Computational Statistics and Data Analysis 56 (2012) 2816–2829

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

EM algorithms for multivariate Gaussian mixture models with truncated and censored data

Gyemin Lee [a,*], Clayton Scott [a,b]

[a] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 48109, USA

[b] Department of Statistics, University of Michigan, Ann Arbor, MI, 48109, USA

ARTICLE INFO

Article history:
Received 9 September 2010
Received in revised form 4 March 2012
Accepted 6 March 2012
Available online 12 March 2012

Keywords:
Multivariate Gaussian mixture model
EM algorithm
Truncation
Censoring
Multivariate truncated Gaussian distribution

ABSTRACT

We present expectation–maximization (EM) algorithms for fitting multivariate Gaussian mixture models to data that are truncated, censored or truncated and censored. These two types of incomplete measurements are naturally handled together through their relation to the multivariate truncated Gaussian distribution. We illustrate our algorithms on synthetic and flow cytometry data.

1. Introduction

This paper addresses the problem of fitting Gaussian mixture models on censored and truncated multivariate data. Censoring and truncation arise in numerous applications, for reasons such as fundamental limitations of measuring equipment, or from experimental design. Data are said to be censored when the exact values of measurements are not reported. For example, the needle of a scale that does not provide a reading over 200 kg will show 200 kg for all the objects that weigh more than the limit. Data are said to be truncated when the number of measurements outside a certain range is not reported. For example, an electronic component manufacturer can limit the test duration to 100 h for life time tests. A data collector might provide only the survival times of the components that failed during the test, but not the number of components tested. In these cases, it is often natural to seek the statistical characteristics of the original (uncensored and untruncated) data instead of the observed (censored or truncated) data.

This work is motivated by the analysis of flow cytometry data. Flow cytometry is an essential tool in the diagnosis of diseases such as acute leukemias, chronic lymphoproliferative disorders, and malignant lymphomas (Shapiro, 1994; Brown and Wittwer, 2000). A flow cytometer measures antigen-based markers associated to cells in a cell population. The analysis of flow cytometry data involves dividing cells into subpopulations and inspecting their characteristics. This clustering process, called gating, is performed manually in practice. To automate gating, researchers recently have been investigating mixture models (Boedigheimer and Ferbas, 2008; Chan et al., 2008; Lo et al., 2008; Pyne et al., 2009; Lakoumentas et al., 2009).

However, a flow cytometer measures a limited range of signal strength and records each marker value within a fixed range, such as between 0 and 1023. If a measurement falls outside the range, then the value is replaced by the nearest

* Corresponding author. Tel.: +1 734 763 5228; fax: +1 734 763 8041.
E-mail addresses: gyemin@umich.edu (G. Lee), cscott@eecs.umich.edu (C. Scott).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.03.003

legitimate value; that is, a value smaller than 0 is censored to 0 and a value larger than 1023 is censored to 1023. Moreover, a large portion of cell measurements can be truncated from recording by the judgment of an operator. Therefore, mixture model fitting that does not account for censoring and truncation can result in biases in the parameter estimates and poor gating results. In flow cytometry, a mixture model fitting algorithm should take censoring and truncation into account to avoid biases. Here we present an expectation–maximization (EM) algorithm to fit a multivariate mixture model while accounting for both censoring and truncation.

When censored and truncated data are from an exponential family, Dempster et al. (1977) suggested using the EM procedure to find the maximum likelihood estimate. Atkinson (1992) derived an EM algorithm for a finite mixture of two univariate normal distributions when data are right-censored. Chauveau (1995) also studied a mixture model of univariate censored data, and presented an EM algorithm and its stochastic version. McLachlan and Jones (1988) developed an EM algorithm for univariate binned and truncated data. Cadez et al. (2002) extended the development of McLachlan and Jones (1988) to multivariate case and applied to bivariate blood sample measurements for diagnosis of iron deficiency anemia. To our knowledge, previous work has not addressed censored multivariate data, or continuous (not binned) truncated multivariate data. Furthermore, censoring and truncation have been treated separately. As we will show below, the development of the truncated data EM algorithm and the censored data EM algorithm are closely related to the truncated multivariate Gaussian distribution (Tallis, 1961; Manjunath and Wilhelm, 2009) and we handle these two problems together under the same framework.

Our algorithms make use of recent methods (Drezner and Wesolowsky, 1989; Genz, 2004; Genz and Bretz, 1999, 2002) for evaluating the cumulative distribution function of a multivariate Gaussian. These algorithms run slower as the dimension increases but, when combined with modern computing resources, they can be used successfully in the kinds of lower dimensional settings where mixture methods tend to be applied. Our MATLAB implementation is available online at http://www.eecs.umich.edu/~cscott/code/tcem.zip.

In the following, we briefly review the standard EM algorithm in Section 2. Then we consider truncation (Section 3), censoring (Section 4) and both truncation and censoring (Section 5). We derive EM algorithms for each case and discuss how these algorithms improve the standard EM algorithm. We discuss the initialization and the termination of the EM algorithms in Section 6. Experimental results are reported in Sections 7 and 8 concludes.

2. The standard EM algorithm

In a mixture model, the probability density function of an observation is

$$f(\mathbf{y};\Theta)=\sum_{k=1}^{K}\pi_k f_k(\mathbf{y};\theta_k) \tag{1}$$

where π_k are positive mixing weights summing to one, f_k are component density functions parameterized by θ_k, and $\Theta=(\pi_1,\ldots,\pi_K,\theta_1,\ldots,\theta_K)$ is the collection of all model parameters. Each observation is assumed to be from one of the K components. A common choice of the component density is a multivariate normal with mean $\boldsymbol{\mu}_k$ and covariance Σ_k. Given a set of independent observations $\mathbf{y}^{1:N}:=\{\mathbf{y}^1,\ldots,\mathbf{y}^N\}$ in $\mathcal{Y}\subseteq\mathbb{R}^d$, the objective is to fit such a model to the data.

The EM algorithm proposed by Dempster et al. (1977) is a widely-applied approach for finding the maximum likelihood estimate of a mixture model (Biernacki et al., 2006). In the EM procedure, the unknown true association of each observation to a component is considered missing, and the expected likelihood of the "complete" data is maximized. Let $\mathbf{z}^n\in\{0,1\}^K$ be the membership indicator variable such that $z_k^n=1$ if $\mathbf{y}^n$ is generated from f_k and 0 otherwise. Then the complete log-likelihood function becomes

$$\begin{aligned}\mathcal{L}(\Theta)&=\sum_n\sum_k z_k^n[\ln\pi_k+\ln f_k(\mathbf{y}^n)]\\&=\sum_n\sum_k z_k^n\left[\ln\pi_k-\frac{1}{2}\ln|\Sigma_k|-\frac{1}{2}\mathrm{tr}(\Sigma_k^{-1}(\mathbf{y}^n-\boldsymbol{\mu}_k)(\mathbf{y}^n-\boldsymbol{\mu}_k)^T)\right]+\text{const}\end{aligned} \tag{2}$$

where tr denotes the trace operator of a matrix. The EM algorithm first computes $\mathcal{Q}(\Theta;\Theta^{old})=\mathbb{E}[\mathcal{L}(\Theta)|\mathbf{y}^{1:N};\Theta^{old}]$ (E step) and then finds a new Θ such that $\Theta^{new}=\arg\max_\Theta\mathcal{Q}(\Theta;\Theta^{old})$ (M step). The EM algorithm repeats the E step and M step and updates Θ each iteration. An acclaimed property of the EM algorithm is that each round the value of the log-likelihood monotonically increases (Hastie et al., 2001). The E step simplifies to computing the conditional probabilities

$$\langle z_k^n\rangle:=p(z_k^n=1|\mathbf{y}^n;\Theta^{old})=\frac{\pi_k f_k(\mathbf{y}^n)}{\sum_l\pi_l f_l(\mathbf{y}^n)}.$$

In the M step, we have an update rule in closed form:

$$\widehat{\pi}_k=\frac{1}{N}\sum_n\langle z_k^n\rangle, \tag{3}$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{\sum_n \langle z_k^n \rangle \mathbf{y}^n}{\sum_n \langle z_k^n \rangle}, \tag{4}$$

$$\widehat{\Sigma}_k = \frac{\sum_n \langle z_k^n \rangle (\mathbf{y}^n - \widehat{\boldsymbol{\mu}}_k)(\mathbf{y}^n - \widehat{\boldsymbol{\mu}}_k)^T}{\sum_n \langle z_k^n \rangle}. \tag{5}$$

The EM algorithm alternates between the E step and the M step until convergence.

When truncation and/or censoring occur, however, the true values of $\mathbf{y}_n$ are not always available and the blindfold use of the standard EM algorithm can result in undesirable parameter estimates.

3. The truncated data EM algorithm

Truncation restricts the observation to a subset $\mathcal{Y}_T \subseteq \mathcal{Y}$. Thus, the data points outside $\mathcal{Y}_T$ are not available for estimation of Θ. For example, in clinical flow cytometry, cells with low forward scatter (FS) value are not of much pathological interest and are often dropped during data collection to save data storage space. Hence, all the recorded forward scatter values are always greater than a truncation level chosen by an operator.

Here we assume that the observation window $\mathcal{Y}_T$ is a hyper-rectangle in $\mathbb{R}^d$ with two vertices $\mathbf{s} = (s_1, \ldots, s_d)^T$ and $\mathbf{t} = (t_1, \ldots, t_d)^T$ on the diagonal opposites such that every observed event satisfies $\mathbf{s} \leq \mathbf{y}^n \leq \mathbf{t}$. These inequalities are element-wise, and $s_i = -\infty$ and $t_i = \infty$ mean no truncation below and above, respectively, on the ith coordinate.

