About the estimation of reinforced urn processes under left-truncation and right-censoring

A Preprint

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

Pasquale Cirillo
ZHAW School of Law and Management
Zurich University of Applied Sciences
Switzerland
ciri@zhaw.ch

Cornelis W. Oosterlee
Mathematical Institute, Utrecht University
The Netherlands
c.w.oosterlee@uu.nl

November 29, 2021

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 using artificial and actual data.

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

1 Introduction

Real-life data is often censored and/or truncated. For example we have right-censoring, when, in a medical study, patients are observed over a limited period of time, and the event of interest—next stage of a disease or even death—may occur after the observation period. And we can also have left-truncation, because many patients can only be diagnosed when the disease is already in an advanced stage, and knowledge about the first evolutions is not given. Censoring and truncation are quite common in survival studies, education studies, engineering and risk management [Angrist et al., 2006, Cheng and Cirillo, 2018, Klein and Moeschberger, 2003, Shen and Yan, 2008, Antonio et al., 2015], among others.

In the univariate case, truncation and censoring have been extensively studied (see for example Wang [1987] for a very nice analysis of these phenomena on likelihood estimation), but 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, e.g. 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, possibly generating negative probabilities. Interesting works are also 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, which 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. Shen and Yan [2008] showed that bivariate truncation complicates things further. In such a case, the univariate Kaplan-Meier (KM) estimator [Kaplan and Meier, 1958] should not be used for the marginal survival functions, since it is not consistent. Thus not only the joint distribution, but also the marginals become difficult to handle.

Since left-truncation and right-censoring fall under the umbrella of incomplete data, many authors have opted for an alternative approach, using the (Expectation-Maximization) EM algorithm of Dempster et al. [1977]. Some relevant results in the univariate framework are available in Pruitt [1991a], van der Laan [1994], Antonio et al. [2015] and references therein. The literature is far more limited for bivariate distributions: Nandi and Dewan [2010] is the first work to approach the problem of bivariate modelling under right-censoring through EM. And, to the best of our knowledge, nobody has so far proposed an extension of the EM algorithm to include bivariate left-truncation.

In this work, we propose a nonparametric implementation of the EM algorithm that captures both bivariate right-censoring and left-truncation [1]. To do so we rely on the Reinforced Urn Process (RUP) of Muliere et al. [2000], for both the univariate and the bivariate settings. For the latter, we focus on the bivariate RUP (B-RUP) model of Bulla et al. [2007], to deal with the non-negative linear dependence between two variables. However, as it shall be clear later, our algorithms can be adapted to more general constructions, as far as they are build on RUPs.

RUPs are neutral-to-the-right processes [Doksum, 1974], and they have been employed in many applications involving right-censoring, for example in the context of Wrong-Way-Risk modelling [Cheng and Cirillo, 2019], or in annuity pricing [Souto Arias and Cirillo, 2021]. However, they have not yet been used to model left-truncation. On top of that, the study and use of RUPs has been so far restricted to the Bayesian community, and all applications have relied on extensive simulations–mainly Markov Chain Monte Carlo as in Bulla et al. [2007] or Peluso et al. [2015]. In our opinion, this is mainly due to the lack of reliable estimation tools for the (hyper-)parameters of the different models[2]. 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, e.g. Galavotti [2001], Galavotti et al. [2008], Jeffreys [1946]. However, for researchers who do not feel at ease with subjectivity and prior elicitation, limitations in the estimation of the model parameters directly from the data can be a deterrent. Our EM algorithm offers a solution.

This said, we also believe that it is important to valorize one of the main features of RUPs: the possibility of combining expert judgements, in the form of some a priori, and empirical data. This can be of great help for those applications in which data can be imprecise or missing, but experts have important "gut feelings" (for a discussion see Cheng and Cirillo [2018]). That is why, besides a more standard EM algorithm, we also propose a novel Expectation-Reinforcement (ER) algorithm, which exploits the natural reinforcement mechanism of RUPs to complement the EM part.

The structure of the paper develops as follows. In Section 2 we briefly revisit the concepts of right-censoring and left-truncation, and how they relate to the variables of interest in both the univariate and bivariate cases. Section 3 summarizes the theory of Reinforced Urn Processes, as well as the bivariate construction (B-RUP) of Bulla et al. [2007]. The EM algorithm for RUPs under left-truncated and right-censored data is described in Section 4, while Section 4.4 discusses the new ER algorithm. Simulation results and an application on actual data are presented in Section 5 to discuss performances, while Section 6 concludes the paper.

2 Left-truncation and right-censoring

Let $\mathbf{X} = (X_1, ..., X_n)$ and $\mathbf{C} = (C_1, ..., C_n)$ be identically and independently distributed (i.i.d) 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 $\mathbb{1}_{\{.\}}$ the indicator function. That is, we observe the minimum between the censoring variable and our target variable, plus and indicator telling us which of the two we observe.

Let $\mathbf{T} = (T_1, ..., T_n)$ be another i.i.d. sequence with distribution F_T, independent of F_X. 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 for $x_i < t_i$ we do not observe anything, we cannot even realize its existence, that is, there is no information in our data about x or t for $x < t$. 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, the truncated sample will be highly biased with respect to the original

[1] In the following, when referring to "censoring" and "truncation," we always imply "right-censoring" and "left-truncation", unless explicitly stated otherwise.

[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 indeed a minority.

underlying distributions. Such a bias is so relevant that, according to Wang [1987], truncated data can also be classified as selection-biased data.

Of course, there is also the case when 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. 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[3] between 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

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

Equation (1) reduces to the KM estimator in the absence of truncation [Kaplan and Meier, 1958], and to the product-limit estimator of Lynden-Bell [1971] without censoring.

Let us now consider the bivariate situation. Following Shen and Yan [2008], we assume that the joint survival function S_{XY} of the variables (X, Y) is independent from the truncation and censoring variables (T^X, C^X, T^Y, C^Y). Observations consist of two triplets (x^*, t^X, δ^X) and (y^*, t^Y, δ^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 (1). However, when there exists dependence, the likelihood must be modelled jointly, and under left-truncation and right-censoring it takes the form:

$$L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\delta^X}, \boldsymbol{\delta^Y}, \mathbf{T}^X, \mathbf{T}^Y) = \sum_{i=1}^{n}[\log(\mathbb{P}^*(x_i^*, y_i^*|\delta_i^X, \delta_i^Y)) - \log(\mathbb{P}(X \geq t_i^X, Y \geq t_i^Y))], \tag{2}$$

with

$$\mathbb{P}^*(x, y|\delta^X, \delta^Y) = \begin{cases} \mathbb{P}(X = x, Y = y) & \text{if } \delta^X = 1 \text{ and } \delta^Y = 1, \\ \mathbb{P}(X > x, Y = y) & \text{if } \delta^X = 0 \text{ and } \delta^Y = 1, \\ \mathbb{P}(X = x, Y > y) & \text{if } \delta^X = 1 \text{ and } \delta^Y = 0, \\ \mathbb{P}(X > x, Y > y) & \text{if } \delta^X = 0 \text{ and } \delta^Y = 0. \end{cases} \tag{3}$$

For ease of notation, from now on we will write $L(\mathbf{X}, \mathbf{Y})$ when referring to the bivariate likelihood in Equation (2), with the conditioning on censoring and truncation always implied.

With the acronym LTRC, we will indicate left-truncated and/or right-censored observations.

2.1 Assumptions for modelling

To make notation and computations easier, while preserving enough generality and applicability, we will make some assumptions on censoring and truncation. However, 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 employed under different assumptions.

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. This situation arises for example when studying coupled lifetimes, where T^X and T^Y denote the age at which each individual enters the study [Frees et al., 1996, Souto Arias and Cirillo, 2021]. 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. $\epsilon - \epsilon_0$, with ϵ_0 a maximum threshold in the age difference. This assumption is heavily inspired by experiments where the truncation variable plays a temporal role[4].

Second, we set $C^i = T^i + \Delta$, for $i \in \{X, Y\}$. Going back to the coupled lifetimes example, both members of the couple will be monitored exactly for the same amount of time in case censoring occurs, which explains using the same r.v. for C^X and C^Y. The 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. This second assumption is also implicitly giving a temporal meaning to the censoring variable, and it basically means that the target variable grows linearly with time.

[3] A similar situation occurs, for example, in lifetime follow-up studies [Klein and Moeschberger, 2003], where T is the age at which an individual joins the study and C is the age at which they drop out. Since it is not possible to drop out 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, again, since most of the times T has a temporal interpretation [Klein and Moeschberger, 2003], the other form is here preferred.

3 Univariate and Bivariate Reinforced Urn Processes

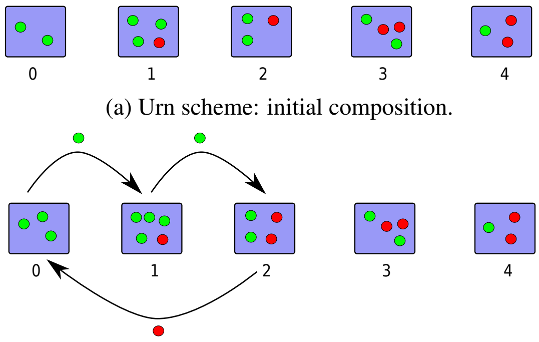
Before recalling the mathematical details of the Reinforced Urn Process of Muliere et al. [2000], we give an intuitive representation of this process using a sequence of Polya urns [Mahmoud, 2008], in the two-color case.

