Statistica Sinica **29** (2019), 963-984
doi:https://doi.org/10.5705/ss.202016.0488

AN ASSEMBLY AND DECOMPOSITION APPROACH FOR CONSTRUCTING SEPARABLE MINORIZING FUNCTIONS IN A CLASS OF MM ALGORITHMS

Guo-Liang Tian, Xi-Fen Huang and Jinfeng Xu

Southern University of Science and Technology,
Yunnan Normal University and The University of Hong Kong

Abstract: The *minorization–maximization* (MM) principle provides a powerful tool for optimization in statistical applications. A challenging and subjective issue in developing an MM algorithm is to construct an appropriate minorizing function. For numerical convenience, our (AD) approach to constructing the minorizing function as the sum of separable univariate functions yields general class of MM algorithms. We employ the assembly technique (A-technique) and the decomposition technique (D-technique). The A-technique introduces a bank of complemental assembly functions which are often the building blocks of various MM algorithms. The D-technique decomposes the objective function into three parts and separately minorizes them. We illustrate the utility of the proposed approach in multiple applications. Numerical experiments demonstrate its advantages.

Key words and phrases: Case II interval censored data, complemental assembly, compound zero-inflated, transmission tomography, truncation.

1. Introduction

With the popular EM algorithm as its special case, the *minorization–maximization* (MM) principle (Becker, Yang and Lange (1997); Lange, Hunter and Yang (2000)) is an important and useful tool for optimization problems and has a broad range of applications in statistics because of its conceptual simplicity, ease of implementation and numerical stability.

Let Y_{obs} denote the observed data, $\ell(\boldsymbol{\theta}|Y_{obs})$ the log-likelihood function, the vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^{\top} \in \boldsymbol{\Theta}$, and $\boldsymbol{\Theta}$ the parameter space. The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}|Y_{obs})$, optimizes an objective function. The MM principle provides a general tool for constructing iterative algorithms with monotone convergence (Hunter and Lange (2004)). A minorizing function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is first constructed to satisfy

$$Q\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big) \leqslant \ell(\boldsymbol{\theta}|Y_{obs}obs), \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta} \quad \text{and} \quad Q\big(\boldsymbol{\theta}^{(t)}\big|\boldsymbol{\theta}^{(t)}\big) = \ell\big(\boldsymbol{\theta}^{(t)}\big|Y_{obs}\big), \tag{1.1}$$

where $\boldsymbol{\theta}^{(t)}$ denotes the t-th approximation of $\hat{\boldsymbol{\theta}}$. Note that $Q\big(\,\cdot\,\big|\boldsymbol{\theta}^{(t)}\big)$ function lies under $\ell(\cdot|Y_{obs})$ and is tangent to it at the point $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. The surrogate function $Q\big(\,\cdot\,\big|\boldsymbol{\theta}^{(t)}\big)$ is then maximized to obtain

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big) \tag{1.2}$$

as the $(t+1)$-th approximation of the $\hat{\boldsymbol{\theta}}$.

Since

$$\ell\big(\boldsymbol{\theta}^{(t+1)}\big|Y_{obs}\big) \geqslant Q\big(\boldsymbol{\theta}^{(t+1)}\big|\boldsymbol{\theta}^{(t)}\big) \geqslant Q\big(\boldsymbol{\theta}^{(t)}\big|\boldsymbol{\theta}^{(t)}\big) = \ell\big(\boldsymbol{\theta}^{(t)}\big|Y_{obs}\big), \tag{1.3}$$

the MM iteration (1.2) possesses the *ascent property*, driving the target function $\ell(\boldsymbol{\theta}|Y_{obs})$ uphill. The MM principle can be dated back to (Ortega and Rheinboldt (1970 p.253–255)) and the acronym MM was first given by (Hunter and Lange (2000a)).

Due to its versatility and desirable properties, MM algorithms have been developed for quantile regressions (Hunter and Lange (2000b)), the proportional odds model (Hunter and Lange (2002)), the Bradley–Terry model (Hunter (2004)), variable selection (Hunter and Li (2005); Yen (2011)), discriminant analysis (Lange and Wu (2008)), discrete multivariate distributions (Zhou and Lange (2010b)), the dominant mode (Zhou and Lange (2010a)), constrained estimation (Mkhadri, N'Guessan and Hafidi (2010)), sparse logistic PCA (Lee and Huang (2013)), distance majorization (Chi, Zhou and Lange (2014)), geometric and signomial programming (Lange and Zhou (2014)), the generalized heron problem (Chi and Lange (2014)). Lange, Chi and Zhou (2014) recently gives an excellent overview.

In developing an MM algorithm, it is sometimes quite challenging to construct an appropriate minorizing function. This is often done case by case and involves a subjective use of Jensen's inequality or convexity. In practice, especially in high-dimensional situations, it is numerically appealing to construct a surrogate function as the sum of separable univariate functions since it bypasses the difficulty of multi-dimensional optimization. In this paper, we propose a new *assembly and decomposition* (AD) approach for constructing a surrogate function as the sum of separable univariate functions in a general class of MM algorithms. The AD method employs the A-technique and the D-technique. The A-technique introduces the notions of assemblies and complemental assemblies to guide the minorization process as they are the building blocks of the MM algorithms in many applications. The D-technique decomposes the objective function (or more generally some intermediate minorization function) into three parts and then minorizes them separately.

The rest of the paper is organized as follows. Section 2 briefly reviews various forms of the Jensen's inequality as it plays an important role in the minorization. The AD method is introduced in Section 3. Section 4 uses four representative examples to illustrate the AD machinery in constructing a surrogate function as the sum of separable univariate functions. In Section 5, we investigate the theoretical behaviors of the AD MM algorithms such as the local and global convergences. Numerical experiments are conducted in Section 6 to assess their practical performance. Some concluding remarks are given in Section 7.

2. Jensen's Inequality

Let $\varphi(\cdot)$ be a concave function. If X is a random variable taking values in the domain of $\varphi(\cdot)$, then Jensen's inequality states that

$$\varphi\big(E(X)\big) \geqslant E\big\{\varphi(X)\big\}, \tag{2.1}$$

provided that both $E(X)$ and $E[\varphi(X)]$ exist.

(a) The continuous version of Jensen's inequality is

$$\varphi\bigg(\int_{\mathbb{X}} \tau(x)\cdot g(x)\,\mathrm{d}x\bigg) \geqslant \int_{\mathbb{X}} \varphi\big(\tau(x)\big)\cdot g(x)\,\mathrm{d}x, \tag{2.2}$$

where $\mathbb{X}\subset\mathbb{R}$, $\tau(\cdot)$ is a real function, and $g(\cdot)$ is a density on $\mathbb{X}$.

(b) The discrete version of Jensen's inequality is

$$\varphi\left(\sum_{i=1}^n \alpha_i x_i\right) \geqslant \sum_{i=1}^n \alpha_i\varphi(x_i), \tag{2.3}$$

where $\alpha_i \geqslant 0$ and $\sum_{i=1}^n \alpha_i = 1$. Note that the right-hand side of (2.3), i.e., $\sum_{i=1}^n \alpha_i\varphi(x_i)$, is *completely additively separable.*

(c) Let $\psi(\cdot)$ be a convex function. The supporting hyperplane inequality is

$$\psi(x) \geqslant \psi(x_0) + (x-x_0)\psi'(x_0). \tag{2.4}$$

3. The Assembly–Decomposition (AD) Method

Suppose that F is a function of n variables $x_1,\ldots,x_n$. We say that F is completely additively separable if there exists univariate functions $f_1,\ldots,f_n$ such that $F(x_1,\ldots,x_n)=\sum_{i=1}^n f_i(x_i)$. We aim to construct a surrogate function as the sum of separable univariate functions in a general class of MM algorithms.

3.1. The assembly technique

We first define eight function families. Any function in each family can be written as a linear combination of assemblies (basis functions) which we refer to as complemental assemblies. By introducing the notions of assemblies and complemental assemblies, the A-technique guides the direction in which the minorization process should be worked.

1) The log-generalized-gamma function family $\mathsf{LGG}_k(\theta)$. A function $g_1(\theta) \in \mathsf{LGG}_k(\theta)$, if $g_1(\theta) = c_0 + c_1 \log(\theta) + c_2\big(-\theta^k\big), \theta \in \mathbb{R}_+$, where $c_0 \in \mathbb{R}$, $c_1, c_2 \geqslant 0$ and $k \in \mathbb{N}_+$. Two complemental assemblies: $\big\{\log(\theta), -\theta^k\big\}$. When $k = 1$, it reduces to the log-gamma function family, denoted by $\mathsf{LG}(\theta)$. When $k = 2$, it reduces to the log-Rayleigh function family, denoted by $\mathsf{LR}(\theta)$.

2) The log-beta function family $\mathsf{LB}(\theta)$. A function $g_2(\theta) \in \mathsf{LB}(\theta)$, if $g_2(\theta) = c_0 + c_1 \log(\theta) + c_2 \log(1-\theta), \theta \in [0, 1]$, where $c_0 \in \mathbb{R}$ and $c_1, c_2 \geqslant 0$. Two complemental assemblies: $\{\log(\theta), \log(1-\theta)\}$.