The probability density function after truncation is given by $g(\mathbf{y}) = f(\mathbf{y}) / \int_{\mathbf{s}}^{\mathbf{t}} f(\mathbf{y}')d\mathbf{y}'$ for $\mathbf{y} \in [\mathbf{s}, \mathbf{t}]$ and by $g(\mathbf{y}) = 0$ otherwise. Then it can be easily seen that $g(\mathbf{y})$ is also a mixture

$$g(\mathbf{y}) = \sum_{k=1}^{K} \eta_k g_k(\mathbf{y}) \mathbf{1}_{[\mathbf{s},\mathbf{t}]}(\mathbf{y}) \tag{6}$$

with mixing weights η_k and component density functions g_k:

$$\eta_k = \pi_k \frac{\int_{\mathbf{s}}^{\mathbf{t}} f_k(\mathbf{y})d\mathbf{y}}{\int_{\mathbf{s}}^{\mathbf{t}} f(\mathbf{y})d\mathbf{y}} \quad \text{and} \quad g_k(\mathbf{y}) = \frac{f_k(\mathbf{y})}{\int_{\mathbf{s}}^{\mathbf{t}} f_k(\mathbf{y}')d\mathbf{y}'} \mathbf{1}_{[\mathbf{s},\mathbf{t}]}(\mathbf{y}). \tag{7}$$

The indicator function $\mathbf{1}_A(\mathbf{y})$ equals one if $\mathbf{y} \in A$ and zero otherwise. Hence, the component density functions g_k are truncated versions of the original component density functions f_k.

Proceeding similarly as in Section 2, we can express the complete data log-likelihood as

$$\begin{aligned} \mathcal{L}_T(\Theta) &= \sum_n \sum_k z_k^n [\ln \eta_k + \ln g_k(\mathbf{y}^n)] \\ &= \sum_n \sum_k z_k^n \left[\ln \eta_k + \ln f_k(\mathbf{y}^n) - \ln \int_{\mathbf{s}}^{\mathbf{t}} f_k(\mathbf{y})d\mathbf{y} \right]. \end{aligned} \tag{8}$$

Recall that $\mathbf{z}^n$ are the component membership indicator variables. It is conceivable to define the complete data differently by treating the unknown number of truncated measurements as a random variable and including it into the complete data (Dempster et al., 1977). However, this approach requires to make an additional assumption on the distribution of the new parameter. It might generalize our approach, but it can be sensitive to the choice of the distribution of the number of truncated sample points (Gelman, 2004).

The E step applied to (8) requires us to compute

$$\begin{aligned} \mathcal{Q}_T(\Theta; \Theta^{old}) &= \mathbb{E}[\mathcal{L}_T(\Theta) | \mathbf{y}^{1:N}; \Theta^{old}] \\ &= \sum_n \sum_k \langle z_k^n \rangle \left[\ln \eta_k + \ln f_k(\mathbf{y}^n) - \ln \int_{\mathbf{s}}^{\mathbf{t}} f_k(\mathbf{y})d\mathbf{y} \right]. \end{aligned}$$

The main difference from (2) is the terms of normalizing factors, $\ln \int_{\mathbf{s}}^{\mathbf{t}} f_k(\mathbf{y})d\mathbf{y}$, which do not complicate the E step of the EM algorithm, and whose calculation is discussed below. Thus, the E step is simply computing the posterior probability that $\mathbf{y}^n$ belongs to component k

$$\langle z_k^n \rangle := p(z_k^n = 1 | \mathbf{y}^n) = \frac{\eta_k g_k(\mathbf{y}^n)}{\sum_l \eta_l g_l(\mathbf{y}^n)} = \frac{\pi_k f_k(\mathbf{y}^n)}{\sum_l \pi_l f_l(\mathbf{y}^n)}. \tag{9}$$

As the last equality indicates, this posterior remains unchanged as if $\mathbf{y}^n$ in the truncated data is from the entire sample space $\mathcal{Y}$. Then the M step computes $\widehat{\Theta}$ that maximizes $\mathcal{Q}_T(\Theta; \Theta^{old})$, which is found by taking the derivatives of $\mathcal{Q}_T(\Theta; \Theta^{old})$ with

respect to each η_k, $\boldsymbol{\mu}_k$ and Σ_k, and setting to zero. Since η_k should satisfy $\sum_k \eta_k = 1$, a Lagrange multiplier is used to find the maximizer. Using (A.6) and (A.7) in Appendix A.2 to calculate the derivatives of the normalizing factors, we have the following M step equations:

$$\widehat{\eta}_k = \frac{1}{N}\sum_n \langle z_k^n \rangle, \tag{10}$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{\sum_n \langle z_k^n \rangle \mathbf{y}^n}{\sum_n \langle z_k^n \rangle} - \mathbf{m}_k, \tag{11}$$

$$\widehat{\Sigma}_k = \frac{\sum_n \langle z_k^n \rangle (\mathbf{y}^n - \widehat{\boldsymbol{\mu}}_k)(\mathbf{y}^n - \widehat{\boldsymbol{\mu}}_k)^T}{\sum_n \langle z_k^n \rangle} + H_k \tag{12}$$

where $\widehat{\eta}_k$, $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\Sigma}_k$ denote the new parameters and

$$\mathbf{m}_k = \mathcal{M}^1(\mathbf{0}, \Sigma_k; [\mathbf{s} - \boldsymbol{\mu}_k, \mathbf{t} - \boldsymbol{\mu}_k]), \tag{13}$$

$$H_k = \Sigma_k - \mathcal{M}^2(\mathbf{0}, \Sigma_k; [\mathbf{s} - \boldsymbol{\mu}_k, \mathbf{t} - \boldsymbol{\mu}_k]). \tag{14}$$

The notations $\mathcal{M}^1(\boldsymbol{\mu}, \Sigma; [\mathbf{s}', \mathbf{t}'])$ and $\mathcal{M}^2(\boldsymbol{\mu}, \Sigma; [\mathbf{s}', \mathbf{t}'])$ in (13) and (14) indicate the first and second moments of a Gaussian with mean $\boldsymbol{\mu}$ and covariance Σ when it is truncated to a hyper-rectangle with vertices $\mathbf{s}'$ and $\mathbf{t}'$. We discuss the computational aspects of these moments in Appendix A. Comparing (10)–(12) with the standard EM equations (3)–(5) shows that the updates for truncated data are similar to those for untruncated data except the correction terms $\mathbf{m}_k$ and H_k.

The original component weight π_k can be recovered from (7). The normal integrals can be evaluated with the help of computational tools for evaluating the multivariate normal cumulative distribution function. Our implementation relies on `mvncdf` function in the MATLAB 7.9.0 statistics toolbox, which uses algorithms developed by Drezner and Wesolowsky (1989) and by Genz (2004) for bivariate and trivariate Gaussian. The toolbox uses a quasi-Monte Carlo integration algorithm developed by Genz and Bretz (1999, 2002) for four or more dimensional Gaussians.

4. The censored data EM algorithm

As discussed above, truncation excludes data points from the dataset, and the number of data points falling outside the measuring range remains unknown. On the contrary, censoring retains such data points while their exact locations remain unknown.

In the following, we investigate censoring on a hyper-rectangle, in which each data point $\mathbf{y}^n$ is censored above at $\mathbf{b} = (b_1, \ldots, b_d)^T$ and below at $\mathbf{a} = (a_1, \ldots, a_d)^T$.[1] Let $\mathcal{Y}_0, \mathcal{Y}_1, \ldots, \mathcal{Y}_C$ be a partition of the overall sample space $\mathcal{Y}$. If $\mathbf{y}^n \in \mathcal{Y}_0$, we observe the exact values of $\mathbf{y}^n$. When $\mathbf{y}^n \in \mathcal{Y}_c$, $c > 0$, however, censoring occurs and the true values are modified so that

$$x_i^n = y_i^n \mathbf{1}_{[a_i, b_i]}(y_i^n) + a_i \mathbf{1}_{(-\infty, a_i)}(y_i^n) + b_i \mathbf{1}_{(b_i, \infty)}(y_i^n), \quad \forall i, \forall n$$

are observed. Therefore, instead of $\mathbf{y}^{1:N}$, we obtain a set of observations $\mathbf{x}^{1:N}$, which satisfy $a_i \le x_i^n \le b_i$ for $i = 1, \ldots, d$ and $n = 1, \ldots, N$. Note that $a_i = -\infty$ means no censoring below and $b_i = \infty$ means no censoring above.

Chauveau (1995) also studied the analysis of censored data. The difference is that in his setup $\mathbf{x}^n = c$ if $\mathbf{y}^n \in \mathcal{Y}_c$, $c > 0$, whereas in ours some coordinates can preserve their exact values. Furthermore, while his primary concern remained on univariate data, our focus extends to multivariate data.

As described above, unless $\mathbf{y}^n \in \mathcal{Y}_0 = \prod_{i=1}^d [a_i, b_i]$, one or more coordinates are censored and its original location is lost. However, we can infer which partition $\mathbf{y}^n$ belongs to from $\mathbf{x}^n$, the censored observation of $\mathbf{y}^n$, by noting when $x_i^n = a_i$ or b_i. Since each data vector may have different censoring patterns, let the censored and uncensored coordinates be indexed by m_n and o_n, respectively, so that y_i^n, $i \in m_n$, are censored values and y_i^n, $i \in o_n$, are observed values. Then $\mathbf{y}^n$ can be divided into the form $\mathbf{y}^n = \begin{bmatrix} \mathbf{y}_{o_n}^n \\ \mathbf{y}_{m_n}^n \end{bmatrix}$ where $\mathbf{y}_{m_n}^n = (y_i^n, i \in m_n)^T$ and $\mathbf{y}_{o_n}^n = (y_i^n, i \in o_n)^T$ denote the censored and uncensored components of $\mathbf{y}^n$. Note that this does not imply that the vector is arranged to have this pattern and should be understood as a notational convenience. Then the likelihood of $\mathbf{x}^n$ is

$$f(\mathbf{x}^n) = f(\mathbf{y}^n) \quad \text{for } \mathbf{y}^n \in \mathcal{Y}_0 \tag{15}$$

$$\begin{aligned} f(\mathbf{x}^n) &= \int_{\mathcal{X}_{c_n}} f(\mathbf{y}_{m_n}, \mathbf{y}_{o_n}^n) d\mathbf{y}_{m_n} \\ &= f(\mathbf{x}_{o_n}^n) \int_{\mathcal{X}_{c_n}} f(\mathbf{y}_{m_n} | \mathbf{x}_{o_n}^n) d\mathbf{y}_{m_n} \quad \text{for } \mathbf{y}^n \in \mathcal{Y}_{c_n}, c_n > 0 \end{aligned} \tag{16}$$

