Exemplar Based Mixture Models with Censored Data

Masahiro Kohjima MASAHIRO.KOHJIMA.EV@HCO.NTT.CO.JP
Tatsushi Matsubayashi TATSUSHI.MATSUBAYASHI.TB@HCO.NTT.CO.JP
Hiroyuki Toda HIROYUKI.TODA.XB@HCO.NTT.CO.JP
NTT Service Evolution Laboratories, NTT Corporation, Yokosuka, Japan

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

In this paper, we propose a method that can handle censored data, data collected under the condition that the exact value is recorded only when the value is within a certain range, abbreviated information is recorded otherwise. It is known that existing methods that use mixture models with censored data suffer from (i) the existence of local optimum solutions and (ii) the need to compute the statistics of truncated distributions for parameter estimation. Our proposal, exemplar based censored mixture model (EBCM), overcomes these two difficulties at once by adopting the exemplar based model approach. The effectiveness of EBCM is confirmed by experiments on synthetic and real world dat sets.

Keywords: exemplar based model, convex clustering, mixture model, censoring, censored data, survival analysis, multivariate survival analysis

1. Introduction

Censored data is the data collected under the condition that the exact value is recorded only when the value is within a certain range; partial information is recorded otherwise Kleinbaum and Klein (2010); Crowder (2012). The censoring occurs for various reasons such as limits of observation period or the of measurement range of sensors. The former is a typical scenario in survival analyses, which treat the lifetimes of devices, humans and so on. Figure 1(a) shows an example of the "time to failure" data of devices. Since observation periods are limited, some devices are still working at the end of the period. Thus, we know only that their time to failure is larger than the elapsed time of testing. The latter scenario reflects sensor limits. For example, if the water level of a river exceeds the observable range of a water gauge, we know only that the level exceeded the maximum value of the gauge. Considering that survival analysis is now the key to various real world problems such as customer lifetime modeling Rosset et al. (2002), click log analysis Chin and Street (2012), and quality analysis of advertisements Barbieri et al. (2016) and that many types of (censored) data are collected by sensors due to the popularity of Internet of Things (IoT), it is clear that the importance of censored data analysis will continue to increase.

Similar to the analysis of standard (not censored) data, the use of mixture models is one promising approach to censored data analysis. For example, mixture models can capture the multimodal structure of the probability distribution such as the time to failure distribution, whose constituents include initial failure and aging failure distributions. To estimate the parameters of mixture models we can apply expectation maximization (EM) Dempster et al.

(1977) based algorithm, EM for censored mixture models (EMCM) Chauveau (1995), or the variational Bayes (VB) Attias (1999); Jordan et al. (1999) based algorithm, VB for censored mixture models (VBCM) Kohjima et al. (2018). However, these two approaches have two problems. The first is that EMCM and VBCM convergence readily fall into local optima since convergence is very sensitive to the initial settings. To ensure adequate coverage, multiple, different initial points must be examined. The second is the necessity of computing the statistics of *truncated* distributions. Since no analytic solution is available, numerical computation is needed, the cost of which increases with the number of dimensions. Since both EMCM and VBCM are iterative algorithms that compute the statistics multiple times, some alternative is essential.

In this study, we propose the exemplar-based censored mixture model (EBCM). EBCM solves the two above difficulties at once, i.e., parameter estimation is assured of converging on global optima and computation of the statistics of the truncated distribution is not required. The idea behind EBCM is to use the approach called the exemplar-based model (EBM) or convex clustering Lashkari and Golland (2008). EBM puts the mixture model's components on (possibly all) data points and considers only the mixing ratio parameter as the parameter to be estimated. This approach converts the objective log-likelihood function into a convex function with no local optima. Furthermore, since EBCM does not need the component parameter, there is no need to compute the statistics of truncated distributions. Thus EBCM solves the two key problems simultaneously. EBCM can also be regarded as a generalization of EBM that permits the input of censored data.

We design component distributions used in EBCM for survival analysis. Since survival data usually have values greater than (or equal to) zero, Gaussian distributions, which support negative and positive values, may not be appropriate. We develop component distribution candidates by using inverse Gaussian and Gamma distributions. We also confirm the effectiveness of EBCM by experiments on both synthetic and real data sets.

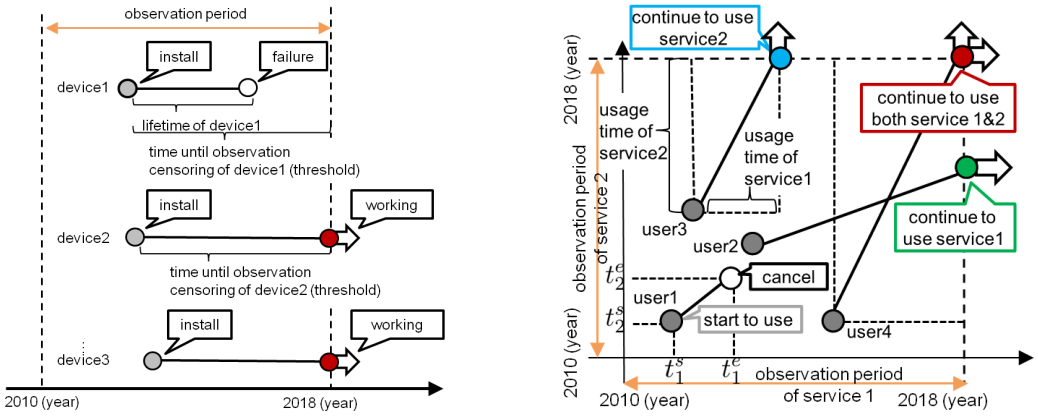
The contributions of this paper are summarized below:

- We develop EBCM, an exemplar-based model that can deal with censored data. The EBCM algorithm converges to the global minimum and does not calculate the moments of truncated distribution.
- We also design component distribution(s) of EBCM for survival analysis. Since survival data usually have values greater (or equal) zero, the constructed distributions should lie only on the positive line.
- We confirm the effectiveness of EBCM by numerical experiments using both synthetic data and real survival data.

The rest of this paper is organized as follows. We introduce related works in § 2 and a description and definition of censored data are given in §3. §4 presents EBCM and its algorithm. The distributions created for survival analysis are also provided. §5 details our experimental evaluation. Finally, §6 concludes the paper.

2. Related Works

Extensions of machine learning models and algorithms for survival analysis have been widely studied in the machine learning and data mining communities Kohjima et al.



(a) 1-dimensional censored data (b) 2-dimensional censored data

Figure 1: Example of censored data. (a) presents one-dimensional censored data representing the failure time of devices. The gray points represent installation times and the white points represent the failure times. The red points represent data points whose failure times were not observed due to observation censoring. (b) presents two-dimensional censored data of user service usage time. The gray points indicate the start time of using the services and the white points represent the cancellation time. The green/blue/red points indicate the data points whose cancellation times were not observed due to the end of observation period.

