Master's Thesis

# Simultaneous Likelihood Estimation for Normal Mixture Distributions and Sparse Precision Matrix

Guidance

Associate Professor Nobuo YAMASHITA

Kazuki MATSUDA

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University


February 2014

**Abstract**

In this paper we consider both clustering and graphical modeling for given data. The clustering is the task of grouping of the data, while the graphical modeling provides a conditional dependence structure among variables in the data. In this paper we suppose that the data obeys a mixture of normal distributions. Then, we may apply the existing methods based on the maximum likelihood estimation, that is, the Expectation Maximization (EM) algorithm and the $L_1$ regularized maximum likelihood estimation. The EM algorithm provides clusters such that each cluster obeys a single normal distribution. The $L_1$ regularized maximum likelihood estimation finds a sparse precision matrix whose nonzero element represents a dependency of the corresponding variables. It assumes that the data obeys a single normal distribution. Thus we may apply it for each cluster given by the EM algorithm. However, this procedure estimates two different mixture distributions by two algorithms, which should be the same.

In this paper we propose a simultaneous estimation model for mixture distributions and a sparse precision matrix from the given data. We first formulate a maximization problem of the log likelihood function of mixture distribution with the $L_1$ regularized term of the precision matrix. We then propose a coordinate descent method for solving the problem. The proposed method is a generalization of the EM algorithm. We present some numerical results that show the validity of the proposed model.

# Contents

# 1 Introduction

With the developments in information science, vast amounts of information have been flooding in society. In such a situation, we must take advantage of a large amount of information using the data processing techniques. Thus it is important to obtain useful information efficiently. As such techniques, there have been studied machine learning, data mining, and a stochastic model.

In this paper we focus on clustering and graphical modeling as the data processing technique. The clustering is the task of grouping of the data, while the graphical modeling provides a conditional dependence structure among variables in the data. These techniques are based on statistics that estimates the nature and structure of the information source from the given data, and provide valuable knowledge on the data with complex structure such as mixture distributions.

In this paper we consider the clustering for normal mixture distributions, which are piled up multiple normal distributions. It assumes that what distribution observations arise from is unknown. Then, since observations is incomplete, it is difficult to estimate the mixture distributions by clustering .

The Expectation Maximization (EM) algorithm is the useful technique for estimating the complex distribution such as mixture distributions. The algorithm is an iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models, where the model depends on unobserved latent variables [1]. In general,it is difficult to maximize the likelihood for margin distribution of observations and latent variables. Thus, the EM algorithm calculates the conditional expectation of the log-likelihood function using the current estimate for the parameters, and then maximize a expectation function. In practice, the objective function arising from a clustering is usually non-convex. Fortunately the expectation function in the EM algorithm is concave [2]. Therefore we can obtain the estimation for the parameters explicitly. In [4], the EM algorithm for normal mixture distributions can be interpreted as a block coordinate descent method. Then we can show a global convergence for the EM iterations.

In this paper we also consider the graphical modeling for investigating the dependence structure among variables. By using the graphical modeling, we can visualize the structure and obtain the essential dependency. A Graphical Gaussian Model (GGM) is a representative example to graphical modeling. GGM describes the liner graph among variables which obey a normal distribution. Although we describe the detail in section 2.2, the nonzero element of a precision matrix represents the dependency of corresponding variables. Thus, for the graphical modeling, we need to estimate the sparsity of a precision matrix.

The $L_1$ regularized maximum likelihood estimation finds a sparse precision matrix whose non-diagonal elements is partly zero. Recently, the $L_1$ regularized maximum likelihood estimation has been studied intensively, and various estimation models are proposed, [7, 8]. In [9], the simultaneous estimation model for multiple precision matrices corresponding to the set of variables. However, since it is assumed that these matrices have the same structure, this model is not appropriate for the different data set.

In this paper we propose a simultaneous estimation model for mixture distributions and a sparse precision matrix from the given data. We first formulate a maximization problem of the log likelihood function of mixture distribution with the $L_1$ regularized term of the precision matrix. We then propose a coordinate descent method for solving the problem. The proposed method is a generalization of the EM algorithm. The proposed model includes the following advantages.

- It enables us to provide the clusters for normal mixture distributions and estimate the graphical model for each cluster.
- It can be Applied to the model that have the unobserved latent variable.

The log likelihood function of mixture distribution with the $L_1$ regularized term of the precision matrix is concave and its maximization problem is a log-determinant semidefinite programing. The problem is solved by interior point method or coordinate descent method.

The remainder of the paper is organized as follows. In Section 2, we introduce the details of the clustering for mixture distributions, the EM algorithm, and the $L_1$ maximum likelihood estimation for GGM. Then we propose the simultaneous estimation model for normal mixture distributions and sparse precision matrix, and formulate the log-likelihood function with $L_1$ regularized term of the precision matrix in Section 3. We also show the global convergence for the proposed model. In Section 4, we report some results of numerical experiments that show the validity of the proposed model. Finally, we give concluding remark in Section 5.

# 2 Preliminaries

In this section, we introduce the clustering for mixture distributions by using EM algorithm. We also explain normal graphical model estimation (GGM).

## 2.1 Mixture distributions and clustering

We first formulate the clustering for mixture distributions including latent variables.

Let $z \in \mathbb{R}^m$ be a latent vector. The latent vector $z$ is a 1-$of$-$m$ expression, where only one variable $z_i$ in $z_1, ..., z_m$ is one, and the others are zero. The latent variable indicates which clusters each data is observed from. For example when a data $x$ is observed from the $i$th cluster, we set $z_i = 1$ and $z_j = 0\,(i \neq j)$. Moreover, if multiple observations $x_1, ..., x_n$ are given, we may define the latent variable $z_k$ corresponding to each data $x_k$. Now we formulate a marginal distribution $p(z)$ and its conditional distribution $p(x|z)$. In addition, $p(z)$ is determined by mixture coefficient $\alpha_i$ as follows.

$$p(z_i = 1) = \alpha_i,$$

where $\alpha_i$ is probability that $x$ arise from the $i$th cluster. Note that $\alpha \in \Omega$, where

$$\Omega = \left\{ \alpha = (\alpha_1, ..., \alpha_m)^T : \sum_{i=1}^{m} \alpha_i = 1,\ \alpha_i \geq 0,\ i = 1, \ldots, m \right\}.$$

