Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 11, No. 4, pp. 669–686, October 2015

A BLOCK COORDINATE DESCENT METHOD FOR MAXIMUM LIKELIHOOD ESTIMATION PROBLEMS OF MIXTURE DISTRIBUTIONS*

Yuya Yamakawa and Nobuo Yamashita

Abstract: In this paper, we propose a block coordinate descent method for a maximum likelihood estimation problem of mixture distributions. The maximum likelihood estimation problem considered in the paper consists of not only a log-likelihood function but also some proper convex functions such as the L_1 regularization and/or indicator functions. Then, we may estimate parameters in distributions with some regularizations and constraints on parameters. Especially, a parameter estimation with lower constraints on mixture coefficients is the main contribution of this paper. Since the problem has the additional convex function, it cannot be solved by the usual EM algorithm. Thus, we consider a block coordinate descent method as a solver of the problem, and show its global convergence by using the results in Tseng (J. Optim. Theory Appl. 109–3, 2001). Moreover, we discuss concrete implementations when a distribution is Gaussian. In particular, we propose efficient methods for a maximum likelihood estimation for special cases.

Key words: *maximum likelihood estimation, mixture distributions, mixture coefficient, block coordinate descent method, regularization, EM algorithm*

Mathematics Subject Classification: *62-07, 62F30, 90C30, 49M27*

1 Introduction

When some observational data $\{x_i\} \subset \mathbf{R}^d$ are supposed to obey a certain distribution $p(x|\theta)$ with parameters $\theta \in \Theta$, it is important to estimate valid parameters $\overline{\theta}$ from the data. A maximum likelihood estimation is one of the estimation procedure of the parameters θ. In this paper, we focus on the maximum likelihood estimation of parameters in mixture distributions, which is frequently used in statistics, machine learning and clustering [1, 11]. The EM algorithm is known to be one of the most powerful method for the estimation [5, 7, 13, 16], and it have been studied actively even in recent years.

The EM algorithm is an iterative method which consists of the Expectation Step (E-Step) and the Maximization Step (M-Step). The E-Step calculates an expectation of the likelihood and the M-Step maximizes the expectation with respect to parameters. Redner and Walker [13] considered the EM algorithm for the estimation of mixture distributions and gave some concrete calculations of the E-Step and M-Step for various distributions, such as Gaussian mixtures. Moreover, they discussed its theoretical convergence property. Hathaway [7] interpreted the EM algorithm as a special case of block coordinate descent methods.

*This work was partially supported by JSPS KAKENHI Grant Number 25330025.

Recently, many researchers actively study a maximum likelihood estimation with regularization methods. For example, when we add the L_1 regularization $\|\theta\|_1$ to the likelihood function, we may choose important parameters in the model. The L_1 regularization is used for a sparse precision matrix selection in Gaussian mixtures [9, 18]. Since the (i, j)-th element of the precision matrix expresses the relation between the i-th and the j-th probability variables of x, the sparse precision matrix plays a critical role in the graphical modeling [6]. Ruan, Yuan and Zou [14] proposed the EM algorithm for Gaussian mixtures with the L_1 regularization, and succeeded in estimating parameters with sparse precision matrices.

In this paper, we first define a maximum likelihood estimation problem, whose objective function consists of not only a log-likelihood function but also some proper convex functions. If we exploit the L_1 regularization term $\|\theta\|_1$ and/or an indicator function δ_S of a constraint set $S \subset \Theta$ as the additional convex function, we can estimate parameters with the regularization and/or the constraint $\theta \in S$. Especially, the parameter estimation with lower constraints on mixture coefficients is the main contribution of this paper. Thanks to such constraints, we can obtain some theoretically and practically nice properties. Meanwhile, the estimation problem considered in the paper is more general than that in [14].

The estimation problem in the paper might not be solved by the usual EM algorithm. Then, we consider a block coordinate descent (BCD) method. At each iteration of a BCD method, the objective function is minimized among a few parameters while all the other parameters are fixed.

Since the log-likelihood function is not separable for each parameter in the mixture distributions, we first construct a separable problem related to the original one. Then, we apply a BCD method to the separable problem, where the block corresponds to a set of parameters in the single distribution. Tseng [15] showed that a BCD method for a nondifferentiable minimization problem has the global convergence property under some reasonable conditions. Using his result, we prove the global convergence of the proposed BCD method when we add certain lower bound constraints on the mixture coefficients. In addition, we discuss efficient implementations for some concrete problems, such as the maximum likelihood estimation with box constraints on mixture coefficients.

The present paper is organized as follows. In Section 2, we introduce a maximum likelihood estimation for mixture distributions. In particular, we present a general maximum likelihood estimation problem that has a log-likelihood function and some proper convex functions, such as the L_1 regularization and/or indicator functions of constraint sets with respect to parameters. In Section 3, we present a BCD method for the maximum likelihood estimation problem, and discuss its global convergence. In Section 4, we discuss how to solve subproblems in the BCD method for some special cases. In Section 5, we report some numerical results for the maximum likelihood estimation problems with some constraints. Finally, we make some concluding remarks in Section 6.

Throughout this paper, we use the following notations. Let p and q be positive integers. For a vector $v \in \mathbf{R}^p$ and a matrix $M \in \mathbf{R}^{p\times q}$, v_i denotes the i-th element of the vector v, and M_{ij} denotes the (i, j)-th element of the matrix M. The superscript $\top$ denotes the transposition of a vector or a matrix. For a square matrix $M \in \mathbf{R}^{p\times p}$, $\mathrm{tr}(M)$ denotes the trace of the matrix M, and norm $\|\cdot\|_1$ is defined by $\|M\|_1 := \sum_{i=1}^{p}\sum_{j=1}^{p}|M_{ij}|$. The notation $\mathbf{S}^p$ denotes a set of $p \times p$ real symmetric matrices. For $M \in \mathbf{S}^p$, $M \succ 0$ ($M \succeq 0$) means that M is positive (semi)definite. Moreover, for $A, B \in \mathbf{S}^p$, $A \succ B$ ($A \succeq B$) means that $A - B \succ 0$ ($A - B \succeq 0$). For a positive semidefinite matrix $A \in \mathbf{S}^p$, $A^{\frac{1}{2}}$ denotes the positive semidefinite matrix such that $A = A^{\frac{1}{2}}A^{\frac{1}{2}}$. For $d_1, \ldots, d_p \in \mathbf{R}$, $\mathrm{diag}(d_1, \ldots, d_p)$ denotes a diagonal matrix whose (i, i)-th element is d_i. For $x \in \mathbf{R}$, $\exp(x)$ denotes e^x, where e is Napier's constant.

2 The Maximum Likelihood Estimation for Mixture Distributions

In this section, we introduce a maximum likelihood estimation problem for mixture distributions.

Assume that probability variables $x \in \mathbf{R}^d$ obey a probability distribution $p(x)$. If $p(x)$ is expressed as the weighted linear combination on m distributions $p_i(x|\theta_i)$:

$$p(x) := \sum_{i=1}^{m} \alpha_i p_i(x|\theta_i),$$

$p(x)$ is called a *mixture distribution*, where $p_i(x|\theta_i)$ is called a mixture component, θ_i denotes the parameters of the i-th mixture component $p_i(x|\theta_i)$, and $\alpha_i \in \mathbf{R}^m$ is called a *mixture coefficient* satisfying

$$\sum_{i=1}^{m} \alpha_i = 1, \quad \alpha_i \geq 0, \quad i = 1, \ldots, m.$$

We express a mixture distribution with parameters (α, θ) by

$$p(x|\alpha, \theta) := \sum_{i=1}^{m} \alpha_i p_i(x|\theta_i), \tag{2.1}$$

where $\alpha := [\alpha_1, \ldots, \alpha_m]^\top$ and $\theta := [\theta_1, \ldots, \theta_m]$.

Suppose that we have observational data $X := [x_1, \ldots, x_n] \in \mathbf{R}^{d \times n}$. Then, we wish to model data X using the mixture distribution $p(x|\alpha, \theta)$. To this end, we consider an estimation of parameters (α, θ).