[1] For univariate data ($d = 1$), left and right censoring are the usual terms.

where the integration is only over the censored coordinates, and $\mathcal{X}_{c_n}$ denote the corresponding integration range. For example, if $x_1^n = a_1$ and $x_2^n = b_2$ while other elements are strictly between **a** and **b**, then $\mathcal{Y}_{c_n} = (-\infty, a_1) \times (b_2, \infty) \times \prod_{i=3}^{d}[a_i, b_i]$, $\mathcal{X}_{c_n} = (-\infty, a_1) \times (b_2, \infty)$ and (16) becomes

$$f(\mathbf{x}^n) = f(\mathbf{x}^n_{o_n}) \int_{-\infty}^{a_1} \int_{b_2}^{\infty} f(y_1, y_2|\mathbf{x}^n_{o_n}) dy_2 dy_1.$$

To invoke the EM machinery, we first compute the expected complete log-likelihood

$$\begin{aligned}\mathcal{Q}_C(\Theta; \Theta^{old}) &= \mathbb{E}[\mathcal{L}(\Theta)|\mathbf{x}^{1:N}; \Theta^{old}] \\ &= \mathbb{E}\left[\sum_n \sum_k z_k^n \left[\ln \pi_k - \frac{1}{2}\ln|\Sigma_k| \right.\right. \\ &\quad \left.\left. - \frac{1}{2}\mathrm{tr}\left(\Sigma_k^{-1}\left(\begin{bmatrix}\mathbf{y}^n_{o_n}\\ \mathbf{y}^n_{m_n}\end{bmatrix} - \boldsymbol{\mu}_k\right)\left(\begin{bmatrix}\mathbf{y}^n_{o_n}\\ \mathbf{y}^n_{m_n}\end{bmatrix} - \boldsymbol{\mu}_k\right)^T\right)\right] \middle| \mathbf{x}^{1:N}; \Theta^{old}\right].\end{aligned}$$

Hence, we need to find posterior probabilities, $p(z_k^n = 1|\mathbf{x}^n)$, and conditional expectations, $\mathbb{E}[z_k^n \mathbf{y}^n_{m_n}|\mathbf{x}^n] = p(z_k^n = 1|\mathbf{x}^n)\,\mathbb{E}[\mathbf{y}^n_{m_n}|\mathbf{x}^n, z_k^n = 1]$ and $\mathbb{E}[z_k^n \mathbf{y}^n_{m_n}\mathbf{y}^{n\,T}_{m_n}|\mathbf{x}^n] = p(z_k^n = 1|\mathbf{x}^n)\,\mathbb{E}[\mathbf{y}^n_{m_n}\mathbf{y}^{n\,T}_{m_n}|\mathbf{x}^n, z_k^n = 1]$.

The posterior probability is

$$\langle z_k^n\rangle := p(z_k^n = 1|\mathbf{x}^n) = \frac{\pi_k f_k(\mathbf{x}^n)}{\sum_l \pi_l f_l(\mathbf{x}^n)}, \tag{17}$$

and it can be computed by (15) or (16). When one or more coordinates are censored, $f_k(\mathbf{x}^n) = f_k(\mathbf{y}^n_{o_n}) \int_{\mathcal{X}_{c_n}} f_k(\mathbf{y}_{m_n}|\mathbf{y}^n_{o_n}) d\mathbf{y}_{m_n}$; thus; it is a product of the probability density function and the cumulative distribution function of Gaussians of lower dimensions, and can be evaluated as explained in Appendix A.1.

The conditional expectations are taken with respect to $f_k(\mathbf{y}_m|\mathbf{x})$. Because $f_k(\mathbf{y}_m|\mathbf{y}_o)$ is a normal density function[2] and satisfies

$$f_k(\mathbf{y}_m|\mathbf{x}) = f_k(\mathbf{y}_m|\mathbf{y}_o, \mathbf{y} \in \mathcal{Y}_c) = \frac{f_k(\mathbf{y}_m|\mathbf{y}_o)}{\int_{\mathcal{X}_c} f_k(\mathbf{y}_m|\mathbf{y}_o) d\mathbf{y}_m} \mathbf{1}_{\mathcal{X}_c}(\mathbf{y}_m), \tag{18}$$

the conditional density $f_k(\mathbf{y}_m|\mathbf{x})$ is a truncated normal density function over $\mathcal{X}_c$. Then we can calculate the following sufficient statistics of $\mathcal{Q}_C$:

$$\begin{aligned}\langle \mathbf{y}^n_{m_n}|k\rangle &:= \mathbb{E}[\mathbf{y}^n_{m_n}|\mathbf{x}^n, z_k^n = 1] = \mathbb{E}[\mathbf{y}^n_{m_n}|\mathbf{y}^n_{o_n}, \mathbf{y}^n \in \mathcal{Y}_{c_n}, z_k^n = 1] \\ &= \mathcal{M}^1(\boldsymbol{\mu}^k_{m_n|o_n}, \Sigma^k_{m_n|o_n}; \mathcal{X}_{c_n}),\end{aligned} \tag{19}$$

$$\begin{aligned}\langle \mathbf{y}^n_{m_n}\mathbf{y}^{n\,T}_{m_n}|k\rangle &:= \mathbb{E}[\mathbf{y}^n_{m_n}\mathbf{y}^{n\,T}_{m_n}|\mathbf{x}^n, z_k^n = 1] = \mathbb{E}[\mathbf{y}^n_{m_n}\mathbf{y}^{n\,T}_{m_n}|\mathbf{y}^n_{o_n}, \mathbf{y}^n \in \mathcal{Y}_{c_n}, z_k^n = 1] \\ &= \mathcal{M}^2(\boldsymbol{\mu}^k_{m_n|o_n}, \Sigma^k_{m_n|o_n}; \mathcal{X}_{c_n})\end{aligned} \tag{20}$$

where $\boldsymbol{\mu}^k_{m_n|o_n}$ and $\Sigma^k_{m_n|o_n}$ are the mean and covariance of $f_k(\mathbf{y}^n_{m_n}|\mathbf{y}^n_{o_n})$. Recall that $\mathcal{M}^1$ and $\mathcal{M}^2$ denote the first and second moments of a truncated normal distribution (see Appendix A.1).

Next, we maximize $\mathcal{Q}_C$ with respect to Θ. Again using a Lagrange multiplier, maximization with respect to π_k gives

$$\widehat{\pi}_k = \frac{1}{N}\sum_n \langle z_k^n\rangle. \tag{21}$$

Similarly, maximization with respect to $\boldsymbol{\mu}_k$ and Σ_k leads to

$$\widehat{\boldsymbol{\mu}}_k = \frac{\sum_n \langle z_k^n\rangle \begin{bmatrix}\mathbf{y}^n_{o_n}\\ \langle \mathbf{y}^n_{m_n}|k\rangle\end{bmatrix}}{\sum_n \langle z_k^n\rangle}, \tag{22}$$

$$\widehat{\Sigma}_k = \frac{\sum_n \langle z_k^n\rangle S_k^n}{\sum_n \langle z_k^n\rangle} \tag{23}$$

[2] If $\mathbf{y} = (\mathbf{y}_m^T, \mathbf{y}_o^T)^T$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance Σ, then the conditional distribution of its partition, $\mathbf{y}_m|\mathbf{y}_o$, is also normally distributed with mean $\boldsymbol{\mu}_{m|o} = \boldsymbol{\mu}_m + \Sigma_{m,o}\Sigma_{o,o}^{-1}(\mathbf{y}_o - \boldsymbol{\mu}_o)$ and covariance $\Sigma_{m|o} = \Sigma_{m,m} - \Sigma_{m,o}\Sigma_{o,o}^{-1}\Sigma_{o,m}$.

where

$$S_k^n = \left(\begin{bmatrix} \mathbf{y}_{o_n}^n \\ \langle \mathbf{y}_{m_n}^n | k \rangle \end{bmatrix} - \widehat{\boldsymbol{\mu}}_k \right) \left(\begin{bmatrix} \mathbf{y}_{o_n}^n \\ \langle \mathbf{y}_{m_n}^n | k \rangle \end{bmatrix} - \widehat{\boldsymbol{\mu}}_k \right)^T + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & R_k^n \end{bmatrix}, \tag{24}$$

$$R_k^n = \langle \mathbf{y}_{m_n}^n \mathbf{y}_{m_n}^{n\,T} | k \rangle - \langle \mathbf{y}_{m_n}^n | k \rangle \langle \mathbf{y}_{m_n}^n | k \rangle^T. \tag{25}$$

Notice that these Eqs. (21)–(23) resemble the update Eqs. (3)–(5) of the standard EM algorithm. In the censored data EM algorithm, the censored elements of $\mathbf{y}^n$ are replaced by the conditional means $\langle \mathbf{y}_{m_n}^n | k \rangle$ and the sample covariance correction R_k^n. When none of the data points are censored, these update equations are equivalent to the standard EM algorithm.

5. The truncated and censored data EM algorithm

In this section, we consider truncation and censoring together and present an EM procedure that encompasses the algorithms above.

