EXPECTATION-MAXIMIZATION FOR UNIVARIATE AND BIVARIATE REINFORCED URN PROCESSES UNDER LEFT-TRUNCATION AND RIGHT-CENSORING

A PREPRINT

Luis Souto Arias*
Centrum Wiskunde & Informatica
The Netherlands
luis.souto.arias@cwi.nl

Pasquale Cirillo
M Open Forecasting Center and Institute For the Future
University of Nicosia
Cyprus
cirillo.p@unic.ac.cy

Cornelis W. Oosterlee
Centrum Wiskunde & Informatica
The Netherlands
c.w.oosterlee@cwi.nl

October 20, 2020

ABSTRACT

Reinforced Urn Processes (RUPs) represent a flexible class of Bayesian nonparametric models suitable for dealing with possibly right-censored and left-truncated observations. A reliable estimation of their hyper-parameters is however missing in the literature. We therefore propose an extension of the Expectation-Maximization (EM) algorithm for RUPs, both in the univariate and the bivariate case. Furthermore, a new methodology combining EM and the prior elicitation mechanism of RUPs is developed: the Expectation-Reinforcement algorithm. Numerical results showing the performance of both algorithms are presented for several analytical examples as well as for a large data set of Canadian annuities.

***K*eywords** Reinforced urn process · Expectation-Maximization · Bivariate Survival Function · Censoring

1 Introduction

Reinforced Urn Processes (RUPs) represent an important class of Bayesian nonparametric models [Walker and Muliere, 1997, Muliere, Secchi, and Walker, 2000], with many applications in biostatistics, engineering, finance and other fields of use of survival analysis. (e.g. see Amerio, Muliere, and Secchi [2004], Cheng and Cirillo [2018, 2019], Cirillo, Hüsler, and Muliere [2010], Cirillo, Hüsler, and Muliere [2013], Giudici, Muliere, and Mezzetti [2003], Muliere, Secchi, and Walker [2003], Peluso, Mira, and Muliere [2015]). As the name suggests, the building blocks of RUPs are urns, Polya-like urns in particular [Mahmoud, 2008].

As shown in Muliere et al. [2000, 2003], a RUP is able to generate neutral-to-the-right [Doksum, 1974] random distributions, like for example the beta-Stacy process of Walker and Muliere [1997]. This feature makes RUPs an interesting tool to deal with right-censored observations.

A construction based on RUPs has been introduced in Bulla, Muliere, and Walker [2007] to deal with right-censored bivariate phenomena, in which the dependence among the two components of the system is linear. An application for the modeling of wrong way risk in the credit risk framework has been recently proposed in Cheng and Cirillo [2019].

*Corresponding author.

A useful characteristic of reinforced urn processes is the possibility of combining expert judgements, in the form of some a priori, and empirical data. Thanks to reinforcement, a RUP is able to actually learn from data, in a way that may recall machine learning approaches [Cheng and Cirillo, 2018, 2019].

Despite their strong potential, the study and use of RUPs has been restricted to the Bayesian community until now, and all applications have relied on more or less extensive simulations (e.g. Markov Chain Monte Carlo as in Bulla et al. [2007], Peluso et al. [2015]). This is due to–in our opinion–the lack of reliable estimation tools for the parameters and the hyper-parameters of the different models[2].

While for a Bayesian statistician the definition of a proper a priori is a natural and fundamental step of the analysis, and in case of real ignorance one can always rely on "non-informative" solutions [de Finetti, 2017, Galavotti, 2001, Galavotti et al., 2008, Jeffreys, 1946], for researchers who do not feel at ease with extreme subjectivity and prior elicitation, the lack of a way of estimating the parameters directly from the data can be a deterrent.

Therefore, in this paper, we address the problem by proposing a way of estimating RUPs parameters and hyper-parameters from data, using the Expectation Maximization (EM) algorithm [Dempster, Laird, and Rubin, 1977], already applied to simpler urn models and the related processes (e.g. Bouguila and Ziou [2006], Figueiredo and Jain [2000]), but also proposing a novel Expectation-Reinforcement (ER) algorithm, which exploits the natural reinforcement mechanism of RUPs to enhance the EM part. We deal with this relevant estimation problem not only for the most common univariate case, but also for bivariate constructions à la Bulla et al. [2007].

In doing that, we also introduce some interesting extensions of the reinforced urn processes themselves, for example by showing that they are conjugate not only under right-censoring [Muliere et al., 2000], but also under left-truncation, another relevant characteristic of many actual data sets.

The structure of the paper is as follows. In Section 2 we revisit the concepts of censoring and truncation, and how they relate to our variables of interest in both the univariate and bivariate cases. Section 3 introduces briefly the theory of Reinforced Urn Processes, as well the bivariate RUP (B-RUP) of Bulla et al. [2007]. The Expectation-Maximization algorithm for RUPs under left-truncated and right-censored data is described in Section 4, while Section 5 discusses the new Expectation-Reinforcement algorithm. Some simulation results and an application on real data are presented in Section 6, and Section 7 concludes the paper.

2 Incomplete data: truncation and censoring

In many applications, data is only available in an incomplete form due to the nature of the experiment being conducted. For example, in medical studies where patients are observed over a limited period of time, the event of interest—next stage of a disease or even death—may not occur during the observation period. In this case we say that observations are right-censored. Furthermore, since in many cases patients do not start being observed from the time they are born, this data will also be left-truncated in case the missing time window is of interest, like in survival studies and risk management [Shen and Yan, 2008, Antonio et al., 2015].

Although truncation and censoring have both been extensively studied when there is only one target variable–see Wang [1987] for a very nice separate and joint analysis of these phenomena on likelihood estimation–much less progress has been made in higher dimensions, even just for the bivariate case.

The bivariate case is indeed of special relevance for the many applications it has, in which at least one variable is subject to either truncation or censoring (see Liuquan and Haobao [2001]). Among the first proposals for a bivariate nonparametric estimator under bivariate censoring we find Campbell and Földes [1982], Dabrowska [1988], Pruitt [1991a] and Pruitt [1993]. However, as shown in Pruitt [1991b], these estimators fail to be monotone in specific cases, generating negative probabilities.

Interesting works are those of Shen and Yan [2008] and Gribkova and Lopez [2015]. The former develops an iterative method to estimate a generalization of the Dabrowska and Campbell and Földes estimators that includes the effect of left-truncation, while the latter uses a nonparametric estimator via random weights, first defined in Lopez [2012], to compute a nonparametric copula.

As shown in Shen and Yan [2008], bivariate truncation complicates things further. In such a case, the famous univariate Kaplan-Meier (KM) estimator [Kaplan and Meier, 1958] should not be used for the marginal survival functions, since it would not be consistent. Thus not only the joint distribution, but also the marginals become difficult to estimate. Given

[2]It should be stressed that the lack of estimation tools for urn models is a more general problem. Urn processes have been mainly approached from a probabilistic point of view, while statistical inference has always been marginal. Some important exceptions exist, e.g. Feng et al. [2017], Line and Philippe [2017] or Marcaccioli and Liva [2019], but they are limited.

that left-truncation and right-censoring fall under the umbrella of incomplete data, it is no wonder that many authors have opted for a different approach using the EM algorithm of Dempster et al. [1977] to tackle the problem. Some relevant results in the univariate framework are available in Pruitt [1991a], van der Laan [1994], Antonio et al. [2015] and the many references therein. However, the literature is far more scarce for two-dimensional distributions: Nandi and Dewan [2010] is the first work to approach the problem of bivariate modelling under right-censoring through EM. Furthermore, to the best of our knowledge, nobody has proposed so far an extension of the EM algorithm to include bivariate left-truncation.

2.1 Right-censoring

Let $\mathbf{X} = (x_1, ..., x_n)$ and $\mathbf{C} = (c_1, ..., c_n)$ be i.i.d. (identically and independently distributed) observations with independent distributions F_X and F_C, respectively.

If right-censoring occurs, we observe the pair (x_i^*, δ_i), where $x_i^* = \min(x_i, c_i)$ and $\delta_i = \mathbb{1}_{\{x_i^* = x_i\}}$ for $i = 1, ..., n$, with $\mathbb{1}_{\{\cdot\}}$ the indicator function. That is, we observe the minimum of the censoring variable and our target variable, plus and indicator telling us which of the two we observe. Note that, although the underlying distributions F_X and F_C are in principle independent, our observations of X and C may have some dependence due to the sampling mechanism. This problem was originally addressed in Kaplan and Meier [1958], a seminal work introducing the well-known nonparametric estimator.

The KM estimator maximizes the conditional likelihood of the observations of X, given the observations of C:

$$L(\mathbf{X}^*|\boldsymbol{\delta}) = \sum_{i=1}^{n} [(1 - \delta_i) \log(\mathbb{P}(X > x_i^*)) + \delta_i \log(\mathbb{P}(X = x_i^*))]. \tag{1}$$

Simply notice that this general setting of random censoring naturally includes some simpler situations with deterministic censoring, sometimes present in clinical trials [Rosenberger and Lachin, 2004].

2.2 Left-truncation

Let again $\mathbf{X} = (x_1, ..., x_n)$ and $\mathbf{T} = (t_1, ..., t_n)$ be i.i.d. samples with independent distributions F_X and F_T, respectively. When left-truncation occurs, we observe the pair (x_i, t_i), for $i = 1, ..., n$, if $x_i \geq t_i$, and nothing otherwise.

Since we do not observe anything in the second case, we cannot even realize its existence, that is, there is no information in our data about x or t for $t > x$. This suggests that a truncated observation provides even less information than a censored one, since for cases where $\mathbb{P}(T \leq X)$ is small, our truncated sample will be highly biased with respect to the original underlying distributions. Such a bias is so relevant that, according to Wang [1987], truncated data can also be classified as selection-biased data.

The first nonparametric estimator of the survival probability function that maximizes the likelihood of X conditioned on the observations of T–known as the product-limit (PL) estimator–was developed in Lynden-Bell [1971], and further studied in Wang, Jewell, and Tsai [1986], Woodroofe [1985], among others.

The conditional likelihood under left-truncation is given by

$$L(\mathbf{X}|\mathbf{T}) = \sum_{i=1}^{n} [\log(\mathbb{P}(X = x_i)) - \log(\mathbb{P}(X \geq t_i))]. \tag{2}$$

Note that, at most, we can infer $S_X^*(\cdot) = S_X(\cdot)/S_X(T^*)$ from the data, where S_X is the survival probability function of X, and T^* is the minimum value observed in $\mathbf{T}$.

As per right-censoring, also left-truncation can be deterministic: think of a survival study in which all individuals have the same age (x_0) when joining. In such a situation, the truncation variable would be deterministic, with a degenerate distribution at x_0. Furthermore, no information could be derived about what happens for $X < x_0$, so that $T^* = x_0$.

2.3 Left-truncation and right-censoring together

Of course, it can also be the case that both truncation and censoring occur at the same time, complicating things further. In such a case it is usually assumed [Cox and Oakes, 1984, Wang, 1987] that the variable of interest X is independent from both T and C, the truncation and censoring variables respectively. In a similar situation what one observes is

the triplet (x^*, t, δ), with x^* and δ defined as before if $T \leq X$, and nothing otherwise. As in Wang [1987], we further assume that $\mathbb{P}(T \leq C) = 1$, hinting towards the existence of dependence between[3] T and C.

A nonparametric estimator for S_X was first introduced in Cox and Oakes [1984] and further studied in Tsai, Jewell, and Wang [1987], as a generalization of the previous estimators, and it reduces to the KM estimator in the absence of truncation, and to the product-limit estimator of Lynden-Bell [1971] without censoring.

Clearly, the conditional likelihood now takes a more general expression including both Equations (1) and (2) as special cases:

$$L(\mathbf{X}^*|\boldsymbol{\delta}, \mathbf{T}) = \sum_{i=1}^{n} [(1-\delta_i)\log(\mathbb{P}(X > x_i^*)) + \delta_i \log(\mathbb{P}(X = x_i^*)) - \log(\mathbb{P}(X \geq t_i))]. \tag{3}$$

2.4 The bivariate case

So far, we have seen the effect of truncation and censoring for one variable of interest. Things get obviously more complicated when we consider a bivariate situation. Following Shen and Yan [2008], we assume that the joint survival function of the target variables, S_{XY}, is independent from the truncation and censoring variables (T^X, C^X, T^Y, C^Y).

The observations consist of two triplets as per Subsection 2.3, with (x^*, t^X, δ) the triplet for variable X, and (y^*, t^Y, ϵ) the triplet for variable Y. If the target variables X and Y are independent, we can write the joint conditional likelihood as the product of the marginal likelihoods, both defined as in Equation (3). However, when there exists a dependence, the likelihood must be modelled jointly, and under left-truncation and right-censoring it takes the form:

$$L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\delta}, \boldsymbol{\epsilon}, \mathbf{T}^X, \mathbf{T}^Y) = \sum_{i=1}^{n} [\log(\mathbb{P}^*(x_i^*, y_i^*|\delta_i, \epsilon_i)) - \log(\mathbb{P}(X \geq t_i^X, Y \geq t_i^Y))], \tag{4}$$

with

$$\mathbb{P}^*(x, y|\delta, \epsilon) = \begin{cases} \mathbb{P}(X = x, Y = y) & \text{if } \delta = 1 \text{ and } \epsilon = 1, \\ \mathbb{P}(X > x, Y = y) & \text{if } \delta = 0 \text{ and } \epsilon = 1, \\ \mathbb{P}(X = x, Y > y) & \text{if } \delta = 1 \text{ and } \epsilon = 0, \\ \mathbb{P}(X > x, Y > y) & \text{if } \delta = 0 \text{ and } \epsilon = 0. \end{cases} \tag{5}$$

For ease of notation, from now on we will write $L(\mathbf{X}, \mathbf{Y})$ when referring to the bivariate likelihood in Equation (4), with the conditioning on censoring and truncation always implied.

