

Research



Cite this article: Souto Arias LA, Cirillo P, Oosterlee CW. 2023 Estimation of reinforced urn processes under left-truncation and right-censoring. *R. Soc. Open Sci.* **10**: 221223. <https://doi.org/10.1098/rsos.221223>

Received: 19 September 2022

Accepted: 20 February 2023

Subject Category:

Mathematics

Subject Areas:

statistics/computational mathematics/applied mathematics

Keywords:

reinforced urn process, expectation–maximization, bivariate survival function, left-truncation, right-censoring

Author for correspondence:

Luis A. Souto Arias
e-mail: l.a.soutoarias@uu.nl

Estimation of reinforced urn processes under left-truncation and right-censoring

Luis A. Souto Arias¹, Pasquale Cirillo² and Cornelis W. Oosterlee¹

¹Mathematical Institute, Utrecht University, Utrecht, The Netherlands

²ZHAW School of Law and Management, Zurich University of Applied Sciences, Zurich, Switzerland

ID: LASA, 0000-0003-2360-6884

We propose a non-parametric estimator for bivariate left-truncated and right-censored observations that combines the expectation–maximization algorithm and the reinforced urn process. The resulting expectation–reinforcement algorithm allows for the inclusion of experts’ knowledge in the form of a prior distribution, thus belonging to the class of Bayesian models. This can be relevant in applications where the data is incomplete, due to biases in the sampling process, as in the case of left-truncation and right-censoring. With this new approach, the distribution of the truncation variables is also recovered, granting further insight into those biases, and playing an important role in applications like prevalent cohort studies. The estimators are tested numerically using artificial and empirical datasets, and compared with other methodologies such as copula models and the Kaplan–Meier estimator.

1. Introduction

We deal with bivariate left-truncated and right-censored (LTRC) data in a non-parametric way, combining reinforced urn processes (RUP) [1]—in particular, the bivariate construction (B-RUP) of Bulla *et al.* [2]—and the expectation–maximization (EM) algorithm. We call this combination the expectation–reinforcement (ER) algorithm, which enjoys the optimization properties of the EM algorithm, while maintaining the Bayesian formulation of the RUP. The ER approach proves preferable over the classic EM when reliable experts’ knowledge about the phenomenon under study is available, which can then be included as a prior distribution. This feature enables the correction of biases present in the data, especially when dealing with extreme and rare

events, as well as the possibility of inputting plausible information about future trends and other not-yet-observed characteristics.

LTRC observations are frequent and relevant in several fields of science, such as medicine, survival studies, education studies, engineering and risk management [3–5]. For instance, right-censoring appears in most medical studies, as patients are observed over a limited period of time, and the event of interest—e.g. the next stage of a disease—may occur after the observation period. Left-truncation, conversely, occurs when patients join the study with the disease already in an advanced stage, giving rise to missed information about the earlier stages [6].

The problem of non-parametric estimation in the bivariate LTRC setting is not new, and among the first proposals to tackle this problem one can find [7–9] and [10]. However, as shown in [11], these estimators may fail to be monotone, thus possibly generating negative probabilities. More recent works are those of Shen & Yan [5] and Gribkova & Lopez [12]. The former develops a complex iterative method to estimate a generalization of the Dabrowska and Campbell and Földes estimators, which includes the effect of left-truncation. A simplified version of this algorithm that does not require iteration was proposed in [13]. In spite of this numerical advantage, the new estimator developed in [13] can be sensitive with respect to some of its parameters, producing inaccurate results when not chosen properly. Gribkova & Lopez [12] use a non-parametric estimator via random weights, first defined in [14], to compute a non-parametric copula in the presence of bivariate right-censoring when there is no left-truncation. The case of bivariate left-truncation is particularly interesting, since, as explained in [5], in this case the Kaplan–Meier (KM) and product-limit estimators [15,16] are not consistent estimators of the marginal survival functions. Specifically, in the bivariate setting with two target variables (X , Y), as shown in [5], the product-limit estimator computes the marginal distribution of X conditioned on the truncation event of Y , which does not correspond to the true marginal distribution of X . A similar reasoning applies to the estimation of the marginal distribution of Y . Therefore, not only the joint distribution, but also the marginal distributions are difficult to determine in the presence of multivariate left-truncation.