(2018); Pölsterl et al. (2015); Kiaee et al. (2016); Fernández et al. (2016); Ranganath et al. (2016); Grob et al. (2018). Fernandez et al. developed the Gaussian process based method Fernández et al. (2016) and Kiaee et al. extended the relevance vector machine Kiaee et al. (2016). Ranganath et al. and Grob et al. focus on deep neural nets Ranganath et al. (2016); Grob et al. (2018). Our study also follows this context and its proposal is based on the exemplar-based model Lashkari and Golland (2008).

In the parameter estimation of mixture models with censored data, the conventional approach makes it necessary to compute the cumulative density function (CDF) of the component distribution and the statistics of the truncated distribution (distribution changed to take values only in a certain range). This is because the latent variable indicating the true value, which was not observed due to observation censoring, is introduced, and this variable follows the truncated distribution defined using CDF. Although many studies have been published (e.g., Tallis (1961); Genz and Bretz (1999); Genz (2004); BG and Wilhelm (2009)), the ability to compute the statistics of the truncated distribution is not adequately developed. For example, only the one-dimensional truncated normal distribution and the truncated exponential distribution are implemented in the statistical module scipy.stats. Also, even if they are provided (e.g. "R"s multivariate truncated normal distribution package, "tmvtnorm" Wilhelm and Manjunath (2010)), they often require numerical calculations internally. CDF also require numerical calculations and its computational cost increases with the dimension number. For example, MathWorks's numerical calculation software MATLAB®, uses the adaptive quadrature method Genz (2004) for the calcula-

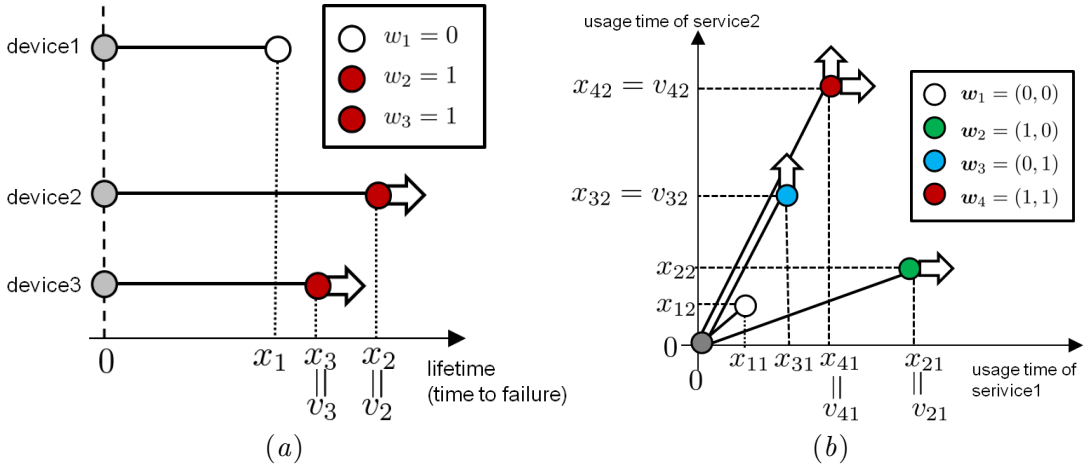


Figure 2: Survival time representation of (a) 1d censored data and (b) 2d censored data shown in Fig. 1).

tion of the cumulative probability of the multidimensional normal distribution if it is three or less, for four or more, the quasi-Monte Carlo integration algorithm must be used Genz and Bretz (1999) [1]. Clearly it is desirable to avoid repeating these calculations. As mentioned above, we constructed a method to avoid repeated numerical calculations by using an example-based model approach.

Advanced studies on the exemplar base model (EBM) can be found in the literature, and some algorithms can find global optimal solutions while avoiding local optima Lashkari and Golland (2008); Sugiyama et al. (2010, 2012). However, none of the above methods can use censored data as input. Our solution, EBCM, is an extension of Lashkari and Golland's approach Lashkari and Golland (2008).

3. Censored Data

3.1. Description

We describe the censored data using the common example of time to failure analysis, see Fig. 1(a). The time of device installation and the time of failure are recorded. For device 1, both the installation time and failure time are recorded. In contrast, for devices 2 and 3, the failure times are not recorded because the observation period ended without them failing. Therefore, the censored data consist of the values (time of failure) for devices 1 while for devices 2 and 3, the entries are written as right censored data.

Although the example uses one-dimensional censored data, the focus of this study includes censored data with more than two dimensions. Figure 1(b) shows an example of two-dimensional censored data; it presents usage periods (contract period) of two services such as movie streaming and music streaming. The time when a user starts to use a service and the time when a user cancels a service are recorded. For user 1, the start and cancel-

1. https://www.mathworks.com/help/stats/mvncdf.html

lation time of service 1, t_1^s and t_1^e, and that of service 2, t_2^s and t_2^e, are observed. However, the time at which user 2 cancels service 1 is not recorded since he/she is still using it at the end of the observation period. Similarly, user 3's cancellation time of service 2 and user 4's cancellation time of services 1 and 2 are not recorded. So the data contain three types of censored situations as multiple dimensions are involved. In general, there are $2^{d_x} - 1$ types of censored situations for d_x-dimensions.

3.2. Definition

This subsection formally defines censored data. We consider that the data are processed and the value represent a survival time such as time to failure and time to cancellation. Figure 2 shows the survival time representation of the data in Fig 1. We use Fig. 2(*b*) as an example in the explanation below.

We denote censored data as $\mathcal{D}=\{\boldsymbol{x}_i, \boldsymbol{w}_i\}_{i=1}^n$. $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ are d_x-dimensional vectors. x_{ij} is user i's usage time of service j and w_{ij} indicates whether the time at which user i's cancelled service j are recorded or not; when recorded we write $(w_{ij} = 0)$, and when not recorded we write $(w_{ij} = 1)$. Similarly, let $\boldsymbol{v}_i \in \mathbb{R}^{d_x}$ be the threshold of user i. v_{ij} is the observation period (length) of user i for service j. We also assume that x_{ij} is set to the threshold, i.e., $x_{ij} = v_{ij}$, when the value is censored $(w_{ij} = 1)$. Although we use the example of censored data for survival analysis, censored data collected by some sensors with limited observed range can be handled in an analogous manner.

4. Exemplar-Based Censored Mixture model (EBCM)

This section provides the proposed method.