3.1 Urn representation

Assume we have $M + 1$ Polya urns, where the j-th urn U_j, $j = 0, 1, ..., M$, initially contains $\omega_j > 0$ green (G) balls and $\beta_j > 0$ red (R) balls. The only exception is urn U_0, which only has green balls. The dynamics of the RUP are governed by the following rules:

1. The process starts from Urn 0, which is the first urn to be sampled with reinforcement.
2. Every time we pick a ball, we note the color, and we reinforce the urn, that is we put the ball back together with an extra ball of the same color. This increases the probability of picking that color again in the future.
3. If the color of the ball is green, we move forward to the next urn, and repeat from step 2. On the other hand, if the color is red we go back to Urn 0, and the process starts anew.

Notice that from Urn 0 we necessarily move to Urn 1, while from all the other urns, we can move forward or jump back to Urn 0.



(a) Urn scheme: initial composition.

(b) Urn scheme: updated composition after the trajectory $(\bullet G, \bullet G, \bullet R)$.

Figure 1: Representation of the RUP as a series of Pólya urns. After each sampling the urns are updated in a way that reinforces the probability of a given sampling.

In Figure 1b we see an example of a possible trajectory $(\bullet G, \bullet G, \bullet R)$ and the resulting urn composition after each sampling. Notice that the probability of observing the same path in the next cycle has increased, thanks to the reinforcement mechanism, which allows for learning.

If the data we want to model includes right-censoring and left-truncation, the previous graphic can be modified as follows. In case of a right-censored observation, we do not include the red ball, with everything else unchanged. If the sample is left-truncated with value j, we start drawing from urn U_j, instead of U_0. It is clear that in this way we can reproduce Equation (4).

In the simple example above we have assumed that reinforcement (the number of extra balls added after each drawing) is unitary for the sake of clarity, but 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, i.e. the initial compositions of the urns.

3.2 Some more formality

We start by defining the mathematical backbone of the RUP: the beta-Stacy process of Walker and Muliere [1997], to which we refer for all detail.

The beta-Stacy process (BSP) is a random distribution that can sample discrete (discrete BSP) or continuous (continuous BSP) distributions. It can be seen as a generalization of the Dirichlet process [Ferguson, 1973], and it represents a very flexible tool for Bayesian nonparametrics [Hjort et al., 2010]. In what follows we focus on the discrete BSP.

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 LTRC sample $(\boldsymbol{X}_n^*, \boldsymbol{T}_n, \boldsymbol{\delta}_n)$, with $\boldsymbol{X}_n^* = \{x_n^*, n \geq 1\}$, the sequence $\boldsymbol{X}_n = \{x_n, n \geq 1\}$ is a RUP with reinforcement r if

$$\hat{S}(x) = \mathbb{P}(X_{n+1} > x | \boldsymbol{X}_n^*, \boldsymbol{T}_n, \boldsymbol{\delta}_n) = \prod_{j=0}^{x} \left[1 - \frac{\beta_j + r \cdot m_j^*(\boldsymbol{X}_n^*, \boldsymbol{\delta}_n)}{\beta_j + \omega_j + r \cdot s_j(\boldsymbol{X}_n^*, \boldsymbol{T}_n)} \right], \tag{4}$$

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

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 LTRC data[5]. Note that, although Equation (4) allows to work with (discretized) float numbers, we focus only on positive integers, so that the dummy variable j in the product takes jumps of size one.

Using the previous urn example, we can notice that $\hat{S}(x)$ in Equation (4) corresponds to the probability of selecting at least x consecutive green balls, starting from urn U_0. The couples $\{\beta_j, \omega_j\}$ play the role of the initial compositions of the urns, while the functions $m_j^*(\boldsymbol{x}_n, \boldsymbol{d}_n)$ and $s_j(\boldsymbol{x}_n, \boldsymbol{t}_n)$ determine the extra number of green and total balls added to each urn, respectively.

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

$$\frac{\beta_j}{\beta_j + \omega_j} = \frac{G(j) - G(j-1)}{1 - G(j-1)}, \qquad j \in \mathbb{N} \tag{5}$$

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

Although the choice for the couples $\{\beta_j, \omega_j\}$ in terms of $G(\cdot)$ is not unique if we simply want Equation (5) to hold, here we follow the steps of Walker and Muliere [1997], assuming

$$\beta_j = c_j G(\{j\}), \qquad \omega_j = c_j(1 - G(j)), \qquad c_j \in \mathbb{R}^+, j \in \mathbb{N}, \tag{6}$$

with c_j denoting the strength of belief in our prior knowledge and $G(\{j\}) = \mathbb{P}_G(X = j)$. The name "strength of belief" comes from the fact that, for high values of c_j, it will be difficult for the posterior distribution to deviate from the a priori, unless one has large amounts of data. On the contrary, when $c_j \to 0$, we recover the KM estimator of Cox and Oakes [1984] from Equation (4), and the a priori plays no role whatsoever.

Finally, observe that the roles of the strength of belief parameters c_j and of the reinforcement parameter r are actually opposite. 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 (B-RUP) of Bulla et al. [2007]

Assume we observe couples of data of the form $((\boldsymbol{X}_n^*, \boldsymbol{T}_n^X, \boldsymbol{\delta}_n^X), (\boldsymbol{Y}_n^*, \boldsymbol{T}_n^Y, \boldsymbol{\delta}_n^Y))$, 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 variables of interest X and Y. As before $\boldsymbol{T}_n^X$ and $\boldsymbol{T}_n^Y$ are the truncation processes for X and Y, respectively.

A flexible yet simple way of modelling the dependence between X and Y is to consider the one-factor construction of Bulla et al. [2007], basen on three independent components: one common and two idiosyncratic factors for X and Y.

[5] The RUP was originally defined only for right-censored observations [Muliere et al., 2000, Walker and Muliere, 1997]. In Appendix A we prove that the RUP is also conjugated to left-truncated observations, and Equation (4) still defines a proper RUP.

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 left-truncation and right-censoring. Now, set

$$\begin{aligned} X &= A + B, \\ Y &= A + C. \end{aligned} \tag{7}$$

The dependence between X and Y clearly relies entirely on A, hence, conditioned on this common process, X and Y are independent. A straightforward calculation yields

$$\text{Cov}(X_{n+1}, Y_{n+1} | \boldsymbol{A}_n, \boldsymbol{B}_n, \boldsymbol{C}_n) = \text{Var}(A_{n+1} | \boldsymbol{A}_n), \quad n \geq 1, \tag{8}$$

where $\boldsymbol{A}_n = \{A_n, n \geq 1\}$ is a sample of size n from A, and similarly for $\boldsymbol{B}_n$ and $\boldsymbol{C}_n$. Notice that the dependence between X and Y is therefore linear and non-negative.

Bulla et al. [2007] showed that the sequence $\{(X_n, Y_n), n \geq 1\}$ is exchangeable, thus, by the de Finetti theorem, there exists a joint random distribution function F_{XY}, conditionally on 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}. For our purposes, the most relevant feature of the model is the one-factor construction. As we will see in the next section, this will allow us to derive a simple and efficient iterative method to estimate the parameters of the model.

4 The Expectation-Maximization (EM) and the Expectation-Reinforcement (ER) Algorithms

The EM algorithm is the de-facto-standard when dealing with incomplete data. It was originally introduced in Dempster et al. [1977], although particular instances of the algorithm had already been developed in Turnbull [1976]. From a mathematical point of view, the algorithm requires the computation of the expectation of the complete likelihood $\log f(\mathbf{x}|\theta)$ at each iteration, conditioned on the observed incomplete data $\mathbf{y}$, and the estimates of the parameters from the previous iteration. Given the many meanings of the term “incomplete", the EM algorithm comes in many flavours in the literature [McLachlan and Krishnan, 2008]. Here, as said before, when we talk about incomplete data, we always refer to observations that suffer from left-truncation and/or right-censoring.

In the following, to improve readability, we will use the following short notation:

$$p_X(x) \coloneqq \mathbb{P}(X = x|\theta), \tag{9}$$

and

$$p_X^{[k]}(x) \coloneqq \mathbb{P}(X = x|\theta^{[k]}), \tag{10}$$

where θ is the vector of parameters (of the B-RUP) and the superscript k denotes a quantity computed at the k-th iteration of the EM algorithm. We also introduce the survival function $S_X(x) \coloneqq \mathbb{P}(X > x|\theta)$, and $S_X^{[k]}(x)$ analogously.

4.1 Univariate case

Let X be a r.v. on the positive integers such that

$$S_X(j) = \prod_{i=0}^{j} \frac{G_i}{N_i}, j \geq 0. \tag{11}$$

Following the urn representation of Section 3, G_j denotes the number of green balls in urn U_j, and N_j the total number of balls in the same urn. These pairs (G_i, N_i) (or rather, their ratio) are the variables we wish to calibrate via the EM algorithm[6].

