Sparse Gaussian Graphical Model with Missing Values

Shinsuke Uda, Hiroyuki Kubota
Medical Institute of Bioregulation, Kyushu University
Fukuoka, JAPAN
uda, kubota@bioreg.kyushu-u.ac.jp

***Abstract*—Recent advances in measurement technology have enabled us to measure various omic layers, such as genome, transcriptome, proteome, and metabolome layers. The demand for data analysis to determine the network structure of the interaction between molecular species is increasing. The Gaussian graphical model is one method of estimating the network structure. However, biological omics data sets tend to include missing values, which is conventionally handled by preprocessing. We propose a novel method by which to estimate the network structure together with missing values by combining a sparse graphical model and matrix factorization. The proposed method was validated by artificial data sets and was applied to a signal transduction data set as a test run.**

I. Introduction

Life phenomena involve molecular interactions in various layers, such as genome, transcriptome, proteome, and metabolome layers. Recent advances in measurement technology have enabled us to measure various omic layers, and the demand for data analysis is increasing. It is necessary to determine the network structure of the interaction between molecular species from the data.

The approaches to analyze such data sets are roughly classified into the knowledge based and the data driven method. The knowledge based method reconstructs the network structure from data base, which stores information on molecular interactions, and it is useful if database is available and reliable for the life phenomena of interest[1]. On the other hand, if database is not available or reliable, the data driven method is demanded. One of the main methods is information theoretical approach, which is roughly described as examining statistical independence or conditional independence between molecular species, and has been applied to data sets of various layers, such as transcriptome and metabolome [2], [3]. Whereas the information theoretic approach can be widely applied, because it is not necessary to assume specific probabilistic distributions, the following disadvantages may occur: it requires high computational costs and large samples size, and the construction of estimation model integrating prior knowledge is not easy. However, if we assume Gaussianity, the disadvantages can be diminished in stead of loss of generality with respect to probabilistic distributions.

The Gaussian graphical model [4] is adopted for estimating the network structure. Let $X_i, i = 1, \ldots, m$ be random variables that obey a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, where X_i corresponds to the measurement data of molecular species i. In the Gaussian graphical model, the edge appearance of a network between nodes i and j is determined by the conditional independence of X_i and X_j given $\{X_k | k \neq i, j, \}$ under the assumption of Gaussianity. In addition, let $\boldsymbol{\sigma}$ be the concentration matrix. Then, we have

$$X_i \perp X_j | X_{k \in \nu \setminus i,j} \Leftrightarrow \sigma^{ij} = 0$$

where $\nu = \{1, \ldots, m\}$ and $\nu \setminus i, j$ denotes a set in which i, j is removed from ν. If $\boldsymbol{\Sigma}$ is not degenerated, then $\boldsymbol{\sigma} = \boldsymbol{\Sigma}^{-1}$. Thus, if $S^{ij} \neq 0$, the edge appears between nodes i and j, and so we infer that there is an interaction between the molecular species corresponding to nodes i and j.

In biological data acquisition, obtaining large-sample-size data sets is difficult because of experimental costs and limitations. Thus, it is desirable to process data sets for which $n < m$. Furthermore, even if $n > m$, if n is close to m, determining the network structure using a straightforward method, such as the estimation of the inverse covariance matrix from data sets, is difficult because off-diagonal entries rarely become 0, due to the large variance of the estimator of σ^{ij}. However, if the concentration matrix has numerous 0 entries, i.e., if the concentration matrix is sparse, we can estimate it effectively by L1 norm regularization, even when $n < m$. The framework is referred to as a sparse graphical model or graphical lasso [5] and has been applied biological data sets[6], [7], [8].

Biological data sets, such as omics measurements, frequently have missing values. Conventionally, these missing values are excluded from data sets through preprocessing of each variable or sample. Alternatively, missing values are replaced with the mean value of the variable. However, estimating the concentration matrix and missing values together would yield a more accurate estimation of network structure, because there are relationships between the missing values and the concentration matrix. We herein propose a novel estimation method using a sparse graphical model, which simultaneously estimates missing values. In the proposed method, a partial correlation matrix is used instead of a concentration matrix. The partial correlation matrix is obtained by normalizing the concentration matrix for each entry as $\frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$.

The rest of this paper is organized as follows. In section 2, we will formulate the problem settings and describe an algorithm. The algorithm has some heuristics, thus, we examine the efficacy of the algorithm by a numerical experiment for cases where the exact network structure is known, and apply the algorithm to a biological data set. The numerical results will be given in section 3. In section 4, we describe conclusion.

ISSN 2305-7254

II. METHODS

A. Estimation model

A sample vector $\boldsymbol{x} \in \Re^m$ is identically and independently generated from the Gaussian distribution $\mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and the data sets $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are acquired. For simplicity, we set $\mu = 0$ without loss of generality. The partial correlation coefficient between x_i and x_j is denoted by ρ_{ij}, and the regression coefficient when x_i regresses to $\{x_j | j \neq i\}$ is denoted by b_{ij}. Using the relationships between ρ_{ij} and b_{ij}, such that $b_{ij} = \rho^{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}$, ρ_{ij} is obtained by the minimization problem

$$\min_{\boldsymbol{\rho}} \left\{ \sum_{i,j} \left(x_{ij} - \sum_{k \neq j} \rho^{jk} \sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}} x_{ik} \right)^2 \right\}, \tag{1}$$

where $\{\sigma_{ij}\}$ is the concentration matrix of $\boldsymbol{x}$. As such, the partial correlation matrix $\{\rho^{ij}\}$ is sparse, i.e., the elements of $\{\rho^{ij}\}$ have many 0s, we can replace the minimization problem of Eq. 1 with

$$\min_{\boldsymbol{\rho}} \left\{ \sum_{i,j} \left(x_{ij} - \sum_{k \neq j} \rho^{jk} \sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}} x_{ik} \right)^2 + \gamma \sum_{i>j} |\rho_{ij}| \right\} \tag{2}$$

by adding the regularization term of the L1 penalty [9] for ρ^{ij} [5], [6]. The hyperparameter γ regulates the sparsity of $\{\rho^{ij}\}$. As γ increases, $\{\rho^{ij}\}$ becomes more sparse. In the case of $m > n$, obtaining $\{\rho^{ij}\}$ by the inverse of covariance matrix or Eq. 1 is difficult. However, Eq. 2 can be solved under the assumption of sparsity for $\{\rho^{ij}\}$. Generally, interactions between molecular species in biology are sparse rather than dense, because one molecular species appears to interact with at most tens of other molecular species, whereas the total number of molecular species is in the thousands.