4.1. Models

Our proposed method, exemplar-based censored mixture model (EBCM), is constructed on (standard) mixture models. The probability distribution of mixture models is defined as

$$\text{(Standard) Mixure Model}: \quad f(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum\nolimits_{k=1}^{K} \theta_k \psi(\boldsymbol{x}|\eta_k) = \boldsymbol{\theta}^T \boldsymbol{\psi}(\boldsymbol{x}|\boldsymbol{\eta}), \tag{1}$$

where K is the number of components and $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^K$ is the mixing ratio, ψ is the component distribution and $\boldsymbol{\eta} = \{\eta_k\}_{k=1}^K$ are the component parameters. When the component is Gaussian, the component $\psi_{\mathcal{N}}$ is written as follows:

$$\psi_{\mathcal{N}}(\boldsymbol{x}|\boldsymbol{\mu}_k, \sigma^2) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^{d_x}} \exp\Big(-\frac{\|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2}{2\sigma^2}\Big), \tag{2}$$

where $\boldsymbol{\mu}_k$ and σ represent the mean and standard deviation, respectively. Here we consider the component is Gaussian; the details of using other distribution types, e.g., inverse Gaussian distribution $\mathcal{IG}$ and gamma distribution $\mathcal{G}$, are provided at the end of this section.

$$\mathcal{IG}(x|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\Big(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\Big), \quad \mathcal{G}(x|a, b) = \frac{x^{a-1}\exp(-x/b)}{b^a \Gamma\big(a\big)}. \tag{3}$$

In (standard) mixture models, both mixing ratio $\boldsymbol{\theta}$ and component parameter $\boldsymbol{\eta}$ are the parameters to be estimated. On the other hand, EBCM has only the mixing ratio as its parameter as it uses the approach of exemplar-based model (EBM) Lashkari and Golland (2008). This is done by putting the components on the (possibly all) data points [2]. Given (non-censored) standard data $\{\boldsymbol{x}_k\}_{k=1}^n$, the EBM constructs a model with Gaussian component as

$$\text{EBM (Gaussian)}: \quad f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum\nolimits_{k=1}^{n} \theta_k \psi_{\mathcal{N}}(\boldsymbol{x}|\boldsymbol{x}_k, \sigma^2).$$

Note that the number of components K is the number of data n and the mean parameters correspond to the data points. Since the mean parameter is removed, there is no need to estimate it. By treating the standard deviation σ as the hyperparameter that is estimated by e.g., cross-validation, this formulation allows us to build a model whose only parameter is its mixing ratio.

We closely followed the EBM approach in building EBCM. The difference is we reflect the nature of the censored data; when the value is censored ($w_{ij} = 1$), although the value of x_{ij} is set to the threshold v_{ij} by definition, its unobserved *true* value is larger than the threshold. Keeping this in mind, given censored data $\{\boldsymbol{x}_k, \boldsymbol{w}_k\}_{k=1}^n$, we build the model of EBCM with Gaussian component as

$$\text{EBCM (Gaussian)}: \quad f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum\nolimits_{k=1}^{n} \theta_k \psi_{\mathcal{N}}(\boldsymbol{x}|\boldsymbol{z}_k, \sigma^2), \tag{4}$$

where we define the variable $\{\boldsymbol{z}_k\}_{k=1}^n$ as $z_{kj} = x_{kj}$ if $w_{kj} = 0$, and $z_{kj} = x_{kj} + \epsilon_{kj}$ otherwise. Note that ϵ_{kj} is a positive random variable (arbitrary) and we use the standard exponential distribution for its generation in an experiment described later. This procedure allows us to put the component at a value greater than the threshold.

4.2. Generative Process

The generative process of censored data $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{w}_i\}_{i=1}^n$ using the model of EBCM shown in Eq. (4) consists of 2 steps. We use the notation $\psi_k(\boldsymbol{x})$ to indicate the k-th component distribution; it allows us to ignore the choice of the distribution while avoiding any dependency on hyperparameters such as standard deviation.

In the first step, given threshold $\boldsymbol{v}_i$, variable $\boldsymbol{w}_i$ is drawn by the following distribution:

$$\begin{aligned} P(\boldsymbol{w}_i|\boldsymbol{\theta}) = F(\boldsymbol{w}_i|\boldsymbol{\theta}; \boldsymbol{v}_i) &= \int_{\boldsymbol{v}_i^c}^{\infty} \Big\{ \int_{-\infty}^{\boldsymbol{v}_i^o} f(\boldsymbol{x}_i|\boldsymbol{\theta}) d\boldsymbol{x}_i^o \Big\} d\boldsymbol{x}_i^c \\ &= \sum\nolimits_{k=1}^{K} \theta_k \Big[\int_{\boldsymbol{v}_i^c}^{\infty} \Big\{ \int_{-\infty}^{\boldsymbol{v}_i^o} \psi_k(\boldsymbol{x}) d\boldsymbol{x}^o \Big\} d\boldsymbol{x}^c \Big], \end{aligned} \tag{5}$$

where we define $\boldsymbol{x}_i^o$ and $\boldsymbol{v}_i^o$ as the elements of the observed dimension of $\boldsymbol{x}_i$ and $\boldsymbol{v}_i$, i.e., $\boldsymbol{x}_i^o = \{x_{ij}|w_{ij} = 0\}$,$\boldsymbol{v}_i^o = \{v_{ij}|w_{ij} = 0\}$. Similarly, we define $\boldsymbol{x}_i^c$ and $\boldsymbol{v}_i^c$ as $\boldsymbol{x}_i^c = \{x_{ij}|w_{ij} = 1\}$, $\boldsymbol{v}_i^c = \{v_{ij}|w_{ij} = 1\}$. Note that the integral in the equation can be computed using the cumulative density function (CDF) of the component distribution if necessary. As shown

2. If the number of data is large, just randomly chosen data points may be used.

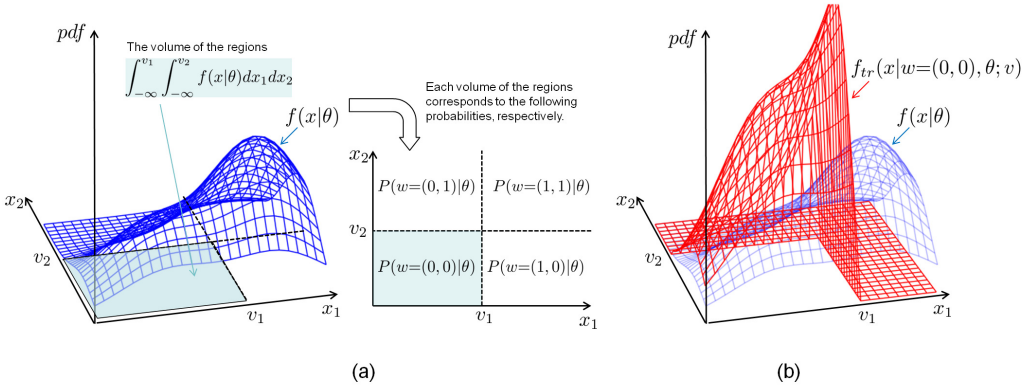


Figure 3: generative process (a) probability of w and (b) that of x given $w = (0, 0)$.

in Fig. 3 (a), $P(\boldsymbol{w}|\boldsymbol{\theta})$ corresponds to the volume of model f in the region divided by the threshold $\boldsymbol{v}$.