In the rest of the paper we will focus on a particular case of bivariate left-truncation and right-censoring, general enough to be useful in many applications, but more manageable from a notational and computational point of view.

First, we assume that T^X and T^Y are dependent through the relation $T^Y = T^X + \epsilon - \epsilon_0$, where $\epsilon_0 \geq 0$, and ϵ is a random variable (r.v.) that only takes values on the positive integers[4]. A similar situation arises for example when studying coupled lifetimes [Frees, Carriere, and Valdez, 1996], where T^X and T^Y denote the age at which each individual enters the study. Clearly, in this case the difference between T^X and T^Y is given by the difference in ages between the individuals, which would be modelled by the r.v. $\epsilon - \epsilon_0$, being ϵ_0 a maximum threshold in the age difference.

Second, we hypothesize that $C^i = T^i + \Delta$, for $i = X, Y$. Going back to the coupled lifetimes example, both members of the couple will be monitored exactly for the same amount of time in the case censoring occurs, which explains using the same r.v. for C^X and C^Y. This new r.v. Δ serves the purpose of modelling the observation period, so that the age at which the individuals start the study plus the time under observation trivially gives the age at which they leave the study.

Although these are the assumptions that will be followed throughout the rest of this work, we would like to emphasize that the methodology we propose in Section 4 is of more general nature and it can also be adapted and used under different assumptions.

[3] A similar situation verifies itself, for example, in lifetime follow-up studies, where T is the age at which an individual joins the study and C is the age at which they drop from it. Since it is not possible to drop from a study without first joining it, the condition $\mathbb{P}(T \leq C) = 1$ is trivially met.

[4] It would be more general to define $T^Y = T^X + \epsilon$, allowing ϵ to take negative values, but since most of the times T has a temporal interpretation, and also for simplicity, the other form is here preferred.

3 Univariate and Bivariate Reinforced Urn Processes

The discrete Reinforced Urn Process (RUP) was first described in Muliere et al. [2000], as a generalization of the Generalized Polya Sequence of Walker and Muliere [1997], in which the results of Blackwell and MacQueen [1973] and Ferguson [1973] are extended to right-censored data. In Muliere et al. [2003] a continuous-time RUP is proposed.

Always in Walker and Muliere [1997], the famous beta-Stacy process, a neutral-to-the-right [Doksum, 1974] generalization of the Dirichlet process [Ferguson, 1973] is introduced in its discrete and continuous versions. The beta-Stacy process is a random distribution that, depending on its definition, can sample discrete or continuous distributions, therefore it represents an extremely flexible tool for Bayesian nonparametrics [Hjort et al., 2010]

Definition 3.1 (Walker and Muliere [1997])**.** A random distribution function F is a discrete beta-Stacy process with jumps at $j \in \mathbb{N}_0$ and parameters $\{\beta_j, \omega_j\}_{j\in\mathbb{R}^+}$, if there exist mutually independent random variables $\{V_j\}_{j\in\mathbb{N}_0}$, each beta distributed with parameters (β_j, ω_j), such that the random mass assigned by F to $\{j\}$, written $F(\{j\})$, is given by $V_j \prod_{i<j}(1 - V_i)$.

Following Walker and Muliere [1997], we introduce couples $\{\beta_j, \omega_j\} \in \mathbb{R}^+ \times \mathbb{R}^+$, with $j \in \mathbb{N}_0$, such that $\beta_j, \omega_j \geq 0$, $\beta_j + \omega_j > 0$, and $\lim_{n\to\infty} \prod_{j=0}^{n} \frac{\omega_j}{\beta_j+\omega_j} = 0$. Then, given a beta-Stacy process F with parameters $\{\beta_j, \omega_j, j\}$, and a right-censored sample $(\boldsymbol{X}_n^*, \boldsymbol{\delta}_n)$, with $\boldsymbol{X}_n^* = \{x_n^*, n \geq 1\}$, the sequence $\boldsymbol{X}_n = \{x_n, n \geq 1\}$ is a RUP if

$$\hat{S}(x) = \mathbb{P}(X_{n+1} > x | \boldsymbol{X}_n^*, \boldsymbol{\delta}_n) = \prod_{j=0}^{x} \left[1 - \frac{\beta_j + m_j^*(\boldsymbol{X}_n^*, \boldsymbol{\delta}_n)}{\beta_j + \omega_j + s_j(\boldsymbol{X}_n^*)}\right], \tag{6}$$

where $m_j^*(\boldsymbol{x}_n, \boldsymbol{d}_n) = \sum_{i=1}^{n} \mathbb{1}_{\{x_i=j, d_i=1\}}$ is the number of exact observations at $x = j$, and $s_j(\boldsymbol{x}_n) = \sum_{i=1}^{n} \mathbb{1}_{\{j\leq x_i\}}$ is the number of censored observations at j.

By defining $\beta_j^* = \beta_j + m_j^*(\boldsymbol{x}_n, \boldsymbol{d}_n)$ and $\omega_j^* = \omega_j + s_j(\boldsymbol{x}_n, \boldsymbol{t}_n) - m_j^*(\boldsymbol{x}_n, \boldsymbol{d}_n)$, we obtain a new beta-Stacy process F^* with parameters $\{\beta_j^*, \omega_j^*, j\}$, which means that the beta-Stacy process is conjugate to right-censored data[5] [Muliere et al., 2000, Walker and Muliere, 1997].

3.1 RUPs and left-truncation

Before we expand on the properties of the RUP, we will prove that this process is not only conjugate to right-censoring but also to left-truncation. As mentioned before, this process is a generalization of the Dirichlet process, and its underlying distribution is actually the generalized Dirichlet distribution of Connor and Mosimann [1969]:

$$\mathbb{P}(p_X(1), p_X(2), \cdots, p_X(M) | \boldsymbol{\beta}, \boldsymbol{\omega}) \propto (p_X(M))^{\omega_{M-1}-1} \prod_{i=0}^{M-1} \left[(p_X(i))^{\beta_i-1} (S_X(i-1))^{\omega_{i-1}-(\beta_i+\omega_i)}\right], \tag{7}$$

with[6] $p_X(i) = \mathbb{P}(X = i)$, $S_X(i) = \mathbb{P}(X > i)$, (β_i, ω_i) defined as before, and where we assume that there are no more observations after time M.

Notice that now we can rewrite the log-likelihood in Equation (3) as

$$L(p_X(1), p_X(2), ..., p_X(M) | \boldsymbol{X}_n) \propto \prod_{i=0}^{M} (p_X(i))^{n_i} \prod_{i=0}^{M-1} (S_X(i))^{r_i - l_{i+1}}, \tag{8}$$

with n_k the number of uncensored observations at k, r_k the number of censored observations at k, and l_k the number of samples truncated at k.

Given that the posterior distribution is proportional to the product of Equations (7) and (8), i.e. the product of the prior distribution and the likelihood, it follows immediately that, by setting $q_k = \sum_{i=k+1}^{M} n_i + \sum_{i=k}^{M} r_i - \sum_{i=1}^{k} l_i$, and if F is a discrete beta-Stacy process with parameters $\{\beta_k, \omega_k\}$, the posterior distribution of F given the left-truncated and right-censored (LTRC) data $(\boldsymbol{X}_n^*, \boldsymbol{\delta}_n, \boldsymbol{T}_n)$ is also a generalized Dirichlet distribution with parameters

$$\beta_k^* = \beta_k + n_k, \qquad \omega_k^* = \omega_k + q_k, \tag{9}$$

which completes the proof.

[5] Note that, although Equation (10) allows to work with (discretized) float numbers, we focus only on the positive integers, so that the dummy variable j on the productory takes jumps of size one.

[6] Please observe that, since we work with discrete distributions with jumps of size one, $\mathbb{P}(X \geq i) = S_X(i-1)$

Since the RUP is conjugate to left-truncation, we can immediately extend Equation (6) to include this effect:

$$\hat{S}(x) = \mathbb{P}(X_{n+1} > x | \boldsymbol{X}_n^*, \boldsymbol{T}_n, \boldsymbol{\delta}_n) = \prod_{j=0}^{x} \left[1 - \frac{\beta_j + m_j^*(\boldsymbol{X}_n^*, \boldsymbol{\delta}_n)}{\beta_j + \omega_j + s_j(\boldsymbol{X}_n^*, \boldsymbol{T}_n)} \right], \tag{10}$$

with $s_j(\boldsymbol{x}_n, \boldsymbol{t}_n) = \sum_{i=1}^{n} \mathbb{1}_{\{t_i \leq j \leq x_i\}}$ the number of observations censored at j under left-truncation.

Finally notice that, according to the assumptions made in Section 2, we can also model the truncation variable T as a RUP itself, with no censoring and *right*-truncation. Although the product-limit estimator under right-truncation is different in nature from (10), even when $\beta_j = \omega_j = 0$ [Wang, 1987]–and this could cast doubts about the RUP representation of T–this problem disappears once we estimate the whole, unbiased sample via the EM algorithm in Section 4. For the moment, we just notice that, given the methodology we propose in the next sections, it is necessary to be able to model the truncation variable as a RUP.

3.2 Urn representation

Following Muliere et al. [2000], we now show that Equation (10) can be obtained using a series of Polya urns, something very useful for the rest of the work.

Assume we have $M + 1$ Polya urns [Mahmoud, 2008], where the j-th urn U_j, $j = 0, 1, ..., M$, initially contains $\omega_j > 0$ white balls and $\beta_j > 0$ black balls. The only exception is urn U_0, which only possesses white balls. This starting urn composition is nothing more than a way of eliciting our a priori Muliere et al. [2000].

We start by drawing a ball from U_0. The only color we can sample is white. We put back the sampled ball, add an extra white ball and move to sample urn U_1. If in U_1 the sampled ball is white, we put it back, add an extra ball of the same color, and go sampling urn U_2. If the ball sampled from U_2 is also white, then we add a white ball to U_2 and draw from U_3, and so on. Every sampled white ball makes us move one step further in the sequence of urns. But if the sampled ball is black, say in urn U_2, we add an extra black ball to U_2, set $X_1 = 1$, and we start drawing again from U_1. If the next black ball is drawn, say, at U_{20}, then we set $X_2 = 20$, and start again from U_0. This sampling scheme defines a reinforced random walk on the state space of the $M + 1$ urns: every extraction of a black ball determines a cycle for the process $\{X_n\}$, which is nothing more than a RUP. Note in fact that, after n samplings of X, the probability distribution of X_{n+1} will indeed be given by Equation (10), if no truncation nor censoring are considered. In Muliere et al. [2000], the properties of this urn construction, including the conditions for recurrence are studied in detail.

If the data we want to model include right-censoring and left-truncation, the sampling mechanism can be modified as follows. In case of a right-censored value j, under the k-th cycle of the RUP, we add one white ball up to urn U_j (included), but no black ball in any of the urns. If, conversely, under the k-th cycle of the RUP, the sample is left-truncated in j, we start drawing from urn U_j, instead of U_0. It is clear that in this way we can fully reproduce Equation (10).

Observe that, under this sampling mechanism, each time a ball is drawn, the probability of sampling a ball of the same color in a given urn increases. And this will affect the RUP the next time it will visit the same urn. Moreover, although in this example we have assumed that the reinforcement (the number of extra balls added after each drawing) is equal to 1, one can easily add a parameter r controlling how many balls are added to each urn after sampling. Clearly, the higher r, the quicker the RUP will change according to sampling, while it will barely move from the initial urn compositions for $r \to 0$. The importance of r is discussed in several papers, e.g. Cheng and Cirillo [2018], Cirillo et al. [2010], Peluso et al. [2015]. The calibration of r allows us to control for how much a RUP should learn from data via reinforcement, and how much we actually trust our a priori.

An important characteristic of the beta-Stacy process as a random distribution, inherited from the Dirichlet process, is that its trajectories can be centered around a certain probability distribution $G(\cdot)$, which in Bayesian nonparametric estimation plays the role of the prior distribution. As shown in Walker and Muliere [1997], a necessary condition for this property to hold is that

$$\frac{\beta_j}{\beta_j + \omega_j} = \frac{G(j) - G(j-1)}{1 - G(j-1)}, \qquad j \in \mathbb{N} \tag{11}$$

where $G(j) = \mathbb{P}_G(X \leq j)$ is the probability that X is at most j under the prior G.

In the wake of Walker and Muliere [1997], Muliere et al. [2000] and Bulla et al. [2007], in what follows we assume

$$\beta_j = c_j G(\{j\}), \qquad \omega_j = c_j (1 - G(j)), \qquad c_j \in \mathbb{R}^+, j \in \mathbb{N}, \tag{12}$$

with c_j denoting the strength of belief in our prior knowledge and $G(\{j\}) = \mathbb{P}_G(X = j)$. Notice once again that this implies that our a priori is easily elicited via the initial urn compositions [Muliere et al., 2000].

If we choose a high value for c_j, large amounts of data will be required in order for the posterior distribution to deviate from our a priori. On the contrary, when $c_j \to 0$, we recover from Equation (10) the KM estimator of Cox and Oakes [1984].

Finally, observe that the roles of the degree of belief parameters c_j and the reinforcement r are actually opposite: increasing both by the same amount leaves the updating mechanism unchanged. In fact, it is possible to fix one of them and just work with the remaining one, and choosing one or another (or both) is just a matter of taste in the calibration.

3.3 The Bivariate RUP of Bulla et al. [2007]

Let us now quickly discuss the bivariate RUP construction of Bulla et al. [2007], also known as the B-RUP model.

