打ち切りデータに対する混合モデルのオンラインEM法の導出と大規模集客イベントにおける到着時間分布推定

幸島　匡宏† 　清武　寛† 　松林　達史† 　塩原　寿子† 　戸田　浩之†

† 日本電信電話株式会社 NTT サービスエボリューション研究所
〒 239-0847 神奈川県横須賀市光の丘 1-1
E-mail:
†{kohjima.masahiro,kiyotake.hiroshi,matsubayashi.tatsushi,shiohara.hisako,toda.hiroyuki}@lab.ntt.co.jp

あらまし　本研究では, 有名アーティストの音楽ライブや人気スポーツの国際試合などの大規模集客イベントにおける観客の会場到着時間の分布をイベント当日に逐次収集されるデータから推定する問題を考える. これらのイベントでは, チケット事前販売数や会場規模に基づき総来場者数に関する情報が利用できるため, 到着済みの観客の到着時間だけでなく, ある時点で「あとどのくらい未到着の観客がいるか」という情報が利用できる. そこで本研究では, 上記の問題が打ち切りデータから確率モデルのパラメタをオンラインに推定する問題と捉えることができることを示し, 混合モデルのパラメタをオンラインに推定する手法を提案する. 提案手法を人工データ, 実データに適用した実験結果について報告する.

キーワード　打ち切りデータ, 混合モデル, EM アルゴリズム, オンラインアルゴリズム

1. はじめに

有名アーティストの音楽ライブや人気スポーツの国際試合などの大規模集客イベントにおいては, 会場周辺に数万人規模の観客が集まる. 来場予定の観客がどの時間帯にどの程度やってくるのか, という観客の会場到着時間の分布を, 当日に収集されるデータを利用して精度良く推定することができれば, イベント運営に貢献できると期待される. たとえば, ライブ/試合開始時間直前の入場ピークが予測されれば, それに備えて入場チェックスタッフの数を増員する, という判断をイベント運営者が下せるようになる.

そこで本研究では, 上記大規模集客イベントにおける観客の会場到着時間の分布をイベント当日に逐次収集されるデータから推定する問題を考える. これらのイベントでは, チケット事前販売数や会場規模に基づき総来場者数に関する情報が利用できるため, 到着済みの観客の到着時間だけでなく, ある時点で「あとどのくらい未到着の観客がいるか」という情報が利用できる. この情報は, 未到着の観客の人数分の「到着時間がその時点以降である」ことを表すデータとして表現できる. したがって, 利用するデータは打ち切りデータ (censored data) と呼ばれる, 観測値がある閾値以上のデータについては値が観測されず, 閾値以上であるという情報しか得られないデータとなる (図 1).

イベント当日にデータを逐次収集する状況では, 時間経過につれて新たに到着した来場者のデータが観測でき, データが時々刻々と更新されていく. このような状況で到着時間分布のパラメタを推定するうえでは, 新たに到着したデータを反映して逐次パラメタを更新する, オンラインアルゴリズム [1] [2] [3] が有用であると考えられる.

そこで本研究では, 時々刻々と更新されていく打ち切りデータから到着時間分布を表現する混合モデルのパラメタをオンラインに推定するアルゴリズム, オンライン EMCM アルゴリズム (online EM algorithm for Censored Mixture models) を提案する. 図 2 に提案手法のイメージを表す. 提案手法は, 新たに観測されたデータを反映して, 到着分布の推定結果を逐次更新する手法である. なお, 到着時間分布のモデルに混合モデルを採用したのは, イベント開始前にアーティストグッズやユニフォームなどの物販購入をする, イベント開始ちょうどに間に合うようにする, など観客の目的に応じて観客の到着時間分布は多峰性を持つと考えられるからである. 提案手法は, 各データの所属コンポーネントを表す潜在変数と閾値以上であったために観測されなかった値を表す潜在変数の 2 種類の潜在変数を導入したデータ生成過程を考えることで導かれる.

人工データと実データを用いた実験により, 提案手法の有効

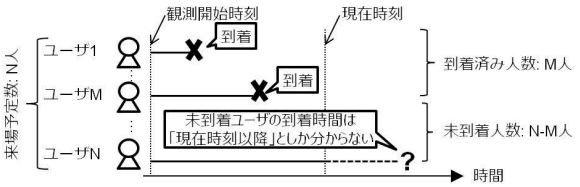


図 1: 到着時間データの打ち切りデータとしての表現

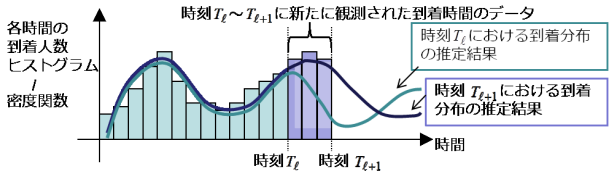


図 2: 提案手法によるオンライン推定.

性を検証した. テストデータに関する対数尤度を評価指標に利用し, 提案手法が, 未到着の観客に関する情報を利用しない既存手法を上回る性能を示すことを確認した.

本稿の構成は次の通りである. § 2 で関連研究, § 3 で提案手法の基となる混合モデルによる打ち切りデータの生成過程とバッチ型の EM アルゴリズムについて述べる. § 4 で提案手法である打ち切りデータを用いたオンライン EM アルゴリズムを示す. § 5 でその有効性を実験的に検証し, § 6 でまとめる.

2. 関 連 研 究

打ち切りデータは, 本研究の設定で考える到着時間データ以外にも, 病気の発症や人の生存期間に関する臨床データ [4] や, 電話回線利用者の契約期間データ [5], E コマースサイト利用者のサービス利用履歴データ [6] など様々なものが存在する. 我々の提案手法は上記いずれのデータにも適用可能である.

提案手法は, 打ち切りデータからパラメタを推定する際に, 潜在変数を導入して EM アルゴリズムを適用する, というアプローチに基づき導出される. このアプローチは古くから存在するものであり [7], 正規分布や指数分布などのパラメタを推定するアルゴリズムが例えば文献 [8] などで示されている.

打ち切りデータから混合モデルのパラメタを (バッチ的に) 推定するという問題に対しては, EMCM アルゴリズム (Expectation-Maximization for Censored Mixture models) [9] と呼ばれる方法が提案されており, 混合指数分布の推定アルゴリズムが導出されている. 我々が提案するオンライン EMCM アルゴリズムという名は上記アルゴリズム名を参考にしたものである. また, 打ち切りデータに対する混合正規分布のバッチ型 EM アルゴリズムは文献 [10] で導出されている. 打ち切りデータに対する混合モデルの推定においては, 所属コンポーネントを表す潜在変数と閾値以上であったために観測されなかった値に関する潜在変数の 2 種類の潜在変数を導入することが鍵となる. 本研究においても, この 2 種類の潜在変数を用いた打ち切りデータの生成過程を考え, オンラインアルゴリズムを導出する.