In the second step, if at least one element is observed, i.e., $\boldsymbol{w}_i \neq \mathbf{1}$[3], variable $\boldsymbol{x}_i$ is drawn as follows:

$$P(\boldsymbol{x}_i|\boldsymbol{w}_i \neq \mathbf{1}, \boldsymbol{\theta}) = \delta(\boldsymbol{x}_i^c - \boldsymbol{v}_i^c) f_{tr}(\boldsymbol{x}_i^o|\boldsymbol{w}_i, \boldsymbol{\theta}; \boldsymbol{v}_i), \tag{6}$$

where $\delta(\cdot)$ is the delta function and f_{tr} is the truncated distribution

$$f_{tr}(\boldsymbol{x}_i^o|\boldsymbol{w}_i, \boldsymbol{\theta}; \boldsymbol{v}_i) = \begin{cases} \frac{g(\boldsymbol{x}_i^o|\boldsymbol{w}_i, \boldsymbol{\theta}; \boldsymbol{v}_i)}{F(\boldsymbol{w}_i|\boldsymbol{\theta}; \boldsymbol{v}_i)} & (\text{if } \boldsymbol{x}_i^o \leq \boldsymbol{v}_i) \\ 0 & (\text{otherwise}) \end{cases}$$

where

$$g(\boldsymbol{x}_i^o|\boldsymbol{w}_i, \boldsymbol{\theta}; \boldsymbol{v}_i) = \int_{\boldsymbol{v}_i^c}^{\infty} f(\boldsymbol{x}_i|\boldsymbol{\theta}) d\boldsymbol{x}_i^c = \sum_{k=1}^{K} \theta_k \Psi_{ik} = \boldsymbol{\theta}^T \boldsymbol{\Psi}_i(\boldsymbol{x}_i^o, \boldsymbol{w}_i, \boldsymbol{v}_i), \tag{7}$$

$$\Psi_{ik} = \Psi_k(\boldsymbol{x}_i^o, \boldsymbol{w}_i, \boldsymbol{v}_i) = \int_{\boldsymbol{v}_i^c}^{\infty} \psi_k(\boldsymbol{x}_i) d\boldsymbol{x}_i^c. \tag{8}$$

The integral of Ψ_{ik} can be computed using CDF as well. If the component is Gaussian (Eq. (2)),

$$\Psi_{ik}^{\mathcal{N}} = \mathcal{N}(\boldsymbol{x}_i^o|\boldsymbol{z}_k^o(\boldsymbol{w}), \sigma^2) \int_{\boldsymbol{v}_i^c}^{\infty} \mathcal{N}(\boldsymbol{x}_i^c|\boldsymbol{z}_k^c(\boldsymbol{w}), \sigma^2) d\boldsymbol{x}_i^c,$$

where $\boldsymbol{z}_k^o(\boldsymbol{w})$ and $\boldsymbol{z}_k^c(\boldsymbol{w})$ are the vectors extracted from $\boldsymbol{z}_k$, $\boldsymbol{z}_k^o(\boldsymbol{w}) = \{z_{kj}|w_j = 0\}$ and $\boldsymbol{z}_k^c(\boldsymbol{w}) = \{z_{kj}|w_j = 1\}$, respectively. Figure 3 (b) shows an example of distribution truncation where all values are less than threshold ($\boldsymbol{w} = \mathbf{0}$). This distribution has non-zero values

3. $\mathbf{1}$ is the vector with all one elements

only in the region which is smaller than threshold value. If all elements are unobserved by censoring, i.e., $\boldsymbol{w}_i = \mathbf{1}$, all elements of $\boldsymbol{x}_i$ equal threshold $\boldsymbol{v}_i$:

$$P(\boldsymbol{x}_i|\boldsymbol{w}_i = \mathbf{1}, \boldsymbol{\theta}) = \delta(\boldsymbol{x}_i - \boldsymbol{v}_i). \tag{9}$$

Summarizing the above processes, the probability of censored data $\mathcal{D}$ given (mixing ratio) parameter θ can be written as

$$\begin{aligned}
&P(\mathcal{D}|\boldsymbol{\theta}) = \textstyle\prod_{i=1}^{n} P(\boldsymbol{x}_i|\boldsymbol{w}_i, \boldsymbol{\theta}) P(\boldsymbol{w}_i|\boldsymbol{\theta}) \\
&= \prod_{\{i|w_i=\mathbf{1}\}} F(\boldsymbol{w}_i|\boldsymbol{\theta}; \boldsymbol{v}_i)\delta(\boldsymbol{x}_i - \boldsymbol{v}_i) \prod_{\{i'|w_{i'}\neq\mathbf{1}\}} F(\boldsymbol{w}_{i'}|\boldsymbol{\theta}; \boldsymbol{v}_{i'}) f_{tr}(\boldsymbol{x}_{i'}^o|\boldsymbol{w}_{i'}, \boldsymbol{\theta}; \boldsymbol{v}_{i'})\delta(\boldsymbol{x}_{i'}^c - \boldsymbol{v}_{i'}^c) \\
&= \Big\{\textstyle\prod_{i=1}^{n} g(\boldsymbol{x}_i^o|\boldsymbol{w}_i, \boldsymbol{\theta}; \boldsymbol{v}_i)\Big\}\Big\{\prod_{\{i|w_i\neq\mathbf{0}\}} \delta(\boldsymbol{x}_i^c - \boldsymbol{v}_i)\Big\}.
\end{aligned}$$

The objective function is derived by taking the negative logarithm and removing the constant terms:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log g(\boldsymbol{x}_i^o|\boldsymbol{w}_i, \boldsymbol{\theta}; \boldsymbol{v}_i) = -\sum_{i=1}^{n} \log\Big(\boldsymbol{\theta}^T \boldsymbol{\Psi}_i(\boldsymbol{x}_i^o, \boldsymbol{w}_i, \boldsymbol{v}_i)\Big).$$

4.3. Algorithm

The algorithm is derived by minimizing the objective function $\mathcal{L}(\boldsymbol{\theta})$. Since θ is the mixing ratio parameter that satisfies the sum to one constraint, it can be formulated as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \quad \text{s.t.} \quad \theta_k \geq 0, \ \sum_{k=1}^{n} \theta_k = 1.0.$$

The algorithm is derived by the method of Lagrange multipliers. Let us define the Lagrange function with the Lagrange multiplier Λ as

$$\mathcal{F}(\boldsymbol{\theta}, \Lambda) = \mathcal{L}(\boldsymbol{\theta}) - \Lambda\Big(\sum_{k=1}^{n} \theta_k - 1\Big).$$