3) The log-extended-beta function family $\mathsf{LEB}(\theta)$. A function $g_3(\theta) \in \mathsf{LEB}(\theta)$, if $g_3(\theta) = c_0 + c_1 \log(\theta) + c_2 \log(1-\theta) + c_3(-\theta), \theta \in [0, 1]$, where $c_0 \in \mathbb{R}$ and $c_1, c_2, c_3 \geqslant 0$. Three complemental assemblies: $\{\log(\theta), \log(1-\theta), -\theta\}$. When $c_3 = 0$, it reduces to the log-beta function family.

4) The log-inverted-beta function family $\mathsf{LIB}(\theta)$. A function $g_4(\theta) \in \mathsf{LIB}(\theta)$, if $g_4(\theta) = c_0 + c_1 \log(\theta) + (c_1 + c_2)\{-\log(\theta+1)\}, \theta \in (0, \theta_0)$, where $\theta_0 \doteq \big\{c_1 + \sqrt{c_1(c_1+c_2)}\big\}/c_2$, $c_0 \in \mathbb{R}$ and $c_1, c_2 > 0$. Two complemental assemblies: $\{\log(\theta), -\log(\theta+1)\}$.

5) The log-extended-gamma function family $\mathsf{LEG}(\theta)$. A function $g_5(\theta) \in \mathsf{LEG}(\theta)$, if $g_5(\theta) = c_0 + c_1 \log(\theta) + c_2(-\theta) + c_3 \log(\theta+1), \theta \in \mathbb{R}_+$, where $c_0 \in \mathbb{R}$, $c_1, c_2 > 0$ and $c_3 \geqslant 0$. Three complemental assemblies: $\{\log(\theta), -\theta, \log(\theta+1)\}$. When $c_3 = 0$, it reduces to the log-gamma function family.

6) The log-inverted-gamma function family $\mathsf{LIG}(\theta)$. A function $g_6(\theta) \in \mathsf{LIG}(\theta)$, if $g_6(\theta) = c_0 + c_1\{-\log(\theta)\} + c_2(-1/\theta), \theta \in (0, \theta_0), \theta_0 \doteq 2c_2/c_1$, where $c_0 \in \mathbb{R}$ and $c_1, c_2 > 0$. Two complemental assemblies: $\{-\log(\theta), -1/\theta\}$.

7) The log-Gumbel-maximum function family $\mathsf{LGM}(\theta)$. A function $g_7(\theta) \in \mathsf{LGM}(\theta)$, if $g_7(\theta) = c_0 + c_1\big(-\mathrm{e}^{-c_2\theta}\big) + c_3(-\theta), \theta \in \mathbb{R}$, where $c_0 \in \mathbb{R}$ and $c_1, c_2, c_3 > 0$. Two complemental assemblies: $\{-\mathrm{e}^{-c_2\theta}, -\theta\}$.

8) The log-Dirichlet function family $\mathsf{LD}_q(\boldsymbol{\theta})$. A function $g_8(\boldsymbol{\theta}) \in \mathsf{LD}_q(\boldsymbol{\theta})$, if $g_8(\boldsymbol{\theta}) = c_0 + \sum_{j=1}^q c_j \log(\theta_j), \boldsymbol{\theta} \in \mathbb{T}_q \hat{=} \{\boldsymbol{\theta}\colon\ \theta_j \geqslant 0, \boldsymbol{\theta}^\top \mathbf{1}_q = 1\}$, where $c_0 \in \mathbb{R}$ and $c_j \geqslant 0$. There are q complemental assemblies: $\{\log(\theta_j)\}_{j=1}^q$, extending the log-beta function family.

Assembling the eight function families, we obtain a bank of assemblies and complemental assemblies:

$$\mathbb{B} = \left\{\pm \log(\theta),\ \log(1-\theta),\ \pm\log(\theta+1);\ -\theta^k,\ -\theta^{-1};\ -\mathrm{e}^{-c_2\theta}\right\}. \tag{3.1}$$

As we aim to separate the surrogate function into the sum of separable univariate functions and these complemental assemblies are the building blocks for the MM algorithms in many applications. We use this bank to guide our separation process. We work towards decomposing the surrogate function into the sum of separable univariate functions each of which comes from this bank of assemblies. In principle, we can include additional function families and their complemental assemblies into this bank to make it more complete and the A-technique is hence more versatile. However, as shown in multiple examples in Section 4, the bank with eight function families already gives enough guidance in designing MM algorithms with separable univariate functions and closed-form updating formulas.

3.2. The decomposition technique

3.2.1. Decomposition of the target function into three parts

Assume that as guided by the A-technique, the target function $\ell(\boldsymbol{\theta}|Y_{obs})$ can be written as

$$\ell(\boldsymbol{\theta}|Y_{obs}) = \ell_0(\boldsymbol{\theta}) + \sum_{i=1}^{n_1} \ell_{1i}\left(\boldsymbol{a}_i^\top \boldsymbol{h}_i(\boldsymbol{\theta})\right) + \sum_{i=1}^{n_2} \ell_{2i}(f_i(\boldsymbol{\theta})), \tag{3.2}$$

where

- $\ell_0(\boldsymbol{\theta}) = \sum_{i=1}^q \ell_{0i}(\theta_i)$ is completely additively separable, each $\ell_{0i}(\cdot)$ is a univariate function,
- $\ell_{1i}(\cdot)$ is a univariate *concave* function, $\boldsymbol{a}_i = \left(a_{i1}, \ldots, a_{ip_i}\right)^\top$, $\{h_{ij}(\boldsymbol{\theta})\}_{j=1}^{p_i}$ may be non-linear,
- $\ell_{2i}(\cdot)$ is a univariate *convex* function, and each $f_i(\cdot)$ is a linear combination of complemental assemblies.

The D-technique, or essentially the equation (3.2), decomposes the target function into three parts. Each part is dealt with differently. For the first part, the separation is already done. For the second part, by (2.3),

$$Q_{1i}\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big) = \sum_{j=1}^{p_i} \frac{a_{ij}h_{ij}\big(\boldsymbol{\theta}^{(t)}\big)}{\boldsymbol{a}_i^\top \boldsymbol{h}_i\big(\boldsymbol{\theta}^{(t)}\big)}\,\ell_{1i}\left(\frac{\boldsymbol{a}_i^\top \boldsymbol{h}_i\big(\boldsymbol{\theta}^{(t)}\big)}{h_{ij}\big(\boldsymbol{\theta}^{(t)}\big)}\cdot h_{ij}(\boldsymbol{\theta})\right) \tag{3.3}$$

minorizes the concave function $\ell_{1i}\big[\boldsymbol{a}_i^\top \boldsymbol{h}_i(\boldsymbol{\theta})\big]$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ if for any $i \in \{1, \ldots, n_1\}$, $a_{ij}h_{ij}\big(\boldsymbol{\theta}^{(t)}\big) \geqslant 0$ for all $j = 1, \ldots, p_i$. For the third part, by (2.4),

$$Q_{2i}\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big) = \ell_{2i}\left(f_i\big(\boldsymbol{\theta}^{(t)}\big)\right) + \left\{f_i(\boldsymbol{\theta}) - f_i\big(\boldsymbol{\theta}^{(t)}\big)\right\}\ell'_{2i}\left(f_i\big(\boldsymbol{\theta}^{(t)}\big)\right), \tag{3.4}$$

minorizes the convex function $\ell_{2i}[f_i(\boldsymbol{\theta})]$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ for any $i \in \{1, \ldots, n_2\}$. After combining the three parts, the surrogate function

$$Q_{12}\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big) = \ell_0(\boldsymbol{\theta}) + \sum_{i=1}^{n_1} Q_{1i}\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big) + \sum_{i=1}^{n_2} Q_{2i}\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big) \tag{3.5}$$

minorizes the target function $\ell(\boldsymbol{\theta}|Y_{obs})$.

For each part, the corresponding surrogate function is the sum of separable univariate functions. The surrogate function (3.5) may not be the sum of separable functions itself since the univariate parameters in three parts may not be the same. But as illustrated in our examples in Section 4, in those situations, it is much easier to further minorize (3.5) and construct a surrogate function as the sum of separable univariate functions. It may also be of interest to ask how general the D-technique is. The A-D method is generally applicable when (3.2) holds for some intermediate minorizing function.