Truncation reduces the sample space from $\mathcal{Y}$ to $\mathcal{Y}_T$. By restricting the partition regions $\mathcal{Y}_c$ and the integration ranges $\mathcal{X}_c$ to this reduced sample space $\mathcal{Y}_T$, we can see that the likelihood of an observation $\mathbf{x}$ is $g(\mathbf{x}) = f(\mathbf{x}) / \int_{\mathbf{s}}^{\mathbf{t}} f(\mathbf{y}) d\mathbf{y}$ where the numerator $f(\mathbf{x})$ is defined by (15) and (16). Then this truncated distribution $g(\mathbf{x})$ is a mixture $g(\mathbf{x}) = \sum_{k=1}^K \eta_k g_k(\mathbf{x})$ with mixing weights η_k and component density functions g_k, where

$$\eta_k = \pi_k \frac{\int_{\mathbf{s}}^{\mathbf{t}} f_k(\mathbf{y}) d\mathbf{y}}{\int_{\mathbf{s}}^{\mathbf{t}} f(\mathbf{y}) d\mathbf{y}} \quad \text{and} \quad g_k(\mathbf{x}) = \frac{f_k(\mathbf{x})}{\int_{\mathbf{s}}^{\mathbf{t}} f_k(\mathbf{y}) d\mathbf{y}}. \tag{26}$$

The E step of the EM algorithm begins by finding the expectation

$$\begin{aligned} \mathcal{Q}_{TC}(\Theta; \Theta^{old}) &= \mathbb{E}[\mathcal{L}_T(\Theta) | \mathbf{x}^{1:N}; \Theta^{old}] \\ &= \mathbb{E}\left[\sum_n \sum_k z_k^n \left[\ln \eta_k + \ln f_k(\mathbf{y}^n) - \ln \int_{\mathbf{s}}^{\mathbf{t}} f_k(\mathbf{y}) d\mathbf{y} \right] \middle| \mathbf{x}^{1:N}; \Theta^{old} \right] \end{aligned}$$

conditional on the observed data. This involves computing the posterior probabilities

$$\langle z_k^n \rangle := p(z_k^n = 1 | \mathbf{x}^n) = \frac{\eta_k g_k(\mathbf{x}^n)}{\sum_l \eta_l g_l(\mathbf{x}^n)} = \frac{\pi_k f_k(\mathbf{x}^n)}{\sum_l \pi_l f_l(\mathbf{x}^n)}$$

and the conditional expectations $\langle \mathbf{y}_{m_n}^n | k \rangle := \mathbb{E}[\mathbf{y}_{m_n}^n | \mathbf{x}^n, z_k^n = 1]$ and $\langle \mathbf{y}_{m_n}^n \mathbf{y}_{m_n}^{n\,T} | k \rangle := \mathbb{E}[\mathbf{y}_{m_n}^n \mathbf{y}_{m_n}^{n\,T} | \mathbf{x}^n, z_k^n = 1]$ with respect to $g_k(\mathbf{y}_m | \mathbf{x})$. Since $g_k(\mathbf{y}_m | \mathbf{x})$ satisfies

$$g_k(\mathbf{y}_m | \mathbf{x}) = g_k(\mathbf{y}_m | \mathbf{y}_o, \mathbf{y} \in \mathcal{Y}_c) = \frac{g_k(\mathbf{y}_m | \mathbf{y}_o)}{\int_{\mathcal{X}_c} g_k(\mathbf{y}_m | \mathbf{y}_o) d\mathbf{y}_m} = \frac{f_k(\mathbf{y}_m | \mathbf{y}_o)}{\int_{\mathcal{X}_c} f_k(\mathbf{y}_m | \mathbf{y}_o) d\mathbf{y}_m}$$

from (26) and this equals (18), we can deduce that the sufficient statistics $\langle z_k^n \rangle$, $\langle \mathbf{y}_{m_n}^n | k \rangle$ and $\langle \mathbf{y}_{m_n}^n \mathbf{y}_{m_n}^{n\,T} | k \rangle$ retain the same forms of (17), (19) and (20) in Section 4.

In the M step, we find the new parameters $\widehat{\eta}_k$, $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\Sigma}_k$ that maximize $\mathcal{Q}_{TC}$. To take account of the constraint $\sum_k \eta_k = 1$, a Lagrange multiplier is used in the maximization with respect to η_k. Combining with the quantities computed in the E step, we obtain the following update equations

$$\widehat{\eta}_k = \frac{1}{N} \sum_n \langle z_k^n \rangle, \tag{27}$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{\sum_n \langle z_k^n \rangle \begin{bmatrix} \mathbf{y}_{o_n}^n \\ \langle \mathbf{y}_{m_n}^n | k \rangle \end{bmatrix}}{\sum_n \langle z_k^n \rangle} - \mathbf{m}_k, \tag{28}$$

$$\widehat{\Sigma}_k = \frac{\sum_n \langle z_k^n \rangle S_k^n}{\sum_n \langle z_k^n \rangle} + H_k \tag{29}$$

where the correction terms $\mathbf{m}_k$ and H_k are given in (13) and (14), and the matrix S_k^n is given in (24). Recall that the original component weight π_k can be obtained from η_k through (26) as in Section 3. The remarks on mean and covariance updates in the truncated data EM algorithm and the censored data EM algorithm naturally lead to an observation that (28) and (29) have the combined forms of (11) and (22), and, respectively, (12) and (23).

Table 1
The true parameters and the estimated parameters are compared for the univariate data experiments. We repeated the experiment ten times, and presented the averages and the standard errors of parameter estimates.

	True	Standard EM on uncensored	Standard EM	Truncated and censored EM
1-dim (a)				
μ	3	15.34 ± 0.14	16.80 ± 0.09	2.87 ± 1.31
σ^2	400	105.42 ± 2.18	133.12 ± 1.62	410.65 ± 27.59
1-dim (b)				
μ	-8	12.37 ± 0.13	12.90 ± 0.16	-5.83 ± 1.77
σ^2	400	86.14 ± 2.19	98.65 ± 3.06	352.66 ± 28.84
1-dim (c)				
μ_1	-3	2.08 ± 0.11	2.06 ± 0.10	-1.01 ± 0.36
μ_2	15	13.64 ± 0.11	14.46 ± 0.10	14.89 ± 0.10
σ_1^2	20	2.31 ± 0.23	2.25 ± 0.21	11.68 ± 1.33
σ_2^2	20	14.19 ± 0.69	17.28 ± 0.70	21.79 ± 1.10
π_1	0.6	0.27 ± 0.03	0.24 ± 0.03	0.47 ± 0.09

6. Initialization and termination of EM algorithms

The initialization is an important issue because the result from an EM algorithm is often sensitive to the initial parameter setting. We suggest to proceed as follows to initialize the parameters for the presented EM algorithms. First, perform the k-means clustering algorithm multiple times with different starting points. Next, compute the mixture model parameter and the corresponding complete data log-likelihood from each k-means clustering result. Finally, choose the parameter that achieves the largest log-likelihood as an initial parameter estimate for the truncated and censored data EM algorithm.

To check the convergence of the EM algorithms, we compute the expectation of the complete data log-likelihood function after each iteration. We terminate the EM algorithms when the relative change of the expected complete data log-likelihood falls below 10^{-10}, or when the number of iterations reaches a fixed number, say 500.

In the following section, we use the described initialization and termination methods for experiments.

7. Experiments and results

We present experimental results to demonstrate the algorithms described above. In the following, we describe experiments on univariate and bivariate synthetic data, and on multi-dimensional flow cytometry data.

7.1. Synthetic data

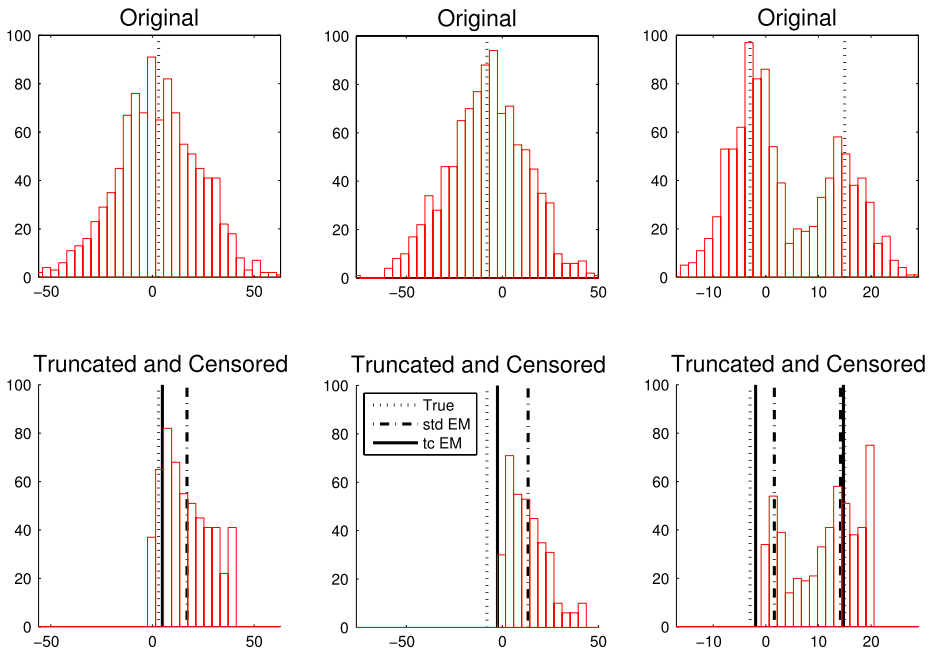
In each experiment, we generated datasets from a known distribution and performed censoring and truncation. On the censored and truncated data, we trained mixture models using the standard EM algorithm and the truncated and censored version of the EM algorithm. We also ran the standard EM algorithm on the set of data points in $\mathcal{Y}_0$, that is, the observations that were not censored.

We first investigated three cases of one dimensional data. In cases (a) and (b), 1000 data points were drawn from a Gaussian (a single component mixture) with means 3 and -8, respectively, and with the same standard deviation 20. Values smaller than 0 were discarded (truncation) and those greater than 40 were set to 40 (censoring). Among 1000 data points, the uncensored data points in $\mathcal{Y}_0$ were about 50% in case (a) and 30% in case (b).

Fig. 1 (a) and (b) show the histograms of each case before and after censoring and truncation. In the figure, the true mean and the estimates are also drawn. As can be seen, the standard EM algorithm always tries to find mean estimates between 0 and 40. Thus, when the true mean is outside this range, the discrepancy between the estimates and the ground-truth can be arbitrarily large. On the other hand, the proposed algorithm finds better estimates. Table 1 compares the true parameter values and the estimated parameter values, and numerically supports this observation.

