole dataset. So according to the obtained results, the dataset does not follow the multinominal distribution. Finally based on Figure 12 in Appendix A and Table 3, the logarithm of glucosylceramide is used in the present study which is denoted as "GluCer".

The violation of multivariate normality does not cause a serious problem in the regression anal-
ysis due to the mentioned assumptions concerning regression analysis in section 3.1. However, the multivariate normality assumption is the first question in SEM analysis. Furthermore, the scatter plot of glucosylceramide against each methylation site is displayed in Figure 14 (in Appendix A). In most of them except of the first and third ones, the fitted regression line covers the data well. Maybe using truncation can help to get better results in some cases e.g. glucosylceramide level versus methylation 2, 5, 6 and 9.

To look at the methylation sites deeply, the correlations between them were calculated. These are found in Table 2. Obviously, there is a high correlation between most of the methylation sites.

Table 2: Correlation Matrix of Methylation Sites

<table>
<thead>
<tr>
<th></th>
<th>Met1</th>
<th>Met2</th>
<th>Met3</th>
<th>Met4</th>
<th>Met5</th>
<th>Met6</th>
<th>Met7</th>
<th>Met8</th>
<th>Met9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met2</td>
<td>0.340</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met3</td>
<td><strong>0.684</strong></td>
<td>0.334</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met4</td>
<td><strong>0.450</strong></td>
<td>0.357</td>
<td><strong>0.419</strong></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met5</td>
<td>0.117</td>
<td>-0.069</td>
<td>0.130</td>
<td><strong>0.454</strong></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met6</td>
<td>0.282</td>
<td>0.158</td>
<td>0.309</td>
<td><strong>0.673</strong></td>
<td><strong>0.701</strong></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met7</td>
<td><strong>0.426</strong></td>
<td>0.286</td>
<td>0.398</td>
<td><strong>0.907</strong></td>
<td><strong>0.550</strong></td>
<td><strong>0.760</strong></td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met8</td>
<td>-0.063</td>
<td>-0.001</td>
<td>-0.076</td>
<td>-0.254</td>
<td>-0.155</td>
<td>-0.149</td>
<td>-0.312</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Met9</td>
<td>-0.097</td>
<td>-0.033</td>
<td>-0.148</td>
<td>-0.338</td>
<td>-0.323</td>
<td>-0.309</td>
<td><strong>-0.438</strong></td>
<td><strong>0.409</strong></td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 3: Test of Univariate Normality

<table>
<thead>
<tr>
<th>Variable</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Skewness and Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Z-Score</td>
<td>P-Value</td>
<td>Z-Score</td>
</tr>
<tr>
<td>glu_cer</td>
<td>10.068</td>
<td>0.000</td>
<td>6.253</td>
</tr>
<tr>
<td>GluCer</td>
<td>0.795</td>
<td>0.426</td>
<td>-0.545</td>
</tr>
<tr>
<td>Met1</td>
<td>12.797</td>
<td>0.000</td>
<td>5.950</td>
</tr>
<tr>
<td>Met2</td>
<td>-6.935</td>
<td>0.000</td>
<td>4.407</td>
</tr>
<tr>
<td>Met3</td>
<td>14.639</td>
<td>0.000</td>
<td>8.007</td>
</tr>
<tr>
<td>Met4</td>
<td>-4.193</td>
<td>0.000</td>
<td>-2.074</td>
</tr>
<tr>
<td>Met5</td>
<td>3.089</td>
<td>0.002</td>
<td>-1.794</td>
</tr>
<tr>
<td>Met6</td>
<td>3.819</td>
<td>0.000</td>
<td>-3.869</td>
</tr>
<tr>
<td>Met7</td>
<td>-2.170</td>
<td>0.030</td>
<td>-6.999</td>
</tr>
<tr>
<td>Met8</td>
<td>-2.028</td>
<td>0.043</td>
<td>4.050</td>
</tr>
<tr>
<td>Met9</td>
<td>-0.238</td>
<td>0.812</td>
<td>1.119</td>
</tr>
<tr>
<td>Age</td>
<td>0.937</td>
<td>0.349</td>
<td>-17.085</td>
</tr>
</tbody>
</table>

Note: glu_cer ≡ original value of glucosylceramide level and GluCer ≡ logarithm value of glucosylceramide
Table 4: Test of Multivariate Normality

<table>
<thead>
<tr>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Skewness and Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>Z-Score</td>
<td>P-Value</td>
</tr>
<tr>
<td>16.227</td>
<td>30.897</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Note: The result has gotten from "tests of multivariate normality" in LISREL.

5.2 Regression Analysis

Our study purpose is to find if the methylation sites influence on glucosylceramide or vice versa or if one unknown variable influences on both methylation and glucosylceramide. To investigate these relations, the employed analysis techniques are multivariate regression, multiple regression and structural equation modeling.

5.2.1 Multivariate Regression

In multivariate regression analysis, the glucosylceramide level and methylation sites are considered as the explanatory and response variables respectively. In other words, the influence of glucosylceramide level on each methylation site is tested. As we would expect, all of the 9 tests are significant thus the null hypothesis that the glucosylceramide level does not effect on methylation sites is strongly rejected at 5% level. This result is a direct outcome coming from the second criteria which has been mentioned in the Pre-study knowledge part. Moreover, the result of multivariate regression analysis is shown in Table 5. Then, the represented result simply shows the estimated regression coefficients for each response, and the model summary is the same as we would obtain by performing separate least-squares regressions for the nine responses.

5.2.2 Multiple Regression

In multiple regression analysis, the glucosylceramide level and methylation sites are considered as the response and explanatory variables respectively. So, the effect of methylation sites on the glucosylceramide level is the main problem in this step. The result of multiple regression analysis is shown in Table 6. It is indicated that intercept, methylation 2, 6 and 9 are significant at 5% level without considering the effect of "sex" and "Age" of each individual. The significant variables in the presence of "sex" and "Age" at 5% and 10% levels are intercept, methylation 2, 6, sex, Age and methylation 3, 8, 9 respectively.