If we have a series of i.i.d. observations $\boldsymbol{X} = \{x_i, 1 \leq i \leq n\}$ generated according to Equation (11), the sample log-likelihood is:

$$L(\mathbf{X}) = L(\mathbf{X}|\theta) = \sum_{i=1}^{n} \left(\log(N_{x_i} - G_{x_i}) - \log(N_{x_i}) + \sum_{j=0}^{x_i - 1} (\log(G_j) - \log(N_j)) \right). \tag{12}$$

[6] Notice that defining the pair (G_i, N_i) is mathematically redundant. Given Equation (11), only the ratio between G_i and N_i is actually relevant. The reason for this choice is that it helps us maintain the urn interpretation of Section 3, particularly useful for the results of Subsection 4.4.

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

$$G_j = \frac{s_{j+1}(\boldsymbol{X})}{s_j(\boldsymbol{X})} N_j, \tag{13}$$

where $s_j(\boldsymbol{X}) = \sum_{i=1}^{n} \mathbb{1}_{\{j \leq x_i\}}$ was defined in Section 3. If we now plug Equation (13) into Equation (11), we end up with the KM estimator in the absence of censoring. It is also straightforward to prove that, in the case of right-censoring and left-truncation, we recover the estimator of Cox and Oakes [1984]. After all, these estimators were specifically built to maximize the likelihood of these type of observations! This also corresponds to the RUP estimator of Equation (4), when we ignore the contribution of the prior distribution elicited by the pairs (β_j, ω_j).

Although the parameters' estimates can already be obtained through classical maximum likelihood estimation (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 MLE solution is far from trivial.

Estimating censored observations via EM is a well-known and extensively studied problem [Turnbull, 1976], and so it poses no major difficulty. The hard part is to do the same for left-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 truncated (biased) observations. In our case, the biased data is that for which $T \leq X$, and the complete sample is the one considering as well the cases in which $T > X$.

Unfortunately, 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

$$M = \frac{n}{\mathbb{P}(T \leq X)}, \tag{14}$$

and consider every value of (X, T) for which $T > X$ holds[7]. Therefore, given an 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 complete log-likelihood:

- $T \leq X$ and X uncensored:
$$L_1(X, \delta, T) = \sum_{i=1}^{n_0} [\log(p_X(x_i)) + \log(p_T(t_i))]. \tag{15}$$
- $T \leq X$ and X right-censored:
$$L_2(X, \delta, T) = \sum_{i=n0+1}^{n} [\mathbb{E}(\log(p_X(X))|X > x_i) + \log(p_T(t_i))]. \tag{16}$$
- $T > X$:
$$L_3(X, \delta, T) = M^{[k]} \sum_{x=0}^{\infty} \sum_{t=x+1}^{\infty} [\log(p_X(x)) + \log(p_T(t))]\, p_X^{[k]}(x)\, p_T^{[k]}(t), \tag{17}$$
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
$$\mathbb{E}(\log(p_X(X))|X > x_i) = \sum_{j=x_0+1}^{\infty} \log(p_X(j))\, p_X^{[k]}(j|X > x_i). \tag{18}$$
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, one can prove that

$$G_j^X = \frac{\sum_{i=1}^{n_0} S_X^{[k]}(j|x_i) + \sum_{i=n_0+1}^{n} S_X^{[k]}(j|X > x_i) + S_X^{[k]}(j|T > X)}{\sum_{i=1}^{n_0} S_X^{[k]}(j-1|x_i) + \sum_{i=n_0+1}^{n} S_X^{[k]}(j-1|X > x_i) + S_X^{[k]}(j-1|T > X)} N_j^X, \tag{19}$$

and

$$G_j^T = \frac{\sum_{i=1}^{n} S_T^{[k]}(j|t_i) + S_T^{[k]}(j|T > X)}{\sum_{i=1}^{n} S_T^{[k]}(j-1|t_i) + S_T^{[k]}(j-1|T > X)} N_j^T. \tag{20}$$

[7] This clearly requires modelling the truncated variable too, which not only increases the dimensionality of the problem, but, in many applications, one cannot infer any meaningful result from it, making it of no particular interest.

4.2 One-factor model without LTRC

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,

$$\log p(A = a, B = b, C = c|\theta) = \log p_A(a) + \log p_B(b) + \log p_C(c), \tag{21}$$

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

$$Q_A(\theta|\theta^{[k]}) = \sum_{a=0}^{x\wedge y} \log p_A(a)\; p_A^{[k]}(a|x, y). \tag{22}$$

A few comments about this equation are in order. First, the lower limit of the summation in Equation (22) is zero because we work with nonnegative processes. Secondly, the upper limit is the minimum of x and y, since A cannot be bigger than X or Y. This is a direct consequence of Equation (7) and nonnegativity.

Further, given that

$$p_A(a|x, y) = \frac{p_A(a)p_B(x-a)p_C(y-a)}{p_{XY}(x, y)}, \tag{23}$$

we can write Equation (22) as

$$Q_A(\theta|\theta^{[k]}) = \sum_{a=0}^{x\wedge y} \log p_A(a) \frac{p_A^{[k]}(a)p_B^{[k]}(x-a)p_C^{[k]}(y-a)}{p_{XY}^{[k]}(x, y)}. \tag{24}$$

However, we will keep using Equation (22), as it is easier to read and 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 (11) with respect to its parameters will be used extensively throughout the rest of this section:

$$\frac{\partial \log p_A(a)}{\partial G_j^A} = \begin{cases} 0 & \text{if } j > a \\ \frac{-1}{N_j^A - G_j^A} & \text{if } j = a \\ \frac{1}{G_j^A} & \text{if } j < a. \end{cases} \tag{25}$$

Combining Equations (25) and (22), we get

$$\frac{\partial Q_A(\theta|\theta^{[k]})}{\partial G_j^A} = \frac{1}{G_j^A} S_A^{[k]}(j|x, y) - \frac{1}{N_j^A - G_j^A} p_A^{[k]}(j|x, y). \tag{26}$$

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

$$G_j^A = \frac{S_A^{[k]}(j|x, y)}{S_A^{[k]}(j-1|x, y)} N_j^A, \tag{27}$$

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

Going back to the urn representation of Subsection 3.1, the ratio W_j^A/N_j^A is defined as the probability of picking a green ball in the j-th urn, conditioned on the fact that we have reached that urn. In probabilistic terms this can be expressed as

$$\frac{G_j^A}{N_j^A} = p_A(A > j|A \geq j) = \frac{S_A(j)}{S_A(j-1)}. \tag{28}$$

Comparing Equations (28) and (27), 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. Finally, if instead of one observation, we have a sample of size n, Equation (27) becomes:

$$G_j^A = \frac{\sum_{i=1}^n S_A^{[k]}(j|x_i, y_i)}{\sum_{i=1}^n S_A^{[k]}(j-1|x_i, y_i)} N_j^A. \tag{29}$$

The solutions for G_j^B and G_j^C are completely analogous.

4.3 The general model for left-truncated and right-censored (LTRC) data

We can now extend the results of previous sections to account for bivariate LTRC data. Following the steps of Subsection 4.1, this means computing the parameters for the target and truncation variables simultaneously. We start deriving the expressions for the target variables A, B and C, then we move to the truncation variables T and ϵ. For the target variables, the log-likelihood will be a combination of the censoring and truncation components. In the case of bivariate right-censoring, we further 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.

Situation a) is trivial to compute, so we move to case b). We will attach 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 also define analogously (x, y^*). Below we only show the distribution for (x^*, y), since (x, y^*) is solved similarly. The conditional probabilities for A, B and C are:

$$p_A(a|x^*, y) = p_A(a)S_B(x-a)p_C(y-a), \qquad \text{if } a \leq y, \tag{30}$$

$$p_B(b|x^*, y) = p_B(b) \sum_{a=x-b}^{y} p_A(a)p_C(y-a), \qquad \text{if } b \geq x-y, \tag{31}$$

$$p_C(c|x^*, y) = p_C(c)p_A(y-c)S_B(x-y+c), \qquad \text{if } c \leq y. \tag{32}$$

Contrary to the setting of Subsection 4.2, now we can observe asymmetries arise in the formulas of B and C. The reason is that, even if X is censored, A cannot be bigger than Y if it is an exact observation, and vice versa.