Since left-truncation and right-censoring fall under the umbrella of incomplete data, many authors have opted for an alternative approach, using the EM algorithm of Dempster *et al.* [17]. Some relevant results in the univariate framework are available in [4] and references therein. Conversely, the literature on bivariate distributions is scarce, focusing mainly on right-censoring (see [9,18,19]). In fact, to our knowledge, there are no proposals in the literature of the EM algorithm that include bivariate left-truncation.

It should be noted that the authors in [20] have already shown that the B-RUP is able to model bivariate censoring in practical applications, comparing its performances to parametric approaches based on copulas [21,22]. However, the approach followed in [20], based on Markov Chain Monte Carlo (MCMC) simulation, is limited to the modelling of right-censoring, omitting the effect of left-truncation.

In fact, in many applications, relevant information can be extracted from the distribution of the truncation variables (see [23]). This is the case for prevalent cohort studies where, under certain assumptions, the truncation distribution corresponds to the disease distribution prior to recruitment. The proportion of truncated observations is another quantity of interest. In medical studies, for example, this quantity corresponds to the number of individuals that have died due to a specific disease prior to recruitment.

Our contribution contains, therefore, three main novelties, which can be summarized as follows:

- We offer an effective treatment of non-parametric distributions under bivariate LTRC data, overcoming some difficulties of the existing literature. The approach we propose for LTRC data can be easily adapted to the simpler situations of just LT or RC data, and to distributions in higher dimensions with minor adaptations in the methodology.
- We present an approach for the explicit estimation of RUPs in particular, and urn models in general. In the literature, urn models [24,25] have been mainly approached from a probabilistic point of view, and statistical inference has always been marginal, with some exceptions such as [26–28].
- To fully exploit the Bayesian properties of RUPs [29], we introduce the ER algorithm, which maintains the convergence properties of the standard EM algorithm but also enables the inclusion of experts' knowledge.

The paper is structured as follows. In §2, we briefly revisit the concepts of right-censoring and left-truncation and introduce the modelling assumptions. Section 3 summarizes the basic theory of RUPs, as well as the bivariate one-factor construction (B-RUP) of Bulla *et al.* [2], and their adaptation to

LTRC data. The EM and the ER algorithms for RUPs are described in §4, and complemented by an error analysis. Section 5 contains the performances of the algorithms using both simulated and empirical data. Finally, §6 concludes the paper.

2. Left-truncation and right-censoring

Let $X_n = (x_1, \dots, x_n)$ and $C_n = (c_1, \dots, c_n)$ be identically and independently distributed (i.i.d) observations with distribution functions F_X and F_C , respectively. F_X and F_C are assumed independent.

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 $i = 1, \dots, n$, with $\mathbb{1}_{\{\cdot\}}$ the indicator function. That is, we observe the minimum of the censoring variable and the target variable, plus an indicator telling us which of the two we observe.

Let $T_n = (t_1, \dots, t_n)$ 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 $i = 1, \dots, n$, if $x_i \geq t_i$, and nothing otherwise.

Since for $x_i < t_i$ nothing is observed, the data contains no information about X or T for $X < T$. This suggests that a truncated observation provides 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 underlying distribution. For this reason, according to Wang [30], truncated data can also be classified as selection-biased data.

In the general case of LTRC observations, it is usually assumed [30,31] that the variable of interest X is independent of both T and C . In this situation what one observes is the triplet (X_n^*, T_n, δ_n) , with X^* and δ defined as before if $T \leq X$, and nothing otherwise. As in [30], we further assume that $\mathbb{P}(T \leq C) = 1$, indicating that T and C are not independent.¹ In this case, the log-likelihood of the sample can be written as

$$L(X_n^*, \delta_n, T_n) = \sum_{i=1}^n [(1 - \delta_i) \log(\mathbb{P}(X > x_i^*)) + \delta_i \log(\mathbb{P}(X = x_i^*)) - \log(\mathbb{P}(X \geq t_i))]. \quad (2.1)$$