After constructing the linear models and based on the correlation matrix, existence of multicollinearity among methylation sites is detected. Multicollinearity can occur in the multiple regression model due to existence of correlation among independent variables. According to diagnostic multicollinearity tests that are described in subsection 3.2, the correlation matrix, VIF values and eigenvalues are presented in Table 2, 7 and 8 respectively. The results demonstrate that:
Table 5: Multivariate regression result

| Regression Model | Estimate | Std. Error | t-value | Pr(>|t|) |
|------------------|----------|------------|---------|----------|
| Met1 ~ GluCer    | $\beta_0$ 0.03385, $\beta_1$ 0.00647 | 2.318 | 0.0208 |
| Met2 ~ GluCer    | $\beta_0$ 0.039262, $\beta_1$ 0.004509 | 6.149 | 1.33e-09 |
| Met3 ~ GluCer    | $\beta_0$ -0.053499, $\beta_1$ 0.009787 | 22.867 | <2e-16 |
| Met4 ~ GluCer    | $\beta_0$ 0.070357, $\beta_1$ 0.004507 | 7.189 | 1.73e-12 |
| Met5 ~ GluCer    | $\beta_0$ 0.163783, $\beta_1$ 0.009787 | 12.66 | <2e-16 |
| Met6 ~ GluCer    | $\beta_0$ 0.041676, $\beta_1$ 0.003269 | 14.90 | <2e-16 |
| Met7 ~ GluCer    | $\beta_0$ 0.117932, $\beta_1$ 0.004513 | 11.58 | <2e-16 |
| Met8 ~ GluCer    | $\beta_0$ 0.062326, $\beta_1$ 0.005414 | 13.81 | <2e-16 |
| Met9 ~ GluCer    | $\beta_0$ 0.313806, $\beta_1$ 0.002692 | 7.316 | 7.21e-13 |

Note: $\beta_0$ and $\beta_1$ express the intercept and slope of the simple linear regression respectively.

1. The correlation matrix reveals high correlation between some regressors which have been highlighted in Table 2.

2. Since two of the VIF values are large, the existence of multicollinearity is obvious. Thus, the VIFs can help identify which regressors are involved in the multicollinearity. The VIF and $\sqrt{VIF}$ values are shown in Table 7.

3. There is one small eigenvalue close to zero, a sign of serious multicollinearity. Moreover, the $\kappa$ value, which is shown in Table 8, is large (>30) which indicates severe multicollinearity.

After all of the above examinations, it is obvious that there exists a strong multicollinearity among methylation sites.

Multivariate statistical techniques such as factor analysis and principal components or techniques such as ridge regression are often employed to "solve" the problem of multicollinearity. The eigenvectors and eigenvalues of the correlation matrix in Table 8 are used to demonstrate factor analysis, principal components and ridge regression.

### 5.3 SEM Analysis

Before starting the SEM analysis, the original dataset should be screened for the multivariate normality problem. As is explained in section 5.1, the observed variables do not follow the
### Table 6: Multiple regression result

| Estimate | Std. Error | t-value | Pr(>|t|) | Regression with sex and age | Estimate | Std. Error | t-value | Pr(>|t|) | Regression without sex and age |
|----------|------------|---------|----------|-----------------------------|----------|------------|---------|----------|--------------------------------|
| $\beta_0$ | 1.2280 | 0.2089 | 5.877 | 6.58e-09 | 1.2532 | 0.21774 | 5.756 | 1.31e-08 |
| $\beta_1$ | -0.0722 | 0.2596 | -0.278 | 0.78095 | -0.19028 | 0.26946 | -0.706 | 0.48034 |
| $\beta_2$ | 1.4039 | 0.2966 | 4.732 | 2.71e-06 | 1.88617 | 0.30094 | 6.268 | 5.65e-10 |
| $\beta_3$ | 0.2802 | 0.1662 | 1.686 | 0.0923 | 0.24130 | 0.17315 | 1.394 | 0.16392 |
| $\beta_4$ | 0.8459 | 0.6106 | 1.385 | 0.1644 | 0.01997 | 0.62196 | 0.032 | 0.97439 |
| $\beta_5$ | 0.1505 | 0.2133 | 0.706 | 0.4806 | 0.29817 | 0.21950 | 1.358 | 0.17480 |
| $\beta_6$ | 1.5920 | 0.6077 | 2.619 | 0.0090 | 2.07789 | 0.63007 | 3.298 | 0.00103 |
| $\beta_7$ | 0.7983 | 0.6874 | 1.161 | 0.2459 | 1.11512 | 0.71304 | 1.564 | 0.11831 |
| $\beta_8$ | -1.0039 | 0.5538 | -1.813 | 0.0703 | -0.73115 | 0.57600 | -1.269 | 0.20475 |
| $\beta_9$ | -0.9607 | 0.5148 | -1.866 | 0.0625 | -1.22518 | 0.53474 | -2.291 | 0.02226 |
| $\beta_{10}$ | -0.0711 | 0.0223 | -3.185 | 0.0015 | 0.0040 | 0.0005 | 7.084 | 3.57e-12 |

Note: The linear regression is expressed as: $\text{GluCer} \sim \beta_0 + \beta_1 \text{Met}1 + \ldots + \beta_9 \text{Met}9 + \beta_{10} \text{sex} + \beta_{11} \text{Age}$

### Table 7: VIF and $\sqrt{VIF}$ values

<table>
<thead>
<tr>
<th>Met1</th>
<th>Met2</th>
<th>Met3</th>
<th>Met4</th>
<th>Met5</th>
<th>Met6</th>
<th>Met7</th>
<th>Met8</th>
<th>Met9</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIF</td>
<td>2.0682</td>
<td>1.3190</td>
<td>1.9918</td>
<td><strong>6.2584</strong></td>
<td>2.1854</td>
<td>3.3891</td>
<td><strong>8.5473</strong></td>
<td>1.2680</td>
</tr>
<tr>
<td>$\sqrt{VIF}$</td>
<td>1.4381</td>
<td>1.1485</td>
<td>1.4113</td>
<td><strong>2.5016</strong></td>
<td>1.4783</td>
<td>1.8409</td>
<td><strong>2.9235</strong></td>
<td>1.1260</td>
</tr>
</tbody>
</table>

multinomial distribution. Then, ML will produce incorrect parameter estimates (i.e., the assumption of a linear model is invalid). Thus, other methods such as ML with robust standard errors and $\chi^2$ (e.g., Bentler, [9]) should be used. The robust ML method, which has been chosen in this study, needs to calculate an asymptotic covariance matrix instead of a covariance matrix which can be obtained by a data screening subprogram of LISREL (PRELIS).

### 5.3.1 SEM analysis with observed variables

As mentioned in subsection 3.3, the first step of SEM is to specify the model. The employed information, which can help to define a reasonable model biologically and statistically, is represented here.