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

$$p_A(a|x^*, y^*) = p_A(a)S_B(x-a)S_C(y-a), \qquad \text{if } a \geq 0, \tag{33}$$

$$p_B(b|x^*, y^*) = p_B(b) \sum_{a=x-b}^{\infty} p_A(a)S_C(y-a), \qquad \text{if } b \geq 0, \tag{34}$$

$$p_C(c|x^*, y^*) = p_C(c) \sum_{a=x-c}^{\infty} p_A(a)S_B(x-a), \qquad \text{if } c \geq 0. \tag{35}$$

Let us know consider truncation. First we define the truncation event as

$$\mathscr{A} = (T \leq X, T + \epsilon \leq Y + \epsilon_0). \tag{36}$$

The unobserved region is that for which $\mathscr{A}^{\mathsf{c}}$, the complement 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

$$p(\cdot) = p(\cdot|\mathscr{A})p(\mathscr{A}) + p(\cdot|\mathscr{A}^{\mathsf{c}})p(\mathscr{A}^{\mathsf{c}}), \tag{37}$$

it is easier to compute all quantities conditioned on $\mathscr{A}$ and then use Equation (37) to condition on its complement. For that we also need the probability of the truncation event, which can be computed as

$$p(\mathscr{A}) = \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} p_{XY}(x,y)p(T \leq x, T + \epsilon \leq y + \epsilon_0), \tag{38}$$

with

$$p(T \leq x, T + \epsilon \leq y + \epsilon_0) = \sum_{t=0}^{x} p_T(t)p_\epsilon(\epsilon \leq y + \epsilon_0 - t). \tag{39}$$

Furthermore, the probabilities of each variable, conditioned on the truncation event, are

$$p_A(a|\mathscr{A}) = \frac{p_A(a)}{p(\mathscr{A})} \sum_{x=a}^{\infty}\sum_{y=a}^{\infty} p_B(x-a)p_C(y-a)p(T \leq x, T + \epsilon \leq y + \epsilon_0), \tag{40}$$

$$p_B(b|\mathscr{A}) = \frac{p_B(b)}{p(\mathscr{A})} \sum_{x=b}^{\infty}\sum_{y=x-b}^{\infty} p_C(y-x+b)p_A(x-b)p(T \leq x, T + \epsilon \leq y + \epsilon_0), \tag{41}$$

$$p_C(c|\mathscr{A}) = \frac{p_C(c)}{p(\mathscr{A})} \sum_{y=c}^{\infty} \sum_{x=y-c}^{\infty} p_B(x-y+c)p_A(y-c)p(T \le x, T+\epsilon \le y+\epsilon_0), \tag{42}$$

$$p_T(t|\mathscr{A}) = \frac{p_T(t)}{p(\mathscr{A})} \sum_{x=t}^{\infty} \sum_{y=(t-\epsilon_0)^+}^{\infty} p_{XY}(x,y)p_\epsilon(\epsilon \le y+\epsilon_0 - t), \tag{43}$$

$$p_\epsilon(e|\mathscr{A}) = \frac{p_\epsilon(e)}{p(\mathscr{A})} \sum_{x=0}^{\infty} \sum_{y=(e-\epsilon_0)^+}^{\infty} p_{XY}(x,y)p_T(T \le x \wedge (y+\epsilon_0 - e)). \tag{44}$$

Now we have all the ingredients to compute the optimal parameters of the current iteration. Without loss of generality, assume we arrange our n observations so that the first n_0 observations are uncensored. In the next $n_X - n_0$, censoring only affects X and, in the following $n_Y - n_X$ couples, censoring only occurs for Y. Finally, in the last $n - n_Y$ observations, censoring affects both variables. As in Subsection 4.2, we arrive at the following expression for the parameters of A:

$$G_j^A = \frac{{}^{n_0}_{0}S_A^{[k]}(j) + {}^{n_X}_{n_0}S_A^{[k]}(j) + {}^{n_Y}_{n_X}S_A^{[k]}(j) + {}^{n_Y}_{n}S_A^{[k]}(j) + (M^{[k]} - n)\, S_A^{[k]}(j|\mathscr{A}^\mathsf{c})}{{}^{n_0}_{0}S_A^{[k]}(j-1) + {}^{n_X}_{n_0}S_A^{[k]}(j-1|) + {}^{n_Y}_{n_X}S_A^{[k]}(j-1) + {}^{n_Y}_{n}S_A^{[k]}(j-1) + (M^{[k]} - n)\, S_A^{[k]}(j-1|\mathscr{A}^\mathsf{c})} N_j^A, \tag{45}$$

where ${}^{n_2}_{n_1}S_A^{[k]}(j) = \sum_{i=n_1+1}^{n_2} S_A^{[k]}(j|x_i, y_i)$, and $M^{[k]}$ is the estimated sample size at the current iteration via Equation (14). As before, the expressions for G_j^B and G_j^C are completely analogous and therefore omitted.

For the truncation variables T and ϵ, expressions are simpler, since we do not have to account for any censoring. We show here the solution for G_j^T only, for G_j^ϵ is obtained in a similar way:

$$G_j^T = \frac{\sum_{i=1}^n \mathbb{1}_{\{t_i > j\}} + (M^{[k]} - n)\, S_T^{[k]}(j|\mathscr{A}^\mathsf{c})}{\sum_{i=1}^n \mathbb{1}_{\{t_i \ge j\}} + (M^{[k]} - n)\, S_T^{[k]}(j-1|\mathscr{A}^\mathsf{c})} N_j^T, \tag{46}$$

where $\boldsymbol{T} = (t_1, ..., t_n)$ is the observed sample of T. It is worth stressing that, in the absence of truncation, Equation (46) returns the standard KM estimator. Therefore, Equation (46) can be interpreted as the expected value of the KM estimator given the missing data and our current estimator of p_T.

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 one-dimensional distributions, we can use the EM algorithm to compute separately the expected KM estimates of the relevant variables given the incomplete observations.

4.4 The Expectation-Reinforcement algorithm and experts' judgements

The last result of Subsection 4.3 combined with Equation (4) inspired us to consider the inclusion of the reinforcement mechanism of RUPs into the EM algorithm, offering the possibility of embedding prior knowledge and experts' judgements 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 (6), 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 [Embrechts et al., 2003], or by epistemic uncertainty [Shackle, 1955, Taleb, 2007], the use of some experts' intuition can be extremely useful to improve the performance of the model.

Moreover, nonparametric estimators 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 to 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.

To illustrate how the algorithm works, we show an example where the estimates of the ER algorithm are computed for variables A and T from the EM estimators in Equations (45) and (46). Since, given the EM estimator, the procedure is the same for the other variables considered, their explicit ER derivation is omitted for the sake of space.

Let us assume that A is a RUP with prior distribution given by the pairs $\{\beta_j^A, \omega_j^A\}$, for $j = 0, ..., M$, through Equation (6). Now we follow the steps described in Subsection 4.3 to obtain the EM estimates at each iteration, but instead of using Equation (45) to update the number of balls in each urn, the updating mechanism under the ER algorithm becomes

$$G_j^A = \omega_j^A + r\left[{}_{0}^{n_0}S_A^{[k]}(j) + {}_{n_0}^{n_X}S_A^{[k]}(j) + {}_{n_X}^{n_Y}S_A^{[k]}(j) + {}_{n}^{n_Y}S_A^{[k]}(j) + \left(M^{[k]} - n\right)S_A^{[k]}(j|\mathscr{A}^{\mathsf{c}})\right], \tag{47}$$

$$N_j^A = \beta_j^A + \omega_j^A + r\left[{}_{0}^{n_0}S_A^{[k]}(j-1) + {}_{n_0}^{n_X}S_A^{[k]}(j-1|) + {}_{n_X}^{n_Y}S_A^{[k]}(j-1) + {}_{n}^{n_Y}S_A^{[k]}(j-1) + \left(M^{[k]} - n\right)S_A^{[k]}(j-1|\mathscr{A}^{\mathsf{c}})\right]. \tag{48}$$

Following a similar procedure for the truncation variable T, we arrive at

$$G_j^T = \omega_j^T + r\left[\sum_{i=1}^{n} \mathbb{1}_{\{t_i > j\}} + \left(M^{[k]} - n\right)S_T^{[k]}(j|\mathscr{A}^{\mathsf{c}})\right], \tag{49}$$

$$N_j^T = \beta_j^T + \omega_j^T + r\left[\sum_{i=1}^{n} \mathbb{1}_{\{t_i \geq j\}} + \left(M^{[k]} - n\right)S_T^{[k]}(j-1|\mathscr{A}^{\mathsf{c}})\right], \tag{50}$$

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.

By giving different values to the reinforcement and belief parameters[8], we can control which component dominates. Similarly to the behaviour shown in Section 3, 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 equal the prior distribution.

Interestingly, Equations (49) and (50) can be interpreted as the estimation of two different data sets combined: 1)the actual observations $\boldsymbol{T}$ (multiplied r times), which correspond to the last term at the right-hand side of Equations (49) and (50); and 2) a fictitious data set $\hat{\boldsymbol{T}}$, chosen in a way such that

$$\omega_j = \sum_{i=1}^{\hat{n}} \mathbb{1}_{\{\hat{t}_i > j\}}, \tag{51}$$

and

$$\beta_j = \sum_{i=1}^{\hat{n}} \mathbb{1}_{\{\hat{t}_i = j\}}, \tag{52}$$

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.

An Expectation-Reinforcement pseudocode can be found in Algorithm 1.

[8] By Equation (6) the strength of belief is already implicit in ω and β.