Then $p(z)$ can be written as

$$p(z) = \prod_{i=1}^{m} {\alpha_i}^{z_i}, \tag{2.1}$$

Now we assumes that the data $x$ obeys a normal distribution. Then, when $z_i = 1$, $p(x|z_i = 1)$ is given by

$$p(x|z_i = 1) = \mathcal{N}(x|\mu_i, \Sigma_i).$$

Note that each cluster $i$ of data corresponds to a single normal distribution $\mathcal{N}(\mu_\rangle, \pm_\rangle)$ Similar to (**??**p(z))), $p(x|z)$ similarly for normal mixture distributions is given as

$$p(x|z) = \prod_{i=1}^{m} \mathcal{N}(x|\mu_i, \Sigma_i)^{z_i}. \tag{2.2}$$

Then, we can give the marginal distribution of $x$ with (2.1) and (2.2) from the joint distribution $p(z)p(x|z)$.

$$p(x) = \sum_{z} p(z)p(x|z) = \sum_{i=1}^{m} \alpha_i \mathcal{N}(x|\mu_i, \Sigma_i). \tag{2.3}$$

Now we consider the estimation the $\alpha_i$, $\mu_i$ and $\Sigma_i$ for $i = 1, ..., m$. Since the observations are given, the mixture distributions function can be regarded as the likelihood function about $\alpha_i$ and random variables $(\mu_i, \Sigma_i)$. Thus we can write the likelihood function as

$$L(\alpha, \mu, \Sigma) = \sum_{i=1}^{m} \alpha_i \mathcal{N}(x|\mu_i, \Sigma_i). \tag{2.4}$$

When $\alpha_i$ is given by the maximum likelihood estimation, we can determine latent variables for each observation. Therefore we can give these observations the information of whether they arise from which clusters. That is why it is possible to realize a clustering for observations.

In next subsection, to estimate parameters for mixture distributions including latent variables, we introduce the EM algorithm.

### 2.1.1 Expectation Maximization algorithm

The EM algorithm is a technique for parameter estimations in a statical model including an unobserved incomplete data. Since the EM algorithm is simple, it is easy to implement. The main iteration of the EM algorithm are described as follows.

- Formulate the log-likelihood function for the parameters if observations and its distribution are given.

- Calculate the conditional expectation to the log-likelihood with the current parameter estimate (Expectation Step, E-Step).
- Find the maximizer of the conditional expectation (Maximization Step, M-Step).

Let an observed data and an unobserved latent variable be $X$ and $Z$. Then $p(X|\theta)$ is the probability distribution of $X$, where $\theta$ is the parameter that represents the probability distribution. Now we formulate the log-likelihood function for $\theta$.

$$L(\theta) = \log p(X|\theta) = \log \left\{ \sum_{Z} p(X,\, Z|\theta) \right\}.$$

When the model is the mixture distributions, we do not know which clusters data $X$ is observed from, and hence we cannot obtain the complete data $(X, Z)$. Thus, We cannot maximize the log-likelihood function $L(\theta)$ directly. Instead it assumes that the latent variable is given, and we consider the conditional expectation to the likelihood with the given parameter (E-Step). Moreover we obtain the maximum likelihood estimate by maximizing the conditional expectation (M-Step).

The EM algorithm alternates the E-Step and the M-Step, and finally maximize the log-likelihood function. The following is the detail.

**E-Step**

Formulate the conditional expectation.

$$Q(\theta|\, \theta^t) \;=\; E_z[L(\theta)\,|\,X,\, \theta^t].$$

**M-Step**

Find the maximizer $\hat{\theta}$ of $Q(\theta|\, \theta^t)$. Update $\hat{\theta}$ to $\theta^{t+1}$

### 2.1.2 The EM algorithm for normal mixture distributions

We explain the case when we apply EM algorithm to normal mixture distributions.

Given $n$ independent and identically-distributed (i.i.d.) observations $x_1, ..., x_n \in \mathbb{R}^d$ drawn from a $d$-dimensional normal distribution $\mathcal{N}_i(\mu_i,\, {\Lambda_i}^{-1})$, the probability density function is given by

$$p(x_k\,|\,\alpha, \mu, \Sigma) \;=\; \sum_{i=1}^{m} \alpha_i p_i(x_k\,|\,\mu_i, \Sigma_i), \tag{2.5}$$

where $\alpha = (\alpha_1, ..., \alpha_m)^T$, $\mu = (\mu_1, ..., \mu_m)$, $\Sigma = (\Sigma_1, ..., \Sigma_m)$. Moreover $\mu_i$ is a mean of the $i$th cluster, and $\Sigma_i$ is a covariance matrix, which is symmetric semidefinite.
Then we represent the joint distributions for the observations $X = (x_1, ..., x_n)$ as :

$$P(X\,|\,\alpha, \mu, \Sigma) \;\;=\;\; \prod_{k=1}^{n} p(x_k\,|\,\alpha, \mu, \Sigma). \tag{2.6}$$

When we obtain the observations $X$, the log-likelihood function for the parameters $\mu$, $\Sigma$ is given by

$$\begin{aligned} L(\alpha, \mu, \Sigma) &= \log P(X \,|\, \alpha, \mu, \Sigma) \\ &= \sum_{k=1}^{n} \log p(x_k \,|\, \alpha, \mu, \Sigma). \end{aligned} \tag{2.7}$$

For the maximum likelihood estimation for $\alpha, \mu, \Sigma$, we formulate the following optimization problem.

$$\begin{aligned} &\text{maximize} \quad L(\alpha, \mu, \Sigma) \\ &\text{subject to} \quad \alpha \in \Omega,\ \Sigma_i \succeq 0,\ i = 1, \ldots, m. \end{aligned}$$

Unfortunately it is quite difficult to solve this problem via a direct optimization solver. Thus we can calculate the conditional expectation to the log-likelihood function based on the given parameters $\alpha^t$, $\mu^t$, $\Sigma^t$ at E-Step.

$$\begin{aligned} Q(\alpha,\, \mu,\, \Sigma) &= E\left[L(\alpha,\, \mu,\, \Sigma) \mid x,\, \alpha^t,\, \mu^t,\, \Sigma^t\right] \\ &= \sum_{k=1}^{n} \sum_{i=1}^{m} \gamma_{ik} \log \alpha_i p_i(x_k \,|\, \mu_i, \Sigma_i), \end{aligned} \tag{2.8}$$

where $\gamma_{ik}$ is defined as follows.

$$\gamma_{ik}^t = \frac{\alpha_i^t p_i(x_k \,|\, \mu_i^t, \Sigma_i^t)}{p(x_k \,|\, \alpha^t,\, \mu^t,\, \Sigma_1^t, ..., \Sigma_m^t)}. \tag{2.9}$$

At M-Step, we maximize $Q(\alpha,\, \mu,\, \Sigma)$ with respect to $\alpha, \mu, \Sigma$ variables. The maximizers of $Q(\alpha,\, \mu,\, \Sigma)$ are obtained via Lagrange multiplier method to each variables. Then these maximizers can be obtained explicitly by using $\gamma_{ik}^t$ as follows and the following is just like at current iteration.

$$\alpha_i^{t+1} = \frac{N_i^t}{n}, \quad \mu_i^{t+1} = \frac{1}{N_i^t} \sum_{k=1}^{n} {}^t\gamma_{ik}\, x_k. \tag{2.10}$$

$$\Sigma_i^{t+1} = \frac{1}{N_i^t} \sum_{k=1}^{n} \gamma_{ik}^t\, (x_k - \mu_i^t)(x_k - \mu_i^t)^T, \tag{2.10'}$$

where $N_i$ is given by

$$N_i^t = \sum_{k=1}^{n} \gamma_{ik}^t. \tag{2.11}$$

From the above, once $\gamma_{ik}$ is obtained, we can calculate the maximizers of $Q(\alpha,\, \mu,\, \Sigma)$ at each iteration. We describe the details of the EM algorithm.