This result is also validated in case (c) in which data points were drawn from a two-component mixture model as illustrated in Fig. 1 (c). One component with weight 0.6 is centered at -3 and the other component with weight 0.4 is centered at 15 with a common variance 20. The dataset was truncated below at 0 and censored above at 20. Most of the data points from the component on the left were truncated, and nearly 50% of data points were uncensored in each realization. While both algorithms accurately estimated the positive component, the deviations of mean estimates to the true means are evident for the negative component. As shown in Table 1, the variance estimates of the proposed method are also much more close to the ground-truth.

Next we compared the algorithms on multivariate datasets. Two experiments were designed with three-component bivariate Gaussian mixtures. In both cases, an observation (x_1, x_2) was limited to a rectangular window $[0, 25] \times [0, 25]$. In case (a), all three component centers were located within the window ($\pi = (0.5, 0.2, 0.3)$, $\mu_1 = (3, 3)$, $\mu_2 = (13, 3)$, $\mu_3 = (20, 20)$, $\Sigma_1 = \text{diag}(20, 5)$, $\Sigma_2 = \text{diag}(5, 20)$, $\Sigma_3 = [20, 10; 10, 20]$). On the contrary, in case (b), two centroids were located outside the window ($\pi = (0.5, 0.2, 0.3)$, $\mu_1 = (-3, 3)$, $\mu_2 = (10, -1)$, $\mu_3 = (20, 20)$, $\Sigma_1 = \text{diag}(20, 5)$, $\Sigma_2 = \text{diag}(5, 20)$, $\Sigma_3 = [20, 10; 10, 20]$). After 1000 data points were drawn in each case, data points with $x_1 < 0$ were truncated,



(a) A Gaussian at 3. (b) A Gaussian at −8. (c) Two Gaussians.

Fig. 1. Experiments on 1-dimensional synthetic data. The original data histograms (top) are significantly different from the observed data histograms (bottom) when truncation and censoring occur. All data are truncated at 0 and right-censored at 40 (a),(b) or 20 (c). Dotted lines indicate the true means of each Gaussian component. Solid lines and dash–dot lines are mean estimates from the truncated and censored data EM algorithm and the standard EM algorithm, respectively.

and all other values outside the observation window were censored. More than 100 data points were truncated and about 700 data points remained uncensored in case (a), and nearly 400 data points were truncated and about 400 data points remained uncensored in case (b). The data points after truncation and censoring are depicted in Fig. 2. In the figure, level contours are displayed to compare the estimated distributions to the true distribution. The figure also shows the results when the standard EM algorithm is applied on a subset after the censored data points are excluded from the dataset (standard EM on uncensored). The differences between algorithms are most conspicuous in case (b) where the estimates from the truncated and censored data EM algorithm significantly outperform the estimates from the standard EM algorithms.

To quantitatively evaluate model estimates, we computed Kullback–Leibler (KL) divergences

$$KL(p \parallel q) = \mathbb{E}_p[\log p - \log q] \approx \frac{1}{N_e}\sum_{n=1}^{N_e}[\log p(\mathbf{x}^n) - \log q(\mathbf{x}^n)]$$

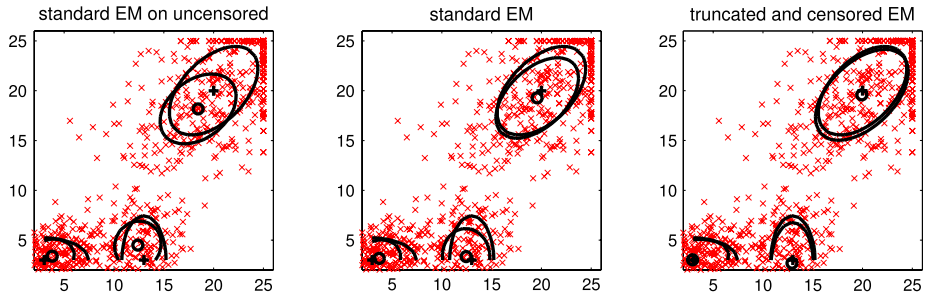
between the known true distribution p and estimated distribution q, where the expectation is approximated by a sample mean over $N_e = 10{,}000$ data points drawn from p. The KL divergence is non-negative and equals to zero if and only if the estimated distribution is the same as the true distribution. We repeated experiments on the ten different samples and averaged the resulting KL divergences. The computed KL divergences are reported in Table 2. For all the investigated datasets, the estimated distributions from the proposed method show significantly smaller KL divergences. Therefore, the truncated and censored data EM algorithm successfully corrects the biases that exist in the standard EM algorithm.

7.2. Application to flow cytometry data

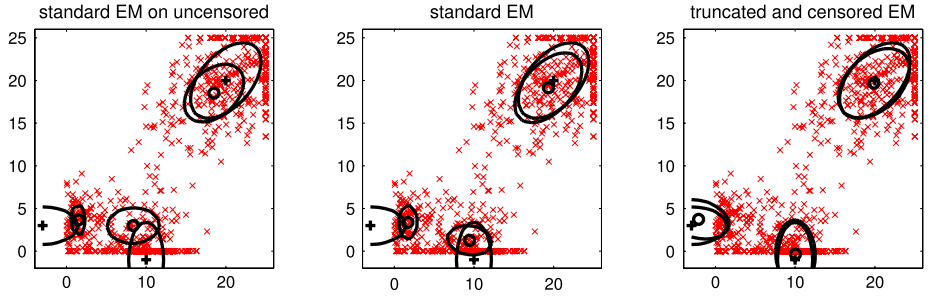
We now discuss a real world application. As explained earlier, this work is motivated by flow cytometry analysis. A flow cytometer measures multiple antigen-based markers associated with cells in a cell population.

In practice, clinicians usually rely on rudimentary tools to analyze flow cytometry data. They select a subset of one, two or three markers and diagnose by visually inspecting the one dimensional histograms or two or three dimensional scatter plots. To facilitate the analysis, the clinicians often select and exclude cell subpopulations. This process is called "gating" and usually performed manually by thresholding and drawing boundaries on the scatter plots. It is labor-intensive and time-consuming, and limits productivity. Moreover, the results of gating vary by user experience, and replicating the same results by others is difficult.

These difficulties have recently motivated interest in automatic and systematic gating methods. Although standard techniques have not been established yet, mathematical modeling of cell populations with mixture models is favored by many researchers due to the possibility of direct analysis in multi-dimensional spaces. In flow cytometry, many different



(a) Three Gaussian components with centroids at (3, 3), (13, 3) and (20, 20).



(b) Three Gaussian components with centroids at (−3, 3), (10, −1) and (20, 20).

Fig. 2. Experiments on 2-dimensional synthetic data. The solid ellipses and 'o's are level-curves and centroids of each component estimate. The dashed ellipses and '+'s are for true mixture components. Small crosses represent data points in the truncated and censored data. x_2 is censored at 0 and 25, and x_1 is truncated at 0 and censored at 25.

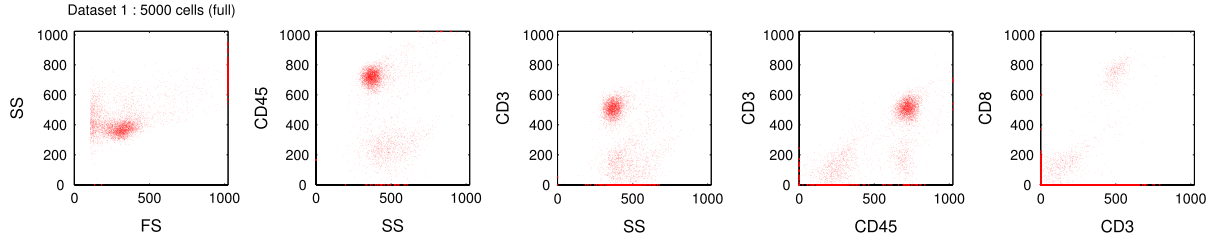
Table 2
The KL divergences between the true densities and the estimated densities are computed for each synthetic dataset. The averages and standard errors across ten samples are reported in the table. The proposed algorithm outperforms the standard EM algorithm.

	Standard EM on uncensored	Standard EM	Truncated and censored EM
1-dim (a)	1.46 ± 0.03	1.15 ± 0.02	0.02 ± 0.01
1-dim (b)	3.48 ± 0.09	3.07 ± 0.09	0.07 ± 0.03
1-dim (c)	3.82 ± 0.07	3.56 ± 0.07	0.35 ± 0.17
2-dim (a)	0.37 ± 0.02	0.50 ± 0.02	0.03 ± 0.01
2-dim (b)	4.05 ± 0.36	4.21 ± 0.36	0.59 ± 0.20

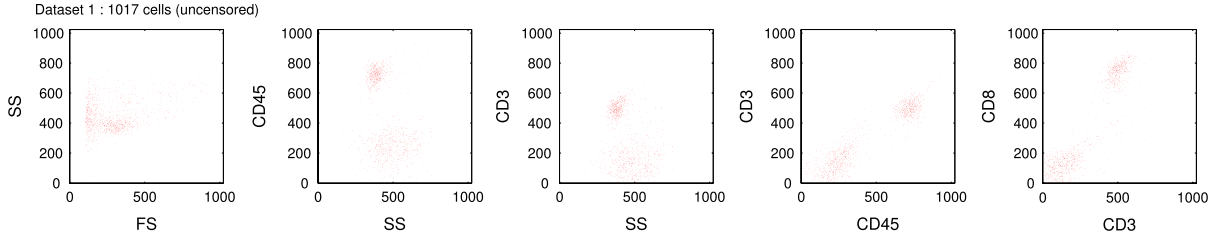
kinds of mixture models have been proposed, some involving more complicated component distributions than others. In Boedigheimer and Ferbas (2008) and Chan et al. (2008), Gaussian mixtures are used to model cell populations. The use of a mixture of t-distributions combined with a Box–Cox transformation is studied in Lo et al. (2008). A more recent study reported successful applications of a mixture of skew normal distributions or skew t-distributions in Pyne et al. (2009). The domain knowledge of field experts is sometimes incorporated in the mixture model (Lakoumentas et al., 2009). However, while truncation and censoring are present in flow cytometry data, we note that these issues have not been explicitly addressed in the literature. We demonstrate fitting Gaussian mixture models to flow cytometry data with the algorithms proposed in this paper. We believe our approach can be extended to other mixture families like those mentioned above, but such extensions are beyond the scope of this work.