The joint probability for the observational data $X := [x_1, \ldots, x_n] \in \mathbf{R}^{d \times n}$ is given by

$$P(X|\alpha, \theta) := \prod_{k=1}^{n} p(x_k|\alpha, \theta).$$

We call $P(X|\alpha, \theta)$ a *likelihood*. Moreover, a maximizer (α^*, θ^*) of a likelihood is called a *maximum likelihood estimator*. In what follows, an estimation of parameters means that we obtain a maximum likelihood estimator. Since a maximization problem of a likelihood is difficult in general, we usually maximize the following log-likelihood function:

$$L(\alpha, \theta) := \log P(X|\alpha, \theta) = \sum_{k=1}^{n} \log \left(\sum_{i=1}^{m} \alpha_i p_i(x_k|\theta_i) \right).$$

We sometimes want to maximize the log-likelihood function $L(\alpha, \theta)$ with regularizations and/or constraints on some parameters in (α, θ). Thus, we consider the following maximization problem:

$$\begin{array}{ll} \text{maximize} & L(\alpha, \theta) - f(\alpha, \theta), \\ \text{subject to} & \alpha \in \Omega^\ell, \ \theta_i \in \Theta_i, \ i = 1, \ldots, m, \end{array} \tag{2.2}$$

where the function $f : \Omega^\ell \times \Theta \to \mathbf{R}$ is proper convex and lower semicontinuous, and the sets Ω^ℓ and Θ are defined by

$$\Omega^\ell := \left\{ \alpha \in \mathbf{R}^m \;\middle|\; \sum_{i=1}^{m} \alpha_i = 1, \ \alpha_i \geq \ell_i, \ i = 1, \ldots, m \right\}, \quad \Theta := \Theta_1 \times \ldots \times \Theta_m,$$

and the set Θ_i is a parameter space of θ_i and $\ell = [\ell_1, \ldots, \ell_m] \in \mathbf{R}^m$ is the constant vector such that $\ell_i \in [0, 1]$ $(i = 1, \ldots, m)$ and $\sum_{i=1}^m \ell_i < 1$. In what follows, we call problem (2.2) a maximum likelihood estimation problem. Note that the function f is regarded as a generalization of the L_1 regularization and indicator functions of constraint sets. Note also that $\ell = 0$ in [7,13,14,16]. To the author's best knowledge, this is the first time to consider the lower bounds $\alpha_i \geq \ell_i$ $(i = 1, \ldots, m)$ in the maximum likelihood estimation for mixture distributions. As seen in Sections 3 and 5, the lower bounds with $\ell_i > 0$ $(i = 1, \ldots, m)$ bring in both theoretically and practically nice effects.

We now give two concrete cases of problem (2.2).

Example 1. The maximum likelihood estimation with constraints on mixture coefficients

We discuss the maximum likelihood estimation with constraints on mixture coefficients. We assume that the mixture coefficients satisfy $\alpha_i \in [\ell_i, u_i]$ $(i = 1, \ldots, m)$, where $\ell_i, u_i \in (0, 1]$ $(i = 1, \ldots, m)$, $\sum_{i=1}^m \ell_i < 1$ and $\sum_{i=1}^m u_i \geq 1$. Then, we may define the function f of problem (2.2) as

$$f(\alpha, \theta) := \begin{cases} 0 & (\alpha \in \Gamma) \\ +\infty & (\alpha \notin \Gamma) \end{cases}, \quad \Gamma := \{\ \alpha \in \Omega^\ell \mid \alpha_i \leq u_i,\ i = 1, \ldots, m\ \}.$$

As described above, the constraints $\ell_i \leq \alpha_i$ $(i = 1, \ldots, m)$ play a critical role in the theoretical and practical aspects. In the theoretical aspect, these constraints enable us to show the global convergence of the BCD method proposed in Section 3. In the practical aspect, these constraints bring in some valid parameter estimations when the amount of the observational data is small.

Example 2. The maximum likelihood estimation with the L_1 regularization for Gaussian mixtures

Suppose that the distributions $p_i(x|\theta_i)$ $(i = 1, \ldots, m)$ in (2.1) are Gaussian distributions:

$$\mathcal{N}(x|\mu_i, \Lambda_i^{-1}) := \frac{\sqrt{\det \Lambda_i}}{(2\pi)^{d/2}} \exp\left[-\frac{1}{2}(x - \mu_i)^\top \Lambda_i (x - \mu_i)\right], \quad i = 1, \ldots, m,$$

where μ_i and Λ_i denote a mean vector and a precision matrix which is the inverse of a covariance matrix. Then, $\theta_i = [\mu_i, \Lambda_i]$ $(i = 1, \ldots, m)$. Friedman, Hastie and Tibshirani [6] and Lu [10] proposed the maximum likelihood estimation with the L_1 regularization. We apply such ideas to the maximum likelihood estimation for mixture distributions. Then, we may consider the following problem:

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{k=1}^n \log\left(\sum_{i=1}^m \alpha_i \mathcal{N}(x_k|\mu_i, \Lambda_i^{-1})\right) - \sum_{i=1}^m \rho_i \|\Lambda_i\|_1, \\ \text{subject to} & \alpha \in \Omega^0,\ \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I,\ i = 1, \ldots, m, \end{array} \tag{2.3}$$

where $\rho_i, \underline{\lambda}_i, \overline{\lambda}_i$ $(i = 1 \ldots, m)$ are constants such that $\rho_i \in [0, \infty)$, $\underline{\lambda}_i \in [0, \infty)$, $\overline{\lambda}_i \in (0, \infty]$ and $\underline{\lambda}_i < \overline{\lambda}_i$. We allow $\overline{\lambda}_i$ to be $+\infty$. Note that $\underline{\lambda}_i = 0$ and $\overline{\lambda}_i = \infty$ in [6,10].

Thanks to the L_1 regularization term $\sum_{i=1}^m \rho_i \|\Lambda_i\|_1$, we can obtain a maximum likelihood estimator with sparse precision matrices. The sparse precision matrix plays an important

role in the graphical modeling. For these details, see [6, 9, 10, 14]. Problem (2.3) is written as (2.2) with $\Theta_i = \mathbf{R}^d \times \mathbf{S}^d$ $(i = 1, \ldots, m)$,

$$f(\alpha, \theta) := \sum_{i=1}^{m} f_i(\Lambda_i), \quad f_i(\Lambda_i) := \begin{cases} \rho_i \|\Lambda_i\|_1 & (\ \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I\ ) \\ +\infty & (\ \text{otherwise}\ ), \end{cases} \quad i = 1, \ldots, m.$$

3 A Block Coordinate Descent Method for the Maximum Likelihood Estimation Problem and its Global Convergence

In this section, we present a BCD method solving the maximum likelihood estimation problem (2.2). To this end, we first construct a separable problem suitable to the proposed BCD method. Next, we give conditions under which the proposed BCD method has the global convergence property.

If a BCD method is directly applied to problem (2.2), then it may solve the following subproblems at each step:

$$\begin{aligned}
\alpha^{t+1} &:= \operatorname*{argmax}_{\alpha \in \Omega^\ell} \left\{ L(\alpha, \theta^t) - f(\alpha, \theta^t) \right\}, \\
\theta_1^{t+1} &:= \operatorname*{argmax}_{\theta_1 \in \Theta_1} \left\{ L(\alpha^{t+1}, \theta_1, \theta_2^t, \ldots, \theta_m^t) - f(\alpha^{t+1}, \theta_1, \theta_2^t, \ldots, \theta_m^t) \right\}, \\
\theta_2^{t+1} &:= \operatorname*{argmax}_{\theta_2 \in \Theta_2} \left\{ L(\alpha^{t+1}, \theta_1^{t+1}, \theta_2, \theta_3^t, \ldots, \theta_m^t) - f(\alpha^{t+1}, \theta_1^{t+1}, \theta_2, \theta_3^t, \ldots, \theta_m^t) \right\}, \\
&\vdots \\
\theta_m^{t+1} &:= \operatorname*{argmax}_{\theta_m \in \Theta_m} \left\{ L(\alpha^{t+1}, \theta_1^{t+1}, \ldots, \theta_{m-1}^{t+1}, \theta_m) - f(\alpha^{t+1}, \theta_1^{t+1}, \ldots, \theta_{m-1}^{t+1}, \theta_m) \right\},
\end{aligned}$$

where the superscript t denotes the t-th iteration. We see that the subproblems cannot be solved in parallel because the log-likelihood function L included in (2.2) has the weighted linear combination of the probability density function in the antilogarithm part. Thus, we construct a separable problem associated with (2.2) in order to solve subproblems in parallel.