打ち切りデータではない, いわゆる “通常”のデータに対するオンライン EM アルゴリズムは文献 [1] [2] [3] でその正当性や性質が調べられている. しかしながら, 打ち切りデータに対するオンライン EM アルゴリズムに関してはこれまで十分に研究が進められてこなかった. 本研究の貢献は, 打ち切りデータに対するオンライン EM アルゴリズムを導出し, その有効性を実験的に検証するところにある.

3. 打切りデータに対する混合モデル

3.1 デ ー タ

データの中で値がある既知の閾値 $C \in \mathbb{R}$ 以上となるデータについては値が分からず, 閾値以上であることしかわからない, 右側打ち切りされたデータが得られているとする. 総データ数を N, 値が観測されたか否かを表す変数を $W = \{w_i\}_{i=1}^N$ $(w_i \in \{0, 1\})$ と書く. i 番目データが値を観測されたことを $w_i = 1$, 値が観測されなかったことを $w_i = 0$ と表すこととする. さらに値が観測されたデータの集合とその総数をそれぞれ $\mathcal{I}_o$, M $(\leqq N)$ と書き, 観測された値をまとめて変数 $X = \{x_i\}_{i \in \mathcal{I}_o}$ $(x_i \in \mathbb{R})$ で表す. 閾値 C は既知であり, X, W の 2 つが観測変数である. 上記の定義を含む以後の説明で用いる記号は表 1 にまとめてある. 図 1 に示す到着時間のデータでは, 総データ数 N が来場予定者数, 観測されたデータ数 M が到着済みの観客数を表し, X が到着済みの観客の到着時間を表すと考えれば良い.

表 1: 本論文で用いる記号一覧

記号	意味
N	総データ数
M	値が観測されたデータ数 (観測データ数)$(M \leqq N)$
w_i	i 番目データが観測されたか否かを表す観測変数
x_i	i 番目データの値 (観測変数)
y_j	値が観測されなかった j 番目データの値 (潜在変数)
z_i	i 番目データの所属コンポーネントを表す潜在変数
π_k	k 番目コンポーネントの混合比パラメタ
μ_k	k 番目コンポーネントの平均パラメタ
σ_k	k 番目コンポーネントの標準偏差パラメタ
γ_{ik}	値が観測された i 番目データにおける コンポーネント k の負担率 (式 (11))
$\eta_k(C)$	閾値 C 以上のため値が観測されなかったデータ におけるコンポーネント k の負担率 (式 (12))
$\nu_k(C)$	閾値 C で切断されたコンポーネント k の 1 次モーメント (式 (13))
$\xi_k(C)$	閾値 C のコンポーネント k の 2 次モーメント (式 (14))
M_k	k 番目コンポーネントに属す観測データ数 (式 (17))
N_k	k 番目コンポーネントに属す全データ数 (式 (17))
S_{k1}	観測値 x_i と負担率 γ_{ik} の積の和 (式 (18))
S_{k2}	観測値の 2 乗 x_i^2 と負担率 γ_{ik} の積の和 (式 (18))
U_{k1}	$\nu_k(C)$ と負担率 $\eta_k(C)$ の積の和 (式 (19))
U_{k2}	$\xi_k(C)$ と負担率 $\eta_k(C)$ の積の和 (式 (20))

3.2 モ デ ル

前節の打ち切りデータが混合モデルによって生成されたと考える. 混合モデルの確率密度関数は一般に次の式で定義される.

$$P\left(x|\theta = \{\boldsymbol{\pi}, \boldsymbol{\varphi}\}\right) = \sum\nolimits_{k=1}^{K} \pi_k f(x|\varphi_k). \qquad (1)$$

K はコンポーネント数, $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$, $\boldsymbol{\varphi} = (\varphi_1, \cdots, \varphi_K)$ がモデルのパラメタを表す. π_k, φ_k はそれぞれ k 番目のコンポーネントの混合比とコンポーネントのパラメタを表す. 本稿では特にコンポーネントとして正規分布を採用した場合を考える (注1). 正規分布の確率密度関数は平均 μ_k と標準偏差 σ_k の 2 種類コンポーネントのパラメタ $\varphi_k = \{\mu_k, \sigma_k\}$ を用いて, 次の式で与えられる.

$$f(x|\varphi_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\Big(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\Big). \qquad (2)$$

また, 以後正規分布の累積密度関数を関数 F で表す.

$$F(C|\varphi_k) = \int_{-\infty}^{C} f(x|\varphi_k)dx. \qquad (3)$$

(注1): 本稿の議論は指数型分布族に属す確率分布や, 対数正規分布など, 他の確率分布の混合を考える場合でもほぼ同様に成り立つ.

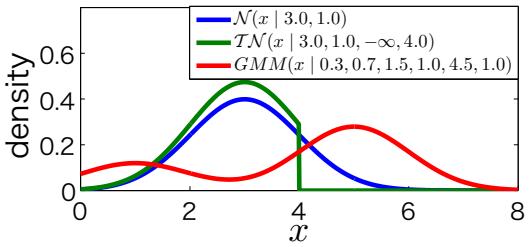


図 3: 正規分布, 切断正規分布, 混合正規分布の確率密度関数

混合モデルによる右側打ち切りされたデータの生成過程は 4 ステップから成る. まず, (i 番目) データが所属するコンポーネントを表す潜在変数 $z_i = (z_{i1}, \cdots, z_{iK})$ (注2) が, 多項分布

$$P(z_i|\boldsymbol{\pi}) = \mathrm{Mult}(z_i|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{ik}} \tag{4}$$

に従い生成される. 次に, 値が観測されるか否かを表す観測変数 w_i がベルヌーイ分布

$$\begin{aligned} P(w_i|z_i, \boldsymbol{\varphi}) &= \prod_{k=1}^{K} \mathrm{Bernoulli}(w_i|F(C|\varphi_k))^{z_{ik}} \\ &= \prod_{k=1}^{K} \Big(F(C|\varphi_k)^{w_i} \big(1 - F(C|\varphi_k)\big)^{1-w_i} \Big)^{z_{ik}} \end{aligned} \tag{5}$$

に従って生成される. さらに, $w_i = 1$, すなわち観測可能となったデータは, 観測変数 $x_i \in (-\infty, C)$ が切断正規分布

$$P(x_i|w_i = 1, z_i, \boldsymbol{\varphi}) = \prod_{k=1}^{K} \mathcal{TN}(x_i|\varphi_k, -\infty, C)^{z_{ik} w_i} \tag{6}$$

に従い生成される. なお, 切断正規分布 $\mathcal{TN}(x|\mu, \sigma^2, a, b)$ は図 3 に示すように範囲 $[a, b]$ 以外では 0 となる以下の確率密度関数で定義される.

$$\mathcal{TN}(x|\mu, \sigma^2, a, b) = \frac{f(x|\mu, \sigma^2)}{F(b|\mu, \sigma^2) - F(a|\mu, \sigma^2)} \tag{7}$$

最後に, $w_i = 0$, すなわち観測不可能となったデータは, 潜在変数 y_i が切断正規分布

$$P(y_i|w_i = 0, z_i, \boldsymbol{\varphi}) = \prod_{k=1}^{K} \mathcal{TN}(y_i|\varphi_k, C, \infty)^{z_{ik}(1-w_i)} \tag{8}$$

に従い生成される. 式 (6)(8) で切断正規分布の取りうる範囲が違うことに注意されたい. 以上を全てのデータ $i = 1, \cdots, N$ に関して繰り返すことで, 観測変数 X, W と潜在変数 Z, Y が生成される. 上記をまとめたデータ生成過程は次の通りである.