We assume that the data matrix $\boldsymbol{X}$, which consists of the data set $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, includes the missing values generated at rate r_{miss}. Each element of the data matrix is discriminated as an observed or missing value

$$x_{ij} = \begin{cases} y_{ij} & \text{for } i,j \in \mathcal{D} \text{ (observed)} \\ z_{ij} & \text{for } i,j \in \bar{\mathcal{D}} \text{ (missed)}, \end{cases}$$

where $\mathcal{D}$ represents the index set of elements of observed values. Here, $\boldsymbol{X}$ is reproduced as the product of two matrices as

$$\boldsymbol{X} = \boldsymbol{P}\boldsymbol{Q}^T$$

where the sizes of $\boldsymbol{P}$ and $\boldsymbol{Q}$ are $n \times k, m \times k$, respectively. Here, $\boldsymbol{P}$ and $\boldsymbol{Q}$ are estimated by minimizing the sum of the squared errors between the observed values and the elements of the matrix product:

$$\min_{\boldsymbol{p},\boldsymbol{q}} \sum_{i,j \in \mathcal{D}} \frac{1}{2} \left(y_{ij} - \sum_h p_{ih} q_{jh} \right)^2. \tag{3}$$

The missing value is estimated as

$$z_{ij} = \sum_h p_{ih} q_{jh}$$

for $i,j \in \bar{\mathcal{D}}$. The estimation technique of the missing value is based on matrix factorization applied to a collaboration filter [10].

The dimensionality of k is determined by the rank of $\boldsymbol{X}$. As such, $\boldsymbol{X}$ has strong correlations between variables, and the distribution of $\boldsymbol{x}$ is biased to the subspace, so that $\boldsymbol{X}$ tends to have small singular values. In practice, generally, data matrices of biological data usually have small singular values, because the molecular species form clusters, which have similar measurement values. Thus, we consider that $\boldsymbol{X}$ can be approximated by a low-rank matrix in biological data analysis.

We propose a novel method for estimating a sparse partial correlation matrix with missing values by combining a sparse graphical model and matrix factorization. The proposed method corresponds to approximately solving the following minimization problem:

$$\begin{aligned} \min_{\boldsymbol{p},\boldsymbol{q},\boldsymbol{\rho}} \Bigg\{ & \sum_{i,j \in \mathcal{D}} \frac{1}{2} \left(y_{ij} - \sum_h p_{ih} q_{jh} \right)^2 \\ & + \frac{\beta}{2} \sum_{i,j} \left(x_{ij} - \sum_{k \neq j} \rho^{jk} \sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}} x_{ik} \right)^2 + \gamma \sum_{i>j} |\rho_{ij}| \Bigg\} \\ & \text{subject to } \; x_{ij} = \sum_h p_{ih} q_{jh} \quad \text{for } \; i,j \in \bar{\mathcal{D}}. \end{aligned} \tag{4}$$

The second and third terms of Eq. 4 are interpreted as regularization terms that depend on $\boldsymbol{\rho}$ for Eq. 3.

B. Interpretation as a probabilistic model

The proposed method can be interpreted as the maximum a posteriori (MAP) estimation by a probabilistic model. The posterior distribution of $\boldsymbol{\rho}, \boldsymbol{p}, \boldsymbol{q}, \boldsymbol{z}$ given observed data $\boldsymbol{y}$ is

$$p(\boldsymbol{\rho}, \boldsymbol{p}, \boldsymbol{q}, \boldsymbol{z} | \boldsymbol{y}) \;\; = \;\; p(\boldsymbol{\rho} | \boldsymbol{p}, \boldsymbol{q}, \boldsymbol{x}) p(\boldsymbol{p}, \boldsymbol{q} | \boldsymbol{y}) p(\boldsymbol{z}). \tag{5}$$

By Bayes theorem,

$$\begin{aligned} p(\boldsymbol{p}, \boldsymbol{q} | \boldsymbol{y}) \;\; & \propto \;\; p(\boldsymbol{y} | \boldsymbol{p}, \boldsymbol{q}) p(\boldsymbol{p}, \boldsymbol{q}) \\ & \propto \;\; p(\boldsymbol{y} | \boldsymbol{p}, \boldsymbol{q}) \\ & \propto \;\; \prod_{i,j \in \mathcal{D}} \exp \left[-\frac{1}{2\sigma_y^2} (y_{ij} - \sum_h p_{ih} q_{jh})^2 \right] \end{aligned}$$

where $p(\boldsymbol{p}, \boldsymbol{q}) =$ constant. We set $\sigma_y^2 = 1$ without loss of generality. Moreover,

$$p(\boldsymbol{\rho} | \boldsymbol{p}, \boldsymbol{q}, \boldsymbol{x}) \;\; \propto \;\; p(\boldsymbol{x} | \boldsymbol{\rho}, \boldsymbol{p}, \boldsymbol{q}) p(\boldsymbol{\rho}).$$

The likelihood function is approximated by

$$\begin{aligned} & p(\boldsymbol{x} | \boldsymbol{\rho}, \boldsymbol{p}, \boldsymbol{q}) \approx \textstyle\prod_{i>j}^{n,m} p(x_{ij} | \{x_{ik}\}_{k \neq j}, \{\rho^{jk}\}_{k \neq j}) \\ & \propto \textstyle\prod_{i>j}^{n,m} \exp \left[-\frac{1}{2V_1} \left(x_{ij} - \sum_{k \neq j} \rho^{jk} \sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}} x_{ik} \right)^2 \right] \quad . \end{aligned}$$