To this end, we assume that the function f is separable with respect to $\alpha, \theta_1, \ldots, \theta_m$, that is, it is written as

$$f(\alpha, \theta) = f_0(\alpha) + \sum_{i=1}^{m} f_i(\theta_i), \tag{3.1}$$

where f_0 is a lower semicontinuous and proper convex function for adding some constraints on mixture coefficients α_i $(i = 1 \ldots, m)$, and f_i $(i = 1, \ldots, m)$ are also lower semicontinuous and proper convex functions for adding some constraints on parameters θ_i $(i = 1, \ldots, m)$, respectively.

Then, we consider the following minimization problem instead of problem (2.2):

$$\begin{array}{ll} \text{minimize} & F(W, \alpha, \theta), \\ \text{subject to} & W \in M,\ \alpha \in \Omega^\ell,\ \theta_i \in \Theta_i,\ i = 1, \ldots, m, \end{array} \tag{3.2}$$

where

$$F(W, \alpha, \theta) := D(W, \alpha, \theta) + f_0(\alpha) + \sum_{i=1}^{m} f_i(\theta_i),$$

$$M := \left\{ W \in \mathbf{R}^{m\times n} \;\middle|\; 0 \le W_{ik},\ \sum_{i=1}^{m} W_{ik} = 1,\ k = 1,\dots,n \right\},$$

and the function $D : M \times \Omega^\ell \times \Theta \to \mathbf{R}$ is defined by

$$D(W,\alpha,\theta) := \sum_{i=1}^{m}\sum_{k=1}^{n} W_{ik}\{\log W_{ik} - \log\alpha_i - \log p_i(x_k|\theta_i)\}. \tag{3.3}$$

Note that the decision variables of problem (3.2) are α, θ and W. Note also that, if we apply a BCD method to problem (3.2), then the objective function of (3.2) is separable for α and θ_i $(i = 1,\dots,m)$ when W is fixed. These details are discussed in Section 4.

Now we mention that we can obtain a solution of (2.2) if we apply a BCD method to problem (3.2). Let $g : \Omega^\ell \times \Theta \to \mathbf{R}$ be defined by

$$g(\alpha,\theta) := \min_{W\in M} D(W,\alpha,\theta). \tag{3.4}$$

Note that for each (α,θ), the function $D(\cdot,\alpha,\theta)$ is strictly convex on the compact set M, and hence the right-hand side of (3.4) has the unique minimizer. The next lemma shows that $g(\alpha,\theta) = -L(\alpha,\theta)$, i.e., problem (2.2) is equivalent to

$$\begin{aligned} &\text{minimize} && g(\alpha,\theta) + f_0(\alpha) + \sum_{i=1}^{m} f_i(\theta_i), \\ &\text{subject to} && \alpha \in \Omega^\ell,\ \theta_i \in \Theta_i,\ i = 1,\dots,m. \end{aligned} \tag{3.5}$$

The equivalence is implicitly given in [7]. Here, we give its proof for the completeness of the paper.

Lemma 3.1. *For each $\alpha \in \Omega^\ell$ and $\theta \in \Theta$, $g(\alpha,\theta) = -L(\alpha,\theta)$.*

Proof. Let W^* be a solution of $\min_{W\in M} D(W,\alpha,\theta)$. The KKT conditions for $\min_{W\in M} D(W,\alpha,\theta)$ are written as

$$\sum_{i=1}^{m} W^*_{ik} = 1, \quad \log W^*_{ik} + 1 - \log\alpha_i p_i(x_k|\theta_i) - u^*_k = 0, \quad i = 1,\dots,m,\ k = 1,\dots,n,$$

where u^*_k is a Lagrange multiplier for $\sum_{i=1}^m W^*_{ik} = 1$. Then,

$$W^*_{ik} = \alpha_i p_i(x_k|\theta_i)\exp(u^*_k - 1), \quad i = 1,\dots,m,\ k = 1,\dots,n. \tag{3.6}$$

It further follows from $\sum_{i=1}^m W^*_{ik} = 1$ that

$$1 = \sum_{i=1}^{m} W^*_{ik} = \exp(u^*_k - 1)\sum_{i=1}^{m}\alpha_i p_i(x_k|\theta_i) = \exp(u^*_k - 1)p(x_k|\alpha,\theta), \quad k = 1,\dots,n,$$

and hence

$$\exp(u^*_k - 1) = \frac{1}{p(x_k|\alpha,\theta)}, \quad k = 1,\dots,n. \tag{3.7}$$

Thus, (3.6) and (3.7) yield that

$$W^*_{ik} = \frac{\alpha_i p_i(x_k|\theta_i)}{p(x_k|\alpha,\theta)}, \quad i = 1,\dots,m,\ k = 1,\dots,n. \tag{3.8}$$

Moreover, we have

$$\begin{aligned} g(\alpha,\theta) &= D(W^*,\alpha,\theta) \\ &= \sum_{i=1}^{m}\sum_{k=1}^{n} W^*_{ik}\left\{\log\frac{\alpha_i p_i(x_k|\theta_i)}{p(x_k|\alpha,\theta)} - \log\alpha_i p_i(x_k|\theta_i)\right\} \\ &= \sum_{i=1}^{m}\sum_{k=1}^{n} W^*_{ik}\left\{\log\alpha_i p_i(x_k|\theta_i) - \log p(x_k|\alpha,\theta) - \log\alpha_i p_i(x_k|\theta_i)\right\} \\ &= -\sum_{k=1}^{n}\left(\sum_{i=1}^{m} W^*_{ik}\right)\log p(x_k|\alpha,\theta) \\ &= -L(\alpha,\theta), \end{aligned}$$

where the last equality follows from $\sum_{i=1}^{m} W^*_{ik} = 1$. □

From Lemma 3.1, problem (2.2) is equivalent to problem (3.5), that is, their global solutions coincide. Moreover, (W,α,θ) is a global optimum of (3.2) if and only if (α,θ) is a global optimum of (3.5).

Remark 3.2. Lemma 3.1 does not state that (α,θ) is a stationary point of (3.5) when (W,α,θ) is a stationary point of (3.2).

We now apply a BCD method to problem (3.2). Let (α^t,θ^t) be given. The BCD method first solves the right-hand side of (3.4), that is,

$$W^t := \operatorname*{argmin}_{W\in M} D(W,\alpha^t,\theta^t).$$

From (3.8) in the proof of Lemma 3.1, the solution W^t is given by

$$W^t_{ik} = \frac{\alpha^t_i p_i(x_k|\theta^t_i)}{p(x_k|\alpha^t,\theta^t)}, \quad i = 1,\dots,m,\ k = 1,\dots,n. \tag{3.9}$$

Next, it solves the following subproblems with respect to α and θ_i $(i = 1,\dots,m)$ independently:

$$\operatorname*{minimize}_{\alpha\in\Omega^\ell} \quad -\sum_{i=1}^{m}\sum_{k=1}^{n} W^t_{ik}\log\alpha_i + f_0(\alpha), \tag{3.10}$$

$$\operatorname*{minimize}_{\theta_i\in\Theta_i} \quad -\sum_{k=1}^{n} W^t_{ik}\log p_i(x_k|\theta_i) + f_i(\theta_i). \tag{3.11}$$

Note that the functions $f_0, f_1,\dots,f_m$ are given by (3.1). Summing up the above discussion, the BCD method is described as follows.

Algorithm 3.3.

Step 0. Choose an initial point $(\alpha^0,\theta^0) \in \mathbf{R}^m \times \Theta$, and set $t := 0$.

Step 1. Calculate W^t by (3.9).

Step 2. Obtain a solution α^{t+1} to problem (3.10).