**Algorithm 1** *Choose starting parameters* $\alpha^0$, $\mu^0$, $\Sigma^0$.
*Let* $t = 0$.

**Step1** *At the current iteration,*

**E-Step** *compute* $\gamma_{ik}^t$ *from* (2.9) *by* ($\alpha^t$, $\mu^t$, $\Sigma^t$).

**M-Step** *Compute the next iterate parameters* ($\alpha^{t+1}$, $\mu^{t+1}$, $\Sigma^{t+1}$) *from* (2.10) *by obtained* $\gamma_{ik}^t$ *at E-Step.*

**Step2** *Terminate the iteration if a certain termination condition holds. Otherwise go to Step1.*

### 2.1.3 Another Interpretation of the EM algorithm

We can give an another interpretation [4] to the Steps of the EM algorithm for normal mixture distributions. Then the following function is defined in [4].

$$D(W,\, \alpha,\, \mu,\, \Sigma) \,=\, \sum_{i=1}^{m} \sum_{k=1}^{n} W_{ik}(\log W_{ik} - \log \alpha_i p_i(x_k \,|\, \mu_i, \Sigma_i)), \tag{2.12}$$

where $W$ is a variable such that

$$W \,\in\, vM \,=\, \left\{ W \in \mathbb{R}^{mn} \,:\, 0 \leq W_{ik} \leq 1,\, \sum_{i=1}^{m} W_{ik} \,=\, 1,\, \sum_{k=1}^{n} W_{ik} \,>\, 0 \right\}.$$

By using this function, we can rewrite the steps of the EM algorithm.

**Algorithm 2** *Choose starting parameters* $\alpha^0$, $\mu^0$, $\Sigma^0$.
*Let* $t = 0$.

**Step1** *At the current iteration,*

**Step1-1** *Let* ($\alpha^t$, $\mu^t$, $\Sigma^t$) *be fixed. Minimize* $D(W, \alpha^t, \mu^t, \Sigma^t)$ *over* $W \in M$. *The minimizer* $\hat{W_{ik}}$ *is obtained explicitly, and is as follows:*

$$\hat{W_{ik}} \,=\, \frac{\alpha_i^t p_i(x_k \,|\, \mu_i^t, \Sigma_i^t)}{p(x_k \,|\, \alpha^t,\, \mu^t,\, \Sigma_1^t, ..., \Sigma_m^t)} \tag{2.13}$$

**Step1-2** *Update* $\hat{W_{ik}}$ *to* $W_{ik}^{t+1}$. *we obtain* ($\alpha^{t+1}$, $\mu^{t+1}$, $\Sigma^{t+1}$) *by minimizing* $D(W^{t+1}, \alpha, \mu, \Sigma)$ *with respect to* $\alpha \in \Omega$, $\mu$, $\Sigma_i \succeq 0$.

**Step2** *Terminate the iterate if*

$$D(W^t, \alpha^t, \mu^t, \Sigma^t) - D(W^{t+1}, \alpha^{t+1}, \mu^{t+1}, \Sigma^{t+1}) < \tau,$$

*where* $\tau$ *is a positive constant.*

Here comparing (2.13) and (2.9), The right hand of (2.13) corresponds to that of (2.9). From this fact, we see the following relation.

**Remark 1** *In* Algorithm 1 *and* Algorithm 2, $\hat{W}$ *from* (2.13) *is equivalent to* $\gamma_{ik}^t$ *from* (2.9)

This means that we regard Step 1-1 as the E-Step of of determining the conditional expectation $Q(\alpha, \mu, \Sigma)$ of the complete data log-likelihood function. For a fixed $W^{t+1}$ Step 1-2 is equivalent to the M-Step of maximizing $Q(\alpha, \mu, \Sigma)$.

Moreover note that $D(W^{t+1}, \alpha, \mu, \Sigma) = -L(\alpha, \mu, \Sigma)$ when $W^{t+1}$ is fixed. Thus $(\alpha^{t+1}, \mu^{t+1}, \Sigma^{t+1})$ are the minimizers of $D(W^{t+1}, \alpha, \mu, \Sigma)$. It follows that the step of the EM algorithm is viewed as the step in the block coordinate descent method for minimizing $D(W, \alpha, \mu, \Sigma)$ with respect to $W$ and $\alpha, \mu, \Sigma$. Then it is clear that

$$\begin{aligned} D(W^t, \alpha^{t-1}, \mu^{t-1}, \Sigma^{t-1}) &\leq D(W^t, \alpha^t, \mu^t, \Sigma^t) \\ &\leq D(W^{t+1}, \alpha^t, \mu^t, \Sigma^t). \end{aligned}$$

Note that $D(W, \alpha, \mu, \Sigma)$ is convex with respect to $W$ and $\alpha, \mu, \Sigma$, respectively. Moreover the algorithm 2 has a global convergence property.

## 2.2 Graphical Gaussian model

The Graphical Gaussian Model (GGM) is a graphical interpretation of the structural dependency among variables that obey a normal distribution. This graph structure has edges and nodes. The edges represent the conditional dependence among the two variables, while the nodes is corresponding to variables.

The conditional independence is an index that shows the dependency among variables. It assumes that $n$ variables $R_1, R_2, ..., R_n$ obeys a single normal distribution. If the conditional independence between $R_i$ and $R_j$ is satisfied, $R_i$ is conditionally independent of $R_j$ for given the other variables.

Estimation of the conditional independence is equivalent to estimation of a sparse precision matrix corresponding to variables that obeys the normal distribution. The precision matrix is the inverse of covariance matrix, that is, $\Lambda = \Sigma^{-1} \in \mathbb{R}^{d \times d}$. The following relation satisfied the elements of the sparse precision matrix and conditional independence to variables is important for our study.