- **Correlation matrices**

  According to Tables 2 and 9, the observed variables can be clustered based on their correlation as follows: $\text{GluCer} \rightarrow \{\text{Met}4\}, \text{Met}6 \rightarrow \{\text{Met}3\}, \text{Met}1 \rightarrow \{\text{Met}4\}, \text{Met}3 \rightarrow \{\text{Met}1, \text{Met}4\}$.
Table 8: Eigenvectors and Eigenvalues of the correlation matrix

<table>
<thead>
<tr>
<th>Eigenvectors</th>
<th>Eigenvalues</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>-0.2962</td>
</tr>
<tr>
<td>$t_2$</td>
<td>-0.4560</td>
</tr>
<tr>
<td>$t_3$</td>
<td>0.1194</td>
</tr>
<tr>
<td>$t_4$</td>
<td>0.7255</td>
</tr>
<tr>
<td>$t_5$</td>
<td>-0.1532</td>
</tr>
<tr>
<td>$t_6$</td>
<td>-0.4291</td>
</tr>
<tr>
<td>$t_7$</td>
<td>0.0839</td>
</tr>
<tr>
<td>$t_8$</td>
<td>-0.0210</td>
</tr>
<tr>
<td>$t_9$</td>
<td>-0.0277</td>
</tr>
<tr>
<td>$\lambda_i$</td>
<td>3.9017</td>
</tr>
</tbody>
</table>

Kappa value: 53.084

Met

\[ Met_4 \rightarrow \begin{cases} \{Met_1, Met_5\}, & Met_5 \rightarrow \begin{cases} \{Met_4\}, & Met_6 \rightarrow \begin{cases} \{Met_4\}, & Met_7 \rightarrow \begin{cases} \{Met_1, Met_4\}, & Met_5 \rightarrow \begin{cases} \{Met_6\}, & Met_6 \rightarrow \begin{cases} \{Met_7\}, \end{cases} \end{cases} \end{cases} \end{cases} \end{cases} \end{cases} \]

Met8 \rightarrow \{Met_9\}, Met9 \rightarrow \{Met_7, Met_8\}

Table 9: Correlation Matrix of whole dataset

<table>
<thead>
<tr>
<th></th>
<th>GluCer</th>
<th>Met1</th>
<th>Met2</th>
<th>Met3</th>
<th>Met4</th>
<th>Met5</th>
<th>Met6</th>
<th>Met7</th>
<th>Met8</th>
<th>Met9</th>
<th>sex</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>GluCer</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met1</td>
<td>0.229</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met2</td>
<td>0.316</td>
<td>0.340</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met3</td>
<td>0.265</td>
<td>0.648</td>
<td>0.334</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met4</td>
<td>0.436</td>
<td>0.450</td>
<td>0.357</td>
<td>0.419</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met5</td>
<td>0.310</td>
<td>0.117</td>
<td>-0.069</td>
<td>0.130</td>
<td>0.454</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met6</td>
<td>0.439</td>
<td>0.282</td>
<td>0.158</td>
<td>0.309</td>
<td>0.673</td>
<td>0.701</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met7</td>
<td>0.468</td>
<td>0.426</td>
<td>0.286</td>
<td>0.398</td>
<td>0.907</td>
<td>0.550</td>
<td>0.760</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met8</td>
<td>-0.172</td>
<td>-0.063</td>
<td>-0.001</td>
<td>-0.076</td>
<td>-0.254</td>
<td>-0.155</td>
<td>-0.149</td>
<td>-0.312</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met9</td>
<td>-0.270</td>
<td>-0.097</td>
<td>-0.033</td>
<td>-0.148</td>
<td>-0.338</td>
<td>-0.323</td>
<td>-0.309</td>
<td>-0.438</td>
<td>0.409</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td>-0.111</td>
<td>-0.070</td>
<td>-0.044</td>
<td>-0.057</td>
<td>-0.033</td>
<td>0.062</td>
<td>-0.005</td>
<td>-0.013</td>
<td>0.004</td>
<td>0.050</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.266</td>
<td>-0.114</td>
<td>0.101</td>
<td>-0.077</td>
<td>-0.073</td>
<td>0.170</td>
<td>0.132</td>
<td>0.012</td>
<td>0.060</td>
<td>-0.051</td>
<td>0.030</td>
<td>1.00</td>
</tr>
</tbody>
</table>

- Selection model based on stepwise regression
  By using the stepwise regression technique (step procedure from R software), the best model based on minimum AIC criteria is found when glucosylceramide level, methylation 1, ... and methylation 9 are regressed on the remaining 9 regressors separately. In
each model, one out of 10 variables is assumed to be a dependent or regressed variable and the other 9 variables are the regressors or predictors. Then the choice of predictive variables is carried out automatically by R. The results are shown in Table 10.

According to the obtained results, the glucosylceramide level influences on methyla-

<table>
<thead>
<tr>
<th>Response variable</th>
<th>Suggested model</th>
</tr>
</thead>
<tbody>
<tr>
<td>glucosylceramide</td>
<td>GluCer \sim Met2 + Met3 + Met4 + Met6 + Met8 + Met9 + sex + Age</td>
</tr>
<tr>
<td>methylation 1</td>
<td>Met 1 \sim Met2 + Met3 + Met6 + Met7 + Met9 + Age</td>
</tr>
<tr>
<td>methylation 2</td>
<td>Met 2 \sim GluCer + Met1 + Met3 + Met4 + Met5 + Met8 + Age</td>
</tr>
<tr>
<td>methylation 3</td>
<td>Met 3 \sim GluCer + Met1 + Met2 + Met6 + Met9 + Age</td>
</tr>
<tr>
<td>methylation 4</td>
<td>Met 4 \sim GluCer + Met1 + Met2 + Met7 + Met9 + Age</td>
</tr>
<tr>
<td>methylation 5</td>
<td>Met 5 \sim Met2 + Met3 + Met6 + Met7 + Met9 + sex + Age</td>
</tr>
<tr>
<td>methylation 6</td>
<td>Met 6 \sim GluCer + Met1 + Met3 + Met5 + Met7 + Met8 + Met9 + Age</td>
</tr>
<tr>
<td>methylation 7</td>
<td>Met 7 \sim Met1 + Met4 + Met5 + Met6 + Met8 + Met9 + Age</td>
</tr>
<tr>
<td>methylation 8</td>
<td>Met 8 \sim GluCer + Met2 + Met6 + Met7 + Met9 + Age</td>
</tr>
<tr>
<td>methylation 9</td>
<td>Met 9 \sim GluCer + Met1 + Met3 + Met4 + Met5 + Met6 + Met7 + Met8 + sex</td>
</tr>
</tbody>
</table>

So based on the above discussed knowledge, glucosylceramide level only effects on methyla-

tion 2, 3, 4, 6, 8, and 9; "sex" has only effect on methylation 5, 9 and glucosylceramide level; "Age" effects on all of the methylation sites except of methylation 9 and glucosylceramide level.