Here we present the analysis of two flow cytometry datasets. These datasets were provided by the Department of Pathology at the University of Michigan. Each cell contains five marker readings. The markers in the first dataset are FS, SS, CD3, CD8 and CD45. These are intended for finding T-cells, a type of white blood cells, in the blood sample. The second dataset includes FS, SS, CD20, CD5 and CD45, and these markers are for identifying B-lymphocytes. The forward scatter (FS) threshold is set at approximately 100, and cells with low FS values are truncated from these datasets. Each dataset also underwent censoring so that it includes no marker values out of the range from 0 to 1023. The censoring was severe in these datasets, and only 20% and 40% of total observed cells are uncensored as can be seen in the scatter plots in Figs. 3 and 4. Furthermore, for each dataset, the distribution of all cells is significantly different from the distribution of the uncensored cells. When we exclude censored cells, the cluster of $CD45^+CD3^-$ cells are lost in the first dataset (Fig. 3 fourth column) and the cluster of $CD45^+CD20^+$ cells are lost in the second dataset (Fig. 4 fourth column). Thus, analysis based exclusively on the uncensored cell population can be misleading.

We modeled the cell population with a Gaussian mixture and fitted the model to the observed 5000 cells using the standard EM algorithm and the truncated and censored version of EM algorithm. We chose a six-component model ($K = 6$)

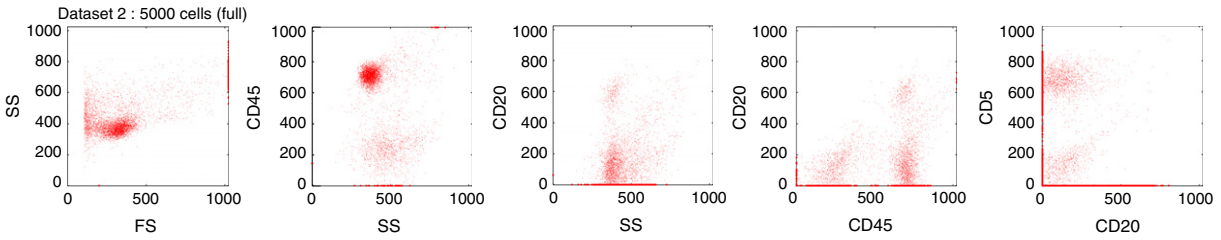


(a) Scatter plots of all observed cells.

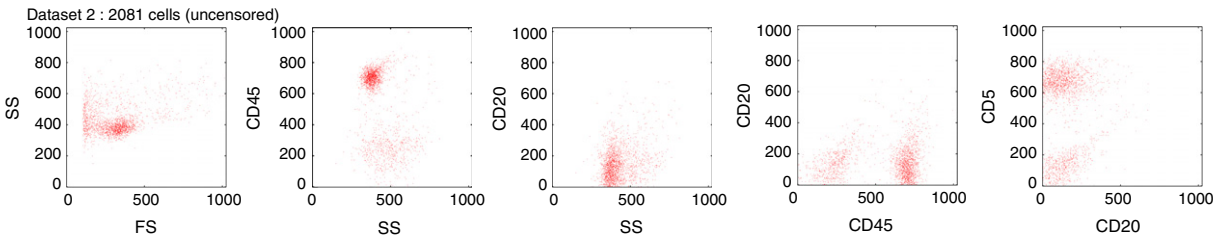


(b) Scatter plots of uncensored cells.

Fig. 3. The first flow cytometry dataset has markers FS, SS, CD3, CD8 and CD45. These markers are chosen to investigate the *T*-cells in the blood sample of a patient. Only 20% of cells are uncensored. The $CD45^{+}CD3^{-}$ subpopulation is missing in (b).



(a) Scatter plots of all observed cells.



(b) Scatter plots of uncensored cells.

Fig. 4. The second dataset includes FS, SS, CD20, CD5 and CD45 for *B*-cell population. The uncensored cells are 40% of the total observed cells. Scatter plots show that the $CD45^{+}$ $CD20^{+}$ cells are missing among the uncensored cells.

since, from these datasets, we expect to find several types of cells such as lymphocytes, which mostly consist of *T*-cells and *B*-cells, lymphoblasts, and small populations of granulocytes and monocytes. We treated each cell as a point in 5-dimensional space. The *k*-means algorithm is first performed to initialize each EM algorithm. The convergence is determined when the relative change of the log-likelihood is less than 10^{-10} or the number of iterations reaches 500. Fig. 5 shows the evolution of the log-likelihood of the truncated and censored data EM on the flow cytometry datasets. The value increases sharply in the first dozens of steps and then converges. The average times per iteration were 0.01 seconds for the standard EM and 2.50 seconds for the truncated and censored data EM under Windows 7 system equipped with two Intel(R) Xeon(R) 2.27 GHz processors and RAM 12 GB. We repeated this process with 10 different starting points based on different runs of the *k*-means algorithm, and presented the results that achieved the highest log-likelihood.

The mixture models fitted by the standard EM and the truncated and censored data EM are shown in Figs. 6 and 7. In the first dataset, both algorithms generated similar estimates of lymphocyte populations (components 1, 2, and 3), which are the primary interest in the flow cytometry data analysis. On the other hand, the results for lymphoblasts are different (component 4 in Fig. 6(a) and components 4, 5 in Fig. 6(b)). Because a large number of lymphoblasts were truncated or

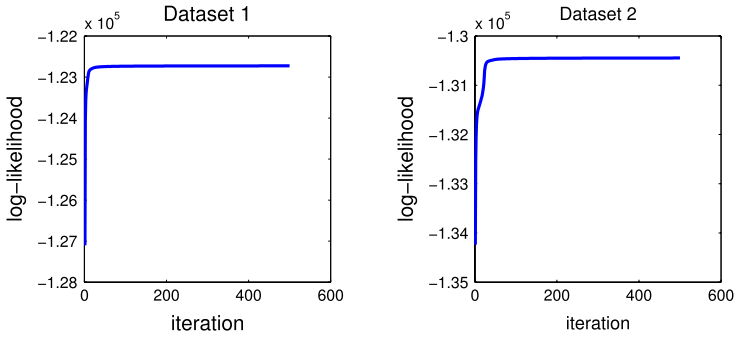
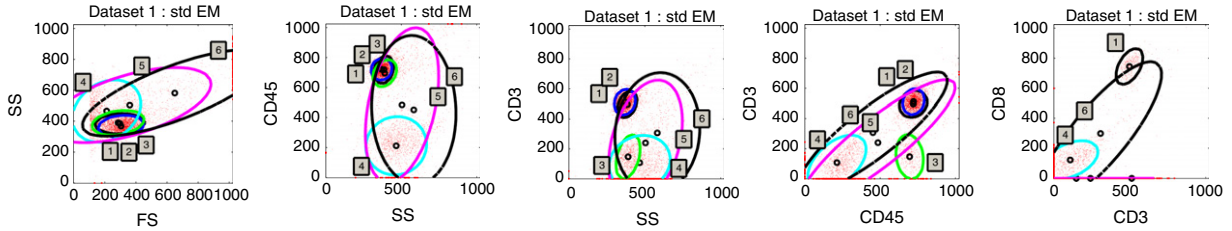
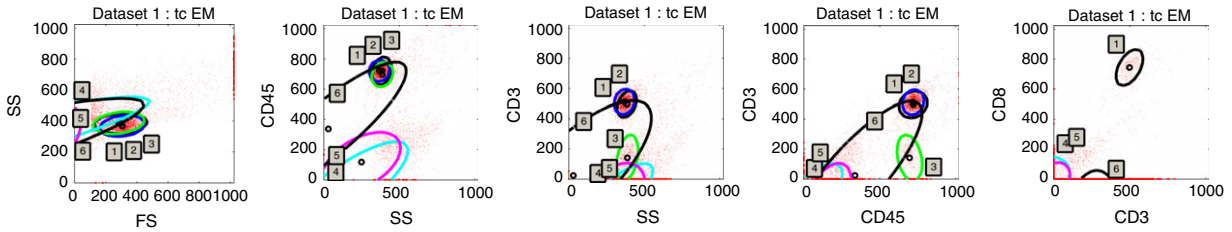


Fig. 5. The truncated and censored data EM algorithm terminates when the log-likelihood value changes less than a predefined constant or the number of iterations reaches 500.



(a) Standard EM algorithm.



(b) Truncated and censored EM algorithm.

Fig. 6. For the first flow cytometry dataset, the mixture model fits using the standard EM algorithm and the truncated and censored EM algorithm are shown. The level contour and centroid 'o' of each component are indicated and labeled. Lymphocyte populations (components 1, 2, and 3) were found well by both algorithms.

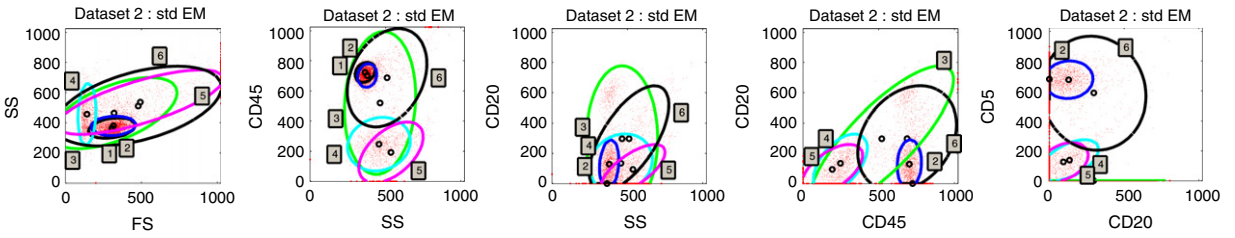
censored, the component centers from the truncated and censored data EM algorithm were located outside the observation window. In the second flow cytometry dataset, the key difference is that the standard EM failed to find the *B*-lymphocytes ($CD45^+CD20^+$) while component 3 in Fig. 7(b) clearly identified the *B*-cells. The truncated and censored data EM also estimated that the centers of components 1 and 2 have negative CD20 values because a large amount of $CD45^+CD20^-$ lymphocytes were censored.

