Ph.D. Dissertation of Natural Sciences

# Inference of mixed graphical models in 2 groups with Markov random field model and its application

## 마르코프 랜덤 필드 모형을 이용한 2개 집단의 혼합 그래프 모형 추정 및 적용

August 2022

Graduate School of Natural Sciences
Seoul National University
Bioinformatics Major

박 재 현 / Jaehyun Park

# Inference of mixed graphical models in 2 groups with Markov random field model and its application

원 성 호 / Sungho Won

Submitting a Ph.D. Dissertation of
Natural Sciences
July 2022

Graduate School of Natural Sciences
Seoul National University
Bioinformatics Major

박 재 현 / Jaehyun Park

Confirming the Ph.D. Dissertation written by
Jaehyun Park
August 2022

|  |  |  |
|---|---|---|
| Chair | 박 태 성 | (Seal) |
| Vice Chair | 원 성 호 | (Seal) |
| Examiner | 유 연 주 | (Seal) |
| Examiner | 이 우 주 | (Seal) |
| Examiner | 김 정 수 | (Seal) |

# Abstract

## Background

Large datasets with a huge number of variables or subjects, such as multi−omics data, have been widely generated recently. Many of these datasets are mixed type including both numeric and categorical variables, which makes their analyses difficult. In some studies, the networks underlying the large dataset may be of interest. There have been several methods that are suggested for the inference of the networks, but most of them can be used only for a single type of data or single class cases.

## Objective

The objective of the study is to develop and propose a new method, named fused MGM (FMGM), that infers network structures underlying mixed data in 2 groups, with assumptions that both the networks and the differences are sparse. Also, statistical analyses including the proposed method were conducted to find biological markers of the atopic dermatitis (AD) and underlying network structures from multi−omics data of 6−month−old infants.

## Methods

For FMGM, the statistical models of the networks are based on pairwise Markov random field model, and the penalty functions implement the main assumption that the networks in 2 groups and their differences are sparse. Fast proximal gradient method (PGM) was used for the optimization of the target function. The extension of FMGM that allows the inclusion of prior knowledges, named prior−induced FMGM (piFMGM), was also developed. The performance of the method was measured with synthetic datasets that simulate power−law network structures. Also, the multi−omics profiles of 6−month−old infants were analyzed. The profiles include host gene transcriptome (N=199), intestinal microbial compositions (N=197), and predicted intestinal microbial functions (N=98; 84 in common).

For the analysis, differential analysis with limma and network inference with FMGM were applied.

## Results

From the analysis of simulated 2-class datasets, generated from simulated scale-free networks, FMGM showed superior performances especially in terms of F1-scores compared to the previous method inferring the networks one by one (0.392 & 0.546). FMGM performed better not only in inferring the differences (0.217 & 0.410), but also in inferring the networks (0.492 & 0.572). Utilizing prior information with piFMGM obtained slightly better F1-scores from the inference of networks (0.572 & 0.589), and from the inference of the difference (0.410 & 0.423). As a result, the overall performance showed slight improvement (0.546 & 0.562). From the inference of networks from 6-month-old infants’ AD data, 10 pairs of variables were shown to have different correlations by disease statuses, including host expression of *LINC01036* and *MIR4788* and abundance of microbial genes related to carotenoid biosynthesis and RNA degradation.

## Conclusions

The proposed method, FMGM inferred the network structures in 2 classes better than the previous method. Inclusion of prior information in piFMGM may be useful in more accurate inference of networks, but since the change was subtle, additional studies may be conducted to improve it. Network inference revealed several markers of AD such as microbial genes related to carotenoid biosynthesis and RNA degradation, suggesting a number of possible underlying metabolisms related to AD such as oxidative stress and microbial RNA balance.

**Keyword** ： network inference, statistical method, Markov random field model, multi-omics, atopic dermatitis (AD)
**Student Number** ： 2016-20461

# Table of Contents

# Chapter 1. Introduction

## 1.1. Study Background

Recently, by the advance of data generation technologies and the decline of the cost, large datasets with massive number of features or observations have been actively generated from various fields. The typical examples include biological 'omics' data that is generated for entire biological aspects, including whole genomics profile of DNA or the quantified measurements of whole proteomics [1], and typically contains thousands or millions of variables.

In many former studies, a single type of omics data was focused and analyzed. For example, genome-wide association studies (GWAS) focused on the association between genetic variants and common diseases [2-4]. These single-omics studies have found important markers for the diseases, but using only one kind of data may fail to detect interactions between different biological aspects, possibly leading to spurious signals or false negatives.

Combining the omics data from multiple sources, referred as 'multi-omics' analyses, can be useful in clarifying more complex and diverse systems under the diseases. Thus, the interest in analyzing multi-omics data has been increasing recently [5, 6]. Large-sized data like multi-omics often includes both numeric and categorical variables, e.g., gene transcription profiles and clinical classifications. This heterogeneous nature of mixed data makes the analysis challenging, and there have been numerous efforts to integrate datasets with multiple variable types or to make statistical inference from them [5].

Scientists are often interested in finding the structure of underlying network from the large data, or identifying all of the direct interactions between variables. Previously proposed methods for this purpose have their own strength and are widely used. Most studies focus on inferring the network structure without considering groups. However, in some cases, inference of the networks in multiple classes with only a small proportion of difference may be desirable. A typical example is biological omics profiles of patients and controls of a specific disease. Biological omics profiles and their networks are expected to remain the same, and certain differences may be responsible for disease status. Gene-gene or protein-protein interactions, such as epistasis [7], can be an instance of the interplay between biological variables that can be related to phenotypes of interest. For example, a previous study reported that *APOE* and *TOMM40*, whose SNP loci were significantly related to the risk of late-onset Alzheimer's disease risk in the Russian population, are correlated at the SNP, gene, and protein levels [8]. In such a scenario, a separate estimation of the networks for each group may be feasible. Moreover, most interplay is expected to be the same between the groups, and the similarity between them needs to be considered.

## 1.2. Prior Works

Graphical models are widely used to model the network structures. The model parameters represent the correlation between each pair of the variables conditioned on the others and can be expressed as edges in the network. Many of the previous methods using graphical models focus on the data with a single variable type or a single group.

### Gaussian Graphical Models

One of the common methods to infer the network, graphical LASSO [9], makes an inference of the precision matrix (the inverse of the variance matrix) from data with Gaussian variables. Elements in the precision matrix can be interpreted as the interaction of two variables adjusting on all of the others, which can be viewed as a direct edge between two nodes in a network. The method assumes the precision matrix would be sparse, or having small proportion of the edges to be non−zero, and uses LASSO−type penalty to the precision matrix.

$$P(\Theta) = \lambda \sum_{i \neq j} \left| \theta_{ij} \right|$$

$\Theta = \{\theta_{ij}\}$ denotes the precision matrix and $\lambda \geq 0$ means the regularization parameter. The larger $\lambda$ is, the sparser the resulting precision matrix becomes. Due to its simplicity and the flexibility of the penalty, it is widely used in the analysis of the network structure. However, it can be applied only to a single Gaussian dataset.

Danaher *et al.* [10] suggested a new method, joint graphical LASSO, that can estimate precision matrices for multiple classes. In addition to the penalties in the original graphical LASSO, joint graphical LASSO also uses convex penalty terms that make some characteristics shared among the networks. Danaher *et al.* presented two kinds of penalty functions: one called the fused graphical LASSO (FGL) [11],

$$P\left(\Theta^{(1)}, \dots, \Theta^{(K)}\right) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} \left| \theta_{ij}^{(k)} \right| + \lambda_2 \sum_{k < k'} \sum_{i,j} \left| \theta_{ij}^{(k)} - \theta_{ij}^{(k')} \right|$$

and another called the group graphical LASSO (GGL) [12],

$$P\left(\Theta^{(1)}, \dots, \Theta^{(K)}\right) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} \left| \theta_{ij}^{(k)} \right| + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{k=1}^{K} {\theta_{ij}^{(k)}}^2}$$

$\Theta^{(k)} = \left\{\theta_{ij}^{(k)}\right\}$ represents the precision matrix in the $k$−th class. GGL induces the structural similarities between the networks, while FGL gives a stronger constraint that leads to the similar values of the elements in the precision matrices. Although the joint graphical LASSO is restricted to the Gaussian data, it successfully extended the network inference method to the problem of multiple classes.

### Mixed Graphical Models

Mixed graphical models (MGMs) that aim to model a mixture of discrete and continuous variables are also considered by several researchers[①]. The main bottleneck of the MGMs is that they require high-dimensional integrals to calculate the likelihoods, and the previous works tried to overcome the problem (1) by separate regression analyses for each variable with all of the others as independent variables, or (2) by using pseudolikelihood functions that are relatively easy to formulate.

The approach using separate regression is intuitive in terms of the interpretation and offers flexibility. Examples of methods using separate regression include performing LASSO [14] or random forests [15] to one variable with all of the others as predictors. The latter approach uses pseudolikelihood functions, which is the product of all of the conditional likelihood functions of each of the variables given all of the others. An example of the approach is using a pairwise Markov random field model [16]. Pseudolikelihoods are known to provide consistent estimators [17, 18].

The previous finding suggests that the likelihood or pseudolikelihood-based methods show better empirical performances compared to the separate regression approaches [16].

[①] The idea of using a graphical model was previously suggested by Lauritzen and Wermuth in 1989 13. Lauritzen, S.L. and N. Wermuth, *Graphical models for associations between variables, some of which are qualitative and some quantitative.* The annals of Statistics, 1989: p. 31-57.. However, this model requires the parameters for continuous variables to be different by all of the possible combinations of discrete variables, which requires huge number of parameters to estimate. Thus, this idea is skipped in this paper.

Moreover, since pseudolikelihood−based methods use parametric models, they are relatively easier to extend to 2−class problems. Therefore, this research decided to extend a pseudolikelihood approach of pairwise Markov random field model for mixed data, suggested by Lee and Hastie [16]. The likelihood of the suggested pairwise Markov random field model simplifies to multivariate Gaussian distribution in continuous−only cases and usual discrete pairwise Markov random field in discrete−only cases. The method uses maximizing pseudolikelihood with penalty terms for edge weights to induce the sparsity of the network.

The drawback of the method is that it applies the same regularization parameters in the penalty function regardless of the types of the edges. Sedgewick *et al.* [19] found that this may cause the method to detect too many continuous−continuous edges while finding insufficient number of edges connected to discrete variables, and suggested giving separate penalization parameters to each of the edge types. This showed better results than using a single parameter in several indices such as recall, F1−score, and Matthew's canonical correlation. Still, this method is limited to single class problems.

## 1.3. Purpose of Research

The main purpose was to develop a method named fused mixed graphical model (FMGM) with mixed data that can be used for detecting the network associated with classes and to extend previous methods so that the new method can be used to detect networks associated with classes. This method extends the framework by Sedgewick *et al.* [19], which applies different regularization amount to each type of edge. It assumes that the difference in networks between classes is sparse and that sparsity is adjusted by incorporating a similar penalty function suggested for the joint graphical LASSO [10].

In chapter 2, a proposed method FMGM is introduced with the likelihood functions of the model and the penalty functions defined and formulated. The target function composed of the pseudolikelihood and the penalties has to be minimized, and fast proximal gradient method (PGM) [20, 21] is utilized for this problem. The details of the algorithm, with the calculation procedures, and the application to the simulated data is also explained in chapter 2. Chapter 3 suggests modifications to include prior information of the networks for the inference. The method induces additional penalty terms that allow different amount of penalization to the edges with prior information. Methods that evaluate the reliability of prior information and that set penalization parameters for edges with priors are also explained. In chapter 4, the multi-omics data from cohort study is analyzed with traditional and proposed methods, including the network inference. This data includes gene transcriptome and gut microbiome information from 6-months-old atopic dermatitis (AD) patients and controls. The discussions regarding limitations, improvements, and future extensions of the framework are suggested in chapter 5.

# Chapter 2. Network Inference of 2-class Mixed Data

## 2.1. Introduction

Several methods have been developed to infer the underlying correlations from large datasets such as omics-scale data in the form of graphs [1]. Parametric and non-parametric methods are used and the former includes precision matrices of Gaussian models, such as graphical LASSO [9] and joint graphical LASSO [10]. Some of these methods, such as GRaFo [15] and GRNBoost [22], utilize non-parametric frameworks, such as random forests [23].

The pairwise Markov random field model is a probabilistic model for graphs. Lee and Hastie [16] suggested a likelihood function for a mixed data pairwise graphical model combining a multivariate Gaussian model and discrete pairwise Markov random field, which can be simplified to either the former or the latter if numeric or categorical data are uniquely present, respectively. The conditional dependencies between variables are parameterized into the edge weights in the likelihood function, and the weight nullity corresponds to the conditional independence and lack of edges in the graph. The main bottleneck of this model is that the likelihood function requires multidimensional integration, which is difficult to calculate in practice. Lee and Hastie [16] overcame this problem by using a product of fully conditioned likelihoods or pseudolikelihood. The conditional likelihoods for the numerical and categorical variables become linear regression and multinomial distributions, respectively.

Sparsity is generally considered in network inference. The main reasons for this include (1) handling a larger number of features than the sample size and (2) shrinking uninformative edges from the resulting model to zero. In the first proposal [16], a penalty function composed of matrix norms with a single penalization parameter is used for sparsity. This method does not reflect the possible heterogeneity between types of variables connected to each edge; therefore, Sedgewick *et al.* [19] extended the framework and used a penalty function that uses different penalization parameters to edge types (continuous−continuous, continuous−discrete, discrete−discrete). These methods can successfully model the mixed−type data parametrically, but they do not consider multi−class cases and model only a single network.

Therefore, in this chapter, a method to infer networks from 2 group data, named fused MGM, will be defined and explained. Section 2.2 explains some notations that are needed to describe data and the model. Section 2.3 derives the target function of the method by following derivations from previous methods, and section 2.4

explains how to optimize the target function with fast proximal gradient method (PGM). The algorithm is implemented in R programming language, as described in section 2.5. Section 2.6 showed the processes and the results of simulated data analyses and compared the results with the previous method. Section 2.7 demonstrates the application of FMGM to a real data from cohort, and the improvements and limitations of the method is discussed in section 2.8.

## 2.2. Notations

It was assumed that there were two different groups and the method was designed to infer the network structures of both groups. $N_m$ independent observations were assumed in group $m$ and each observation was indexed with $n_m$. The total number of observations was $N = N_1 + N_2$. $p$ numerical variables and $q$ categorical variables were assued, which were used to infer the networks. The numerical and categorical variable vectors from the observation $n_m$ were denoted respectively as $x^{(n_m)}$ and $y^{(n_m)}$, and the $s-$th numerical and $r-$th categorical variables were denoted as $x_s^{(n_m)}$ and $y_r^{(n_m)}$, respectively. The level number for $y_r^{(n_m)}$ was denoted with $L_r$.

## 2.3. Model Formulation

Pairwise Markov Random Field Model and Its Pseudolikelihood

Lee and Hastie [16] proposed a pairwise Markov random field that was used to model the likelihood function as follows:

$$
\begin{aligned}
&\log p\left(x^{(n_m)}, y^{(n_m)}; \Theta^{(m)}\right) \\
&\qquad = \sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}^{\left(\Theta^{(m)}\right)} x_s^{(n_m)} x_t^{(n_m)} + \sum_{s=1}^{p} \alpha_s^{\left(\Theta^{(m)}\right)} x_s^{(n_m)} \\
&\qquad + \sum_{s=1}^{p}\sum_{j=1}^{q} \rho_{sj}^{\left(\Theta^{(m)}\right)}\left(y_j^{(n_m)}\right) x_s^{(n_m)} + \sum_{j=1}^{q}\sum_{r=1}^{q} \phi_{rj}^{\left(\Theta^{(m)}\right)}\left(y_r^{(n_m)}, y_j^{(n_m)}\right) \\
&\qquad - \log Z\left(\Theta^{(m)}\right)
\end{aligned}
$$

where $\Theta^{(m)} = \left(\alpha_s^{\left(\Theta^{(m)}\right)}, \beta_{st}^{\left(\Theta^{(m)}\right)}, \rho_{sj}^{\left(\Theta^{(m)}\right)}, \mathrm{vec}\left(\phi_{rj}^{\left(\Theta^{(m)}\right)}\right)'\right)'$ denotes a vector of parameters that models the network in group $m$, and $\mathbf{vec}$ is the vec operator. $\beta_{st}^{\left(\Theta^{(m)}\right)}$ is the edge weight between the $s-$th and the $t-$th numerical variable nodes, $\alpha_s^{\left(\Theta^{(m)}\right)}$ corresponds to the node potential of the $s-$th numerical variable, $\rho_{sj}^{\left(\Theta^{(m)}\right)}$ denotes the correlation between the $s-$th numerical and the $j-$th categorical variables, and $\phi_{rj}^{\left(\Theta^{(m)}\right)}$ denotes the edge weights between the $r-$th and the $j-$th categorical variable nodes. $\rho_{sj}^{\left(\Theta^{(m)}\right)}$ is an $L_j$ dimensional vector of length with the $l-$th element $\rho_{sj}^{\left(\Theta^{(m)}\right)}(l)$, and indicates the correlation between variables $s$ and $j$. If all elements are zero, they are conditionally independent. Likewise, $\phi_{rj}^{\left(\Theta^{(m)}\right)}$ is an $L_r \times L_j$ matrix with the $(l,k)-$th element $\phi_{rj}^{\left(\Theta^{(m)}\right)}(l,k)$, and variables $r$ and $j$ are conditionally independent if all of the elements are 0.

Directly maximizing the likelihood function is difficult because the calculation of the partition function $Z\left(\Theta^{(m)}\right)$ requires multidimensional integration. Alternatively, the likelihood could be replaced with a pseudolikelihood function, the conditional distribution function product for each variable [17, 18]. Minimizing the pseudolikelihood function is computationally efficient and provides a consistent estimator. The negative log of the pseudolikelihood equation was formulated as follows:

$$f_m\left(\Theta^{(m)}\right) = -\frac{1}{N}\sum_{n_m}\left(\sum_{s=1}^{p}\log p\left(x_s^{(n_m)}\middle|x_{\setminus s}^{(n_m)}, y^{(n_m)}; \Theta^{(m)}\right) + \sum_{r=1}^{q}\log p\left(y_r^{(n_m)}\ \middle|x^{(n_m)}, y_{\setminus r}^{(n_m)}; \Theta^{(m)}\right)\right)$$

$$f(\Theta) = \sum_{m=1}^{2} f_m\left(\Theta^{(m)}\right)$$

The function was divided by the number of observations to make the pseudolikelihood log scale and the penalty function equivalent. For continuous variables, their conditional distribution was assumed to follow a Gaussian distribution, with the mean in the form of linear regression model and the variance is equal to the $\beta_{ss}^{(\Theta^{(m)})}$ reciprocal.

$$p\left(x_s^{(n_m)}\middle|x_{\setminus s}^{(n_m)}, y_r^{(n_m)}, \Theta^{(m)}\right) = \sqrt{\frac{\beta_{ss}^{(\Theta^{(m)})}}{2\pi}}\exp\left(-\frac{\beta_{ss}^{(\Theta^{(m)})}}{2}\left(a_s^{(m)} - x_s^{(n_m)}\right)^2\right)$$

$$a_s^{(m)} = \frac{\alpha_s^{(\Theta^{(m)})} + \sum_j \rho_{sj}^{(\Theta^{(m)})}\left(y_j^{(n_m)}\right) - \sum_{t\neq s}\beta_{st}^{(\Theta^{(m)})} x_t^{(n_m)}}{\beta_{ss}^{(\Theta^{(m)})}}$$

For categorical variables, a multinomial distribution with $L_r$ levels was considered for conditional distribution.

$$p\left(y_r^{(n_m)}\ \middle|x^{(n_m)}, y_{\setminus r}^{(n_m)}; \Theta^{(\Theta^{(m)})}\right) = p_r^{(m)}(l) = \frac{\exp\left(b_r^{(m)}\left(y_r^{(n_m)}\right)\right)}{\sum_{l=1}^{L_r}\exp\left(b_r^{(m)}(l)\right)}$$

$$b_r^{(m)}\left(y_r^{(n_m)}\right) = \sum_s \rho_{sr}^{(\Theta^{(m)})}\left(y_r^{(n_m)}\right)x_s + \phi_{rr}^{(\Theta^{(m)})}\left(y_r^{(n_m)}, y_r^{(n_m)}\right) + \sum_{j\neq r}\phi_{rj}^{(\Theta^{(m)})}\left(y_r^{(n_m)}, y_j^{(n_m)}\right)$$

Lee and Hastie had shown that the negative pseudolikelihood log is jointly convex if $\beta_{ss}^{(\Theta^{(m)})} > \mathbf{0}$ [16].