First scenario: the suggested stepwise regression model, where the response variable is gluco-
sylceramide level, is selected for further model specification. Then, two different models are specifed as:

1. Based on Table 10 the red colored methylation sites for which the influence of glucosylceramide level is significant, are chosen as the endogenous variables. Moreover, "Age" has influence on all methylation sites which confirms our pre-knowledge; only methyla-
tion 9 is effected by "sex".

2. All methylation sites are considered as endogenous variables and "Age" and "sex" as the exogenous variables.

In the second scenario, the simplest meaningful model is tested. Based on previous knowledge and tests, it is reasonable to believe that "sex" and "Age" have influence on glucosylceramide level. Then, a model without methylation sites is investigated to find out which of them might be added as an effective variable on glucosylceramide level.
Table 11: Different SEM models

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Specified model</th>
<th>Modified model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First scenario: Model 1</strong></td>
<td>GluCer = Met2 Met3 Met4 Met6 Met8 Met9 sex Age Met2 = Age Met3 = Age Met4 = Age Met6 = Age Met8 = Age Met9 = sex</td>
<td>GluCer = Met2 Met3 Met4 Met6 Met8 Met9 sex Age Met2 = Met4 Age Met3 = GluCer Age Met4 = Met8 Age Met6 = Age Met8 = Age Met9 = Met6 sex Error Terms of Met6 and Met9 be correlated Error Terms of Met8 and Met9 be correlated Error Terms of Met2 and Met4 be correlated Error Terms of Met8 and Met6 be correlated</td>
</tr>
<tr>
<td><strong>Model 2</strong></td>
<td>GluCer = Met2 Met3 Met4 Met6 Met8 Met9 sex Age Met1 = Age Met2 = Met4 Met5 Age Met3 = GluCer Age Met4 = Age Met5 = Met6 Met9 sex Age Met6 = Met7 Age Met7 = Age Met8 = Met9 Age Met9 = sex</td>
<td>GluCer = Met2 Met3 Met4 Met6 Met8 Met9 sex Age Met1 = GluCer Met3 Age Met2 = Met4 Met5 Age Met3 = GluCer Age Met4 = Age Met5 = Met6 Met9 sex Age Met6 = Met7 Age Met7 = Age Met8 = Met9 Age Met9 = Met1 Met4 Met7 sex Error Terms of Met4 and Met7 be correlated Error Terms of Met1 and GluCer be correlated Error Terms of Met8 and Met9 be correlated Error Terms of Met8 and Met6 be correlated</td>
</tr>
<tr>
<td><strong>Second scenario: Model 3</strong></td>
<td>GluCer = sex Age Met2 = Age Met3 = Age Met4 = Age Met6 = Age Met8 = Age Met9 = sex</td>
<td>GluCer = Met4 sex Age Met2 = Age Met3 = Met2 Met4 Age Met4 = Met2 Met8 Met9 Age Met6 = GluCer Met4 Age Met8 = Age Met9 = sex Error Terms of Met8 and Met9 be correlated Error Terms of Met4 and GluCer be correlated Error Terms of Met8 and Met6 be correlated</td>
</tr>
</tbody>
</table>
After running the specified models, some modification indices are suggested by the program in each step. These modification indices, as explained before in subsection 3.3, help to reduce the chi-square values and improve the specified model. Every time after applying the modification indices, it is decided to accept the model or continue to modify it by looking at goodness of fit indices. Modified process are stopped as soon as the model is accepted according to goodness of fit statistics. In this study, the final modified models that are accepted according to the goodness of fit statistics are presented in Table 11.

By looking at the final accepted models, the influence of methylation 4 on glucosylceramide level is obvious in all three models. Also, the error terms of methylation 8 and 9 are correlated in all models which means that methylation 8 and 9 would be correlated even after removing the effects of their explanatory variables. For instance in model 3, methylation 8 and 9 are correlated even after removing the effect of "Age" and "sex". Moreover, these properties are common between model 1 and 2:

- glucosylceramide level influences on methylation 3
- methylation 4 influences on methylation 2
- existence of correlation between error term of methylation 6 and 8

The models have been assessed by the goodness of fit indices during the assessment process. The obtained results of goodness of fit measures that have been used in this study are presented in Table 12. A lot of goodness of fit statistics are reported by the program (Jöreskog [16]). The employed goodness of fit measures in this study are chi-square statistics and its degree of freedom, p-value, the RMSEA and its confidence interval, the SRMR, the CFI and one parsimony fit index such as the PNFI. These indices are the most insensitive to sample size, model misspecification and parameter estimates. All of the reported values in Table 12 are accepted according to their thresholds.

To interpret the models, *reliability* or the squared multiple correlation is a useful aspect.

| Table 12: Goodness of Fit Statistics in the case of SEM analysis with observed variables |
|-------------------------------------|--------|----------|--------|-------------------------------|--------|--------|--------|--------|
| chi-square | df     | p-value  | RMSEA  | confidence interval of RMSEA | SRMR   | CFI    | PNFI   |
| Model 1    | 21.11  | 13       | 0.071  | 0.030 (0.0 ; 0.053)          | 0.026  | 0.99   | 0.36   |
| Model 2    | 36.56  | 31       | 0.23   | 0.016 (0.0 ; 0.034)          | 0.020  | 1.00   | 0.47   |
| Model 3    | 24.28  | 16       | 0.084  | 0.028 (0.0 ; 0.049)          | 0.026  | 0.99   | 0.44   |

The squared multiple correlation ($R^2$) is interpreted as the reliability. The result of the model reliability is illustrated in Table 13. $R^2$ is a measure of the strength of the linear relationship e.g. the most strength of the linear relationship in model 2 belongs to

\[ Met6 = Met7 + Age \]

which is represented as $R^2 = 0.59$.  

26
The estimates of the parameters for all models in matrix form are represented in Appendix B. The comparison of the presented models based on AIC, CAIC and ECVI values are shown in Table 14. As can be seen from the table, the values of AIC and CAIC are appear in descending order across Model 2, Model 1 and Model 3. So, the smallest values of AIC and CAIC belong to Model 3. Moreover, Model 1 and 3 have the smallest value of ECVI. Therefore, Model 3 is preferred in terms of the AIC, CAIC and ECVI criteria. However, the differences of AIC and CAIC values between Model 3 and 1 are not so large.