3.2.2. Double minorization

In practice, we often encounter the situation where

$$\ell(\boldsymbol{\theta}|Y_{obs}) = \ell_0(\boldsymbol{\theta}) - \sum_{i=1}^{n_3} b_i \log\left(1 - \boldsymbol{a}_i^\top \boldsymbol{h}_i(\boldsymbol{\theta})\right), \tag{3.6}$$

(e.g., see Examples (4.3) in Section 4) where the $\{\ell_0(\boldsymbol{\theta}), \boldsymbol{a}_i, \boldsymbol{h}_i(\boldsymbol{\theta})\}$ are defined in (3.2), $b_i > 0$ and $0 < \boldsymbol{a}_i^\top \boldsymbol{h}_i(\boldsymbol{\theta}) < 1$. The second part of (3.6) does not take the form of (3.2). In this situation, we provide another technique referred to as double minorization. By applying (2.3) with $n = 2$ to the $\log(\cdot)$ function, we have

$$-\log\left(1 - \boldsymbol{a}_i^\top \boldsymbol{h}_i(\boldsymbol{\theta})\right) \geqslant c_i^{(t)} + \frac{1 - \omega_i^{(t)}}{\omega_i^{(t)}}\log\left(\boldsymbol{a}_i^\top \boldsymbol{h}_i(\boldsymbol{\theta})\right) \mathrel{\hat{=}} Q_{3i}\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big), \tag{3.7}$$

where

$$c_i^{(t)}= - \log \omega_i^{(t)} - \frac{1-\omega_i^{(t)}}{\omega_i^{(t)}} \log\left(1-\omega_i^{(t)}\right) \quad \text{and} \quad \omega_i^{(t)} = 1 - \boldsymbol{a}_i^\top \boldsymbol{h}_i\left(\boldsymbol{\theta}^{(t)}\right). \tag{3.8}$$

Therefore, the surrogate function

$$\begin{aligned} Q_3\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) &= \ell_0(\boldsymbol{\theta}) + \sum_{i=1}^{n_3} b_i Q_{3i}\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) \\ &= c + \ell_0(\boldsymbol{\theta}) + \sum_{i=1}^{n_3} \frac{b_i\left(1-\omega_i^{(t)}\right)}{\omega_i^{(t)}} \log\left(\boldsymbol{a}_i^\top \boldsymbol{h}_i(\boldsymbol{\theta})\right) \end{aligned} \tag{3.9}$$

minorizes the target function $\ell(\boldsymbol{\theta}|Y_{obs})$ in (3.6). Here (3.2) holds for the intermediate minorizing function $Q_3\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right)$.

4. Applications

We use four examples to illustrate our method. To our knowledge, the algorithms based on the AD method are new and not previously reported.

4.1. Poisson model for transmission tomography

Suppose that there are n detectors and Y_i is the transmission measurement of the i-th detector. We consider the model

$$Y_i \overset{\text{ind.}}{\sim} \text{Poisson}\left(r_i + s_i \mathrm{e}^{-[\boldsymbol{A}\boldsymbol{\pi}]_i}\right), i = 1, \ldots, n,$$

where r_i is the mean number of background counts of the i-th detector, s_i is the blank scan counts, $\boldsymbol{A} = \left(a_{ij}\right)$ is the $n \times q$ system matrix with $\boldsymbol{a}_i^\top = \left(a_{i1}, \ldots, a_{iq}\right)$ as its i-th row (i.e., $\boldsymbol{A}^\top = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$), $[\boldsymbol{A}\boldsymbol{\pi}]_i \mathrel{\hat{=}} \boldsymbol{a}_i^\top \boldsymbol{\pi}$ denotes the i-th line integral of the attenuation map $\boldsymbol{\pi} = \left(\pi_1, \ldots, \pi_q\right)^\top$, with π_j the unknown attenuation coefficient in the j-th pixel and q the number of pixels. The EM algorithm does not exhibit a closed-form solution for the M-step (Lange and Carson (1984); Fessler (2000)). (Lange and Fessler (1995)) considered a special case where $r_i = 0$ and hence the log likelihood function is concave. Let $Y_{obs} = \{y_i\}_{i=1}^n$ be the observations and $\{r_i, s_i, a_{ij}\}$ known nonnegative constants. The log likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\pi}|Y_{obs}) \;&=\; c + \sum_{i=1}^n \left\{-s_i \exp\left(-\boldsymbol{a}_i^\top \boldsymbol{\pi}\right)\right\} + \sum_{i=1}^n \left\{y_i \log\left(r_i + s_i \mathrm{e}^{-\boldsymbol{a}_i^\top \boldsymbol{\pi}}\right)\right\} \\ &\mathrel{\hat{=}}\; c + \sum_{i=1}^n \ell_{1i}\left(\boldsymbol{a}_i^\top \boldsymbol{\pi}\right) + \sum_{i=1}^n \ell_{2i}(g_i(\boldsymbol{\pi})). \end{aligned} \tag{4.1}$$

It is easy to check that (3.2) holds. For $\ell_{1i}(\cdot)$ of $\boldsymbol{a}_i^\top \boldsymbol{\pi}$, we construct weight

$\omega_{ij} = a_{ij}/\sum_{j=1}^{q} a_{ij}$ and write $\boldsymbol{a}_i^\top \boldsymbol{\pi} = \sum_{j=1}^{q} \omega_{ij}\{\omega_{ij}^{-1} a_{ij}(\pi_j - \pi_j^{(t)}) + \boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\}$. By (3.3), we obtain the minorizing function

$$Q_{1i}(\boldsymbol{\pi}|\boldsymbol{\pi}^{(t)}) = -\sum_{j=1}^{q} s_i \omega_{ij} \exp\left(-\omega_{ij}^{-1} a_{ij}\left(\pi_j - \pi_j^{(t)}\right) - \boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right). \tag{4.2}$$

For $\ell_{2i}(\cdot)$ of $g_i(\boldsymbol{\pi}) = \boldsymbol{a}_i^\top \boldsymbol{\pi}$, by (3.4), we obtain the minorizing function

$$\begin{aligned} Q_{2i}(\boldsymbol{\pi}|\boldsymbol{\pi}^{(t)}) &= y_i \log\left(r_i + s_i \exp\left(-\boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right)\right) \\ &\quad - \left(\boldsymbol{a}_i^\top \boldsymbol{\pi} - \boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right) \frac{y_i s_i \exp\left(-\boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right)}{r_i + s_i \exp\left(-\boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right)}. \end{aligned} \tag{4.3}$$

Therefore, the overall surrogate function

$$\begin{aligned} Qv(\boldsymbol{\pi}|\boldsymbol{\pi}^{(t)}) &= c + \sum_{i=1}^{n} Q_{1i}(\boldsymbol{\pi}|\boldsymbol{\pi}^{(t)}) + \sum_{i=1}^{n} Q_{2i}(\boldsymbol{\pi}|\boldsymbol{\pi}^{(t)}) \\ &= c_1 + \sum_{j=1}^{q}\sum_{i=1}^{n} Q_{3,ij}(\pi_j|\boldsymbol{\pi}^{(t)}) \end{aligned} \tag{4.4}$$

minorizes $\ell(\boldsymbol{\pi}|Y_{obs})$ in (4.1), where

$$\begin{aligned} & Q_{3,ij}(\pi_j|\boldsymbol{\pi}^{(t)}) \\ &= -s_i \omega_{ij} \exp\left(-\omega_{ij}^{-1} a_{ij}(\pi_j - \pi_j^{(t)}) - \boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right) + \frac{a_{ij} y_i s_i \exp\left(-\boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right)}{r_i + s_i \exp\left(-\boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right)}(-\pi_j) \\ & \in \mathsf{LGM}(\pi_j). \end{aligned}$$

The parameters in (4.4) are separated and updated by

$$\pi_j = \pi_j^{(t)} - \frac{\sum_{i=1}^{n} a_{ij} s_i \left\{\exp\left(-\boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right) - y_i \exp\left(-\boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right)/r_i + s_i \exp\left(-\boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right)\right\}}{-\sum_{i=1}^{n} a_{ij}^2 s_i \omega_{ij}^{-1} \exp\left(\boldsymbol{a}_i^\top \boldsymbol{\pi}^{(t)}\right)}. \tag{4.5}$$

4.2. Multivariate compound zero-inflated generalized Poisson distribution

Let $Z_0 \sim \text{Bernoulli}\,(1-\phi_0)$, $\mathbf{x} = (X_1, \ldots, X_m)^\top$, $X_i \sim \text{ZIGP}(\phi_i, \lambda_i, \pi_i)$ for $i = 1, \ldots, m$, and $(Z_0, X_1, \ldots, X_m)$ be mutually independent. A random vector $\mathbf{y} = (Y_1, \ldots, Y_m)^\top$ follows a multivariate compound *zero-inflated generalized Poisson* distribution if

$$\mathbf{y} \stackrel{\text{d}}{=} Z_0\,\mathbf{x} = \begin{cases} \mathbf{0}, \text{ with probability } \phi_0, \\ \mathbf{x}, \text{ with probability } 1-\phi_0, \end{cases}$$

where $\phi_0 \in [0,1)$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^\top \in [0,1)^m$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top \in \mathbb{R}_+^m$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^\top \in [0,1)^m$. We write $\mathbf{y} \sim \text{CZIGP}_m(\phi_0, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi})$. The joint pmf of $\mathbf{y}$ is

$$\gamma_1^{I(\boldsymbol{y}=\mathbf{0})} \times \{(1-\phi_0)\gamma_2\}^{I(\boldsymbol{y}\neq\mathbf{0})} \tag{4.6}$$

where

$$\gamma_1 = \phi_0 + (1-\phi_0)\prod_{i=1}^m \left\{\phi_i + (1-\phi_i)\mathrm{e}^{-\lambda_i}\right\},$$

$$\gamma_2 = \prod_{i=1}^m \left\{\phi_i + (1-\phi_i)\mathrm{e}^{-\lambda_i}\right\}^{I(y_i=0)} \left\{(1-\phi_i)\frac{\lambda_i(\lambda_i+\pi_i y_i)^{y_i-1}\mathrm{e}^{-\lambda_i-\pi_i y_i}}{y_i!}\right\}^{I(y_i>0)}.$$