Assume we observe couples of data of the form $((\boldsymbol{X}_n^*, \boldsymbol{T}_n^X, \boldsymbol{\delta}_n), (\boldsymbol{Y}_n^*, \boldsymbol{T}_n^Y, \boldsymbol{\epsilon}_n))$, where $\boldsymbol{X}_n = \{X_n, n \geq 1\}$ and $\boldsymbol{Y}_n = \{Y_n, n \geq 1\}$ are the possibly censored observations corresponding to the target variables X and Y. As before $\boldsymbol{T}_n^X$ and $\boldsymbol{T}_n^Y$ are the truncation processes for X and Y, respectively[7].

A flexible yet simple way of modelling the dependence between X and Y is to consider a one-factor construction, playing with three independent components: one common factor for X and Y jointly, and two idiosyncratic factors for X and Y apart.

Let A, B and C be independent RUPs with parameters (β_j^A, ω_j^A), (β_j^B, ω_j^B) and (β_j^C, ω_j^C) for $j \in \mathbb{N}_0$. These RUPs can be subject to truncation and censoring.

Now, set

$$\begin{aligned} X &= A + B, \\ Y &= A + C. \end{aligned} \tag{13}$$

The dependence between X and Y clearly relies on A, and therefore, conditioned on this common process, X and Y are independent. Again, this is nothing more than a simple one-factor model defined via RUPs.

A straightforward calculation yields

$$\text{Cov}(X_{n+1}, Y_{n+1} | \boldsymbol{A}_n, \boldsymbol{B}_n, \boldsymbol{C}_n) = \text{Var}(A_{n+1} | \boldsymbol{A}_n), \quad n \geq 1, \tag{14}$$

where $\boldsymbol{A}_n = \{A_n, n \geq 1\}$ is a sample of size n from A, and similarly for $\boldsymbol{B}_n$ and $\boldsymbol{C}_n$.

In Bulla et al. [2007] it is shown that the sequence $\{(X_n, Y_n), n \geq 1\}$ is exchangeable, and therefore by the de Finetti theorem there exists a joint random distribution function F_{XY} for which the elements of $(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ are independent and identically distributed according to F_{XY}. We refer to Bulla et al. [2007] for a detailed explanation of the many probabilistic properties of F_{XY}.

4 Expectation-Maximization

Expectation-Maximization (EM) is the de facto algorithm when dealing with incomplete data. It was originally introduced in Dempster et al. [1977], although particular instances of the algorithm had already been developed before that, e.g. in Turnbull [1976] for univariate censored and truncated data.

As noted in Dempster et al. [1977], the distinction between incomplete and complete data implies the existence of two spaces: $\mathscr{X}$ and $\mathscr{Y}$. The realizations of the observed data belong to $\mathscr{Y}$, while there exists a many-to-one mapping between $\mathscr{X}$ and $\mathscr{Y}$. In other words, for one realization $\mathbf{y}$ from $\mathscr{Y}$, there are several $\mathbf{x}$'s in $\mathscr{X}$ verifying the equation $\mathbf{y}(\mathbf{x}) = \mathbf{y}$. Using the same notation of Dempster et al. [1977], we denote this subset as $\mathscr{X}(\mathbf{y})$.

If we define $f(\mathbf{x})$ as the p.d.f (probability density function) of the complete data $\mathbf{x}$, we can obtain the p.d.f of $\mathbf{y}$ by integrating $f(\mathbf{x})$ over the subset $\mathscr{X}(\mathbf{y})$:

$$g(\mathbf{y}|\theta) = \int_{\mathscr{X}(\mathbf{y})} f(\mathbf{x}|\theta)\, d\mathbf{x}, \tag{15}$$

with $\theta \in \Theta$ being an instance of the parameters, and $\Theta \subset \mathbb{R}^m$ the parameter space.

With this distinction, the main idea behind the EM algorithm is to use the observations from $\mathbf{y}$ to compute the expected value of the log-likelihood of $\mathbf{x}$ ($\log f(\mathbf{x}|\theta)$) in an iterative procedure, instead of working with the incomplete log-likelihood, $\log g(\mathbf{y}|\theta)$, directly. Working with the complete data usually makes the model more tractable and easier

[7] In this case we also need to model T^X and T^Y as RUPs. In particular, given the assumptions of Section 2, we need to define a RUP for T and ϵ, with $T^X = T$ and $T^Y = T + \epsilon$.

to interpret. Furthermore, there are many cases for which the maximization of the complete log-likelihood can be computed analytically, while the maximization of $\log g(\mathbf{y}|\theta)$ is usually not straightforward.

From a mathematical point of view, the algorithm requires the computation of the expectation of $\log f(\mathbf{x}|\theta)$ at each iteration, conditioned on the observed data $\mathbf{y}$ and the estimates of the parameters from the previous iteration.

The following scheme summarizes the core of the EM algorithm:

1. Initialization: start with an initial estimate $\theta^{[0]}$ of the parameters.
2. Expectation step (E-step): compute the expectation of the complete log-likelihood with our current estimate of $\theta^{[k]}$:
$$Q(\theta|\theta^{[k]}) = \mathbb{E}[\log f(\mathbf{x}|\theta)|\mathbf{y}, \theta^{[k]}]. \tag{16}$$
The first argument in Q denotes the parameters that we want to estimate in the $(k+1)$-th iteration, while the second argument denotes a fixed and known set of parameters, obtained by the same procedure of the previous iteration.
3. Maximization (M-step): find the values for θ that maximize $Q(\theta|\theta^{[k]})$, i.e.
$$\theta^{[k+1]} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{[k]}). \tag{17}$$
4. Repeat from step 2 until convergence is reached.

Noticed that, as proved by Dempster et al. [1977], the EM algorithm presents a monotonic increase of the likelihood at each iteration, and while its convergence rate is usually slow, there are several proposals in the literature to improve this rate (e.g. Varadhan and Roland [2008] and references therein). However, the EM does not achieve global convergence, and therefore running the algorithm with several initial guesses is recommended[8].

Given the wide range of meanings for the term “incomplete", the EM algorithm appears in different versions in the literature. Here, when we talk about incomplete data, we always refer to observations that suffer from left-truncation and/or right-censoring.

For the bivariate case, there is a further level of incompleteness that is included through the one factor construction of Equation (13). In fact, even if we observe the pairs (X, Y) without truncation nor censoring, the model would still be “incomplete", given that what we really need are the observations for the RUPs A, B and C.

Before entering into the core of the paper, we introduce some notation to facilitate the reading:

$$p_X(x) \coloneqq \mathbb{P}(X = x|\theta), \tag{18}$$

while

$$p_X^{[k]}(x) \coloneqq \mathbb{P}(X = x|\theta^{[k]}), \tag{19}$$

where $\theta^{[k]}$ is defined as before in the context of the EM iterations. We also introduce the survival probability function $S_X(x) \coloneqq \mathbb{P}(X > x|\theta)$, with $S_X^{[k]}(x)$ defined analogously.

4.1 The EM algorithm for a single RUP

Let X be a r.v. on the positive integers generated by the following distribution:

$$S_X(j|\theta) = \prod_{i=0}^{j} \frac{W_i}{N_i}, j \geq 0, \tag{20}$$

where $\theta = \{(W_i, N_i), 0 \leq i \leq \infty\}$ are the hyperparameters of the distribution[9].

This distribution is a special case of the RUP distribution in Equation (10) in the absence of reinforcement. In fact, if $r = 0$, and defining $W_i = \omega_i$, $N_i = \beta_i + \omega_i$, the equivalence becomes evident. Also notice that, in the urn view, W_j denotes the number of white balls in urn U_j, and N_j the total number of balls in the same urn. In what follows,

[8] In Wu [1983], it is shown that the algorithm may not even converge to a local maximum, depending on conditions that in practice are very difficult to validate. For an extensive discussion on this topic we direct the reader to Wu [1983], Boyles [1983] and Nettleton [1999].

[9] Notice that, indeed, defining the pair (W_i, N_i) is somewhat redundant since, given the shape of (20), only the ratio between them is relevant. The reason for this choice is already hinted in Section 3, and will become more clear in Section 5.

we will not take the pairs (W_i, N_i) to come from a given a priori, but to be the parameters to be calibrated via the EM algorithm.

If we have a series of observations $\boldsymbol{X} = \{x_i, 1 \leq i \leq n\}$ generated i.i.d by (20), the sample log-likelihood is:

$$L(\mathbf{X}|\theta) = \sum_{i=1}^{n} \left(\log(N_{x_i} - W_{x_i}) - \log(N_{x_i}) + \sum_{j=0}^{x_i - 1} (\log(W_j) - \log(N_j)) \right). \tag{21}$$

If we derive this expression with respect to W_i we obtain the optimal values:

$$W_j = \frac{s_{j+1}(\boldsymbol{X})}{s_j(\boldsymbol{X})} N_j, \tag{22}$$

where $s_j(\boldsymbol{X}) = \sum_{i=1}^{n} \mathbb{1}_{\{j \leq x_i\}}$ was already defined in Section 3. If we now plug Equation (22) into Equation (20), we end up with the KM estimator in the absence of censoring. Moreover, given Equation (3) for the likelihood under left-truncation and right-censoring, it is obvious that if we compute the optimal parameters W_j in this case, and plug them back in Equation (20), we recover the estimator of Cox and Oakes [1984]. This result is, in fact, a trivial consequence of Equation (10). Just notice that, since the KM estimator maximizes the likelihood in Equation (3) with respect to the LTRC data, if we want to optimize Equation (10) so to also maximize Equation (3), we just need to omit the contribution of the prior distribution given by the pairs (β_j, ω_j).

Although the optimal parameters of the RUP distribution can already be obtained through the classic MLE, implementing the EM version of this problem will help us highlight some of the key points that we will use for the bivariate setting, where a nonparametric solution via MLE is far from trivial.

How do we approach this problem using the EM algorithm instead of the usual product-limit estimator? Estimating censored observations via EM is a well-known problem [Turnbull, 1976]. Regarding truncation, in Dempster et al. [1977] and McLachlan and Jones [1988], it was already explained that we need to estimate the whole sample from our biased observations. In our case the biased data is that for which $T \leq X$–being X and T the target and truncation variables, respectively–and the complete sample is the one considering also observations where $T > X$ holds.

However, by the very definition of truncation, there is no information at all in our data set about the region $T > X$. At most, what we can do is estimate the size of the whole sample M as

$$M = \frac{n}{\mathbb{P}(T \leq X)}, \tag{23}$$

and consider every value of (X, T) for which $T > X$ holds. Therefore, given a LTRC sample where the first n_0 observations are uncensored and the other $n - n_0$ are right-censored, we distinguish three clear components in our likelihood:

- $T \leq X$ and X uncensored:

$$L_1(X, \delta, T) = \sum_{i=1}^{n_0} [\log(p_X(x_i)) + \log(p_T(t_i))]. \tag{24}$$

- $T \leq X$ and X right-censored:

$$L_2(X, \delta, T) = \sum_{i=n0+1}^{n} [\mathbb{E}(\log(p_X(X))|X > x_i) + \log(p_T(t_i))]. \tag{25}$$

- $T > X$:

$$L_3(X, \delta, T) = M^{[k]} \sum_{x=0}^{\infty} \sum_{t=x+1}^{\infty} [\log(p_X(x)) + \log(p_T(t))]\, p_X^{[k]}(x)\, p_T^{[k]}(t), \tag{26}$$

with $M^{[k]} = \dfrac{n}{p^{[k]}(T \leq X)}$, where $p^{[k]}(T \leq X)$ is the probability of truncation using the estimates of the k-th iteration, and

$$\mathbb{E}(\log(p_X(X))|X > x_i) = \sum_{j=x_0+1}^{\infty} \log(p_X(j))\, p_X^{[k]}(j|X > x_i). \tag{27}$$

Notice that in the second likelihood component the truncation variable is not affected by censoring and that $L_3(X, \delta, T)$ is not conditioned on the data because there are no observations that fall in that region.

Now we only need to apply the maximization step with respect to $L(X,T) = L_1(X,T) + L_2(X,T) + L_3(X,T)$ in order to complete the model. Observe that the maximization in this case is very simple since we can treat the likelihoods of X and T separately. In fact, it is not difficult to prove that:

$$W_j^X = \frac{\sum_{i=1}^{n_0} S_X^{[k]}(j|x_i) + \sum_{i=n_0+1}^{n} S_X^{[k]}(j|X > x_i) + S_X^{[k]}(j|T > X)}{\sum_{i=1}^{n_0} S_X^{[k]}(j-1|x_i) + \sum_{i=n_0+1}^{n} S_X^{[k]}(j-1|X > x_i) + S_X^{[k]}(j-1|T > X)} N_j^X, \tag{28}$$

and

$$W_j^T = \frac{\sum_{i=1}^{n} S_T^{[k]}(j|t_i) + S_T^{[k]}(j|T > X)}{\sum_{i=1}^{n} S_T^{[k]}(j-1|t_i) + S_T^{[k]}(j-1|T > X)} N_j^T. \tag{29}$$

In the next sections we extend this result to the bivariate scenario via the one-factor model of Equation (13). We will also assume that every r.v. that we define is generated by an urn distribution like (20), unless otherwise specified.

4.2 The EM algorithm in a one factor model with RUPs

For ease of explanation, let us assume that there is no truncation nor censoring at first. Thanks to the one-factor construction, we can write the complete likelihood as the product of the likelihoods of A, B and C. That is,

$$\log p(A = a, B = b, C = c|\theta) = \log p_A(a) + \log p_B(b) + \log p_C(c), \tag{30}$$

which greatly simplifies computations, since we can work on each component separately. For example, using the EM algorithm, the complete likelihood of A given an observation (x, y) of X and Y is given by

$$Q_A(\theta|\theta^{[k]}) = \sum_{a=0}^{x \wedge y} \log p_A(a)\ p_A^{[k]}(a|x, y), \tag{31}$$

where the lower limit of the summation in Equation (31) is zero, for we work with nonnegative processes, while the upper limit is the minimum between x and y, since, due to Equation (13), A cannot be bigger than X or Y.