5.3.2 SEM analysis with observed and latent variable

Based on the selection criteria of the 9 methylation sites, we would expect them to be independent. However, in my analyses it was seen that they are highly correlated. Therefore, we spent some time on finding out why. The first finding explanation was that for 4 sites (methylation 4 - methylation 7) the probes used for determining methylation levels, matched to multiple locations in the genome. Interestingly, all these four mapped to the mitochondria. Based on our findings there are some possible explanations for what we are really measuring with these four probes.

1. That they are truly measuring the methylation level at four different chromosomes as was the intention. This is quite unlikely since the correlation between two of the sites are larger than 0.9.

2. That they are measuring the methylation level of the mitochondria. Then all four probes are measuring the same thing. This is the most possible case.
3. That they are measuring the ratio of nuclear (by the chromosomal position the prove were intended to measure) to mitochondria DNA. E.g. it is assumed that the chromosomal position that they were intending to measure had a methylation level of 20% and the mitochondrial position that they are also measuring has a different methylation level (e.g. 80%). If they have 20 copies of the mitochondria DNA per copy of the chromosome then it would be expected the measured value to be \( \frac{20\% \times 1 + 80\% \times 20}{21} \).

However, for another individual with only 2 copies of the mitochondria DNA per copy of the chromosome the measured value will be \( \frac{20\% \times 1 + 80\% \times 2}{3} \).

Based on the above explanations, measurements of the mitochondria is a likely explanation for the correlation of methylation 4, 5, 6 and 7. The mitochondria is considered as a latent variable; then the relationships among the observed variables and the latent variable are of interest. The observed variables are represented as methylation 4, ..., 7, "sex", "Age" and glucosylceramide level.

The specified model is shown in Figure 6. By convention, in the presented path diagram the measured (observed) variables are shown in boxes and unmeasured (latent) variable in an ellipse. The gray and blue boxes represent the exogenous measured and endogenous variables respectively. The left and right parts of the path diagram are defined by equations 1 and 2 in section 3.3. So, there is one latent variable and 7 observed variables. Associated with each observed variable is a small one way arrow indicating an error term. The curved two-way arrow represents correlation or association between pairs of variables.

To investigate the relationships among the mentioned observed variables and mitochondria, the flowchart in Figure 7 is applied.

### 5.3.2.1 Model considering age and sex influences

After running the specified model in Figure 6 and applying the suggested modification indices, the obtained model is not yet acceptable in terms of the goodness of fit statistics. To improve the model, there are two options: (1) adding a correlation between the error term of "sex" and methylation 5, (2) ignoring the direct path from "sex" to the mitochondria. According to t-values, the affection of "sex" on mitochondria is not significant; on the other hand, the modification index suggests adding the correlation between the error term of "sex" and methylation 5.

One of the fundamental assumptions in SEM is that the error terms of the independent variable and the dependent variable in each relationship cannot be correlated. Existence of correlation between the error terms of independent (exogenous) and dependent (endogenous) variables causes the omitted variables bias problem. The omitted variables bias happens because it is often the independent variables in the model account for only a fraction of the variation and covariation for the dependent variable because the other effective independent variables are not included in the model. In other words, omitted variables are variables that significantly influence on the dependent variable and should be included in the model, but are excluded. Violation of this fundamental assumption leads to biased and inconsistent estimates of the structural coefficients in the linear equations.
1. **Model considering correlation between error terms of sex and Met 5** Consider the model with adding a correlation between the error term of dependent variable (methylation 5) and independent variable (sex). The existence of a correlation between error term of "sex" and the error term of dependent variable ("Met 5") can be interpreted such that there is a variable that causes both the "Met 5" and "sex" and that variable is not included in the model. Before and after adding the correlation between the error terms, the t-value shows that "sex" is not a significant variable at 5% level. In other words, the influence of gender on the mitochondria is not significant in the present model.

2. **Model without considering sex** In the model without considering gender affection, t-values show that all of the parameters are significant and unlike the former model there is no correlation between error term of dependent and independent variables.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\lambda}_1$</th>
<th>$\hat{\lambda}_2$</th>
<th>$\hat{\lambda}_3$</th>
<th>$\hat{\lambda}_4$</th>
<th>$\hat{\lambda}_5$</th>
<th>$\hat{\gamma}_1$</th>
<th>$\hat{\gamma}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1: correlated error terms (sex and Met5)</td>
<td>1</td>
<td>0.12</td>
<td>0.05</td>
<td>0.20</td>
<td>0.21</td>
<td>4.55</td>
<td>-0.08</td>
</tr>
<tr>
<td>1.2: without sex affection</td>
<td>1</td>
<td>0.12</td>
<td>0.05</td>
<td>0.20</td>
<td>0.21</td>
<td>4.43</td>
<td>-</td>
</tr>
</tbody>
</table>
The conceptual plots of both models are displayed in Figures 8 and 9. Their estimated parameters and goodness of fit indices are shown in Tables 15 and 16 respectively. In both diagrams, the gray straight path from mitochondria to glucosylceramide indicates scaling the latent variable.

The latent variables are unobservable and unmeasurable; so they have no natural scale. Then to define the model properly, both the origin and the unit of measurement of each latent variable must be defined. According to Jöreskog [16], one way to set the scale of the latent variable is fixing a non-zero coefficient (usually one) in the relationship for one of its observed indicators, that has been used in this study. This defines the unit for each latent variable in relation to one of the observed variables, which is the reference variable. In practice, one chooses as reference the observed variable, which, in some sense, best represents the latent variable. In the present study, the glucosylceramide level is considered as a reference variable. The coefficient of the "GluCer" on "mitochon" ($\lambda_1$) will not be estimated and will be fixed equal to one. However,
the coefficient of the rest of the observed indicators will be estimated. The comparison criteria of the model without "sex" influence (model 1.2) and the model that considers the correlation between the error term of "sex" and methylation 5 (model 1.1) are shown in Table 21. Based on this table, the model without considering the influence of gender on mitochondria is preferable because the smallest values of AIC, CAIC and ECVI belong to this model.

Figure 8: Conceptual plot of model 1.1

Figure 9: Conceptual plot of model 1.2

5.3.2.2 Model without considering age influence Another interesting aspect of the relationship between methylation $4, ..., 7$ and glucosylceramide level as the observed variables and mitochondria as the latent variable is investigating the impact of gender solely without the presence of the age affection. To find out, two situations are considered that can eliminate the effect of age.
1. **Adjusting model by "Age"**: Because the range of "Age" is so broad (from 15 to 96 years old) it can be considered as a continuous variable and its effect can be adjusted. At first, each variable (i.e. all of the methylation sites and the glucosylceramide level, not sex) is regressed on "Age" separately and then the variable values are replaced by the corresponding residuals from the different regressions.