データ生成過程

潜在変数 Y, Z と観測変数 W, X の生成

- 全てのデータ $i = 1, \cdots, N$ について

（1） z の生成, $z_i|\boldsymbol{\pi} \sim \mathrm{Mult}(\boldsymbol{\pi})$

（2） w の生成, $w_i|z_i, \boldsymbol{\varphi} \sim \mathrm{Bernoulli}(\{F(C|\varphi_{z_i})\})$

（3） x の生成, $x_i|z_i, w_i = 1, \boldsymbol{\varphi} \sim \mathcal{TN}(x_i|\varphi_{z_i}, -\infty, C)$

（4） y の生成, $y_i|z_i, w_i = 0, \boldsymbol{\varphi} \sim \mathcal{TN}(y_i|\varphi_{z_i}, C, \infty)$

(注2)：i 番目のデータが第 k 番目コンポーネントに属するならば $z_{ik} = 1$, それ以外の $k' \neq k$ については $z_{ik'} = 0$.

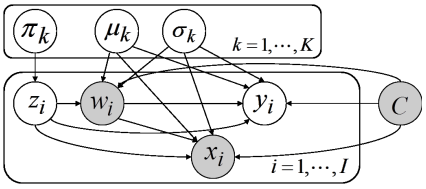


図 4: グラフィカルモデル. 影付きのノードが観測変数を表す.

Algorithm 1 打ち切りデータに対する混合ガウスモデルのバッチ型 EM アルゴリズム

Input: $\boldsymbol{X}, \boldsymbol{W}$: 入力データ, C: (右側) 打ち切り値, K: 混合数
Output: $\theta = \{\pi_k, \mu_k, \sigma_k\}_{k=1}^{K}$: 尤度関数を最大化するパラメタ

```
パラメタ {π_k, μ_k, σ_k} の初期化
repeat
    //E ステップ
    for i = 1 to M do
        負担率 γ_ik を式 (11) に従い計算.
    end for
    負担率 η_k(C) と ν_k(C), ξ_k(C) を式 (12)(13)(14) に従い計算.
    //M ステップ
    各統計量 M_k, N_k, S_k1, S_k2, U_k1, U_k2 を式 (17)(18)(19)(20) に従い計算.
    上記統計量を用いてパラメタ θ を式 (21)(22)(23) に従い更新.
until 収束条件が満たされるまで繰り返し
```

以後表記の簡便さのため, 生成されたデータは $1 \leqq i \leqq M$ で $w_i = 1$, $M + 1 \leqq j \leqq N$ では $w_j = 0$ となるように並び替えてあるとする. このとき, 式 (4)(5)(6)(8) を用いて完全データの尤度関数は次の式で与えられる.

$$\begin{aligned} P(X, W, Y, Z|\theta) &= \Big(\prod_{i=1}^{N} P(z_i|\boldsymbol{\pi}) P(w_i|z_i, \boldsymbol{\varphi}) \Big) \\ &\cdot \Big(\prod_{i=1}^{M} P(x_i|w_i, z_i, \boldsymbol{\varphi}) \Big) \Big(\prod_{j=M+1}^{N} P(y_j|w_j, z_j, \boldsymbol{\varphi}) \Big) \end{aligned} \tag{9}$$

モデルのグラフィカルモデルを図 4 に示す. なお, 上記の式から潜在変数 Y, Z を周辺化して消去すると観測変数 X, W のみの生成確率は下記の式で与えられることが分かる.

$$\begin{aligned} P(X, W|\theta) &= \int \int P(X, W, Y, Z|\theta) dY dZ \\ &= \Big(\prod_{i=1}^{M} P(x_i|\theta) \Big) \cdot \Big(\int_{C}^{\infty} P(y|\theta) dy \Big)^{N-M} \end{aligned} \tag{10}$$

上記の式は後の実験の評価指標として利用する.

3.3 バッチ型 EM アルゴリズム

Expectation-Maximization (EM) アルゴリズムは, 潜在変数を含むモデルの推定に広く利用される手法である (e.g., [11]). 潜在変数の事後確率の算出とそれを用いた期待値の計算からなる E ステップと, $\mathcal{Q}$ 関数と呼ばれる, 対数尤度関数を潜在変数の事後確率に関して平均した関数を最大化する M ステップの 2 ステップからなる.

本モデルの E ステップにおいては, 観測値が得られた場合の事後確率 $P(z_i|x_i, w_i = 1, \theta)$ と得られなかった場合の $P(z_i, y_i|w_i = 0, \theta)$ の 2 つが必要となり, これらはそれぞれ以下の式で与えられる.

$$P(z_i|x_i, w_i = 1, \theta) = \frac{P(z_i, x_i|w_i = 1, \theta)}{P(x_i|w_i = 1, \theta)} = \frac{\pi_k f(x_i|\varphi_k)}{\sum_{k'} \pi_k f(x_i|\varphi_{k'})},$$

$$P(z_j, y_j|w_j = 0, \theta) = P(z_j|w_j = 0, \theta) P(y_j|w_j = 0, z_j, \boldsymbol{\varphi})$$

$$P(z_j|w_j = 0, \theta) = \frac{P(z_j, w_j = 0|\theta)}{P(z_j = 0|\theta)} = \frac{\pi_k \big(1 - F(C|\varphi_k)\big)}{\sum_{k'} \pi_{k'} \big(1 - F(C|\varphi_{k'})\big)}.$$

上記の事後確率を用いて, 各データ x_i に対する各コンポーネント k の負担率 γ_{ik} と, 値が観測されなかったデータに対する負担率とモーメント $\eta_k(C), \nu_k(C), \xi_k(C)$ を計算できる.