Setting the partial derivative equal to zero yields

$$\frac{\partial \mathcal{F}(\boldsymbol{\theta})}{\partial \theta_k} = 0 \Leftrightarrow \Lambda = -\sum_{i=1}^{n} \frac{\Psi_k(\boldsymbol{x}_i^o, \boldsymbol{w}_i, \boldsymbol{v}_i)}{\boldsymbol{\theta}^T \boldsymbol{\Psi}_i(\boldsymbol{x}_i^o, \boldsymbol{w}_i, \boldsymbol{v}_i)}.$$

Multiplying both side of this equation by θ_k on and taking a summation w.r.t. k, $\Lambda = -n$. Then, the optimum solution satisfies

$$\theta_k = \frac{1}{n}\Big\{\sum_{i=1}^{n} \frac{\theta_k \Psi_k(\boldsymbol{x}_i^o, \boldsymbol{w}_i, \boldsymbol{v}_i)}{\boldsymbol{\theta}^T \boldsymbol{\Psi}_i(\boldsymbol{x}_i^o, \boldsymbol{w}_i, \boldsymbol{v}_i)}\Big\}. \tag{10}$$

The parameter estimation algorithm for EBCM iterates the update of $\boldsymbol{\theta}$ following Eq. (10). Note that when data is (non-censored) standard data, the update equation reduces to that of EBM. Therefore, EBCM is a natural generalization of EBM that can handle censored data. The proposed algorithm converges to the *global* optimum solution. Moreover, following the work by Lashkari and Golland Lashkari and Golland (2008), A clipping procedure can be

added to raise the convergence speed of this algorithm. This is done, in each iteration, by setting $\theta_k = 0$ if θ_k is smaller than some given threshold α and re-normalizing $\boldsymbol{\theta}$ to satisfy the sum to one constraint. In a later experiment, we set $\alpha = 10^{-3}/n$.

The *global* convergence of the algorithm is confirmed as follows. Let us define the *responsibility* term $\boldsymbol{\gamma}_i = \{\gamma_{ik}\}_{k=1}^n$ as

$$\gamma_{ik} = \frac{\theta_k \Psi_k(\boldsymbol{x}_i^o, \boldsymbol{w}_i, \boldsymbol{v}_i)}{\boldsymbol{\theta}^T \boldsymbol{\Psi}_i(\boldsymbol{x}_i^o, \boldsymbol{w}_i, \boldsymbol{v}_i)}.$$

This allows update equation Eq. (10) to be written as $\theta_k = \sum_{i=1}^n \gamma_{ik}/n$, which is exactly same update equation of the mixing ratio in the EM algorithm Bishop (2006). Therefore, the monotone decreasing characteristic of the objective function by this update can be shown in the same manner. In addition, our objective function, $\mathcal{L}(\theta)$, is a convex function unlike standard mixture models using EM. This can be confirmed by checking the Hessian of the objective function, $\mathcal{H} = \{\mathcal{H}_{kk'}\}$, which is defined as

$$\mathcal{H}_{kk'} = \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_{k'}} = \sum\nolimits_{i=1}^n \gamma_{ik}\gamma_{ik'}.$$

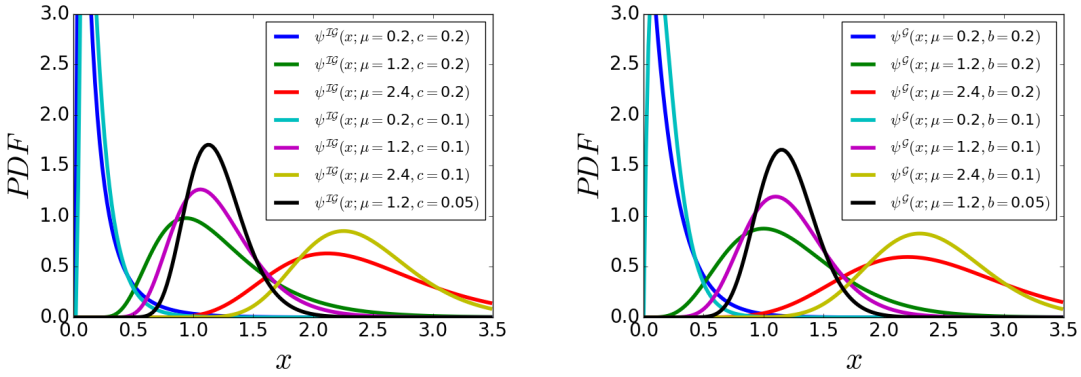
Since for arbitrary $\boldsymbol{\theta}$, $\boldsymbol{\theta}^T \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T \boldsymbol{\theta} = \|\boldsymbol{\theta}^T \boldsymbol{\gamma}_i\|^2 \geq 0$ holds and the sum of the positive definite matrices is also positive definite, $\mathcal{H}$ is a positive definite matrix. Therefore, the objective function is convex and the solution reached by this algorithm is the *global* optimum of the objective function.

4.4. Component Distributions for Survival Analysis

At the end of this section, we state our choice for the component distribution for survival analysis. The data values in survival analysis, which represent e.g., the time to failure or the usage time of services, are larger than or equal zero. Therefore, Gaussian distributions which support both negative and positive values may not be appropriate. We present examples using inverse Gaussian and Gamma distributions. Using a well-known distribution is of practical use since the PDF and CDF of the component distribution needs to be computed before running the algorithm (See Eq. (10)).

To use the inverse Gaussian and Gamma distributions, it is necessary to decide how to arrange the distribution with respect to the data points since the average and the mode do not match, unlike Gaussian distributions [4]. So we adopt the following approach: (i) re-parametrize the inverse Gaussian and Gamma distributions using mean parameter μ and set the component distributions so that the mean parameter matches the data point. (ii) introduce standard deviation like hyperparameters (b and c) so that the mode equals the mean when taking the hyperparameter limit to 0. This approach allows us to use inverse Gaussian and Gamma distributions as Gaussian-like distributions.

4. Studies on kernel density estimation (KDE) used asymmetric kernels such as Beta kernel Chen (1999), Gamma kenel Chen (2000), and inverse Gaussian kernel Scaillet (2004). Inverse Gaussian kernel and the estimator given (non-censored) data $\{x_i\}$ are defined as follows: $K_{\mathcal{IG}(x,1/c)}(\mu) = \mathcal{IG}(\mu|x, 1/c)$, $\hat{f}_{\mathcal{IG}}(x) = \frac{1}{n}\sum_{i=1}^n K_{\mathcal{IG}(x,1/c)}(x_i)$. Note that the x in the estimator $\hat{f}_{\mathcal{IG}}(x)$ corresponds to the 1st parameter of inverse Gaussian $\mathcal{IG}(\mu|x, 1/c)$. We do not to use these kernels as component distributions since their CDF is not easily computed; it requires the integral w.r.t. parameter.