8. Discussion

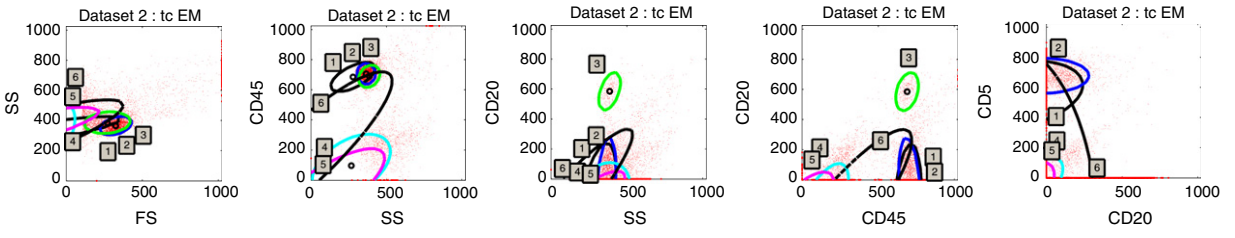
In this paper, we addressed the problem of fitting multivariate Gaussian mixture models to truncated and censored data. We presented EM algorithms and showed that their computation can be achieved using the properties of truncated multivariate normal distributions. Simulation results on synthetic datasets showed that the proposed algorithm corrects for the biases caused by truncation and censoring, and significantly outperforms the standard EM algorithm. We also applied the truncated and censored data EM algorithm to automatic gating of flow cytometry data and compared the gating results to the standard EM algorithm. Our results suggest that the proposed algorithm can be effective in identifying clinically important cell populations in flow cytometry data analysis.

Although these algorithms can be readily applied to lower dimensional data, they depend on methods for evaluating a multivariate normal cumulative distribution function, and the algorithms can run slower as the dimension increases. However, ever-growing computing power will lower the hindrance to using these algorithms in the future.

Several criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been proposed to select the number of components for finite mixture models. However, they are based on complete



(a) Standard EM algorithm.



(b) Truncated and censored EM algorithm.

Fig. 7. The results for the second flow cytometry dataset are displayed. While the standard EM result failed to find the $CD45^{+}CD20^{+}$ *B*-lymphocytes, the truncated and censored EM found this cell population (component 3 in (b)).

measurements, and it is not straightforward to use them for censored and truncated data. We envision that extending these model selection criteria to the truncated and censored data is an interesting problem for future work.

Another future direction is developing stochastic versions of the truncated and censored data EM algorithm. The stochastic EM (SEM) algorithms are known to be less susceptible to the initialization. While (Chauveau, 1995) proposed SEM algorithms for censored data, his focus was on univariate data. Therefore, the SEM for multivariate mixture models would be an interesting next step.

Acknowledgments

The authors are grateful to Dr. Lloyd Stoolman, Usha Kota and Dr. William Finn, University of Michigan Department of Pathology, for data and helpful discussions. This work was supported in part by NSF Award No. 0953135. G. Lee was partially supported by the Edwin R. Riethmiller Fellowship.

Appendix. Truncated multivariate normal

We consider here some key properties of truncated multivariate normal distributions used in this paper. The first two moments are derived and the derivatives of normal integrals are related to the moments.

A.1. First and second moments

Tallis (1961) derived the moment generating function of a standardized and truncated normal distribution. Then he derived the first and the second moments from the moment generating function. Here we extend his approach to a normal distribution with arbitrary mean and covariance that is truncated above and below, and show that we can simplify the computation of the first two moments. We note that a similar derivation appeared in Manjunath and Wilhelm (2009).

Let $\mathbf{X} \in \mathbb{R}^d$ be normally distributed with a probability density function $\phi_d(\mathbf{x}; \mathbf{0}, \Sigma)$ where

$$\phi_d(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

We consider the nonzero mean case later in this section. Suppose a truncation of $\mathbf{X}$ below at $\mathbf{a}$ and above at $\mathbf{b}$ and denote

$$\alpha = P(\mathbf{a} \le \mathbf{X} \le \mathbf{b}) = \int_{\mathbf{a}}^{\mathbf{b}} \phi_d(\mathbf{x}; \mathbf{0}, \Sigma) d\mathbf{x} = \Phi_d(\mathbf{a}, \mathbf{b}; \mathbf{0}, \Sigma)$$

where the inequality is component-wise and $\Phi_d(\mathbf{a}, \mathbf{b}; \mathbf{0}, \Sigma)$ denotes the normal integration over the rectangle with vertices $\mathbf{a}$ and $\mathbf{b}$. Then the moment generating function is

$$m(\mathbf{t}) = \frac{1}{\alpha}\int_{\mathbf{a}}^{\mathbf{b}} \exp(\mathbf{t}^T\mathbf{x})\phi_d(\mathbf{x}; \mathbf{0}, \Sigma) d\mathbf{x} = \frac{\exp\left(\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right)}{\alpha}\int_{\mathbf{a}-\Sigma\mathbf{t}}^{\mathbf{b}-\Sigma\mathbf{t}} \phi_d(\mathbf{x}; \mathbf{0}, \Sigma) d\mathbf{x}. \quad (A.1)$$

We can find the first moment and the second moment from (A.1). We first differentiate (A.1) with respect to t_i and evaluate at $\mathbf{t} = 0$. Then

$$\alpha \left.\frac{\partial m(\mathbf{t})}{\partial t_i}\right|_{\mathbf{t}=0} = \alpha \mathbb{E}[X_i] = \sum_{k=1}^{d} \sigma_{i,k}(F_k(a_k) - F_k(b_k)) \tag{A.2}$$

where $\sigma_{i,k} = [\Sigma]_{i,k}$ and

$$\begin{aligned} F_k(x) &= \int_{\mathbf{a}_{-k}}^{\mathbf{b}_{-k}} \phi_d(\mathbf{x}; \mathbf{0}, \Sigma) d\mathbf{x}_{-k} \\ &= \phi_1(x; 0, \sigma_{k,k}) \int_{\mathbf{a}_{-k}}^{\mathbf{b}_{-k}} \phi_{d-1}(\mathbf{x}_{-k}; \boldsymbol{\mu}_{-k|k}(x), \Sigma_{-k|k}) d\mathbf{x}_{-k} \\ &= \phi_1(x; 0, \sigma_{k,k}) \Phi_{d-1}(\mathbf{a}_{-k}, \mathbf{b}_{-k}; \boldsymbol{\mu}_{-k|k}(x), \Sigma_{-k|k}). \end{aligned}$$

Here we used $-k$ to denote the set of elements $\{1, \ldots, (k-1), (k+1), \ldots, d\}$ other than the kth. The conditional mean and covariance are $\boldsymbol{\mu}_{-k|k}(x) = \Sigma_{-k,k}\Sigma_{k,k}^{-1} x$ and $\Sigma_{-k|k} = \Sigma_{-k,-k} - \Sigma_{-k,k}\Sigma_{k,k}^{-1}\Sigma_{k,-k}$.

Taking the derivatives of (A.1) with respect t_i and t_j at $\mathbf{t} = 0$ gives the second moment

$$\begin{aligned} \alpha \left.\frac{\partial^2 m(\mathbf{t})}{\partial t_i \partial t_i}\right|_{\mathbf{t}=0} &= \alpha \mathbb{E}[X_i X_j] \\ &= \alpha\sigma_{i,j} + \sum_{k=1}^{d} \frac{\sigma_{i,k}\sigma_{j,k}}{\sigma_{k,k}}(a_k F_k(a_k) - b_k F_k(b_k)) \\ &\quad + \sum_{k=1}^{d} \sigma_{i,k} \sum_{q \neq k} \left(\sigma_{j,q} - \frac{\sigma_{k,q}\sigma_{j,k}}{\sigma_{k,k}}\right) [F_{k,q}(a_k, a_q) + F_{k,q}(b_k, b_q) - F_{k,q}(a_k, b_q) - F_{k,q}(b_k, a_q)] \end{aligned} \tag{A.3}$$

where

$$\begin{aligned} F_{k,q}(x_k, x_q) &= \int_{\mathbf{a}_{-(k,q)}}^{\mathbf{b}_{-(k,q)}} \phi_d(x_k, x_q, \mathbf{x}_{-(k,q)}; \mathbf{0}, \Sigma) d\mathbf{x}_{-(k,q)} \\ &= \phi_2(x_k, x_q, ; \mathbf{0}, \Sigma_{(k,q),(k,q)}) \int_{\mathbf{a}_{-(k,q)}}^{\mathbf{b}_{-(k,q)}} \phi_{d-2}(\mathbf{x}_{-(k,q)}; \boldsymbol{\mu}_{-(k,q)|(k,q)}(x_k, x_q), \Sigma_{-(k,q)|(k,q)}) d\mathbf{x}_{-(k,q)} \\ &= \phi_2(x_k, x_q, ; \mathbf{0}, \Sigma_{(k,q),(k,q)}) \Phi_{d-2}(\mathbf{a}_{-(k,q)}, \mathbf{b}_{-(k,q)}; \boldsymbol{\mu}_{-(k,q)|(k,q)}(x_k, x_q), \Sigma_{-(k,q)|(k,q)}). \end{aligned}$$

Likewise, $-(k, q)$ indicates the set of elements except the kth and qth elements. $\boldsymbol{\mu}_{-(k,q)|(k,q)}(x_k, x_q)$ and $\Sigma_{-(k,q)|(k,q)}$ are also defined similarly.