$$\gamma_{ik} = \frac{\pi_k f(x_i|\varphi_k)}{\sum_{k'} \pi_k f(x_i|\varphi_{k'})}, \tag{11}$$

$$\eta_k(C) = \frac{\pi_k \big(1 - F(C|\varphi_k)\big)}{\sum_{k'} \pi_{k'} \big(1 - F(C|\varphi_{k'})\big)}, \tag{12}$$

$$\nu_k(C) = \mathbb{E}_{y_j|k}\big[y_j\big] = \mu_k + \sigma_k^2 \frac{f(C|\varphi_k)}{1 - F(C|\varphi_k)}, \tag{13}$$

$$\xi_k(C) = \mathbb{E}_{y_j|k}\big[y_j^2\big] = \mu_k^2 + \sigma_k^2 \Big\{1 + \frac{(C + \mu_k) f(C|\varphi_k)}{1 - F(C|\varphi_k)}\Big\}. \tag{14}$$

ただし, $\mathbb{E}_{y_j|k}$ は事後確率 $P(y_j|w_j = 0, z_j = k, \varphi)$ の出方に関する平均を表し, 切断正規分布の 1 次と 2 次モーメントの結果を利用している. 詳細は Appendix を参考にされたい. また, 以後閾値 C が明らかなときには式 (12)(13)(14) を η_k, ν_k, ξ_k などと省略して表記する. これらを用いると M ステップで最大化する $\mathcal{Q}$ 関数は以下の式で表現される.

$$\mathcal{Q}(\theta, \theta^{old}) = \mathbb{E}_{Z,Y|X,W,\theta^{old}}\big[\log P(X, W, Y, Z|\theta)\big] \tag{15}$$

$$= \sum_{k=1}^{K} N_k \log \pi_k + \sum_{k=1}^{K} -\frac{N_k}{2} \log(2\pi\sigma_k^2) + \sum_{k=1}^{K} -\frac{1}{2\sigma_k^2}\Big\{S_{k2} - 2S_{k1}\mu_k + M_k\mu_k^2\Big\} + \sum_{k=1}^{K} -\frac{1}{2\sigma_k^2}\Big\{U_{k2} - 2U_{k1}\mu_k + (N_k - M_k)\mu_k^2\Big\}. \tag{16}$$

ただし,

$$M_k = \sum_{i=1}^{M} \gamma_{ik}, \quad N_k = M_k + (N - M)\eta_k, \tag{17}$$

$$S_{k1} = \sum_{i=1}^{M} \gamma_{ik} x_i, \quad S_{k2} = \sum_{i=1}^{M} \gamma_{ik} x_i^2, \tag{18}$$

$$U_{k1} = \sum_{j=M+1}^{N} \eta_k \nu_k = (N_k - M_k)\nu_k, \tag{19}$$

$$U_{k2} = \sum_{j=M+1}^{M} \eta_k \xi_k = (N_k - M_k)\xi_k. \tag{20}$$

偏微分をゼロと置いて解くと $\mathcal{Q}$ 関数を最大化するパラメタは

$$\pi_k^{new} = \frac{N_k}{N}, \tag{21}$$

$$\mu_k^{new} = \frac{1}{N_k}\Big(S_{k1} + U_{k1}\Big), \tag{22}$$

$$(\sigma_k^{new})^2 = \frac{1}{N_k}\Big(S_{k2} + U_{k2}\Big) - \frac{1}{N_k^2}\Big(S_{k1} + U_{k1}\Big)^2, \tag{23}$$

で与えられる. これにより打ち切りデータに対する混合モデルのバッチ型 EM アルゴリズムが求められた. Algorithm 1 に手続きをまとめる. E ステップ, M ステップによってパラメタの更新を繰り返し, 各反復において, 対数尤度関数は単調増加し, (局所) 最適解への収束が保証される. また, 値が観測されないデータが存在せず $N = M$ が成立するとき, 全ての k で $U_{k1} = U_{k2} = 0$ となり, 打ち切りデータでない通常のデータに対する混合モデルの EM アルゴリズムと一致する.

4. 提案手法: オンライン EMCM アルゴリズム

Algorithm 1 のバッチ型 EM アルゴリズムは, E ステップでメモリの全データに対して負担率を計算し, それらを用いて M ステップで各統計量 $M_k, N_k, S_{k1}, S_{k2}, U_{k1}, U_{k2}$ の計算を繰り返す. これはすなわち, 値の観測されたデータの値全てをメモリに保持し, 各反復でこの値全体を利用することとなる. それに対して, 我々の提案するアルゴリズム, オンライン EMCM アルゴリズム (online Expectation-Maximization algorithm for Censored Mixture models) は, データ全てをメモリに保持する必要がなく, 新たに観測されたデータのみを利用して, 負担率や統計量を計算してパラメタ更新を行うことが可能である. これによりバッチ型アルゴリズムと比較して高速かつ省メモリでパラメタ更新が可能となる.

4.1 逐次更新型オンライン EM アルゴリズム

まずはじめに新たにデータが観測されるたびにパラメタを更新する, 逐次更新型のアルゴリズムについて述べる. 提案アルゴリズムの導出は, 値が観測されたデータに関する統計量 M_k, S_{k1}, S_{k2} が逐次的な形で書けることを利用する.

あるパラメタ θ が固定された状態で, $i = 1, \cdots, t-1$ までの $M^{(t-1)}(= t-1)$ 個の観測データを用いて計算された統計量 $M_k^{(t-1)}, S_{k1}^{(t-1)}, S_{k2}^{(t-1)}$ がすでに得られているとする (注3). このとき, 観測データ数の 1 つ増えた $i = 1, \cdots, t$ までの $M^{(t)}(= t)$ 個の観測データを用いて計算された統計量 $M_k^{(t)}, S_{k1}^{(t)}, S_{k2}^{(t)}$ は, t 番目データ x_t の値を用いて計算される負担率 γ_{tk} を用いて

$$M_k^{(t)} = M_k^{(t-1)} + \gamma_{tk}, \tag{24}$$

$$S_{k1}^{(t)} = S_{k1}^{(t-1)} + \gamma_{tk} x_t, \quad S_{k2}^{(t)} = S_{k2}^{(t-1)} + \gamma_{tk} x_t^2, \tag{25}$$

と計算できる. よって, 観測データ数が $M^{(t)}$ である時の上記以外の統計量も式 (17)(19)(20) より以下の式で求められる.

$$N_k^{(t)} = M_k^{(t)} + (N - M^{(t)})\eta_k(C), \tag{26}$$

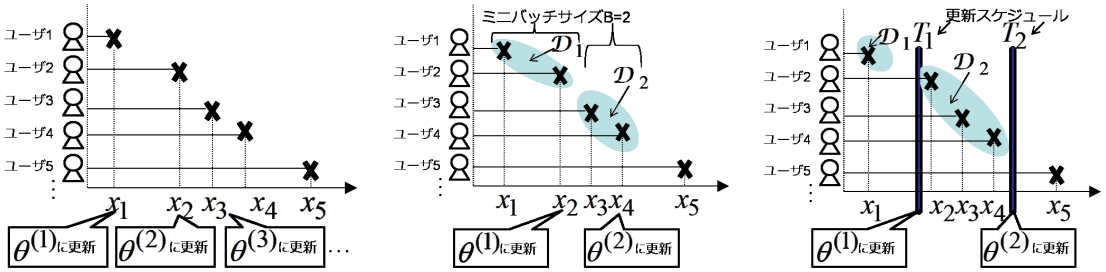
$$U_{k1}^{(t)} = (N_k^{(t)} - M_k^{(t)})\nu_k(C), \tag{27}$$

$$U_{k2}^{(t)} = (N_k^{(t)} - M_k^{(t)})\xi_k(C). \tag{28}$$

したがって, 統計量 $M_k^{(t)}, N_k^{(t)}, S_{k1}^{(t)}, S_{k2}^{(t)}, U_{k1}^{(t)}, U_{k2}^{(t)}$ が得られたため, 式 (21)(22)(23) によって t 番目データ x_t を反映させた混合モデルのパラメタを得ることができる.