Step 3. For each $i \in \{1, \ldots, m\}$, obtain a solution θ_i^{t+1} to problem (3.11).

Step 4. If an appropriate termination criterion is satisfied, then stop. Otherwise, set $t := t + 1$ and go to Step 1.

Next, by using the result of [15], we give conditions under which Algorithm 3.3 has the global convergence property. To this end, we provide definitions of a stationary point and a coordinatewise minimum point of the following problem:

$$\text{minimize} \quad F(\xi_1, \ldots, \xi_r) := D(\xi_1, \ldots, \xi_r) + \sum_{i=1}^{r} f_i(\xi_i), \tag{3.12}$$

where $D : \mathbf{R}^{n_1+\cdots+n_r} \to \mathbf{R} \cup \{\infty\}$ is differentiable and $f_i : \mathbf{R}^{n_k} \to \mathbf{R} \cup \{\infty\}$ $(i = 1, \ldots, r)$ are nondifferentiable. We say that z is a stationary point of (3.12) if $z \in \mathrm{dom}F := \{\ \xi \in \mathbf{R}^{n_1+\cdots+n_r} \mid F(\xi) < \infty\ \}$ and

$$F'(z; d) := \liminf_{\tau \downarrow 0} \frac{F(\xi + \tau d) - F(\xi)}{\tau} \geq 0 \quad \text{for all } d \in \mathbf{R}^{n_1+\cdots+n_r}.$$

We say that z is a coordinatewise minimum point of (3.12) if $z \in \mathrm{dom}F$ and

$$F(z + (0, \ldots, d_k, \ldots, 0)) \geq F(z) \quad \text{for all } d_k \in \mathbf{R}^{n_k} \text{ and } k = 1, \ldots, r,$$

where $(0, \ldots, d_k, \ldots, 0) \in \mathbf{R}^{n_1+\cdots+n_r}$ denotes the vector whose k-th coordinate block is d_k and whose other coordinates are zero. In addition, we say that F is hemivariate if F is not constant on any line segment of $\mathrm{dom}F$, that is, if there exist no distinct points $\xi, \zeta \in \mathrm{dom}F$ such that $\tau\xi + (1-\tau)\zeta \in \mathrm{dom}F$ and $F(\xi) = F(\tau\xi + (1-\tau)\zeta)$ for all $\tau \in [0, 1]$.

Theorem 3.4. *Suppose that Algorithm 3.3 generates an infinite sequence* $\{(W^t, \alpha^{t+1}, \theta^{t+1})\}$. *Suppose also that the sequence* $\{(W^t, \alpha^{t+1}, \theta^{t+1})\}$ *has an accumulation point* $(\overline{W}, \overline{\alpha}, \overline{\theta})$. *If the following conditions (i)–(iv) hold, then* $(\overline{W}, \overline{\alpha}, \overline{\theta})$ *is a stationary point of problem (3.2).*

(i) *The functions* f_0 *and* f_i $(i = 1, \ldots, m)$ *are lower semicontinuous and proper convex.*

(ii) *For each* $x_k \in \{x_1, \ldots, x_n\}$ *and* $i \in \{1, \ldots, m\}$, *the function* $p_i(x_k|\cdot)$ *is continuous on* Θ_i *and* $p_i(x_k|\theta_i) > 0$ *for all* $\theta_i \in \Theta_i$. *Moreover, the function* $-\log p_i(x_k|\cdot)$ *is convex and hemivariate on* Θ_i.

(iii) *There exists* $\underline{\alpha} \in \mathbf{R}^m$ *such that* $\alpha_i^t \geq \underline{\alpha}_i > 0$ $(i = 1, \ldots, m)$.

(iv) *The function* D *is differentiable at* $(\overline{W}, \overline{\alpha}, \overline{\theta})$.

Proof. From assumption (iii), we may replace D in problem (3.2) with $\overline{D}$ defined by

$$\overline{D}(W, \alpha, \theta) := \begin{cases} D(W, \alpha, \theta) & \text{if } W \in M,\ \alpha \in \Omega^{\underline{\alpha}},\ \theta_i \in \Theta_i\ (i = 1, \ldots, m), \\ +\infty & \text{otherwise.} \end{cases}$$

Then, from assumption (ii), $\mathrm{dom}\overline{D} := \{\ (W, \alpha, \theta) \in M \times \Omega^{\underline{\alpha}} \times \Theta \mid \overline{D}(W, \alpha, \theta) < \infty\ \} = M \times \Omega^{\underline{\alpha}} \times \Theta_1 \times \ldots \times \Theta_m$. Thus, (C2) of [15, Proposition 5.1] holds. We also have (B1)–(B3)

of [15, Proposition 5.1] from assumptions (i) and (ii). Then [15, Proposition 5.1] shows that $(\overline{W}, \overline{\alpha}, \overline{\theta})$ is a coordinatewise minimum point of (3.2), that is,

$$\begin{aligned}
&F(W, \overline{\alpha}, \overline{\theta}) \geq F(\overline{W}, \overline{\alpha}, \overline{\theta}) \quad \text{for all } W \in M,\\
&F(\overline{W}, \alpha, \overline{\theta}) \geq F(\overline{W}, \overline{\alpha}, \overline{\theta}) \quad \text{for all } \alpha \in \Omega^{\alpha},\\
&F(\overline{W}, \overline{\alpha}, \theta_1, \overline{\theta}_2, \ldots, \overline{\theta}_m) \geq F(\overline{W}, \overline{\alpha}, \overline{\theta}) \quad \text{for all } \theta_1 \in \Theta_1,\\
&\qquad\qquad \vdots\\
&F(\overline{W}, \overline{\alpha}, \overline{\theta}_1, \ldots, \overline{\theta}_{m-1}, \theta_m) \geq F(\overline{W}, \overline{\alpha}, \overline{\theta}) \quad \text{for all } \theta_m \in \Theta_m.
\end{aligned}$$

It then follows from assumption (iv) that $(\overline{W}, \overline{\alpha}, \overline{\theta})$ is a stationary point of (3.2). □

Remark 3.5. For the global convergence, we should get exact solutions of subproblems (3.10) and (3.11). As shown in Section 4, we can get them in some special cases.

We now discuss when assumptions (i)–(iv) of Theorem 3.4 hold.

(1) When we employ indicator functions on closed convex sets and/or the L_1 regularization as f_0 and f_i $(i = 1, \ldots, m)$, assumption (i) holds.

(2) Some distributions, such as a logistic distribution, satisfy assumption (ii). For these details, see [3, Chapter 7].

(3) If we have a large amount of the observational data, assumption (iii) holds in many cases. Moreover, if ℓ_i $(i = 1, \ldots, m)$ in problem (2.2) are positive, then assumption (iii) certainly holds. Section 4.1 will present a solution method of problem (3.10) with the constraints.

(4) When assumptions (ii) and (iii) hold, $\overline{W}_{ik}$ is positive from (3.9). Thus, assumption (iv) holds when the function $p_i(x_k|\cdot)$ is differentiable at $\bar{\theta}_i$.

Unfortunately, a Gaussian distribution $\mathcal{N}(x|\mu_i, \Lambda_i^{-1})$ does not satisfy the convexity assumption in (ii). However, under some reasonable assumptions, we can construct a global convergent BCD method for Gaussian mixtures. Note that $\theta_i = [\mu_i, \Lambda_i]$ $(i = 1, \ldots, m)$ for Gaussian mixtures. In addition, we use the notations $\mu := [\mu_1, \ldots, \mu_m]$ and $\Lambda := [\Lambda_1, \ldots, \Lambda_m]$. We assume that the function f_i is separable with respect to μ_i and Λ_i for all $i \in \{1, \ldots, m\}$, that is,

$$f_i(\theta_i) = f_i^{\mu}(\mu_i) + f_i^{\Lambda}(\Lambda_i), \quad i = 1, \ldots, m, \tag{3.13}$$

where f_i^{μ} and f_i^{Λ} are lower semicontinuous and proper convex for all $i \in \{1, \ldots, m\}$. Then, we execute the following two steps instead of Step 3 in Algorithm 3.3.