The prior distribution is set as

$$\begin{aligned} p(\rho^{ij}) \;\; & = \;\; \frac{1}{2\gamma_{ij}} \exp[-\gamma |\rho^{ij}|] \\ p(\boldsymbol{z}) \;\; & = \;\; \prod_{i,j \in \bar{\mathcal{D}}} \delta \left(z_{ij} - \sum_h p_{ih} q_{jh} \right). \end{aligned}$$

Thus, by Eq. 6,

$$p(\boldsymbol{\rho},\boldsymbol{p},\boldsymbol{q},\boldsymbol{z}|\boldsymbol{y}) \approx \frac{1}{Z}\exp\left[-\frac{1}{2}\sum_{i,j\in\mathcal{D}}(y_{ij}-\sum_h p_{ih}q_{jh})^2 - \frac{1}{2V_1}\sum_{i>j}^{n,m}\left(x_{ij}-\sum_{k\neq j}\rho^{jk}\sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}}x_{ik}\right)^2 - \gamma\sum_{i>j}^{n,m}|\rho^{ij}|\right] \cdot \prod_{i,j\in\bar{\mathcal{D}}}\delta\left(z_{ij}-\sum_h p_{ih}q_{jh}\right) \quad (6)$$

is obtained, where Z is the normalization constant. The MAP estimation of Eq. 6 corresponds to the minimization problem of Eq. 4.

C. Algorithm

The minimization problem of Eq. 4 is rewritten as

$$\min_{\boldsymbol{p},\boldsymbol{q},\boldsymbol{\rho}}\left\{\sum_{i,j\in\mathcal{D}}\frac{1}{2}\left(y_{ij}-\sum_h p_{ih}q_{jh}\right)^2 + \frac{\beta}{2}\sum_{i,j}\left(x_{ij}-\sum_{k\neq j}\rho^{jk}\sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}}x_{ik}\right)^2 + \gamma\sum_{i>j}|\rho_{ij}|\right\} + \sum_{i,j\in\bar{\mathcal{D}}}\lambda_{i,j}\left(x_{ij}-\sum_h p_{ih}q_{jh}\right) \quad (7)$$

by the Lagrange multiplier. The minimization problem of Eq. 7 is divided into two subproblems. By fixing $\boldsymbol{\rho}$ of Eq. 7, we define

$$f(\boldsymbol{p},\boldsymbol{q};\boldsymbol{\rho}) = \sum_{i,j\in\mathcal{D}}\frac{1}{2}\left(y_{ij}-\sum_h p_{ih}q_{jh}\right)^2 + \frac{\beta}{2}\sum_{i,j}\left(x_{ij}-\sum_{k\neq j}\rho^{jk}\sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}}x_{ik}\right)^2. \quad (8)$$

By fixing $\boldsymbol{p},\boldsymbol{q}$, we define

$$g(\boldsymbol{\rho};\boldsymbol{p},\boldsymbol{q}) = \frac{1}{2}\sum_{i,j}\left(x_{ij}-\sum_{k\neq j}\rho^{jk}\sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}}x_{ik}\right)^2 + \gamma\sum_{i>j}|\rho_{ij}|. \quad (9)$$

The minimization problem of Eq. 7 is replaced by the following two minimization subproblems:

$$\min_{\boldsymbol{p},\boldsymbol{q}} \quad f(\boldsymbol{p},\boldsymbol{q};\boldsymbol{\rho}), \quad (10)$$

$$\min_{\boldsymbol{\rho}} \quad g(\boldsymbol{\rho};\boldsymbol{p},\boldsymbol{q}), \quad (11)$$

which are solved numerically in alternate shifts like EM algorithm until convergence or upper limit number of iteration. By differentiating the objective function of the minimization problem given by Eq. 7 with respect to λ_{ij}, we obtain

$$z_{ij} = \sum_h p_{ih}q_{jh}, \quad (12)$$

and substitute the values into Eq.8.

The minimization subproblem given by Eq.10 is numerically solved by the stochastic gradient method due to local minima. We define

$$\epsilon_{ij} = \frac{1}{2}\left(y_{ij}-\sum_h p_{ih}q_{jh}\right)^2 + \frac{\beta}{2}\left(x_{ij}-\sum_{k\neq j}\rho^{jk}\sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}}x_{ik}\right)^2 + \gamma|\rho_{ij}|.$$

When $\{i,k\}\in\mathcal{D}$ for $k=1,\ldots,m$,

$$\frac{\partial\epsilon_{ij}}{\partial p_{il}} = -\left(y_{ij}-\sum_h p_{ih}q_{jh}\right)q_{jl},$$

when $\{i,k\}\in\bar{\mathcal{D}}$ for $\exists k\neq j, \{i,j\}\in\mathcal{D}$,

$$\frac{\partial\epsilon_{ij}}{\partial p_{il}} = -\left(y_{ij}-\sum_h p_{ih}q_{jh}\right)q_{jl} - \beta\left(y_{ij}-\sum_{k\neq j}\rho^{jk}\sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}}x_{ik}\right)\cdot\left(\sum_{k\neq j,(i,k)\in\bar{\mathcal{D}}}\sum_{k\neq j}\rho^{jk}\sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}}q_{kl}\right),$$

when $\{i,j\}\in\bar{\mathcal{D}}$,

$$\frac{\partial\epsilon_{ij}}{\partial q_{jb}} = -\left(y_{ij}-\sum_h p_{ih}q_{jh}\right)p_{ib},$$

when $a\neq j, \{i,a\}\in\bar{\mathcal{D}}$,

$$\frac{\partial\epsilon_{ij}}{\partial q_{ab}} = -\beta\left(y_{ij}-\sum_{k\neq j}\rho^{jk}\sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}}x_{ik}\right)\rho^{ja}\sqrt{\frac{\sigma^{aa}}{\sigma^{jj}}}p_{ib}.$$

The derivatives yield the updated formula

$$p_{il}^{(t+1)} := p_{il}^{(t)} + \eta\frac{\partial\epsilon_{ij}^{2,(t)}}{\partial p_{il}}$$

$$q_{ab}^{(t+1)} := q_{ab}^{(t)} + \eta\frac{\partial\epsilon_{ij}^{2,(t)}}{\partial q_{ab}}.$$

for the minimization subproblem of Eq. 10, where η is a step size.