提案アルゴリズムは上記の手続きを新たな観測データが手に入るたびに行うことでパラメタを推定する手法である. 提案アルゴリズムを Algorithm 2 に示す. バッチ型の手法 (Algorithm 1) が E ステップで全ての観測データの負担率を計算しているのに対し, 提案アルゴリズムでは新しい観測データのみに関して負担率を計算し, パラメタを更新する. よって, 全ての観測データ

(注3) : $M_k^{(t-1)} = \sum_{i=1}^{t-1} \gamma_{ik}$, $S_{k1} = \sum_{i=1}^{t-1} \gamma_{ik} x_i$, $S_{k2} = \sum_{i=1}^{t-1} \gamma_{ik} x_i^2$.



(a) 逐次更新型　(b) ミニバッチ型　(c) スケジュール型

図 5: 提案する 3 種類のオンライン EM アルゴリズム.

Algorithm 2 打ち切りデータに対するオンライン EM アルゴリズム (逐次更新型)

Input: N: データ数, K: 混合数
Output: $\{\theta^{(t)}\}_{t=1}^{N}$

1: パラメタ $\{\pi_k^{(0)}, \mu_k^{(0)}, \sigma_k^{(0)}\}$, 統計量 $M_k^{(0)}, S_{k1}^{(0)}, S_{k2}^{(0)}$ の初期化
2: 観測データ数を初期化 $M^{(0)} \leftarrow 0$
3: **for** $t = 1$ to N **do**
4: 　データ x_t を取得. 観測データ数を更新: $M^{(t)} \leftarrow M^{(t-1)} + 1$
5: 　閾値を更新: $C \leftarrow x_t$
6: 　//E ステップ
7: 　負担率 γ_{tk} を式 (11) に従い計算.
8: 　負担率 $\eta_k(C)$ と $\nu_k(C), \xi_k(C)$ を式 (12)(13)(14) に従い計算.
9: 　//M ステップ
10: 　各統計量 $M_k^{(t)}$, $N_k^{(t)}$, $S_{k1}^{(t)}$, $S_{k2}^{(t)}$, $U_{k1}^{(t)}$, $U_{k2}^{(t)}$ を式 (24)(25)(26)(27)(28) に従い計算.
11: 　パラメタ $\theta^{(t)}$ を式 (21)(22)(23) に従い更新.
12: **end for**

を保持する必要なく, パラメタの更新が可能となっている.

また, Algorithm 2 の 5 行目の閾値更新処理は, 到着分布推定において, 新しい観測データ x_t の到着が次のユーザの到着時間は x_t 以降であること, すなわち閾値 C が x_t に変更されたことを反映するための処理である.

4.2 ミニバッチ型&スケジュール型アルゴリズム

4.1 節では, データが新たに到着するたびパラメタ更新を行う (a) 逐次更新型アルゴリズムを示した (図 5a). しかし, データ到着時の毎回のパラメタ更新は必須でなく, 図 5b5c に示すように, データ観測と更新のタイミングの異なるアルゴリズムを導出できる. したがってこの節では, 4.1 節の (a) 逐次更新型に加えて, (b) ミニバッチ型, (c) スケジュール型の 2 種類のアルゴリズムを示す.

まず, ミニバッチ型について説明する. この方法では, あらかじめパラメタ更新までに蓄えるデータの数 B(これをミニバッチサイズと呼ぶ) を定めておき, この数のデータが蓄えられた時点でパラメタ更新を行う. E ステップにおいて, 蓄えられたデータ B 個のデータ全ての負担率を計算する点が逐次更新型との違いとなる.

$$M_k^{(\ell)} = M_k^{(\ell-1)} + \sum\nolimits_{t \in \mathcal{D}_\ell} \gamma_{tk}, \tag{29}$$

$$S_{k1}^{(\ell)} = S_{k1}^{(\ell-1)} + \sum\nolimits_{t \in \mathcal{D}_\ell} \gamma_{tk} x_t, \tag{30}$$

$$S_{k2}^{(\ell)} = S_{k2}^{(\ell-1)} + \sum\nolimits_{t \in \mathcal{D}_\ell} \gamma_{tk} x_t^2, \tag{31}$$

観測されなかったデータに関する統計量の更新式は下記のように逐次更新型のものと全く同一である.

$$N_k^{(\ell)} = M_k^{(\ell)} + (N - M)\eta_k(C), \tag{32}$$

$$U_{k1}^{(\ell)} = (N_k^{(\ell)} - M_k^{(\ell)})\nu_k(C), \tag{33}$$

$$U_{k2}^{(\ell)} = (N_k^{(\ell)} - M_k^{(\ell)})\xi_k(C). \tag{34}$$

あとは上記統計量を用いて M ステップでパラメタを更新すればよい. 提案アルゴリズムの手続きを Algorithm 3 にまとめる. アルゴリズム中の記号 $\lfloor \cdot \rfloor$ は入力値を越えない整数を返す床関数を表す.

次に (c) スケジュール型について説明する. スケジュール型アルゴリズムは (b) ミニバッチ型とほぼ同様であり, 違いは, ミニバッチサイズを決めるのではなく, パラメタ更新時刻を定めておくというところにある. これにより, 例えば, 観測データが一定間隔でまとめて入手されるような状況にアルゴリズムが適用できる. スケジュール型における更新式はミニバッチ型の場合と同一である. 提案アルゴリズムの手続きを Algorithm 4 にまとめる. 後の実験ではこのスケジュール型アルゴリズムを利用する. なお, 上記 3 種類のアルゴリズムの更新方法をミックスさせた方法, たとえばミニバッチと更新スケジュールの両方を利用する方法も同様に構築可能であるが, 割愛する.

5. 実　　験

この章では提案手法の有効性を人工データとイベント来場者データを用いて実験的に検証する.