### Penalty Functions for Sparse Networks

Penalty functions based on vector or matrix norms can be used to induce a network sparsity. Lee and Hastie added the following terms to the penalty:

$$g(\Theta^{(m)}) = \lambda\left(\sum_{t<s}\left|\beta_{st}^{(\Theta^{(m)})}\right| + \sum_{s,j}\left\|\rho_{sj}^{(\Theta^{(m)})}\right\|_2 + \sum_{r<j}\left\|\phi_{rj}^{(\Theta^{(m)})}\right\|_F\right)$$

$\lambda$ is the regularization parameter, $\|\cdot\|_2$ indicates the $l_2$−norm of a vector, and $\|\cdot\|_F$ means the Frobenius norm of a matrix. $l_2$− and Frobenius norms were used instead of $l_1$ −norms because the corresponding variables are mutually independent if and only if all of the elements in the vector or the matrix are 0.

Sedgewick *et al.* [19] modified the penalty by granting different regularization parameters for different edge types.

$$g(\Theta^{(m)}) = \lambda_{cc}\sum_{t<s}\left|\beta_{st}^{(\Theta^{(m)})}\right| + \lambda_{cd}\sum_{s,j}\left\|\rho_{sj}^{(\Theta^{(m)})}\right\|_2 + \lambda_{dd}\sum_{r<j}\left\|\phi_{rj}^{(\Theta^{(m)})}\right\|_F$$

These penalties induce sparsity to separate networks. To induce the sparsity to the difference, the penalty terms regarding the differences in the network edges were added. Consequently, the subdifferentiable penalty function was set as follows:

$$\begin{aligned} g(\Theta) = {} & \lambda_{cc}\sum_m\sum_{t<s}\left|\beta_{st}^{(\Theta^{(m)})}\right| + \lambda_{cd}\sum_m\sum_{s,j}\left\|\rho_{sj}^{(\Theta^{(m)})}\right\|_2 + \lambda_{dd}\sum_m\sum_{r<j}\left\|\phi_{rj}^{(\Theta^{(m)})}\right\|_F \\ & + \lambda'_{cc}\sum_{t<s}\left|\beta_{st}^{(\Theta^{(1)})} - \beta_{st}^{(\Theta^{(2)})}\right| + \lambda'_{cd}\sum_{s,j}\left\|\rho_{sj}^{(\Theta^{(1)})} - \rho_{sj}^{(\Theta^{(2)})}\right\|_2 \\ & + \lambda'_{dd}\sum_{r<j}\left\|\phi_{rj}^{(\Theta^{(1)})} - \phi_{rj}^{(\Theta^{(2)})}\right\|_F \end{aligned}$$

The final form of the target function to minimize is as follows:

$$
\begin{aligned}
\bar{\bar{l}}_{\lambda}(\Theta|x,y) &= f(\Theta)+g(\Theta) \\
&= \frac{1}{N}\sum_{m}\left(-\sum_{n_m}\left(\sum_{s=1}^{p}\log p\left(x_s^{(n_m)}\middle|x_{\backslash s}^{(n_m)},y^{(n_m)};\Theta^{(m)}\right)\right.\right. \\
&\qquad\left.\left.+\sum_{r=1}^{q}\log p\left(y_r^{(n_m)}\middle|x^{(n_m)},y_{\backslash r}^{(n_m)};\Theta^{(m)}\right)\right)\right) \\
&+\lambda_{cc}\sum_{m}\sum_{t<s}\left|\beta_{st}^{(\Theta^{(m)})}\right|+\lambda_{cd}\sum_{m}\sum_{s,j}\left\|\rho_{sj}^{(\Theta^{(m)})}\right\|_2+\lambda_{dd}\sum_{m}\sum_{r<j}\left\|\phi_{rj}^{(\Theta^{(m)})}\right\|_F \\
&+\lambda'_{cc}\sum_{t<s}\left|\beta_{st}^{(\Theta^{(1)})}-\beta_{st}^{(\Theta^{(2)})}\right|+\lambda'_{cd}\sum_{s,j}\left\|\rho_{sj}^{(\Theta^{(1)})}-\rho_{sj}^{(\Theta^{(2)})}\right\|_2 \\
&+\lambda'_{dd}\sum_{r<j}\left\|\phi_{rj}^{(\Theta^{(1)})}-\phi_{rj}^{(\Theta^{(2)})}\right\|_F
\end{aligned}
$$

## 2.4. Optimization with Fast Proximal Gradient Method

To minimize the target function, fast proximal gradient method (PGM) was used [20, 21]. As suggested by Beck and Teboulle [20], the backtracking step size-rule and monotonicity was implemented. For the details of the method, please refer to the original literature.

First, below terms are defined:

$$
\begin{aligned}
F(\Theta) &\coloneqq f(\Theta)+g(\Theta) \\
Q_L(\Theta,\Lambda) &\coloneqq f(\Lambda)+\langle\Theta-\Lambda,\nabla f(\Lambda)\rangle+\frac{L}{2}\|\Theta-\Lambda\|^2+g(\Theta) \\
p_L(\Lambda) &\coloneqq \arg\min_{\Theta}\{Q_L(\Theta,\Lambda)\} \\
&= \arg\min_{\Theta}\left\{g(\Theta)+\frac{L}{2}\left\|\Theta-\left(\Lambda-\frac{1}{L}\nabla f(\Lambda)\right)\right\|^2\right\}
\end{aligned}
$$

where $\nabla f(\Lambda)$ denotes the first partial derivative of $f$ at the point $\Lambda$. The updated parameter values in the $k$-th step were denoted as $\Theta_{(k)}$, and the initial value was defined as $\Theta_{(0)}$.

$L$ is the estimate of Lipschitz constant of $f$ and its update at the $k$-th step was denoted by $L_{(k)}$. The $L_{(k)}$ update was performed by backtracking, where the smallest non-negative integer $i_{(k)}$ that satisfies the following equation was sought:

$$
\begin{aligned}
&\bar{L} = \eta^{i_{(k)}} L_{(k-1)} \\
&F\left(p_{\bar{L}}(\Lambda_{(k)})\right) \leq Q_L\left(p_{\bar{L}}(\Lambda_{(k)}), \Lambda_{(k)}\right)
\end{aligned}
$$

Here, $\eta > 1$ is a multiplier, and $L_{(k)}$ is set as $L_{(k)} = \eta^{i_{(k)}} L_{(k-1)}$. Then, the parameters were updated as follows:

$$
\begin{aligned}
K_{(k)} &= p_{L_{(k)}}(\Lambda_{(k)}) \\
t_{(k+1)} &= \frac{1 + \sqrt{1 + 4t_{(k)}^2}}{2} \\
\Theta_{(k)} &= \arg\min\{F(\Theta) : \Theta = K_{(k)}, \Theta_{(k-1)}\} \\
\Lambda_{(k+1)} &= \Theta_{(k)} + \left(\frac{t_{(k)}}{t_{(k+1)}}\right)(K_{(k)} - \Theta_{(k)}) + \left(\frac{t_{(k)} - 1}{t_{(k+1)}}\right)(\Theta_{(k)} - \Theta_{(k-1)})
\end{aligned}
$$

where $t_{(k)}$ determines the weights by which previous estimates are reflected.

The iteration stopped if $\Theta_{(k)}$ converged. The convergence could be determined using (1) the rooted mean squared deviation of all of the parameter values, or (2) the difference in the target function values. Because the algorithm forces $\Theta_{(k)}$ not to change in some steps to guarantee monotonicity, the number of iterations in a row that satisfied the convergence conditions was set to be larger than 1 (default: 3).

### Calculating the First Derivative of the Pseudolikelihood

Proximal gradient method included the calculation of $p_L(\Theta)$. If $\Lambda^* = \Lambda - \frac{1}{\mathrm{L}} \nabla f(\Lambda)$ was denoted, the calculation of $p_L(\Theta)$ was equivalent to minimizing $g(\Theta) + \frac{L}{2} \|\Theta - \Lambda^*\|^2$. The calculation of $\Lambda^*$ required the first partial derivatives of $f$, which have closed forms as follows:

$$\frac{\partial f_m\left(\Theta^{(m)}\right)}{\partial \alpha^{\left(\Theta^{(m)}\right)}} = \frac{1}{N}\sum_{n_m}\left(a_s^{(m)} - x_s^{(n_m)}\right)$$

$$\frac{\partial f_m\left(\Theta^{(m)}\right)}{\partial \beta_{ss}^{\left(\Theta^{(m)}\right)}}$$

$$= -\frac{N_m}{2N\beta_{ss}^{\left(\Theta^{(m)}\right)}} - \frac{1}{2}\sum_{n_m}\left(a_s^{(m)} - x_s^{(n_m)}\right)^2 - \sum_{n_m} x_s^{(n_m)}\left(a_s^{(m)} - x_s^{(n_m)}\right)$$

$$\frac{\partial f_m\left(\Theta^{(m)}\right)}{\partial \beta_{st}^{\left(\Theta^{(m)}\right)}} = -\frac{1}{N}\left(\sum_{n_m} x_t^{(n_m)}\left(a_s^{(m)} - x_s^{(n_m)}\right) + \sum_{n_m} x_s^{(n_m)}\left(a_t^{(m)} - x_t^{(n_m)}\right)\right)$$

$$\frac{\partial f_m\left(\Theta^{(m)}\right)}{\partial \rho_{sj}^{\left(\Theta^{(m)}\right)}}(l)_{l=1,\ldots,L_j}$$

$$= \frac{1}{N}\sum_{n_m}\left(I\left(l = y_j^{(n_m)}\right)\left(a_s^{(m)} - x_s^{(n_m)}\right) + \left(p_j^{(m)}(l) - I\left(l = y_j^{(n_m)}\right)\right)x_s^{(n_m)}\right)$$

$$\frac{\partial f_m\left(\Theta^{(m)}\right)}{\partial \phi_{rr}^{\left(\Theta^{(m)}\right)}}(l,l)_{l=1,\ldots,L_r} = \frac{1}{N}\sum_{n_m}\left(p_r^{(m)}(l) - I\left(l = y_r^{(n_m)}\right)\right)$$

$$\frac{\partial f_m\left(\Theta^{(m)}\right)}{\partial \phi_{rj}^{\left(\Theta^{(m)}\right)}}(l,k)_{\substack{l=1,\ldots,L_r\\ k=1,\ldots,L_j}}$$

$$= \frac{1}{N}\sum_{n_g}\left(p_r^{(m)}(l) - I\left(l = y_r^{(n_m)}\right)\right)I\left(k = y_j^{(n_m)}\right)$$

$$+ \sum_{n_m} I\left(l = y_r^{(n_m)}\right)\left(p_j^{(m)}(k) - I\left(k = y_j^{(n_m)}\right)\right)$$

## Fixed Point Approach: Non−diagonal Cases

After calculating $\Lambda^*$, the function to minimize in group $m$ was written as follows:

$$\frac{L}{2}\left(\beta_{st}^{\left(\Theta^{(m)}\right)} - \beta_{st}^{\left(\Lambda^{*(m)}\right)}\right)^2 + \lambda_{cc}\left|\beta_{st}^{\left(\Theta^{(m)}\right)}\right| + \lambda'_{cc}\left|\beta_{st}^{\left(\Theta^{(1)}\right)} - \beta_{st}^{\left(\Theta^{(2)}\right)}\right|$$

$$\frac{L}{2}\sum_{l=1}^{L_j}\left(\rho_{sj}^{\left(\Theta^{(m)}\right)}(l) - \rho_{sj}^{\left(\Lambda^{*(m)}\right)}(l)\right)^2 + \lambda_{cd}\left\|\rho_{sj}^{\left(\Theta^{(m)}\right)}\right\|_2 + \lambda'_{cd}\left\|\rho_{sj}^{\left(\Theta^{(1)}\right)} - \rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2$$

$$\frac{L}{2}\sum_{l,k}\left(\phi_{rs}^{\left(\Theta^{(m)}\right)}(l,k)-\phi_{rs}^{\left(\Lambda^{*(m)}\right)}(l,k)\right)^2+\lambda_{dd}\left\|\phi_{rs}^{\left(\Theta^{(m)}\right)}\right\|_F+\lambda'_{dd}\left\|\phi_{rs}^{\left(\Theta^{(1)}\right)}-\phi_{rs}^{\left(\Theta^{(2)}\right)}\right\|_F$$

Because the functions included the sum of norm functions, the simultaneous minimization of all parameters might be challenging. Instead, the minimization step concerning the parameters in one group was performed first, followed by minimization with the other group. Minimization according to the group was iterated until the parameters converged.

For the edges between numeric variables, it was relatively easy to solve the minimization problem because the function had the form of a quadratic term and absolute value terms. However, the edges that included categorical variables were too complicated to directly determine the minimization point. To solve the problem, another method similar to Weiszfeld's approach for the Fermat-Weber location problem was used [24, 25]. Because a similar approach could be applied to $\phi_{rs}^{\left(\Theta^{(m)}\right)}$, the following statements focused on $\rho_{sj}^{\left(\Theta^{(m)}\right)}$, the weights of the edges between continuous-discrete variables.

First, a set of non-smooth points $\mathrm{B}\left(\rho_{sj}^{\left(\Theta^{(m)}\right)}\right):=\{0\}\cup\left\{\rho_{sj}^{\left(\Theta^{(m')}\right)};m'\neq m\right\}$ was defined. The function values at each point in $\mathrm{B}\left(\rho_{sj}^{\left(\Theta^{(m)}\right)}\right)$ were calculated, and the point with the smallest value was denoted as $\rho_{sj}^{(\min)}$. Second, $\rho_{sj}^{(\min)}$ was determined judged if it satisfied the following inequation:

$$\lambda'_{cd}\geq\left\|L\left(\rho_{sj}^{\left(\Lambda^{*(m)}\right)}-\rho_{sj}^{(\min)}\right)+\sum_{\rho_{sj}^{*}\in\mathrm{B}\left(\rho_{sj}^{\left(\Theta^{(m)}\right)}\right)\backslash\rho_{sj}^{(\min)}}\lambda'_{cd}\frac{\rho_{sj}^{*}-\rho_{sj}^{(\min)}}{\left\|\rho_{sj}^{*}-\rho_{sj}^{(\min)}\right\|_2}\right\|_2$$

On the left-handed side, $\lambda'_{cd}$ was replaced with $\lambda_{cd}$ if $\rho_{sj}^{(\min)}=0$.

Similarly, on the right-handed side, $\lambda'_{cd}$ was replaced with $\lambda_{cd}$ when $\rho^{*}_{sj} = \mathbf{0}$. If the inequality held, $\rho_{sj}^{(\mathrm{min})}$ was considered the minimization point. This was a modification of the criterion proposed by Katz and Vogl [26]. The derivation of the criterion is described in Appendix.

If the inequality was not satisfied, the fixed-point approach was used. Since the target function is convex, the minimization point satisfied the condition that the first derivative was equal to zero at the point:

$$L\left(\rho_{sj}^{\left(\Theta^{(m)}\right)} - \rho_{sj}^{\left(\Lambda^{*(m)}\right)}\right) + \lambda_{cd}\frac{\rho_{sj}^{\left(\Theta^{(m)}\right)}}{\left\|\rho_{sj}^{\left(\Theta^{(m)}\right)}\right\|_2} + \lambda'_{cd}\frac{\rho_{sj}^{\left(\Theta^{(m)}\right)} - \rho_{sj}^{\left(\Theta^{(m')}\right)}}{\left\|\rho_{sj}^{\left(\Theta^{(1)}\right)} - \rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2} = \mathbf{0}$$

where $m' = 2$ if $m = 1$ and vice versa. This condition was equivalent to the following statement:

$$\rho_{sj}^{\left(\Theta^{(m)}\right)}\left(L + \frac{\lambda_{cd}}{\left\|\rho_{sj}^{\left(\Theta^{(m)}\right)}\right\|_2} + \frac{\lambda'_{cd}}{\left\|\rho_{sj}^{\left(\Theta^{(1)}\right)} - \rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}\right) = L\rho_{sj}^{\left(\Lambda^{*(m)}\right)} + \frac{\lambda'_{cd}\rho_{sj}^{\left(\Theta^{(m')}\right)}}{\left\|\rho_{sj}^{\left(\Theta^{(1)}\right)} - \rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}$$

$$\rho_{sj}^{\left(\Theta^{(m)}\right)} = \frac{L\rho_{sj}^{\left(\Lambda^{*(m)}\right)} + \dfrac{\lambda'_{cd}\rho_{sj}^{\left(\Theta^{(m')}\right)}}{\left\|\rho_{sj}^{\left(\Theta^{(1)}\right)} - \rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}}{L + \dfrac{\lambda_{cd}}{\left\|\rho_{sj}^{\left(\Theta^{(m)}\right)}\right\|_2} + \dfrac{\lambda'_{cd}}{\left\|\rho_{sj}^{\left(\Theta^{1}\right)} - \rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}}$$

If the values of $\rho_{sj}^{\left(\Theta^{(m)}\right)}$ at step $p = 1,2,\dots$ were set as $\rho_{sj}^{[p]}$, the $p$-th step of the fixed-point approach was as follows:

$$\rho_{sj}^{[p+1]} = \frac{L\rho_{sj}^{\left(\Lambda^{*(m)}\right)} + \frac{\lambda'_{cd}\rho_{sj}^{\left(\Theta^{(m')}\right)}}{\left\|\rho_{sj}^{[p]} - \rho_{sj}^{\left(\Theta^{(m')}\right)}\right\|_2}}{L + \frac{\lambda_{cd}}{\left\|\rho_{sj}^{[p]}\right\|_2} + \frac{\lambda'_{cd}}{\left\|\rho_{sj}^{[p]} - \rho_{sj}^{\left(\Theta^{(m')}\right)}\right\|_2}}$$

The iteration continued until $\rho_{sj}^{[p]}$ converged. In the implementation, the initial value was set as the weighted sum of the points in $\mathrm{B}\left(\rho_{sj}^{\left(\Theta^{(m)}\right)}\right)$ as follows:

$$\rho_{sj}^{[1]} = \frac{\sum_{m' \neq m} \lambda'_{cd}\rho_{sj}^{\left(\Theta^{(m')}\right)}}{\lambda_{cd} + \sum_{m' \neq m} \lambda'_{cd}}$$

### Fixed Point Approach: Diagonal Cases

The main drawback of the approach that minimized the function according to the parameters in each of the group was that the iteration could stop on the non-zero diagonal point, or $\beta_{st}^{\left(\Theta^{(1)}\right)} = \beta_{st}^{\left(\Theta^{(2)}\right)} \neq 0$, even if the point was not a true minimization point. For example, if the values in function were set as $\beta_{st}^{\left(\Lambda^{*(1)}\right)} = \beta_{st}^{\left(\Lambda^{*(2)}\right)} = 0.1$ in $L = 5$, and $\lambda_{cc} = \lambda'_{cc} = 0.8$: the iteration stopped at $\beta_{st}^{\left(\Theta^{(1)}\right)} = \beta_{st}^{\left(\Theta^{(2)}\right)} = 0.1$ although the true minimization point was $\beta_{st}^{\left(\Theta^{(1)}\right)} = \beta_{st}^{\left(\Theta^{(2)}\right)} = 0$. Therefore, an additional minimization with $\beta_{st}^{\left(\Theta^{(1)}\right)} = \beta_{st}^{\left(\Theta^{(2)}\right)}$ fixation was performed if the iteration was halted at the diagonal point. The statements below focus on $\rho_{st}^{\left(\Theta^{(m)}\right)}$, the edges between continuous variables. However, the similar approaches could be applied to $\rho_{sj}^{\left(\Theta^{(m)}\right)}$ and $\phi_{rs}^{\left(\Theta^{(m)}\right)}$.