**Definition 2.1 (Conditional independence)** *Suppose that n variables* $R_1, R_2, ..., R_n$ *obeys a single normal distribution* $\mathcal{N}(\mu, \Lambda^{-1})$. *Then* $R_i$ *and* $R_j$ *are conditional independent given the other variables if and only if* $\Lambda_{ij}$ *is equal to zero.*

In brief, if most of variables are conditional independent, the precision matrix is sparse. However the precision matrix estimated by the maximum likelihood estimation is usually dense, and therefore essential dependency among variables is not clear.

In order to estimate a sparse precision matrix, the $L_1$ regularized maximum likelihood estimation have been proposed. This estimation is based on the following optimization with the $L_1$ regularized term of the precision matrix $\Lambda$.

$$\begin{array}{ll} \text{minimize} & L(\Lambda) + \rho\,||\Lambda||_1 \\ \text{subject to} & \Lambda \succeq 0, \end{array}$$

where $\rho$ is a positive constant, and $L(\Lambda)$ is log-likelihood function of the single normal distribution. It assumes that observations $x$ is given, $L(\Lambda)$ is just as follows.

$$L(\Lambda) = \mathrm{tr}(\Lambda\hat{\Sigma}) \ - \ \log\det(\Lambda),$$

where $\hat{\Sigma}$ is the sample covariance matrix. It is revealed that if a precision matrix $\Lambda$ is semidefinite symmetric, this function is convex [11]. Therefore the above $L_1$ regularized problem is a convex optimization.

# 3 Simultaneous Estimation Model

In this section, we propose the simultaneous estimation model for mixture distributions and sparse precision matrices. We realize a clustering for mixture distributions by the EM algorithm, and estimate these precision matrices by using the $L_1$-regularized term.

## 3.1 The Proposed model

As in the previous section, given $n$ i.i.d. observations $x_1, ..., x_n \in \mathbb{R}^d$ drawn from a $d$-dimensional normal distribution $\mathcal{N}_i(\mu_i, {\Lambda_i}^{-1})$, the probability density function is as follows.

$$p_i(x_k \,|\, \mu_i, {\Lambda_i}^{-1}) = \frac{1}{(2\pi)^{\frac{n}{2}}\,|{\Lambda_i}^{-1}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_k-\mu_i)^T\Lambda_i(x_k-\mu_i)\right). \tag{3.14}$$

where the precision matrix $\Lambda = \Sigma^{-1}$ is the inverse covariance matrix, and is symmetric semidefinite matrix. Then a family of mixture distributions of $m$ clusters is given by

$$\begin{aligned} p(x_k \,|\, \alpha, \mu, {\Lambda_i}^{-1}) &= \sum_{i=1}^{m} \alpha_i p_i(x_k \,|\, \mu_i, {\Lambda_i}^{-1}) \\ &= \sum_{i=1}^{m} \frac{\alpha_i}{(2\pi)^{\frac{n}{2}}\,|{\Lambda_i}^{-1}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_k-\mu_i)^T\Lambda_i(x_k-\mu_i),\right) \end{aligned} \tag{3.15}$$

where $\alpha \in \Omega$ is defined before.
Moreover, from (2.7), we consider the following log-likelihood function to joint distribution for observations $X = (x_1, ..., x_n)$,

$$\begin{aligned} L(\alpha, \mu, \Lambda_1^{-1}, ..., \Lambda_m^{-1}) &= \log P(X \,|\, \alpha, \mu, \Lambda_1^{-1}, ..., \Lambda_m^{-1}) \\ &= \sum_{k=1}^{n} \log p(x_k \,|\, \alpha, \mu, \Lambda_1^{-1}, ..., \Lambda_m^{-1}). \end{aligned} \tag{3.16}$$

Since $1 \,/\, |{\Lambda_i}^{-1}|^{\frac{1}{2}} = |\Lambda_i|$, the left-hand can be regarded as expression as a variable $\Lambda_i$. In what follows, we view (3.16) as $L(\alpha, \mu, \Lambda_1, ..., \Lambda_m)$ instead of $L(\alpha, \mu, \Lambda_1^{-1}, ..., \Lambda_m^{-1})$. There let $\Lambda = (\Lambda_1, ..., \Lambda_m)$. Then we can rewrite

$$L(\alpha, \mu, \Lambda) = \sum_{k=1}^{n} \log \left\{ \sum_{i=1}^{m} \frac{\alpha_i |\Lambda_i|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left( -\frac{1}{2}(x_k - \mu_i)^T \Lambda_i (x_k - \mu_i) \right) \right\}. \tag{3.17}$$

However, even if we maximize this likelihood function directly to estimate parameters, precision matrices $\Lambda_i (i = 1, ..., m)$ do not have sparsity patterns because these covariance matrices are dense in general.

Therefore, we introduce the $L_1$-regularized norm of precision matrix is introduced to (3.17) so that these matrices is sparsity. By maximizing the log-likelihood function with $L_1$-regularized norm, we can realize a clustering for mixture distributions and estimate the sparse precision matrix simultaneously. Then optimization problem that is based on the $L_1$-regularized maximum likelihood method is as follows.

$$\begin{aligned} \text{(P)} \quad \text{maximize} \quad & L(\alpha, \mu, \Lambda) - \sum_{i=1}^{m} \rho \, ||\Lambda_i||_1 \\ \text{subject to} \quad & \alpha \in \Omega, \, \Lambda_i \succeq 0, \; i = 1, \ldots, m, \end{aligned}$$

where $\rho$ is a nonnegative weight, and $||\Lambda||_1 = \sum_{i=1}^{n} \sum_{j=1}^{n} |\Lambda_{ij}|$. The parameter $\rho$ controls the trade-off between the goodness-of-fit and sparsity of $\Lambda_i$. As discussed in the previous section, after calculating the conditional expectation at E-Step, we get the following function.

$$\begin{aligned} \bar{Q}(\alpha, \, \mu, \, \Lambda) &= E\left[ L(\alpha, \, \mu, \, \Lambda) - \sum_{i=1}^{m} \rho \, ||\Lambda_i||_1 \,|\, X, \, \alpha^t, \, \mu^t, \, \Lambda^t \right] \\ &= \sum_{k=1}^{n} \sum_{i=1}^{m} \gamma_{ik} \log \alpha_i p_i(x_k \,|\, \mu_i, \Lambda_i) - \sum_{i=1}^{m} \rho \, ||\Lambda_i||_1, \end{aligned} \tag{3.18}$$

where $\gamma_{ik}$ is the same as (2.9).

Here the following theorem gives the estimation values of parameters that are maximizers of $\bar{Q}(\alpha, \, \mu, \, \Lambda)$ at M-Step.

**Theorem 3.1** *The maximizers* $(\alpha^{t+1}, \, \mu^{t+1})$ *of* $\bar{Q}(\alpha, \, \mu, \, \Lambda)$ *are given by* (2.10).