In the case of non-negative integers, a non-parametric estimator for the survival function, S_X , which maximizes the log-likelihood defined in equation (2.1) was introduced in [31] and further studied in [16], is given by

$$S_X := \mathbb{P}(X > x | X_n^*, T_n, \delta_n) = \prod_{j=0}^x \left[1 - \frac{m_j(X_n^*, \delta_n)}{s_j(X_n^*, T_n)} \right], \quad (2.2)$$

where $m_j(x_n, \delta_n) = \sum_{i=1}^n \mathbb{1}_{\{x_i = j, \delta_i = 1\}}$ is the number of exact observations at j , and $s_j(x_n, t_n) = \sum_{i=1}^n \mathbb{1}_{\{t_i \leq j \leq x_i\}}$ is the number of censored plus uncensored observations at j under left-truncation. Equation (2.1) reduces to the well-known KM estimator in the absence of truncation [15], and to the product-limit estimator of Lynden-Bell [32] without censoring.

Remark. In this article, we work with non-negative integers unless otherwise stated. This is done for simplicity, and the extension from the non-negative integers to a discrete set of the real line is straightforward.

In the bivariate situation, we assume that the joint survival function S_{XY} of the variables (X, Y) is independent of the truncation and censoring variables (T^X, C^X, T^Y, C^Y) (see [5]). Observations consist of two triplets $(X_n^*, T_n^X, \delta_n^X)$ and $(Y_n^*, T_n^Y, \delta_n^Y)$. If the target variables X and Y are independent, we can write the joint conditional log-likelihood as the sum of the marginal log-likelihoods, both defined as in equation (2.1). However, when there exists dependence, the log-likelihood of the bivariate LTRC sample takes the form:

$$L(X_n, Y_n | \delta_n^X, \delta_n^Y, T_n^X, T_n^Y) = \sum_{i=1}^n [\log(\mathbb{P}^*(x_i^*, y_i^* | \delta_i^X, \delta_i^Y)) - \log(\mathbb{P}(X \geq t_i^X, Y \geq t_i^Y))], \quad (2.3)$$

¹This condition is satisfied, for example, in lifetime follow-up studies [6], where T is the age at which an individual joins the study and C is the age at which they drop out.

with

$$\mathbb{P}^*(x, y | \delta^X, \delta^Y) = \begin{cases} \mathbb{P}(X = x, Y = y) & \text{if } \delta^X = 1 \text{ and } \delta^Y = 1, \\ \mathbb{P}(X > x, Y = y) & \text{if } \delta^X = 0 \text{ and } \delta^Y = 1, \\ \mathbb{P}(X = x, Y > y) & \text{if } \delta^X = 1 \text{ and } \delta^Y = 0, \\ \mathbb{P}(X > x, Y > y) & \text{if } \delta^X = 0 \text{ and } \delta^Y = 0. \end{cases} \quad (2.4)$$

2.1. Modelling assumptions

As mentioned in [23], it is not clear how to estimate the joint distribution $H_{T,C}$ of the censoring and truncation variables (T, C) , in the case of LTRC observations, which is, however, necessary for the algorithms developed in §4. For that purpose, Wang [23] makes the assumption, in the univariate setting, that $C = T + \Delta$, where T and Δ are independent random variables (r.v). Under this assumption, the distribution $H_{T,C}$ can be expressed as

$$dH_{T,C}(t, c) = dH_T(t) dH_\Delta(c - t), \quad (2.5)$$

where $H_T(\cdot)$ and $H_\Delta(\cdot)$ are the probability distributions of T and Δ , respectively.

Remark. Note that the censoring assumption $C = T + \Delta$ implies that censoring is at the end of the follow up, either because the study has finished or because the participant dropped out of the study. This is the case in many practical situations and, in particular, it is true for the annuity problem studied in [20,21], which we treat in §5.2.