(a) inverse Gaussian (b) Gamma

Figure 4: Candidates of component distributions using (a) inverse Gaussian and (b) Gamma distribution for survival analysis

Following this approach, we use the following inverse Gaussian based and Gamma based component distributions:

$$\psi_{\mathcal{IG}}(\boldsymbol{x};\boldsymbol{\mu},c)=\prod_{j=1}^{d_x}\mathcal{IG}(x_j|\mu_j,c\mu_j^2),\quad \psi_{\mathcal{G}}(\boldsymbol{x};\boldsymbol{\mu},b)=\prod_{j=1}^{d_x}\mathcal{G}(x_j|b\mu_j,b).$$

The shapes and statistics are shown in Fig. 4(a), 4(b) and Table 1. From Table 1, we can confirm that the modes converge to μ by taking the hyperparameter limit to 0:

$$\lim_{c\to 0}(\mu^2+9c^2/4)^{\frac{1}{2}}-3c/2=\mu,\quad \lim_{b\to 0}\mu-b=\mu.$$

Following Eq. (4), EBCM with inverse Gaussian/Gamma distributions can be constructed as follows:

$$\text{EBCM (inverseGaussian)}:\quad f(\boldsymbol{x}|\boldsymbol{\theta})=\sum_{k=1}^{n}\theta_k\psi_{\mathcal{IG}}(\boldsymbol{x}|\boldsymbol{z}_k,c)$$
$$\text{EBCM (Gamma)}:\quad f(\boldsymbol{x}|\boldsymbol{\theta})=\sum_{k=1}^{n}\theta_k\psi_{\mathcal{G}}(\boldsymbol{x}|\boldsymbol{z}_k,b)$$

For parameter estimation, Ψ_{ik} is required before running the algorithm. The term Ψ_{ik} for inverse Gaussian/Gamma can be computed as follows:

$$\Psi_{ik}^{\mathcal{IG}}=\prod_{\{j|w_{ij}=0\}}\mathcal{IG}(x_j|z_{kj},cz_{kj}^2)\prod_{\{\ell|w_{i\ell}=1\}}\int_{v_\ell}^{\infty}\mathcal{IG}(x_\ell|z_{k\ell},cz_{k\ell}^2)dx_\ell$$
$$\Psi_{ik}^{\mathcal{G}}=\prod_{\{j|w_{ij}=0\}}\mathcal{G}(x_j|bz_{kj},b)\prod_{\{\ell|w_{i\ell}=1\}}\int_{v_\ell}^{\infty}\mathcal{G}(x_\ell|bz_{k\ell},b)dx_\ell$$

Note that, although we show the example of component distribution using Gamma and inverse Gaussian, EBCM can use any type of distribution and parameterization.

Table 1: Examples of component distributions and their statistics.

distribution	expectation	variance	mode
$\mathcal{N}(x\|\mu,\sigma^2)$	μ	σ^2	μ
$\mathcal{IG}(x\|\mu,\lambda)$	μ	μ^3/λ	$\mu[(1+9\mu^2/4\lambda^2)^{\frac{1}{2}}-3\mu/2\lambda]$
$\mathcal{G}(x\|a,b)$	ab	ab^2	$(a-1)b \quad (\text{if } a \geq 1)$
$\psi^{\mathcal{IG}}(x;\mu,c)=\mathcal{IG}(x\|\mu,\mu^2/c)$	μ	μc	$(\mu^2+9c^2/4)^{\frac{1}{2}}-3c/2$
$\psi^{\mathcal{G}}(x;\mu,b)=\mathcal{G}(x\|\mu/b,b)$	μ	μb	$\mu-b \quad (\text{if } \mu \geq b)$

5. Experiments

5.1. Setting

This section confirms the effectiveness of EBCM.

Synthetic Data: We prepared synthetic data set (synth), that follows a *true* gaussian mixture distribution. We set the true number of components to $K^*=2$ and the component parameters to $\boldsymbol{\mu}_1=(-2,-2)$, $\boldsymbol{\mu}_2=(2,-2)$, $\boldsymbol{\mu}_3=(-2,2)$, $\boldsymbol{\mu}_4=(2,2)$, The standard deviation was set to $\sigma^*=1.0$. We randomly generated 5 pairs of training and test data using this true distribution. The threshold used in censoring was set to $\boldsymbol{v}_i=(1.5,1.5)$ for all i. Note that test data for evaluation was generated without censoring. The training data and test data consisted of 100 and 1000 items, respectively. Generated data are shown in Fig. 5(a) and 5(b).

Censored Benchmark Data: We used Old Faithful Geyser Data (geyser) [5] and Fisher's iris data (iris) [6]. geyser represents the waiting time and duration until the geyser spout in Yellowstone National Park, USA. iris is taken from Fisher's article. These two are not censored data and so we covert them by setting 3/4 quantile point of each dimension as the threshold value $\boldsymbol{v}_i$ for all i. The y axis value of geyser has been multiplied by 0.1 to match the scale of the x and y axes. Original and censored data are shown in Fig. 6(a) and 6(b). We prepared 5 datasets by dividing the data into five, using 80% of the data for training and the remaining 20% for testing.

Real Survival Data: We also used two publicity available survival data sets, Crowder's repeated response times data (rrt) and paired response times data (prt), both of which are presented in Table 6.1 and 6.2 in Crowder (2012). These data are collected to investigate the effect on the brain of lead absorption by young children living in a traffic-clogged city. rrt contains the response time of rats (in second), (t_1,t_2,t_3,t_4), when they are exposed to a harmless sensory stimulus at 0, 15, 30, and 60 minutes after administration of a drug. The values are censored at 10 seconds. prt also contain response time of rats, (t_1,t_2), before and after administration of the same drug. The values are censored at 250 seconds. The value of prt has been multiplied by 0.01 to match the scale. We prepared five data sets by dividing the data into five, using 80% of the data for training and the remaining 20% for testing.

Baseline: We compare EBCM with two existing methods, EBM and KDE. EBM Lashkari and Golland (2008) is the exemplar-based model, which is the basis of the proposed method.

5. https://www.stat.cmu.edu/~larry/all-of-statistics/
6. https://archive.ics.uci.edu/ml/datasets/Iris

Table 2: Results of EBCM (proposed method) and baseline. Average and standard deviation of negative log-likelihood are shown. Smaller values are better.