If $\beta_{st}^{\left(\Theta^{(1)}\right)} = \beta_{st}^{\left(\Theta^{(2)}\right)}$, the target function was reduced to the following:

$$\frac{L}{2}\left(\left(\beta_{st}^{(\Theta^{(1)})}-\beta_{st}^{\left(\Lambda^{*(1)}\right)}\right)^2+\left(\beta_{st}^{(\Theta^{(1)})}-\beta_{st}^{\left(\Lambda^{*(2)}\right)}\right)^2\right)+2\lambda_{cc}\left|\beta_{st}^{(\Theta^{(1)})}\right|$$

This function has a minimization point of zero if the following inequality was satisfied [26]:

$$2\lambda_{cc}\geq L\left|\beta_{st}^{\left(\Lambda^{*(1)}\right)}+\beta_{st}^{\left(\Lambda^{*(2)}\right)}\right|$$

The proof of the criterion is in Appendix. Otherwise, a fixed−point approach was used. The first derivative of the reduced function was as follows:

$$L\left(\left(\beta_{st}^{(\Theta^{(1)})}-\beta_{st}^{\left(\Lambda^{*(1)}\right)}\right)+\left(\beta_{st}^{(\Theta^{(1)})}-\beta_{st}^{\left(\Lambda^{*(2)}\right)}\right)\right)+2\lambda_{cc}\frac{\beta_{st}^{(\Theta^{(1)})}}{\left|\beta_{st}^{(\Theta^{(1)})}\right|}=0$$

The above equation was equivalent to

$$\beta_{st}^{(\Theta^{(1)})}=\frac{L\dfrac{\beta_{st}^{\left(\Lambda^{*(1)}\right)}+\beta_{st}^{\left(\Lambda^{*(2)}\right)}}{2}}{L+\dfrac{\lambda_{cc}}{\left|\beta_{st}^{(\Theta^{(1)})}\right|}}$$

Thus, the $p$−th step was as follows:

$$\beta_{st}^{[p+1]}=\frac{L\dfrac{\beta_{st}^{\left(\Lambda^{*(1)}\right)}+\beta_{st}^{\left(\Lambda^{*(2)}\right)}}{2}}{L+\dfrac{\lambda_{cc}}{\left|\beta_{st}^{[p]}\right|}}$$

## Determination of Regularization Parameters with StEPS

For the determination of the penalization parameters, we used stable edge-specific penalty selection (StEPS) [19], a modification of stability approach to regularization selection (StARS) [27]. From a dataset of $n$ samples, $N$ subsamples of size $b$ were drawn without replacement. The model was fitted for each subsample with a single penalization parameter $\lambda$. For the edge between variables $s$ and $t$, the fraction of subsamples that contained non-zero edge, $\hat{\theta}_{st}(\lambda)$, was calculated. Edge instability, the empirical probability of disagreement of having a non-zero edge at each $\lambda$ value, was defined as follows:

$$\hat{\xi}_{st}(\lambda) = 2\hat{\theta}_{st}(\lambda)\left(1 - \hat{\theta}_{st}(\lambda)\right)$$

The total instability for each type was calculated as follows:

$$\hat{D}_{cc}(\lambda) = \frac{\sum_{cc} \hat{\xi}_{st}(\lambda)}{\binom{p}{2}}$$
$$\hat{D}_{cd}(\lambda) = \frac{\sum_{cd} \hat{\xi}_{st}(\lambda)}{pq}$$
$$\hat{D}_{dd}(\lambda) = \frac{\sum_{dd} \hat{\xi}_{st}(\lambda)}{\binom{q}{2}}$$

The calculated instabilities were monotonized to prevent selecting dense graphs with low instability as follows:

$$\overline{D}_{cc}(\lambda) = \sup_{\lambda \le t} \hat{D}_{cc}(t)$$
$$\overline{D}_{cd}(\lambda) = \sup_{\lambda \le t} \hat{D}_{cd}(t)$$
$$\overline{D}_{dd}(\lambda) = \sup_{\lambda \le t} \hat{D}_{dd}(t)$$

From the largest $\lambda$ value, the value was reduced until the threshold was reached where the threshold was defined a priori.

## 2.5. Code Implementation

The numerical optimization procedures were implemented using the R programming language (https://www.R-project.org/). The number of edges to be optimized was $O(p^2+q^2)$; therefore, the computational time would be very long if each edge was estimated individually. Therefore, the code uses 'bigmemory', 'bigalgebra', and 'biganalytics' packages [28, 29] so that the optimization could be parallelized according to edges.

The initial value of the estimate of Lipschitz constant $L_{(0)}$ and multiplier $\eta$ was set to one and two, respectively. Furthermore, to accelerate the procedures, a small positive number $\alpha < 1$ (default: 0.9) was multiplied by the estimate in each step before the backtracking step.

The number and size of the subsamples in StEPS were determined beforehand. The default values in the implementation were $N = 20$ and $\mathrm{b} = 10\sqrt{n}$ as suggested by Liu *et al* [27].

## 2.6. Simulated Data Analysis

### Data Generation

The simulation data were generated using previously published methods [10, 19]. One hundred variables, 50 of which were normally distributed numeric variables and the others were categorical variables with 4 levels each, were randomly divided into five equal-sized variable sets. Scale-free networks were generated for each of the variable sets based on the method used by Bollobás *et al.* [30] Briefly, the generation started with only 2 variables connected, and the edges were iteratively added until all of the 20 variables in the set were connected. In each step, two non-zero-degree nodes were connected by a probability of 0.3, or a node with 0 degree was connected to a non-zero-degree node by probability of 0.7. Non-zero-degree nodes were selected by their probabilities proportional to their degrees. The directionality and duplicated edges were ignored in the resulting networks. An example of a generated network is shown in Figure 1.

The networks from two classes were simulated. One of the five networks was randomly selected and removed from the first class and the other was selected again and removed from the second class. Each class has four distinct networks, three of them overlapping. For each class, 250 observations were generated using Markov chain Monte-Carlo method. The weight of each edge, $w_{st}$, was randomly sampled from a uniform distribution ranging from 0.5 to 0.8. For each of the $\beta_{st}$, the value was set to $w_{st}$ and the sign was randomly sampled with even probability. To ensure that the matrix was positive definite, the diagonal elements were set to be the largest value among the sums of the absolute values of edge weights connected to each of the nodes. For $\rho_{st}$, the values were permuted from $[-w_{st}, -.5w_{st}, .5w_{st}, w_{st}]$. For $\phi_{st}$, one parameter in each column and each row was set to be $w_{st}$ and the rest was set to be $-w_{st}$. The generation of simulated networks and datasets were repeated for 50 times.

## Data Analysis

The suggested method, fused MGM (FMGM), was run for 50 datasets that were repeatedly simulated using the procedures described above. For regularization parameters, seven values ranging from 0.08 to 0.32, which were equally spaced on a log2-scale, were tested using StEPS with the default parameters. The smallest values with average estimation instabilities below the threshold for each edge type were used.

Additionally, we used a previous method by Sedgewick *et al.* [19] for each class. The causalMGM package is currently unavailable in

R[②], but because FMGM is reduced to causalMGM in one-class cases, the same code implemented in R was used. The same values as the suggested method were used in StEPS, with the same criteria used to determine the penalization parameters.

## Results

An inference result overview is presented in Figure 2. The accuracy (the sum of true positives and true negatives divided by the total number of edges) was fairly high (0.988 in overall) for all cases, as expected by the low true edge number and the nature of the network estimation methods (Table 1). The average F1 score and Matthew's correlation coefficient was 0.546 and 0.548, respectively, with relatively high precision (0.627) but relatively low recall (0.495; Figure 3). This result indicated that FMGM was relatively conservative in estimating the networks and differences. The performance was better in estimating the networks than the difference (F1 score 0.572 vs. 0.410).

Compared to the previous method, the overall performance was improved (Table 1; F1 score 0.546 and 0.392). FMGM showed a better performance in the detection of inter-network differences (F1 scores 0.410 and 0.217), as expected due to the penalty function inducing the sparsity of the network differences. Surprisingly, FMGM showed better performance also in inferring the network edges themselves (F1 score 0.572 and 0.492). The previous method had a higher recall compared to the proposed method (0.587 vs. 0.495), but the crucial decrease in the precision (0.344 vs. 0.627) was the main reason of the lower F1 score.

[②] (Jan 2022) The package is downloadable, but the required Java code is not able to be downloaded due to "403 forbidden" error.

## 2.7. Real Data Analysis: DNA Methylation Data

For demonstration, FMGM was applied to cohort data. This data includes omics data representing DNA methylation statuses and several clinical variables related to the methylation.

### Data Description

Study subjects were from Mothers and Children's Environmental Health (MOCEH) multicenter prospective cohort study [31]. Participants enrolled in health centers in Ulsan, Ewha, and Dankook university from 2006 to 2010, with informed consents at the first prenatal visit. Pregnant women over 18 years old, with the pregnancies under 20 weeks, living near the study sites without moving plan within a year of screening, and without cognitive or mental defect were able to participate.

DNA CpG site methylation data was collected from 384 umbilical cord blood samples with HumanMethylationEPIC BeadChip (Illumina Inc, San Diego, CA, USA). The data acquisition method followed the manufacturer's instructions and a previously described publications [32, 33]. Preprocessing of the methylation data was performed with ewastools R package [34]. Data points having detection p-values over 0.01 were masked as missing, and dye bias was corrected with RELIC with the Theil-Sen estimator [35]. Beta values representing the ratios between methylated and unmethylated signals were calculated afterwards.

Leukocyte compositions, used in the quality control and analyses steps, were estimated from beta values by Houseman's method [36] with a reference panel by Salas *et al.* [37]. Identifying with optimal reference libraries (IDOL) was used to select CpG sites for deconvolution [38]. The estimated proportions of CD4+, CD8+, natural killer cells, monocytes, granulocytes, B cells, and nucleated red cells were retained for each sample. This process was performed with an extension of Bioconductor package minfi [39].

Ewastools package [34] was also used in the quality control steps. Participants were excluded by the following criteria: (1) 17 BeadArray control metrics described in the BeadArray Controls Reporter Software Guide (support.illumina.com) with the recommended cutoffs, (2) the inconsistency between normalized probe intensities on X & Y chromosomes and the reported sexes, (3) outliers (mean log odds of being outlier calculated from SNP probes bigger than −4) or duplicates from genotypes, (4) outliers from principal component (PC) scores calculated with autosomal data, and (5) outliers from estimated leukocyte compositions. Also, methylation sites with the proportion of missing values larger than 3% were excluded.

Clinical variables in the data included delivery method, children's sex, survey of maternal smoking during pregnancy, nulliparity, maternal education, gestational age in weeks, and maternal pre-pregnancy body mass index (BMI). Delivery method, children's sex, maternal smoking, and nulliparity were binomial variables, while maternal education is a three-level multinomial variable (~high school graduation, junior or community college graduation, and college or university graduation~). Gestational ages and prenatal BMI were numeric variables.

Mothers with some of the symptoms that can act as indicators for caesarean section were discarded before analyses. The exclusion criteria were gestational diabetes, pregnancy induces hypertension including pre-eclampsia, or fetal macrosomia. After the exclusion, the total number of 300 samples were used in the analyses.

### Data Analysis

As a group variable, delivery methods (caesarean section or not) were used. As a feature selection, linear model regression with limma

[40] was conducted beforehand for methylation data with the adjustment of clinical variables and leukocyte compositions, and 20 CpG sites with the smallest p−values were utilized in the downstream analyses.

Fused MGM was applied to the inference of networks by delivery methods. Continuous variables include the methylation profiles, gestational ages, pre−pregnancy BMI, leukocyte compositions, and ten genotype PC scores, and discrete variables include children's sex, maternal smoking statuses, nulliparity, and maternal education. For the decision of penalization parameters, StEPS with ten values evenly spaced on log2−scale from 0.08 to 0.64 were conducted.

## Results

Table 2 shows the summary of clinical variables in each group. Regarding the number of the variables, all of the covariates did not show significant difference between two groups (Wilcoxon rank−sum tests or Pearson's chi−squared tests).

In both groups, among 528 possible pairs of variables, 66 (12.5%) had non−zero correlations (Figure 4). Among clinical variables, maternal BMI had significantly negative correlation with cg08959771 ($8.541\times10^{-4}$), and infant sex was related to the methylation statuses of cg13136220 (∓ 0.114). According to MRC−IEU EWAS catalog [41], cg13136220 was significantly related to sex, age, and their interaction [42]. The validation may be needed to check if the correlation between cg13136220 and age is mediated by sex.

Interestingly, there was only one difference of networks between 2 groups. The negative correlation between the methylation statuses of cg14642773 and cg07582043 had different size (−0.013 in C−section group, −0.018 in non−C−section group). cg14642773, located on position 75,888,474 on chromosome 18 and S−shore of CpG island ranging from 75,886,577 to 75,886,801, was reported to be significantly related to age [42]. However, cg07582043 had no annotation and did not have any nearby genes or previously reported relationships. Thus, the additional study regarding cg07582043 may be needed to clarify the biological meaning of the result.

## 2.8. Discussion

In this chapter, a new method named fused MGM (FMGM) was proposed and demonstrated with simulated and real datasets. FMGM showed better performance in terms of precisions and F1-scores, compared to separate inference by class with the previous method by Sedgewick *et al.* [19].

Type 1 errors were alleviated much better in FMGM than in the previous method. It was also found that FMGM showed much better performance in inferring the networks themselves. This may be because the penalty function in FMGM makes the resulting networks similar, and a network from one class can be considered as 'reference information' in the inference of a network in another class. Thus, the inference precision may benefit from the optimization complementary property.

The synthesized data firstly used a setting that the networks from 2 classes differ by whole sub-networks. This setting was derived from the paper of joint graphical LASSO [10], but the setting does not cover cases where only one or more 'edges' are different. It may be possible that, if the proportion of differential edges is too low, FMGM may show poor performance compared to the former case, since the differences are likely to be shrunken to zero since all of the penalties with the same type are bound with a same parameter. Joint graphical LASSO and its derivations [43, 44] used simulated data with the settings where subnetworks or matrix blocks differ, but they did not handle the cases with only a few different edges. Using additional parameters may handle these cases, but it will make the computation too complicated and can suffer from overfitting problem.

A pilot study to test if FMGM could handle the cases if the networks are different by a few ‘edges’ was conducted. All of the procedures were same, but only 20% of the edges in each sub-network were excluded from each class, and for simplicity, data with only two sub-networks (20 variables × 2) were simulated. The results are in Table 3. Even though the simulation was simplified and the number of repetitions was only 10, FMGM still showed inference performance superior to the separate inference with causalMGM (F1 scores 0.609 vs. 0.504). Future studies may check if these results are still valid for larger number of variables with more repetitions. Also, other algorithmic bypass or methods tailored for these cases may be considered or developed in further studies.

The main limitation of FMGM is the that it is computationally intensive; a single run of FMGM for 500 observations with 100 variables took 3-4h (CPU Intel Xeon 12Core 24 threads, RAM DDR4 16G; Figure 5). The most computationally intensive part is the calculation of $p_L(\Lambda)$, which involves the sequential optimization of parameters in the two groups. The parameters in two classes must be optimized individually until they converge, and this requires a considerable amount of time. Using looser cutoffs for convergence or a smaller iteration number may avoid this problem. However, it may affect the overall inference accuracy. In fact, the observation showed that the looser cutoff resulted in worse final target function values (Figure 6). Thus, the cutoff should be set to find the appropriate balance between time consumption and inference accuracy. In our implementation, the cutoff was set empirically to a rooted mean squared deviation (RMSD) of $10^{-5}$, since this cutoff did not show a significantly longer running time but had better target function values (Figure 5 & 6) compared to looser cutoffs. Another solution is to implement the method in faster computer languages, such as C++.

StEPS [19] was utilized for regularization parameter selection.

StEPS assumes independence of the edges by type, which is violated in the FMGM setting. If the weight of a specific edge is fixed at zero in both classes, the difference also shrinks to zero. Because the regularization parameter values are determined without considering the dependence, the values corresponding to the difference can become large, and the actual inference can be more conservative than expected. Because of this dependence, setting the StEPS instability cutoff to 0.1 was also considered, but the performance was not appreciably different from the original settings (results not shown). Further studies that resolve this violation and develop a parameter selection method that consideres dependence should be performed.

Moreover, in this study, StEPS was not compared with other parameter selection methods such as Akaike information Criterion (AIC) or cross-validation (CV). This was because there were six penalization parameters in the penalty function of the suggested method, and it is practically unfeasible to test all of the combinations of the candidate values. For example, if there are four candidate values, the number of combinations to be tested should be $4^6$=4,096. StEPS could reduce the computational burden significantly, but it is still difficult to conclude that StEPS performs well enough to replace those methods. Sedgewick et al. (2016) showed that StEPS showed better results than traditional methods such as AIC for single class cases, where there are three regularization parameters, but it is also not sure if this can be extended to the method in this paper, where six parameters are needed. Therefore, future studies may be conducted to compare the performance of StEPS and other parameter selection methods.

# Chapter 3. Integration of Prior Information for Network Inference

## 3.1. Introduction

In some cases, the prior information may be available and researchers want to incorporate it. The existing databases, such as Kyoto Encyclopedia of Gene and Genomes (KEGG) pathways [45-47] or STRING database [48, 49], can act as a hint of inference of the network structures. To achieve this goal, Manatakis *et al.* [50] suggested a new method named prior-induced MGM (piMGM) that extends a framework by Sedgewick *et al.* [19]. This method was built to accept multiple prior sources and give different scores to each of the sources based on their reliabilities.

In this section, an extension of FMGM to make use of the pre-existing information will be explained. Section 3.2 re-defines the penalty function of the network parameters, and section 3.3 provides the method how to determine the regularization parameters corresponding to newly added terms. Section 3.4 and 3.5 describes the results of applying the method to the simulated and the real data, respectively, and the limitations and future works are discussed in section 3.6.

## 3.2. Use of Separate Parameter for Prior Information

Prior-induced MGM (piMGM) proposed by Manatakis *et al.* [50] estimated a probabilistic graph while incorporating prior information from multiple sources with different reliabilities. piMGM used the following penalty terms instead:

$$\lambda_{cc}^{np} \sum_{t<s} \left| \beta_{st}^{\left(\left(\Theta^{(m)}\right)\right)} \right|^{np} + \lambda_{cc}^{wp} \sum_{t<s} \left| \beta_{st}^{\left(\left(\Theta^{(m)}\right)\right)} \right|^{wp}$$
$$+\lambda_{cd}^{np} \sum_{s,j} \left\| \rho_{sj}^{\left(\left(\Theta^{(m)}\right)\right)} \right\|_2^{np} + \lambda_{cd}^{wp} \sum_{s,j} \left\| \rho_{sj}^{\left(\left(\Theta^{(m)}\right)\right)} \right\|_2^{wp}$$

$$+\lambda_{dd}^{np}\sum_{r<j}\left\|\phi_{rj}^{\left(\left(\Theta^{(m)}\right)\right)}\right\|_F^{np}+\lambda_{dd}^{wp}\sum_{r<j}\left\|\phi_{rj}^{\left(\left(\Theta^{(m)}\right)\right)}\right\|_F^{wp}$$

Here, each of the edge types have two penalty terms; one corresponds to the edges that do not have prior information ($np$), another corresponds to the edges with the prior information ($wp$). For non−prior edges, the same method of StEPS can be utilized. For edges with the prior, Manatakis *et al.* suggested a procedure of several steps to determine the value of the parameters, as described in the section 3.3.