The minimization subproblem is numerically solved by the shooting method [11]. We define

$$\check{x} = (x_{1,1},\ldots,x_{n,1},x_{1,2},\ldots,x_{n,2},\cdots,x_{1,m},\ldots,x_{m,m})^T$$

and column vector

$$\chi_{(i,j)} = \left(0,\ldots,0,\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\boldsymbol{X}_{1:n,j}^T,0,\ldots,0,\sqrt{\frac{\sigma^{ii}}{\sigma^{jj}}}\boldsymbol{X}_{1:n,i}^T,0\ldots,0\right)^T,$$

which has $\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\boldsymbol{X}_{1:n,j}^T$ and $\sqrt{\frac{\sigma^{ii}}{\sigma^{jj}}}\boldsymbol{X}_{1:n,i}^T$ at the ith and jth blocks, respectively, where $\boldsymbol{X}_{1:n,i} = (x_{1,i}, \ldots, x_{n,i})^T$. Let $\boldsymbol{\chi}$ be nm by the $m(m-1)/2$ matrix denoted by

$$\boldsymbol{\chi} = \left(\chi_{(1,2)}, \ldots, \chi_{(m-1,m)}\right).$$

The minimization subproblem of Eq. 11 is rewritten as

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\|\check{x} - \boldsymbol{\chi}\boldsymbol{\theta}\|_2^2 + \gamma|\boldsymbol{\theta}|$$

as linear regression with the L1 regularization term, where $\boldsymbol{\theta} = (\rho^{1,2}, \ldots, \rho^{m-1,m})^T$. The initial value of $\boldsymbol{\theta}$ is set by

$$\begin{aligned}\theta_j^{(0)} &= \arg\min_{\theta_j} \frac{1}{2}\|\check{x} - \theta_j\chi_j\|_2^2 + \gamma|\theta_j| \\ &= \text{sign}\left(\check{x}^T\chi_j\right)\frac{\left(|\check{x}^T\chi_j| - \gamma\right)_+}{\chi_j^T\chi_j}\end{aligned}$$

where $\boldsymbol{\chi}_j$ is the jth column vector of $\boldsymbol{\chi}$ and $(a)_+ = aI(a > 0)$. The updated formula for the minimization subproblem of Eq. 11 is derived by the shooting method [11], as follows:

$$\begin{aligned}\theta_j^{(t+1)} &= \arg\min_{\theta_j} \frac{1}{2}\|\check{x} - \sum_{i\neq j}\theta_i^{(t)}\chi_i - \theta_j\chi_j\|_2^2 + \gamma|\theta_j| \\ &= \text{sign}\left(\frac{(\epsilon_\theta^{(t)})^T\chi_j}{\chi_j^T\chi_j} + \theta_j^{(t)}\right)\left(\left|\frac{(\epsilon_\epsilon^{(t)})^T\chi_j}{\chi_j^T\chi_j} + \theta_j^{(t)}\right| - \frac{\gamma}{\chi_j^T\chi_j}\right)_+ \end{aligned} \quad (13)$$

where $\boldsymbol{\epsilon}_\theta^{(t)} = \check{x} - \boldsymbol{\chi}\boldsymbol{\theta}^{(t)}$. Here, σ^{ii} is estimated using the sum of the squared errors

$$\sigma^{ii} = \left(\frac{1}{n}\left\|\chi_i - \sum_{j\neq i}\theta_{ij}\chi_j\right\|_2^2\right)^{-1}.$$

If the each solution of minimization subproblems of Eq. 10 and 11 decreases the objective function of Eq. 4, the objective function sequentially decrease. However, the solution of minimization subproblems of Eq. 10 dose not always decrease the objective function because of the stochastic gradient method. Thus, the objective function does not always decrease.

The pseudo code for the proposed method is denoted in algorithm1.

Algorithm 1

$\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{\rho}, \boldsymbol{\sigma} \Leftarrow$ Initial values
repeat
  for i, l **do**
    $\boldsymbol{p}_{il} \Leftarrow \boldsymbol{p}_{il} + r\frac{\partial\epsilon_{ij}}{\partial p_{il}}$
    update $\boldsymbol{\epsilon}$
  end for
  for a, b **do**
    $\boldsymbol{q}_{ab} \Leftarrow \boldsymbol{q}_{ab} + r\frac{\partial\epsilon_{ij}}{\partial p_{ab}}$
    update $\boldsymbol{\epsilon}$
  end for
  for $i > j$ **do**
    update $\rho_{i,j}$ by Eq. 13
    $\sigma^{ii} \Leftarrow \left(\frac{1}{n}\left\|\chi_i - \sum_{j\neq i}\theta_{ij}\chi_j\right\|_2^2\right)^{-1}$
  end for
until convergence for $\boldsymbol{\rho}$ or upper limit number of iteration

III. Results

A. Artificial data

We evaluated the performance of the proposed method using artificial data. The data matrix $\boldsymbol{X}$ is generated from a normal distribution with a mean of $\mathbf{0}$, which has a sparse partial correlation matrix, and about $m/2$ of its singular values are approximately less than $1/2$ of maximum singular value. After generating a complete data matrix, which has no missing values, we replaced the small rnm values of the complete data matrix with missing values, where r is the ratio of missing values to the total number nm of entries of $\boldsymbol{X}$. In the biological data measurement, small values tend to be missing because they have a smaller signal-to-noise ratio than large values.