**Proof** *Since the* $L_1$*-regularized term of* $\Lambda_i$ *is independent of the parameters* $\alpha$, $\mu$, *it is clear that* $\alpha^{t+1} \, \mu^{t+1}$ *obtained at* M-Step *are written as* (2.10). □

Considering the above theorem, the steps of EM algorithm for (P) are rewritten as follows.

**Algorithm 3** *Choose starting parameters* $\alpha^0$, $\mu^0$, $\Lambda^0$. *Let* $\rho = (\rho_1, \rho_2, \ldots, \rho_m)^T$ *be a given parameter. Let* $t = 0$.

**Step1** *Compute* $\gamma_{ik}^t$ *from* (2.9) *by* ($\alpha^t$, $\mu^t$, $\Lambda^t$).

**Step2**

**Step2-1** *Compute the next iterate parameters* ($\alpha^{t+1}$, $\mu^{t+1}$) *from* (2.10) *by obtained* $\gamma_{ik}^t$ *at Step1.*

**Step2-2** *Compute* $\Lambda^{t+1}$ *by solving solve the below problem for* $\Lambda_i \succeq 0$.

$$\begin{aligned} (\mathrm{Q}') \quad & \text{maximize} \quad \bar{Q}(\alpha^{t+1}, \mu^{t+1}, \Lambda) \\ & \text{subject to} \quad \Lambda_i \succeq 0,\ i = 1, \ldots, m. \end{aligned}$$

**Step3** *Terminate the iteration if a certain termination criteria holds. Otherwise go to Step1.*

Note that the objective function of (Q') is written as

$$\bar{Q}(\alpha^{t+1}, \mu^{t+1}, \Lambda) = \sum_{i=1}^{m} \left\{ N_i^t \left( \log\left(\det\Lambda_i\right) - \mathrm{tr}(\Lambda_i \hat{\Sigma}_i) \right) - \rho \, ||\Lambda_i||_1 \right\}, \tag{3.19}$$

where $\hat{\Sigma}$ is defined by

$$\hat{\Sigma_i} = \frac{1}{N_i^t} \sum_{k=1}^{n} \gamma_{ik}^t \, (x_k - \mu_i^{t+1})(x_k - \mu_i^{t+1})^T.$$

If $\Lambda_i$ is symmetric semidefinite matrix, $\log\left(\det\Lambda_i\right) - \mathrm{tr}(\Lambda_i \hat{\Sigma})$ in (3.19) is concave on $\Lambda_i$. Moreover since the $L_1$-regularized term is convex, we say that $\bar{Q}(\alpha, \mu, \Lambda)$ is concave with respect to $\Lambda_i$. Thus by changing the sign of the objective function (3.19), we can regard (Q') as the convex semidefinite programming problem, and its problem has a unique solution.

## 3.2 Global Convergence

Here we show the global convergence of of Algorithm 3 for the proposed model (P). We introduce the following function that has the $L_1$-regularized term on $D(W, \alpha, \mu, \Sigma)$.

$$\bar{D}(W, \alpha, \mu, \Lambda) \; = \; \sum_{i=1}^{m} \sum_{k=1}^{n} W_{ik} (\log W_{ik} - \log \alpha_i p_i(x_k \,|\, \mu_i, \Lambda_i)) \; + \; \sum_{i=1}^{m} \rho \, ||\Lambda_i||_1.$$

By applying the block coordinate descent method to the minimization of the above function, we can minimize $\bar{D}(W, \alpha, \mu, \Lambda)$ with respect to $W$ and $\alpha, \mu, \Lambda$, respectively. Then we give the following theorem.

**Theorem 3.2** *Let* $(\alpha^t,\, \mu^t,\, \Lambda^t)$ *be fixed. Then the minimizer* $\bar{W_{ik}}$ *of* $\bar{D}(W,\, \alpha^t,\, \mu^t,\, \Lambda^t)$ *over* $W \in M$ *is given as* (2.13).

**Proof** *Note that* $\sum_{i=1}^{m} \rho||\Lambda_i^t||_1$ *is regarded as a constant because* $\Lambda_i^t$ *is fixed. Thus, ignoring the constant term that does not depend on* $W$*, the minimization of* $\bar{D}(W,\, \alpha^t,\, \mu^t,\, \Lambda^t)$ *over* $W$ *is equivalent to the minimization of* $D(W,\, \alpha^t,\, \mu^t,\, \Lambda^t)$ *from* (2.12). □

It follows from Remark 1 that

$$\bar{W_{ik}} \;=\; \frac{\alpha_i^t p_i(x_k \,|\, \mu_i^t, \Lambda_i^t)}{p(x_k \,|\, \alpha^t,\, \alpha^t,\, \mu^t,\, \Lambda^t)}.$$

Besides, the next theorem explains the relation between the updated parameters $(\alpha^t,\, \mu^t)$ and $\bar{D}(W^t,\, \alpha,\, \mu,\, \Lambda^t)$.

**Theorem 3.3** *The parameters* $\alpha^{t+1},\, \mu^{t+1}$ *by* (2.10) *are the minimizer of* $\bar{D}(W^{t+1},\, \alpha,\, \mu,\, \Lambda^t)$ *with respect to* $\alpha$ *and* $\mu$*, respectively.*

**Proof** *Since the* $L_1$*-regularized term is not dependent of* $\alpha$ *and* $\mu$*, this theorem is shown by considering* $\alpha^{t+1}, \mu^{t+1}$ *are optimal solutions for* $D(W^{t+1},\, \alpha,\, \mu,\, \Lambda^t)$ *from* (2.12). □

Now suppose that $W^{t+1},\, \alpha^t,\, \mu^t$ are fixed. Now suppose that fix $(W^{t+1},\, \alpha^t,\, \mu^t)$. Ignoring the constant term and substituting $\bar{D}(W,\, \alpha,\, \mu, \Lambda)$ to $W^{t+1},\, \alpha^t,\, \mu^t$, we obtain

$$\bar{D}(W^{t+1},\, \alpha^{t+1},\, \mu^{t+1},\, \Lambda) = \sum_{i=1}^{m} \left\{ N_i^t \left( \mathrm{tr}(\Lambda_i \hat{\Sigma_i}) - \log\left(\det \Lambda_i\right) \right) + \rho_i |||\lambda_i||_1 \right\}.$$

Note that $\bar{D}(W^{t+1},\, \alpha^{t+1},\, \mu^{t+1}, \Lambda)$ is also convex with respect to $\Lambda_i$. Thus, as $\bar{D}(W, \alpha, \mu, \Lambda)$ is convex function with respect to $W$, $\alpha$, $\mu$, and $\Lambda$, respectively. The approach by the block coordinate descent method has global convergence properties, Algorithm 3 also has global convergence.