From our knowledge, there are several databases that can work as prior information for network structures, but no information for the difference of the networks by the phenotype is currently available. Therefore, the penalty terms with prior were applied only to the edge weight norms, and the final form of the penalty function is as follows:

$$\begin{aligned} g(\Theta) = {} & \lambda_{cc}^{np}\sum_{m}\sum_{t<s}\left|\beta_{st}^{\left(\Theta^{(m)}\right)}\right| + \lambda_{cd}^{np}\sum_{m}\sum_{s,j}\left\|\rho_{sj}^{\left(\Theta^{(m)}\right)}\right\|_2 + \lambda_{dd}^{np}\sum_{m}\sum_{r<j}\left\|\phi_{rj}^{\left(\Theta^{(m)}\right)}\right\|_F \\ & +\lambda_{cc}^{wp}\sum_{m}\sum_{t<s}\left|\beta_{st}^{\left(\Theta^{(m)}\right)}\right| + \lambda_{cd}^{wp}\sum_{m}\sum_{s,j}\left\|\rho_{sj}^{\left(\Theta^{(m)}\right)}\right\|_2 + \lambda_{dd}^{wp}\sum_{m}\sum_{r<j}\left\|\phi_{rj}^{\left(\Theta^{(m)}\right)}\right\|_F \\ & +\lambda_{cc}'\sum_{t<s}\left|\beta_{st}^{\left(\Theta^{(1)}\right)}-\beta_{st}^{\left(\Theta^{(2)}\right)}\right| + \lambda_{cd}'\sum_{s,j}\left\|\rho_{sj}^{\left(\Theta^{(1)}\right)}-\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2 \\ & +\lambda_{dd}'\sum_{r<j}\left\|\phi_{rj}^{\left(\Theta^{(1)}\right)}-\phi_{rj}^{\left(\Theta^{(2)}\right)}\right\|_F \end{aligned}$$

The new method will be denoted as prior−induced FMGM (piFMGM).

## 3.3. Determination of Regularization Parameters

### The Reliability of Prior Information

Denote there are $R$ prior sources $\{t_1, t_2, \dots, t_R\}$, and each source have the prior information vector $M(t_r) = \{m_{st}(t_r)\}$, whose element means the probability of each edge to exist. If the prior information for an edge was invalid, the value of the corresponding element was set to null. The set of edges with the non−null elements in $M(t_r)$ was denoted as $wp(t_r)$, and all of the edges with their corresponding prior is denoted as $wp = \bigcup_{t_r} wp(t_r)$. Also, a set of sources that contain prior information about an edge between variables $s$ and $t$ was marked as $T_{st}$.

From $N$ subsamples, the expected number of subsamples that contain the edge between elements $s$ and $t$, based on the prior source $t_r$, was equal to

$$\psi_{st}(t_r) = m_{st}(t_r) \cdot N$$

If the empirical probability of the existence of the edge between elements $s$ and $t$ was denoted as $P_{st} = N_{st}/(N \cdot J)$, where $N_{st}$ was the number of subsamples including the edge and $J$ was the number of regularization parameters tested, the number of appearances was assumed to follow a Binomial distribution $B(N, P_{st})$. Since Gaussian distribution could approximate the Binomial distributions well [51], the approximation was carried out to make the burden of calculation lower. The mean and variance of the Gaussian distribution was denoted as $\mu_{st} = P_{st}N$ and $Var_{st} = P_{st}(1 - P_{st})N$.

The confidence on the source $t_r$ was calculated as follows:

$$\tau(t_r) = \frac{\sum_{st \in wp(t_r)} |\psi_{st}(t_r) - \mu_{st}|}{|wp(t_r)|}$$

The larger value of $\tau(t_r)$ means the divergence between the empirical mean estimation and the prior source estimation is larger, leading to less confidence of the corresponding source.

Since different sources might have information of different numbers of edges, the calculation of $\tau(t_r)$ could be affected.

Manatakis *et al.* alleviated the problem by estimating the empirical null distribution by permutation, similarly to GSEA [52]. In other words, given a prior source, the labels of the variables were randomly permuted and $\tau(t_r)$ were repeatedly calculated to produce the distribution. To make the scale of $\tau(t_r)$ invariant to the size of the prior sources, $\tau(t_r)$ was divided by the mean of the empirical null distribution. Also, the null distribution could be used to calculate the accordance of the prior source with the data. The proportion of permuted values smaller than $\tau(t_r)$ was the empirical P−value of the prior source $t_r$, and the smaller p−value meant that the network properly embodies the source.

### Integrating Prior Information from Multiple Sources

The model for the prior information from source $t_r$ could be expressed as a Gaussian distribution $N(\psi_{st}(t_r), \tau(t_r)^2)$. For multiple prior sources of the edge between $s$ and $t$, the prior distributions were combined as follows:

$$\Phi_{st} = \sum_{t_r \in T_{st}} w(t_r) \cdot N(\psi_{st}(t_r), \tau(t_r)^2)$$

$$w(t_r) = \frac{1/\tau(t_r)}{\sum_{t_r \in T_{st}} 1/\tau(t_r)}$$

The weight assumed that prior information with smaller variance is more reliable. The mixture was then approximated with a single Gaussian distribution that minimizes the Kullback−Leibler (KL) divergence from the mixture. The mean and the variance of the approximated Gaussian distribution were [53]

$$\psi_{st}^{KL} = \sum_{t_r \in T_{st}} w(t_r) \cdot \psi_{st}(t_r)$$

$$(\tau_{st}^{KL})^2 = \sum_{t_r \in T_{st}} w(t_r) \cdot \left(\tau(t_r)^2 + \left(\psi_{st}^{KL} - \psi_{st}(t_r)\right)^2\right)$$

The data−driven model was assumed to be Binomial, which was then approximated with a Gaussian distribution $N(\mu_{st}, Var_{st})$. By the Bayes rule, the posterior was also a Gaussian distribution $N(\mu_{st}^{*}, Var_{st}^{*})$ with the parameters

$$\mu_{st}^{*} = \frac{\psi_{st}^{KL} Var_{st} + \mu_{st}(\tau_{st}^{KL})^2}{(\tau_{st}^{KL})^2 + Var_{st}}$$

$$Var_{st}^{*} = \frac{(\tau_{st}^{KL})^2 \cdot Var_{st}}{(\tau_{st}^{KL})^2 + Var_{st}}$$

### Selecting the Regularization Parameter

For each tested regularization parameter $\lambda$, the number of subsamples $\bar{\theta}_{st}(\lambda)$ that contain the edge between $s$ and $t$ in the estimated network was counted. The probability of the appearance was calculated as

$$\alpha_{\mathrm{st}}(\lambda) = \int_{\bar{\theta}_{st}(\lambda)-\epsilon}^{\bar{\theta}_{st}(\lambda)+\epsilon} N(x|\mu_{st}^{*}, Var_{st}^{*})dx$$

where $\epsilon$ was a small arbitrary positive value (default 0.05 in the implementation). The probabilities were calculated for all of the edges with the priors, and the following score was calculated:

$$\mathrm{score}(\lambda) = \sum_{st \in wp} \alpha_{\mathrm{st}}(\lambda) \cdot \left(1 - 2\hat{\xi}_{st}(\lambda)\right)$$

The value of the score function increases as the probability of the edge appearance $\alpha_{\mathrm{st}}(\lambda)$ increases and as the instability of the inference $\hat{\xi}_{st}(\lambda)$ decreases. The value of the regularization parameter was set as $\lambda$ that maximizes the score.

## 3.4. Simulated Data Analysis

### Data Description

The same data and networks from section 2.5 were used in this simulation. The prior information was generated using the existing networks, with a method slightly modified from [50].

Six prior networks per simulation were generated, and three of them were selected as ‘reliable’ priors. For each prior, the number of edges to be included in the prior was set to be a random number between 10 and $N$, where $N$ is a number of edges existing in any of the networks in two classes. In three reliable priors, 100%/80%/60% of the prior edges were sampled from the edges existing in at least one of the true networks, and the others were randomly chosen among the edges not in both of the networks. The information of these edges was set to be real numbers between 0.6 and 1. For unreliable sources, the edges were selected with the same procedure as reliable ones, and the information was randomly set from 0 to 1.

### Data Analysis

The parameter selection procedure was run with same candidate values from section 2.5. For edges without priors, the regularization parameters were set to be the smallest values with the instabilities under the threshold, which is the same method as section 2.5. For edges with priors, the value with the highest score, calculated as in section 3.2, for each edge type was selected.

### Results

The estimated confidence values of reliable priors were consistently lower than those of unreliable priors (Figure 7). Since lower values indicate larger confidence, the results were consistent with the simulation settings. For reliable sources, the values became larger if the proportion of true edges went lower, indicating less true information in the prior (and simultaneously more false prior) leads to the lower confidence. Therefore, the indices properly reflect the reliability of the prior sources

piFMGM showed slight improvement in the performance compared to the classic FMGM (F1 score 0.546 & 0.562; Table 1). The F1 scores were higher for piFMGM when inferring the network structures (0.572 & 0.589), mainly because of slightly higher recall in piFMGM (0.538 & 0.562. Also, the inference of the difference of networks had similar trend (0.410 & 0.423), with subtle difference in precision (0.643 & 0.643) but slightly better recall (0.322 & 0.337). Therefore, including prior sources could increase the performance of the inferences, although the effects were slight.

## 3.5. Real Data Analysis: Multi-Omics Data from Asthma Patients

Multi−omics datasets collected from Asan Medical Center (AMC) were used for demonstration of piFMGM. The datasets include transcriptome, proteome, and DNA methylome profiles from blood cells along with several clinical variables.

### Data Description

Total 294 asthma patients from AMC had omics profiles of gene transcription, protein, and DNA methylation. Transcriptome data was obtained for 500 samples with RNA−sequencing technique. Trimmomatic 0.39 [54] was used to remove reads with bad qualities, and the refined data was mapped onto human hg19 reference genome with HISAT 2.2.1 [55, 56]. The alignments were sorted with SAMtools 1.9 [57], and the read counts of total 28,720 genes were acquired with HTSeq 1.99.2. Genes with not enough amount of expression for analyses, total expression level lower than one−tenth of the sample size, were discarded, and 22,765 genes were used in the analysis. For linear modeling, the read counts were normalized with trimmed mean of M−values (TMM) and voom transformed [58].

Proteomic profiles for 404 samples were obtained with sequential window acquisition of all theoretical mass spectra (SWATH−MS), with quantification with the spectral library obtained by basic−pH fractionation. Proteins with the missing rates under 70% were retained, and missing values were imputed with imputeLCMD [59, 60].

Cytosine methylation profiles in genome−wide scale was observed with methylated DNA immunoprecipitation sequencing (MeDIP−seq) technique for 539 samples. Reads with poor qualities were removed with Trimmomatic 0.39 [54]. The remaining reads were mapped on hg19 human reference genome with Bismark aligner v0.23.0 [61], and SAMtools 1.9 [57] was used for sorting the alignments. The duplicated alignments were deduplicated with Bismark with default options. Similarly to RNA−seq data, the gene−wise read counts were obtained, and TMM normalization and voom transformation [58] were conducted.

### Data Analysis

For each of the omics datasets, limma [40] was first applied and top 20 features were selected for the further analysis. Sex, age, and BMI were used as adjusting covariates.

PiFMGM was applied to the data of 294 subjects with all of the omics information and clinical variables. Blood eosinophil levels exceeding 300 or not were used as a group variable, and 158 out of 294 exceeded the criterion. For prior information, STRING protein−protein interaction database [49] and BioGRID database [62]. For STRING database, pre−calculated scores based on evidence channels were directly used as priors, and the priors were applied to proteomics data only. For BioGRID database, only interaction validated by multiple sources were used with ‘hard’ priors of 1. This prior database was applied to each of the omics datasets separately. In StEPS for parameter determination, ten values from 0.08 to 0.64, evenly spaced on log2 scale, were tested.

## Results

Table 4 summarizes clinical variables used in the analyses. Since age in the group with lower eosinophil level was significantly older (Wilcoxon rank-sum test p-value 0.0149), the results may be confounded by the effect of ages, and the interpretation requires caution.

The overview of the inference results is depicted in Figure 8. Compared to correlations between variables from different omics profiles, only 16 (1.33%) of which were non-null in both groups, correlations within same omics data had much higher densities of 45.2% (transcriptome), 31.1% (proteome), and 31.6% (methylome), respectively. Features in proteome and CpG methylome showed relatively high proportion of non-zero interactions (15 out of 400 in both groups, 3.75%).

Within-omics interactions also showed higher densities in the inference results of difference. 1.05% of edges between transcriptome variables, 8.95% of edges between proteome variables, and 5.79% of edges between CpG methylome variables were different between groups. Among between-omics edges, KRT5 protein - *C6orf25* methylation and C4BPB protein - *RP11-15E18.6* methylation had different amount of associations (Table 5). According to the previous study, keratin is suspected to be involved in eosinophilic esophagitis [63], and *C16orf25* codes *MPIG68* gene which is located in the major histocompatibility complex (MHC) class III region and is a member of the immunoglobulin (Ig) superfamily [64]. The interaction between 2 components are not reported yet, but since both of them are related to immune system activities, it may be hypothesized that the activities are balanced in low-eosinophil group. C4BPB protein is related to innate immune system, but its relationship with eosinophil is still unclear, and the function of *RP11-15E18.6* is not revealed yet.

Eight edges had prior information, and those with higher STRING scores or having BioGRID information tended to have larger amount of correlation (Table 6). Only one of them, APOA1 and APOD proteins, showed different amount of correlations. The negative correlation became larger in the group with higher eosinophil. Both proteins are apolipoproteins, and APOA1 was reported to reduce eosinophil migration and regulate the related diseases [65]. Therefore, it may be assumed that the larger negative association induces the lower amount of apolipoproteins, and thus it causes higher eosinophil level.

## 3.6. Discussion

In this chapter, the inclusion of prior sources was implemented with methods similar to prior-induced MGM (piMGM) by Manatakis *et al.* [50]. According to the results, this method successfully reflected sources with higher reliabilities and ignored unreliable sources. Moreover, compared to FMGM without priors, piFMGM showed slightly better F1 scores, though the amount of increase was subtle. In details, all of the indices, except the precision in inferring the differences, showed improved results compared to the original FMGM.

It was tried to induce prior information also to the differences, since it may improve the performance of piFMGM further. However, the database for the difference of networks by phenotypes does not exist as our knowledge. Future studies may focus on improving prior-induced methods by constructing the database of network differences, for diverse variables and in diverse cases.

# Chapter 4. Multi-Omics Data Analysis of Atopic Dermatitis (AD)

In this chapter, multi−omics data from 6−months infants with the disease information of atopic dermatitis (AD) is analyzed. The analyses include inference of the underlying networks for each of the disease statuses. This analysis was the extension of a previously published study [66].

Section 4.1 explains the background of the study. Section 4.2 describes the sources and the structures of omics datasets, and the statistical methods for analyses are explained in section 4.3.

## 4.1. Background

Atopic dermatitis (AD) is characterized by overactivation of the immune systems and consequential inflammation of skin barriers, and associated with various symptoms such as abnormalities of epidermal lipid and protease [67−69]. AD shows great variability between the patients by several reasons, such as the interaction between internal factors e.g. genetic variations and external factors e.g. surrounding environment [70].

For deeper insights into the inter−patient variability of AD, several studies tried to combine omics data with the disease information. Mostly, previous studies have used a single type of omics data. For example, some studies performed genome−wide association studies (GWAS) and succeeded to find genetic variants possibly related to AD [2−4], while other studies deciphered the correlation between gut microbial colonization and a risk of the disease [71, 72]. These single omics studies were successful to search for the candidates of biological markers of AD. However, using more than one type of omics data is anticipated to reveal interplays between the features and therefore the underlying pathogenic processes.

In this chapter, multi-omics data of 6-month-old infants collected from Cohort for Childhood Origin of Asthma and allergic diseases (COCOA) [73] was analyzed. Gene transcriptome profiles, represented by microarray data, and microbial compositional and functional profiles, obtained with sequencing techniques, were used. The feature selection with sparse DIABLO was firstly performed, and the construction of disease prediction model and network inference analysis were conducted with the selected variables.

## 4.2. Data Description

The Cohort for Childhood Origin of Asthma and allergic diseases (COCOA) is a prospective birth-cohort study of Korean inner-city population, started September 2007 [73]. 6-months children were investigated annually with standardized follow-up assessments until 10 years of ages, without regarding the development of allergic diseases.

Ninety-five children with atopic dermatitis (AD) were selected from the cohort, and controls were selected using propensity scores calculated based on age, sex, and feeding type. The data used in this study comprised biological multi-omics data and several clinical covariates for the selected samples. Multi-omics data included gene transcriptome profiles and intestinal microbial profiles in 6-month-old children.

### Gene Transcriptome Data by Microarray

Gene transcriptome data was generated from fecal samples of

199 participants using microarray on GeneChip® Human Gene 2.0 ST Arrays (Thermo Fisher Scientific, Inc., Waltham, MA, USA) by Macrogen, Inc (Geumcheon-gu, Seoul, Republic of Korea). Colonocytes collected from fecal samples were used for total RNA extraction. Signal intensities of 44,625 probes were generated and normalized using the Robust Multi-chip Average (RMA) method. A total of 30,980 probes were annotated with their corresponding genes and used for downstream analyses.

### Microbial Compositional Data by 16S rRNA Gene

Microbial compositional profile information was obtained using 16S rRNA amplicons from the stool samples [74]. Two groups, composed of 149 and 48 participants, were included in microbial data generation separately, with 454 pyrosequencing and Illumina MiSeq platforms, respectively. Seventy-six genera were commonly observed from both of the platforms and used in the analyses.

Quality control was performed for each sequencing platform. Genera were removed if (1) the genus read count for all subjects was less than 0.05% of the total read counts or (2) the proportion of subjects with at least one read count for the genus was less than 25% of the total subjects. The Li *et al.* [75] and Bokulich *et al.* [76] criteria were used, while the actual cutoff values were decided empirically. After quality control, log values of counts per million (log-CPM) transformation were applied to each subject using the R package edgeR [77]. Centering and scaling were conducted to adjust batch effects using sequencing.

### Microbial Functional Data with Metagenome Shotgun

Microbial functions were profiled via whole-metagenome sequencing using stool samples [74]. Similar to 16S rRNA amplicon sequencing, metagenome sequencing was performed for two separate datasets with 58 and 40 subjects. The Nesoni highthroughput sequencing data analysis toolset (ver. 0.127) Nesoni clip tool was used to remove the Illumina adapter sequences and sequences with lengths < 150 bp in each pair or Q scores < 20. Moreover, human DNA sequences were removed using BBMap with the reference human genome.

Functional profiles were annotated for sequences using Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology IDs (KO) with HUMAnN2 utility scripts. After annotation, the reads per kilobase (RPK) values for each pathway were retained. A total of 361 pathways were commonly observed from two datasets, and for each dataset, pathways with subjects having at least one read count less than 25% of the total subjects were excluded. The remaining data were log−CPM transformed with edgeR [77]. To alleviate the batch effect, centering and scaling were performed separately before combining.

### Clinical Covariates

Clinical covariates for the statistical analyses include sex, delivery method, feeding method, and family history. MissForest [78] was used to fill in the missing information. The clinical covariate summary statistics are shown in Table 7 with the comparisons using chi−squared tests performed with the R package coin [79]. None of the covariates showed significant differences between patients and controls (chi−squared test, $p > 0.05$).

### Subject Description

84 subjects had full profiles of all omics information and clinical variables, 38 of which were patients with AD. Microbial composition profiles for 37 and 47 participants were generated using 454 pyrosequencing and Illumina MiSeq, respectively, and functional profiles for 56 and 28 participants were obtained separately.

## 4.3. Statistical Analysis

### Differential Analysis

As a pilot study, biological markers that significantly differ by disease statuses were searched for each omics data. The differential tests were performed with Linear Models for Microarray Data (limma) [40], and the resulting p-values were adjusted by Benjamini-Hochberg's method [80].

### Feature Selection

To lower the burden of the computation and focus on more reliable variables, the feature selection was conducted before the network analysis. Sparse DIABLO with a single component was used for feature selection using the R package mixOmics [81, 82]. The design matrix was set according to the original publication, and the number of variables to be selected from each omics dataset was determined using the internal tuning function (range: 5-50). Clinical variables did not undergo this process.