人工データ: 人工データは, 3.2 節のデータ生成過程を用いてデータ数 $N = 1000$ のデータを作成した. データを生成される際の (真の) パラメタと, 提案アルゴリズムのパラメタの初期値は確率的に設定した. 具体的には, 混合数 $K = 3$ とし, 混合比 π は $\alpha = 5.0$ の対称ディリクレ分布, 平均パラメタ $\mu = (\mu_1, \mu_2, \mu_3)$ は, それぞれ平均 1.0, 4.5, 7.0, 分散 0.3 の正規分布にしたがって生成した. 分散パラメタは真のパラメタ, 初期値ともに $\sigma_1 = \sigma_2 = \sigma_3 = 1.0$ とした. 生成したパラメタに対応する確率分布を図 6a に示す. パラメタとデータの生成をそれぞれ 10 回繰り返すことで, 10 種類の学習用データとテスト用

Algorithm 3 打ち切りデータに対するオンライン EM アルゴリズム (ミニバッチ型)

Input: N: データ数, K: 混合数, B: ミニバッチサイズ
Output: $\{\theta^{(\ell)}\}_{\ell=1}^{L}$
パラメタ $\{\pi_k^{(0)}, \mu_k^{(0)}, \sigma_k^{(0)}\}$, 統計量 $M_k^{(0)}, S_{k1}^{(0)}, S_{k2}^{(0)}$ の初期化
観測データ数を初期化 $M \leftarrow 0$
for $\ell = 1$ to $\lfloor N/B \rfloor$ **do**
  データ $\mathcal{D}_\ell$ を取得. 観測データ数を更新: $M^{(t)} \leftarrow M^{(t-1)} + B$
  閾値を更新: $C \leftarrow \max_{t \in \mathcal{D}_\ell} x_t$
  //E ステップ
  全ての $t \in \mathcal{D}_\ell$ の負担率 γ_{tk} を式 (11) に従い計算.
  負担率 $\eta_k(C)$ と $\nu_k(C), \xi_k(C)$ を式 (12)(13)(14) に従い計算.
  //M ステップ
  各統計量 $M_k^{(\ell)}, N_k^{(\ell)}, S_{k1}^{(\ell)}, S_{k2}^{(\ell)}, U_{k1}^{(\ell)}, U_{k2}^{(\ell)}$ を式 (29)(30)(31)(32)(33)(34) に従い計算.
  パラメタ $\theta^{(\ell)}$ を式 (21)(22)(23) に従い更新.
end for

Algorithm 4 打ち切りデータに対するオンライン EM アルゴリズム (スケジュール型)

Input: N: データ数, K: 混合数, L: 更新回数, 更新スケジュール: $T_1, T_2, \cdots, T_L$
Output: $\{\theta^{(\ell)}\}_{\ell=1}^{L}$
パラメタ $\{\pi_k^{(0)}, \mu_k^{(0)}, \sigma_k^{(0)}\}$, 統計量 $M_k^{(0)}, S_{k1}^{(0)}, S_{k2}^{(0)}$ の初期化
観測データ数を初期化 $M \leftarrow 0$
for $\ell = 1$ to L **do**
  データ $\mathcal{D}_\ell$ を取得. 観測データ数を更新: $M^{(t)} \leftarrow M^{(t-1)} + |\mathcal{D}_\ell|$
  閾値を更新: $C \leftarrow T_\ell$
  //E ステップ
  全ての $t \in \mathcal{D}_\ell$ の負担率 γ_{tk} を式 (11) に従い計算.
  負担率 $\eta_k(C)$ と $\nu_k(C), \xi_k(C)$ を式 (12)(13)(14) に従い計算.
  //M ステップ
  各統計量 $M_k^{(\ell)}, N_k^{(\ell)}, S_{k1}^{(\ell)}, S_{k2}^{(\ell)}, U_{k1}^{(\ell)}, U_{k2}^{(\ell)}$ を式 (29)(30)(31)(32)(33)(34) に従い計算.
  パラメタ $\theta^{(\ell)}$ を式 (21)(22)(23) に従い更新.
end for

データを作成した. テストデータのデータ数は 100000 とした.

イベントデータ: イベント来場者のデータには, 2017 年 11 月 29 日に開催された「YOYOGI CANDLE 2020」(注4) イベントで収集されたデータを利用する. このイベントは, NTT ドコモ代々木ビルに空手選手の演舞や専用アプリの投稿コメントをリアルタイムにプロジェクションマッピングするものであり, 18:25-19:00 と 20:25-21:00 の 2 回に渡って実施された (注5). 本実験で利用するデータは, 上記イベントの観覧エリアの 1 つに到着または通過した人数を 1 分間隔でカウントしたものである. データを図 6b に示す. 水色の塗りつぶされた領域がイベントの実施時間を表している. アルゴリズムの適用にあたり, モデルの混合数は $K = 2$ とし, 混合数と平均パラメタの初期値を $\pi_1 = \pi_2 = 0.5$, $\mu_1 = 25.0, \mu_2 = 145.0$ と設定した. 平均パラメタの初期値はイベント開始時刻に対応する. 分散パラメタの初期値は $\sigma_1 = \sigma_2 = 10.0$ とした. また, 本イベントは来場する総観客数を事前に知ることができないイベントであるため, データの到着/通過人数の総和を総データ数 N として利用した. データを 5 分割し, 全体の 8 割を学習用データ, 残りの 2 割をテスト用データとする 5 種類のデータセットを作成した.

比較手法: 提案手法 (Proposed) の比較手法には, 値の観測されないデータに関する情報 (未到着の観客に関する情報) を用いない "通常"のオンライン EM アルゴリズム (Baseline) を採用した. 両者ともにスケジュール型のアルゴリズム (Algorithm 4) を利用し, 更新スケジュールの間隔 $|T_{\ell+1} - T_\ell|$ は, 人工データを用いた実験では 0.4, イベントデータでは 5.0 と設定した. 提案手法とモデルの混合数と初期値は同じものを利用した.

評価指標: 評価指標にはテストデータに対する対数尤度を利用する. ただし, 学習データが打ち切りデータであることを考慮し, 閾値 C で右打ち切りされたデータから推定したモデルパラメタ θ に対しては, テストデータも学習データと同一の閾値 C で打ち切られるよう変換したうえで得たテストデータ X^{test}, W^{test} に対する対数尤度 (Test Average Censored Log-Likelihood, Test-ACLL と呼ぶこととする) を以下の式で計算した.

$$\text{TestACLL} = \frac{1}{N_{test}} \log P(X^{\text{test}}, W^{\text{test}}|\theta) \tag{35}$$
$$= \frac{1}{N_{test}} \sum_{i=1}^{M_{test}} \log P(x_i^{\text{test}}|\theta) + \left(1 - \frac{M_{test}}{N_{test}}\right) \log \int_C^\infty P(x|\theta) dx.$$

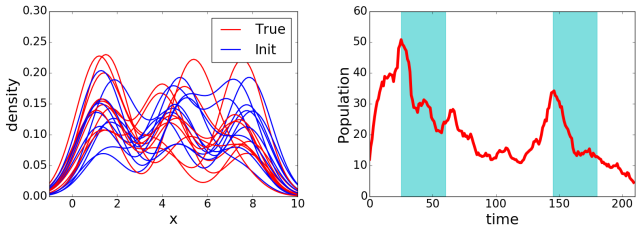
ただし, N_{test} は (観測されないデータを含んだ) 全テストデータ数, M_{test} は観測されたテストデータ数を表す. TestACLL は値が大きいほどモデルが正しく真の確率分布を表現できていることを表す. 打ち切りデータでない, 通常のテストデータに対する対数尤度では右辺の第 1 項のみを計算するが, TestACLL では右辺第 2 項で "観測できないが存在する"データに対する生成確率を同時に評価している. パラメタを更新する毎に上記指標を計算し各手法を比較する.