	EBCM	EBM 1	EBM 2	KDE 1	KDE 2
synth	**4.22** $\pm$ 0.05	5.51 $\pm$ 0.44	4.33 $\pm$ 0.01	5.15 $\pm$ 0.25	4.35 $\pm$ 0.02
geyser	**2.13** $\pm$ 0.08	3.08 $\pm$ 0.23	2.32 $\pm$ 0.16	2.55 $\pm$ 0.17	2.18 $\pm$ 0.15
iris	**2.01** $\pm$ 0.18	3.12 $\pm$ 0.56	2.09 $\pm$ 0.35	3.01 $\pm$ 0.50	2.16 $\pm$ 0.30
rrt	**6.84** $\pm$ 1.30	8.36 $\pm$ 1.86	6.90 $\pm$ 1.30	8.30 $\pm$ 1.89	**6.84** $\pm$ 1.34
prt	**0.40** $\pm$ 0.10	0.57 $\pm$ 0.18	0.43 $\pm$ 0.09	0.52 $\pm$ 0.15	0.46 $\pm$ 0.09

KDE is kernel density estimation Bishop (2006). We selected these two because they don't have component parameters and thus are candidate alternatives to the EM-based iterative algorithm Chauveau (1995). Since EBM and KDE cannot handle censored data, we applied two heuristics: (i) EBM-1 and KDE-1 which only use the data whose all elements are observed, i.e., $\{\boldsymbol{x}_i | \boldsymbol{w}_i = \boldsymbol{0}\}$ and (i) EBM-2 and KDE-2 which treat the censored values as observed values. We use Gaussian component in common for synthetic and censored benchmark data. Hyperparameter σ was set to $0.1, 0.25$ and 0.25 for synth, geyser and iris based on a preliminary experiment.

Evaluation Metric: To evaluate predictive performance, we use the negative log likelihood metric. Since we knew the *true* value (the value which would be observed if the threshold $\boldsymbol{v}$ was infinitely large) of the censored data in both the synthetic data and censored benchmark data sets, we use it as test data $\{y_i^{test}\}_{i=1}^{n_{test}}$. The negative log likelihood for test data is computed as follows:

$$\mathcal{L}_{test} = -\frac{1}{n_{test}} \sum\nolimits_{\ell=1}^{n_{test}} \log\{ f(y_\ell^{test} | \theta) \},$$

For real survival data, since test data are also censored, we use the following negative log-likelihood for the censored data as the performance metric:

$$\mathcal{L}_{c-test} = -\frac{1}{n_{test}} \sum\nolimits_{\ell=1}^{n_{test}} \log\Big\{ \int_{v_\ell^c} f(x_i^{test} | \theta) d\boldsymbol{x}_\ell^c \Big\},$$

where $\{x_\ell^{test}, w_\ell^{test}\}_{\ell=1}^{n_{test}}$ is the censored test data.

5.2. Result

Quantitative Evaluation: The results of a quantitative evaluation of the experiment are shown in Table 2. It can be seen that EBCM offer better performance than the other methods examined. It is considered that this is because the proposed method is designed to use an objective function that considers the censored data. Table 3 also shows the results of EBCM on real survival data for various component distributions and hyperparameters. Although the use of Gamma component achieves the best performance, the other distributions also able to achieve performance close to the best value. This validates the effectiveness of EBCM regardless of the choice of the component distribution.

Table 3: Results of EBCM under various component distributions and hyperparameters. Average and standard deviation of negative log-likelihood are shown. Smaller values are better.

	Gaussian component			Gamma component		InvGauss component	
	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1.0$	$b = 0.05$	$b = 0.1$	$c = 0.05$	$c = 0.1$
rrt	96.99±46.9	8.10±2.17	6.90±1.30	8.15±2.30	**6.84**±1.30	8.21±2.44	6.88±1.37
prt	0.42±0.14	1.29±0.10	2.17±0.04	**0.40**±0.10	0.74±0.08	0.44±0.11	0.85±0.08

Qualitative Evaluation: The estimated probability densities in the synthetic data experiment are shown in Fig. 5. Estimated results of the proposed method (Fig. 5(c)) accurately capture the peaks (mode) of the true probability distribution. On the other hand, in the estimation result of EBM-1 (Fig. 5(d)), only the position of the lower left peak in the figure is captured because only the observed data is used. Also, for EBM-2 (Fig 5(e)), although a structure with four peaks is captured, the positions of the upper left, lower right, and upper right peaks are closer to the origin than their true positions since the threshold is used as the observed value. By using censored data appropriately, the proposed method can accurately estimate the true distribution which contributes to the performance improvement. Figure 6 also shows that EBCM estimates the upper right component more precisely than the baseline methods.

6. Conclusion

In this study, we proposed EBCM, a generalized variant of the example-based model that can analyze censored data. The proposed method can guarantee convergence to a global optimum solution and can estimate parameters without the need to perform repeated numerical computation of the statistics of truncated distribution. The effectiveness of the proposed method was confirmed by experiments using synthetic data and real survival data. Future directions of this study include construction of an online algorithm Cappé and Moulines (2009).

References

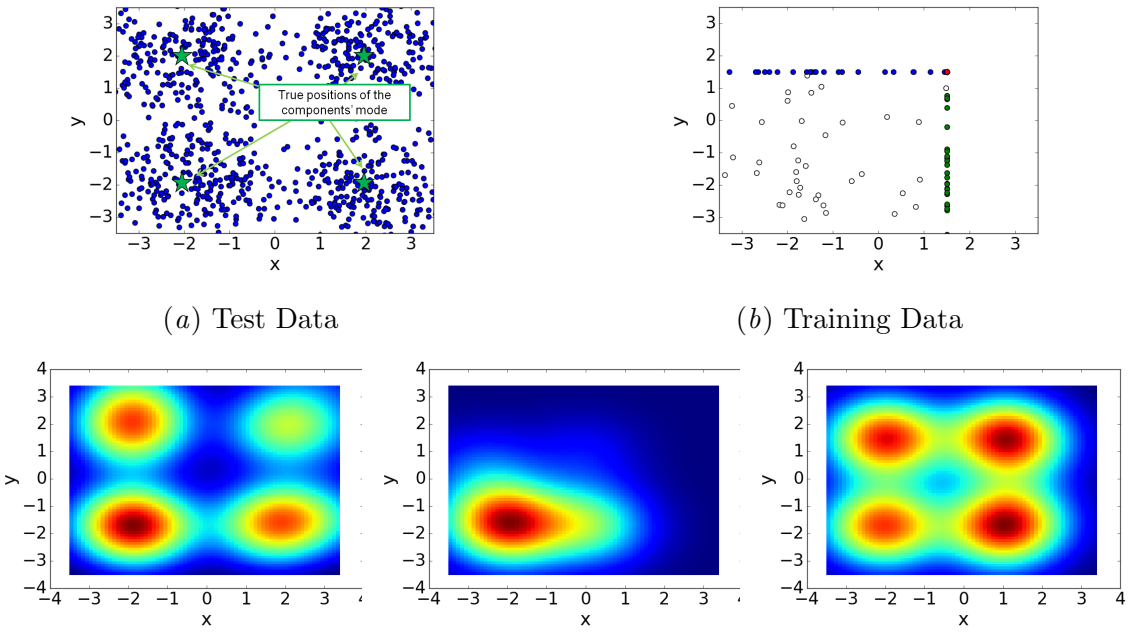
Hagai Attias. Inferring parameters and structure of latent variable models by variational bayes. In *UAI*, pages 21–30, 1999.