# 4 Numerical experiment

In this section, we conduct numerical experiments to evaluate the validity of the proposed model. We make sure that the precision matrix estimated by the proposed model has sparse pattern. All computations were carried out on a machine with Intel(R) Atom(TM) 1.86GHz CPU and 2.00GB memory, and we implement all codes in MATLAB8.0.0 (R2012b).

In experiments, we maximize the log-likelihood function with the $L_1$ regularized term of the precision matrix by applying the block coordinate descent method to Algorithm 3. Then at M-Step, we must the following problem $(i = 1, ..., m)$ to obtain $\Lambda^{t+1}$ from (3.19).

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}(\Lambda_i \hat{\Sigma}_i) - \log(\det \Lambda_i) + \rho ||\Lambda_i||_1 \\ \text{subject to} & \Lambda_i \succeq 0. \end{array}$$

[10] shows that this problem is solved efficiently by applying interior point method. Thus we use the same method to solve this problem.

In this paper, we first generate the observations from the sample mixture distributions whose the parameters are $\tilde{\mu_i}$, $\tilde{\Lambda_i}\,(i = 1, ..., m)$. $\tilde{\mu_i} \in \mathbb{R}^d$ is a vector $((i-1)\eta, 0, ..., 0)^T$, which $\eta$ is a positive constant and indicates how far the centers of each cluster is. In other words, the larger *eta* is, the farer the centers of each cluster are. The sample precision matrix $\tilde{\Lambda_i} \in \mathbb{R}^{d \times d}$ is the sparse matrix that is generated arbitrarily. We use k-means method to set the staring iterates $\alpha^0$, $\mu^0$, $\Lambda^0$. Then we compute $\mu^0$ to the observations by k-means, and we compute the covariance matrices $\bar{\Sigma}_i (i = 1, ..., m)$ of the clusters divided by k-means. There, to ensure that $\Lambda_i^0$ is positive definite, we set $\Lambda_i^0 = (\bar{\Sigma}_i + I)^{-1}$, where $I$ is an unit matrix. Moreover given the random number to $\gamma_{ik}^0 \in M$, we compute $\alpha^0$ from (3.19). Now we compare the estimated precision matrices $\Lambda_i$ to the sample precision matrices $\tilde{\Lambda_i}$. Then we set $m = 2$ so that we can compare them easily. We also set up $\tilde{\Lambda_1}$ like a arrow matrix, $\tilde{\Lambda_2}$ like a tridiagonal matrix.

*Experiment 1* : We conduct the experiments by changing the value of $\rho$. The following figures illustrate what the nonzero elements of the precision matrices are visualized when we set $d = 20$, $n = 2000$, $\eta = 5$. Figure 1 presents the sample precision matrices, while Figure 2. 3. 4 present the estimated precision matrices. We note that all the estimated matrices are the same structures as the sample precision matrices. Thus the proposed model estimates the exact precision matrices relatively. In particular, the number of the nonzero elements in the estimated matrices is fewer than that of the arrow matrix $\tilde{\Lambda_1}$ by the $L_1$ regularized term. However, on the other hand, the number of the nonzero elements increases for the estimated matrices to the tridiagonal matrix $\tilde{\Lambda_2}$. This shows the effect by the parameter $\rho$. If the value of $\rho$ is larger, the accuracy of estimate is worse, while if the value is smaller, the effect of the $L_1$ regularized term is smaller.

*Experiment 2* : When we fix the parameter $\rho = 0.1$, we conduct the experiments by changing the value of $\eta$. Then we use the same sample precision matrix on *Experiment 1*. The following Figure 5. 6. 7 illustrate the estimated precision matrices on $\eta = 0$, 5, 10. We can see the same result on Figure 6. 7. There we examine the value of $\gamma_{ik}^0$ that is responsibility for $x_k$ when we set $\eta = 5$, 10. If $\gamma_{ik}^0$ is 1 or 0, the observation $x_k$ belongs to an one cluster completely. On $\eta = 5$, $\gamma_{ik}^0$ is not 1 or 0 for every $k = 1, ..., n$, while on $\eta = 10$, $\gamma_{ik}^0$ is 1 or 0 for all $k = 1, ..., n$. Thus,the same result on Figure 6. 7 shows that the clustering by EM algorithm is well on $\eta = 5$. Moreover we note that the estimated precision matrix corresponding the arrow matrix on Figure 5 is sparser than the other Figure 6, 7. Considering the centers of each clusters coincides when $\eta = 0$, we can see that our estimation is well relatively. However, on any Figures, the estimated precision