In the bivariate LTRC setting, we are interested in the joint distribution H of two censoring (C^X, C^Y) and two truncation (T^X, T^Y) variables. Therefore, the censoring assumption becomes $C^i = T^i + \Delta$, for $i \in \{X, Y\}$. Observe that we take the same Δ for T^X and T^Y , thus entailing that both participants end the follow up at the same time. This is the case when the end of the follow up matches the end of the study. On the other hand, it could happen that the two participants drop out at different times during the study. In that case the source of dependence between participants is lost, diminishing the relevance of that particular observation. Another possibility, that we do not pursue in this work, is to set $C^X = \Delta + T^X + C^Y$ (see [5,13]). In this case, assuming Δ and T^X non-negative, it is clear that $\mathbb{P}[C^Y \leq C^X] = 1$, which can be relevant in scenarios where we are sure that one participant will drop out earlier than the other one. The methodology proposed in §4 can also be applied under this assumption, with minor modifications.

Finally, we make the truncation assumption $T^Y = T^X + \epsilon$, where T^X and ϵ are independent random variables. This situation arises for example in survival studies, where T^X and T^Y denote the age at which each individual enters the study [20,21]. In this case, the r.v. ϵ corresponds to the age difference between the individuals. Another example where this assumption is verified is in paediatric AIDS cohort studies (see [13]), where ϵ denotes the difference between the time of infection of the mother and the time of birth of the child.

3. Univariate and bivariate reinforced urn processes

The RUP and B-RUP models define random processes in one and two dimensions, respectively. In particular, we use the B-RUP to model bivariate LTRC data in a non-parametric way. Since the B-RUP is composed of several RUPs, a brief explanation of the univariate RUP model is also needed to keep this article self-contained. The reader is referred to Walker & Muliere [1], Bulla *et al.* [2], Arias & Cirillo [20] and Muliere *et al.* [33] and the references therein for an extended and more thorough analysis of the theoretical properties of these models.

3.1. The reinforced urn process

An interesting property of the RUP is that it can be easily visualized as a series of Pólya urns with balls of two different colours. Thus, assume we have $M + 1$ Pólya urns [25], where the j th urn U_j , $j = 0, 1, \dots, M$, 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 number of balls does not need to be an integer.

- (i) The process starts in U_0 .

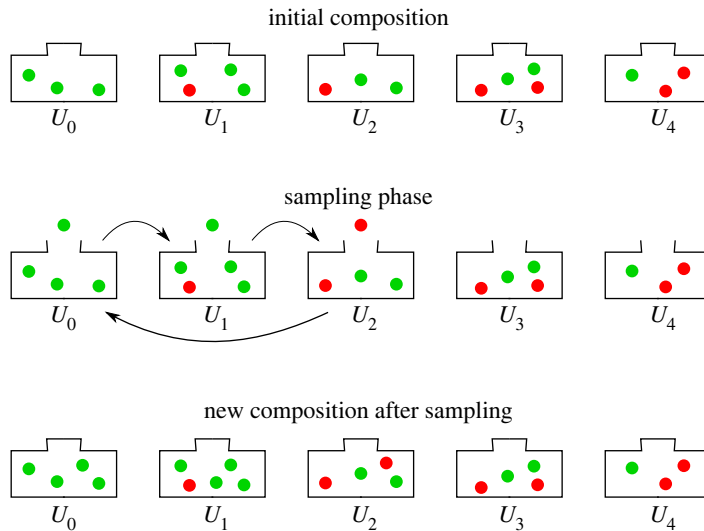


Figure 1. Representation of the RUP as a series of Pólya urns. The balls outside of the urns indicate which colour has been sampled. After each sampling, the urns are updated in a way that reinforces the probability of observing the same event again.

- (ii) For $j = 0, 1, \dots, M$, randomly pick a ball from urn U_j , look at its colour, put the ball back into the urn, and add $r > 0$ balls of the same colour.
- (iii) If the colour of the chosen ball is green, move forward to urn U_{j+1} , and repeat from step 2. If the colour is red, go back to step 1.

As an example, in figure 1 we see the result of picking (green G ball, green G ball, red R ball) and the resulting urn compositions, if we assume $r = 1$. In many applications, the event of interest is usually the sampling of a red ball. In the context of survival studies, this corresponds to the event of death, while the urn at which this happens corresponds to the lifetime of an object or an individual (if each urn denotes a different year, the observed lifetime in figure 1 would be 2 years). Hence, by observing different lifetimes and reinforcing the urns accordingly, the RUP is able to learn from the data and produce an accurate predictive distribution of the event of interest.