### Network Inference

With the selected features and clinical variables for all subjects, network inference according to disease status was conducted and FMGM was applied. StEPS was run to determine the regularization parameters with ten values ranging from 0.08 to 0.64 evenly spaced on a log2-scale.

## 4.4. Results

### Differential Analysis

Among host genes, none of them showed the differential test p-values below 0.05 after Benjamini-Hochberg correction. If the cutoff was modified to 0.1, 61 genes were found to be differentially expressed. However, gene set analysis with the differential genes, performed by Kolmogorov-Smirnov test with elim algorithm [83], could not find any enriched GO terms in biological process (BP) category. The gene set analysis was conducted with R package topGO [84]. The gene list is presented in Table 8.

For microbial profiles, no genera had the adjusted p-value under 0.1. On the other hand, the differential analysis of metagenome shotgun data reported that microbial genes related to carotenoid biosynthesis (ko00906) showed significant difference with lower abundance in AD patients than in controls (p-value $3.46 \times 10^{-6}$; 0.00125 after Benjamini-Hochberg adjustment).

### Network Inference by Disease Statuses

5 variables from microarray, 10 variables from microbial genera, and 25 variables from microbial functional pathways were selected by sparse DIABLO. Inference results are shown in Figure 9. Of the 946 pairs, 118 and 115 showed non-zero interactions in controls and patients, respectively, but only ten pairs differed between the two networks (Table 9). Among the differential edges, five showed a difference larger than $10^{-3}$. Two interactions, *LINC01036-MIR4788* and *Veillonella*-ko00311 (penicillin and cephalosporin biosynthesis), were negatively correlated in controls, but these correlations disappeared in AD group. Other three pairs, *Raoultella-Cronobacter*, ko00906 (carotenoid biosynthesis)-ko03018 (RNA degradation), and ko00906-ko04066 (HIF-1 signaling pathway), were negatively correlated in both groups, but the correlation was weaker in patients with AD. Among these, ko04066 is a pathway related to humans, and the related result might be spurious.

## 4.5. Discussion

In this chapter, biological networks of multi−omics data were considered. For multi−omics data, transcriptome by microarray, microbial composition by 16s rRNA gene sequencing, and potential microbial functions by metagenome shotgun method were used.

It has previously been shown that several microbes in guts are closely related to AD. For instance, Penders and colleagues [71, 72] discovered that some microbe species such as *Escherichia coli* and *Clostridium difficile* were related to the presence of atopic eczema. Moreover, *Clostridium* cluster I was known to act as a mediator between birth mode or birth order and AD [71].

Microbial pathways from metagenome shotgun data were considered rather than microbial genes, which match the level of human gene expression data. Pathway information was used as it was assumed that the comprehensive functions of gene groups may be associated with human gene expressions. The effect of microbial genes in gene levels may be investigated in future works, though more samples are required because of the sparsity of the data.

Among microbial gene groups, genes related to the carotenoid biosynthesis pathway were observed as the most differential biomarker of AD. Carotenoid was reported to function as a protector against oxidative damage and antimicrobial agents [85]. The previous findings indicate that oxidative stress has a crucial effect on the exacerbation of AD [86]. Carotenoid synthesized from intestinal microbiome may alleviate the disease as an antioxidant. A previous study reported ingesting antioxidant nutrients, including β−carotene, significantly lowered the risk of AD in young children [87]. Also, the effect of microbial carotenoid on the host immune response could play a role. The excessive activation of Th2/Th22−deviated immune response is responsible for AD [88]. β−carotene is converted by epithelial and intestinal dendritic cells into retinoic acid (RA), which is suggested to favor the generation of regulatory T cells and inhibit the induction of inflammatory Th2 cells in mesenteric lymph node [89]. Thus, carotenoid generated by microbiome may work as a precursor of Th2 inhibitor and lower the risk of AD by suppressing the inflammatory immune response.

From network analyses of the patients and controls, only small proportion of variable pairs showed difference between groups. *LINC01036* and *MIR4788* expressions were negatively correlated in controls, which disappeared for patients. None of the genes had reported functions or annotated diseases. Since the functions of many lncRNAs and miRNAs are still unknown, downstream analysis of these genes may give new insights into atopic eczema functions and pathogenesis.

*Veillonella* was found to be dominant in one of the main enterotypes in breast- and mixed-fed infants, although the prevalence of AD did not seem to be significantly high or low compared with another enterotypes in both groups [74]. The abundance of *Veillonalla* did not differ by disease status but was negatively associated with microbial genes related to the biosynthesis of antibiotics (ko00311: Penicillin and cephalosporin biosynthesis). This relationship disappeared in patients with AD, suggesting that AD pathogenesis might be related to the loss of balance between antibiotics and specific intestinal genera.

Similarly, negative interactions between *Raoultella* and *Cronobacter* spp. were weakened in patients with AD. No genus has been associated with atopic eczema. However, some features of microbes may be related to disease status. A species that belongs to *Raoultella* genus, *R. ornithinolytica*, converts histidine to histamine [90], a molecule that contributes to inflammatory responses. *Cronobacter* is related to diverse healthcare-related infections, such as neonatal meningitis [91]. Therefore, a weakened balance between these two genera could affect the health status of the host, although the actual consequences need to be studied.

The number of microbial genes related to RNA degradation (ko03018) was negatively correlated with the genes annotated with carotenoid biosynthesis (ko00906). As previously described, some carotenoids act as antioxidants that reduce oxidative damage caused by reactive oxygen species [92]. Since microbes can perform quality control by degrading oxidated RNA via several proteins such as MutT and PNPase [93], low amounts of carotenoid series could be related to high oxidative stress of microbial RNA molecules, requiring more microbial genes related to RNA degradation. Thus, disturbances in RNA quality control might be related to AD occurrence, and future analyses could focus on the validation of the results.

There are several limitations in the approaches of this study. First, limited number of clinical covariates were available. In the analyses, only four clinical variables were considered in the model, and none of those showed significant differences between cases and controls. There were some variables expected to be crucial for the atopic dermatitis, such as cord blood indices like IgE [94], and if those variables were available, more meaningful results and findings might be newly found. Second, only 83 samples had the complete profiles of the omics data and were available for the study. Statistical analyses with larger samples may have improved the accuracy of the inference, with the anticipated improvement of better recall.

# Chapter 5. Conclusion

In this dissertation, fused MGM (FMGM), a method to infer networks from mixed data in two classes, was proposed. The method uses likelihood functions of pairwise Markov random fields and penalty functions that make the networks and their differences sparse. The fast proximal gradient method (PGM) was used to solve the optimization problem, and the internal minimization problem was resolved using a fixed-point approach. The method showed superior performance in terms of lower type 1 errors and better F1 scores compared to inferring networks according to classes separately. Surprisingly, FMGM showed higher performance not only in inferring the differences but also in inferring the network structures, and this may be due to the complementary property of the inference processes. FMGM requires six regularization parameters to be determined, but the StEPS method requires much less effort. To the best of our knowledge, this is the first parametric approach to the simultaneous inference of networks in multiple classes.

Similar framework to prior-induced MGM (piMGM) [50] was used for introduction of prior sources to FMGM. The score metrics calculated with posterior probabilities and inference instabilities successfully gave more weights to sources with higher reliabilities. When the priors were included in the inference, the performance was slightly better than running FMGM without utilizing prior information. However, since the change was slight, further studies may be needed to improve the performance, such as constructing prior sources of network differences.

The multi-omics analysis of 6-month infants, from COCOA cohort study, was also conducted and the results were shown. By differential analysis, it was successful to find biological markers that may be helpful in prediction atopic dermatitis (AD), such as expression of host *EARS2* gene and microbial genes annotated with carotenoid biosynthesis. By applying FMGM to the data, some interactions that disappeared or weakened in AD patients were found, such as host genes *LINC01036-MIR4788* and microbial genes related to carotenoid biosynthesis and RNA degradation.

Future studies may be conducted to improve the method and analysis. The most important problem of FMGM was its time consumption, so algorithmic solutions or using other languages for the implementation may be sought. Also, in the setting of FMGM, assumption of independence between penalties in StEPS is not satisfied, and the future works may cope with this problem by adjusting StEPS or finding more appropriate threshold. The lack of diversity of simulation settings is also the shortcoming of this study; more simulations in diverse settings can be performed. In piFMGM, the conservative nature of inter-network inference was a main issue. Outside the method, finding or building databases of difference of biological networks by various phenotypes can be studied. Inside the method, solving the conservativeness and improving the overall performance can be a main subject of the future studies. In the multi-omics analysis of atopic dermatitis, larger number of samples will be helpful in the performance of prediction models, by using more omics information or interactions, and precise inference of networks. Also, downstream analyses that replicates or validates the findings may be conducted.

# Appendix

## Proof of Criterion Judging a Minimization Point

This section derives the criterion that judges a non-smooth point as a minimization point in a fixed-point approach described in section 2.3. The section focused on non-diagonal cases for continuous-discrete edges, and similar approaches could be applied to other cases.

The function to minimize was expressed as follows:

$$f\left(\rho_{sj}^{(\Theta^{(1)})}\right) = \frac{L}{2}\sum_{l=1}^{L_1}\left(\rho_{sj}^{(\Theta^{(1)})}(l) - \rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l)\right)^2 + \lambda_{cd}\left\|\rho_{sj}^{(\Theta^{(1)})}\right\|_2 + \lambda'_{cd}\left\|\rho_{sj}^{(\Theta^{(1)})} - \rho_{sj}^{(\Theta^{(2)})}\right\|_2$$

$z = \{z(l)\}_{l=1,\dots,L_1}$ was taken as an arbitrary unit vector, that is, $\|z\|_2 = 1$. If $\rho_{sj}^{(\min)} = 0$, the rate of change at $\rho_{sj}^{(\min)}$ along the direction of $z$ could be calculated as follows:

$$f\left(\rho_{sj}^{(\min)} + tz\right) = f(tz)$$

$$= \frac{L}{2}\sum_{l=1}^{L_1}\left(tz(l) - \rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l)\right)^2 + \lambda_{cd}\|tz\|_2 + \lambda'_{cd}\left\|tz - \rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2$$

$$\frac{\partial}{\partial t} f\left(\rho_{sj}^{(\min)} + tz\right)$$

$$= L\sum_{l=1}^{L_1} z(l)\left(tz(l) - \rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l)\right) + \lambda_{cd}$$

$$+ \lambda_{cd}^{'}\frac{\sum_{l=1}^{L_1} z(l)\left(tz(l) - \rho_{sj}^{\left(\Theta^{(2)}\right)}(l)\right)}{\left\|tz - \rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}$$

$$\frac{\partial}{\partial t} f\left(\rho_{sj}^{(\min)} + tz\right)\Big|_{t\downarrow 0} = L\sum_{l=1}^{L_1} z(l)\left(-\rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l)\right) + \lambda_{cd} + \lambda'_{cd}\frac{\sum_{l=1}^{L_1} z(l)\left(-\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)\right)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}$$

$$= \lambda_{cd} - L\sum_{l=1}^{L_1} z(l)\rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l) - \frac{\sum_{l=1}^{L_1}\lambda'_{cd} z(l)\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}$$

$$= \lambda_{cd} - z\cdot\left(L\rho_{sj}^{\left(\Lambda^{*(1)}\right)} + \frac{\lambda'_{cd}\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}\right)$$

where ・ is an inner product operator. Thus, the descent was greatest in the following direction:

$$R = \frac{L\rho_{sj}^{\left(\Lambda^{*(1)}\right)} + \frac{\lambda'_{cd}\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}}{\left\|L\rho_{sj}^{\left(\Lambda^{*(1)}\right)} + \frac{\lambda'_{cd}\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}\right\|_2}$$

If $z$ was replaced with $R$, the change rate was equal to the following:

$$\lambda_{cd} - \left\| L\rho_{sj}^{\left(\Lambda^{*(1)}\right)} + \frac{\lambda'_{cd}\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2} \right\|$$

$\rho_{sj}^{(\mathrm{min})} = \mathbf{0}$ was the minimizing point if the descent was non−negative, and this was equivalent to the criterion in section 2.3 where $\rho_{sj}^{(\mathrm{min})}$ was replaced with a zero vector. Therefore, the criterion was proven in the case where $\rho_{sj}^{(\mathrm{min})} = \mathbf{0}$.

Similarly, if $\rho_{sj}^{(min)} = \rho_{sj}^{\left(\Theta^{(2)}\right)}$, the descent along a unit vector $z$ was derived as follows:

$$\begin{aligned}
f\left(\rho_{sj}^{(\mathrm{min})} + tz\right) &= f\left(\rho_{sj}^{\left(\Theta^{(2)}\right)} + tz\right) \\
&= \frac{L}{2}\sum_{l=1}^{L_1}\left(\rho_{sj}^{\left(\Theta^{(2)}\right)}(l) + tz(l) - \rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l)\right)^2 + \lambda_{cd}\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)} + tz\right\|_2 \\
&\quad + \lambda'_{cd}\|tz\|_2 \\
\frac{\partial}{\partial t} f\left(\rho_{sj}^{(\mathrm{min})} + tz\right) & \\
&= L\sum_{l=1}^{L_1} z(l)\left(\rho_{sj}^{\left(\Theta^{(2)}\right)}(l) + tz(l) - \rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l)\right) \\
&\quad + \lambda_{cd}\frac{\sum_{l=1}^{L_1} z(l)\left(\rho_{sj}^{\left(\Theta^{(2)}\right)}(l) + tz(l)\right)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)} + tz\right\|_2} + \lambda_{cd}^{'}
\end{aligned}$$

$$\begin{aligned}
\left.\frac{\partial}{\partial t} f\left(\rho_{sj}^{(\mathrm{min})} + tz\right)\right|_{t\downarrow 0} &= L\sum_{l=1}^{L_1} z(l)\left(\rho_{sj}^{\left(\Theta^{(2)}\right)}(l) - \rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l)\right) + \lambda_{cd}\frac{\sum_{l=1}^{L_1} z(l)\left(\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)\right)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2} + \lambda'_{cd} \\
&= \lambda_{cd} - z \cdot \left(L\left(\rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l) - \rho_{sj}^{\left(\Theta^{(2)}\right)}(l)\right) - \frac{\lambda_{cd}\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}\right)
\end{aligned}$$

The descent would be the steepest in the following direction:

$$R = \frac{L\left(\rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l) - \rho_{sj}^{\left(\Theta^{(2)}\right)}(l)\right) - \frac{\lambda_{cd}\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}}{\left\|L\left(\rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l) - \rho_{sj}^{\left(\Theta^{(2)}\right)}(l)\right) - \frac{\lambda_{cd}\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}\right\|_2}$$

If $z$ was replaced with $R$, the descent was reduced to the following:

$$\lambda_{cd} - \left\|L\left(\rho_{sj}^{\left(\Lambda^{*(1)}\right)}(l) - \rho_{sj}^{\left(\Theta^{(2)}\right)}(l)\right) - \frac{\lambda_{cd}\rho_{sj}^{\left(\Theta^{(2)}\right)}(l)}{\left\|\rho_{sj}^{\left(\Theta^{(2)}\right)}\right\|_2}\right\|_2$$

If the value was non−negative, $\rho_{sj}^{(min)} = \rho_{sj}^{\left(\Theta^{(2)}\right)}$ was the minimization point. Thus, this criterion also held in the case of $\rho_{sj}^{(min)} = \rho_{sj}^{\left(\Theta^{(2)}\right)}$.

# Bibliography

1. Hawe, J.S., F.J. Theis, and M. Heinig, *Inferring interaction networks from multi-omics data.* Frontiers in genetics, 2019. **10**: p. 535.
2. Hirota, T., et al., *Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population.* Nature genetics, 2012. **44**(11): p. 1222.
3. Sun, L.-D., et al., *Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population.* Nature genetics, 2011. **43**(7): p. 690.
4. Paternoster, L., et al., *Meta-analysis of genome-wide association studies identifies three new risk loci for atopic dermatitis.* Nature genetics, 2012. **44**(2): p. 187.
5. Subramanian, I., et al., *Multi-omics data integration, interpretation, and its application.* Bioinformatics and biology insights, 2020. **14**: p. 1177932219899051.
6. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease.* Genome biology, 2017. **18**(1): p. 1-15.
7. Mackay, T.F., *Epistasis and quantitative traits: using model organisms to study gene–gene interactions.* Nature Reviews Genetics, 2014. **15**(1): p. 22-33.
8. Bocharova, A., et al., *Association and Gene–Gene Interactions Study of Late-Onset Alzheimer's Disease in the Russian Population.* Genes, 2021. **12**(10): p. 1647.
9. Friedman, J., T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso.* Biostatistics, 2008. **9**(3): p. 432-441.
10. Danaher, P., P. Wang, and D.M. Witten, *The joint graphical lasso for inverse covariance estimation across multiple classes.* Journal of the Royal Statistical Society. Series B, Statistical methodology, 2014. **76**(2): p. 373.
11. Hoefling, H., *A path algorithm for the fused lasso signal approximator.* Journal of Computational and Graphical Statistics, 2010. **19**(4): p. 984-1006.
12. Yuan, M. and Y. Lin, *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006. **68**(1): p. 49-67.
13. Lauritzen, S.L. and N. Wermuth, *Graphical models for associations between variables, some of which are qualitative and some quantitative.* The annals of Statistics, 1989: p. 31-57.
14. Meinshausen, N. and P. Bühlmann, *High-dimensional graphs and variable selection with the lasso.* The annals of statistics, 2006. **34**(3): p. 1436-1462.
15. Fellinghauer, B., et al., *Stable graphical model estimation with random forests for discrete, continuous, and mixed variables.* Computational Statistics & Data Analysis, 2013. **64**: p. 132-152.

16. Lee, J. and T. Hastie. *Structure learning of mixed graphical models*. in *Artificial Intelligence and Statistics*. 2013. PMLR.

17. Besag, J., *Statistical analysis of non-lattice data.* Journal of the Royal Statistical Society: Series D (The Statistician), 1975. **24**(3): p. 179-195.

18. Besag, J., *Spatial interaction and the statistical analysis of lattice systems.* Journal of the Royal Statistical Society: Series B (Methodological), 1974. **36**(2): p. 192-225.

19. Sedgewick, A.J., et al., *Learning mixed graphical models with separate sparsity parameters and stability-based model selection.* BMC bioinformatics, 2016. **17**(5): p. 307-318.

20. Beck, A. and M. Teboulle, *Gradient-based algorithms with applications to signal recovery.* Convex optimization in signal processing and communications, 2009: p. 42-88.

21. Combettes, P.L. and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in *Fixed-point algorithms for inverse problems in science and engineering*. 2011, Springer. p. 185-212.

22. Aibar, S., et al., *SCENIC: single-cell regulatory network inference and clustering.* Nature methods, 2017. **14**(11): p. 1083-1086.

23. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

24. Weiszfeld, E., *Sur le point pour lequel la somme des distances de n points donnés est minimum.* Tohoku Mathematical Journal, First Series, 1937. **43**: p. 355-386.

25. Kuhn, H.W., *A note on Fermat's problem.* Mathematical programming, 1973. **4**(1): p. 98-107.

26. Katz, I.N. and S.R. Vogl, *A Weiszfeld algorithm for the solution of an asymmetric extension of the generalized Fermat location problem.* Computers & mathematics with applications, 2010. **59**(1): p. 399-410.

27. Liu, H., K. Roeder, and L. Wasserman, *Stability approach to regularization selection (stars) for high dimensional graphical models.* Advances in neural information processing systems, 2010. **24**(2): p. 1432.

28. Kane, M., J.W. Emerson, and S. Weston, *Scalable strategies for computing with massive data.* Journal of Statistical Software, 2013. **55**: p. 1-19.

29. Emerson, J.W. and M.J. Kane, *biganalytics: Utilities for 'big.matrix' Objects from Package 'bigmemory'*. 2020.

30. Bollobás, B., et al. *Directed scale-free graphs*. in *SODA*. 2003.

31. Kim, B.-M., et al., *The mothers and children's environmental health (MOCEH) study.* European journal of epidemiology, 2009. **24**(9): p. 573-583.