We determined $m = 20, n = 10$. For the partial correlation matrix of the generative normal distribution, the ratio of nonzero off-diagonal entries to all off-diagonal entries was 0.3. The nonzero or zero off-diagonal entries were randomly determined using a binomial distribution. Each nonzero off-diagonal entry was generated from a uniform distribution with domain $[-1, -0.5] \cup [0.5, 1]$, and the non-zero entries were rescaled in order to assure positive definiteness. We heuristically tuned the partial correlation matrix, so that the ratio of nonzero off-diagonal entries was 0.3 and the ratio of the lower 50% singular values was small. (i) We then set the lower 50% singular values at small values and reconstruct the partial correlation matrix by singular value decomposition. (ii) Each off-diagonal entry, which was changed from zero to nonzero by (i), was set at zero. Here, (i) and (ii) were iterated until the conditions were satisfied. We conducted a numerical experiment involving 560 trials in order to estimate the network structure, i.e., to estimate the nonzero coefficient of partial correlation, and evaluated the performance of estimation by F_1 score between the estimated network structure and the true network structure. The F_1 score is the harmonic mean of precision and recall and is $[0, 1]$. If the F_1 score is 1, the estimated network structure is the same as the true network structure.

We compared two methods with the proposed methods (Fig. 1,2,3,4). First, the missing values and the partial correlation matrix are estimated separately (separate method). The missing values are estimated from $\boldsymbol{P}, \boldsymbol{Q}$ using Eq. 3, and the network structure is estimated using Eq. 2. Second, the missing values are simply replaced with the mean of each variable, and the network structure is estimated using Eq. 2 (mean imputation). We set $k = 10$ for each numerical experiment.

Fig. 1 shows the average F_1 score of 560 trials, when the hyperparameter γ was determined by a grid search of 100 points, so that the $F1$ score was maximized. Thus, the F_1 score is interpreted as the potential performance when γ is most successfully selected. The proposed method provides a higher F_1 score over all missing rates. The F_1 scores for all methods tend to decrease, as the missing rate increases. In some cases, the F_1 scores were larger than those of the complete matrix. This suggests that the estimated missing values may become useful for partial correlation estimation, because the proposed method simultaneously estimates the missing values and the partial correlation matrix. As another possibility, the bias of the

estimated missing values is unexpectedly advantageous for the partial correlation matrix, particularly for the separate method and mean imputation for a missing rate of 0.1. However, we consider that it would generally be difficult to select a γ that exceeds the performance of the complete data.

In practice, γ would be selected by an evaluation function, e.g., information criteria for model selection. We selected γ so as to minimize the sum of the squared errors with the L1 regularization term

$$F(\gamma) \equiv \sum_{i,j} \left(x_{ij} - \sum_{k \neq j} \rho^{jk} \sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}} x_{ik} \right)^2 + \gamma \sum_{i>j} |\rho_{ij}|, \quad (14)$$

which corresponds to a log likelihood function with a prior distribution function (Fig. 2). Although minimizing $F(\gamma)$ is an ad hoc solution, the proposed method provides an F_1 score close to the maximum F_1 score (Fig.1) around a missing rate of $0.1 \sim 0.3$ and a higher F_1 score than for the separate method and mean imputation. In addition, the minimization of $F(\gamma)$ is interpreted as the particular case of type II maximum likelihood estimation, the posterior distribution of which is highly concentrated in the mode.

On the other hand, the BIC-type criterion ([6]) is defined as

$$\begin{aligned} \mathrm{BIC}(\gamma) \quad &= \sum_{j}^{m} n \log \left[\sum_{i} \left(x_{ij} - \sum_{k \neq j} \rho^{jk} \sqrt{\frac{\sigma^{kk}}{\sigma^{jj}}} x_{ik} \right)^2 \right] \\ + \quad &\log n \cdot \sum_{k \neq j} I(\rho^{j,k} \neq 0), \end{aligned}$$

and the F_1 score selected by minimizing the BIC-type criterion (Fig. 3) is lower than the selection by $F(\gamma)$ (Fig. 2).

The root mean squared error (RMSE) between the estimated missing values and the values of complete data of the proposed method is smaller than that of the separate method, except for $r_{\mathrm{miss}} = 0.1$ (Fig. 4). Whereas the RMSE of the mean imputation is smaller than the proposed method and the separate method, the F_1 score of the mean imputation is lower than those, possibly due to substituting same values to missing entries by each variable. Although the RMSE of the separate method is slightly smaller than that of the proposed method at $r_{\mathrm{miss}} = 0.1$, the F_1 score of the proposed method is higher than that of the separate method for the maximum F_1 score (Fig. 1) and selection by Eq. 14 (Fig. 2). These suggest that the smaller RMSE does not necessarily increase the accuracy of the network structure. Furthermore, in the minimization subproblem of Eq. 10, the smaller RMSE between the estimated observed values and the observed values is not necessary for network structure estimation, because the second term of Eq. 8 is regarded as the regularization term for the minimization of the RMSE. This suggests that reconstruction of the data matrix to more closely recover the complete data set is not necessary in order to increase the accuracy of network structure estimation.

B. Signal transduction pathway

The network structure of interaction between molecular species was estimated from the data set [12], which consists

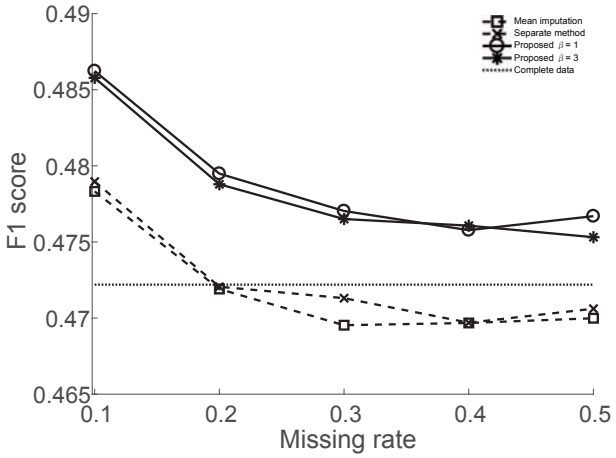

Fig. 1. Maximum F_1 score for γ selection. The average of 560 trials is shown. The solid lines with circle and asterisk symbols show the data obtained by the proposed method for $\beta = 1$ and $\beta = 3$, respectively. The dashed lines with x and square symbols show the data obtained by the separate method and mean imputation, respectively. The dotted line shows the data of the complete data set.