Suppose that $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{\text{iid}}{\sim} \text{CZIGP}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\phi_0, \phi_1, \ldots, \phi_m, \lambda_1, \ldots, \lambda_m, \pi_1, \ldots, \pi_m)^\top$ and $\mathbf{y}_j = \big(Y_{1j}, \ldots, Y_{mj}\big)^\top$ for $j = 1, \ldots, n$. Let $\boldsymbol{y}_j = \big(y_{1j}, \ldots, y_{mj}\big)^\top$ and $Y_{obs} = \{\boldsymbol{y}_j\}_{j=1}^n$ be the observations. Let $n_0 = \sum_{j=1}^n I\big(\boldsymbol{y}_j = \mathbf{0}\big)$. The log-likelihood function

$$\begin{aligned}
&\ell(\boldsymbol{\theta}|Y_{obs}) \\
&= c_0 + n_0 \log\left(\phi_0 + (1-\phi_0)\prod_{i=1}^m \left\{\phi_i + (1-\phi_i)\mathrm{e}^{-\lambda_i}\right\}\right) + (n-n_0)\log(1-\phi_0) \\
&\quad + \sum_{j=1}^n I(\boldsymbol{y}_j \neq 0)\sum_{i=1}^m \Big[I(y_{ji}=0)\log\Big(\phi_i + (1-\phi_i)\mathrm{e}^{-\lambda_i}\Big) \\
&\quad + I(y_{ji}\neq 0)\big\{\log(1-\phi_i) + \log(\lambda_i) + \big(y_{ji}-1\big)\log\big(\lambda_i + \pi_i y_{ji}\big) - \lambda_i - \pi_i y_{ji}\big\}\Big],
\end{aligned}$$

which can be decomposed as

$$\ell(\boldsymbol{\theta}|Y_{obs}) = \ell_0(\boldsymbol{\theta}) + \ell_1(\boldsymbol{\theta}) + \sum_{j=1}^n\sum_{i=1}^m I(\boldsymbol{y}_j \neq 0)\big\{I\big(y_{ji}=0\big)\ell_{2i}(\boldsymbol{\theta}) + I\big(y_{ji}\neq 0\big)\ell_{3ji}(\boldsymbol{\theta})\big\},$$

where c_0 is a constant not involving $\boldsymbol{\theta}$,

$$\begin{aligned}
\ell_0(\boldsymbol{\theta}) &= c_0 + (n-n_0)\log(1-\phi_0) \\
&\qquad + \sum_{j=1}^n I(\boldsymbol{y}_j \neq 0)\sum_{i=1}^m \big[I(y_{ji}\neq 0)\big\{\log(1-\phi_i) + \log(\lambda_i) - \lambda_i - \pi_i y_{ji}\big\}\big],
\end{aligned}$$

$$\ell_1(\boldsymbol{\theta}) = n_0 \log\left(\phi_0 + (1-\phi_0)\prod_{i=1}^m \left\{\phi_i + (1-\phi_i)\mathrm{e}^{-\lambda_i}\right\}\right),$$

$$= n_0 \log\left((1,1)\begin{pmatrix}\phi_0\\(1-\phi_0)\prod_{i=1}^m\left\{\phi_i+(1-\phi_i)\mathrm{e}^{-\lambda_i}\right\}\end{pmatrix}\right) = n_0 \log\left(\boldsymbol{a}^\top \boldsymbol{h}_1(\boldsymbol{\theta})\right),$$

$$\ell_{2i}(\boldsymbol{\theta}) = \log\left(\phi_i+(1-\phi_i)\mathrm{e}^{-\lambda_i}\right),$$

$$= \log\left((1,1)\begin{pmatrix}\phi_i\\(1-\phi_i)\mathrm{e}^{-\lambda_i}\end{pmatrix}\right) = n_0 \log\left(\boldsymbol{a}^\top \boldsymbol{h}_{2i}(\boldsymbol{\theta})\right),$$

$$\ell_{3ij}(\boldsymbol{\theta}) = \big(y_{ji}-1\big)\log\big(\lambda_i+\pi_i y_{ji}\big),$$

$$= \big(y_{ji}-1\big)\log\left(\big(1,y_{ji}\big)\begin{pmatrix}\lambda_i\\\pi_i\end{pmatrix}\right) = \big(y_{ji}-1\big)\log\left(\boldsymbol{a}_{ji}^\top \boldsymbol{h}_{3i}(\boldsymbol{\theta})\right).$$

From this it can be seen that the parameters in $\ell_0(\boldsymbol{\theta})$ are separated and $\ell_0(\boldsymbol{\theta})$ is a linear combination of $1+4m$ assemblies: $\log(1-\phi_0)$ and $\{\log(1-\phi_i), \log(\lambda_i), -\lambda_i, -\pi_i\}_{i=1}^m$, where $\{\log(\lambda_i), -\lambda_i\}$ are a pair of complemental assemblies. It is easy to find that $\ell_1(\boldsymbol{\theta}), \ell_{2i}(\boldsymbol{\theta})$ and $\ell_{3ji}(\boldsymbol{\theta})$ are concave functions of linear combinations $\boldsymbol{a}^\top\boldsymbol{h}_1(\boldsymbol{\theta}), \boldsymbol{a}^\top\boldsymbol{h}_{2i}(\boldsymbol{\theta})$ and $\boldsymbol{a}_{ji}^\top\boldsymbol{h}_{3i}(\boldsymbol{\theta})$, respectively. Thus (3.2) holds and we apply the D-technique. By (2.3), we construct the surrogate functions for $\ell_1(\boldsymbol{\theta}), \ell_{2i}(\boldsymbol{\theta})$ and $\ell_{3ji}(\boldsymbol{\theta})$ separately, , which are combined to be $Q^*\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big)$ for $\ell(\boldsymbol{\theta}|Y_{obs})$,

$$\begin{aligned}
&Q^*\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big)\\
&= \ell_0(\boldsymbol{\theta}) + \frac{n_0\phi_0^{(t)}}{\gamma_1^{(t)}}\log(\phi_0)\\
&+n_0\left(1-\frac{\phi_0^{(t)}}{\gamma_1^{(t)}}\right)\left\{\log(1-\phi_0)+\sum_{i=1}^m \log\left(\phi_i+(1-\phi_i)\mathrm{e}^{-\lambda_i}\right)\right\}\\
&+\sum_{j=1}^n\sum_{i=1}^m I\big(\boldsymbol{y}_j\neq 0\big)I\big(y_{ji}=0\big)\left[\frac{\phi_i^{(t)}}{\beta_i^{(t)}}\log(\phi_i)+\left(1-\frac{\phi_i^{(t)}}{\beta_i^{(t)}}\right)\big\{\log(1-\phi_i)-\lambda_i\big\}\right]\\
&+\sum_{j=1}^n\sum_{i=1}^m I\big(y_{ji}>0\big)\left\{\frac{\lambda_i^{(t)}\big(y_{ji}-1\big)}{\lambda_i^{(t)}+\pi_i^{(t)}y_{ji}}\log(\lambda_i)+\frac{\pi_i^{(t)}y_{ji}\big(y_{ji}-1\big)}{\lambda_i^{(t)}+\pi_i^{(t)}y_{ji}}\log(\pi_i)\right\},
\end{aligned}$$

where $\beta_i^{(t)} = \phi_i^{(t)} + \big(1-\phi_i^{(t)}\big)\mathrm{e}^{-\lambda_i^{(t)}}, i=1,\ldots,m$. We find that we did not completely separate all the parameters since $\sum_{i=1}^m \log\left(\phi_i+(1-\phi_i)\mathrm{e}^{-\lambda_i}\right)$ take the same form as $\ell_{2i}(\boldsymbol{\theta})$. The technique based on (2.3) can be applied to these terms and we can then obtain the completely additively separable function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$

to minimize $\ell(\boldsymbol{\theta}|Y_{obs})$ as follows,

$$Q\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big) = Q\big(\phi_0\big|\boldsymbol{\theta}^{(t)}\big) + \sum_{i=1}^{m}\Big\{Q\big(\phi_i\big|\boldsymbol{\theta}^{(t)}\big) + Q\big(\lambda_i\big|\boldsymbol{\theta}^{(t)}\big) + Q\big(\pi_i\big|\boldsymbol{\theta}^{(t)}\big)\Big\}, \tag{4.7}$$

where

$$Q\big(\phi_0\big|\boldsymbol{\theta}^{(t)}\big) = \frac{n_0\phi_0^{(t)}}{\gamma_1^{(t)}}\log(\phi_0) + \left(n - \frac{n_0\phi_0^{(t)}}{\gamma_1^{(t)}}\right)\log(1-\phi_0) \in \mathsf{LB}(\phi_0),$$