Furthermore, given that

$$p_A(a|x, y) = \frac{p_A(a)p_B(x - a)p_C(y - a)}{p_{XY}(x, y)}, \tag{32}$$

we can write Equation (31) as

$$Q_A(\theta|\theta^{[k]}) = \sum_{a=0}^{x \wedge y} \log p_A(a) \frac{p_A^{[k]}(a)p_B^{[k]}(x - a)p_C^{[k]}(y - a)}{p_{XY}^{[k]}(x, y)}. \tag{33}$$

However, for the sake of readability, we will keep using Equation (31), as it is easier to interpret.

Once we have the expectation of the complete likelihood, we only need to compute the derivatives with respect to the parameters, and equal them to zero to obtain the values of the next iteration. The following expression for the derivative of (20) with respect to its parameters will be used extensively throughout the rest of the paper:

$$\frac{\partial \log p_A(a)}{\partial W_j^A} = \begin{cases} 0 & \text{if } j > a \\ \frac{-1}{N_j^A - W_j^A} & \text{if } j = a \\ \frac{1}{W_j^A} & \text{if } j < a, \end{cases} \tag{34}$$

where, using the intuition developed in Section 3, W_j is the number of white balls[10] in the j-th urn, and N_j is the total number of balls in j. Notice that we assume N_j to be constant, since, given (20), we only need to compute the ratio W_j/N_j in order to completely specify the distribution.

Combining Equations (34) and (31), we obtain

$$\frac{\partial Q_A(\theta|\theta^{[k]})}{\partial W_j^A} = \frac{1}{W_j^A} S_A^{[k]}(j|x, y) - \frac{1}{N_j^A - W_j^A} p_A^{[k]}(j|x, y). \tag{35}$$

[10] Although there is a clear relation between ω_j from Equation (10) and W_j, note that ω_j is the number of white balls given by the prior distribution, while W_j is the total number of white balls in the same urn at every moment.

Equating this last expression to zero and solving for W_j^A yields the value for the next iteration:

$$W_j^A = \frac{S_A^{[k]}(j|x,y)}{S_A^{[k]}(j-1|x,y)} N_j^A, \tag{36}$$

where we have assumed that our processes are discrete with jumps of size one, and thus $p(A \geq j) = p(A > j-1)$, although this can trivially be generalized to jumps of different sizes.

Equation (36) is consistent with the definition of the urn distribution, since the ratio W_j^A/N_j^A is defined as the probability of surviving longer than the j-th urn, conditioned on the fact that the process A has reached it. In probabilistic terms this can be expressed as

$$\frac{W_j^A}{N_j^A} = p_A(A > j | A \geq j) = \frac{S_A(j)}{S_A(j-1)}. \tag{37}$$

Comparing Equations (37) and (36), we can see that the intuition of the ball ratio is preserved through the EM iterations, but now we condition on the incomplete observations (x, y) and on the results of the previous iteration. If instead of one observation we have a sample of size n, Equation (36) becomes:

$$W_j^A = \frac{\sum_{i=1}^n S_A^{[k]}(j|x_i,y_i)}{\sum_{i=1}^n S_A^{[k]}(j-1|x_i,y_i)} N_j^A. \tag{38}$$

The solutions for W_j^B and W_j^C are completely analogous, and therefore they are omitted.

4.3 The EM algorithm for RUPs and right-censoring

When right-censoring occurs in a bivariate model, it is helpful to distinguish 3 different cases: a) both X and Y are uncensored, b) one of them is observed and the other is censored, and c) both of them are censored.

Let us start by considering case b), when only one of the two variables is right-censored. We will add the superscript $*$ to the censored variable, so that the observation (x^*, y) means that Y is observed with value y and X is censored with value x. We define analogously (x, y^*). Below we only show the solution for (x^*, y), since the one for (x, y^*) requires simple changes like dealing with C instead of B.

The conditional probabilities for A, B and C are:

$$p_A(a|x^*, y) = p_A(a) S_B(x-a) p_C(y-a), \qquad \text{if } a \leq y, \tag{39}$$

$$p_B(b|x^*, y) = p_B(b) \sum_{a=x-b}^{y} p_A(a) p_C(y-a), \qquad \text{if } b \geq x-y, \tag{40}$$

$$p_C(c|x^*, y) = p_C(c) p_A(y-c) S_B(x-y+c), \qquad \text{if } c \leq y. \tag{41}$$

Note that, even if X is censored, A cannot be bigger than Y if it is an exact observation, and therefore asymmetries arise in the formulas of B and C.

When both X and Y are censored, i.e. under case c), we consider the couple (x^*, y^*), and the conditional probabilities are

$$p_A(a|x^*, y^*) = p_A(a) S_B(x-a) S_C(y-a), \qquad \text{if } a \geq 0, \tag{42}$$

$$p_B(b|x^*, y^*) = p_B(b) \sum_{a=x-b}^{\infty} p_A(a) S_C(y-a), \qquad \text{if } b \geq 0, \tag{43}$$

$$p_C(c|x^*, y^*) = p_C(c) \sum_{a=x-c}^{\infty} p_A(a) S_B(x-a), \qquad \text{if } c \geq 0. \tag{44}$$

In the following, assume we arrange our n observations so that the first n_0 observations are uncensored. In the next $n_X - n_0$ samples censoring only affects X and, in the next $n_Y - n_X$ couples after that, censoring only happens on Y. Finally, in the last $n - n_Y$ samples censoring occurs on both variables. By a similar procedure as in Section 4.2, we arrive at the following expression for the parameters of A:

$$W_j^A = \frac{{}_{0}^{n_0}S_A^{[k]}(j|\boldsymbol{X},\boldsymbol{Y}) + {}_{n_0}^{n_X}S_A^{[k]}(j|\boldsymbol{X}^*,\boldsymbol{Y}) + {}_{n_X}^{n_Y}S_A^{[k]}(j|\boldsymbol{X},\boldsymbol{Y}^*) + {}_{n}^{n_Y}S_A^{[k]}(j|\boldsymbol{X}^*,\boldsymbol{Y}^*)}{{}_{0}^{n_0}S_A^{[k]}(j-1|\boldsymbol{X},\boldsymbol{Y}) + {}_{n_0}^{n_X}S_A^{[k]}(j-1|\boldsymbol{X}^*,\boldsymbol{Y}) + {}_{n_X}^{n_Y}S_A^{[k]}(j-1|\boldsymbol{X},\boldsymbol{Y}^*) + {}_{n}^{n_Y}S_A^{[k]}(j-1|\boldsymbol{X}^*,\boldsymbol{Y}^*)} N_j^A, \tag{45}$$

where ${}_{n_1}^{n_2}S_A^{[k]}(j|\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=n_1+1}^{n_2} S_A^{[k]}(j|x_i,y_i)$. As before, the expressions for W_j^B and W_j^C are completely analogous and therefore omitted.

When no censoring takes place, $n_0 = n$, and we recover Equation (38) from Equation (45). Note that no information about the censoring processes is required in order to estimate A, B and C from the observed data, and therefore the dimensionality remains unchanged under censoring. We will see next that this is not the case when dealing with truncation.

4.4 The general model for LTRC data

While the literature on how to use EM to tackle random censoring is rather extensive, much less has been said about random truncation. As we shall see, in this case the modelling of the truncated variable is required, which not only increases the dimensionality of our problem, but, in many applications, we cannot infer any meaningful results from the truncation variable, making it of no particular interest.

However, as explained in Section 2, bivariate truncation is sensitive to the dependence between X and Y, to the point that KM-type estimators are no longer reliable to accurately analyse the marginal distributions, and thus target and truncated variables should be modelled jointly. Luckily, under the one factor model and thanks to our truncation assumptions, all processes are independent, and therefore, even if the dimensionality increases, each likelihood can be treated separately.

If we want to extend the methodology introduced in Subsection 4.1 to cope with bivariate distributions, under the truncation assumptions of Section 2, the truncation event becomes

$$\mathscr{A} = (T \leq X, T + \epsilon \leq Y + \epsilon_0). \tag{46}$$

The unobserved region is that for which $\mathscr{A}^{\mathsf{c}}$, the complementary of $\mathscr{A}$, is true. Since $\mathscr{A}^{\mathsf{c}}$ actually consists of three different events–i.e. $(T \leq X, T + \epsilon > Y + \epsilon_0)$, $(T > X, T + \epsilon \leq Y + \epsilon_0)$ and $(T > X, T + \epsilon > Y + \epsilon_0)$–and

$$p(\cdot) = p(\cdot|\mathscr{A})p(\mathscr{A}) + p(\cdot|\mathscr{A}^{\mathsf{c}})p(\mathscr{A}^{\mathsf{c}}), \tag{47}$$

it is easier to compute all quantities conditioned on $\mathscr{A}$ and then use Equation (47) to condition on its complementary.

Finally, we substitute the pair (X, Y) with the triplet (A, B, C) by means of Equation (13), and compute the optimal values for W_j^A and the other processes following the same approach of the previous sections. The only difference is the computation of the conditional probabilities, since in this case the truncation variables are also involved.

Notice that we also need the probability of the truncation event, since otherwise we cannot estimate the whole sample size (M), as per Eq. (23). This probability can be computed as

$$p(\mathscr{A}) = \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} p_{XY}(x,y)p(T \leq x, T + \epsilon \leq y + \epsilon_0), \tag{48}$$

with

$$p(T \leq x, T + \epsilon \leq y + \epsilon_0) = \sum_{t=0}^{x} p_T(t)p_\epsilon(\epsilon \leq y + \epsilon_0 - t). \tag{49}$$

Furthermore, the probabilities conditioned on the truncation event, for every variable, are

$$p_A(a|\mathscr{A}) = \frac{p_A(a)}{p(\mathscr{A})}\sum_{x=a}^{\infty}\sum_{y=a}^{\infty} p_B(x-a)p_C(y-a)p(T \leq x, T + \epsilon \leq y + \epsilon_0), \tag{50}$$

$$p_B(b|\mathscr{A}) = \frac{p_B(b)}{p(\mathscr{A})}\sum_{x=b}^{\infty}\sum_{y=x-b}^{\infty} p_C(y-x+b)p_A(x-b)p(T \leq x, T + \epsilon \leq y + \epsilon_0), \tag{51}$$

$$p_C(c|\mathscr{A}) = \frac{p_C(c)}{p(\mathscr{A})}\sum_{y=c}^{\infty}\sum_{x=y-c}^{\infty} p_B(x-y+c)p_A(y-c)p(T \leq x, T + \epsilon \leq y + \epsilon_0), \tag{52}$$

$$p_T(t|\mathscr{A}) = \frac{p_T(t)}{p(\mathscr{A})}\sum_{x=t}^{\infty}\sum_{y=(t-\epsilon_0)^+}^{\infty} p_{XY}(x,y)p_\epsilon(\epsilon \leq y + \epsilon_0 - t), \tag{53}$$

$$p_\epsilon(e|\mathscr{A}) = \frac{p_\epsilon(e)}{p(\mathscr{A})} \sum_{x=0}^{\infty} \sum_{y=(e-\epsilon_0)^+}^{\infty} p_{XY}(x,y) p_T(T \leq x \wedge (y + \epsilon_0 - e)). \tag{54}$$

In order to include left-truncation in (45), it is easy to verify that we only need to add the term $(M^{[k]} - n)\, S_A^{[k]}(j|\mathscr{A}^\mathsf{c})$ in the numerator and $(M^{[k]} - n)\, S_A^{[k]}(j-1|\mathscr{A}^\mathsf{c})$ in the denominator, where $M^{[k]}$ is the estimated sample size at the kth iteration, according to Eq. (23).

For the processes T and ϵ the expressions simplify, since we do not have to account for any censoring. We show here the solution for W_j^T, while W_j^ϵ is obtained in an analogous way:

$$W_j^T = \frac{\sum_{i=1}^n \mathbb{1}_{\{t_i > j\}} + (M^{[k]} - n)\, S_T^{[k]}(j|\mathscr{A}^\mathsf{c})}{\sum_{i=1}^n \mathbb{1}_{\{t_i \geq j\}} + (M^{[k]} - n)\, S_T^{[k]}(j-1|\mathscr{A}^\mathsf{c})}\, N_j^T, \tag{55}$$

where $\boldsymbol{T} = (t_1, ..., t_n)$ is the observed sample of T. It is worthy to stress that, in the absence of truncation, Equation (55) returns the standard KM estimator. Therefore, Equation (55) can be interpreted as the expected value of the KM estimator given the missing data and our current estimator of p_T.

Moreover, Equation (45) can also be interpreted as the KM estimator for A given the observations of (x, y) and its censored counterparts. Thus, by breaking the bivariate joint distribution into a convolution of one-dimensional processes, we only need to compute separately the expected KM estimates of the relevant variables, given the incomplete observations and the results from the previous iteration.

5 The Expectation-Reinforcement algorithm

The last result of Section 4 combined with Equation (10) inspired us to consider the inclusion of the reinforcement mechanism of RUPs into the EM algorithm, offering the possibility of embedding experts' knowledge into the estimates. Due to the combination of EM and RUPs, we call the new algorithm *Expectation-Reinforcement* (ER) algorithm.