Step 3-1. For each $i \in \{1, \ldots, m\}$, obtain a solution μ_i^{t+1} of the following problem:

$$\underset{\mu_i \in \mathbf{R}^d}{\text{minimize}} \quad -\sum_{k=1}^{n} W_{ik}^t \log \mathcal{N}(x_k|\mu_i, (\Lambda_i^t)^{-1}) + f_i^{\mu}(\mu_i). \tag{3.14}$$

Step 3-2. For each $i \in \{1, \ldots, m\}$, obtain a solution Λ_i^{t+1} of the following problem:

$$\underset{\Lambda_i \succeq 0}{\text{minimize}} \quad -\sum_{k=1}^{n} W_{ik}^t \log \mathcal{N}(x_k|\mu_i^{t+1}, \Lambda_i^{-1}) + f_i^{\Lambda}(\Lambda_i). \tag{3.15}$$

Note that the modified method is also a BCD method. We call it Algorithm 2 in the remainder of the paper.

The next theorem shows the global convergence of Algorithm 2.

Theorem 3.6. *Let* $p_i(x|\theta_i) := \mathcal{N}(x|\mu_i, \Lambda_i^{-1})$, $\theta_i := [\mu_i, \Lambda_i]$ $(i = 1, \ldots, m)$. *Suppose that Algorithm 2 generates an infinite sequence* $\{(W^t, \alpha^{t+1}, \mu^{t+1}, \Lambda^{t+1})\}$. *Suppose also that the sequence* $\{(W^t, \alpha^{t+1}, \mu^{t+1}, \Lambda^{t+1})\}$ *has an accumulation point* $(\overline{W}, \overline{\alpha}, \overline{\mu}, \overline{\Lambda})$. *If the following conditions (i)–(iii) hold, then* $(\overline{W}, \overline{\alpha}, \overline{\mu}, \overline{\Lambda})$ *is a stationary point of problem (3.2).*

(i) *The function* f_0 *is lower semicontinuous and proper convex, and the functions* f_i $(i = 1, \ldots, m)$ *are written as (3.13) with lower semicontinuous and proper convex functions* f_i^{μ}, f_i^{Λ} $(i = 1, \ldots, m)$.

(ii) *There exists* $\underline{\alpha} \in \mathbf{R}^m$ *such that* $\alpha_i^t \geq \underline{\alpha}_i > 0$ $(i = 1, \ldots, m)$.

(iii) *There exists* $\underline{\lambda} \in \mathbf{R}^m$ *such that* $\Lambda_i^t \succeq \underline{\lambda}_i I \succ 0$ $(i = 1, \ldots, m)$.

Proof. Note that $\overline{W}_{ik} > 0$ from (3.9) and assumption (ii). It then follows from assumption (iii) that D is differentiable at $(\overline{W}, \overline{\alpha}, \overline{\mu}, \overline{\Lambda})$. Consequently, we can prove this theorem in a way similar to the proof of Theorem 3.4. □

4 Implementation Issue for Special Cases

In this section, we describe efficient solution methods solving subproblems (3.10), (3.14) and (3.15) for special cases such as Examples 1 and 2 in Section 2.

4.1 The maximum likelihood estimation with constraints on mixture coefficients

We discuss the maximum likelihood estimation with box constraints on mixture coefficients as described in Example 1 of Section 2. Since the update of the mixture coefficients appears only in subproblem (3.10) of Step 2, we only discuss how to solve subproblem (3.10).

Subproblem (3.10) has simple constraints $\sum_{i=1}^m \alpha_i = 1$, $\ell_i \leq \alpha_i \leq u_i$ $(i = 1, \ldots, m)$. There exist efficient methods that solve special convex problems with the constraints in $O(m)$ [4, 12]. Although the objective function in (3.10) is different from those in [4, 12], we can construct an $O(m)$ method for (3.10) by using the ideas of [4, 12]. For the details, see [17, Subsection 5.4.1].

4.2 The maximum likelihood estimation for Gaussian mixtures

Now, we consider the case where mixture components are given by Gaussian distributions, that is, $p_i(x|\theta_i) := \mathcal{N}(x|\mu_i, \Lambda_i^{-1})$, $\theta_i := [\mu_i, \Lambda_i]$ $(i = 1, \ldots, m)$.

The maximum likelihood estimation for Gaussian mixtures is equivalent to problem (3.2) with $\Theta_i := \mathbf{R}^d \times \mathbf{S}^d$ $(i = 1, \ldots, m)$ and

$$f_0(\alpha) := \begin{cases} 0 & (\alpha \in \Omega^0) \\ +\infty & (\alpha \notin \Omega^0), \end{cases} \quad f_i^{\mu}(\mu_i) := 0, \; f_i^{\Lambda}(\Lambda_i) := \begin{cases} 0 & (\Lambda_i \succeq 0) \\ +\infty & (\Lambda_i \not\succeq 0), \end{cases} \quad i = 1, \ldots, m. \tag{4.1}$$

Then, $\alpha_i^{t+1}, \mu_i^{t+1}$ and Λ_i^{t+1} in Steps 2, 3-1 and 3-2 of Algorithm 2 are given by

$$\alpha_i^{t+1} = \frac{N_i^t}{n}, \; \mu_i^{t+1} = \frac{1}{N_i^t} \sum_{k=1}^n W_{ik}^t x_k, \; \Lambda_i^{t+1} = \left(\frac{1}{N_i^t} \sum_{k=1}^n W_{ik}^t (x_k - \mu_i^{t+1})(x_k - \mu_i^{t+1})^\top \right)^{-1} \tag{4.2}$$

where $N_i^t := \sum_{k=1}^n W_{ik}^t$. We see that (4.2) is equivalent to the EM algorithm. Note that the equivalence has already been pointed out in [16].

4.3 The maximum likelihood estimation for Gaussian mixtures with constraints on precision matrices

In this subsection, we consider the maximum likelihood estimation for Gaussian mixtures that has additional constraints on the precision matrices such that $\underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I$ $(i = 1, \ldots, m)$, where $\underline{\lambda}_i, \overline{\lambda}_i \in \mathbf{R}$ $(i = 1, \ldots, m)$ are constants such that $0 < \underline{\lambda}_i < \overline{\lambda}_i$.

In this case, we should replace f_i^Λ in (4.1) with

$$f_i^\Lambda(\Lambda_i) := \begin{cases} 0 & (\ \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I\ ) \\ \infty & (\ \text{otherwise}\ ), \end{cases} \qquad i = 1, \ldots, m.$$

Note that α_i^{t+1} and μ_i^{t+1} are also given by (4.2) because subproblems with respect to α_i and μ_i are same as those in Subsection 4.2. On the other hand, subproblem (3.15) with respect to Λ_i is different, and it is expressed as

$$\begin{array}{ll} \underset{\Lambda_i \in \mathbf{S}^d}{\text{minimize}} & \operatorname{tr}\left(A_i^t \Lambda_i\right) - \log\det\Lambda_i, \\ \text{subject to} & \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I, \end{array} \tag{4.3}$$

where

$$A_i^t := \frac{1}{N_i^t} \sum_{k=1}^n W_{ik}^t (x_k - \mu_i^{t+1})(x_k - \mu_i^{t+1})^\top, \tag{4.4}$$

and N_i^t is given in Subsection 4.2.

Thanks to the constraints $\underline{\lambda}_i I \preceq \Lambda_i$, the condition (iii) of Theorem 3.6 holds. Moreover, if we also add the constraints on mixture coefficients as described in Subsection 4.1, the condition (ii) of Theorem 3.6 also holds. Therefore, such constraints guarantee the global convergence of Algorithm 2.

Now, we discuss how to solve (4.3). As shown below, we can provide a solution of (4.3) analytically. For simplicity, let $A := A_i^t$, $\Lambda := \Lambda_i$, $\underline{\lambda} := \underline{\lambda}_i$ and $\overline{\lambda} := \overline{\lambda}_i$ in the rest of this subsection.