$$\begin{aligned}
Q\big(\phi_i\big|\boldsymbol{\theta}^{(t)}\big) = &\left\{\frac{n_0\phi_i^{(t)}\left(\gamma_1^{(t)} - \phi_0^{(t)}\right)}{\gamma_1^{(t)}\beta_i^{(t)}} + \sum_{j=1}^{n}\frac{I(\boldsymbol{y}_j \neq 0)I(y_{ji}=0)\phi_i^{(t)}}{\beta_i^{(t)}}\right\}\log(\phi_i) \\
&+\left[n_0\left(1-\frac{\phi_0^{(t)}}{\gamma_1^{(t)}}\right)\left(1-\frac{\phi_i^{(t)}}{\beta_i^{(t)}}\right)\right. \\
&\left.+\sum_{j=1}^{n} I(\boldsymbol{y}_j \neq 0)\left\{1 - \frac{\phi_i^{(t)}I(y_{ji}=0)}{\beta_i^{(t)}}\right\}\right]\log(1-\phi_i) \in \mathsf{LB}(\phi_i),
\end{aligned}$$

$$\begin{aligned}
Q\big(\lambda_i\big|\boldsymbol{\theta}^{(t)}\big) = &\left\{\sum_{j=1}^{n}\frac{I(\boldsymbol{y}_j \neq 0)I(y_{ji}>0)(\lambda_i+\pi_i)y_{ji}}{\lambda_i+\pi_i y_{ji}}\right\}\log(\lambda_i) \\
&-\left[n_0\left(1-\frac{\phi_0^{(t)}}{\gamma_1^{(t)}}\right)\left(1-\frac{\phi_i^{(t)}}{\beta_i^{(t)}}\right)\right. \\
&\left.+\sum_{j=1}^{n} I(\boldsymbol{y}_j \neq 0)\left\{1 - \frac{\phi_i^{(t)}I\big(y_{ji}=0\big)}{\beta_i^{(t)}}\right\}\right]\lambda_i \in \mathsf{LG}(\lambda_i),
\end{aligned}$$

$$\begin{aligned}
Q\big(\pi_i\big|\boldsymbol{\theta}^{(t)}\big) = &\left\{\sum_{j=1}^{n}\frac{I(\boldsymbol{y}_j \neq 0)I\big(y_{ji}>0\big)\pi_i^{(t)}y_{ji}\big(y_{ji}-1\big)}{\lambda_i^{(t)}+\pi_i^{(t)}y_{ji}}\right\}\log(\pi_i) \\
&-\left\{\sum_{j=1}^{n} I\big(\boldsymbol{y}_j \neq 0\big)I\big(y_{ji}>0\big)y_{ji}\right\}\pi_i \in \mathsf{LG}(\pi_i).
\end{aligned}$$

The parameters are completely separated and the iteration is explicitly

$$\begin{cases} \phi_0^{(t+1)} = \dfrac{n_0\phi_0^{(t)}}{n\gamma_1^{(t)}}, \\[2ex] \phi_i^{(t+1)} = \dfrac{n_0\phi_i^{(t)}\big(\gamma_1^{(t)}-\phi_0^{(t)}\big)/\gamma_1^{(t)}\beta_i^{(t)} + \sum_{j=1}^n I(\boldsymbol{y}_j\neq 0)I(y_{ji}=0)\phi_i^{(t)}/\beta_i^{(t)}}{\big(n-n_0\phi_0^{(t)}/\gamma_1^{(t)}\big)}, \\[3ex] \lambda_i^{(t+1)} = \dfrac{\sum_{j=1}^n I(\boldsymbol{y}_j\neq 0)I(y_{ji}>0)(\lambda_i+\pi_i)y_{ji}/(\lambda_i+\pi_i y_{ji})}{n_0\big(1-\phi_0^{(t)}/\gamma_1^{(t)}\big)\big(1-\phi_i^{(t)}/\beta_i^{(t)}\big)+\sum_{j=1}^n I(\boldsymbol{y}_j\neq 0)\big(1-\phi_i^{(t)}I(y_{ji}=0)/\beta_i^{(t)}\big)}, \\[3ex] \pi_i^{(t+1)} = \dfrac{\sum_{j=1}^n I(\boldsymbol{y}_j\neq 0)I(y_{ji}>0)\pi_i^{(t)}y_{ji}(y_{ji}-1)/\left(\lambda_i^{(t)}+\pi_i^{(t)}y_{ji}\right)}{\sum_{j=1}^n I(\boldsymbol{y}_j\neq 0)I(y_{ji}>0)y_{ji}}, i=1,\ldots,m. \end{cases} \tag{4.8}$$

4.3. Left-truncated normal distribution

A left-truncated normal distribution, $\mathrm{LTN}\big(\mu,\sigma^2;a\big)$, has the density function

$$f\big(y;\mu,\sigma^2,a,\infty\big) = \frac{1}{c\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)\cdot I(y\geqslant a),$$

where $\big(\mu,\sigma^2\big)$ are two unknown parameters, a is a known constant, $c = 1-\Phi((a-\mu)/\sigma)$, and $\Phi(\cdot)$ is the cdf of the standard normal distribution. Suppose that $Y_1,\ldots,Y_n \overset{\text{iid}}{\sim} \mathrm{LTN}\big(\mu,\sigma^2;a\big)$ and $Y_{obs} = \{y_i\}_{i=1}^n$ are the observations. The log-likelihood function

$$\ell\big(\mu,\sigma^2\big|Y_{obs}\big) = -\frac{n\log(2\pi)}{2} - \frac{n\log\left(\sigma^2\right)}{2} - \sum_{i=1}^n\frac{(y_i-\mu)^2}{2\sigma^2} - n\log\left(1-\Phi\left(\frac{a-\mu}{\sigma}\right)\right).$$

The last term is a special case of the second term in (3.6) with $n_3 = 1$. By (3.7) and (3.8),

$$\begin{aligned} &-n\log\left(1-\Phi\left(\frac{a-\mu}{\sigma}\right)\right) \\ &\geqslant -n\log\omega^{(t)} - ns_1^{(t)}\log\left(1-\omega^{(t)}\right) + ns_1^{(t)}\log\left(\Phi\left(\frac{a-\mu}{\sigma}\right)\right), \end{aligned}$$

where

$$\omega^{(t)} = 1-\Phi\left(\frac{a-\mu^{(t)}}{\sigma^{(t)}}\right) \quad \text{and} \quad s_1^{(t)} = \frac{1-\omega^{(t)}}{\omega^{(t)}}. \tag{4.9}$$

Thus the surrogate function

$$Q^*\big(\mu,\sigma^2\big|\mu^{(t)},\sigma^{2(t)}\big) = c_1^{(t)} - \frac{n\log\left(\sigma^2\right)}{2} - \frac{\sum_{i=1}^n(y_i-\mu)^2}{2\sigma^2}$$

$$+ ns_1^{(t)} \log\left(\Phi\left(\frac{a-\mu}{\sigma}\right)\right) \tag{4.10}$$

minorizes $\ell\big(\mu, \sigma^2 \big| Y_{obs}\big)$ at $\big(\mu, \sigma^2\big) = \big(\mu^{(t)}, \sigma^{2(t)}\big)$, where $c_1^{(t)}$ is a constant not depending on $\big(\mu, \sigma^2\big)$. Directly maximizing (4.10) with respect to $\big(\mu, \sigma^2\big)$ cannot yield closed-form solutions for $\big(\mu^{(t+1)}, \sigma^{2(t+1)}\big)$ due to the presence of $\log[\Phi((a-\mu)/\sigma)]$. To overcome this difficulty, let

$$\tau\big(x; \mu, \sigma^2\big) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

and define the weight function

$$g\left(x; \mu^{(t)}, \sigma^{2(t)}, -\infty, a\right) = \frac{\tau\left(x; \mu^{(t)}, \sigma^{2(t)}\right) \cdot I(x < a)}{\Phi\left(\left(a-\mu^{(t)}\right)/\sigma^{(t)}\right)} = \frac{\tau\left(x; \mu^{(t)}, \sigma^{2(t)}\right) \cdot I(x < a)}{1-\omega^{(t)}}.$$

Applying the integral version of Jensen's inequality to $\log[\Phi((a-\mu)/\sigma)]$, we have

$$\begin{aligned}
&\log\left(\Phi\Big(\frac{a-\mu}{\sigma}\Big)\right) \\
&= \log\left(\int_{-\infty}^{a} \frac{\tau\big(x; \mu, \sigma^2\big)}{g\left(x; \mu^{(t)}, \sigma^{2(t)}, -\infty, a\right)} \cdot g\left(x; \mu^{(t)}, \sigma^{2(t)}, -\infty, a\right) \mathrm{d}x\right) \\
&\overset{(2.2)}{\geqslant} \int_{-\infty}^{a} \log\left(\frac{\tau\big(x; \mu, \sigma^2\big)}{g\left(x; \mu^{(t)}, \sigma^{2(t)}, -\infty, a\right)}\right) \cdot g\left(x; \mu^{(t)}, \sigma^{2(t)}, -\infty, a\right) \mathrm{d}x \\
&= c_2^{(t)} + \int_{-\infty}^{a} \log\left(\tau\big(x; \mu, \sigma^2\big)\right) \cdot g\left(x; \mu^{(t)}, \sigma^{2(t)}, -\infty, a\right) \mathrm{d}x \\
&= c_3^{(t)} - \frac{\log\left(\sigma^2\right)}{2} - \frac{\sigma^{2(t)} + \left(\mu^{(t)} - \mu\right)^2}{2\sigma^2} \\
&\quad + \frac{\sigma^{2(t)}\left(a + \mu^{(t)} - 2\mu\right) g\left(a; \mu^{(t)}, \sigma^{2(t)}, -\infty, a\right)}{2\sigma^2},
\end{aligned} \tag{4.11}$$