5.1 実験結果

定量評価: 人工データとイベントデータを用いた実験の定量評価結果を図 7 に示す. どちらのデータにおいても, 推定を行った時刻によらず提案手法が比較手法よりも高い性能を発揮していることが確かめられる. 特に値の観測されているデータ数が小さい, 早い時刻 (x 軸の原点に近い領域) においてその差は顕著である. これは観測データ数が少ないことによって, 比較手法が観測データに過剰にフィットすることが原因であると考えられる. 提案手法は, 値の観測されていない未到着の観客のデータも利用することで, 推定性能を向上できている.

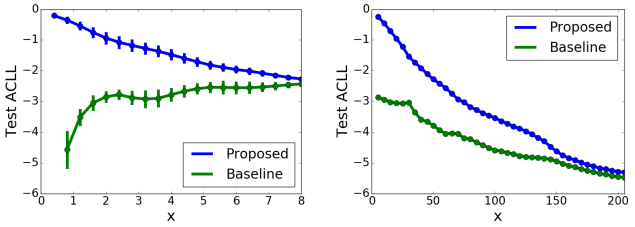
定性評価: 図 8 に人工データを用いた実験における各時刻での推定結果の確率密度関数を示す. 既存手法は提案手法と同一の混合数 $K = 3$ のモデルを利用しているにも関わらず, 単峰であるかのような推定結果を得ている. これは学習の初期時点で値が観測されたデータに過剰にフィットし, 観測データのない領域に生成確率をもつコンポーネントの混合比を小さい値と推

(注4): https://yyg.tokyo
(注5): http://www.ntt.co.jp/topics/pdf/topics_20171130.pdf



(a) 人工データ　　(b) イベントデータ

図 6: (a) データ生成に用いた真の確率分布と初期値パラメタに対応する確率分布. (b) 利用したイベントデータの移動平均.



(a) 人工データ　　(b) イベントデータ

図 7: (a) 人工データ実験と (b) イベントデータ実験による提案手法と比較手法 (通常のオンライン EM) のテストデータに対する対数尤度 (Test ACLL). 値が大きいほど良い.

定した影響が後まで残ることが原因である. それに対し, 提案手法は, 値の観測されていないデータを利用することで, 観測データのみに過剰にフィットすることなく, 学習の最終段階では真の分布をよく表現する推定結果を得ている. イベントデータへの適用結果である図 9 においても, 同様のことが確認できる (注6). さらに, 図 8f, 8g と図 9e, 9f に着目すると, 提案手法は値の観測されていないデータを利用することで, 破線より右側の現在時刻より未来の時刻に中心座標を持つコンポーネントの混合比 (図中の山の高さに相当) を正しく推定することができている. これは, (少なくとも一定時刻まで経過した後には) 提案手法が将来時刻の到着時刻の来場ピークの高さを推定しうることを示している.

6. ま と め

本研究では, 有名アーティストの音楽ライブや人気スポーツの国際試合などのイベントの当日に収集される観客の到着時間に関するデータが打ち切りデータとして表現されることに着目し, 到着時間分布を表現する混合モデルのモデルパラメタをオンラインに推定する手法を提案した. 提案手法は, バッチ型のアルゴリズムと異なり全ての観測データを保持する必要なくパラメタの更新が可能である. 人工データ, 実データを利用した実験によって, 提案手法が未到着の観客に関する情報を利用しない既存技術を上回る性能を示すことを確認した. 実環境での適用と検証が今後の課題である.

(注6): 提案手法による区間 $[a, b]$ の間に到着する人数 $N_{a:b}$ の予測値は, $N_{a:b} = N \times \sum_{k=1}^{K} \pi_k \left\{ F(b|\varphi_k) - F(a|\varphi_k) \right\}$ により求めた.

Appendix: 切断正規分布の性質

"標準"正規分布の確率密度を ϕ, 累積密度を Φ で表す. 文献 [12] より切断正規分布 $\mathcal{TN}(x|\mu, \sigma^2, a, b)$ に従う確率変数 x の平均, 分散はそれぞれ以下で与えられる.

$$\mathbb{E}[x] = \mu + \sigma \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)},$$

$$\mathbb{E}[\{x - \mathbb{E}[x]\}^2] = \sigma^2 \left[1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} \right)^2 \right].$$

ただし, $\alpha = \frac{a-\mu}{\sigma}$, $\beta = \frac{b-\mu}{\sigma}$ とおいた. これより 2 次モーメントも以下の式で与えられることが分かる.

$$\begin{aligned} \mathbb{E}[x^2] &= \mathbb{E}[\{x - \mathbb{E}[x]\}^2] + (\mathbb{E}[x])^2 \\ &= \mu^2 + \sigma^2 + \sigma^2 \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} + 2\mu\sigma \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}. \end{aligned}$$

片側のみ打ち切り, すなわち, $b \to \infty$ のときは 1 次と 2 次のモーメントはそれぞれ

$$\begin{aligned} \mathbb{E}[x] &= \mu + \sigma \frac{\phi(\alpha)}{1 - \Phi(\alpha)} \\ \mathbb{E}[x^2] &= \mu^2 + \sigma^2 + \frac{(\sigma^2\alpha + 2\mu\sigma)\phi(\alpha)}{1 - \Phi(\alpha)}, \\ &= \mu^2 + \sigma^2 + \frac{(a + \mu)\sigma\phi(\alpha)}{1 - \Phi(\alpha)}, \\ &= \mu^2 + \sigma^2 + \frac{(a + \mu)\sigma^2 f(a|\mu, \sigma^2)}{1 - F(a|\mu, \sigma^2)}, \end{aligned}$$

で与えられる. 式 (13)(14) の導出には上記の結果を利用した. なお, 最終行の式変形には以下の標準正規分布と一般の正規分布の密度関数, 累積密度の間に成り立つ関係式

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma}\phi\left(\frac{x - \mu}{\sigma}\right), \; F(c|\mu, \sigma^2) = \Phi\left(\frac{c - \mu}{\sigma}\right) \quad (36)$$

を利用した.