32. Salvini, T.F., et al., *Effects of electrical stimulation and stretching on the adaptation of denervated skeletal muscle: implications for physical therapy.* Brazilian journal of physical therapy, 2012. **16**(3): p. 175-183.

33. Park, J., et al., *Methylation quantitative trait loci analysis in Korean exposome study.* Molecular & Cellular Toxicology, 2020. **16**(2): p. 175-183.

34. Heiss, J.A. and A.C. Just, *Identifying mislabeled and contaminated*

*DNA methylation microarray data: an extended quality control toolset with examples from GEO.* Clinical Epigenetics, 2018. **10**(1): p. 1-9.

35. Xu, Z., et al., *RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip.* BMC genomics, 2017. **18**(1): p. 1-7.
36. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution.* BMC bioinformatics, 2012. **13**(1): p. 1-16.
37. Salas, L.A., et al., *An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray.* Genome biology, 2018. **19**(1): p. 1-14.
38. Koestler, D.C., et al., *Improving cell mixture deconvolution by id entifying o ptimal dna methylation l ibraries (idol).* BMC bioinformatics, 2016. **17**(1): p. 1-21.
39. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.* Bioinformatics, 2014. **30**(10): p. 1363-1369.
40. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
41. Battram, T., et al., *The EWAS Catalog: a database of epigenome-wide association studies.* Wellcome Open Research, 2022. **7**(41): p. 41.
42. Mulder, R.H., et al., *Epigenome-wide change and variation in DNA methylation in childhood: trajectories from birth to late adolescence.* Human molecular genetics, 2021. **30**(1): p. 119-134.
43. Li, Z., T. Mccormick, and S. Clark. *Bayesian joint spike-and-slab graphical lasso*. in *International Conference on Machine Learning*. 2019. PMLR.
44. Yin, H., X. Liu, and X. Kong. *Gaussian Mixture Graphical Lasso with Application to Edge Detection in Brain Networks*. in *2020 IEEE International Conference on Big Data (Big Data)*. 2020. IEEE.
45. Kanehisa, M., *Toward understanding the origin and evolution of cellular organisms.* Protein Science, 2019. **28**(11): p. 1947-1951.
46. Kanehisa, M., et al., *KEGG: integrating viruses and cellular organisms.* Nucleic acids research, 2021. **49**(D1): p. D545-D551.
47. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.
48. Mering, C.v., et al., *STRING: a database of predicted functional associations between proteins.* Nucleic acids research, 2003. **31**(1): p. 258-261.
49. Szklarczyk, D., et al., *The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets.* Nucleic acids research, 2021. **49**(D1): p. D605-D612.
50. Manatakis, D.V., V.K. Raghu, and P.V. Benos, *piMGM: incorporating multi-source priors in mixed graphical models for learning disease networks.* Bioinformatics, 2018. **34**(17): p. i848-i856.
51. Papoulis, A. and S.U. Pillai, *Probability, random variables, and*

*stochastic processes*. 2002: Tata McGraw-Hill Education.

52. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.
53. Runnalls, A.R., *Kullback-Leibler approach to Gaussian mixture reduction.* IEEE Transactions on Aerospace and Electronic Systems, 2007. **43**(3): p. 989-999.
54. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.
55. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements.* Nature methods, 2015. **12**(4): p. 357-360.
56. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nature biotechnology, 2019. **37**(8): p. 907-915.
57. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
58. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome biology, 2014. **15**(2): p. 1-17.
59. Lazar, C., *imputeLCMD: A collection of methods for left-censored missing data imputation*. 2015.
60. Karpievitch, Y.V., A.R. Dabney, and R.D. Smith, *Normalization and missing value imputation for label-free LC-MS analysis.* BMC bioinformatics, 2012. **13**(16): p. 1-9.
61. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* bioinformatics, 2011. **27**(11): p. 1571-1572.
62. Oughtred, R., et al., *The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions.* Protein Science, 2021. **30**(1): p. 187-200.
63. Kiran, K., M.E. Rothenberg, and J.D. Sherrill, *In vitro model for studying esophageal epithelial differentiation and allergic inflammatory responses identifies keratin involvement in eosinophilic esophagitis.* PloS one, 2015. **10**(6): p. e0127755.
64. de Vet, E.C., B. Aguado, and R.D. Campbell, *G6b, a novel immunoglobulin superfamily member encoded in the human major histocompatibility complex, interacts with SHP-1 and SHP-2.* Journal of Biological Chemistry, 2001. **276**(45): p. 42070-42076.
65. Sturm, E.M., E. Knuplez, and G. Marsche, *Role of short chain fatty acids and apolipoproteins in the regulation of eosinophilia-associated diseases.* International Journal of Molecular Sciences, 2021. **22**(9): p. 4377.
66. Park, J., et al., *Multi-omics analyses implicate EARS2 in the pathogenesis of atopic dermatitis.* Allergy, 2021. **76**(8): p. 2602-2604.
67. Elias, P.M. and M. Schmuth, *Abnormal skin barrier in the*

*etiopathogenesis of atopic dermatitis.* Current allergy and asthma reports, 2009. **9**(4): p. 265-272.

68. Leung, D.Y., et al., *New insights into atopic dermatitis.* The Journal of clinical investigation, 2004. **113**(5): p. 651-657.

69. Yamamoto, A., et al., *Stratum corneum lipid abnormalities in atopic dermatitis.* Archives of dermatological research, 1991. **283**(4): p. 219-223.

70. Cork, M.J., et al., *New perspectives on epidermal barrier dysfunction in atopic dermatitis: gene–environment interactions.* Journal of Allergy and Clinical Immunology, 2006. **118**(1): p. 3-21.

71. Penders, J., et al., *Establishment of the intestinal microbiota and its role for atopic dermatitis in early childhood.* Journal of Allergy and Clinical Immunology, 2013. **132**(3): p. 601-607. e8.

72. Penders, J., et al., *Gut microbiota composition and development of atopic manifestations in infancy: the KOALA Birth Cohort Study.* Gut, 2006.

73. Yang, H.-J., et al., *The Cohort for Childhood Origin of Asthma and allergic diseases (COCOA) study: design, rationale and methods.* BMC pulmonary medicine, 2014. **14**(1): p. 109.

74. Lee, M.-J., et al., *Perturbations of gut microbiome genes in infants with atopic dermatitis according to feeding type.* Journal of Allergy and Clinical Immunology, 2018. **141**(4): p. 1310-1319.

75. Li, K., M. Bihan, and B.A. Methé, *Analyses of the Stability and Core Taxonomic Memberships of the Human Microbiome.* PLOS ONE, 2013. **8**(5): p. e63139.

76. Bokulich, N.A., et al., *Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.* Nature Methods, 2012. **10**: p. 57.

77. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

78. Stekhoven, D.J. *missForest: Nonparametric Missing Value Imputation using Random Forest.* 2013; Available from: http://cran.r-project.org/web/packages/missForest/index.html.

79. Hothorn, T., et al., *A Lego system for conditional inference.* The American Statistician, 2006. **60**(3): p. 257-263.

80. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.

81. Singh, A., et al., *DIABLO - an integrative, multi-omics, multivariate method for multi-group classification.* bioRxiv, 2016.

82. Rohart, F., et al., *mixOmics: An R package for 'omics feature selection and multiple data integration.* PLOS Computational Biology, 2017. **13**(11): p. e1005752.

83. Alexa, A., J. Rahnenführer, and T.J.B. Lengauer, *Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.* 2006. **22**(13): p. 1600-1607.

84. Alexa, A. and J.J.R.p.v. Rahnenfuhrer, *topGO: enrichment analysis for gene ontology.* 2010. **2**(0): p. 2010.

85. Kirti, K., et al., *Colorful world of microbes: carotenoids and their applications.* Advances in Biology, 2014. **2014**.

86. Tsukahara, H., et al., *Oxidative stress and altered antioxidant defenses in children with acute exacerbation of atopic dermatitis.* Life sciences, 2003. **72**(22): p. 2509-2516.

87. Oh, S., et al., *Antioxidant nutrient intakes and corresponding biomarkers associated with the risk of atopic dermatitis in young children.* European journal of clinical nutrition, 2010. **64**(3): p. 245.

88. Furue, M., et al., *Atopic dermatitis: immune deviation, barrier dysfunction, IgE autoreactivity and new therapies.* Allergology International, 2017. **66**(3): p. 398-403.

89. Julia, V., L. Macia, and D. Dombrowicz, *The impact of diet on asthma and allergic diseases.* Nature Reviews Immunology, 2015. **15**(5): p. 308.

90. Kanki, M., et al., *Klebsiella pneumoniae produces no histamine: Raoultella planticola and Raoultella ornithinolytica strains are histamine producers.* Applied and Environmental Microbiology, 2002. **68**(7): p. 3462-3466.

91. Holý, O. and S. Forsythe, *Cronobacter spp. as emerging causes of healthcare-associated infection.* Journal of Hospital Infection, 2014. **86**(3): p. 169-177.

92. Chew, B.P. and J.S. Park, *Carotenoid action on the immune response.* The Journal of nutrition, 2004. **134**(1): p. 257S-261S.

93. Seixas, A.F., et al., *Bacterial response to oxidative stress and RNA oxidation.* Frontiers in Genetics, 2021. **12**.

94. Edenharter, G., et al., *Cord blood-IgE as risk factor and predictor for atopic diseases.* Clinical and experimental allergy: journal of the British Society for Allergy and Clinical Immunology, 1998. **28**(6): p. 671.

# Abstract

**연구 배경**

최근 다중 오믹스 자료와 같이 다수의 변수 혹은 관찰을 포함하는 대용량 자료가 광범위하게 생산되고 있다. 이러한 자료는 연속형 및 이산형 변수를 모두 포함하는 혼합형 자료인 경우가 많으며, 이는 자료의 통계적 분석을 어렵게 한다. 특히 기저 네트워크 추론의 경우, 그간 몇몇 통계적 방법들이 제시되어 왔으나, 대부분 변수 유형이 단일하거나 집단이 하나인 경우에 대해서만 적용 가능하다.

**연구 목적**

본 연구에서는 2개 집단의 혼합형 자료로부터 기저 네트워크를 추론하는 방법인 fused MGM (FMGM)을 개발하고 제시하고자 하였다. 이 방법은 네트워크 자체에 더하여 그 차이 역시 전체 자료에 비해 희박한 밀도를 가짐을 가정한다. 또한, 6개월 아동의 다중 오믹스 자료에 이 방법을 포함한 통계적 분석 방법을 적용하여, 아토피성 피부염과 관련된 생물학적 마커 및 기저 네트워크 구조를 찾아내고자 하였다.

**연구 방법**

FMGM은 쌍별 마르코프 랜덤 필드에 기반한 통계적 모형을 사용하며, 벌점 함수를 통해 네트워크 및 차이의 희박함을 유도한다. 목적함수의 최적화에는 고속 근위 경사법을 사용하였다. 또한 FMGM의 추론에 사전 정보를 도입할 수 있도록 하는 사전 정보 유도 FMGM (piFMGM) 역시 개발하였다. 추론 방법의 성능은 역법칙 네트워크 구조를 시뮬레이션한 합성 자료를 통해 측정하였다. 6개월 아동의 다중 오믹스 정보 역시 분석하였으며, 오믹스 정보에는 숙주 유전자 전사체 (N=199), 장내 미생물체 구성 (N=197) 및 장내 미생물 기능 정보 (N=98)가 포함된다 (공통 표본 수 84). 분석에는 선형 모형을 통한 차이 분석과 FMGM을 통한 네트워크 추론을 사용하였다.

**연구 결과**

시뮬레이션한 무척도 네트워크로부터 2개 집단 자료를 생성하여 분석한 결과, 개별 집단에 대해 네트워크를 추론한 결과와 비교하여 FMGM이 더 높은 F1 점수를 나타내어 성능이 더 우수함을 보였다 (0.392 & 0.546). FMGM은 네트워크 간 차이 (0.217 & 0.410)뿐만

아니라 네트워크 자체의 추론에서도 더 우수한 성능을 보였다 (0.492 & 0.572). 사전 정보를 piFMGM을 통해 도입한 경우 전체적인 성능이 미세한 증가를 보였다 (0.546 & 0.562). 네트워크의 추론뿐만 아니라 (0.572 & 0.589), 차이를 추론할 때의 성능 역시 작은 증가세를 띄었다 (0.410 & 0.423). 6개월 아동의 아토피성 피부염 자료로부터 네트워크 추론을 수행한 결과 숙주의 *LINC01036* 및 *MIR4788* 발현, 장내 미생물의 카로티노이드 생합성 및 RNA 분해 관련 유전자 등, 10개 변수 쌍이 피부염 여부에 따른 상관성 차이를 나타냈다.

**결론**

본 연구에서 제시한 방법인 FMGM은 기존 방법에 비해 2개 집단의 혼합형 자료에서 네트워크를 추론할 때 더 좋은 성능을 나타냈다. 사전 정보를 piFMGM을 통해 포함시킬 경우 네트워크 추론의 정확성이 향상되나, 그 차이가 크지 않아 추후 연구에서 이를 발전시키기 위한 방법이 필요할 것으로 보인다. 다중 오믹스 자료의 네트워크 추론 분석을 통해 장내 미생물의 카로티노이드 생합성 또는 RNA 분해 관련 유전자 등 아토피성 피부염과 관련된 생물학적 마커를 복수 발견하였으며, 이는 아토피성 피부염의 기저에 산화 스트레스 또는 미생물 RNA 조절 등이 관련될 수 있음을 제시한다.

**주요어 :** 네트워크 추론, 통계적 방법론, 마르코프 랜덤 필드 모형, 다중 오믹스, 아토피성 피부염
**학번 :** 2016-20461

**Table 1.** Performance measures obtained using the stated network inference methods with the simulated data. Each simulated data contains 100 variables (50 Gaussian & 50 categorical), and 4,950 pairs of variables were subjected to the inference. The average values over 50 repetitions are shown, and the standard deviations are shown in the parenthesis.

| | Accuracy | Precision | Recall | F1−score | Matthews CC |
|---|---|---|---|---|---|
| CausalMGM for each group | | | | | |
| Overall | 0.974 (0.00908) | 0.344 (0.0987) | 0.587 (0.2233) | 0.392 (0.1381) | 0.413 (0.1289) |
| Intra−network | 0.980 (0.00666) | 0.508 (0.1536) | 0.591 (0.2265) | 0.492 (0.1726) | 0.504 (0.1509) |
| Cont−Cont | 0.982 (0.00920) | 0.580 (0.2081) | 0.633 (0.2973) | 0.543 (0.1954) | 0.564 (0.1748) |
| Cont−Disc | 0.984 (0.00707) | 0.606 (0.1655) | 0.753 (0.2556) | 0.607 (0.1987) | 0.626 (0.1670) |
| Disc−Disc | 0.969 (0.01792) | 0.207 (0.3244) | 0.211 (0.3516) | 0.201 (0.3378) | 0.190 (0.3431) |
| Inter−network | 0.963 (0.01529) | 0.148 (0.0687) | 0.573 (0.2242) | 0.217 (0.0846) | 0.267 (0.0982) |
| Cont−Cont | 0.965 (0.01959) | 0.173 (0.1299) | 0.629 (0.3225) | 0.232 (0.1348) | 0.291 (0.1417) |
| Cont−Disc | 0.964 (0.01618) | 0.187 (0.1093) | 0.747 (0.2531) | 0.269 (0.0925) | 0.343 (0.1069) |
| Disc−Disc | 0.959 (0.02423) | 0.060 (0.1222) | 0.194 (0.3598) | 0.090 (0.1788) | 0.092 (0.2096) |
| Fused MGM (FMGM) | | | | | |
| Overall | 0.988 (0.00250) | 0.627 (0.1013) | 0.495 (0.1462) | 0.546 (0.1208) | 0.548 (0.1180) |

| | | | | | |
|---|---|---|---|---|---|
| Intra−network | 0.986 (0.00329) | 0.626 (0.1072) | 0.538 (0.1499) | 0.572 (0.1218) | 0.570 (0.1201) |
| Cont−Cont | 0.990 (0.00420) | 0.698 (0.1175) | 0.778 (0.1908) | 0.722 (0.1256) | 0.725 (0.1252) |
| Cont−Disc | 0.991 (0.00259) | 0.947 (0.0325) | 0.570 (0.1702) | 0.696 (0.1326) | 0.722 (0.1089) |
| Disc−Disc | 0.970 (0.01229) | 0.191 (0.2939) | 0.219 (0.3431) | 0.202 (0.3163) | 0.188 (0.3230) |
| Inter−network | 0.992 (0.00133) | 0.643 (0.1129) | 0.322 (0.1630) | 0.410 (0.1556) | 0.440 (0.1379) |
| Cont−Cont | 0.994 (0.00336) | 0.761 (0.2746) | 0.411 (0.3093) | 0.504 (0.2687) | 0.540 (0.2565) |
| Cont−Disc | 0.993 (0.00176) | 0.784 (0.1470) | 0.374 (0.1703) | 0.489 (0.1767) | 0.527 (0.1567) |
| Disc−Disc | 0.988 (0.00478) | 0.175 (0.3226) | 0.135 (0.2785) | 0.151 (0.2906) | 0.150 (0.2959) |
| Prior−induced FMGM (piFMGM) | | | | | |
| Overall | 0.988 (0.00230) | 0.637 (0.0940) | 0.517 (0.1468) | 0.562 (0.1144) | 0.564 (0.1101) |
| Intra−network | 0.986 (0.00311) | 0.637 (0.0991) | 0.562 (0.1526) | 0.589 (0.1155) | 0.588 (0.1137) |
| Cont−Cont | 0.990 (0.00406) | 0.703 (0.1105) | 0.777 (0.1898) | 0.724 (0.1223) | 0.727 (0.1213) |
| Cont−Disc | 0.992 (0.00268) | 0.934 (0.0382) | 0.626 (0.1806) | 0.733 (0.1301) | 0.752 (0.1093) |
| Disc−Disc | 0.970 (0.01182) | 0.192 (0.2977) | 0.205 (0.3163) | 0.196 (0.3050) | 0.183 (0.3110) |
| Inter−network | 0.992 (0.00112) | 0.643 (0.1161) | 0.337 (0.1588) | 0.423 (0.1457) | 0.450 (0.1288) |
| Cont−Cont | 0.994 (0.00339) | 0.768 (0.2736) | 0.407 (0.3090) | 0.498 (0.2673) | 0.549 (0.2426) |
| Cont−Disc | 0.993 (0.00158) | 0.770 (0.1375) | 0.410 (0.1677) | 0.519 (0.1680) | 0.549 (0.1510) |
| Disc−Disc | 0.988 (0.00436) | 0.157 (0.2985) | 0.118 (0.2445) | 0.134 (0.2661) | 0.132 (0.2698) |

**Table 2.** Summary statistics of clinical covariates included in the network analyses of data from MOCEH cohort. For continuous variables, means and standard deviations are shown. The p−values of the difference between two groups were obtained by chi−squared tests or Wilcoxon rank−sum tests with the R package coin. BMI, body−mass index.

| Variable | Category | Caesarean section (n = 105) | Other delivery methods (n = 195) | P−value |
|---|---|---|---|---|
| Sex | Male | 62 | 92 | 0.0506 |
| | Female | 43 | 103 | |
| Smoking statuses | Smoking | 95 | 175 | 1 |
| | Non−smoking | 10 | 20 | |
| Maternal education | ~ High school | 26 | 41 | 0.3186 |
| | Junior or community college graduation | 13 | 37 | |
| | College or university ~ | 66 | 117 | |
| Nulliparity | Yes | 58 | 117 | 0.4567 |
| | No | 47 | 78 | |
| Maternal BMI | (kg/m$^2$) | 22.51 (2.99) | 21.66 (2.79) | 0.0123 |
| Delivery weight | (g) | 3349.91 (432.53) | 3258.01 (400.30) | 0.1100 |

**Table 3.** Pilot study results for the cases with network different a few edges. The average values over 10 repetitions are shown, and the standard deviations are in the parenthesis. Since the number of edges were too low due to the lower number of variables, the detailed measures by edge types were skipped.