Since problem (4.3) is convex, $\Lambda_i^* \in \mathbf{S}^d$ satisfying the following KKT conditions is an optimal solution:

$$\begin{array}{lll} A - (\Lambda^*)^{-1} + U^* - V^* = 0, & (\underline{\lambda} I - \Lambda^*)U^* = 0, & (\overline{\lambda} I - \Lambda^*)V^* = 0, \\ \underline{\lambda} I \preceq \Lambda^* \preceq \overline{\lambda} I, & 0 \preceq U^*, & 0 \preceq V^*, \end{array} \tag{4.5}$$

where $U^* \in \mathbf{S}^d$ and $V^* \in \mathbf{S}^d$ are Lagrange multipliers for $\underline{\lambda} I \preceq \Lambda^*$ and $\Lambda^* \preceq \overline{\lambda} I$, respectively. We have from (4.5) that Λ^*, U^*, V^* and A commute mutually because U^* and V^* are symmetric matrices. This result and [8, Theorem 1.3.19] yield that Λ^*, U^*, V^* and A are simultaneously diagonalizable, that is, there exists an orthogonal matrix $P \in \mathbf{S}^d$ such that

$$\begin{array}{ll} P^\top \Lambda^* P = \operatorname{diag}(\lambda_1^*, \ldots, \lambda_d^*), & P^\top U^* P = \operatorname{diag}(u_1^*, \ldots, u_d^*), \\ P^\top V^* P = \operatorname{diag}(v_1^*, \ldots, v_d^*), & P^\top A P = \operatorname{diag}(a_1, \ldots, a_d), \end{array}$$

where $\lambda_j^*, u_j^*, v_j^*$ and a_j $(j = 1, \ldots, d)$ are eigenvalues of matrices Λ^*, U^*, V^* and A, respectively. Pre- and post-multiplying (4.5) by $P^\top$ and P, respectively,

$$\begin{array}{lll} a_j - (\lambda_j^*)^{-1} + u_j^* - v_j^* = 0, & (\underline{\lambda} - \lambda_j^*)u_j^* = 0, & (\overline{\lambda} - \lambda_j^*)v_j^* = 0, \\ \underline{\lambda} \le \lambda_j^* \le \overline{\lambda}, & 0 \le u_j^*, & 0 \le v_j^* \end{array} \tag{4.6}$$

for $j = 1, \ldots, d$. Therefore, we have from (4.6) that

$$\Lambda^* = P \ \mathrm{diag}(\lambda_1^*, \ldots, \lambda_d^*) \ P^\top, \quad \lambda_j^* = \begin{cases} \overline{\lambda} & (1/\overline{\lambda} \geq a_j) \\ 1/a_j & (1/\overline{\lambda} \leq a_j \leq 1/\underline{\lambda}) \\ \underline{\lambda} & (1/\underline{\lambda} \leq a_j), \end{cases} \quad j = 1, \ldots, d. \tag{4.7}$$

In order to obtain Λ^*, we may conduct the following procedure. We first get the eigenvalues a_j $(j = 1, \ldots, d)$ and the orthogonal matrix P by diagonalizing A. Next, we calculate Λ^* by (4.7).

4.4 The maximum likelihood estimation for Gaussian mixtures with sparse precision matrices

We also discuss the maximum likelihood estimation for Gaussian mixtures in Subsection 4.3. However, we add the L_1 regularization in order to obtain precision matrices being sparse. In this case, we should replace f_i^Λ in (4.1) with

$$f_i^\Lambda(\Lambda_i) := \begin{cases} \rho_i \|\Lambda_i\|_1 & (\ \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I\ ) \\ +\infty & (\ \text{otherwise}\ ), \end{cases} \quad i = 1, \ldots, m,$$

where $\rho_1, \ldots, \rho_m$ are positive constants.

Note that α_i^{t+1} and μ_i^{t+1} are also given by (4.2) as mentioned in Subsection 4.3. On the other hand, subproblem (3.15) with respect to Λ_i is different, and it is written as

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}\left(A_i^t \Lambda_i\right) - \log\det \Lambda_i + \tau_i^t \|\Lambda_i\|_1, \\ \text{subject to} & \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I, \end{array} \tag{4.8}$$

where A_i^t is given by (4.4). We can obtain the solution Λ_i^{t+1} of problem (4.8) by the existing methods such as [9, 10, 18].

5 Numerical Experiments

In this section, we report two numerical experiments for the models discussed in Subsections 4.1 and 4.3. The program was coded in MATLAB R2010a and run on a machine with an Intel Core i7 920 2.67GHz CPU and 3.00GB RAM.

Experiment 1 for the model discussed in Subsection 4.1

In the Experiment 1, we investigate the validity of the model discussed in Subsection 4.1.

Throughout the Experiment 1, we used the observational data $X = [x_1, \ldots, x_n] \in \mathbf{R}^{1\times n}$ and the test data $\widetilde{X} := [\widetilde{x}_1, \ldots, \widetilde{x}_{10000}] \in \mathbf{R}^{1\times 10000}$ generated by the following Gaussian mixture with $d = 1$ and $m = 5$:

$$p(x) = \frac{1}{5}\mathcal{N}(x|-10, 5) + \frac{1}{5}\mathcal{N}(x|-8, 5) + \frac{1}{5}\mathcal{N}(x|0, 5) + \frac{1}{5}\mathcal{N}(x|8, 5) + \frac{1}{5}\mathcal{N}(x|10, 5). \tag{5.1}$$

For the given observational data X, we estimated parameters of the Gaussian mixture with $d = 1$ and $m = 5$, that is, we solved the following problem by Algorithm 2:

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{k=1}^{n} \log\left(\sum_{i=1}^{5} \alpha_i \mathcal{N}(x_k|\mu_i, \Lambda_i^{-1})\right), \\ \text{subject to} & \displaystyle\sum_{i=1}^{5} \alpha_i = 1,\ \ell_i \leq \alpha_i,\ 0 \leq \Lambda_i,\ i = 1, \ldots, 5. \end{array} \tag{5.2}$$

In the Experiment 1, we estimated parameters by using three models with $\ell_i = 0$ ($i = 1, \ldots, 5$), $\ell_i = 0.1$ ($i = 1, \ldots, 5$) and $\ell_i = 0.15$ ($i = 1, \ldots, 5$) in (5.2).

An initial point $(\alpha^0, \mu^0, \Lambda^0)$ of Algorithm 2 was chosen as follows. We set $\alpha_i^0 = 1$, $\Lambda_i^0 = 1$ ($i = 1, \ldots, 5$). A mean μ^0 was set to the computational result of K-means algorithm (`kmeans`) in MATLAB. Moreover, we stopped Algorithm 2 when

$$\left| D(W^{t+1}, \alpha^t, \mu^t, \Lambda^t) - D(W^t, \alpha^{t-1}, \mu^{t-1}, \Lambda^{t-1}) \right| < 10^{-5},$$

where the function D is defined by (3.3).

Tables 1 and 2 show the results when the number of the observational data is 30 and 100, respectively. In each case, we carried out the maximum likelihood estimation 15 times for 15 different observational data. In two tables, we report the log-likelihoods for the observational data and the test data. Note that we used the same test data $\widetilde{X} \in \mathbf{R}^{1\times 10000}$ in all experiments. Since the amount of the test data is sufficiently large, we may consider that the estimation with bigger log-likelihood for the test data is better than that with small one. For each experiment in Table 1, numbers with boldface type indicate the highest log-likelihood among the various ℓ_i. Furthermore, "$*$" in Tables indicates that Algorithm 2 was stopped by numerical difficulty. The reason for the difficulty is that the mixture coefficient α_i became too small, and hence assumptions (ii) and (iii) in Theorem 3.6 did not hold.