After running the model with the new calculated variable values (the obtained residuals), based on some fit indices that are noticeable in the large sample size situation such as NNFI, PNFI, CFI and standardized RMR, the suggested model is accepted. However, one more modification is suggested to add a correlation between the error term of "sex" and glucosylceramide level. Although after adding this correlation, the p-value of the model is improved the other goodness of fit indices have not changed dramatically. On the other hand, the obtained estimated coefficients in the new model (model considering the correlated error terms) are the same as for the former model (model not considering the correlated error terms). The models without correlated error terms and with correlated error terms are called *model 2.1.1* and *model 2.1.2* respectively.

So, there is not much more gained by adding the correlation between error terms of exogenous and endogenous variables. The estimated parameters and the goodness of fit indices of both models are represented in Tables 17 and 18. Note that, the estimated parameters in model 2.1.2 are biased due to the existence of the omitted variables bias problem.

Moreover, "sex" is not significant neither in the model without considering error term

### Table 17: Estimated parameters of models 2.1.1 and 2.1.2

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\lambda_3$</th>
<th>$\lambda_4$</th>
<th>$\lambda_5$</th>
<th>$\gamma_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>without correlated error terms</td>
<td>1</td>
<td>0.22</td>
<td>0.23</td>
<td>0.14</td>
<td>0.27</td>
<td>0</td>
</tr>
<tr>
<td>with correlated error terms (sex and glucosylceramide)</td>
<td>1</td>
<td>0.22</td>
<td>0.23</td>
<td>0.14</td>
<td>0.27</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 18: Goodness of Fit Statistics of models 2.1.1 and 2.1.2

<table>
<thead>
<tr>
<th>Model</th>
<th>Satorra-Bentler Scaled Chi-Square</th>
<th>Degrees of Freedom</th>
<th>P-value</th>
<th>RMSEA</th>
<th>90% CI of RMSEA</th>
<th>SRMR</th>
<th>CFI</th>
<th>PNFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>without correlated error terms</td>
<td>29.19</td>
<td>6</td>
<td>0.00</td>
<td>0.075</td>
<td>(0.049 , 0.10)</td>
<td>0.044</td>
<td>0.95</td>
<td>0.38</td>
</tr>
<tr>
<td>correlated error terms (sex and glucosylceramide)</td>
<td>10.38</td>
<td>5</td>
<td>0.065</td>
<td>0.040</td>
<td>(0.0 , 0.074)</td>
<td>0.021</td>
<td>.99</td>
<td>0.33</td>
</tr>
</tbody>
</table>

of "sex" and glucosylceramide level nor in the model considering this correlation. This
is confirmed by testing the relationships among four methylation sites (Met 4 - Met7), glucosylceramide level and the mitochondria in the separate gender groups. This model is called model 2.1.3.

Here, it is of interest to test whether the model is similar between different groups (male and female). In other words, the indicators (Met4, ... , Met7 and glucosylceramide) measuring the same underlying factor (mitochondria) in different groups. The obtained results based on p-values indicate that the factor loadings and the error variances of observed variables are almost invariant between female and male groups.

The reliability of different sex groups as female and male can be found in Table 19. This table shows that there is no dramatic difference between the reliabilities of indicators of mitochondria among male and female. Furthermore, the most reliable indicator of mitochondria is methylation 7 in both groups. The conceptual plot of model 2.1.3 is displayed in Figure 10.

### Table 19: Squared Multiple Correlations of model 2.1.3

<table>
<thead>
<tr>
<th>Group</th>
<th>new-GluCer</th>
<th>new-Met4</th>
<th>new-Met5</th>
<th>new-Met6</th>
<th>new-Met7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Female</td>
<td>0.27</td>
<td>0.83</td>
<td>0.34</td>
<td>0.59</td>
<td>1.00</td>
</tr>
<tr>
<td>2: Male</td>
<td>0.27</td>
<td>0.83</td>
<td>0.29</td>
<td>0.58</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: The prefixes of "new" are indicating that the new calculated values in terms of their residuals.

![Figure 10: Conceptual plot of model 2.1.3](image)

2. Model 2.2: **Grouping model by "Age" (Multi-sample structural equation modeling):**

To eliminate the age influence in the model, the data is partitioned to 6 different age groups and then the structural equation modeling analysis is used separately on all of them. All of the six groups have almost equal sample sizes as \( n_1 = 116, n_2 = 116, \)
Figure 11 shows the specified model in each group. There are three sets of parameters in the model: (1) the five factor loadings corresponding to the paths from mitochondria to the observed dependent variables, (2) the six error variances of the observed variables, and (3) the path from sex to mitochondria. It is desirable to investigate to what extent each of these sets of parameters are invariant over groups.

In multi-sample structural equation modeling, at first the equality of factor structures should be tested. The equality of factor structures means that all relationships and all parameters are the same over groups. To test the identicality over the groups:

(a) Test if the factor loadings (coefficients) are equal over the groups. In our study, the result reveals that factor loadings are different for the six groups.
(b) Test if the error variances of observed variables are invariant over the groups. The obtained result shows that they are not equal for six groups.

After modifying the model over the six different age groups, the reliabilities (squared multiple correlations) which are shown in Table 20 are obtained. The table indicates that the most reliable indicator of mitochondria is methylation 7 in group 4 and 6. In other word, the linear relationship of methylation 7 and mitochondria in group 4 and 6 is very strong.

The comparison criteria of the six models are presented in Table 21; although, the comparison of models 2.1.3 and 2.2 with the other models is not correct. Due to, in multi-sample examples the model specifies and fits for each group separately e.g. in model 2.2 we estimated the five
factor loadings in all six groups. Then, the chi-square is a measure of fit of all models in all groups, and, this value cannot be decomposed for each group separately. LISREL reports the global goodness of fit statistics (RMSEA, CFI, AIC, ...) in the output.

It can be seen that the smallest values of AIC and ECVI, between the models 1.1, 1.2, 2.1.1 and 2.1.2, belong to model 1.2. On the other hand, model 2.1.2 has the smallest value of ECVI. However, the comparison criteria values of model 1.2 and 2.1.2 are so close to each other and there is no huge difference in their criteria values. On the other hand, these two models (the model which excludes sex influence on the mitochondria and the model that eliminates the age influence from the model by adjusting it in terms of age and then considering the correlated error between "sex" and glucosylceramide level) can explain the research relationships of 4 probes (Met4 - Met7) and the mitochondria based on the analyzed sample better than the others.