| | Accuracy | Precision | Recall | F1−score | Matthews CC |
|---|---|---|---|---|---|
| CausalMGM for each group | | | | | |
| Overall | 0.954 (0.01677) | 0.485 (0.1439) | 0.582 (0.1250) | 0.504 (0.0828) | 0.497 (0.0786) |
| Intra−network | 0.962 (0.01346) | 0.673 (0.1962) | 0.591 (0.1329) | 0.599 (0.0985) | 0.597 (0.0939) |
| Inter−network | 0.939 (0.02597) | 0.210 (0.0812) | 0.545 (0.1349) | 0.291 (0.0804) | 0.307 (0.0756) |
| Fused MGM (FMGM) | | | | | |
| Overall | 0.969 (0.00389) | 0.601 (0.0505) | 0.624 (0.0592) | 0.609 (0.0279) | 0.595 (0.0295) |
| Intra−network | 0.966 (0.00357) | 0.629 (0.0444) | 0.671 (0.0528) | 0.647 (0.0261) | 0.631 (0.0272) |
| Inter−network | 0.976 (0.00825) | 0.502 (0.1540) | 0.420 (0.1335) | 0.432 (0.0858) | 0.435 (0.0874) |

**Table 4.** Summary statistics of clinical covariates included in the network analyses of data from Asan Medical Center (AMC). For continuous variables, means and standard deviations are shown. The p−values of the difference between two groups were obtained by chi−squared tests or Wilcoxon rank−sum tests with the R package coin.

| Variable | Category | Eosinophil < 300<br>(n = 136) | Eosinophil ≥ 300<br>(n = 158) | P−value |
|---|---|---|---|---|
| Sex | Male | 51 | 64 | 0.6409 |
| | Female | 85 | 94 | |
| Age | (years) | 52.49 (16.55) | 48.16 (15.30) | 0.0149 |
| Body−mass index | (kg/m$^2$) | 25.39 (5.03) | 24.77 (4.37) | 0.1892 |

**Table 5.** List of partial correlations different between asthma patients with low (< 300) and high (≥ 300) eosinophil levels. The study subjects were recruited from Asan Medical Center (AMC).

| Variable 1 | Variable 2 | Interaction in eosinophil < 300 | Interaction in eosinophil ≥ 300 | Difference |
|---|---|---|---|---|
| KRT5 protein | C6orf25 methylation | 0.018 | 0 | 0.018 |
| C4BPB protein | RP11−15E18.6 methylation | 0 | $9.968\times10^{-3}$ | $9.968\times10^{-3}$ |
| LAMB1 protein | sex | 0 | $1.863\times10^{-12}$ | $1.863\times10^{-12}$ |
| KRT5 protein | age | 0 | −0.034 | 0.034 |
| C4BPB protein | BMI | −0.115 | −0.051 | 0.064 |
| RPL7A methylation | BMI | $-2.239\times10^{-3}$ | 0 | $2.239\times10^{-3}$ |

**Table 6.** Network inference results for edges with priors, performed with asthma patients from Asan Medical Center (AMC). NA, not in the corresponding prior source.

| Variable 1 | Variable 2 | STRING prior score | BioGRID prior score | Interaction in eosinophil < 300 | Interaction in eosinophil ≥ 300 | Difference |
|---|---|---|---|---|---|---|
| GRN (RNA) | CTSB (RNA) | NA | 1 | −0.808 | −0.808 | 0 |
| A2M (protein) | IGF1 (protein) | 0.256 | NA | 0 | 0 | 0 |
| A2M (protein) | LCAT (protein) | 0.451 | 1 | 0 | 0 | 0 |
| A2M (protein) | APOD (protein) | 0.331 | NA | 0 | 0 | 0 |
| APOA1 (protein) | LCAT (protein) | 0.985 | 1 | −0.380 | −0.380 | 0 |
| APOA1 (protein) | APOD (protein) | 0.597 | NA | −0.132 | −0.345 | 0.214 |
| APOA1 (protein) | GPLD1 (protein) | 0.445 | 1 | −0.230 | −0.230 | 0 |
| LCAT (protein) | APOD (protein) | 0.265 | NA | 0 | 0 | 0 |

**Table 7.** Summary statistics of clinical covariates included in the multi−omics analyses of data from the COCOA cohort. The p−values of the difference between two groups were obtained by chi−squared tests with the R package coin. AD, atopic dermatitis.

[*]Since only one sample had a family history score of 2, the corresponding sample was excluded from the analyses.

| Variable | Category | Controls (n = 46) | AD patients (n = 38) | P−value |
|---|---|---|---|---|
| Sex | Male | 23 | 27 | 0.0504 |
| | Female | 23 | 11 | |
| Delivery mode | Caesarean | 15 | 11 | 0.718 |
| | Vaginal | 31 | 27 | |
| Feeding type | Breast | 13 | 15 | 0.552 |
| | Mixed | 24 | 17 | |
| | Formula | 9 | 6 | |
| Family history | 0 | 21 | 13 | 0.340 |
| | 1 | 24 | 25 | |
| | 2 | 1[*] | 0 | |

**Table 8.** List of host genes that showed significantly different expression between atopic dermatitis (AD) patient and control 6-months-old infants. The study subjects were gathered from COCOA cohort, and the differential analysis was performed with Linear Models for Microarray Data (limma). Genes with Benjamini-Hochberg adjusted p-values below 0.1 are shown. Mean values represent the average of RMA normalized expression values. P-values were adjusted with Benjamini-Hochberg's method.

| Gene | Effect size | p-value | Adjusted p-value | Mean (AD) | Mean (Control) |
|---|---|---|---|---|---|
| *C16orf72* | 0.297 | $1.73\times10^{-6}$ | 0.054 | 1.791 | 1.496 |
| *OR2J3* | −0.334 | $6.35\times10^{-5}$ | 0.091 | 1.144 | 1.454 |
| *MARCO* | 0.191 | $1.54\times10^{-5}$ | 0.091 | 2.275 | 2.088 |
| *MIR4320* | −0.374 | $1.90\times10^{-5}$ | 0.091 | 1.205 | 1.557 |
| *LOC100288570* | −0.306 | $2.88\times10^{-5}$ | 0.091 | 3.273 | 3.592 |
| *SNORD115-3* | −0.268 | $2.90\times10^{-5}$ | 0.091 | 0.932 | 1.154 |
| *SNORD114-1* | −0.444 | $3.36\times10^{-5}$ | 0.091 | 1.978 | 2.471 |
| *CDK14* | −0.228 | $4.12\times10^{-5}$ | 0.091 | 0.976 | 1.207 |
| *SMIM4* | 0.202 | $4.15\times10^{-5}$ | 0.091 | 2.026 | 1.843 |
| *LOC102723362* | −0.227 | $4.24\times10^{-5}$ | 0.091 | 1.272 | 1.515 |
| *LOC105376789* | 0.211 | $4.38\times10^{-5}$ | 0.091 | 1.428 | 1.236 |
| *SNX32* | 0.207 | $4.99\times10^{-5}$ | 0.091 | 1.428 | 1.237 |
| *FAM19A3* | −0.258 | $5.20\times10^{-5}$ | 0.091 | 1.830 | 2.048 |
| *DAW1* | −0.153 | $5.81\times10^{-5}$ | 0.091 | 1.131 | 1.272 |
| *LINC00266-3* | −0.302 | $5.83\times10^{-5}$ | 0.091 | 1.600 | 1.882 |
| *ACTR5* | −0.295 | $5.97\times10^{-5}$ | 0.091 | 1.104 | 1.384 |
| *LOC105378404* | −0.213 | $6.17\times10^{-5}$ | 0.091 | 1.525 | 1.744 |
| *LINC00637* | 0.233 | $7.25\times10^{-5}$ | 0.091 | 1.717 | 1.488 |
| *NAE1* | −0.167 | $7.64\times10^{-5}$ | 0.091 | 1.010 | 1.193 |

| | | | | | |
|---|---|---|---|---|---|
| *SGOL1−AS1* | 0.158 | $7.94\times10^{-5}$ | 0.091 | 1.349 | 1.188 |
| *EMILIN2* | 0.201 | $8.00\times10^{-5}$ | 0.091 | 1.842 | 1.644 |
| *LMBR1* | −0.184 | $8.17\times10^{-5}$ | 0.091 | 1.190 | 1.382 |
| *RAB8B* | −0.207 | $8.39\times10^{-5}$ | 0.091 | 1.199 | 1.421 |
| *LOC105372482* | −0.197 | $8.41\times10^{-5}$ | 0.091 | 2.127 | 2.313 |
| *CRTC2* | 0.183 | $8.75\times10^{-5}$ | 0.091 | 1.469 | 1.282 |
| *LOC102724148* | −0.483 | $8.82\times10^{-5}$ | 0.091 | 1.840 | 2.289 |
| *LINCR−0002* | 0.223 | $8.91\times10^{-5}$ | 0.091 | 1.791 | 1.592 |
| *IGHM* | 0.314 | $9.17\times10^{-5}$ | 0.091 | 1.501 | 1.189 |
| *TNFSF13* | 0.231 | $9.37\times10^{-5}$ | 0.091 | 1.860 | 1.616 |
| *PKN2* | −0.161 | $9.42\times10^{-5}$ | 0.091 | 0.950 | 1.102 |
| *OR6C70* | −0.293 | $9.77\times10^{-5}$ | 0.091 | 1.082 | 1.376 |
| *LOC105373864* | −0.210 | $9.93\times10^{-5}$ | 0.091 | 1.227 | 1.427 |
| *LOC105371017* | −0.282 | $1.01\times10^{-4}$ | 0.091 | 2.284 | 2.522 |
| *LOC105374166* | −0.184 | $1.04\times10^{-4}$ | 0.091 | 0.875 | 1.057 |
| *TAF1L* | −0.250 | $1.14\times10^{-4}$ | 0.091 | 1.364 | 1.578 |
| *TRAJ48* | −0.398 | $1.16\times10^{-4}$ | 0.091 | 2.039 | 2.342 |
| *TSTD3* | −0.222 | $1.17\times10^{-4}$ | 0.091 | 1.118 | 1.359 |
| *NIPA2* | 0.193 | $1.19\times10^{-4}$ | 0.091 | 1.347 | 1.153 |
| *COL28A1* | −0.211 | $1.25\times10^{-4}$ | 0.091 | 1.190 | 1.376 |
| *BROX* | −0.132 | $1.28\times10^{-4}$ | 0.091 | 1.107 | 1.235 |
| *TRY2P* | 0.151 | $1.31\times10^{-4}$ | 0.091 | 1.052 | 0.914 |
| *LOC105371559* | −0.311 | $1.31\times10^{-4}$ | 0.091 | 1.236 | 1.520 |
| *HNRNPA3P1* | 0.308 | $1.32\times10^{-4}$ | 0.091 | 1.442 | 1.182 |
| *LOC105377043* | −0.261 | $1.34\times10^{-4}$ | 0.091 | 1.686 | 1.908 |
| *HINT1* | 0.183 | $1.35\times10^{-4}$ | 0.091 | 1.347 | 1.170 |
| *PDCD2L* | 0.218 | $1.37\times10^{-4}$ | 0.091 | 3.113 | 2.908 |
| *CRYBA2* | 0.184 | $1.38\times10^{-4}$ | 0.091 | 1.567 | 1.379 |
| *EARS2* | 0.194 | $1.45\times10^{-4}$ | 0.093 | 1.872 | 1.676 |

| | | | | | |
|---|---|---|---|---|---|
| *MIR4659A* | −0.420 | $1.50\times10^{-4}$ | 0.093 | 1.000 | 1.416 |
| *ASXL1* | −0.181 | $1.52\times10^{-4}$ | 0.093 | 1.316 | 1.501 |
| *MIR4534* | −0.353 | $1.57\times10^{-4}$ | 0.093 | 3.406 | 3.742 |
| *CLDN19* | 0.218 | $1.57\times10^{-4}$ | 0.093 | 2.130 | 1.910 |
| *LOC285422* | −0.260 | $1.59\times10^{-4}$ | 0.093 | 1.596 | 1.845 |
| *ANKRD20A4* | −0.306 | $1.65\times10^{-4}$ | 0.095 | 1.665 | 1.966 |
| *IVL* | −0.237 | $1.78\times10^{-4}$ | 0.099 | 1.831 | 2.037 |
| *TRGJP2* | −0.340 | $1.82\times10^{-4}$ | 0.099 | 2.524 | 2.797 |
| *LOC101927907* | −0.209 | $1.83\times10^{-4}$ | 0.099 | 3.484 | 3.665 |
| *LOC105371120* | −0.379 | $1.86\times10^{-4}$ | 0.099 | 1.438 | 1.811 |
| *ZNF669* | −0.229 | $1.92\times10^{-4}$ | 0.099 | 1.708 | 1.908 |
| *IGFL4* | 0.197 | $1.93\times10^{-4}$ | 0.099 | 1.553 | 1.378 |
| *FRMPD2* | 0.254 | $1.97\times10^{-4}$ | 0.100 | 3.778 | 3.520 |

**Table 9**. Conditional correlations between variables that indicate differences between 6−month−old patients with atopic dermatitis (AD) and controls.

| Variable 1 | Variable 2 | Interaction in controls | Interaction in patients | Difference |
|---|---|---|---|---|
| *LINC01036* (NR_126347) | *MIR4788* (NR_039951) | −0.117 | 0 | 0.117 |
| *Veillonella* | ko00311: Penicillin and cephalosporin biosynthesis | $-3.562\times10^{-3}$ | 0 | $3.562\times10^{-3}$ |
| *Raoultella* | *Cronobacter* | −0.676 | −0.507 | 0.168 |
| ko00906: Carotenoid biosynthesis | ko03018: RNA degradation | −0.154 | −0.064 | 0.091 |
| ko00906: Carotenoid biosynthesis | ko04066: HIF−1 signaling pathway | −0.118 | −0.094 | 0.023 |
| Gender | *C16orf72* (NM_014117) | $1.725\times10^{-12}$ | 0 | $1.725\times10^{-12}$ |
| Gender | *C16orf72* (NM_014117) | $2.177\times10^{-12}$ | 0 | $2.177\times10^{-12}$ |
| Delivery | *Clostridium_g4* | $1.275\times10^{-12}$ | 0 | $1.275\times10^{-12}$ |
| Family history | *Clostridium_g4* | 0 | $2.722\times10^{-12}$ | $2.722\times10^{-12}$ |
| Feeding | *Veillonella* | 0 | $1.010\times10^{-12}$ | $1.010\times10^{-12}$ |

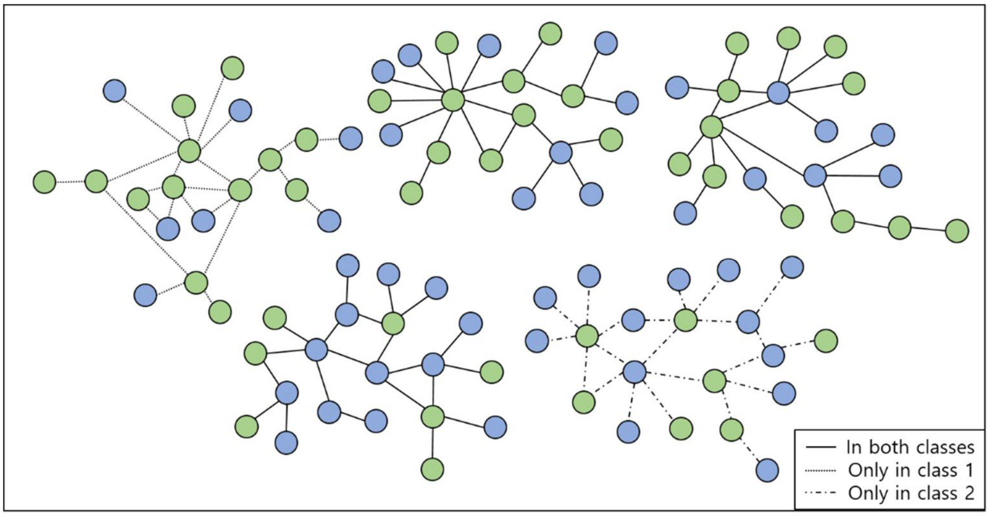

**Figure 1.** An example of simulated scale−free network for simulation data analysis. One of five exclusively connected networks exist only in the first class, and another one exists only in the second class. Blue circles denote numeric variables, and green circles mean categorical variables.

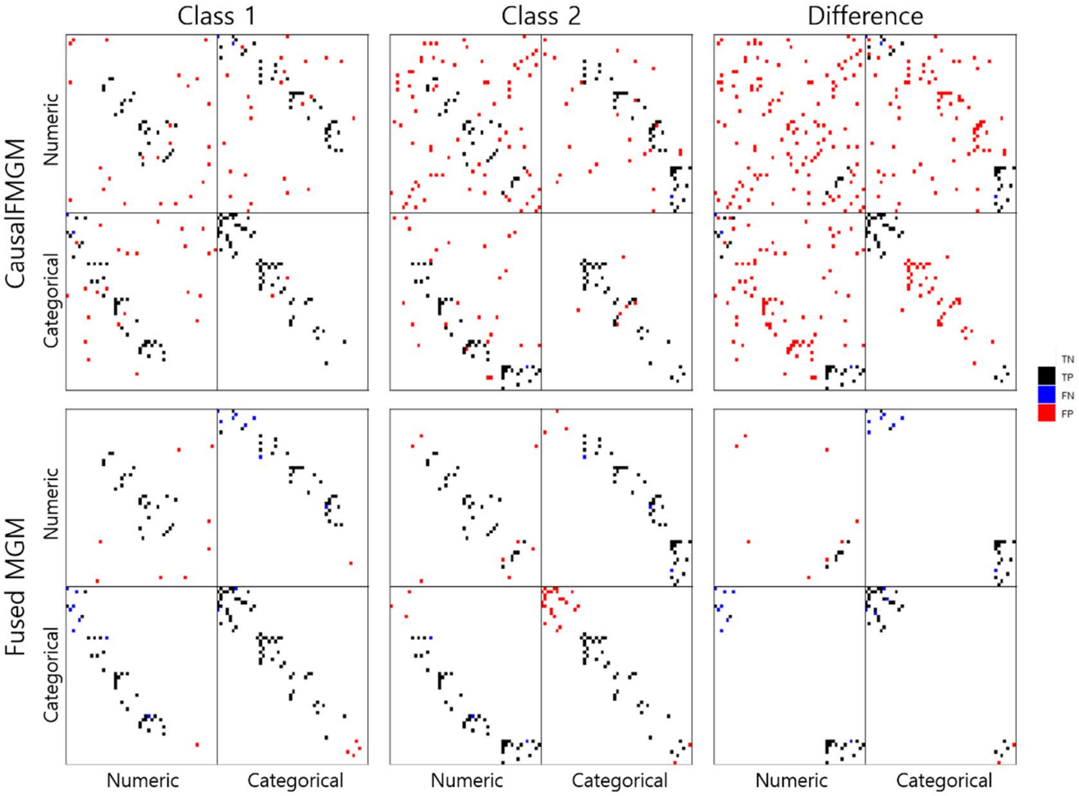

**Figure 2.** Overview of an example of the inference results from simulated datasets, with causalFMGM for each class and fused MGM. Inference results for each network and the difference are shown. Black dots represent true positives, while red and blue dots represent false positives and false negatives, respectively.