The idea is to combine a prior distribution, given by the pairs $\{\beta_j, \omega_j\}$ defined in Equation (12), with the expected values of the KM estimates at each EM iteration, so to obtain a posterior distribution that mixes both experts knowledge and data. Since in many applications the amount of data is rather scarce, especially when considering phenomena characterized by extreme risks and fat tails [McNeil et al., 2015, Taleb, 2007], or by epistemic uncertainty [Shackle, 1955], the use of some experts' intuition can be extremely useful to improve the performance of the model.

Moreover, nonparametric estimators, as the one described before, usually suffer from overfitting [James et al., 2013]. Such a problem occurs when the model calibrates too well to the sample data, making the whole procedure highly sensitive to small variations in the sample properties, and thus reducing its predictive power out-of-sample. Of course, the magnitude of this variation diminishes for large and reliable data sets [Wasserman, 2006], but in many applications such a thing is simply unavailable. For example, in credit risk one requires information about the default event of other companies for a proper model calibration. However, these events are rare by nature, and thus one only has access to a few observations [McNeil et al., 2015].

From the bias-variance trade-off point of view, nonparametric estimators have the smallest possible bias, as they capture all the features of the data set, but their variance can be considerably large since their parameters are very sensitive to small changes in the observations. By embedding the reinforcement mechanism of RUPs into the EM algorithm we can control for this trade-off as follows: for extremely high strengths of belief (and/or a very small reinforcement), the posterior distribution will not be affected by the observations and will therefore tend to coincide with the prior distribution; while in the opposite case, with almost zero strength of belief (and/or strong reinforcement), the posterior will adapt completely to the data. In the first scenario, the variance of the model with respect to different data sets is zero, but the bias will be arbitrarily high if our initial guesses are wrong. In the second scenario, the bias will be as small as possible but the model will be very sensitive the small changes in the data, and therefore the variance will likely be high. Thus the trade-off can be somehow balanced by choosing intermediate values of the strength of belief and reinforcement parameters.

In the following, we show an example where the estimates of the ER algorithm are computed for the truncation variable T of the EM estimators in Equation (55). Since, given the EM estimator, the procedure is the same for the other variables previously considered, their explicit ER derivation is omitted.

Let us assume that T is a RUP with prior distribution given by the pairs $\{\beta_j^T, \omega_j^T\}$, for $j = 0, ..., M$, through Equation (12). Now we follow the steps described in Section 4 to obtain the EM estimates at each iteration, but instead of using

Equation (55) to update the number of balls in each urn, the updating mechanism under the ER algorithm becomes

$$W_j^T = \omega_j^T + r\left[\sum_{i=1}^{n} \mathbb{1}_{\{t_i>j\}} + (M^{[k]} - n)\, S_T^{[k]}(j|\mathscr{A}^{\mathsf{c}})\right], \tag{56}$$

$$N_j^T = \beta_j^T + \omega_j^T + r\left[\sum_{i=1}^{n} \mathbb{1}_{\{t_i\geq j\}} + (M^{[k]} - n)\, S_T^{[k]}(j-1|\mathscr{A}^{\mathsf{c}})\right], \tag{57}$$

where r is the reinforcement parameter and the other quantities are defined as usual, with the superscript $[k]$ indicating that those quantities are computed using the estimators of the k-th iteration of the algorithm.

Note that, by giving different values to the reinforcement and belief parameters[11], we can control which component dominates. Similarly to the behaviour shown in Section 2, when the strength of belief tends to zero, we simply recover the estimates of the EM algorithm, while if we make the reinforcement tend to zero instead, the posterior distribution will be very similar to the prior distribution.

Moreover, Equations (56) and (57) can be interpreted as the estimation of two different data sets combined: the actual observations $\boldsymbol{T}$ (multiplied r times), which correspond to the last term on the right-hand side of Equations (56) and (57); and a fictitious data set $\hat{\boldsymbol{T}}$ chosen in a way such that

$$\omega_j = \sum_{i=1}^{\hat{n}} \mathbb{1}_{\{\hat{t}_i>j\}}, \tag{58}$$

and

$$\beta_j = \sum_{i=1}^{\hat{n}} \mathbb{1}_{\{\hat{t}_i=j\}}, \tag{59}$$

where $\hat{n}$ is the size of the fictitious data, which coincides with the strength of belief c_j, if $c_j = c$ for all j. Notice that, while $\boldsymbol{T}$ may be subject to censoring and truncation, this is not the case for $\hat{\boldsymbol{T}}$.

The ER algorithm can therefore be considered as an instance of the EM algorithm, where one tries to find the parameters that maximize the incomplete likelihood of the new data set $\boldsymbol{T} \cup \hat{\boldsymbol{T}}$, where $\cup$ is to be interpreted as the combination of both data sets (a reasonable assumption in the case of i.i.d samples). This implies that all convergence results derived for the EM algorithm equally apply to the ER algorithm, in the context of the "new" data set.

6 The algorithms at work

In this section we present several examples showing the performance of our algorithms. In Subsection 6.1 we present analytical examples of left-truncation and right-censoring in both one and two dimensions. Then, in Subsection 6.3, we analyze a Canadian data set of annuities widely used by insurers [Frees et al., 1996], and known for its complexity, due to the strong presence of censoring and truncation.

6.1 Analytical examples

The advantage of analytical examples is that, knowing the true underlying distributions, not only we can test the accuracy of our model even when other commonly used benchmarks fail to capture the true solution, but our results are also easily replicable in a controlled environment.

We start with a univariate example where the ER estimator is compared with the KM estimator of Cox and Oakes [1984] and with the true solution. We will see that for high levels of truncation and censoring the KM estimator becomes more erratic on the tails, while the KM estimator still manages to give a smooth curve that captures fairly well the underlying distribution.

This difference will become more pronounced in the bivariate example, where the univariate KM estimates are not reliable under strong truncation, as they do not converge to the marginal distributions.

[11] By Equation (12) the strength of belief is already implicit in ω and β.

	EM	ER^l	ER^m	ER^h
r	–	10^4	$2 \cdot 10^2$	$2 \cdot 10^1$
c^X	0	1	$5 \cdot 10^2$	10^3
c^T	0	1	$5 \cdot 10^2$	10^3

Table 1: Different proposed scenarios defined by the ratio between the belief and reinforcement parameters. The values are used throughout Section 6. The superscripts in the ER columns denote "low", "medium" and "high", respectively, referring to the weight of the belief parameters. Here r is the reinforcement parameter—which, we assume, is the same for every variable—and c^X and c^T are the belief parameters associated to X and T, respectively, introduced into the pairs $\{\beta, \omega\}$ through Equation (12). Notice that, although we are assuming that c^X and c^T are constant for each variable, Equation (12) allows a more general representation where each urn has a different strength of belief.

	Data	KM	ER^l	ER^m	ER^h	Analytical
Mean	65.091	60.467	60.451	60.275	59.625	60.469
Variance	33.121	53.902	54.011	55.340	59.834	53.213

Table 2: Comparison of the mean and variance obtained using: the uncensored data set, the KM estimator, the ER estimator using the scenarios of Table 1, and the analytical solution, respectively. Notice that omitting truncation and censoring greatly underestimates the variance. As expected in this simple example, the best results are provided by the ER^l and KM solutions.

6.1.1 Univariate case

Take X, T and Δ defined as in Section 2, then assume $X \sim \text{Poi}(60)$, $T \sim \text{Poi}(70)$ and $\Delta \sim \text{Poi}(2)$, where $\text{Poi}(\mu)$ denotes a Poisson distribution with parameter μ. Clearly, $\mathbb{P}(T \leq X) \simeq 0.2$, so that only 20% of the whole sample is actually observed. Furthermore, around 70% of the observations are right-censored.

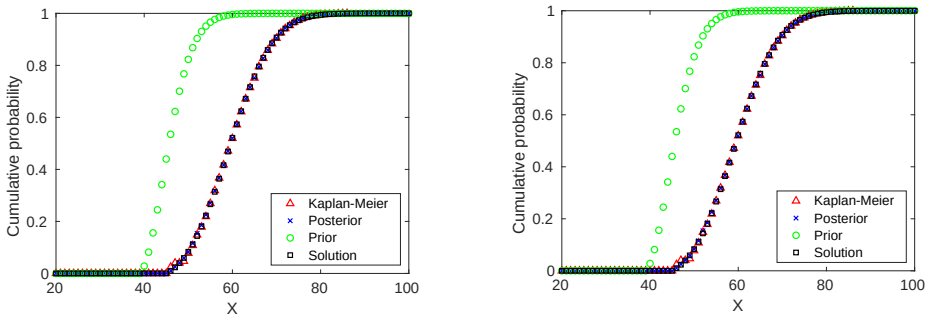
In order to use the ER algorithm, first we identify our variables of interest: X and T. Then we define a RUP for each of them with parameters $\{\beta_j^X, \omega_j^X\}$ and $\{\beta_j^T, \omega_j^T\}$, respectively.

As explained in Section 3, via Equation (12) we can center our RUP on a particular prior distribution G. For this first example we choose $G^T = \text{Poi}(40)$ and $G^X = \text{Poi}(40)$, thus assuming that our prior elicitation is not far from the truth, at least from the point of view of the distributional type. However, notice that both Poissons are far from the actual solution, and that the truncation level is underestimated. Moreover, given our discussion about censoring and truncation in Section 2, the estimated distributions are conditioned on the minimum uncensored observation, which in this example is $X_{min} = 46$.

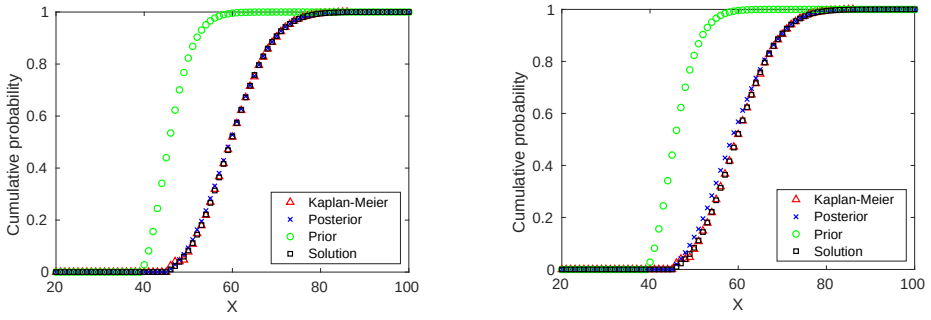
In Figure 1 we present the results obtained from a sample (X, T, δ) of size 10^4 for the values of the belief parameters defined in Table 1, ranging from zero belief—which corresponds to the EM estimator—to a high belief where the posterior distribution is highly influenced by the prior. In the figure it can be observed that, due to the high levels of truncation in the data, the KM estimator presents a stair-case behaviour near the tails, while the EM estimate returns a smooth curve that properly captures the underlying distribution. The same applies for the ER estimator with low strength of belief (Figure 1b). However, as we lean towards our prior, the fit given by the ER estimators starts to worsen. This is an expected result, since we know that our prior does not represent the true distribution. A more quantitative comparison is given in Table 2, where we present the means and variances for several scenarios. Moreover, due to the similarity between the results of the EM and ER^l estimators—with the superscript always referring to the corresponding scenario in Table 1—we will refer to both examples as just ER^l for simplicity.

This behaviour is further emphasized in Figure 2, where QQ-plots for the ER^l and KM estimators against the analytical solution are presented. The vertical lines in the subfigures represent the range in which data were observed, i.e. the minimum and maximum uncensored values of $\boldsymbol{X}^*$. Any inference beyond those lines is not possible since there is no data to compare with.

The QQ-plots were created by generating two samples of size 10^3 each from the ER^l and KM solutions, and comparing them with a sample of the same size from the analytical solution. The same samples are also used to perform permutation tests for the means and variances. We use the difference in means as test statistic for the means, while to compare the variances we use Good's test (see Good [1994], and Baker [1995] for a generalization with unequal sample sizes). The

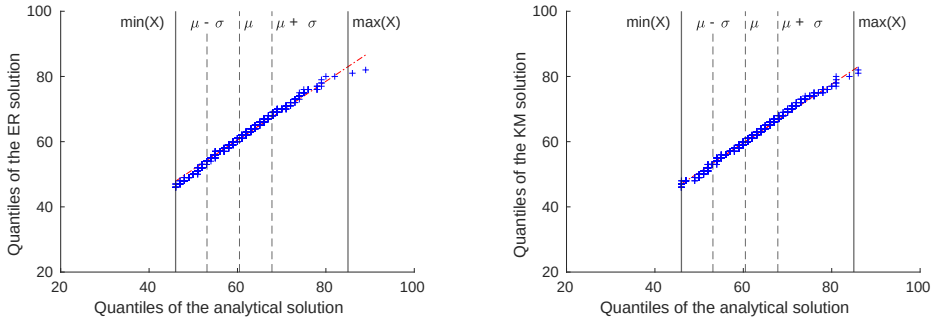


(a) EM. (b) ER^l.



(c) ER^m. (d) ER^h.

Figure 1: Comparison between the Kaplan-Meier estimator (red) and our ER result (blue). The analytical solution is represented in black and the initial estimate in green. Due to the left-truncation effect in the data, the KM estimator presents a stair-case behaviour near the left tail of the distribution, overfitting the data in that area.



(a) QQ-plot: EM solution. (b) QQ-plot: Kaplan-Meier estimator.

Figure 2: QQ-plot comparing our ER^l solution (a) and the KM estimator (b) with the analytical solution. The solid vertical lines, from left to right, correspond to the minimum and maximum uncensored values of $\boldsymbol{X}^*$, respectively. The dashed lines correspond to the mean (μ) and $\pm$ one standard deviation (σ).

results of the test are presented in Table 3, and, as expected, fail to reject the null hypothesis for both the ER^l and KM estimators.