Algorithm 1 ER algorithm pseudocode.

```
Set the belief and reinforcement parameters.
Set a prior via the pairs ${\beta, \omega}$ according to Equation (6).
Choose initial estimates for the first iteration of the EM algorithm.        ▷ E.g. the prior itself.
while $stop\_criteria = False$ do
    Compute the unconditional distributions using Equation (11) with the estimates from the previous iteration.
    Compute $M^{[k]}$ via Equation (14).
    Compute the conditional probabilities from Equations (30)-(32), (33)- (35) or (40)-(44) depending on the type of censoring and truncation.
    Compute the estimates for this iteration using Equations (47)-(48) for the target variables and (49)-(50) for the truncation variables.
end while
```

5 The algorithms at work

Let us test the performance of our algorithms in different settings. We do this first in a controlled environment in Subsection 5.1 using an analytical example for which we know the underlying distribution. Then, in Subsection 5.2, we analyze a Canadian data set of coupled lifetimes, widely used by scholars in actuarial sciences in the context of joint annuity evaluation [Frees et al., 1996, Souto Arias and Cirillo, 2021]. This data set is known for its complexity, due to the strong presence of censoring and truncation. Since we do not know the true underlying distribution for this problem, we compare the results of the ER algorithm with the Frank copula of Frees et al. [1996], which has already shown to give highly satisfactory results.

5.1 Analytical example

We start with a bivariate example where the ER estimator is compared with the KM estimator of Cox and Oakes [1984] for the marginals, and with the true solution. Although we have verified the performance of the ER algorithm for several examples, here we only show one of the most representative cases, based on Poisson distributions (also recalling a similar experiment in Bulla [2005]). In Appendix C we include another bivariate example in which distributions of different nature are combined, while a basic univariate example is given for completeness in Appendix B.

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

Since we also want to study the impact of the a priori on the posterior, we distinguish between two scenarios: *low* strength of belief and *high* strength of belief. In the remaining of this work, we will denote the ER estimator obtained in these scenarios by ER^l and ER^h, respectively. The values for the belief and reinforcement parameters for each case can be found in Table 1. We chose these settings because, here, we are interested in observing how the posterior deviates from the prior for different beliefs. However, in practice, it could be more interesting to attach different strengths of belief to each urn. For example, for the urns associated with the bulk of the data we could use low strengths of belief, and higher values for areas were observations are sparse, if we have some evidence that the data set is biased in those areas due to the lack of observations.

	EM	ER^l	ER^h
r	–	10^4	$2 \cdot 10^1$
c_1	0	1	10^3
c_2	0	1	10^3

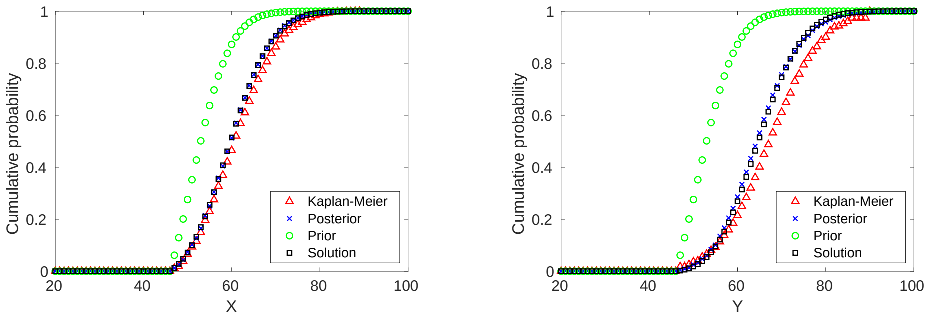
Table 1: Proposed scenarios defined by the ratio between the belief and reinforcement parameters. The values are used throughout Section 5. The superscripts in the ER columns denote "low" and "high", respectively, referring to the weight of the belief parameters. Here r is the reinforcement parameter—which, we assume, is the same for every variable—and c_1 and c_2 are the belief parameters associated to the target and truncation variables, respectively.

Now we initialize the ER algorithm. There are five processes we need to model with RUPs: A, B, C, T and ϵ. We start by giving values to the pairs $\{\beta_j, \omega_j\}$ so that the RUPs center on specific a prioris. Our choices for this example are $G^A =$ Poi(25), $G^B =$ Poi(25), $G^C =$ Poi(25), $G^T =$ Poi(50), $G^\epsilon =$ Poi(10), $\epsilon_0 = 10$. With this choice we are basically 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 $c^A = c^B = c^C = c_1$ and $c^T = c^\epsilon = c_2$, with c_1 and c_2 as per Table 1. For example, in the ER^l scenario we would have $c^A = c^B = c^C = c^\epsilon = c^T = 1$, and $r = 10^4$ (to make the data speak quickly and loudly).

In Table 2, we show for some urns affecting the A process, the comparisons between the initial composition (as per the elicited prior) and the results after the ER estimation, under the l and the h scenario. For the sake of space[9], we only show a small subset of all the urns forming the RUP behind A, and for the same reason we omit those of B and C. Even with a few urns, it is possible to observe the impact of the strengths of belief in estimation.

% Urn	Prior	ER^l	ER^h
0	100.00	100.00	100.00
20	94.01	98.62	96.41
40	59.11	75.23	92.97
60	40.55	17.93	99.99

Table 2: Comparison of some urn compositions for process A before and after estimation (results of B and C are comparable). The table shows the percentage of green balls in each urn under the prior, and after the ER estimation for the two scenarios l and h. Given the large number of urns, we only show the composition for a few selected cases. Notice that, by construction, Urn 0 will stay untouched, as it only contains green balls.



(a) Marginal of X.

(b) Marginal of Y.

Figure 2: 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.

In Figure 2 we show the ER^l estimate for the marginals of X and Y compared to the KM estimator and the true distribution, while in Table 3 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 appropriately capture both curves and moments. Also notice how computing the correlation coefficient from the data (“Raw Data"), while ignoring censoring and truncation, definitely overestimates dependence (0.88 against a true value of 0.64). A further analysis on how good this fit is can be found in Figure 4, where we show the QQ-plots for the ER^l marginals, and Table 4, where we perform permutation tests for the means and the variances. The QQ-plots were created by generating two samples of size 10^3 from each marginal of the ER^l solution, 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).

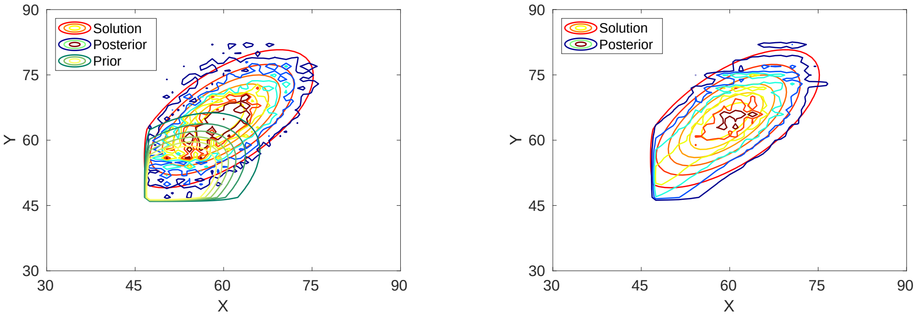
Finally, in Figures 3a and 3b, we give the contour plots of the bivariate ER distribution and the analytical solution, for low and high strengths of belief, respectively. Notice that, as expected, increasing the strength of belief in our a

[9] If needed, all urns compositions are available upon request, also for the other applications discussed in the present paper.

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[10].

	Raw 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 3: Comparison of the means, variances and correlation of X and Y using: the data as they are ignoring censoring and truncation, 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 3: 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).

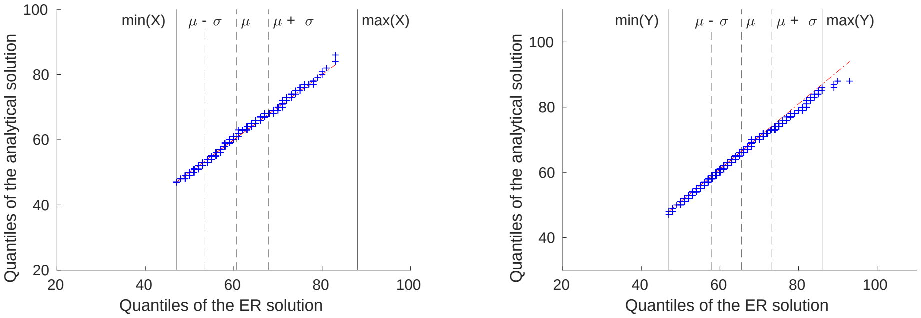
	Value	p (%)	H_0
mean(X^l)	-0.229	46.46	Do not reject
mean(Y^l)	-0.409	25.20	Do not reject
var(X^l)	-5529	94.86	Do not reject
var(Y^l)	-6588	16.86	Do not reject

Table 4: 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.

5.2 Empirical case: coupled lifetimes

We now apply the same methodology to a Canadian data set of coupled lifetimes widely used in the field of joint annuity modeling [Frees et al., 1996, Luciano et al., 2008].

[10] 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 for themselves instead.



(a) QQ-plot: X.

(b) QQ-plot: Y.

Figure 4: 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.

The data consists 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.

Following 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 contracts, and thus they may appear several times on the data sheet. Therefore we remove all repeated entries so that each couple is considered only once. Finally, as in Frees et al. [1996], we condition on couples that 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 significant 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].