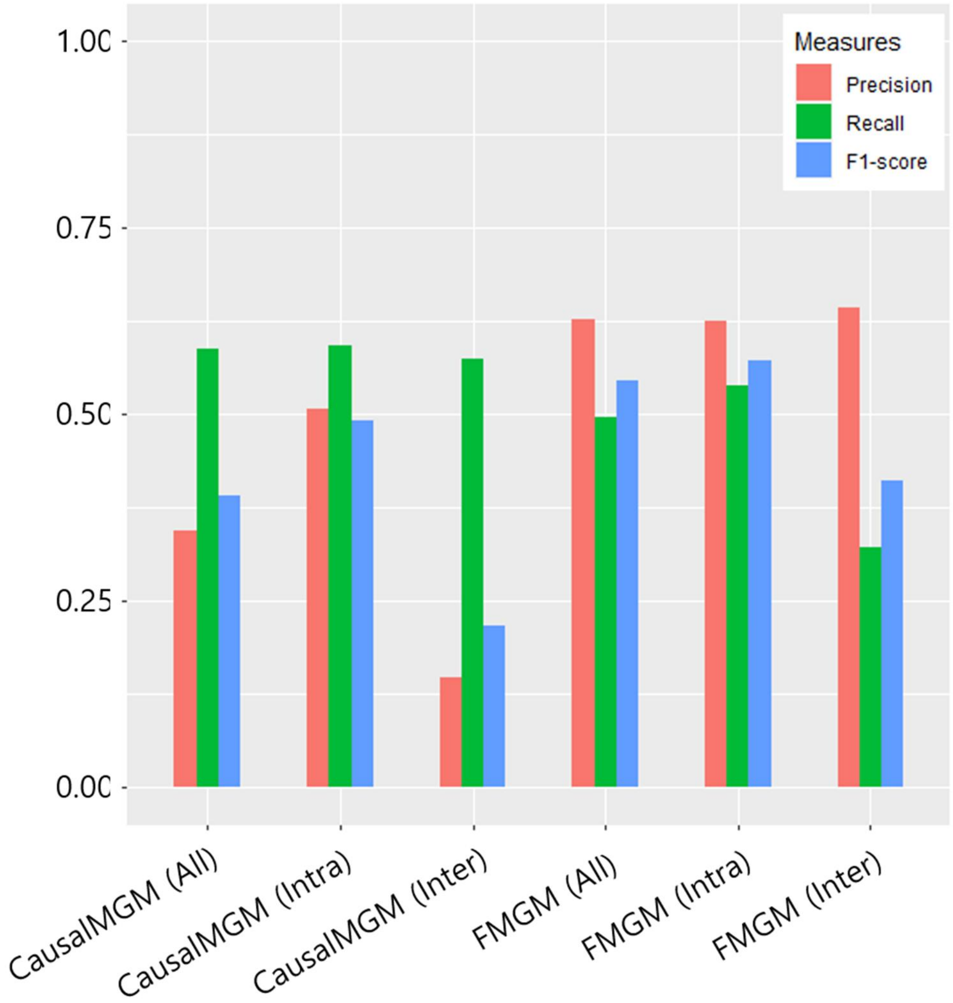

**Figure 3.** Overview of performance measures of causalMGM (left) and fused MGM (right) applied to simulated datasets. Precisions (red), recalls (green), and F1−scores (blue) are shown for overall inference (all), inference of networks (intra), and inference of differences (inter).

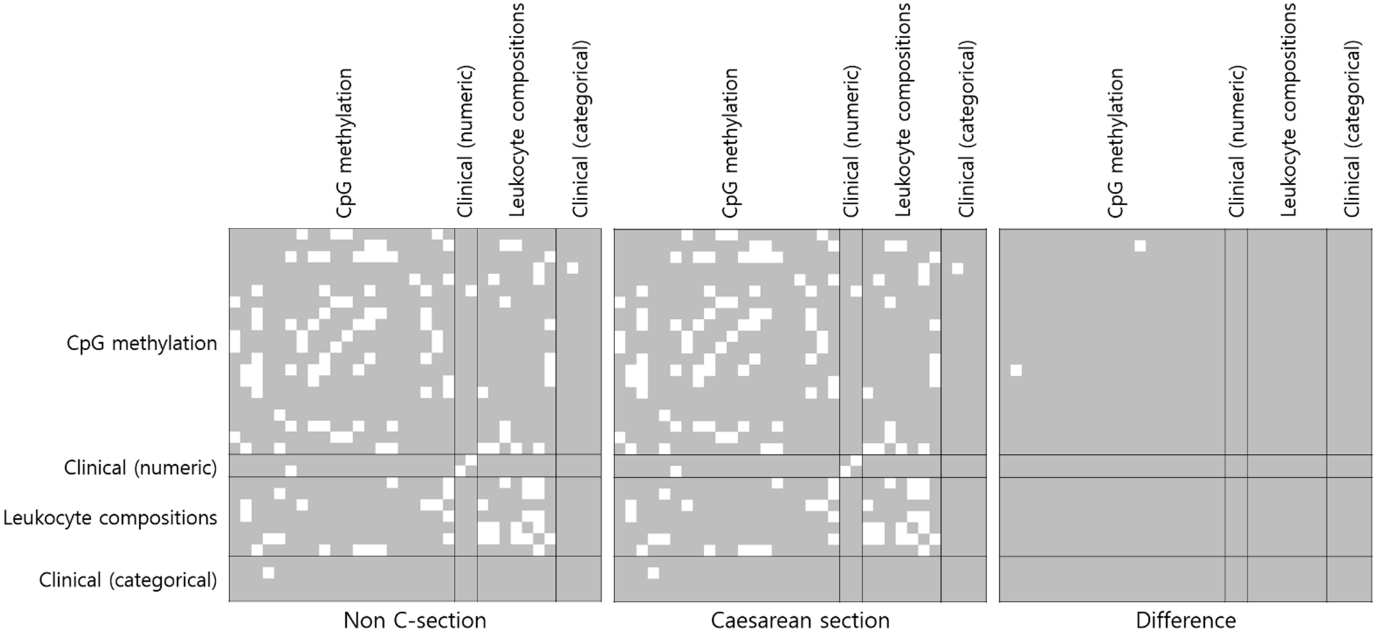

**Figure 4.** Overview of the inference result of networks of variables from mother−child pairs with or without caesarean section. White dots indicate non−zero edges or non−zero differences.

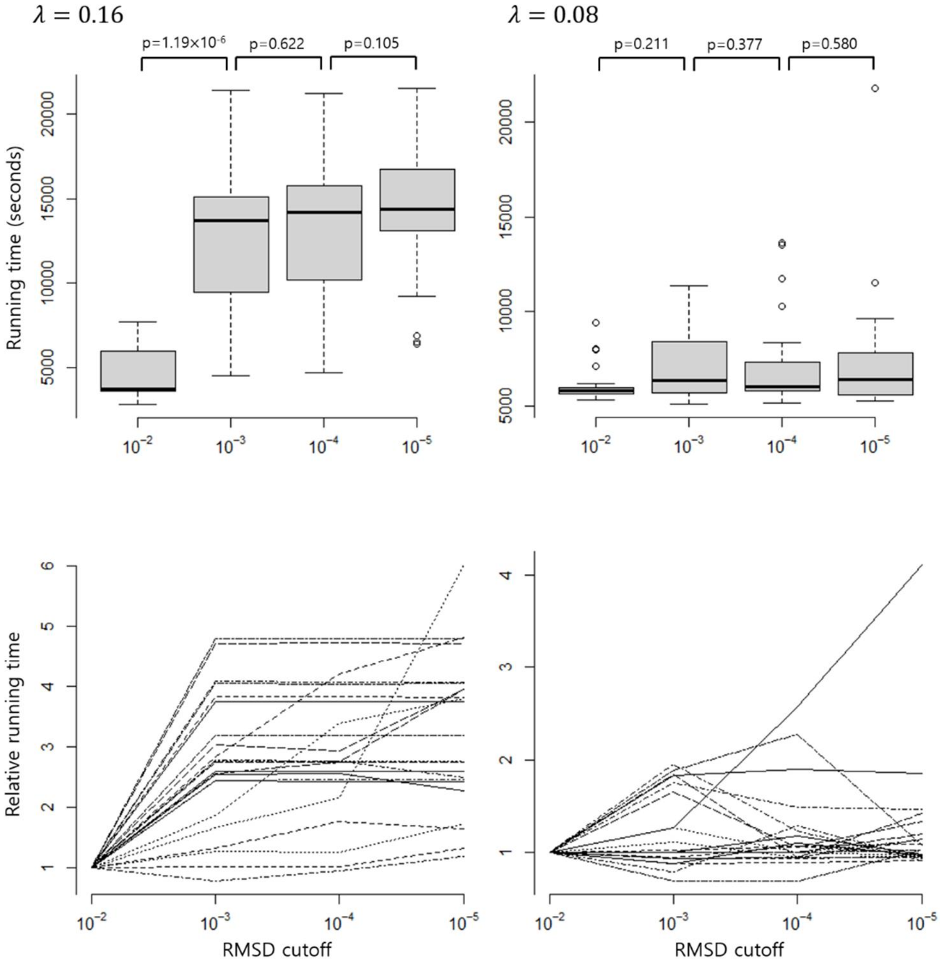

**Figure 5.** Boxplots of running times (in seconds) of FMGM with simulated datasets, with penalization parameters of 0.16 (left) and 0.08 (right), and spaghetti plots of relative running times. Times were calculated with four different rooted mean squared difference (RMSD) cutoffs of convergence in calculating $p_L(\Lambda)$, and the relative running times were calculated by dividing the actual times with the times from the loosest cutoff. Plots were calculated with 25 repetitions. The specs of the working machines were Intel Xeon 12Core 24 threads CPU with DDR4 16G RAM. P−values were calculated with Wilcoxon rank−sum tests of the differences (null hypothesis: the difference is zero).

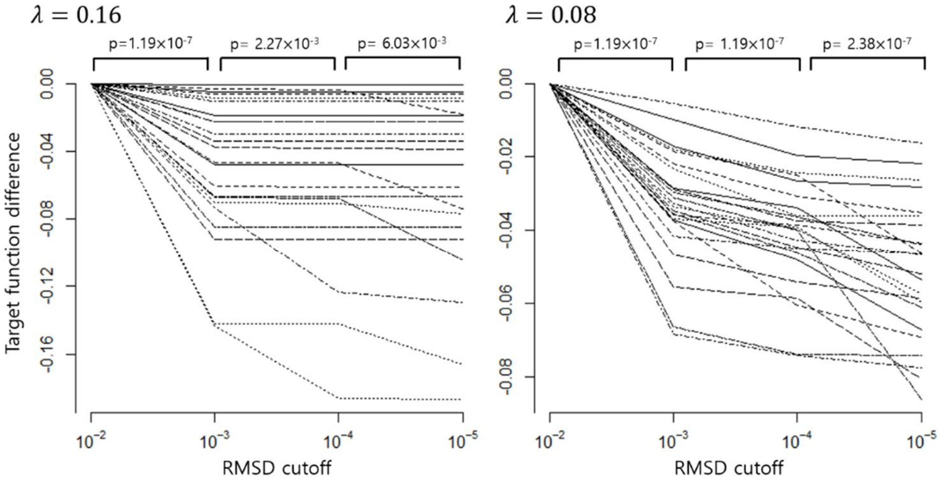

**Figure 6.** Spaghetti plots of differences of final values of target functions, calculated with four different rooted mean squared difference (RMSD) cutoffs of convergence in calculating $p_L(\Lambda)$. The differences were calculated by subtracting actual function values with the values from the loosest cutoff. Plots were calculated with 25 repetitions. P−values were calculated with Wilcoxon rank−sum tests of the differences (null hypothesis: the difference is zero).

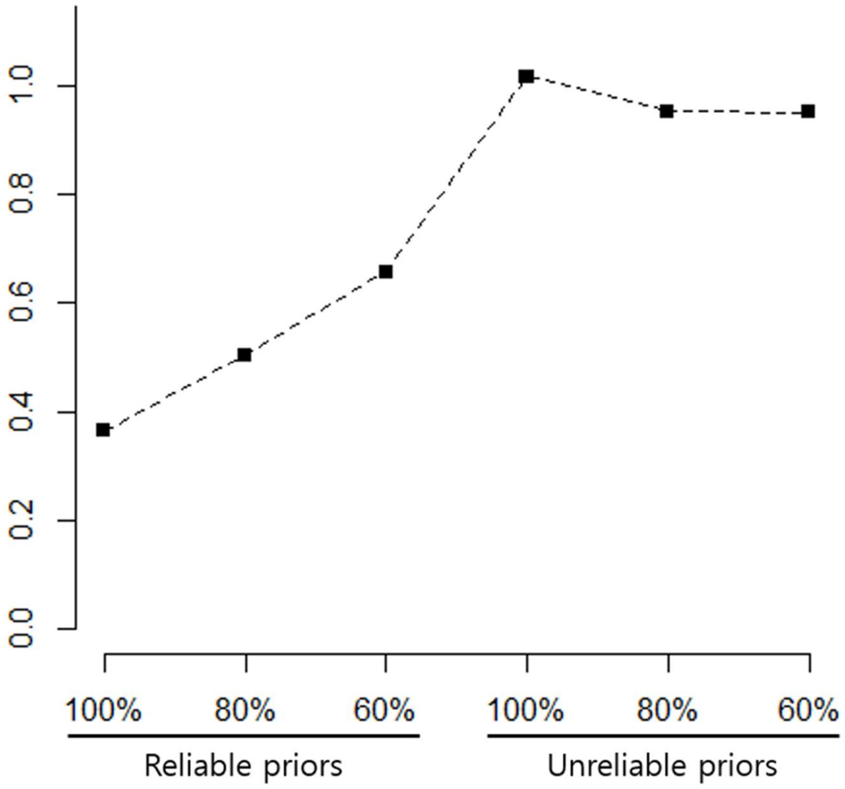

**Figure 7.** Average estimated confidences of prior sources from simulation. The confidences were averaged over all repetitions and groups. Percentages denote the proportions of prior edges included in the true networks. For the definition of ‘reliable’ and ‘unreliable’ priors, please refer to the main text section 3.3.

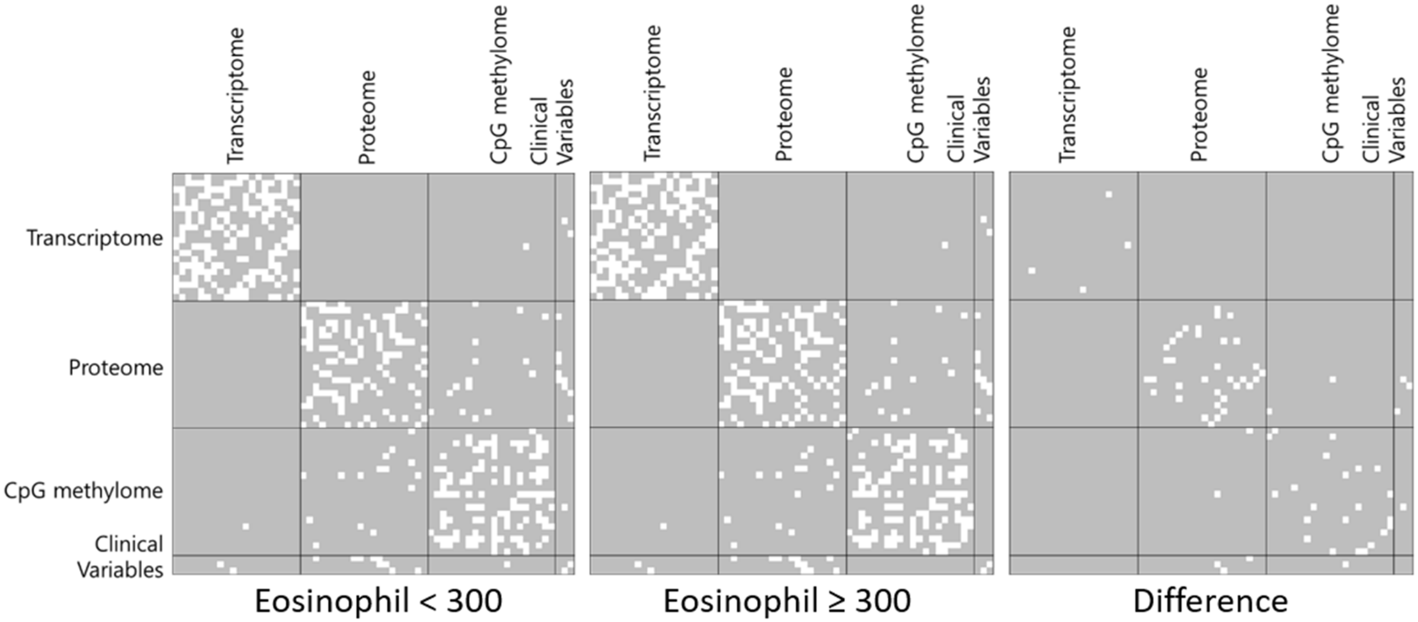

**Figure 8.** Overview of the inference result of networks underlying asthma patients with the eosinophil levels below or above 300, using multi−omics data and clinical variables. White dots indicate non−zero edges or non−zero differences.

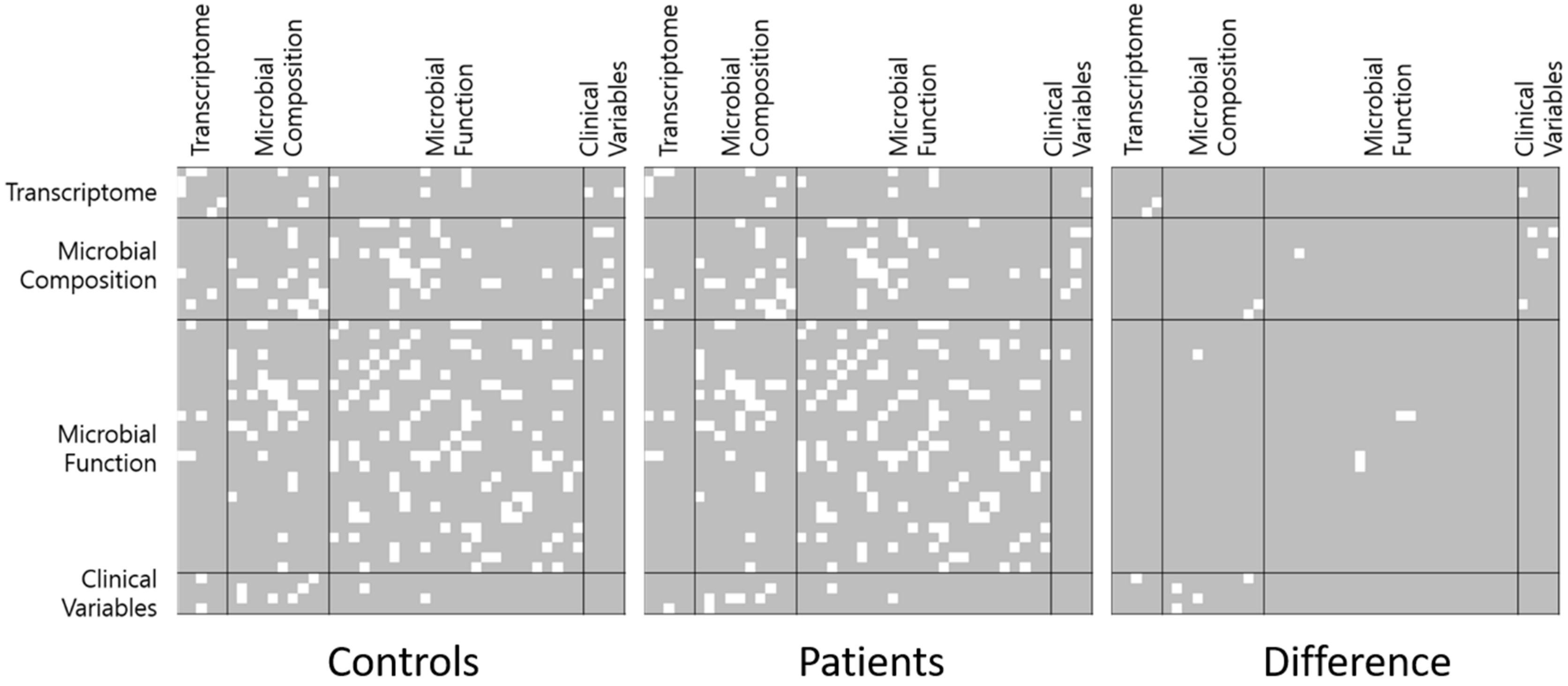

**Figure 9.** Overview of the inference result of networks underlying atopic dermatitis (AD) patients and controls among 6-month-old infants, using multi-omics data and clinical variables. White dots indicate non-zero edges or non-zero differences.