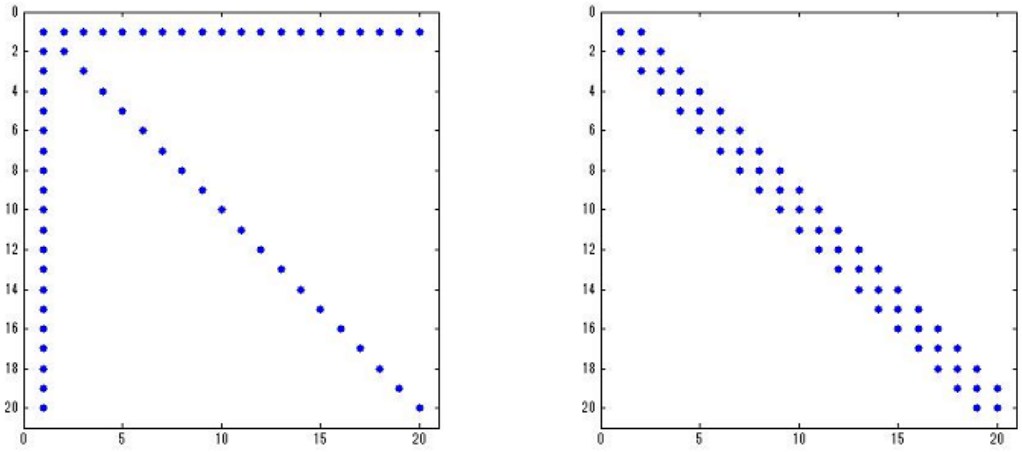

Figure 1: the sample precision matrices

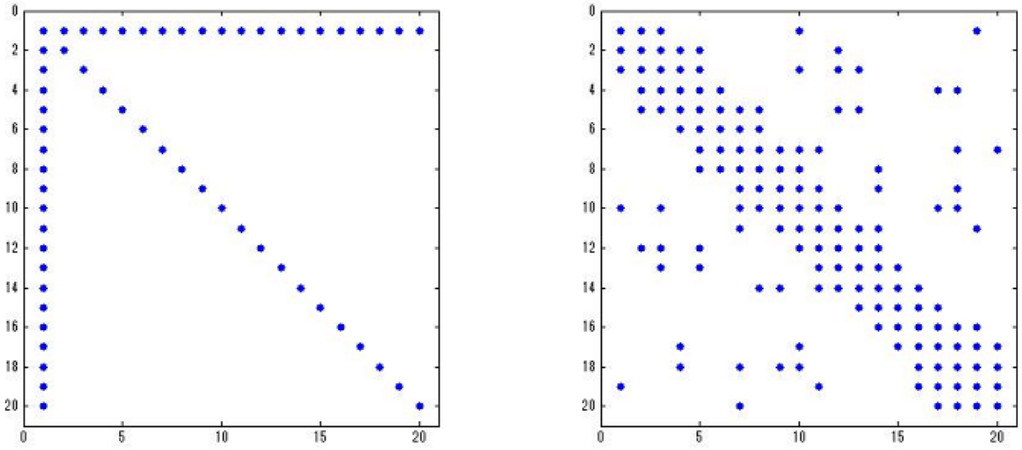

Figure 2: $\rho = 0.05$

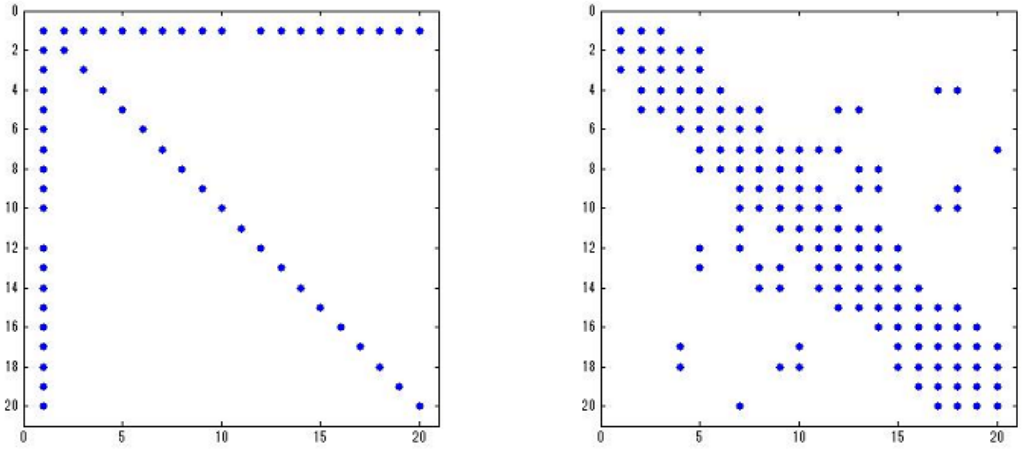

Figure 3: $\rho = 0.1$

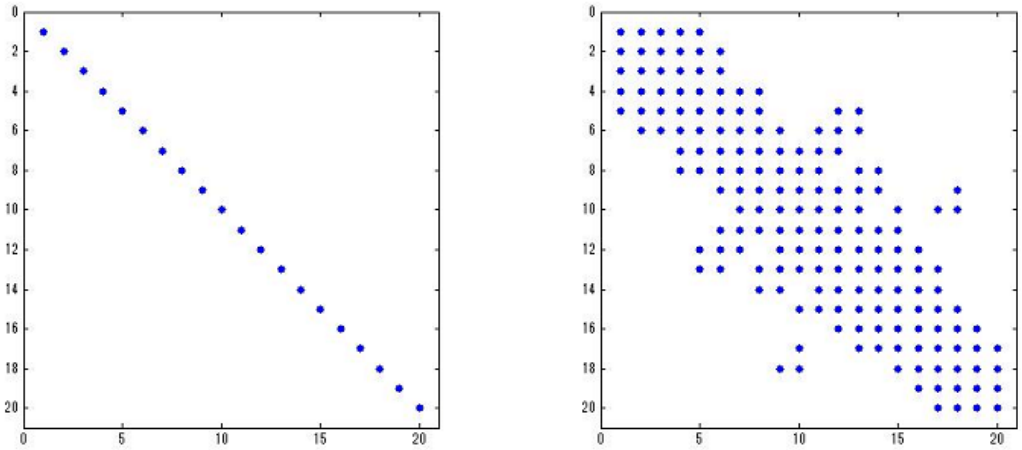

Figure 4: $\rho = 0.3$

matrices corresponding to the tridiagonal matrix $\tilde{\Lambda}_2$ are not the tridiagonal. This results show that the more complex the sample precision matrix is, the worse the accuracy of estimation is.

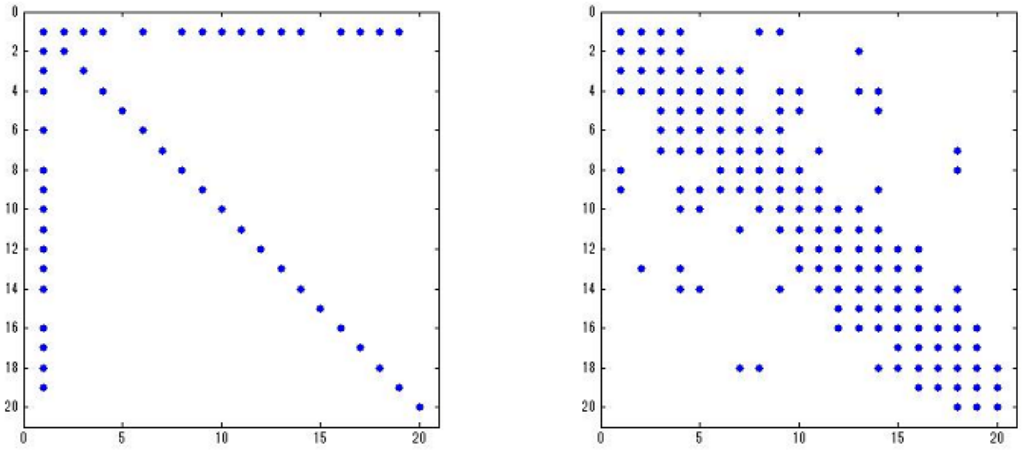

Figure 5: $\eta = 0$

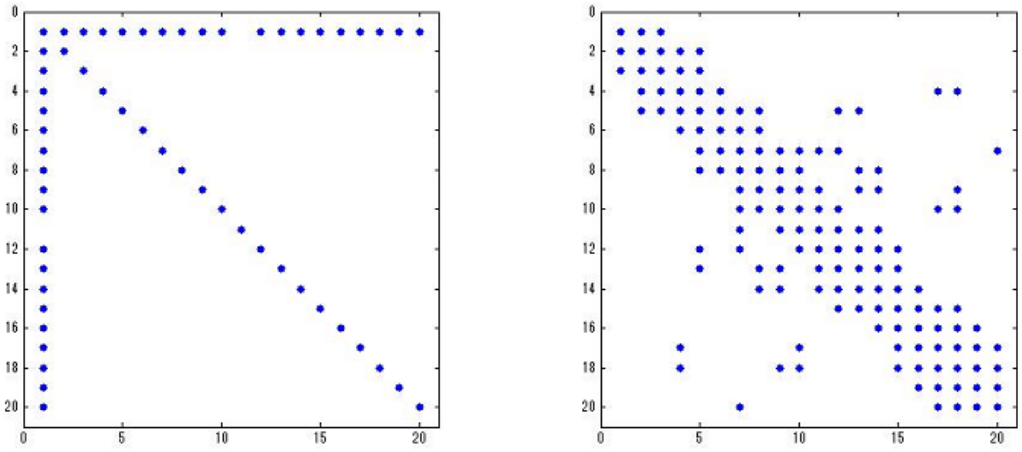

Figure 6: $\eta = 5$

## 5 Conclusion

In this paper, we proposed the simultaneous estimation model for normal mixture distributions and the sparse precision matrix. We also show the global convergence to the