where $c_3^{(t)}$ is a constant not depending on $\big(\mu, \sigma^2\big)$. By (4.10) and (4.11), the surrogate function

$$\begin{aligned}
Q\big(\mu, \sigma^2 \big| \mu^{(t)}, \sigma^{2(t)}\big) = c_4^{(t)} &- \frac{n\left(1 + s_1^{(t)}\right)}{2} \log\left(\sigma^2\right) - \frac{ns_1^{(t)}\left\{\sigma^{2(t)} + \left(\mu^{(t)} - \mu\right)^2\right\}}{2\sigma^2} \\
&+ \frac{ns_1^{(t)} \sigma^{2(t)}\left(a + \mu^{(t)} - 2\mu\right) g\left(a; \mu^{(t)}, \sigma^{2(t)}, -\infty, a\right)}{2\sigma^2} \\
&- \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2},
\end{aligned} \tag{4.12}$$

minorizes $\ell\big(\mu,\sigma^2\big|Y_{obs}\big)$ at $\big(\mu,\sigma^2\big) = \big(\mu^{(t)},\sigma^{2(t)}\big)$, where $c_4^{(t)}$ is a constant not depending on $\big(\mu,\sigma^2\big)$. The MM iterations are explicitly given by

$$\begin{cases} \mu^{(t+1)} &= \dfrac{\bar{y} + s_1^{(t)}\left\{\mu^{(t)} - \sigma^{2(t)} g\left(a;\mu^{(t)},\sigma^{2(t)},-\infty,a\right)\right\}}{1+s_1^{(t)}}, \\ \sigma^{2(t+1)} &= \dfrac{\sum_{i=1}^n \left(y_i - \mu^{(t+1)}\right)^2 + n s_1^{(t)}\delta^{(t)}}{n(1+s_1^{(t)})}, \end{cases} \tag{4.13}$$

where $\delta^{(t)} \mathrel{\hat{=}} \sigma^{2(t)} + \big(\mu^{(t)} - \mu^{(t+1)}\big)^2 - \sigma^{2(t)}\big(a + \mu^{(t)} - 2\mu^{(t+1)}\big) g\big(a;\mu^{(t)},\sigma^{2(t)},-\infty,a\big)$.

4.4. Case Ⅱ interval-censored data

Consider a failure time study that consists of n independent subjects from a homogeneous population with survival function $S(t)$. Let T_i denote the survival time of interest for subject $i, i = 1,\ldots,n$. Suppose that interval-censored data on the T_i's are observed and given by

$$Y_{obs} = \{(L_i, R_i]; i = 1,\ldots,n\},$$

where $T_i \in (L_i, R_i]$. Let $\big\{s_j\big\}_{j=0}^m$ denote the unique ordered elements of $\{0, L_i, R_i;$ $i = 1,\ldots,n\}$. Take $\alpha_{ij} = I(s_j \in (L_i, R_i])$ and $p_j = S\big(s_{j-1}\big) - S\big(s_j\big), i = 1,\ldots,n,$ $j = 1,\ldots,m$. The log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{p}|Y_{obs}) &= \sum_{i=1}^n \log(S(L_i) - S(R_i)) \\ &= \sum_{i=1}^n \log\left(\sum_{j=1}^m \alpha_{ij}p_j\right) \mathrel{\hat{=}} \sum_{i=1}^n \ell_{1i}\left(\sum_{j=1}^m \alpha_{ij}p_j\right), \end{aligned}$$

where $\boldsymbol{p} = (p_1,\ldots,p_m)'$, $\sum_{j=1}^m p_j = 1$, $p_j \geqslant 0$ $(j = 1,\ldots,m)$ and $\ell_{1i}(\cdot) = \log(\cdot)$ is concave. By (3.3), we obtain the minorizing function

$$Q(\boldsymbol{p}|\hat{\boldsymbol{p}}) = \sum_{j=1}^m \left\{ \sum_{i=1}^n \frac{\alpha_{ij}p_j^{(t)}}{\sum_{j=1}^m \alpha_{ij}p_j^{(t)}} \log\left(\frac{\sum_{j=1}^m \alpha_{ij}p_j^{(t)}}{p_j^{(t)}} p_j\right)\right\} \in \mathsf{LD}_m(\boldsymbol{p}). \tag{4.14}$$

It is easy to see that the parameters in $Q(\boldsymbol{p}|\hat{\boldsymbol{p}})$ are separated and the MM iterations are explicitly given by

$$p_j^{(t+1)} = \frac{\sum_{i=1}^n \alpha_{ij}p_j^{(t)} / \sum_{j=1}^m \alpha_{ij}p_j^{(t)}}{\sum_{j=1}^m \sum_{i=1}^n \alpha_{ij}p_j^{(t)} / \sum_{j=1}^m \alpha_{ij}p_j^{(t)}}. \tag{4.15}$$

5. Convergence Properties

We establish the theoretical properties of the proposed AD algorithms such as local convergence, global convergence, and convergence rate. The proofs are relegated to the supplementary materials.

5.1. Local and global convergence

Let $\ell(\boldsymbol{\theta})$ be the function to maximize and $Q\big[\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big]$ be the minorizing function, where $\boldsymbol{\theta}$ is the parameter vector and $\boldsymbol{\theta}^{(t)}$ is its current estimate. Denote the maximizer of $Q[\cdot|\boldsymbol{\theta}]$ by $M(\boldsymbol{\theta})$. Following Proposition 15.3.1 and Proposition 15.4.3 of (Lange (2010)), we first give general and verifiable conditions for proving the local and global convergence of an AD MM sequence.

Proposition 1. *If the minorizing function $Q\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big)$ is strictly concave, then the proposed MM algorithm based on the AD approach is locally attracted to a local optimum $\boldsymbol{\theta}^{\infty}$ at a linear rate equal to the spectral radius of $I-d^{20}Q(\boldsymbol{\theta}^{\infty}|\boldsymbol{\theta}^{\infty})^{-1}d^2\ell$ $(\boldsymbol{\theta}^{\infty})$.*

The mapping functions $\boldsymbol{\theta}^{(t+1)} = M\big(\boldsymbol{\theta}^{(t)}\big)$ of the examples in Section 4 are differentiable and the surrogate functions in (4.4), (4.7) and (4.14) are strictly concave. The local convergence results follow directly by Proposition 1.

Corollary 1. *With an initial value $\boldsymbol{\theta}^{(0)}$, the sequences $\big\{\boldsymbol{\theta}^{(t)}\big\}$ generated by the MM algorithms that update the estimates by* (4.5), (4.8) *and* (4.15), *respectively, are convergent to a local optimal $\boldsymbol{\theta}^{\infty}$.*

A function $f:\mathbb{R}^q \to \mathbb{R}\bigcup\{-\infty,+\infty\}$ is coercive if and only if $f(x)\to+\infty$ as $||x||_2\to+\infty$, where $||\cdot||_2$ denotes the standard Euclidean norm.

Proposition 2. *If $-\ell(\boldsymbol{\theta})$ is coercive, the subset $\big\{\boldsymbol{\theta}\in\Omega : \ell(\boldsymbol{\theta}) \geqslant \ell\big(\boldsymbol{\theta}^{(t)}\big)\big\}$ of parameter domain Ω is compact and all stationary points of $\ell(\boldsymbol{\theta})$ are isolated. The minorizing function $Q\big(\boldsymbol{\theta}\big|\boldsymbol{\theta}^{(t)}\big)$ constructed by the AD approach is strictly concave and differentiable in both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(t)}$. Then the MM sequence $\boldsymbol{\theta}^{(t+1)} = M\big(\boldsymbol{\theta}^{(t)}\big)$ converges to the stationary point of $\ell(\boldsymbol{\theta})$. If $\ell(\boldsymbol{\theta})$ is strictly concave, then the limiting point of $\big\{\boldsymbol{\theta}^{(t)}\big\}$ is the maximum.*

We thus have the following result for the examples in Section 4.