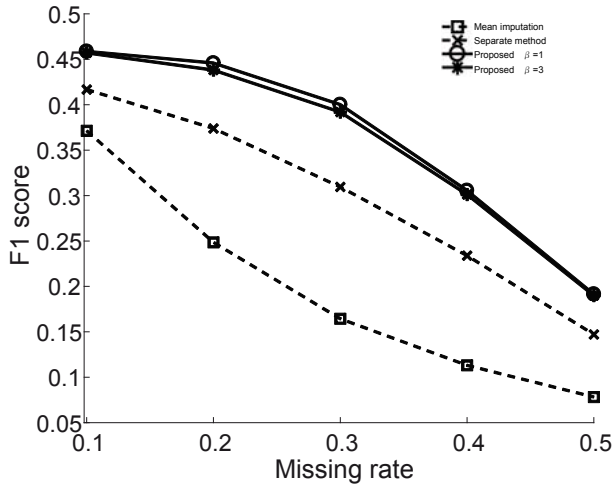

Fig. 2. F_1 score for γ selection by $F(\gamma)$. The average of 560 trials is shown. The solid lines with circle and asterisk symbols show the data obtained by the proposed method for $\beta = 1$ and $\beta = 3$, respectively. The dashed lines with x and square symbols show the data obtained by the separate method and mean imputation, respectively.

of the phosphorylation intensity of the signal transduction molecule: ERK, CREB, p38, and JNK, and the expression intensity of the gene product: c-FOS, EGR1, c-JUN, JUNB, and FOSB, by the proposed methods. The phosphorylation and expression intensities were measured at 60 time points from 0 to 177 minutes in three-minute intervals after stimulating PC 12 cells at 0 minutes by each growth factor, namely, EGF (5 ng/ml, 0.5 ng/ml), NGF (5 ng/ml, 0.5 ng/ml), PACAP (100 nM, 1 nM), or Anisomycin (50 ng/ml). Thus, data sets consisting of 420 measurement conditions and nine molecular species were obtained by combining time point and stimulus data. The nine measured molecular species are considered to be involved in the ERK pathway.

The data set was measured by quantitative image cytometry

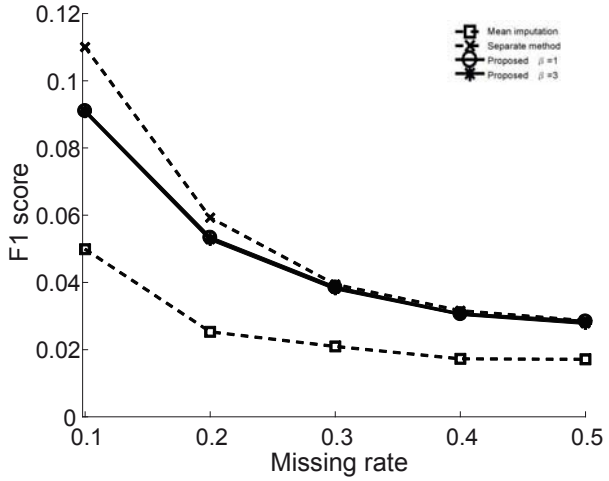

Fig. 3. F_1 score for γ selection by the BIC-type criterion. The average of 560 trials is shown. The solid lines with circle and asterisk symbols show the data obtained by the proposed method for $\beta = 1$ and $\beta = 3$, respectively. The dashed lines with x and square symbols show the data obtained by the separate method and mean imputation, respectively.

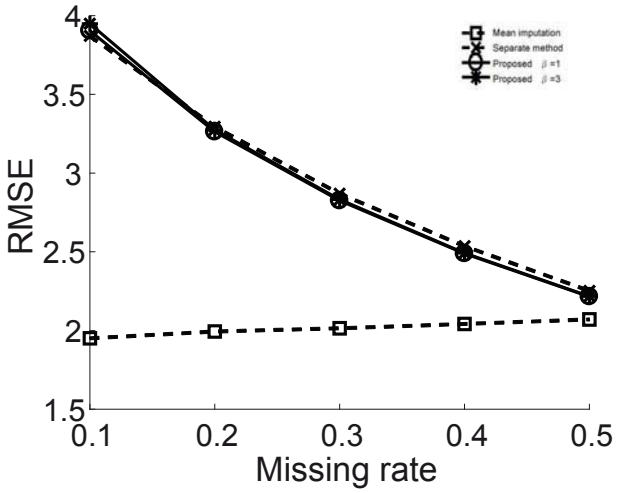

Fig. 4. Root mean squared error between the estimated missing values and the complete data set for the γ selection by $F(\gamma)$. The average of 560 trials is shown. The solid lines with circle and asterisk symbols show the data obtained by the proposed method for $\beta = 1$ and $\beta = 3$, respectively. The dashed line with x and square are separate method and mean imputation, respectively

(QIC) [13]. A data set with a large sample size can be obtained by QIC, because QIC semi-automatically measures samples by liquid-handling robots and processes images of immunostaining. However, data acquisition involving a large sample size by a conventional biological experiment is difficult and expensive. In particular, for the sample size in omics measurement, it is likely that $n < m$. We used the data set on trial to evaluate the performance of proposed method. Thus, we generated a $n = 7, m = 9$ data matrix by randomly selecting seven measurement conditions from among 420 measurement conditions and replaced the lower $r_{\text{miss}} \times 100\%$ values with missing values. Each data set was centralized at mean $\mathbf{0}$ by preprocessing. Here, γ was selected by a grid search of 100 points so as to minimize Eq. 14 and set $k = 3$ and $\beta = 10$. In total, 5,600 trials were performed as part of the numerical experiments for network structure estimation.