Since the performance of the B-RUP on this data set has already been extensively analyzed in Souto Arias and Cirillo [2021], even if without left-truncation and only using Markov Chain Monte Carlo techniques, here we do not focus on the impact that the B-RUP parameters have on the posterior. Rather, we compare the ER estimator with a popular parametric model which also can capture bivariate right-censoring and left-truncation: the Frank copula model defined in Frees et al. [1996].

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 = Poi(35), G^B = Poi(40), G^C = Poi(40), G^T = Poi(80), G^ϵ = Poi(40), $\epsilon_0 = 40$, where ϵ_0 was inferred from the maximum difference in ages in the data set. With this a priori, 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 example (Table 1), apart from the reinforcement of ER^h, for which we now set to $r = 10$.

Regarding the copula model, we select the model from Frees et al. [1996], where the authors use Gompertz distributions for the individual lifetimes, and the Frank copula to model the dependence. The Gompertz distribution is given by

$$\text{Gomp}(x; \mu, \sigma) = 1 - \exp\left(e^{-\frac{\mu}{\sigma}}(1 - e^{\frac{x}{\sigma}})\right), \tag{53}$$

where μ, σ are the location and scale parameter, respectively.

The Frank copula is defined as

$$C(u, v; \alpha) = \frac{1}{\alpha} \log\left(1 + \frac{(e^{\alpha u} - 1)(e^{\alpha v} - 1)}{e^{\alpha} - 1}\right), \tag{54}$$

where u, v are the marginal distributions for the male and female annuitants, respectively, and α is the parameter controlling the dependence. A negative value of α indicates positive dependence, while $\alpha = 0$ means independence [Nelsen, 2006].

We follow the same procedure of Frees et al. [1996] for estimating the model parameters. The interested reader can check the original paper for the methodology, while here we jump directly to the results. In Table 5 we present the optimal parameters obtained via MLE, where (μ_X, σ_X) are the Gompertz estimates for the male annuitants, and

(μ_Y, σ_Y) the estimates for the female annuitants. Since the value of α is highly negative, we expect a strong positive dependence (something that also justified the use of the B-RUP construction in Souto Arias and Cirillo [2021]). In Table 6, we see indeed a Pearson correlation of 0.5. In the same table, we observe a correlation of 0.82 for the “Raw Data", but it is important to stress that this is due to the fact that, in that column, we are explicitly ignoring censoring and truncation, to show how easily we can end up with overestimating correlation.

μ_X	σ_X	μ_Y	σ_Y	α
84.809	9.926	87.575	7.792	-4.081

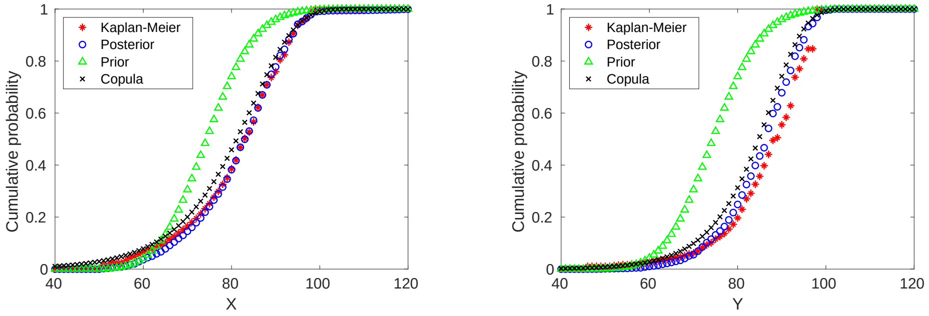
Table 5: Calibration of the copula model via MLE. The subscripts X, Y refer to the marginals of the male and female annuitants, respectively. Notice that α is negative and away from 0, so according to Equation (54) there is positive dependence.

In Figures 5 and 6c we show the marginal and joint distributions, when the strength of belief in the a priori is low. Note that our prior clearly underestimated the average lifetimes, especially for females, and the correlation slightly decreases from 0.47 to 0.40. Furthermore, in Figure 6c we can observe small contours in which the age difference is particularly large. In those cases we cannot be sure whether the observation corresponds to a married couple or a parent-child relationship. In Figure 6d, we show the results when we assume that the a priori is strong and reliable.

Finally, we present some moments and quantiles of the different estimators in Table 6. Column *ER*^{l} accounts for the low strength of belief case, while *ER*^{h} deals with the high belief situation.

	Raw Data	KM	ER^l	ER^h	ER2^l	ER2^h	Copula
Mean(X)	74.514	81.581	81.901	81.785	81.952	82.000	79.597
Q1(X)	70.000	75.000	76.000	75.000	76.000	76.000	72.000
Median(X)	74.000	83.000	83.000	83.000	83.000	83.000	81.000
Q3(X)	79.000	90.000	89.000	89.000	89.000	89.000	87.000
Mean(Y)	74.011	86.989	85.491	84.637	85.246	85.432	83.579
Q1(Y)	69.000	82.000	80.000	79.000	80.000	80.000	77.000
Median(Y)	73.000	89.000	87.000	86.000	87.000	86.000	84.000
Q3(Y)	79.000	94.000	92.000	91.000	92.000	92.000	90.000
Var(X)	52.049	124.393	117.075	112.088	113.935	123.091	160.610
Var(Y)	61.707	99.945	76.866	85.410	77.543	89.931	99.844
Corr(X,Y)	0.820	–	0.401	0.461	0.432	0.398	0.502

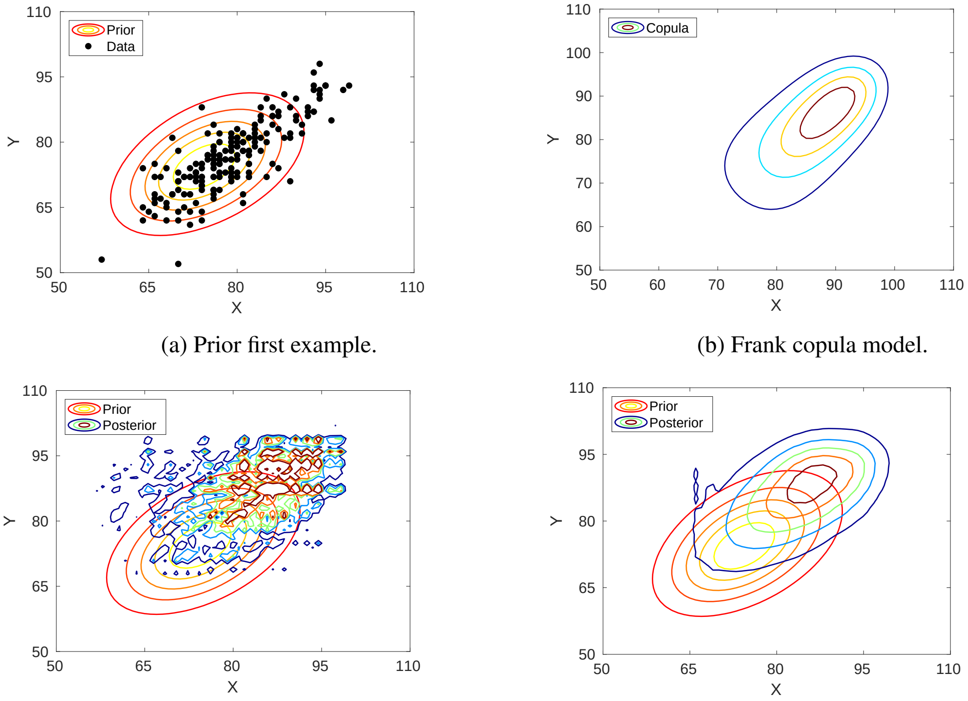
Table 6: Comparison of the means, medians, first and third quartiles, variances and correlation of X and Y. Here the column “Raw Data” refers to the data as they are, if we ignore censoring and truncation, overestimating dependence.



(a) Marginal of X.

(b) Marginal of Y.

Figure 5: Fitting of the marginal distributions of X and Y using the Canadian data set. Initial estimates are represented by the green line, KM estimates are in red and our ER solution in blue.



(a) Prior first example.

(b) Frank copula model.

(c) Low strength of belief.

(d) High strength of belief.

Figure 6: Contour plot for the empirical data set. In the upper row we show a scatter plot of the uncensored samples available in the Canadian data set, together with the distribution obtained by the Frank copula model. In the lower row we present a comparison of the posterior distributions for the low and high belief scenarios, respectively. We can clearly see that uncensored data is not representative enough of the total population, since the bulk of the uncensored data is considerably shifted with respect to the higher density areas of both the copula and ER distributions.