Higher values of r imply a stronger reinforcement, i.e. the composition of the RUP quickly changes via sampling, while for values of r close to zero the RUP requires many iterations in order to move away from its initial composition. The reader is referred to Cirillo *et al.* [34] and Peluso *et al.* [35] for a discussion about the role of r .

If the data include right-censoring and left-truncation, the procedure we have just described can be easily modified. In case of a right-censored observation, the red ball is not included, but everything else is left unchanged. If the sample is left-truncated at value j , the new cycle starts from urn U_j , and not from U_0 .

One of the most interesting features of a RUP is that the probability of its trajectories can be obtained in closed form Muliere *et al.* [33] and Fortini & Petrone [36]. This comes from the fact that, as shown in [33], a RUP generates a random distribution over the space of discrete distributions, and this random distribution is the discrete beta-Stacy process of Walker & Muliere [1].

Definition 3.1 Beta-Stacy process (discrete case). A random distribution function F is a discrete beta-Stacy process with jumps at $j \in \mathbb{N}_0$ and parameters $\{\beta_j, \omega_j\}_{j \in \mathbb{N}_0}$, 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 & Muliere [1], 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 $\lim_{n \rightarrow \infty} \prod_{j=0}^n (\omega_j / (\beta_j + \omega_j)) = 0$.

Then, given a beta-Stacy process F with parameters $\{\beta_j, \omega_j\}_{j \in \mathbb{N}_0}$, and an LTRC sample (X_n^*, T_n, δ_n) , the posterior distribution of the r.v. X given by an RUP with reinforcement r reads

$$\hat{S}(x) := \mathbb{P}(X > x | X_n^*, T_n, \delta_n) = \prod_{j=0}^x \left[1 - \frac{\beta_j + r \cdot m_j(X_n^*, \delta_n)}{\beta_j + \omega_j + r \cdot s_j(X_n^*, T_n)} \right], \quad (3.1)$$

where $m_j(x_n, d_n)$ and $s_j(x_n, t_n)$ are defined as in equation (2.2).

Defining $\beta_j^* = \beta_j + r \cdot m_j^*(x_n, d_n)$ and $\omega_j^* = \omega_j + r \cdot (s_j(x_n, t_n) - m_j^*(x_n, d_n))$, we obtain a new beta-Stacy process F^* with parameters $\{\beta_j^*, \omega_j^*\}_{j \in \mathbb{N}_0}$, which implies that the beta-Stacy process—and hence also the RUP—is conjugate to LTRC data. This is proven formally in the following proposition, in which we extend the results of Muliere *et al.* [33] and Walker & Muliere [1], where the authors only focused on right-censored observations.

Proposition 3.2. *The RUP as defined in equation (3.1) is conjugate to LTRC data.*

Proof. We start by using the fact that the underlying distribution of the RUP is the generalized Dirichlet distribution of Connor & Mosimann [37], i.e.

$$\mathbb{P}(p_X(0), p_X(1), \dots, p_X(x_M) | \beta, \omega) \propto (p_X(x_M))^{\omega_{x_M}-1} \times \prod_{j=0}^{x_M-1} [(p_X(j))^{\beta_j-1} (S_X(j-1))^{\omega_{j-1}-(\beta_j+\omega_j)}], \quad (3.2)$$

where $p_X(\cdot)$ and $S_X(\cdot)$ denote the probability mass function (PMF) and survival function of X , respectively, and x_M is the maximum value X can take.

Under this distribution, the likelihood of the LTRC sample (X_n^*, T_n, δ_n) is given by

$$L(X_n^* | T_n, \delta_n) = \prod_{j=0}^{x_M} (p_X(j))^{m_j(X_n^*, \delta_n)} \prod_{j=0}^{x_M-1} (S_X(j))^{r_j(X_n^*, \delta_n) - l_j(T_n)}, \quad (3.3)$$

where $r_j(x_n, d_n) = \sum_{i=1}^n \mathbb{1}_{\{x_i=j, d_i=0\}}$ is the number of censored observations at j , and $l_j(t_n) = \sum_{i=1}^n \mathbb{1}_{\{t_i \leq j\}}$ is the number of truncated observations at j .