To sum up, the gained results in the presence of latent variable indicates that, the best fitted models to the dataset are obtained in two situations: models 1.2 and 2.1.2.

Table 20: Squared Multiple Correlations of model 2.2

<table>
<thead>
<tr>
<th>Group</th>
<th>GluCer</th>
<th>Met4</th>
<th>Met5</th>
<th>Met6</th>
<th>Met7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Age &lt; 22</td>
<td>0.14</td>
<td>0.74</td>
<td>0.12</td>
<td>0.79</td>
<td>0.29</td>
</tr>
<tr>
<td>2: 22 ≤ Age &lt; 37</td>
<td>0.27</td>
<td>0.27</td>
<td>0.47</td>
<td>0.31</td>
<td>0.67</td>
</tr>
<tr>
<td>3: 37 ≤ Age &lt; 47</td>
<td>0.08</td>
<td>0.71</td>
<td>0.23</td>
<td>0.71</td>
<td>0.28</td>
</tr>
<tr>
<td>4: 47 ≤ Age &lt; 58</td>
<td>0.29</td>
<td>0.24</td>
<td>0.09</td>
<td>0.14</td>
<td>1.00</td>
</tr>
<tr>
<td>5: 58 ≤ Age &lt; 71</td>
<td>0.19</td>
<td>0.78</td>
<td>0.42</td>
<td>0.49</td>
<td>0.26</td>
</tr>
<tr>
<td>6: Age ≥ 71</td>
<td>0.50</td>
<td>0.23</td>
<td>0.17</td>
<td>0.14</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 21: Criteria of models comparison in the case of SEM analysis with observed and latent variable

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>CAIC</th>
<th>ECVI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: With the presence of age and sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1 with correlated error terms (sex and Met5)</td>
<td>53.77</td>
<td>191.86</td>
<td>0.079</td>
</tr>
<tr>
<td>1.2 delete sex</td>
<td>41.68</td>
<td>146.63</td>
<td>0.061</td>
</tr>
</tbody>
</table>

| 2: With the presence of sex                |       |       |       |
| 2.1: adjust age                            |       |       |       |
| 2.1.1 without correlated error terms (sex and GluCer) | 59.19 | 142.04| 0.085 |
| 2.1.2 correlated error terms (sex and GluCer) | 42.38 | 130.75| 0.062 |
| 2.1.3 grouping sex                         | 78.01 | 238.20| 0.11  |

| 2.2: grouping age                          |       |       |       |
|                                            | 230.79| 757.19| 0.34  |

Note: The best fitted models among the first four models are 1.2 and 2.1.2.
6 Conclusion

Investigating the relationships between the selected methylation sites and glucosylceramide level was the main goal in this study. To deal with this question, the employed statistical methods were: multivariate regression analysis, multiple regression analysis and SEM analysis.

In the multivariate regression analysis, dependent and independent variables were defined as the methylation sites and the glucosylceramide level respectively. Based on the selection criteria of methylation sites, there must be an association between glucosylceramide level and each individual methylation site. Therefore, it was not surprising that, the influence of glucosylceramide level on all of the nine methylation sites was significant.

In the multiple regression analysis, the glucosylceramide level and the methylation sites were considered as dependent and independent variables in regression analysis. Based on the selection criteria of methylation sites, it would be expected that the nine methylation sites were uncorrelated. One of the most important results of this part was the existence of multicollinearity between the nine selected methylation sites. The regression result revealed that all of the methylation sites do not have significant influence on the glucosylceramide level. It can be interpreted as a consequence of the multicollinearity.

In the SEM analysis, normality of the dataset was not confirmed by "tests of multivariate normality" in LISREL. This test is necessary before going through the SEM analysis since the default estimation method is Maximum Likelihood. There are some alternative methods such as WLS, DWLS and RML to deal with nonnormality for non-normal variables (Jöreskog [16]). The employed method in this study was RML which is suitable continuous non-normal variables. The SEM analysis was applied in two aspects with and without the presence of latent variables.

In the SEM analysis with only observed variables, the nine methylation sites and glucosylceramide level were considered as the endogenous (or dependent) variables and, sex and age are the exogenous (or independent) variables. To specify the primary casual models, we used previous biological study knowledge and some statistical techniques. The three fitted models resulted in, had minimum discrepancy between the estimated and the sample covariance matrices. The causal processes were represented by the series of structural relations in these three models. Interestingly, the simplest model resulted in best test statistics, and we can therefore conclude that the strong association seen between glucosylceramide and all the individual methylation sites is more likely due to one or few underlaying factors.

Finding the correlation between the nine methylation sites was a motivation to detect if there are some possible unknown factors behind them. We could only find one possible biological factor that could serve as a latent variable for four of the methylation sites (methylation 4, ...,7). However, one interesting finding was that the behavior of "sex" is different in the presence and absence of "Age". Only two out of four models obtain the best fit to the dataset, and those were when deleting "sex" influence alone or at first eliminating "Age" influence and then considering a correlation between error term of "sex" and glucosylceramide. The result showed that
there was no difference between the male and the female groups in adjusted model by "Age". Moreover, when we grouped model by "Age" there was no omitted variables bias problem. The analyses with the latent variable revealed that there is a correlation between the error term of "sex" and some dependent variables in most models. This correlation can be interpreted as the omitted variables bias problem. In other words, "sex" does not account for the whole variation and covariation of the dependent constructs in these specified models. According to these results, it is possible that the specified model might not have considered all required variables associated with the dependent constructs, to describe the complex biological nature of the data. In summary, we have shown that methylation sites, even located on different chromosomes, are highly correlated. Their individual association with glucosylceramide might be partly explained by the fact that the probes measuring the methylation level at four of the sites also maps to the mitochondria. However, it is likely that additional unknown factors, that influence as well DNA methylation as glucosylceramide are also involved, which suggests that more biological research is needed in this area.
References


Figure 12: Histogram of glucosylceramide and log(glucosylceramide)
Figure 13: Histogram of methylation sites
Figure 14: Scatter Plot of methylation sites vs log(glucosylceramide)
## B Estimated parameters of section 5.3.1

The hypothesis model in a compact matrix form is written as:

\[ Y = \alpha + BY + \Gamma X + \zeta \]

where \( \alpha \) is a vector of intercepts terms, \( B \) and \( Y \) are the coefficient matrices associated with the effects of \( Y \) and \( X \) on \( Y \), \( \zeta \) is a vector of random disturbance terms (error terms in equations). The estimate of the parameters:

1. First Scenario

- **Model 1**

\[
\hat{\alpha} = \begin{pmatrix} 0.60 \\ 0.25 \\ -0.35 \\ 0.42 \\ 0.19 \\ 0.21 \\ 0.32 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} 0^* & 2.81 & -2.72 & 2.87 & 3.06 & -0.84 & -1.49 \\ 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\ 0.23 & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\ -0.05 & 0.00 \\ 0^* & 0.00 \\ 0^* & 0.00 \end{pmatrix}, \quad \hat{\Gamma} = \begin{pmatrix} 0^* & 0.14 \\ 0^* & 0.00 \\ 0^* & 0.00 \\ 0^* & 0.00 \\ 0^* & 0.00 \\ 0^* & 0.00 \end{pmatrix}, \quad \hat{\Psi} = \begin{pmatrix} 0^* & -0.24 & 0^* & 0^* \end{pmatrix}
\]

where \( 0^* \) denotes that the corresponding parameters were fixed to 0 when it was defined the model.