From Table 1, we see that the model with $\ell_i = 0$ is better than the models with $\ell_i = 0.1$ and $\ell_i = 0.15$ from the viewpoint of the log-likelihood **for the observational data**. The results are quite natural because the feasible set with $\ell_i = 0$ is larger than those with $\ell_i = 0.1$ or $\ell_i = 0.15$. On the other hand, from the viewpoint of the log-likelihood **for the test data**, the models with $\ell_i = 0.1$ and $\ell_i = 0.15$ are better than the model with $\ell_i = 0$ for many trials. In particular, the model with $\ell_i = 0.15$ tends to be the best. This is because the true mixture coefficient is 0.2 as in (5.1). Moreover, the estimation of the model with $\ell_i = 0$ is overfitting for the small observational data. This can be seen in Figure 1 and Table 3 that present the details of the numerical result for No. 3 in Table 1. Figure 1 (a) and (b) are probability density functions obtained by the models with $\ell_i = 0$ and $\ell_i = 0.15$, respectively. In the both figures, the black dash line indicates the probability density function of the true mixture distribution (5.1), and the black line indicates the estimated probability density function. Table 3 presents the estimated parameters. From Table 3, we see that α_5 and Λ_5^{-1} of the model with $\ell_i = 0$ are very small. Thus, the probability density function value in Figure 1 (a) becomes very large around $\mu_5 = 4.9377$. This phenomenon sometimes occurred when the amount of the observational data is small. See [1, Section 9.2.1] for its details. On the other hand, such a singular phenomenon did not happen on the model with $\ell_i = 0.15$ (Figure 1 (b)).

From Table 2, we do not see big differences in the log-likelihoods for the test data among the models. The reason for these results is that we were able to estimate parameters correctly regardless of the value of ℓ_i because we had sufficient amount of the observational data.

From these results, even if the amount of the observational data is small, the model with ℓ_i close to the true value is expected to avoid the overfitting and find an appropriate estimation.

Experiment 2 for the model discussed in Subsection 4.3

In the Experiment 2, we use the model discussed in Subsection 4.3, and study its effectivity.

In this experiment, we used the observational data $X = [x_1, \ldots, x_n] \in \mathbf{R}^{d\times n}$ and the test data $\widetilde{X} = [\widetilde{x}_1, \ldots, \widetilde{x}_{10000}] \in \mathbf{R}^{d\times 10000}$. These data are generated by the following Gaussian

mixture:

$$p(x) = \sum_{i=1}^{10} \frac{1}{10} \mathcal{N}(x|\hat{\mu}_i, \hat{\Lambda}_i^{-1}),$$

where the elements of $\hat{\mu}_i$ were selected randomly from the interval $[-1, 1]$, and $\hat{\Lambda}_i^{-1}$ $(i = 1, \ldots, 10)$ are selected as follows. First, we generated a matrix $A_i \in \mathbf{R}^d$ $(i = 1, \ldots, 10)$ whose elements are normally distributed with mean 0 and variance 1. Then we set $\hat{\Lambda}_i^{-1} := (A_i^\top A_i)^{\frac{1}{2}}$ $(i = 1, \ldots, 10)$. For the observational data X, we solved the following model by Algorithm 2 in order to estimate parameters α_i, μ_i and Λ_i^{-1} $(i = 1, \ldots, 10)$:

$$\begin{aligned} &\text{maximize} && \sum_{k=1}^{n} \log \left(\sum_{i=1}^{10} \alpha_i \mathcal{N}(x_k|\mu_i, \Lambda_i^{-1}) \right), \\ &\text{subject to} && \sum_{i=1}^{10} \alpha_i = 1,\ 10^{-3} \leq \alpha_i,\ \underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I,\ i = 1, \ldots, 10. \end{aligned} \tag{5.3}$$

In the experiments, we estimated parameters by using two models with $(\underline{\lambda}_i, \overline{\lambda}_i) = (0, \infty)$ and $(\underline{\lambda}_i, \overline{\lambda}_i) = (10^{-3}, 10^3)$ in (5.3). In the following, the models (A) and (B) indicate the model with $(\underline{\lambda}_i, \overline{\lambda}_i) = (0, \infty)$ and $(\underline{\lambda}_i, \overline{\lambda}_i) = (10^{-3}, 10^3)$, respectively.

An initial point $(\alpha^0, \mu^0, \Lambda^0)$ of Algorithm 2 was chosen as follows. We set $\alpha_i^0 = 1$, $\Lambda_i^0 = I$ $(i = 1, \ldots, 10)$, and set μ^0 as the computational result of K-means algorithm (`kmeans`) in MATLAB. Moreover, we used the same termination criterion as the Experiment 1.

Tables 4 and 5 show the results when the dimension d of the observational data X is 10 and 30, respectively. In each case, we conducted the maximum likelihood estimation 10 times by using observational data. In No. 1 of Tables 4 and 5, we exploited the observational data X such that $n = 100$. In the subsequent estimations, we added 100 observational data into the previous ones, and used those data as the observational data. Note that we exploited the same test data in each dimension d. In Tables 4 and 5, we report the log-likelihoods for both the observational and test data divided by the numbers of data, respectively. As with the Experiment 1, " $*$ " indicates that Algorithm 2 was stopped by numerical difficulty.

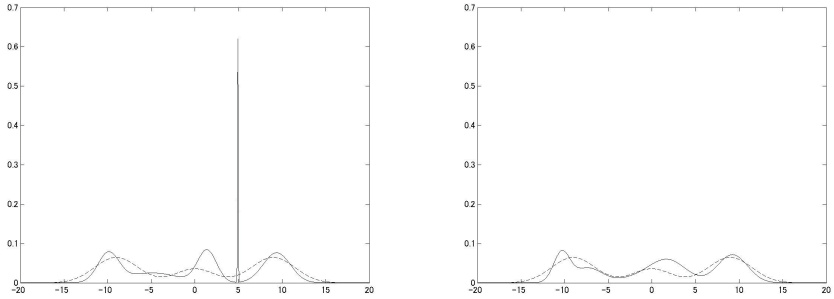
As seen in Table 4 when $d = 10$, we do not see big differences between the both models. On the other hand, as seen in Table 5, the differences appeared between the models (A) and (B). Although the model (A) could not estimate parameters when the amount of the observational data is small, the model (B) could estimate parameters owing to the constraints $\underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I$ $(i = 1, \ldots, m)$.

6 Concluding Remarks

In this paper, we presented a BCD method for the maximum likelihood estimation problem of mixture distributions, where the problem may have regularizations/constraints on the parameters. Moreover, we presented efficient implementations of the BCD method for some special problems. In particular, we gave the $O(m)$ solution method for subproblem (3.10) when the lower constraints $\alpha_i \geq \ell_i$ $(i = 1, \ldots, m)$ exist. In addition, we provided an analytical solution for subproblem (3.15) with the constraint $\underline{\lambda}_i I \preceq \Lambda_i \preceq \overline{\lambda}_i I$. Finally, we conducted the numerical experiments for the models discussed in Subsections 4.1 and 4.3. From the experiments, we see that the models with reasonable constraints yield the valid parameter estimations even if the amount of the observational data is small.

As a future work, we are interested in an inexact version of the proposed BCD method. The proposed method requires that subproblems (3.10) and (3.11) are solved exactly for

Figure 1: Results of No. 3 in Table 1



(a) $\ell_i = 0$ (b) $\ell_i = 0.15$

(a) $\ell_i = 0$ (b) $\ell_i = 0.15$

Table 1: Comparison of log-likelihoods (The amount of data is 30.)

	$\ell_i = 0$		$\ell_i = 0.1$		$\ell_i = 0.15$	
No.	Observation	Test	Observation	Test	Observation	Test
1	-92.6920	-3.8819e+004	-93.9322	-3.8646e+004	-94.0707	**-3.8355e+004**
2	-85.1689	-4.2377e+004	-85.3814	**-4.2056e+004**	-86.7102	-4.2398e+004
3	-89.0016	-3.5282e+004	-94.2949	-3.4429e+004	-94.2979	**-3.4427e+004**
4	-83.7478	-4.8337e+004	-83.9552	-4.8651e+004	-90.0500	**-3.6657e+004**
5	-90.9218	-3.5632e+004	-91.1861	-3.6142e+004	-94.1681	**-3.5011e+004**
6	-87.3364	-3.6083e+004	-89.0853	-3.5634e+004	-93.4515	**-3.3895e+004**
7	-94.0355	-3.4853e+004	-94.1238	-3.4554e+004	-94.3455	**-3.4372e+004**
8	-85.6939	-3.8715e+004	-85.8422	-3.8203e+004	-86.2903	**-3.7063e+004**
9	-93.2788	-3.6004e+004	-97.6769	**-3.3735e+004**	-97.7062	-3.3868e+004
10	-86.7174	-3.8413e+004	-86.9268	-3.8697e+004	-90.2407	**-3.6068e+004**
11	-89.2880	**-3.5906e+004**	-89.4100	-3.6042e+004	-90.0250	-3.6390e+004
12	*	*	-88.0714	-3.8834e+003	-90.1522	**-3.5275e+003**
13	-87.4854	-3.9886e+004	-91.6894	-3.9187e+004	-94.7792	**-3.6992e+004**
14	*	*	-100.3292	-3.3846e+004	-101.2753	**-3.3538e+004**
15	-94.9501	-3.4860e+004	-94.9503	-3.4865e+004	-95.2875	**-3.4847e+004**

Table 2: Comparison of log-likelihoods (The amount of data is 100.)