The results of these 5,600 trials are summarized as the empirical probability of edge appearance p_{emp} (Fig. 5). The edges, which are $p_{\text{emp}} < 0.25$, are ignored due to low reliability and in order to avoid complexity of visualization. The estimated network structure of higher r_{miss} had a nest relation that constitutes a subgraph of the network structure of lower r_{miss} in Fig. 5. Although the nest relation would not be general, it would exhibit interesting tendencies. In addition, the number of edges and p_{emp} becomes small as r_{miss} increases.

Since the ERK pathway has been studied extensively by biologists, a great deal of knowledge on the ERK pathway has been accumulated. In the network structure of $r_{\text{miss}} = 0.05$ (Fig. 5 A), 11 of the 17 edges were reported to correspond to direct interaction between molecular species [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. The remainder of the edges, which are between ERK and EGR1[24], between ERK and CREB[25], between CREB and p38[26], between CREB and c-JUN[20], between CREB and JUNB[27], or between p38 and JNK[28], were reported to corresponded to indirect interaction between molecular species. Note that the existence of indirect interactions does not preclude the possibility of direct interactions. In addition, whereas the edges between JNK and c-JUN[29], between JNK and JUNB[30], and between c-JUN and FOSB[31], were reported to involve direct interaction between molecular species, the p_{emp} for the edges was less than 0.25. Biological knowledge is based on experimental studies conducted under various conditions, such as different cell lines, from various organs or species. Thus, biological knowledge would not be consistent among individuals, and some interactions in biological knowledge possibly may be inactive depending on the organ or species. In the present study, the edges between JNK and c-JUN, between JNK and JUNB, and between c-JUN and FOSB may not be active. In general, identifying the disappearance of interactions in biological knowledge is important for understanding cell systems.

We approximately estimated the network structures by partial correlation under the assumption of Gaussianity while ignoring time ordering. Since the objective is not necessarily to estimate the physical or biochemical interaction between molecular species, we consider that even statistical relationships between molecular species yields information beneficial to understanding life phenomena, particularly for pathways with little knowledge.

IV. Conclusion

We proposed a novel method to estimate a network structure using a sparse partial correlation matrix with missing values. The proposed method was applied and validated by the artificial data set and the signal transduction data set. In the analysis of artificial data, on average, the proposed method exhibited a higher F_1 score between estimated and true network structures than the separate method and mean imputation. In the analysis of signal transduction data, according to biological knowledge, 11 of the 17 edges of $p_{\text{emp}} > 0.25$ estimated using the proposed method have direct interactions, and the remaining six edges have indirect interactions. Moreover, the

estimated network structure having a higher missing rate had a nest relation that constitutes the subgraph of the estimated network structure having a lower missing rate. This suggests that the edges, which have direct or indirect interactions, appear very frequently. Although estimated networks will be rough if the sample size is small, the proposed method, when used in conjunction with biological experiments for validation, is better for investigating unknown interactions. Although the algorithm of proposed method includes some heuristics, the numerical experiment for the artificial and signal transduction data set, shows advantage compared to other methods and yields biologically reasonable interpretation, respectively. In the future, we improve the algorithm in order to be computationally more efficient and intend to apply the proposed method to omics data sets.

ACKNOWLEDGMENTS

The authors would like to thank Ms. Y. Yamauchi for her assistance in compiling the references. The present study was supported by a Grant-in-Aid for Scientific Research on Innovative Areas (Grant Number 16H01551) from the Japan Society for the Promotion of Science (JSPS) .

REFERENCES

[1] K. Yugi, et al., "Reconstruction of insulin signal flow from phosphoproteome and metabolome data.", *Cell Rep.* 8(4), 2014, pp.1171-1183.

[2] Z. Mousavian., et al., "Information theory in systems biology. Part I: Gene regulatory and metabolic networks.",*Seminars in Cell & Developmental Biology*, 51, 2016, pp.3-13.

[3] Z. Mousavian., et al., "Information theory in systems biology. Part II: proteinprotein interaction and signaling networks",*Seminars in Cell & Developmental Biology*, 51, 2016, pp.14-23.

[4] A. P. Dempster, "Covariance Selection.", *Biometrics* 28(1), 1972, pp.157-&.

[5] Meinshausen, N. and P. Buhlmann, "High-dimensional graphs and variable selection with the Lasso." *Annals of Statistics 34(3)*, 2006, pp.1436-1462.

[6] J. Peng, et al., "Partial Correlation Estimation by Joint Sparse Regression Models." *Journal of the American Statistical Association* 104(486), 2009, pp.735-746.

[7] P. Danaher., et al., "The joint graphical lasso for inverse covariance estimation across multiple classes.", *J. R. Stat. Soc. Series B Stat. Methodol.*, 76(2), 2014, pp.373-397.

[8] S. Basu., et al., "Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data", *Bioinformatics*, 33(10), 2017, pp.1545-1553.

[9] R. Tibshirani, "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society Series B-Methodological* 58(1), 1996, pp.267-288.

[10] P. Ott, "Incremental Matrix Factorization for Collaborative Filtering." *Science, Technology and Design 01/2008, Anhalt University of Applied Sciences*, 2008.

[11] W. J. J. Fu, "Penalized regressions: The bridge versus the lasso." *Journal of Computational and Graphical Statistics* 7(3), 1998, pp.397-416.

[12] T.H. Saito, et al., "Temporal Decoding of MAP Kinase and CREB Phosphorylation by Selective Immediate Early Gene Expression." *Plos One* 8(3), 2013.

[13] Y. Ozaki, et al., "A Quantitative Image Cytometry Technique for Time Series or Population Analyses of Signaling Networks." *Plos One* 5(3), 2010.