6.1.2 Bivariate case I

In this section we present a more interesting example with a bivariate setting under high levels of truncation and censoring.

	Value	p (%)	H_0
ER^l (mean)	-0.508	11.09	Do not reject
KM (mean)	0.480	13.88	Do not reject
ER^l (variance)	5851	18.76	Do not reject
KM (variance)	5983	6.01	Do not reject

Table 3: Permutation tests for the ER^l and KM estimators. The first column gives the value of the test statistic, the second column is the p-value, and the last column tells whether we reject or fail to reject the null hypothesis that the distributions agree with the true one with a type-I error of 5%. The number of permutations for each test is 10^5.

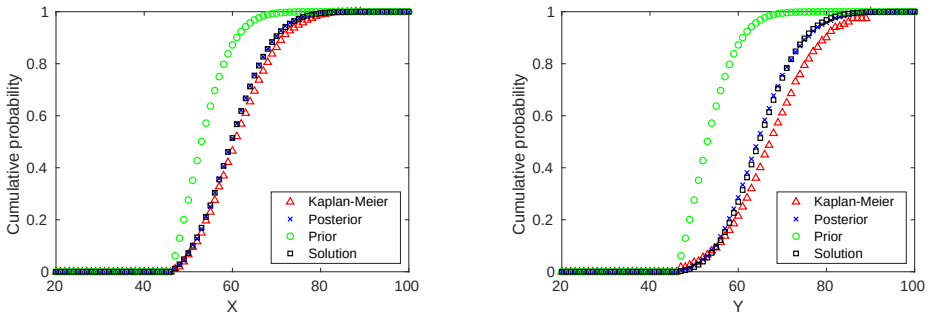
We assume the following: $A \sim \text{Poi}(40)$, $B \sim \text{Poi}(20)$, $C \sim \text{Poi}(25)$, $T \sim \text{Poi}(70)$, $\epsilon \sim \text{Poi}(7)$, $\epsilon_0 = 5$ and $\Delta \sim \text{Poi}(2)$. Thus, in our one-factor construction, the marginal distributions are $X \sim \text{Poi}(60)$ and $Y \sim \text{Poi}(65)$, respectively, and the correlation is $\rho = 0.64$. The sample consists of 10^4 couples of observations.

This time there are five processes that we need to model using RUPS: A, B, C, T and ϵ. As before, we start by giving values to the pairs $\{\beta_j, \omega_j\}$ so that the RUPs center on a specific a priori. Our choices for this example are $G^A = \text{Poi}(25)$, $G^B = \text{Poi}(25)$, $G^C = \text{Poi}(25)$, $G^T = \text{Poi}(50)$, $G^\epsilon = \text{Poi}(10)$, $\epsilon_0 = 10$. Again, we are underestimating the mean values of all processes, including the correlation and the truncation levels. Furthermore, the first uncensored observation occurs for $X = Y = 47$, and therefore every distribution we present will be conditioned accordingly. Finally, regarding the belief and reinforcement parameters, we assume $c^A = c^B = c^C = c^X$ and $c^T = c^\epsilon$, with c^X and c^T as per Table 1. Therefore, if we refer, for example, to the ER^l scenario, then $c^A = c^B = c^C = c^\epsilon = c^T = 1$ and $r = 10^4$, where again we use the same reinforcement for every RUP involved.

In Figure 3 we show our ER^l estimate for the marginals of X and Y compared to the KM estimator and the true distribution, while in Table 4 we compare the first two moments as well as the correlation. The results clearly show that the KM estimators are biased with respect to the underlying distribution, while our results nicely capture both curves and moments.

A more quantitative analysis on how good this fit is can be found in Figure 5—where we show the QQ-plots for the ER^l marginals—and Table 5, where we perform another permutation test for the means and variances (as before, with a sample of size 10^3).

Finally, in Figures 4a and 4b, we give contour plots of our bivariate ER distribution and the analytical solution for the ER^l and ER^h scenarios, respectively. Notice that, as expected, increasing the strength of belief in our a priori makes it more difficult for the data to influence the posterior distribution: since the initial prior does not fit the observations, increasing the strength of belief necessarily worsens the posterior fit[12].



(a) Marginal of X.

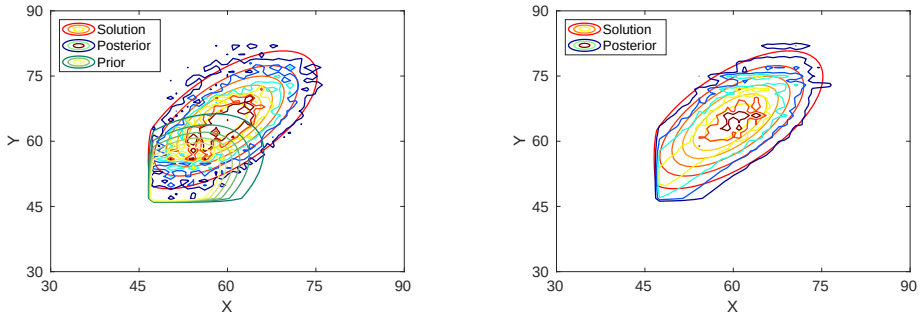
(b) Marginal of Y.

Figure 3: Fitting of the marginal distributions of X (a) and Y (b). Initial estimates are represented by the green line, KM estimates are in red, our ER solution in blue and the underlying distributions in black. Note that the bivariate truncation causes the KM estimators to be biased with respect to the true marginals.

[12]Once again, this should be seen as a plus of RUPs, as it allows for the correction of data problems, under the existence of reliable expert judgements. If no strong a priori is available, it is sufficient to set a very feeble strength of belief and let the data speak instead.

	Data	KM	ER^l	ER^h	Analytical
Mean(X)	65.447	61.699	60.642	59.780	60.661
Mean(Y)	67.689	67.937	65.310	63.960	65.509
Var(X)	31.169	59.799	50.294	57.834	51.358
Var(Y)	34.610	88.299	65.098	70.479	59.619
Corr(X,Y)	0.876	–	0.562	0.661	0.608

Table 4: Comparison of the means, variances and correlation of X and Y using: the uncensored data, the KM estimator, the ER estimator and the analytical solution, respectively. Note how omitting truncation and censoring greatly underestimates the variance. Moreover, because of the bivariate truncation, the KM estimates present a high bias and overestimate the means of both variables, especially in the case of Y, where truncation has a bigger impact. The ER estimator, on the contrary, manages to capture both marginal and joint behaviours better.



(a) ER^l scenario.

(b) ER^h scenario.

Figure 4: Contour plots for low and high strengths of belief. On the left plot, initial prior, final estimate (posterior) and true distribution are represented to show the transition from the wrong initial prior to capturing the behaviour of the true solution. On the right plot, only the analytical solution and the posterior distribution are presented for readability purposes (the prior is the same).

6.2 Bivariate case II

In the last analytical example, for generality, we would like to show the behaviour of our algorithm when the underlying distributions of A, B and C are all of different nature.

Take $A \sim \text{U}(40)$, $B \sim \text{Poi}(35)$, $C \sim \text{Gomp}(40, 6)$, $T \sim \text{Poi}(70)$, $\epsilon \sim \text{Poi}(7)$, $\epsilon_0 = 5$ and $\Delta \sim \text{Poi}(2)$. Here $\text{U}(x)$ denotes a discrete uniform distribution in the range $[0, x]$ and $\text{Gomp}(\mu, \sigma)$ a discrete Gompertz distribution with parameters μ and σ:

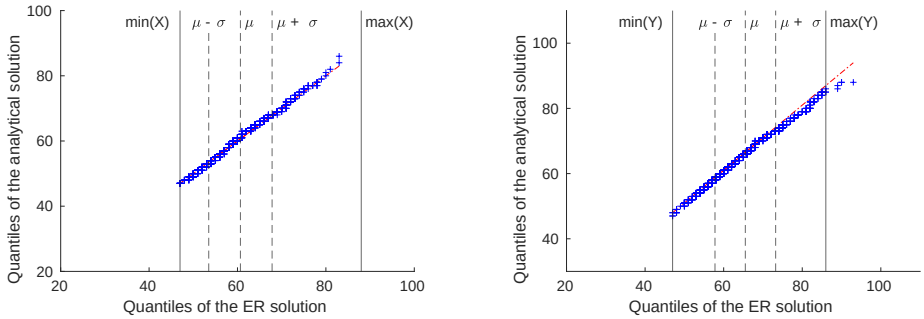
$$\text{Gomp}(x; \mu, \sigma) = 1 - \exp(e^{-\mu/\sigma}(1 - e^{x/\sigma}). \tag{60}$$

Note that since we are working with the discrete version, we are assuming that $p_X(i-1 \leq X < i) = f_X(i-1)$, where f is the Gompertz p.d.f.

The procedure is the same as in the previous example. We start by defining the initial behaviour of our RUPs through the pairs $\{\beta_j, \omega_j\}$. For this example we still use the same type of distribution for A, B and C, but instead of a Poisson distribution we use Gompertz distributions: $G^A = \text{Gomp}(25, 8)$, $G^B = \text{Gomp}(30, 7)$, $G^C = \text{Gomp}(30, 7)$, $G^T = \text{Poi}(50)$, $G^\epsilon = \text{Poi}(10)$, $\epsilon_0 = 10$. Furthermore, we assume the same relationship for the strength of belief parameters as in the previous example, so that the ER^l and ER^h scenarios are defined analogously.

As usual, we condition on the minimum uncensored values for both X and Y. Since the truncation variable has the same distribution as before we expect these values to be similar to the previous example. Indeed, this time we observe $X_{min} = 44$ and $Y_{min} = 43$, and thus we condition the distributions on survival up to these values.

We present the fitting of the ER^l marginals in Figure 6, the contour plots for both ER^l and ER^h in Figure 7, and the computed first two moments in Table 6. The conclusions are very similar to the previous example: the KM estimator is biased with respect the analytical solution due to the presence of bivariate truncation, while the ER^l estimator properly captures both marginals.



(a) QQ-plot: X.

(b) QQ-plot: Y.

Figure 5: QQ-plot comparing our ER solution with the analytical solution for X (a) and Y (b). The solid vertical lines, from left to right, correspond to the minimum and the maximum uncensored values of $\boldsymbol{i}^*$, respectively, for $i = X, Y$. The dashed lines correspond to the mean (μ) and one standard deviation (σ) of each variable.

	Value	p (%)	H_0
mean(X^l)	-0.229	46.46	Do not reject
mean(Y^l)	-0.409	25.20	Do not reject
var(X^l)	-5529	94.86	Do not reject
var(Y^l)	-6588	16.86	Do not reject

Table 5: Permutation test for samples from the ER marginal estimators. The first column gives the value of the test statistic, the second is the p-value, and the last column tells whether we accept or reject the null hypothesis. The number of permutations is 10^5.

The marginals are analyzed via QQ-plots in Figure 8, while Table 7 shows the results of the permutation test with a sample of size 10^3 from both marginals.

Even if the marginals are nicely recovered, the correlation parameter is slightly underestimated, as is clear in Table 6. This is mostly due to the convergence properties of the ER algorithm–inherited from the EM–since it converges to a local minimum that depends on the initial estimate, and therefore running the algorithm with several reasonable prior distributions is a practice we strongly suggest.

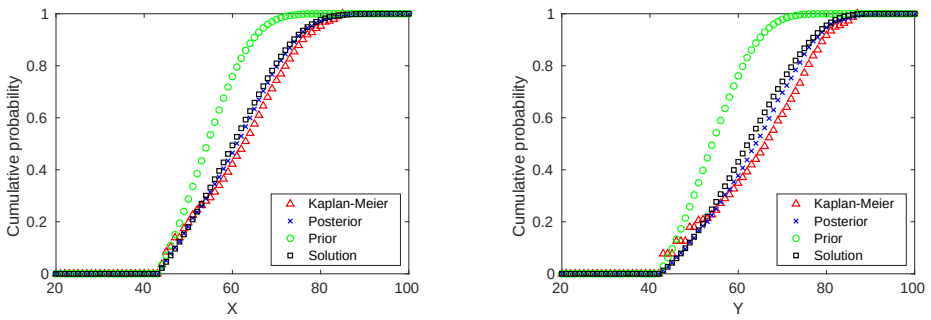
There is also a second reason for this mismatch: the support of the underlying distribution, which is considerably larger than in the previous example. A large support means that more data is needed in order to obtain a representative sample. Thus, even if our estimate fits the observations, it does not mean it will actually fit the underlying distribution. This is a clear sign of overfitting, and it is precisely one of the two situations we mentioned in Section 5, where our algorithm may improve over the original EM algorithm, by properly working with priors. For this reason we also show in Figure 7d the resulting distribution when giving a high strength belief to the a priori. In Table 6, under the column ER^h, we show the corresponding results for the moments.

6.3 A real example

We consider a Canadian data set widely used in the field of joint annuity modeling [Frees et al., 1996, Luciano et al., 2008].

The data consist of almost 15,000 couple of clients of a Canadian insurance company. Each couple has a joint annuity contract with the insurer. For each couple several pieces of information are available: date of contract, date of birth of the two annuitants, date of death if observed, or age at the end of the observation window, incomes, etc.

As in Luciano et al. [2008], we remove same-sex contracts in order to define X as the lifetime of males (most of the first annuitants are male) and Y as the lifetime of females in the couple. In the same paper they also mention that the same couple may have entered into more than one contract, and thus may appear several times on the data sheet. Therefore we remove all repeated entries so that every couple is considered only once. Finally, as in Frees et al. [1996], we condition on couples who are at least 40 years old. This leaves us with a total of 11,421 male-female couples, of which only 197 are completely uncensored. Since the period of observation is 5 years, truncation will also play a big



(a) Marginal of X. (b) Marginal of Y.