Therefore, we can compute the first moment (A.2) and the second moment (A.3) from a density function and a normal integration, which can be evaluated from the cumulative distribution function and are available in many statistical toolboxes (for example, Fortran, R or Matlab). In particular, `mvncdf` function in Matlab 7.9.0 evaluates the multivariate cumulative probability using the methods developed by Drezner and Wesolowsky (1989) and by Genz (2004) for bivariate and trivariate Gaussian. For four or more dimensional Gaussians, it uses a quasi-Monte Carlo integration algorithm developed by Genz and Bretz (1999, 2002).

Note that $\frac{F_k(x)}{\alpha}$ and $\frac{F_{k,q}(x_k,x_q)}{\alpha}$ are univariate and bivariate marginals of X_k and (X_k, X_q).

Now consider a normal distribution $\phi_d(\mathbf{y}; \boldsymbol{\mu}, \Sigma)$ truncated at $\mathbf{a}^*$ and $\mathbf{b}^*$. Then

$$\begin{aligned} \mathcal{M}^1(\boldsymbol{\mu}, \Sigma; [\mathbf{a}^*, \mathbf{b}^*]) &:= \mathbb{E}[\mathbf{Y}] \\ &= \mathbb{E}[\mathbf{X}] + \boldsymbol{\mu}, \end{aligned} \tag{A.4}$$

$$\begin{aligned} \mathcal{M}^2(\boldsymbol{\mu}, \Sigma; [\mathbf{a}^*, \mathbf{b}^*]) &:= \mathbb{E}[\mathbf{Y}\mathbf{Y}^T] \\ &= \mathbb{E}[\mathbf{X}\mathbf{X}^T] + \boldsymbol{\mu}\mathbb{E}[\mathbf{X}]^T + \mathbb{E}[\mathbf{X}]\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T \\ &= \mathbb{E}[\mathbf{Y}]\mathbb{E}[\mathbf{Y}]^T + \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^T \end{aligned} \tag{A.5}$$

where $\mathbb{E}[X_i]$ and $\mathbb{E}[X_i X_j]$ are evaluated at $\mathbf{a} = \mathbf{a}^* - \boldsymbol{\mu}$ and $\mathbf{b} = \mathbf{b}^* - \boldsymbol{\mu}$. In Section 3, we introduced the notations $\mathcal{M}^1$ and $\mathcal{M}^2$ to denote above expectations (A.4) and (A.5).

For example, consider a univariate random variable Y distributed normally with mean μ and variance σ^2. If it is truncated above at 0 (that is, $a^* = -\infty$, $b^* = 0$), then

$$\mathbb{E}[Y] = \mu - \sigma \frac{\phi_1\left(-\frac{\mu}{\sigma}; 0, 1\right)}{\Phi_1\left(-\frac{\mu}{\sigma}; 0, 1\right)},$$

$$\mathbb{E}[Y^2] = \mu^2 + \sigma^2 - \mu\sigma \frac{\phi_1\left(-\frac{\mu}{\sigma}; 0, 1\right)}{\Phi_1\left(-\frac{\mu}{\sigma}; 0, 1\right)}$$

where the fraction of the standard normal density function ϕ and the distribution function Φ is known as the inverse Mills ratio.

A.2. Derivatives

Here we consider the derivatives of $\alpha(\boldsymbol{\mu}, \Sigma) := \int_{\mathbf{a}}^{\mathbf{b}} \phi_d(\mathbf{y}; \boldsymbol{\mu}, \Sigma)d\mathbf{y}$ with respect to $\boldsymbol{\mu}$ and Σ used in the derivation in Section 3, and relate them with the first and the second moments. Taking the derivative of $\alpha(\boldsymbol{\mu}, \Sigma)$ with respect to $\boldsymbol{\mu}$,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}} \int_{\mathbf{a}}^{\mathbf{b}} \phi_d(\mathbf{y}; \boldsymbol{\mu}, \Sigma)d\mathbf{y} &= \int_{\mathbf{a}}^{\mathbf{b}} \Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})\phi_d(\mathbf{y}; \boldsymbol{\mu}, \Sigma)d\mathbf{y} \\ &= \Sigma^{-1}[\alpha \mathcal{M}^1(\boldsymbol{\mu}, \Sigma; [\mathbf{a}, \mathbf{b}]) - \alpha\boldsymbol{\mu}] \\ &= \alpha\Sigma^{-1}\mathcal{M}^1(\mathbf{0}, \Sigma; [\mathbf{a} - \boldsymbol{\mu}, \mathbf{b} - \boldsymbol{\mu}])\end{aligned}$$

where the last equality is from (A.4), so we obtain

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln \int_{\mathbf{a}}^{\mathbf{b}} \phi_d(\mathbf{y}; \boldsymbol{\mu}, \Sigma)d\mathbf{y} = \Sigma^{-1}\mathcal{M}^1(\mathbf{0}, \Sigma; [\mathbf{a} - \boldsymbol{\mu}, \mathbf{b} - \boldsymbol{\mu}]). \tag{A.6}$$

Next if we take the derivative with respect to Σ, we have

$$\begin{aligned}\frac{\partial}{\partial \Sigma} \ln \int_{\mathbf{a}}^{\mathbf{b}} \phi_d(\mathbf{y}; \boldsymbol{\mu}, \Sigma)d\mathbf{y} &= \frac{1}{\alpha}\frac{\partial}{\partial \Sigma} \int_{\mathbf{a}-\boldsymbol{\mu}}^{\mathbf{b}-\boldsymbol{\mu}} \phi_d(\mathbf{y}; \mathbf{0}, \Sigma)d\mathbf{y} \\ &= \frac{1}{\alpha} \int_{\mathbf{a}-\boldsymbol{\mu}}^{\mathbf{b}-\boldsymbol{\mu}} \left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\mathbf{y}\mathbf{y}^T\Sigma^{-1}\right)\phi_d(\mathbf{y}; \mathbf{0}, \Sigma)d\mathbf{y} \\ &= -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\mathcal{M}^2(\mathbf{0}, \Sigma; [\mathbf{a} - \boldsymbol{\mu}, \mathbf{b} - \boldsymbol{\mu}])\Sigma^{-1}\end{aligned} \tag{A.7}$$

where we used the following facts (Magnus and Neudecker, 1999) in the second equality:

$$\frac{\partial}{\partial \Sigma}\mathrm{tr}(\Sigma^{-1}A) = -(\Sigma^{-1}A\Sigma^{-1})^T, \qquad \frac{\partial}{\partial \Sigma}\ln|\Sigma| = (\Sigma^{-1})^T.$$

References

Atkinson, S.E., 1992. The performance of standard and hybrid EM algorithms for ML estimates of the normal mixture model with censoring. Journal of Statistical Computation and Simulation 44, 105–115.

Biernacki, C., Celeux, G., Govaert, G., Langrognet, F., 2006. Model-based cluster and discriminant analysis with the MIXMOD software. Computational Statistics and Data Analysis 51 (2), 587–600. URL http://www.mixmod.org/.

Boedigheimer, M.J., Ferbas, J., 2008. Mixture modeling approach to flow cytometry data. Cytometry Part A 73, 421–429.

Brown, M., Wittwer, C., 2000. Flow cytometry: principles and clinical applications in hematology. Clinical Chemistry 46, 1221–1229.

Cadez, I.V., Smyth, P., MacLachlan, G.J., McLaren, C.E., 2002. Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. Machine Learning 47, 7–34.

Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., Kepler, T., 2008. Statistical mixture modeling for cell subtype identification in flow cytometry. Cytometry Part A 73, 693–701.

Chauveau, D., 1995. A stochastic EM algorithm for mixtures with censored data. Journal of Statistical Planning and Inference 46, 1–25.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38.

Drezner, Z., Wesolowsky, G.O., 1989. On the computation of the bivariate normal integral. Journal of Statistical Computation and Simulation 35, 101–107.

Gelman, A., 2004. Parameterization and bayesian modeling. Journal of the American Statistical Association 99 (466), 537–544.

Genz, A., 2004. Numerical computation of rectangular bivariate and trivariate normal and t probabilities. Statistics and Computing 14, 251–260.

Genz, A., Bretz, F., 1999. Numerical computation of multivariate t probabilities with application to power calculation of multiple contrasts. Journal of Statistical Computation and Simulation 63, 361–378.

Genz, A., Bretz, F., 2002. Comparison of methods for the computation of multivariate t probabilities. Journal of Computational and Graphical Statistics 11, 950–971.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning; Data Mining, Inference and Prediction. Springer Verlag, New York.

Lakoumentas, J., Drakos, J., Karakantza, M., Nikiforidis, G.C., Sakellaropoulos, G.C., 2009. Bayesian clustering of flow cytometry data for the diagnosis of B-chronic lymphocytic leukemia. Journal of Biomedical Informatics 42, 251–261.

Lo, K., Brinkman, R.R., Gottardo, R., 2008. Automated gating of flow cytometry data via robust model-based clustering. Cytometry Part A 73, 321–332.

Magnus, J.R., Neudecker, H., 1999. Matrix Differential Calculus with Applications in Statistics and Econometrics, second ed. Wiley.

Manjunath, B.G., Wilhelm, S., 2009. Moments calculation for the double truncated multivariate normal density (working paper). URL SSRN: http://ssrn.com/abstract=1472153.

McLachlan, G.J., Jones, P.N., 1988. Fitting mixture models to grouped and truncated data via the EM algorithm. Biometrics 44, 571–578.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L.M., Baecher-Allan, C., McLachlan, G.J., Tamayo, P., Hafler, D.A., Jager, P.L.D., Mesirov, J.P., 2009. Automated high-dimensional flow cytometric data analysis. PNAS 106, 8519–8524.

Shapiro, H., 1994. Practical Flow Cytometry, third ed. Wiley-Liss.

Tallis, G.M., 1961. The moment generating function of the truncated multi-normal distribution. Journal of the Royal Statistical Society. Series B (Methodological) 23, 223–229.