- **Model 2**

\[
\hat{\alpha} = \begin{pmatrix} 1.11 \\ -0.01 \\ 0.25 \\ -0.37 \\ 0.30 \\ 0.25 \\ 0.05 \\ 0.26 \\ 0.04 \\ 0.32 \end{pmatrix}, \quad \hat{\Gamma} = \begin{pmatrix} -0.05 \\ 0^* \\ 0^* \\ 0^* \\ 0^* \\ 0^* \\ 0^* \end{pmatrix}, \quad \hat{\Psi} = \begin{pmatrix} 0.00 \end{pmatrix}
\]
Let

For instance in first scenario, Model 1 can be explained in the matrix form as:

\[
\hat{\Psi} = \begin{pmatrix}
0.15 \\
-0.01 & 0.00 \\
0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0.01 \\
0^* & 0^* & 0^* & 0^* & 0.00 \\
0^* & 0^* & 0^* & 0^* & 0^* & 0.00 \\
0^* & 0^* & 0^* & 0^* & 0.00 & 0^* & 0.00 \\
0^* & 0^* & 0^* & 0^* & 0^* & 0.00 & 0^* & 0.00 \\
0^* & 0^* & 0^* & 0^* & 0^* & 0.00 & 0^* & 0.00 & 0.00 & 0.00
\end{pmatrix},
\]

\[
\hat{\Theta} = \begin{pmatrix}
0^* & 0^* & 2.92 & -2.82 & 2.98 & 0^* & 2.42 & 0^* & -1.45 & -2.66 \\
0.04 & 0^* & 0^* & 0.51 & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0.50 & -0.19 & 0^* & 0^* & 0^* & 0^* \\
0.24 & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0.54 & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0.15 & 0^* & 0^* & 0^* & -0.420^* & 0^* \\
0^* & 0.09 & 0^* & 0^* & 0.15 & 0^* & 0^* & 0^* & 0^* & 0^*
\end{pmatrix},
\]

2. Second Scenario

\[
\hat{\alpha} = \begin{pmatrix}0.28 \\
-0.23 \\
0.37 \\
0.02 \\
0.21 \\
0.27 \end{pmatrix}, \quad \hat{\Theta} = \begin{pmatrix}
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0.48 & 0^* & 0.68 & 0^* & 0^* & 0^* \\
0^* & 0.01 & 0^* & 0^* & 0.49 & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\
0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \end{pmatrix},
\]

\[
\hat{\Gamma} = \begin{pmatrix}
-0.04 & 0.01 \end{pmatrix}, \quad \hat{\Psi} = \begin{pmatrix}
0.10 \\
0^* & 0.00 \\
0^* & 0^* & 0.01 \\
0^* & 0.00 \\
0^* & 0.00 \\
0^* & 0.00 \\
0.00 & 0^* \end{pmatrix}, \quad \hat{\Psi} = \begin{pmatrix}
0.00 & 0^* & 0^* & 0.00 \\
0^* & 0.00 & 0^* & 0^* & 0.00 \\
0^* & 0.00 & 0^* & 0^* & 0.00 \\
0.00 & 0^* & 0^* & 0^* & 0^* & 0^* & 0.00 & 0.00 \\
0.00 & 0^* & 0^* & 0^* & 0.00 & 0^* & 0.00 & 0.00 \end{pmatrix}
\]

Note that, as expected, the estimates of \( \mu_x \) and \( \Phi \) are the same as the corresponding elements of computed sample mean vector and covariance matrix for the vector variable \( X \).

For instance in first scenario, Model 1 can be explained in the matrix form as:

Let \( y_1, y_2, \ldots, y_7 \) and \( x_1 \) and \( x_2 \) as:

44
$y_1 = \text{glucosylceramide}, y_2 = \text{methylation 2}, y_3 = \text{methylation 3}, y_4 = \text{methylation 4}, y_5 = \text{methylation 6}, y_6 = \text{methylation 8}, y_7 = \text{methylation 9}$ and $x_1 = \text{sex}, x_2 = \text{Age}$.

\[
\begin{pmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  y_4 \\
  y_5 \\
  y_6 \\
  y_7 \\
  \zeta_1 \\
  \zeta_2 \\
  \zeta_3 \\
  \zeta_4 \\
  \zeta_5 \\
  \zeta_6 \\
  \zeta_7
\end{pmatrix}
= \begin{pmatrix}
  \alpha_1 \\
  \alpha_2 \\
  \alpha_3 \\
  \alpha_4 \\
  \alpha_5 \\
  \alpha_6 \\
  \alpha_7
\end{pmatrix}
+ \begin{pmatrix}
  0 & \beta_{12} & \beta_{13} & \beta_{14} & \beta_{15} & \beta_{16} & \beta_{17} \\
  0 & 0 & 0 & \beta_{24} & 0 & 0 & 0 \\
  \beta_{31} & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & \beta_{75} & 0 & 0
\end{pmatrix}
\begin{pmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  y_4 \\
  y_5 \\
  y_6 \\
  y_7 \\
  \gamma_{11} \\
  \gamma_{12} \\
  \gamma_{22} \\
  \gamma_{32} \\
  \gamma_{42} \\
  \gamma_{52} \\
  \gamma_{62} \\
  \gamma_{71} \\
  0
\end{pmatrix}
+ \begin{pmatrix}
  \gamma_{11} \\
  \gamma_{12} \\
  \gamma_{22} \\
  \gamma_{32} \\
  \gamma_{42} \\
  \gamma_{52} \\
  \gamma_{62} \\
  \gamma_{71} \\
  0
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  x_2
\end{pmatrix}
\]