	$\ell_i = 0$		$\ell_i = 0.1$		$\ell_i = 0.15$	
No.	Observation	Test	Observation	Test	Observation	Test
1	-333.3528	-3.3148e+004	-333.3910	-3.3173e+004	-333.3504	-3.3173e+004
2	-320.6859	-3.3347e+004	-322.3568	-3.2981e+004	-323.1440	-3.3170e+004
3	-321.1681	-3.3312e+004	-321.1681	-3.3312e+004	-321.1942	-3.3346e+004
4	-321.6798	-3.3584e+004	-325.7542	-3.3380e+004	-327.4315	-3.3076e+004
5	-317.9389	-3.3457e+004	-318.2513	-3.3593e+004	-319.8712	-3.3566e+004
6	-321.1656	-3.3005e+004	-321.8685	-3.2909e+004	-321.5855	-3.3038e+004
7	-316.4248	-3.3408e+004	-321.0262	-3.3135e+004	-321.8640	-3.3118e+004
8	-315.8101	-3.4256e+004	-317.1999	-3.3815e+004	-317.8504	-3.4037e+004
9	*	*	-326.2708	-3.3251e+004	-325.2491	-3.3916e+004
10	-316.0259	-3.3110e+004	-316.9840	-3.3029e+004	-316.9493	-3.3088e+004
11	-305.3455	-3.3814e+004	-307.2223	-3.3443e+004	-308.8353	-3.3541e+004
12	-334.4943	-3.3415e+004	-334.9925	-3.3392e+004	-338.3864	-3.3058e+004
13	-307.9100	-3.4562e+004	-307.9106	-3.4554e+004	-308.6720	-3.4483e+004
14	-312.5304	-3.3674e+004	-312.5304	-3.3674e+004	-312.5319	-3.3678e+004
15	-306.6108	-3.4756e+004	-307.6602	-3.4569e+004	-315.3187	-3.3014e+004

Table 3: Results of No. 3 in Table 1

	(a) $\ell_i = 0$			(b) $\ell_i = 0.15$		
i	α_i	μ_i	Λ_i^{-1}	α_i	μ_i	Λ_i^{-1}
1	0.2233	1.3976	1.1864	0.1500	0.0175	7.8016
2	0.1869	-4.8498	8.7129	0.1526	-7.3207	2.6494
3	0.2232	-9.9129	1.4691	0.1764	-10.3462	0.8697
4	0.3003	9.3869	2.4501	0.3074	9.2639	2.9449
5	0.0663	4.9377	0.0018	0.2136	2.0166	3.9377

Table 4: Comparison of log-likelihoods (The dimension is 10.)

		(A) $(\underline{\lambda}_i, \overline{\lambda}_i) = (0, \infty)$		(B) $(\underline{\lambda}_i, \overline{\lambda}_i) = (10^{-3}, 10^3)$	
No.	# of data	Observation	Test	Observation	Test
1	100	*	*	*	*
2	200	*	*	-16.2364	-24.5632
3	300	-17.4153	-22.0561	-17.4153	-22.0561
4	400	-17.7676	-21.1256	-17.7676	-21.1256
5	500	-17.9586	-20.7800	-17.9586	-20.7800
6	600	-18.0046	-20.5797	-18.0046	-20.5797
7	700	-18.1480	-20.2934	-18.1480	-20.2934
8	800	-18.2535	-20.0519	-18.2535	-20.0519
9	900	-18.2848	-19.9797	-18.2848	-19.9797
10	1000	-18.2381	-19.6845	-18.2381	-19.6845

Table 5: Comparison of log-likelihoods (The dimension is 30.)

		(A) $(\underline{\lambda}_i, \overline{\lambda}_i) = (0, \infty)$		(B) $(\underline{\lambda}_i, \overline{\lambda}_i) = (10^{-3}, 10^3)$	
No.	# of data	Observation	Test	Observation	Test
1	100	*	*	*	*
2	200	*	*	*	*
3	300	*	*	-45.2811	-151.2185
4	400	*	*	-55.0462	-85.0240
5	500	-58.2170	-76.7472	-58.2170	-76.7472
6	600	-59.6852	-73.3414	-59.6852	-73.3414
7	700	-60.3356	-71.5572	-60.3356	-71.5572
8	800	-60.6195	-70.5857	-60.6195	-70.5857
9	900	-61.0230	-70.0405	-61.0230	-70.0405
10	1000	-61.5174	-69.2069	-61.5174	-69.2069

global convergence. It is worth constructing the global convergent BCD method that allows inexact solutions of subproblems (3.10) and (3.11).

Acknowledgments

The authors would like to thank two anonymous referees for their valuable comments and constructive suggestions, which have significantly improved the quality of the paper.

References

[1] C.M. Bishop, *Pattern recognition and machine learning*, Springer Science+Business Media, 2006.

[2] M. Blum, R.W. Floyd, V. Pratt, R.L. Rivest and R.E. Tarjan, Time bounds for selection, *J. Comput. Sistem Sci.* **7** (1973) 448–461.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[4] P. Brucker, An $O(n)$ algorithm for quadratic knapsack problems, *Oper. Res. Lett.* **3** (1984) 163–166.

[5] A. P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. Ser. B* **39** (1977) 1–38.

[6] J. Friedman, T. Hastie and R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* **9** (2008) 432–441.

[7] R.J. Hathaway, Another interpretation of the EM algorithm for mixture distributions, *Statist. Probab. Lett.* **4** (1986) 53–56.

[8] R.A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[9] L. Li and K.-C. Toh, An inexact interior point method for L_1-regularized sparse covariance selection, *Math. Program. Comput.* **2** (2010) 291–315.

[10] Z. Lu, Adaptive first-order methods for general sparse inverse covariance selection, *SIAM J. Matrix Anal. A.* **31** (2010) 2000–2016.

[11] R.B. Millar, *Maximum Likelihood Estimation and Inference*, John Wiley & Sons, 2011.

[12] P.M. Pardalos and N. Kovoor, An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds, *Math. Program.* **46** (1990) 321–328.

[13] R.A. Redner and H. F. Walker, Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev.* **26** (1984) 195–239.

[14] L. Ruan, M. Yuan and H. Zou, Regularized parameter estimation in high-dimensional Gaussian mixture models, *Neural Comput.* **23** (2011) 1605–1622.

[15] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, *J. Optim. Theory Appl.* **109** (2001) 475–494.

[16] L. Xu and M. I. Jordan, On convergence properties of the EM algorithm for Gaussian mixtures, *Neural Comput.* **8** (1996) 129–151.

[17] Y. Yamakawa, Studies on optimization methods for nonlinear semidefinite programming problems, Doctoral Thesis, Kyoto University, 2015.

[18] X. Yuan, Alternating direction method for covariance selection models, *J. Sci. Comput.* **51** (2012) 261–273.

Manuscript received 22 January 2015
revised 15 May 2015
accepted for publication 15 May 2015

Yuya Yamakawa
Department of Applied Mathematics and Physics
Graduate School of Informatics, Kyoto University
Kyoto 606-8501, Japan
E-mail address: yamakawa@amp.i.kyoto-u.ac.jp

Nobuo Yamashita
Department of Applied Mathematics and Physics
Graduate School of Informatics, Kyoto University
Kyoto 606-8501, Japan
E-mail address: nobuo@i.kyoto-u.ac.jp