Figure 6: Fitting of the marginal distributions of X (a) and Y (b). Initial estimates are represented by the green line, KM estimates are in red, our ER solution in blue and the underlying distributions in black. Note that the bivariate truncation causes the KM estimators to be biased with respect to the true marginals.

	Data	KM	ER^l	ER^h	Analytical
Mean(X)	67.422	62.363	61.415	61.222	60.974
Mean(Y)	69.482	64.945	63.934	63.504	62.825
Var(X)	36.275	120.680	103.207	100.959	98.986
Var(Y)	41.181	151.079	117.361	115.322	110.041
Corr(X,Y)	0.892	–	0.513	0.594	0.652

Table 6: Comparison of the means, variances and correlation of X and Y using: the data set when omitting censoring and truncation effects, the KM estimator, the ER estimator and the analytical solution, respectively. Note how omitting truncation and censoring greatly underestimates the variance. Moreover, due to bivariate truncation the KM estimates present a high bias and overestimate the means of both variables, specially in the case of Y, where truncation has a bigger impact. The ER estimator, on the contrary, manages to capture the means and variances but underestimates the correlation. The results of increasing the strength of belief for the same prior, in the column ER^h, partially correct this mismatch.

role when determining the underlying distribution. For an extended analysis of this data set we refer to Frees et al. [1996] and Luciano et al. [2008].

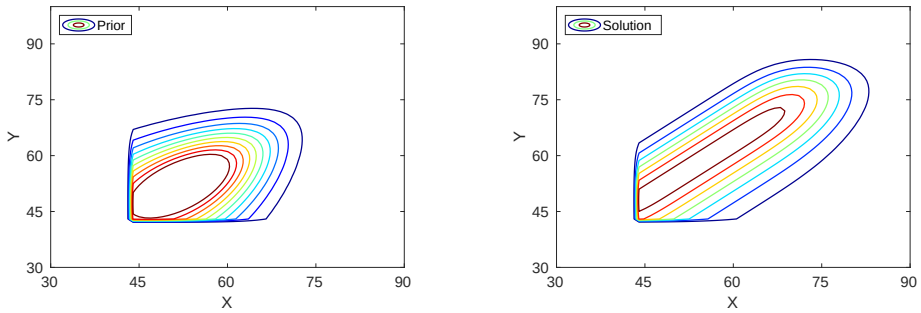
To show the importance of the a priori elicitation, we present the results obtained for two different initial estimates: one using our own invented prior knowledge and another influenced by the data. In both examples we show the impact of increasing the strength of belief.

As usual, we start by defining the initial behaviour of our processes through the pairs $\{\beta_j, \omega_j\}$. For the first example we chose: $G^A = \text{Poi}(35)$, $G^B = \text{Poi}(40)$, $G^C = \text{Poi}(40)$, $G^T = \text{Poi}(80)$, $G^\epsilon = \text{Poi}(40)$, $\epsilon_0 = 40$, where ϵ_0 was inferred from the maximum difference in ages in the data set. With this prior, males and females have the same average lifetime, which is around 75 years, with a standard deviation of almost 9 years. We also assume that the average difference in their ages is 0, with a standard deviation of 6 years. Moreover, we consider the same strength of belief scenarios as in the previous examples–check again Table 1 if needed–apart from the reinforcement[13] of ER^h, for which we now take $r = 10$.

In Figures 9a, 9b and 10c we show the marginals and joint distributions, respectively, for the case where our strength of belief in the a priori is low. Note that we clearly underestimated the average lifetimes, especially for the females, and the correlation slightly decreases from 0.47 to 0.40. Moreover, in Figure 10c we can observe small contours in which the age difference is particularly large. In those cases we cannot be sure if the observation corresponds to a married couple or a parent-child relationship.

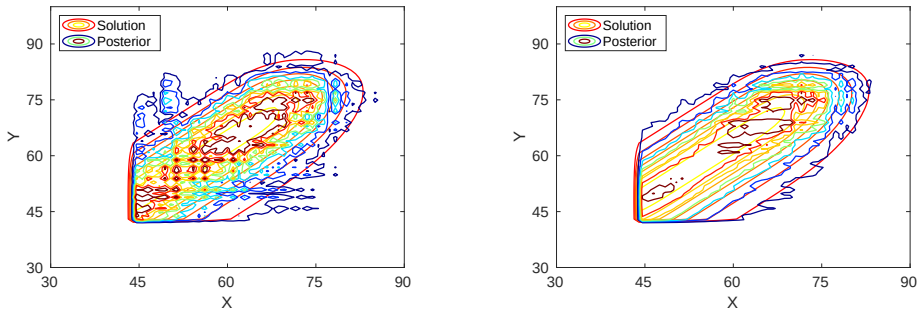
Then, we show the same example when we assume our prior to be strong and correct. This translates into giving a higher value to the strength of belief, i.e. scenario ER^h. The resultant distribution is presented in Figure 10e.

[13]Notice that, given the same values for belief and reinforcement, a larger sample size obviously means more weight on the reinforcement side.



(a) Prior distribution.

(b) Underlying distribution.



(c) ER^l scenario.

(d) ER^h scenario.

Figure 7: Contour plots for low and high strengths of belief. We also show the prior distribution (a) and the original distribution (b) for comparison purposes. On the lower-left plot (c) we show the posterior distribution for low strength of belief and the lower-right plot (d) the posterior distribution when the strength of belief is high. Note how for low strengths of belief many outlier contours appear while by forcing a specific shape with the belief parameters we avoid this behaviour.

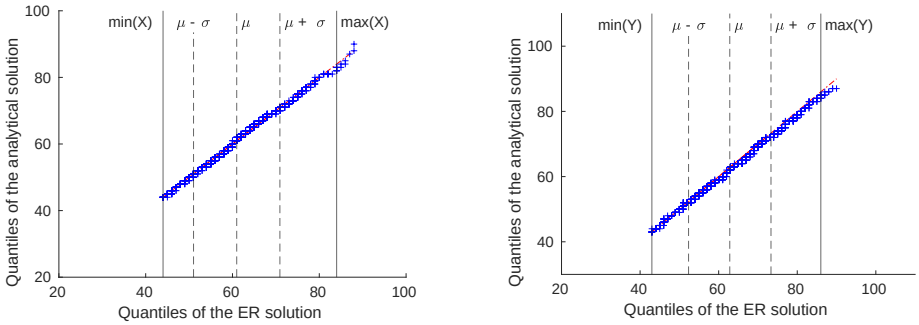
	Value	p (%)	H_0
mean(X^l)	0.135	76.02	Do not reject
mean(Y^l)	0.743	12.58	Do not reject
var(X^l)	8273	73.56	Do not reject
var(Y^l)	9354	11.01	Do not reject

Table 7: Permutation test for samples from the ER^l marginal estimators. The first column gives the value of the test statistic, the second column is the p-value, and the last column is whether we accept or reject the null hypothesis H_0 that the distributions agree with the analytical one. Note that test fails to reject the null hypothesis for both variables at the 5% level. As before, the number of permutations is 10^5.

In Table 8, we show the first two moments. Column *ER*^{l} accounts for the low strength of belief case, while *ER*^{h} deals with the opposite situation. Please observe that, for strong beliefs, not only the correlation has barely changed from the initial estimate (from 0.47 to 0.46), but also the shape of the contours is much less erratic than in the previous case, which makes this distribution more tractable.

Let us now consider a situation in which we try to use a more objective prior, elicited by looking at the data. Notice that, for a Bayesian purist, in this case we should not speak about proper prior and posterior distributions, since the a priori is contaminated by data [de Finetti, 2017, Galavotti, 2001, Galavotti et al., 2008]. However, even in this case, the Bayesian nature of the model can be exploited, and the strength of belief parameters can be used to obtain smoother contours.

Given that we previously underestimated the average lifetimes obtained from the data, our initial estimates this time are: $G^A = \text{Poi}(40)$, $G^B = \text{Poi}(45)$, $G^C = \text{Poi}(45)$, $G^T = \text{Poi}(80)$, $G^\epsilon = \text{Poi}(40)$, $\epsilon_0 = 40$, so that the sample means are for example replicated.



(a) QQ-plot: X. (b) QQ-plot: Y.

Figure 8: QQ-plot comparing our ER solution with the analytical solution for X (a) and Y (b). The solid vertical lines, from left to right, correspond to the minimum value of $\boldsymbol{T}^i$ and the maximum value of $\boldsymbol{i}^*$, respectively, for $i = X, Y$. The dashed lines correspond to the mean (μ) and one standard deviation (σ) of each variable.

	Data	KM	ER^l	ER^h	ER2^l	ER2^h
Mean(X)	74.514	81.581	81.901	81.785	81.952	82.000
Q1(X)	70.000	75.000	76.000	75.000	76.000	76.000
Median(X)	74.000	83.000	83.000	83.000	83.000	83.000
Q3(X)	79.000	90.000	89.000	89.000	89.000	89.000
Mean(Y)	74.011	86.989	85.491	84.637	85.246	85.432
Q1(Y)	69.000	82.000	80.000	79.000	80.000	80.000
Median(Y)	73.000	89.000	87.000	86.000	87.000	86.000
Q3(Y)	79.000	94.000	92.000	91.000	92.000	92.000
Var(X)	52.049	124.393	117.075	112.088	113.935	123.091
Var(Y)	61.707	99.945	76.866	85.410	77.543	89.931
Corr(X,Y)	0.820	–	0.401	0.461	0.432	0.398

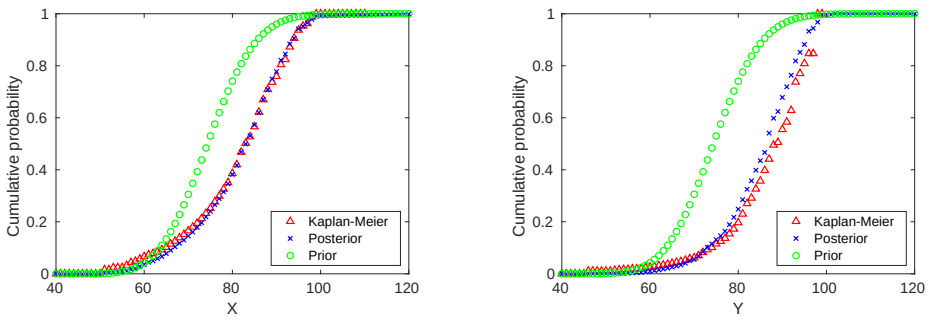
Table 8: Comparison of the means, medians, first and third quartiles, variances and correlation of X and Y using: the uncensored data set, the KM estimator, the ER estimator for the first example under low and high strengths of belief, and the ER estimator for the second example again for low and high strengths of belief, respectively. Note that both examples return different distributions even for low beliefs, which hints towards the existence of local minima. These differences, however, are small and show that there is not a big impact in the posterior distribution for similar priors.

In Figures 9c, 9d and 10d we show the results for low strengths of belief. We observe that there are barely any differences with the previous example for the case of low belief in our prior, except, perhaps, for the correlation parameter–now it is 0.43–and the areas of smaller probability. This is an expected result since, for low strengths of belief, the algorithm tries to fit the data as good as possible. The only way the results could be different is when the initial distributions lead to different local minima.

In Figure 10f, conversely, the posterior distribution when assuming strong belief in our prior is shown. As before, the first two moments are presented in the columns *ER2*^{l} and *ER2*^{h} of Table 8. Notice that, according to Figure 10f, the posterior distribution has barely distanced from the prior. In the previous examples so far, the algorithm always reached a compromise between the prior and the data when increasing the stength of belief. However, since this time the prior was specifically chosen by taking the data into consideration, the resultant posterior is barely affected by the data.

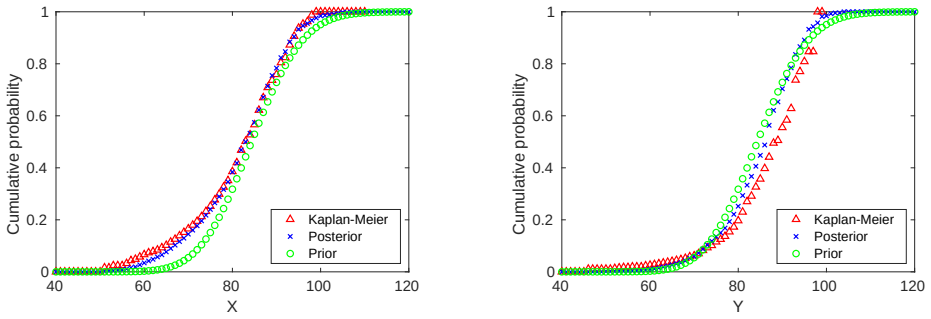
7 Conclusions

In this work we have discussed estimation techniques for Reinforced Urn Processes (RUPs), a flexible class of models in the Bayesian nonparametric literature, to deal with possibly right-censored observations [Muliere et al., 2000]. We have considered both the standard univariate setting, and the one-factor bivariate construction of Bulla et al. [2007],



(a) Marginal of X.

(b) Marginal of Y.



(c) Marginal of X.

(d) Marginal of Y.

Figure 9: Fitting of the marginal distributions of X and Y using the Canadian data set for the first example ((a) and (b),respectively) and for the second example ((c) and (d)). Initial estimates are represented by the green line, KM estimates are in red and our ER solution in blue.

discussing the usefulness of prior elicitation and reinforcement, two constituent features of every RUP. Moreover, we have extended some basic properties of the RUP construction, showing the effects of left-truncation in the data. In particular, we show that conjugacy is maintained.