Since the posterior distribution is proportional to the product of equations (3.2) and (3.3), i.e. the product of the prior distribution and the likelihood, it follows immediately that if F is a discrete beta-Stacy process with parameters $\{\beta_j, \omega_j\}_{j \in \mathbb{N}_0}$, by setting $q_j = \sum_{i=j+1}^{x_M} m_i(X_n^*, \delta_n) + \sum_{i=j}^{x_M} r_i(X_n^*, \delta_n) - \sum_{i=1}^j l_i(T_n)$, the posterior distribution of F given the LTRC data (X_n^*, T_n, δ_n) is also a discrete beta-Stacy process with parameters

$$\beta_j^* = \beta_j + m_j(X_n^*, \delta_n) \quad \text{and} \quad \omega_j^* = \omega_j + q_j, \quad (3.4)$$

and since $q_j = s_j(X_n^*, T_n) - m_j(X_n^*, \delta_n)$, this shows that the RUP is conjugate to LTRC data. ■

Comparing with the urn representation of figure 1, $\hat{S}(x)$ in equation (3.1) corresponds to the probability of selecting at least x consecutive green balls, starting from urn U_0 . For $j=0, 1, 2, \dots$, the couple $\{\beta_j, \omega_j\}$ determines the initial number of red and green balls in urn U_j , respectively, while the functions $m_j^*(x_n, d_n)$ and $s_j(x_n, t_n)$ determine the extra number of green and ‘green plus red’ balls added to each urn over time, respectively.

Being a random distribution, an important characteristic of the (discrete) beta-Stacy process is that its trajectories can be centred around a certain (discrete) probability distribution $F_0(\cdot)$, which plays the role of the prior distribution. As proven in [1], a necessary condition for this to hold is that

$$\frac{\beta_j}{\beta_j + \omega_j} = \frac{F_0(j) - F_0(j-1)}{1 - F_0(j-1)}, \quad j \in \mathbb{N}, \quad (3.5)$$

where $F_0(j) = \mathbb{P}_{F_0}(X \leq j)$ is the probability that X is at most j under the prior F_0 .

Note that, according to equation (3.5), the choice for the couples $\{\beta_j, \omega_j\}_{j \in \mathbb{N}_0}$ is not unique. A common choice in the literature, also adopted in this paper, is to set

$$\beta_j = c_j F_0(\{j\}) \quad \text{and} \quad \omega_j = c_j (1 - F_0(j)), \quad c_j \in \mathbb{R}^+, j \in \mathbb{N}, \quad (3.6)$$

with c_j denoting the strength of belief in the prior knowledge and $F_0(\{j\}) = \mathbb{P}_{F_0}(X = j)$. 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 prior distribution, except with large amounts of data. On the contrary, when $c_j \rightarrow 0$, equation (3.1) reduces to the KM estimator of Cox & Oakes [31], which is unaffected by the choice of F_0 .

Finally, observe that the roles of the strength of belief parameters c_j and of the reinforcement parameter r are actually opposite. It is, therefore, possible to fix one of them and just work with the remaining one.

3.2. The bivariate RUP of Bulla *et al.* [2]

As mentioned in §3.1, a RUP is a one-dimensional process. In order to cope with two dimensions, a bivariate extension (B-RUP) was first proposed in [2], using a one-factor construction.

Assume we have a sample of bivariate LTRC observations of the form $((X_n^*, T_n^X, \delta_n^X), (Y_n^*, T_n^Y, \delta_n^Y))$ about two variables of interest X and Y . As before, T_n^X and T_n^Y are the truncation variables for X and Y , respectively. A simple way of modelling the dependence between X and Y is to consider a one-factor construction, based on three independent components: one common and two idiosyncratic factors for X and Y . Thus, let A , B and C be independent r.v.s and set

$$\text{and} \quad \left. \begin{aligned} X &= A + B \\ Y &= A + C. \end{aligned} \right\} \quad (3.7)$$

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

$$\text{Cov}(X, Y|A, B, C) = \text{Var}(A). \quad (3.8)$$

Therefore, the dependence between X and Y is linear and non-negative.