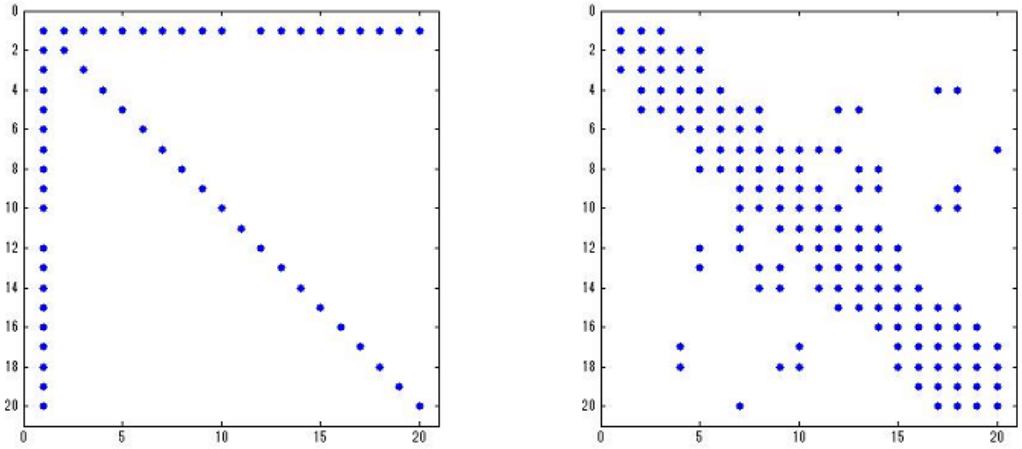

Figure 7: $\eta = 10$

algorithm for proposed model by applying the block coordinate descent method. Moreover we experiment for the proposed model. By changing the parameter $\rho$, we can see the effect of the $L_1$ regularized term, and by changing the position of the center of the sample cluster, we can see the clustering for mixture distributions is well done. However, when the sample sparse precision matrix is complex, the estimation is worse. Therefore it is important to study more accurate estimation for the complex precision matrix. It is also worth to try analyzing the real data with proposed model.

## Acknowledgments

First of all, I would like to express my sincere appreciation to Associate Professor Nobuo Yamashita for his advise. Although I sometimes troubled to him by my faults, he alwayssupported me kindly. It is an honor to have studied under him. I wouldlike to tender my acknowledgements to Assistant Professor Ellen Hidemi Fukuda for her numerous comments and encouragement. In addition, I also deeply thank all members of Yamashita Laboratory for their encouragements. Finally, I specially thank my family and my friends for their constant support.

## References

[1] A. P. Dempster, N. M. Laird and D.B. Rubin : Maximum likelihood from incomplete data via the EM Algorithm, Journal of the Royal Statistical Society B39, 1-38 1977.

[2] R. A. Redner and H. F. Walker : Mixture densities, maximum likelihood and the EM algorithm. SIAM Review, Vol. 26, pp.95-239 1984.

[3] C. F. J. Wu : On the convergence properties of the EM algorithm. Annals of Statistics, Vol. 11, pp. 95-103, 1983.

[4] R. J. Hathaway : Another Interpretation of The EM Algorithm For Mixture Distributions, Statistics & Probability Letters, 4, 2, 53-56 1986.

[5] S. L. Lauritzen, Graphical Models. Carendon Press, Oxford 1996.

[6] A. P. Dempster ; Covariance selection, Biometrics, 28, 1, 157-175 1972

[7] M. Yuan, and Y. Lin : Model selection and estimation in the normal graphical model, Biometrika, 94, 19-35 2007.

[8] J. Friedman, T. Hastie, and R. Tibshirani : Sparse inverse covariance estimation with the graphical lasso, Biostatistics, 9, 3, 432-441 2008.

[9] J. Honorio, and D. Samaras : Multi-task learning of normal graphical models, in Proceedings of the 27th Conference on Machine Learning 2010.

[10] L. Li, K.-C. Toh : An inexact interior point method for L1-regularized sparse covariance selection, Mathematical Programing, Computation, 2, 3-4, 291-315, 2010.

[11] R. Fletcher : A New Variational Result for Quasi-Newton Formulae, SIAM, Optimization, 1, 1, 18-21, 1991.

Master's Thesis

# Simultaneous Likelihood Estimation for Normal Mixture Distributions and Sparse Precision Matrix

Guidance

Associate Professor Nobuo YAMASHITA

Kazuki MATSUDA

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University


February 2014

| | | |
|---|---|---|
| Simultaneous Likelihood Estimation for Normal Mixture Distributions and Sparse Precision Matrix | Kazuki MATSUDA | February 2014 |

# Simultaneous Likelihood Estimation for Normal Mixture Distributions and Sparse Precision Matrix

Kazuki MATSUDA

**Abstract**

In this paper we consider both clustering and graphical modeling for given data. The clustering is the task of grouping of the data, while the graphical modeling provides a conditional dependence structure among variables in the data. In this paper we suppose that the data obeys a mixture of normal distributions. Then, we may apply the existing methods based on the maximum likelihood estimation, that is, the Expectation Maximization (EM) algorithm and the $L_1$ regularized maximum likelihood estimation. The EM algorithm provides clusters such that each cluster obeys a single normal distribution. The $L_1$ regularized maximum likelihood estimation finds a sparse precision matrix whose nonzero element represents a dependency of the corresponding variables. It assumes that the data obeys a single normal distribution. Thus we may apply it for each cluster given by the EM algorithm. However, this procedure estimates two different mixture distributions by two algorithms, which should be the same.

In this paper we propose a simultaneous estimation model for mixture distributions and a sparse precision matrix from the given data. We first formulate a maximization problem of the log likelihood function of mixture distribution with the $L_1$ regularized term of the precision matrix. We then propose a coordinate descent method for solving the problem. The proposed method is a generalization of the EM algorithm. We present some numerical results that show the validity of the proposed model.