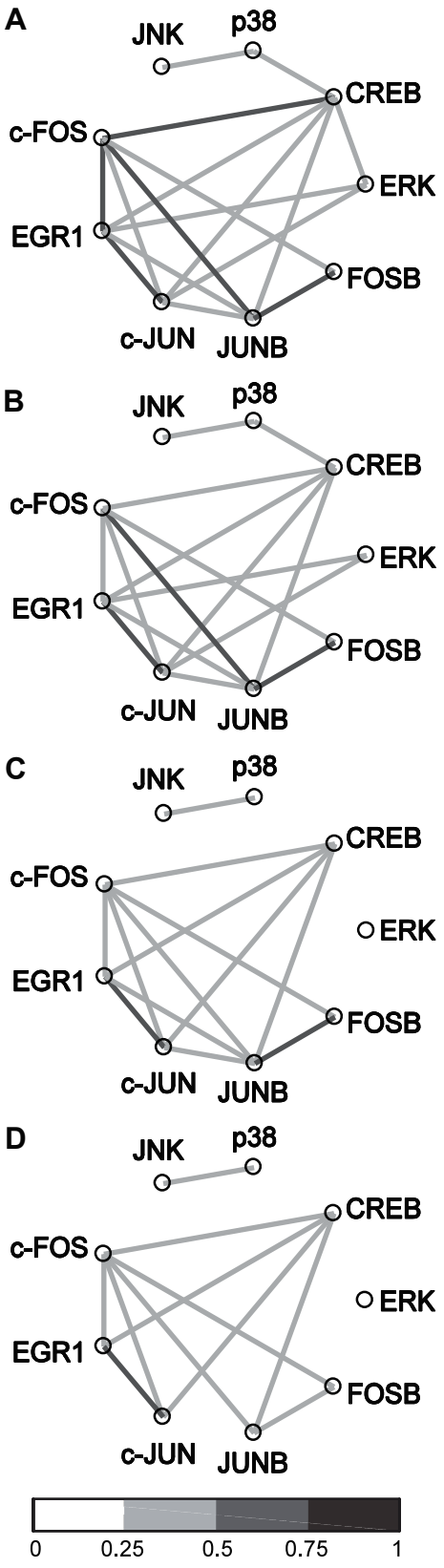

Fig. 5. Empirical probability of edge appearance p_{emp}. (A) $r_{\text{miss}} = 0.05$. (B) $r_{\text{miss}} = 0.1$. (C) $r_{\text{miss}} = 0.15$. (D) $r_{\text{miss}} = 0.2$.

[14] D. Qiao, et al., "Bile acid-induced activation of activator protein-1 requires both extracellular signal-regulated kinase and protein kinase C signaling." *J Biol Chem.*, 275(20), 2000, pp.15090-8.

[15] H. Cha-Molstad, et al., "Cell-type-specific binding of the transcription factor CREB to the cAMP-response element.", *Proc Natl Acad Sci U S A.* 101(37), 2004, pp.13572-7.

[16] TL. Beaumont et al., "Layer-specific CREB target gene induction in human neocortical epilepsy.", *J Neurosci.* 32(41), 2012, pp.14389-401.

[17] W. Sommer et al., "Intrastriatally injected c-fos antisense oligonucleotide interferes with striatonigral but not striatopallidal gamma-aminobutyric acid transmission in the conscious rat.", *Proc Natl Acad Sci U S A.* 93(24), 1996, pp.14134-9.

[18] PM. Chevray and D. Nathans, "Protein interaction cloning in yeast: identification of mammalian proteins that react with the leucine zipper of Jun.", *Proc Natl Acad Sci U S A.* 89(13), 1992, pp.5789-93.

[19] JC. Hsu et al., "Identification of LRF-1, a leucine-zipper protein that is rapidly and highly induced in regenerating liver.", *Proc Natl Acad Sci U S A.* 88(9), 1991, pp.3511-5.

[20] T. Herdegen and JD. Leah, "Inducible and constitutive transcription factors in the mammalian nervous system: control of gene expression by Jun, Fos and Krox, and CREB/ATF proteins.", *Brain Res Brain Res Rev.* 28(3), 1998, pp.370-490.

[21] Y. Levkovitz and JM. Baraban, "A dominant negative Egr inhibitor blocks nerve growth factor-induced neurite outgrowth by suppressing c-Jun activation: role of an Egr/c-Jun complex.", *J Neurosci.* 22(10), 2002, pp.3845-54.

[22] K. Krishnaraju et al., "The zinc finger transcription factor Egr-1 activates macrophage differentiation in M1 myeloblastic leukemia cells.", *Blood.* 92(6), 1998, pp.1957-66.

[23] KG. Mendelson et al., "Independent regulation of JNK/p38 mitogen-activated protein kinases by metabolic oxidative stress in the liver.", *Proc Natl Acad Sci U S A.* 93(23), 1996, pp.12908-13.

[24] E. Camerer et al., "Binding of factor VIIa to tissue factor on keratinocytes induces gene expression.", *J Biol Chem.* 275(9), 2000, pp.6580-5.

[25] S. Impey et al., "Cross talk between ERK and PKA is required for Ca2+ stimulation of CREB-dependent transcription and ERK nuclear translocation.", *Neuron.* 21(4), 1998, pp.869-83.

[26] JM. Swart et al., "Identification of a membrane Ig-induced p38 mitogen-activated protein kinase module that regulates cAMP response element binding protein phosphorylation and transcriptional activation in CH31 B cell lymphomas.", *J Immunol.*, 164(5), 2000, pp.2311-9.

[27] SN. Schiffmann et al., "Adenosine A2A receptors and basal ganglia physiology.", *Prog Neurobiol.* 83(5), 2007, pp.277-92.

[28] ED. Chan and DW. Riches, "IFN-gamma + LPS induction of iNOS is modulated by ERK, JNK/SAPK, and p38(mapk) in a mouse macrophage cell line.", *Am J Physiol Cell Physiol.*, 280(3), 2001, pp.C441-50.

[29] B. Drijard et al., "JNK1: a protein kinase stimulated by UV light and Ha-Ras that binds and phosphorylates the c-Jun activation domain.", *Cell.* 76(6), 1994, pp.1025-37.

[30] SY. Fuchs et al., "c-Jun NH2-terminal kinases target the ubiquitination of their associated transcription factors.", *J Biol Chem.* 272(51), 1997, pp.32163-8.

[31] J. Yen et al., "An alternative spliced form of FosB is a negative regulator of transcriptional activation and transformation by Fos proteins.", *Proc Natl Acad Sci U S A.* 88(12), 1991, pp.5077-81.