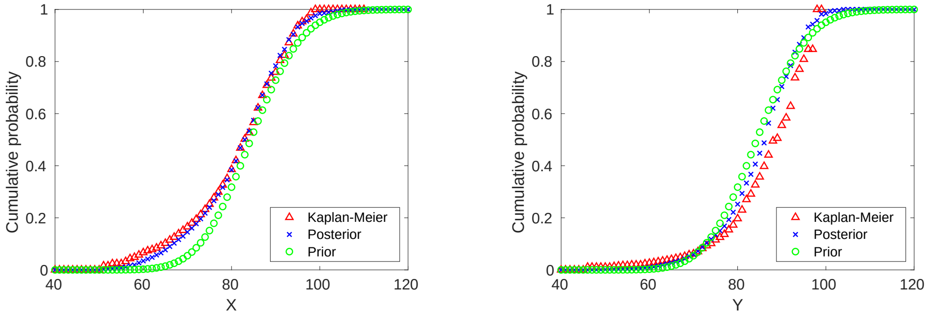
From these experiments we can draw several conclusions. First, both the copula and the ER methodology yield consistent results, showing that the models are similar in performances. Second, ignoring the effects of censoring and truncation highly underestimates the marginal moments of the distribution, while it overestimates the cross moments. This is particularly obvious from the results of Table 6, but also from Figure 6a. Third, the copula model associates a much larger weight to the left-tail of the distribution than the ER model. This seems to be a property of the parametric model in particular, since the data itself does not support such a result. Thus, from a purely data-driven point of view, the ER estimator seems more accurate. However, it is also true that, due to left-truncation, there are not many observations in that range, hinting at the possibility of some bias in the data. If that were actually the case, we could modify the ER distribution by introducing a prior with a heavier left-tail. Finally, both models yield a similar degree of strong positive dependence between the target variables.

Let us now consider a situation in which we use a more objective prior, elicited by looking at the data. 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 replicated.

In Figures 7a, 7b and 8b 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.

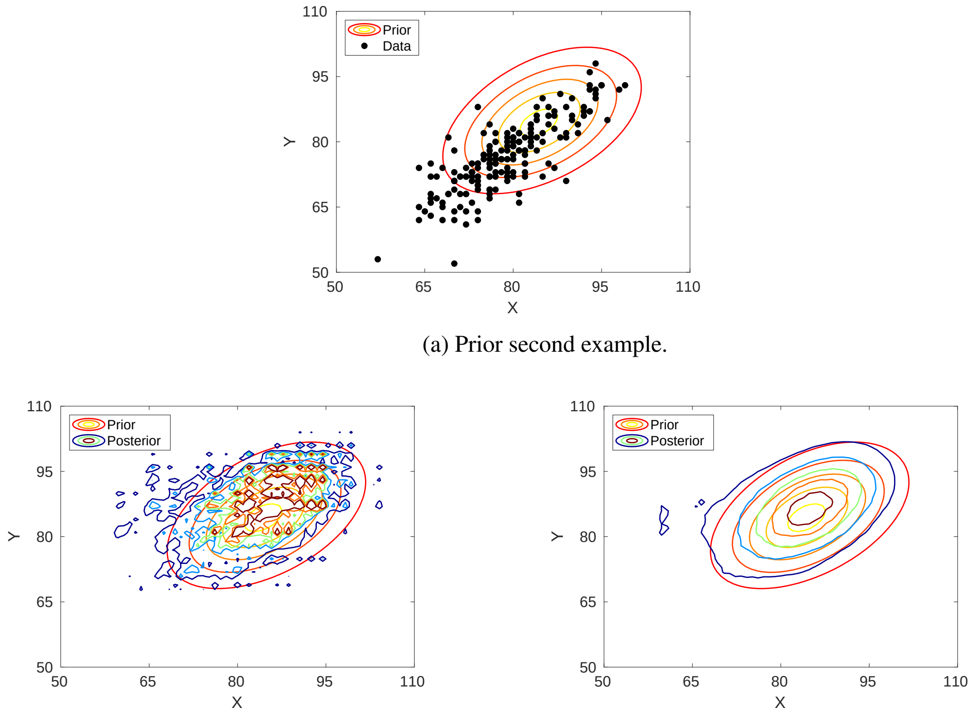
In Figure 8c, conversely, the posterior distribution when assuming strong belief in our prior is shown. As before, the first two moments are presented in the columns $ER2^l$ and $ER2^h$ of Table 6. Notice that, according to Figure 8c, the posterior distribution has barely moved away from the prior. In the previous examples so far, the algorithm always reached a compromise between the prior and the data when increasing the strength of belief. However, since this time the prior was specifically chosen by taking the data into consideration, the resulting posterior is barely affected by the data.



(a) Marginal of X.

(b) Marginal of Y.

Figure 7: Fitting of the marginal distributions of X and Y using the Canadian data set. Initial estimates are represented by the green line, KM estimates are in red and our ER solution in blue.



(a) Prior second example.

(b) Low strength of belief.

(c) High strength of belief.

Figure 8: Contour plot for the empirical data set. In the upper row we show a scatter plot of the uncensored samples available in the Canadian data set. In the lower row we present a comparison with the posterior distribution with low and high beliefs, respectively. We can clearly see that uncensored data is not representative enough of the total population, since the bulk of the uncensored data is considerably shifted with respect to the higher density areas of the posterior.

6 Conclusions

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

As the main goal of our work, for the first time in the literature, 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 Markov Chain Monte Carlo.

The performances of both the EM and the ER algorithms have been tested using artificial and actual left-truncated and right-censored data, showing their superiority with respect to other common alternatives like the estimator of Kaplan and Meier [1958], 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 straightforward in the absence of censoring and truncation (see also Bulla [2005]), but it requires much more work in a realistic and general setting. Moreover, other forms of dependence, far from linearity, can complicate things further.

Computationally, increasing the dimensionality of the problem may generate non-trivial questions to be solved. Nevertheless, we believe that the computational burden can be decreased by implementing schemes that increase the convergence of the EM algorithm, and using high performance techniques such as parallel computing.

Acknowledgment

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

A Proof of Equation (4)

As mentioned before, the RUP is a generalization of the Dirichlet process, and its underlying distribution is actually the generalized Dirichlet distribution of Connor and Mosimann [1969], i.e.

$$\mathbb{P}(p_X(1), p_X(2), \cdots, p_X(M)|\boldsymbol{\beta}, \boldsymbol{\omega}) \propto (p_X(M))^{\omega_{M-1}-1} \prod_{i=0}^{M-1} \left[(p_X(i))^{\beta_i - 1} (S_X(i-1))^{\omega_{i-1}-(\beta_i+\omega_i)} \right], \tag{55}$$

with $p_X(i) = \mathbb{P}(X = i)$, $S_X(i) = \mathbb{P}(X > i)$, (β_i, ω_i) defined as before[11], and where we assume that there are no more observations after time M.

Notice that now we can rewrite the log-likelihood in Equation (1) as

$$L(p_X(1), p_X(2), ..., p_X(M)|\boldsymbol{X}_n) \propto \prod_{i=0}^{M} (p_X(i))^{n_i} \prod_{i=0}^{M-1} (S_X(i))^{r_i - l_{i+1}}, \tag{56}$$

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 (55) and (56), 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

$$\beta_k^* = \beta_k + n_k, \qquad \omega_k^* = \omega_k + q_k, \tag{57}$$

which completes the proof.

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

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

Table 7: 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.

	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 8: 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.

B 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.

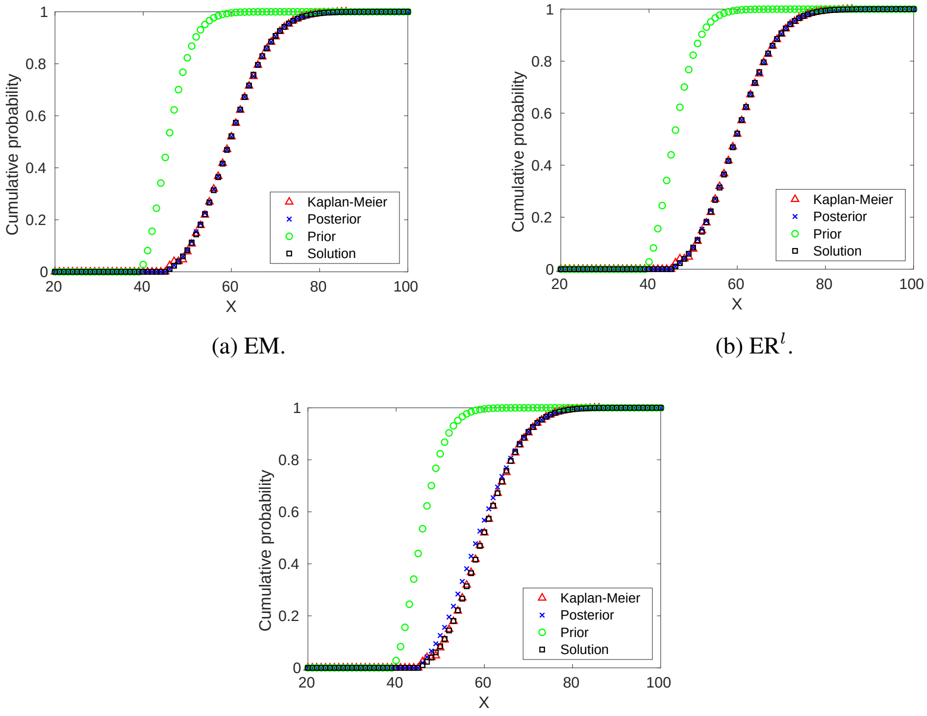
In order to use the ER algorithm, first we identify our variables of interest: X and T. Then we define a RUP for each of them with parameters $\{\beta_j^X, \omega_j^X\}$ and $\{\beta_j^T, \omega_j^T\}$, respectively.