Corollary 2. *If the differentiability and coerciveness of $-\ell(\boldsymbol{\theta})$ hold, all stationary points of $-\ell(\boldsymbol{\theta})$ are isolated and the subsets $\big\{\boldsymbol{\theta}\in\Omega : \ell(\boldsymbol{\theta}) \geqslant \ell\big(\boldsymbol{\theta}^{(t)}\big)\big\}$ of parameter domain Ω are compact, for* examples 1, 2 *and* 4 *in* Section 4*. Then the sequence* *of iterates in* (4.5)*,* (4.8) *and* (4.15) *converge to the stationary points of* $\ell(\boldsymbol{\theta})$*, respectively. If the strict concavity of* $\ell(\boldsymbol{\theta})$ *hold for the three examples, then the sequence of iterates in* (4.5)*,* (4.8) *and* (4.15) *converge to the maximum points of* $\ell(\boldsymbol{\theta})$*, respectively.*

5.2. Convergence rate

The convergence rate is usually used to characterize the convergence behavior of an iterative algorithm. It is well known that the convergence rate of an MM or EM algorithm is, in general, linear. Consider an MM or EM mapping function $M(\boldsymbol{\theta})$ that maximizes the objective function $\ell(\boldsymbol{\theta})$ via the minorizing function $Q\big(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\big)$. If $\big\{\boldsymbol{\theta}^{(t)}\big\}$ converges to some optimal point $\boldsymbol{\theta}^{\infty}$ of $\ell(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta})$ is continuous, then $\boldsymbol{\theta}^{\infty}$ is a fixed point and $\boldsymbol{\theta}^{\infty} = M(\boldsymbol{\theta}^{\infty})$. By Taylor expansion, $\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{\infty} \approx dM(\boldsymbol{\theta}^{\infty})\big(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{\infty}\big)$, where $dM(\boldsymbol{\theta}^{\infty})$ is the differential of the mapping M at $\boldsymbol{\theta}^{\infty}$ and often referred to as the matrix rate of convergence. The spectral radius of $dM(\boldsymbol{\theta}^{\infty})$ is usually defined as the local convergence rate of the sequence $\boldsymbol{\theta}^{(t+1)} = M\big(\boldsymbol{\theta}^{(t)}\big)$. (Mclachlan and Krishnan (2008)) and (Lange (2010)) showed that

$$dM(\boldsymbol{\theta}^{\infty}) = \mathbf{I} - \big\{d^2 Q(\boldsymbol{\theta}^{\infty}|\boldsymbol{\theta}^{\infty})\big\}^{-1} d^2\ell(\boldsymbol{\theta}^{\infty}). \tag{5.1}$$

The formula (5.1) is data-dependent and by the law of large numbers, can be approximated by

$$E[dM(\boldsymbol{\theta}^{\infty})] = \mathbf{I} - \left[E\left\{\frac{d^2 Q(\boldsymbol{\theta}^{\infty}|\boldsymbol{\theta}^{\infty})}{n}\right\}\right]^{-1} E\left\{\frac{d^2\ell(\boldsymbol{\theta}^{\infty})}{n}\right\}. \tag{5.2}$$

The spectral radius of $E\{dM(\boldsymbol{\theta}^{\infty})\}$ characterizes the local convergence rate of the sequence. By (5.2), it relies on how well the expected curvature of the objective function is approximated by that of the minorizing function. A smaller convergence rate implies a faster convergence.

6. Numerical Experiments

We conducted numerical experiments to assess the practical performance of the proposed AD MM algorithms for the examples in Section 4. The simulation was coded in R and run in a desktop in Intel(R) Core(TM) i7-2600 with CPU 3.40 GHz, and the stopping criterion was set to be

$$\frac{\big|\ell\big(\boldsymbol{\theta}^{(t+1)}\big|Y_{\,\text{textitobs}}\big) - \ell\big(\boldsymbol{\theta}^{(t)}\big|Y_{obs}\big)\big|}{\big|\ell\big(\boldsymbol{\theta}^{(t)}\big|Y_{obs}\big)\big| + 1} < 10^{-6}.$$

The first three examples are parametric: the Poisson model for transmission

Table 1. Simulation results for examples 1-3.

Settings	No. of Par.	Method	K	Time	L	MSE	Rate
(q, n)				PET			
$(20, 300)$	20	MM	406	0.5186	-131.72	0.8658	0.9918
$(20, 600)$	20	MM	352	0.8774	-277.34	0.4068	0.9898
$(40, 300)$	40	MM	1106	3.6457	-137.72	2.5496	0.9977
$(40, 600)$	40	MM	795	3.9729	-296.91	1.0368	0.9965
$(60, 300)$	60	MM	3677	14.7528	-81.62	5.0082	0.9994
$(60, 600)$	60	MM	1622	12.3621	-197.98	2.8783	0.9985
(m, n)				CZIGP			
$(50, 200)$	151	MM	26	0.1378	5.2626×10^5	0.1522	0.8937
		EM	60	0.8826	5.2626×10^5	0.2201	0.9368
$(50, 1000)$	151	MM	23	0.3877	2.6281×10^6	0.0491	0.8816
		EM	59	2.4772	2.6281×10^6	0.0497	0.9133
$(100, 200)$	301	MM	26	0.2281	1.0563×10^6	0.1537	0.8966
		EM	60	1.5357	1.0563×10^6	0.2199	0.9409
$(100, 1000)$	301	MM	23	0.6975	5.2552×10^6	0.0491	0.8826
		EM	59	4.4465	5.2552×10^6	0.0445	0.9153
$(200, 200)$	601	MM	26	0.4098	2.1058×10^6	0.1562	0.8991
		EM	60	2.8083	2.1058×10^6	0.2237	0.9451
$(200, 1000)$	601	MM	23	1.2729	1.0536×10^7	0.0487	0.8837
		EM	59	8.0647	1.0536×10^7	0.0488	0.9178
(μ, σ^2, n)				LTN			
$(4, 4, 200)$	2	MM	23	0.0006	-202.87	0.1774	0.6547
	2	EM	47	0.0023	-202.87	0.1752	0.8510
$(4, 4, 500)$	2	MM	23	0.0009	-508.66	0.0704	0.6604
	2	EM	47	0.0024	-508.66	0.0696	0.8548

Note: MSE $= 1/R\sum_{r=1}^{R}||\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0||^2/q$, where q indicates the number of parameters.

tomography (PET), the left-truncated normal distribution (LTN) and the multivariate compound zero-inflated generalized Poisson distribution (CZIGP). We generated R replications from various parameter settings and compared the proposed MM algorithms with the EM algorithms of (Lange and Carson (1984)), (Fessler (2000)) and (Tian, et al. (2018)). The average values of iteration numbers (K), run times (Time) in seconds, the final objective values (L), the mean squred error (MSE) and the convergence rate (Rate) via (5.2) are summarized in Table 1. The MSE is defined as

$$\frac{1}{R}\sum_{r=1}^{R}\frac{\left|\left|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right|\right|^2}{q},$$

where $\boldsymbol{\theta}_0$ denote the true value and q is the dimension of $\boldsymbol{\theta}_0$.

In the PET example, the true coefficient vector $\boldsymbol{\pi}$ consists of q components.

Table 2. Simulation results for case Ⅱ Interval-censored data.

Sample size	Method	K	Time	L	MAE	MSE
$n = 50$	MM	223	0.0712	−52.20	0.2318	0.0664
	EM	957	0.5487	−52.19	0.2321	0.0664
$n = 100$	MM	321	0.3899	−114.64	0.1983	0.0539
	EM	1514	3.4107	−114.63	0.1990	0.0540
$n = 200$	MM	441	3.8954	−241.86	0.1733	0.0476
	EM	2428	37.5129	−241.83	0.1743	0.0476
$n = 500$	MM	604	27.2460	−633.63	0.1502	0.0423
	EM	4303	380.1016	−633.46	0.1511	0.0423

Note: MAE $= \max_{t\in(0,t_0]} |\hat{S}(t) - S(t)|$, MSE $= \int_0^{t_0} (\hat{S}(t) - S(t))^2 dt$.

The first $q/2$ components are 3 and the other $q/2$ components are -2. We chose $q \in \{20, 40, 60\}$ and $n \in \{300, 600\}$. The numerical results show that the Newton-method based EM algorithms of (Lange and Carson (1984)) and (Fessler (2000)) break down in most of 500 replications since there are too many constraints for the parameters. In the meantime, when they converge, the EM algorithms are much slower since each iteration is computationally more expensive as the number of parameters is large. In contrast, the AD MM algorithm works well in these high-dimensional situations with the number of parameters varying from 20 to 60 for a samples of size of 300 to 600.

For the CZIGP example, the m-dimensional vectors $\boldsymbol{\phi}, \boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ were set to be constantly $0.1, 9$, and 0.7, respectively. We chose $\phi_0 = 0.2$, $m \in \{50, 100, 200\}$ and $n \in \{200, 1000\}$. For comparison, we derived the EM algorithm as well and the details are in the supplementary materials. From Table 1, the number of parameters varies from 151 to 601, in these high-dimensional situations, the AD MM algorithm has a smaller convergence rate and is much faster than the EM algorithm, requiring less than one sixth of the computation time.

For the LTN example, the true values of $\left(\mu, \sigma^2\right)$ were set to be $(4, 4)$. We chose $a = 1$, and $n = 200$ or 500. The numerical results also show that the MM algorithm has a smaller convergence rate than the EM algorithm and converges much faster.

For the nonparametric example, case Ⅱ interval-censored data, the number of parameters is of the same magnitude as the sample size. The true survival function was set to be $S(t) = \exp(-0.5t)$. The sample size varied from 50 to 500 and the number of parameters was up to 2,000. We report the the average values of iteration numbers (K), run time (Time) in seconds, the final objective values (L), the maximum absolute error (MAE) and the mean squred error (MSE) of

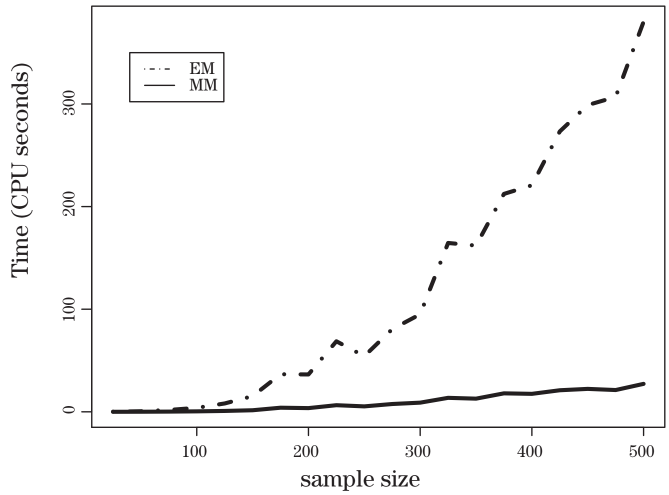

Figure 1. The average run time of Case II Interval-censored Data via EM and MM algorithms based on 500 replications for different sample sizes.