As the main goal of our work, we have provided an explicit Expectation-Maximization (EM) approach, and offered an extension which exploits the reinforcement mechanism of Polya urns, proposing the so-called Expectation-Reinforcement (ER) algorithm. To the best of our knowledge, this is the first systematic attempt to offer estimation algorithms for RUPs, which are generally treated via simulation-based techniques, like MCMC.

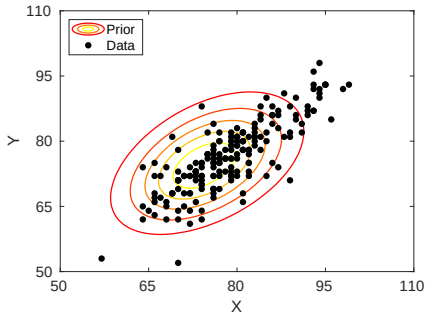
The performances of both the EM and the ER algorithms have been tested using artificial and real data, showing their superiority with respect to other common alternatives like the Kaplan-Meier estimator, especially in the bivariate setting. For what concerns the ER algorithm, the possibility of playing with priors and reinforcement can be a very important point of strength, when dealing with complex data sets, with missing observations, problems of representativeness and fat tails.

Future lines of work involve extending the one-factor model of Bulla et al. [2007], to cope with multivariate situations. In keeping linear dependence, this extension is definitely straightforward in the absence of censoring and truncation, but it requires much more work in a more general setting. Furthermore, thinking about other forms of dependence, far from linearity, can make things extremely complicated.

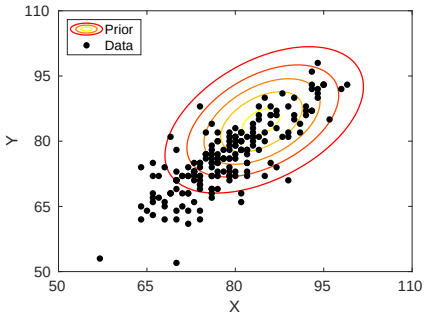
Also computationally, increasing the dimensionality of the problem may generate non trivial questions to be solved. Nevertheless, we believe that the computational burden could be decreased by implementing schemes that increase the convergence of the EM algorithm, and using HPC techniques such as parallelization.

Acknowledgements

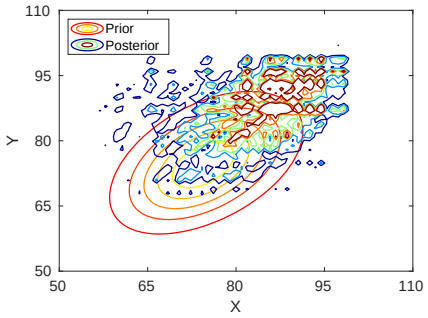
This research has been generously financed by the European Union, under the H2020-EU.1.3.1. MSCA-ITN-2018 scheme, Grant 813261.



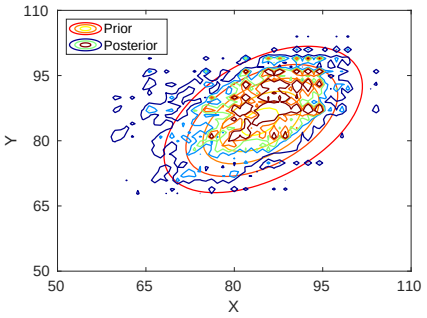
(a) Prior first example.



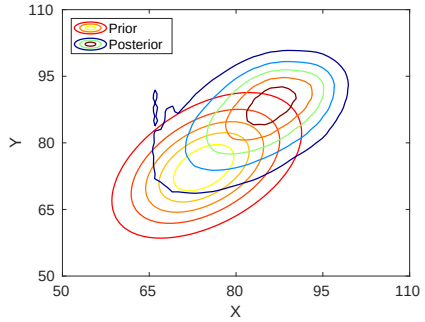
(b) Prior second example.



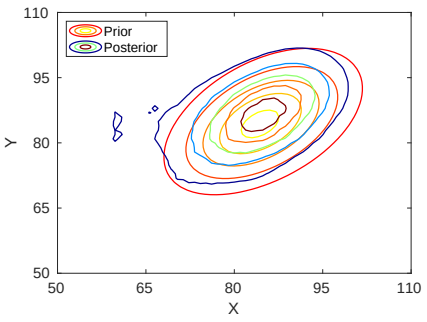
(c) Low strength of belief.



(d) Low strength of belief.



(e) High strength of belief.



(f) High strength of belief.

Figure 10: Contour plot for the empirical data set. In the upper row we show the prior distributions for each example, as well as a scatter plot of the uncensored samples available in the Canadian data set. Note how the uncensored data is clearly not representative enough of the total population since the first prior seems to be a better fit. In the middle row we present a comparison with the posterior distribution with low beliefs, and in the lower row the same results when the beliefs are high. Similar to what happens in the analytical case, increasing the belief parameters works both as a bridge between initial estimate and observations and as a smoothing mechanism.

References

Emanuele Amerio, Pietro Muliere, and Piercesare Secchi. Reinforced Urn Processes for Modeling Credit Default Distributions. *International Journal of Theoretical and Applied Finance*, 7(4):407–423, 2004.

Katrien Antonio, Andrei Badescu, Lan Gong, Sheldon Lin, and Roel Verbelen. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulletin*, 45(3):729–758, 2015.

Rose D. Baker. Two permutation tests of equality of variances. *Statistics and Computing*, 5(4):289–296, 1995.

Davis Blackwell and James B. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1(2): 353–355, 1973.

Nizar Bouguila and Djemel Ziou. A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized Dirichlet mixture. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 15:2657–68, 10 2006. doi: 10.1109/TIP.2006.877379.

R. A. Boyles. On the convergence of the em algorithm. *Journal of the Royal Statistical Society. Series B*, 45(1):47–50, 1983.

Paolo Bulla, Pietro Muliere, and Steven G. Walker. Bayesian Nonparametric Estimation of a Bivariate Survival Function. *Statistica Sinica*, 17(3):427–444, 2007.

G. Campbell and A. Földes. Large sample properties of nonparametric statistical inference. In B. V. Gnredenko, M. L. Puri, and I. Vineze, editors, *Nonparametric statistical inference*, pages 103–122. North-Holland, Amsterdam, 1982.

Dan Cheng and Pasquale Cirillo. A Reinforced Urn Process Modeling of Recovery Rates and Recovery Times. *Journal of Banking & Finance*, 96:1–17, 2018.

Dan Cheng and Pasquale Cirillo. An urn-based nonparametric modeling of the dependence between pd and lgd with an application to mortgages. *Risks*, 7(3):76, 2019.

Pasquale Cirillo, Jürg Hüsler, and Pietro Muliere. A Nonparametric Urn-based Approach to Interacting Failing Systems with an Application to Credit Risk Modeling. *International Journal of Theoretical and Applied Finance*, 41:1–18, 2010.

Pasquale Cirillo, Jürg Hüsler, and Pietro Muliere. Alarm Systems and Catastrophes from a Diverse Point of View. *Methodology and Computing in Applied Probability*, 15(4):821–839, 2013.

Robert J. Connor and James E. Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

D. R. Cox and D. Oakes. *Analysis of survival data*. Chapman & Hall, 1984.

Dorota M. Dabrowska. Kaplan-meier estimate on the plane. *The Annals of Statistics*, 16(4):1475–1489, 1988.

B. de Finetti. *Theory of Probability: A critical introductory treatment*. Wiley Series in Probability and Statistics. Wiley, 2017. ISBN 9781119286370.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

Kjell Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, 2 (2):183–201, 1974.

Yarong Feng, Xing Chen, Liyi Jia, Xiruo Song, and Hosam M. Mahmoud. Estimating the Pólya process. *Communications in Statistics - Theory and Methods*, 46(19):9397–9406, 2017. doi: 10.1080/03610926.2016.1208242.

Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

Mário A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:381–396, 2000.

Edward W. Frees, Jacques Carriere, and Emiliano Valdez. Annuity valuation with dependent mortality. *The Journal of Risk and Insurance*, 63(2):229–261, 1996.

Maria Carla Galavotti. Subjectivism, objectivism and objectivity in bruno de finetti's bayesianism. In David Corfield and Jon Williamson, editors, *Foundations of Bayesianism*, pages 161–174. Springer, Dordrecht, 2001.

Maria Carla Galavotti, H. Hosni, B. de Finetti, and A. Mura. *Philosophical Lectures on Probability: collected, edited, and annotated by Alberto Mura*. Synthese Library. Springer Netherlands, 2008. ISBN 9781402082023.

Paolo Giudici, Pietro Muliere, and Maura Mezzetti. Mixtures of Dirichlet Process Priors for Variable Selection in Survival Analysis. *Journal of Statistical Planning and Inference*, 17:867–878, 2003.

Phillip Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer New York, 1994.

Svetlana Gribkova and Olivier Lopez. Non-parametric copula estimation under bivariate censoring. *Scandinavian Journal of Statistics*, 42(4):925–946, 2015.

Nils Lid Hjort, Chris Holmes, Peter Mueller, and Stephen G. Walker. *Bayesian Nonparametrics*. Cambridge University Press, 2010.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Application in R*. Springer, 2013.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Chloé Le Goff Line and Soulier Philippe. Parameter estimation of a two-colored urn model class. *The International Journal of Biostatistics*, 13(1), 2017.

Sun Liuquan and Ren Haobao. Bivariate estimation with left-truncated data. *Acta Mathematicae Applicatae Sinica*, 17 (2):145–156, 2001.

Olivier Lopez. A generalization of kaplan-meier estimator for analyzing bivariate mortality under right-censoring and left-truncation with applications to model-checking for survival copula models. *Insurance: Mathematics and Economics*, 51(3):505–516, 2012.

Elisa Luciano, Jaap Spreeuw, and Elena Vigna. Modelling stochastic mortality for dependent lives. *Insurance: Mathematics and Economics*, 43:234–244, 2008.

D. Lynden-Bell. A method of allowing for known observational selection in small samples applied to 3cr quasars. *Monthly Notices of the Royal Astronomical Society*, 155(1):95–118, 1971.

Hosam M. Mahmoud. *Pólya urn models*. Chapman & Hall/CRC, 2008.

Riccardo Marcaccioli and Giacomo Liva. A polya urn approach to information filtering in complex networks. *Nature Communications*, 10(745), 2019. doi: 10.1038/s41467-019-08667-3.

G. J. McLachlan and P. N. Jones. Fitting Mixture Models to Grouped and Truncated Data via the EM Algorithm. *Biometrics*, 44(2):571–578, 1988.

Alexander J. McNeil, Ruediger Frey, and Paul Embrechts. *Quantitative Risk Management*. Princeton University Press, Princeton, 2015.

Pietro Muliere, Piercesare Secchi, and Stephen G. Walker. Urn Schemes and Reinforced Random Walks. *Stochastic Processes and their Applications*, 88(1):59–78, 2000. ISSN 03044149.

Pietro Muliere, Piercesare Secchi, and Stephen G. Walker. Reinforced Random Processes in Continuous Time. *Stochastic Processes and their Applications*, 104(1):117–130, 2003.

Swagata Nandi and Isha Dewan. An EM algorithm for estimating the parameters of bivariate Weibull distribution under random censoring. *Computational Statistics & Data Analysis*, 54(6):1559–1569, 2010.

Dan Nettleton. Convergence properties of the EM algorithm in constrained parameter spaces. *The Canadian Journal of Statistics*, 27(3):639–648, 1999.

Stefano Peluso, Antonietta Mira, and Pietro Muliere. Reinforced Urn Processes for Credit Risk Models. *Journal of Econometrics*, 184(1):1–12, 2015.

R. C. Pruitt. Strong consistency of self-consistent estimators: general theory and an application to bivariate survival analysis. Technical Report 543, University of Minnesota, 1991a.

R. C. Pruitt. On negative mass assigned by the bivariate kaplan-meier estimator. *Annals of Statistics*, 19:443–453, 1991b.

R. C. Pruitt. Small sample comparisons of six bivariate survival curve estimators. *Journal of Multivariate Analysis*, 45 (3):147–167, 1993.

W. F. Rosenberger and J. M. Lachin. *Randomization in Clinical Trials: Theory and Practice*. Wiley Series in Probability and Statistics. Wiley, 2004. ISBN 9780471654070.

George Lennox Sharman Shackle. *Uncertainty in Economics and Other Reflections*. Cambridge University Press, Cambridge, 1955.

Pao-Sheng Shen and Ya-Fang Yan. Nonparametric estimation of the bivariate survival function with left-truncated and right-censored data. *Journal of Statistical Planning and Inference*, 138(12):4041–4054, 2008.

Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House, 2007.

Wei-Yann Tsai, Nicholas P. Jewell, and Mei-Cheng Wang. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74(4):883–886, 1987.

Bruce W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295, 1976.

M. J. van der Laan. Modified EM-estimator of the Bivariate Survival Function. *Mathematical Methods of Statistics*, 3 (3):213–243, 1994.

Ravi Varadhan and Christophe Roland. Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353, 2008.

Stephen G. Walker and Pietro Muliere. Beta-Stacy Processes and a Generalization of the Pólya-Urn Scheme. *The Annals of Statistics*, 25(4):1762–1780, 1997.

Mei-Cheng Wang. Product limit estimates: a generalized maximum likelihood study. *Communications in Statistics - Theory and Methods*, 16(11):3117–3132, 1987.

Mei-Cheng Wang, Nicholas P. Jewell, and Wei-Yann Tsai. Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics*, 14(4):1597–1605, 1986.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

M. Woodroofe. Estimating a distribution function with truncated data. *The Annals of Statistics*, 13(1):163–177, 1985.

C. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.