As explained in Section 3, via Equation (6) 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 9 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 9b). 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 7, 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 10, 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 results of the test are presented in Table 8, and, as expected, fail to reject the null hypothesis for both the ER^l and KM estimators.

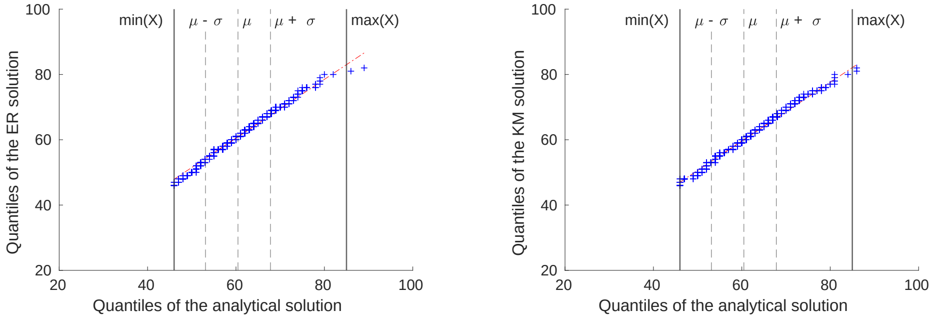


(a) EM.

(b) ER^l.

(c) ER^h.

Figure 9: 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 10: 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 ± one standard deviation (σ).

C Bivariate case II

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

$$\mathrm{Gomp}(x; \mu, \sigma) = 1 - \exp(e^{-\mu/\sigma}(1 - e^{x/\sigma}). \tag{58}$$

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.

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: $G^A = \mathrm{Gomp}(25, 8)$, $G^B = \mathrm{Gomp}(30, 7)$, $G^C = \mathrm{Gomp}(30, 7)$, $G^T = \mathrm{Poi}(50)$, $G^\epsilon = \mathrm{Poi}(10)$, $\epsilon_0 = 10$. 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.

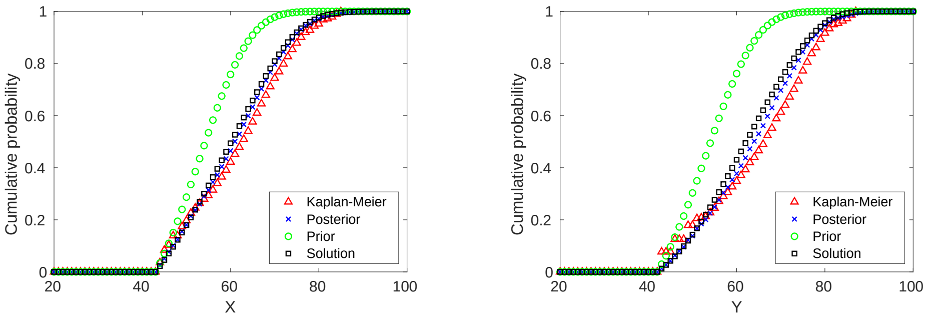
As usual, we condition on the minimum uncensored values for both X and Y. Since the truncation variable has the same distribution as before we expect these values to be similar to the previous example. Indeed, this time we observe $X_{min} = 44$ and $Y_{min} = 43$, and thus we condition the distributions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

We present the fitting of the ER^l marginals in Figure 11, the contour plots for both ER^l and ER^h in Figure 12, and the computed first two moments in Table 9. 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.

The marginals are analyzed via QQ-plots in Figure 13, while Table 10 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 9. 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 4.4, where our algorithm may improve over the original EM algorithm, by properly working with priors. For this reason we also show in Figure 12d the resulting distribution when giving a high strength belief to the a priori. In Table 9, under the column ER^h, we show the corresponding results for the moments.



(a) Marginal of X.

(b) Marginal of Y.

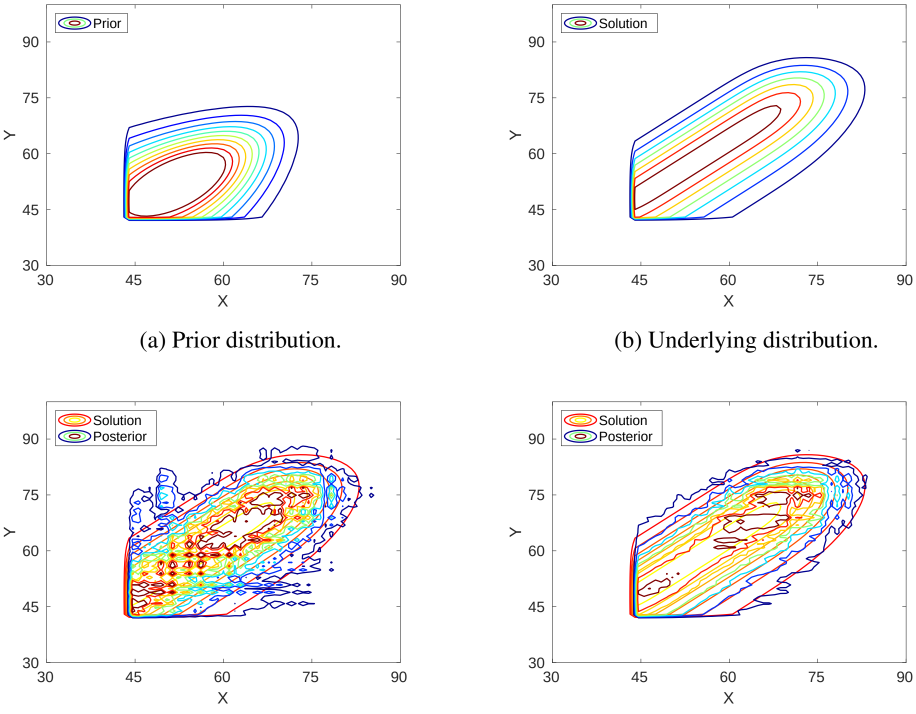
Figure 11: Fitting of the marginal distributions of X (a) and Y (b). Initial estimates are represented by the green line, KM estimates are in red, our ER solution in blue and the underlying distributions in black. Note that the bivariate truncation causes the KM estimators to be biased with respect to the true marginals.

References

Joshua Angrist, Eric Bettinger, and Michael Kremer. Long-term educational consequences of secondary school vouchers: Evidence from administrative records in colombia. *American Economic Review*, 96(3):847–862, June

	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 9: 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.



(a) Prior distribution.

(b) Underlying distribution.

(c) ER^l scenario.

(d) ER^h scenario.

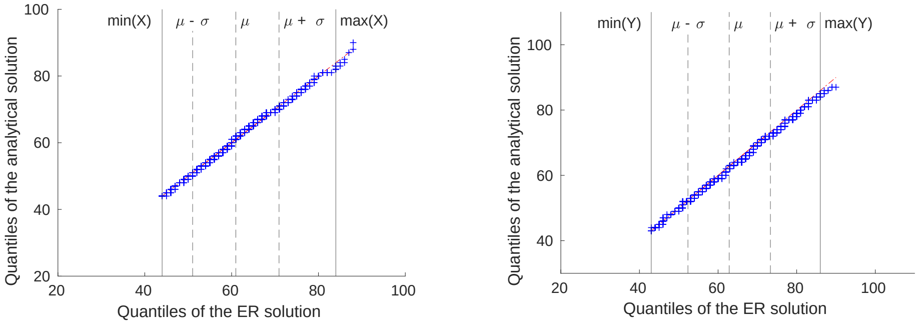
Figure 12: Contour plots for low and high strengths of belief. We also show the prior distribution (a) and the original distribution (b) for comparison purposes. On the lower-left plot (c) we show the posterior distribution for low strength of belief and the lower-right plot (d) the posterior distribution 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.

2006. doi:10.1257/aer.96.3.847. URL https://www.aeaweb.org/articles?id=10.1257/aer.96.3.847.

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.

Paolo Bulla. *Application of Reinforced Urn Processes to Survival Analysis*. PhD thesis, Bocconi University, 2005.



(a) QQ-plot: X.

(b) QQ-plot: Y.

Figure 13: 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.

	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 10: 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.

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.

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.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events for insurance and finance, 2nd edition*, volume 33. Springer Science & Business Media, 2003.

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.

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.

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.

John P. Klein and Melvin L. Moeschberger. *Survival Analysis Techniques for Censored and Truncated Data*. Second edition, 2003.

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.

Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2. ed edition, 2008.

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.

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.

Roger B. Nelsen. *An Introduction to Copulas*. Springer, New York, 2006. URL `http://www.worldcat.org/search?qt=worldcat_org_all&q=0387286594`.

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.

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.

Luis A. Souto Arias and Pasquale Cirillo. Joint and Survivor Annuity Valuation with a Bivariate Reinforced Urn Process. *Insurance: Mathematics and Economics*, 99:174–189, 2021. doi:10.1016/j.insmatheco.2021.04.004.

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.

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.

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