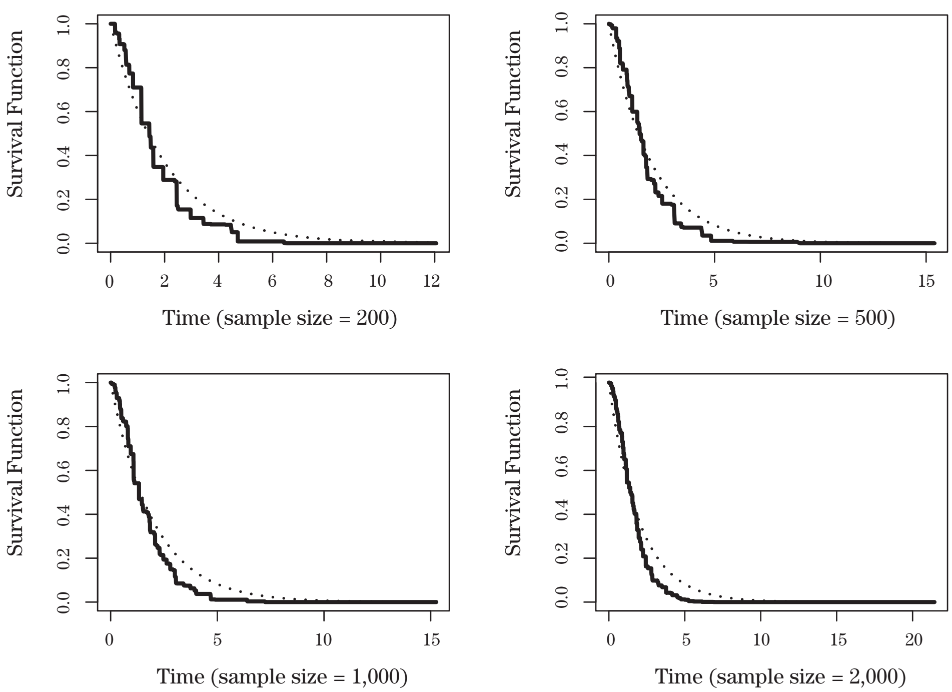

Figure 2. The real lines indicate the true survival function $S(t)$, the dotted lines indicate the estimated survival function $\hat{S}(t)$ via MM algorithm.

MM algorithm and Sun (2006)'s EM algorithm for 500 replications, where the MAE and MSE are defined as

$$\mathrm{MAE} = \max_{t\in(0,t_0]} \left|\hat{S}(t) - S(t)\right|, \mathrm{MSE} = \int_0^{t_0} \left\{\hat{S}(t) - S(t)\right\}^2 dt.$$

The simulation results are summarized in Table 2. The EM and MM algorithms perform similarly well in estimation accuracy. Based on the iteration number and run time, the MM algorithm converges much faster than the EM algorithm. For illustration, we provide a plot in Figure 1 to show the significant difference in the run time between the EM and MM algorithms with the sample size varying from 25 to 500 with step length 25. We also give a plot in Figure 2 to show the difference between the true survival function $S(t)$ and the estimated $\hat{S}(t)$ via the MM algorithm for the different sample sizes.

It is of interest to theoretically compare the convergence rate constants of different minorization schemes and will be pursued in our future work.

Supplementary Materials

The online supplementary material includes the proofs of Propositions 1 and 2, the derivation of the rate matrix for Examples 1-3 and some contents of the old version.

Acknowledgment

The authors are grateful to the Editor, an associate editor, and referees for many helpful comments. This work has been supported by the University of Hong Kong Seed Fund for Basic Research (201611159026), the University of Hong Kong Seed Fund for Translational and Applied Research (201711160015), the University of Hong Kong, Zhejiang Institute of Research and Innovation Seed Fund, the Hong Kong General Research Fund (17308018) and by the National Natural Science Foundation of China (11771199, 71772153 and 11601524).

References

Becker, M.P., Yang, I. and Lange, K. (1997). EM algorithms without missing data. *Statistical Methods in Medical Research* **6**, 38–54.

Chi, E.C. and Lange, K. (2014). A look at the generalized heron problem through the lens of majorization–minimization. *The American Mathematical Monthly* **121**, 95–108.

Chi, E.C., Zhou, H. and Lange, K. (2014). Distance majorization and its applications. *Mathematical Programming, Series A* **146**, 409–436.

De Pierro, A.R. (1995). A modified EM algorithm for penalized likelihood estimation in emission tomography. *IEEE Transactions on Medical Imaging* **14**, 132–137.

Fessler, J.A. (2000). Statistical image reconstruction methods for transmission tomography. *Handbook of Medical Imaging* **2**, 1–70.

Hunter, D.R. (2004). MM algorithms for generalized bradley-terry models. *The Annals of Statistics* **32**, 384–406.

Hunter, D.R. and Lange, K. (2000a). Rejoinder to discussion of "Optimization transfer algorithms using surrogate objective functions." *Journal of Computational and Graphical Statistics* **9**, 52–59.

Hunter, D.R. and Lange, K. (2000b). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics* **9**, 60–77.

Hunter, D.R. and Lange, K. (2002). Computing estimates in the proportional odds model. *Annals of the Institute of Statistical Mathematics* **54**, 155–168.

Hunter, D.R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician* **58**, 30–37.

Hunter, D.R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics* **33**, 1617–1642.

Lange, K. (2010). Numerical Analysis for Statisticians, 2nd Edition. *Statistics and Computing.* Springer, New York.

Lange, K. and Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography* **8**, 306–316.

Lange, K., Chi, E.C. and Zhou, H. (2014). A brief survey of modern optimization for statisticians (with discussions). *International Statistical Review* **82**, 46–89.

Lange, K. and Fessler, J.A. (1995). Globally convergent algorithms for maximum a posteriori transmission tomography. *IEEE Transactions on Image Processing* **4**, 1430–1438.

Lange, K., Hunter, D.R. and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussions). *Journal of Computational and Graphical Statistics* **9**, 1–20.

Lange, K. and Wu, T.T. (2008). An MM algorithm for multicategory vertex discriminant analysis. *Journal of Computational and Graphical Statistics* **17**, 527–544.

Lange, K. and Zhou, H. (2014). MM algorithms for geometric and signomial programming. *Mathematical Programming, Series A* **143**, 339–356.

Lee, S. and Huang, J.Z. (2013). A coordinate descent MM algorithm for fast computation of sparse logistic PCA. *Computational Statistics and Data Analysis* **62**, 26–38.

Mclachlan, G.J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*, 2nd Edition. *Wiley Series in Probability and Statistics, Wiley-Interscience*, John Wiley & Sons. Hoboken, NJ.

Mkhadri, A., N'Guessan, A. and Hafidi, B. (2010). An MM algorithm for constrained estimation in a road safety measure modeling. *Communication in Statistics—Simulation and Computation* **39**, 1057–2010.

Ortega, J.M. and Rheinboldt, W.C. (1970). *Iterative Solutions of Nonlinear Equations in Several Variables.* Academic, New York.

Sun, J.G. (2006). The statistical analysis of interval-censored failure time data. *Statistics for Biology and Health* Springer, Verlag New York.

Tian, G.L., Ju, D., Yuen, K.C. and Zhang, C. (2018). New expectation-maximization-type algorithms via stochastic representation for the analysis of truncated normal data with applications in biomedicine. *Statistical Methods in Medical Research* **27**, 2459–2477.

Yen, T.J. (2011). A majorization–minimization approach to variable selection using spike and slab priors. *The Annals of Statistics* **39**, 1748–1775.

Zhou, H. and Lange, K. (2010a). On the bumpy road to the dominant mode. *Scandinavian Journal of Statistics* **37**, 612–631.

Zhou, H. and Lange, K. (2010b). MM algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics* **19**, 645–665.

Zhou, H. and Zhang, Y.W. (2012). EM vs MM: A case study. *Computational Statistics and Data Analysis* **56**, 3909–3920.

Zhou, H., Alexander, D. and Lange, K. (2011). A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statistics and Computing* **21**, 261–273.

Department of Mathematics, Southern University of Science and Technology, No. 1088, Xueyuan Road, Shenzhen City, Guangdong Province, P. R. China.

E-mail: tiangl@sustc.edu.cn

School of Mathematics, Yunnan Normal University, Kunming, Yunnan, P.R. China.

E-mail: u3003396@connect.hku.hk

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, P.R. China.

E-mail: xujf@hku.hk

(Received November 2016; accepted October 2017)