文　献

[1] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pp. 355–368. Springer, 1998.

[2] Masa-Aki Sato and Shin Ishii. On-line em algorithm for the normalized gaussian network. *Neural computation*, Vol. 12, No. 2, pp. 407–432, 2000.

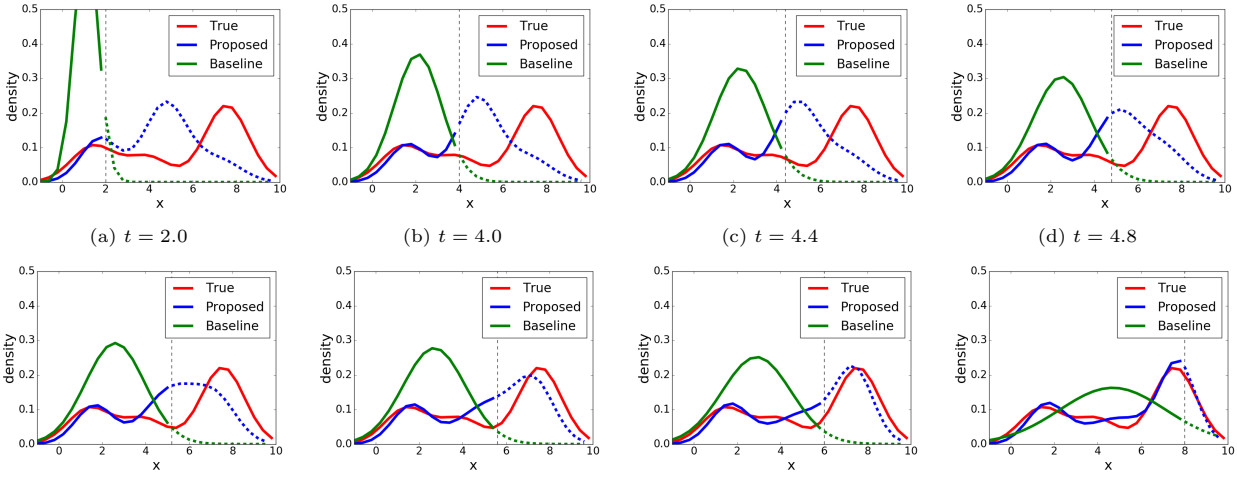
[3] Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 71, No. 3, pp. 593–613, 2009.

[4] サバイバルデータの解析: 生存時間とイベントヒストリデータ. バイオ統計シリーズ. 近代科学社, 2010.

[5] Junxiang Lu. Predicting customer churn in the telecommunications industry—an application of survival analysis modeling using sas. *SAS User Group International (SUGI27) Online Proceedings*, pp. 114–27, 2002.

[6] Shouichi Nagano, Yusuke Ichikawa, Noriko Takaya, Tadasu Uchiyama, and Makoto Abe. Nonparametric hierarchal bayesian modeling in non-contractual heterogeneous survival data. In *Proc. of the 19th ACM SIGKDD int. conf. on Knowledge discovery and data mining*, pp. 668–676, 2013.

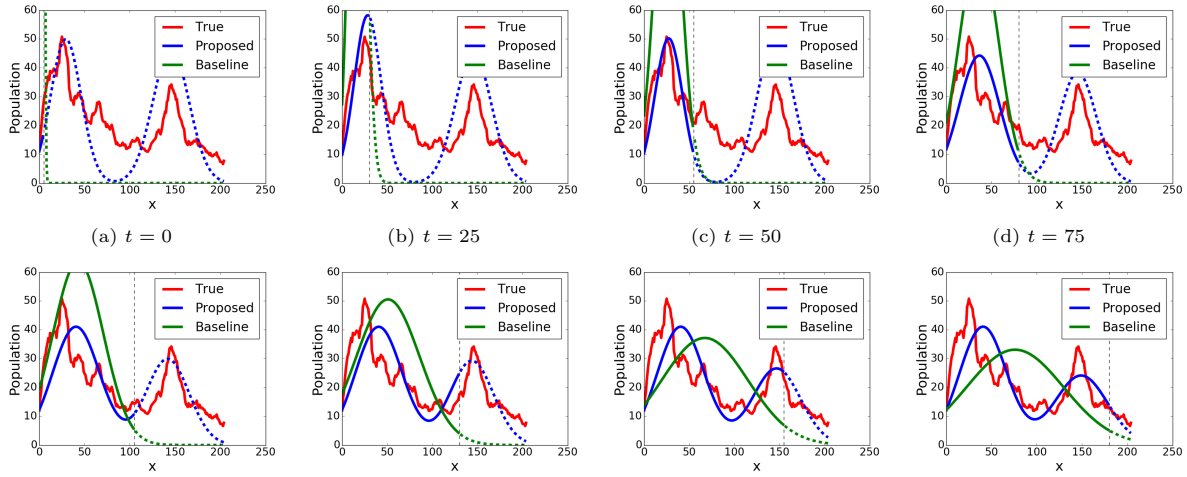
[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin.



(a) $t = 2.0$ (b) $t = 4.0$ (c) $t = 4.4$ (d) $t = 4.8$

(e) $t = 5.2$ (f) $t = 5.6$ (g) $t = 6.0$ (h) $t = 8.0$

図 8: 人工データ実験による提案手法 (Proposed) と比較手法 (Baseline) の時刻 $t = 2.0 \sim 8.0$ における確率密度関数の推定結果の抜粋. 黒の破線はその時点での時刻 (打ち切りの閾値と等しい) を表す.



(a) $t = 0$ (b) $t = 25$ (c) $t = 50$ (d) $t = 75$

(e) $t = 100$ (f) $t = 125$ (g) $t = 150$ (h) $t = 175$

図 9: イベントデータ実験による提案手法 (Proposed) と比較手法 (Baseline) の時刻 $t = 0 \sim 35$ における人数予測の推定結果の抜粋. 黒の破線はその時点での時刻 (打ち切りの閾値と等しい) を表す. True には観測データをスムージングした値を示す.

Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[8] Chanseok Park. Parameter estimation from load-sharing system data using the expectation–maximization algorithm. *IIE Transactions*, Vol. 45, No. 2, pp. 147–163, 2013.

[9] Didier Chauveau. A stochastic em algorithm for mixtures with censored data. *Journal of statistical planning and inference*, Vol. 46, No. 1, pp. 1–25, 1995.

[10] Gyemin Lee and Clayton Scott. Em algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, Vol. 56, No. 9, pp. 2816–2829, 2012.

[11] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[12] NL Johnson, Samuel Kotz, and N Balakrishnan. *Continuous Univariate Probability Distributions,(Vol. 1)*. John Wiley & Sons Inc., NY, 1994.