Nicola Barbieri, Fabrizio Silvestri, and Mounia Lalmas. Improving post-click user engagement on native ads via survival analysis. In *WWW*, pages 761–770, 2016.

Manjunath BG and Stefan Wilhelm. Moments calculation for the double truncated multivariate normal density. 2009. URL https://papers.ssrn.com/abstract_id=1472153.

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

Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B*, 71(3):593–613, 2009.



(*a*) Test Data (*b*) Training Data

(*c*) EBCM (*d*) EBM-1 (*e*) EBM-2

Figure 5: (a) Test and (b) training data example from synthetic data experiments (synth). The white points are data where all values are observed and the green/blue/red points indicate the data points whose x-axis/y-axis/x&y-axis values were not observed due to censoring. The estimated probability densities yielded by (c) proposed method and (d)(e) baseline methods using EBM are shown.

Didier Chauveau. A stochastic em algorithm for mixtures with censored data. *Journal of statistical planning and inference*, 46(1):1–25, 1995.

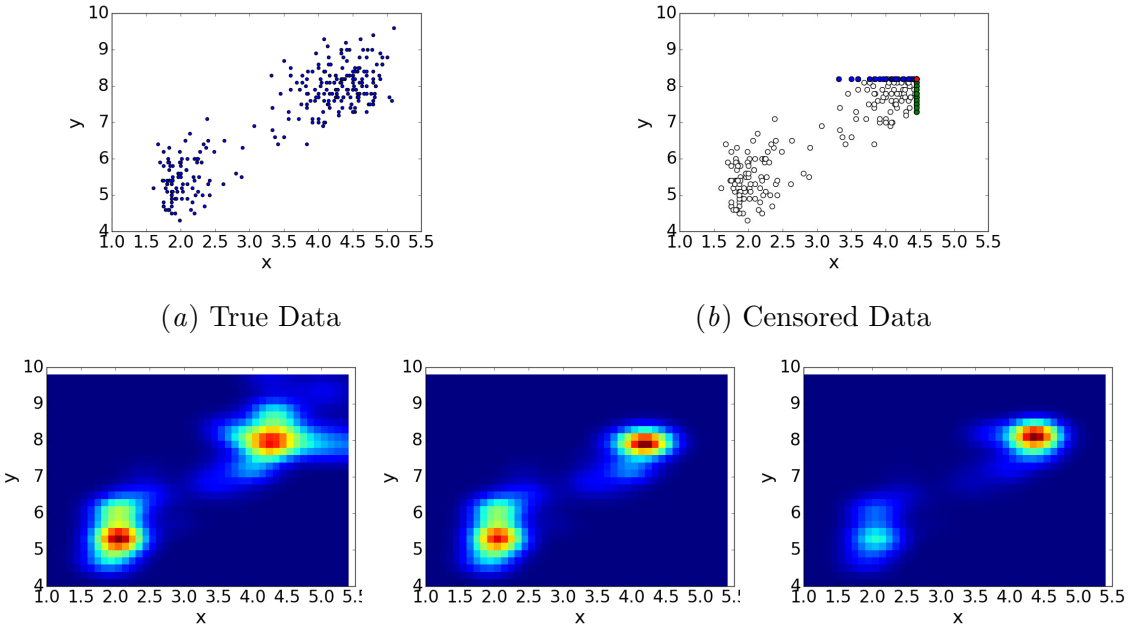
Song Xi Chen. Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2):131–145, 1999.

Song Xi Chen. Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52(3):471–480, 2000.

Si-Chi Chin and W Nick Street. Survival analysis of click logs. In *SIGIR*, pages 1149–1150, 2012.

Martin J Crowder. *Multivariate survival analysis and competing risks*. Chapman and Hall/CRC, 2012.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B*, pages 1–38, 1977.



(*a*) True Data (*b*) Censored Data

(*c*) EBCM (*d*) EBM 1 (*e*) EBM 2

Figure 6: Data (a) before and (b) after censoring in censored benchmark data experiments (geyser). The white points are data where all values are observed and the green/blue/red points indicate the data points whose x-axis/y-axis/x&y-axis values were not observed due to censoring. The estimated probability densities are shown from (c) proposed method and (d)(e) baseline methods using EBM.

Tamara Fernández, Nicolás Rivera, and Yee Whye Teh. Gaussian processes for survival analysis. In *NIPS*, pages 5021–5029, 2016.

Alan Genz. Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, 14(3):251–260, 2004.

Alan Genz and Frank Bretz. Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63(4):103–117, 1999.

Georg L Grob, Ângelo Cardoso, CH Bryan Liu, Duncan A Little, and Benjamin Paul Chamberlain. A recurrent neural network survival model: Predicting web user return time. In *ECMLPKDD*, pages 152–168, 2018.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Farkhondeh Kiaee, Hamid Sheikhzadeh, and Samaneh Eftekhari Mahabadi. Relevance vector machine for survival analysis. *IEEE transactions on neural networks and learning systems*, 27(3):648–660, 2016.

David G Kleinbaum and Mitchel Klein. *Survival analysis*, volume 3. Springer, 2010.

Masahiro Kohjima, Tatsushi Matsubayashi, and Hiroyuki Toda. Variational bayes for mixture models with censored data. In *ECMLPKDD*, pages 605–620, 2018.

Danial Lashkari and Polina Golland. Convex clustering with exemplar-based models. In *NIPS*, pages 825–832, 2008.

Sebastian Pölsterl, Nassir Navab, and Amin Katouzian. Fast training of support vector machines for survival analysis. In *ECMLPKDD*, pages 243–259, 2015.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pages 101–114, 2016.

Saharon Rosset, Einat Neumann, Uri Eick, Nurit Vatnik, and Yizhak Idan. Customer lifetime value modeling and its use for customer retention planning. In *KDD*, pages 332–340, 2002.

Olivier Scaillet. Density estimation using inverse and reciprocal inverse gaussian kernels. *Nonparametric statistics*, 16(1-2):217–226, 2004.

Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Conditional density estimation via least-squares density ratio estimation. In *AISTATS*, pages 781–788, 2010.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Georges M Tallis. The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society. Series B*, pages 223–229, 1961.

Stefan Wilhelm and BG Manjunath. tmvtnorm: A package for the truncated multivariate normal distribution. *sigma*, 2(2), 2010.