The idea of Bulla *et al.* [2] was to model the distributions of A , B and C with RUPs using equation (3.1) in order to exploit the properties of the RUP and avoid the caveats of bivariate non-parametric estimators already mentioned in §1. The resulting joint distribution F_{XY} of X and Y is then obtained through convolutions. We refer to Bulla *et al.* [2] 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 §4, this will enable us to derive a simple and efficient iterative method to estimate F_{XY} .

Remark. The variables A , B and C are artificially created, and therefore there are no observations of these variables in practice. This poses the problem of estimating the distributions of A , B and C given observations of X and Y . We show in §4 that this can be considered as a type of incomplete data problem, and thus suitable for the EM and ER algorithms.

4. The expectation–maximization and the expectation–reinforcement algorithms

The EM algorithm was introduced in [17] to unify several, seemingly different, methodologies that extract information from incomplete data (see, for example, [38]). From a mathematical point of view, the algorithm computes the expectation of the complete log-likelihood $L(x|y, \theta)$ at each iteration, conditioned on the observed incomplete data y , and the estimates of the parameters from the previous iteration θ . Then, it maximizes this expectation in order to find the optimal parameters θ^* for the next iteration. This iteration procedure is repeated until some stopping criterion is met. In the context of this article, the incomplete data corresponds to the triplets $((X_n^*, T_n^X, \delta_n^X), (Y_n^*, T_n^Y, \delta_n^Y))$, while the complete data would be uncensored observations (A_n, B_n, C_n) of the processes that conform the B-RUP. The set of parameters θ thus corresponds to the distributions of A , B and C .

Before explaining the details of the proposed methodology, we introduce some notation to facilitate the reading:

$$p_X(x) := \mathbb{P}(X = x|\theta), \quad (4.1)$$

and

$$p_X^{[k]}(x) := \mathbb{P}(X = x|\theta^{[k]}), \quad (4.2)$$

where θ is again the set of parameters and the superscript k the iteration number in the EM algorithm. We also introduce the survival function $S_X(x) := \mathbb{P}(X > x|\theta)$, and $S_X^{[k]}(x)$ analogously.

As mentioned in §3, the main idea of the B-RUP is to model the distributions of (A, B, C) as individual, independent RUPs. In the following sections, we apply a similar idea, and define the distributions of (A, B, C) in terms of urns. In §4.2, it is then shown how the resulting models are related to the original RUP model.

Since the variables (A, B, C) are independent, we start defining the urn distribution for variable A , with the distributions for variables B and C defined analogously.

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

$$S_A(a) = \prod_{j=0}^a \frac{G_j^A}{N_j^A}, \quad a \in \mathbb{N}_0, \quad (4.3)$$

with the convention $\prod_{j=0}^a = 1$ for $a < 0$. Following the urn representation of §3, G_j^A denotes the number of green balls in urn U_j , and N_j^A the total number of balls in the same urn. These pairs $(G_j^A, N_j^A)_{j \in \mathbb{N}_0}$ are the variables we wish to calibrate via the EM algorithm. In other words, $\theta = (G_j^A, N_j^A, G_j^B, N_j^B, G_j^C, N_j^C)_{j \in \mathbb{N}_0}$. However, in the presence of left-truncation θ needs to be enlarged to account for the truncation variables. This is explained in §4.1.

Remark. Defining the pair (G_i^A, N_i^A) is not really necessary. Given equation (4.3), it is the ratio between G_i^A and N_i^A that characterizes the distribution of A . The pair is relevant, however, in the context of §4.2, as this distinction is what allows the inclusion of prior knowledge and a complete analogy with the RUP.

4.1. Expectation–maximization for LTRC data

For simplicity, we assume first that there is no left-truncation. That is, the observations are the pairs (X_n, δ_n^X) and (Y_n, δ_n^Y) , and we want to estimate the distributions of (A, B, C) . Due to the one-factor construction of equation (3.7), we can write the log-likelihood of (A, B, C) as the sum of the log-likelihoods of A, B and C :

$$\log \mathbb{P}(A = a, B = b, C = c) = \log p_A(a) + \log p_B(b) + \log p_C(c). \quad (4.4)$$

We present only the results for A , with the results for B and C following analogously. The expected complete log-likelihood of A at the $(k+1)$ th iteration, given a single observation $(x, y, \delta^X, \delta^Y)$ of (X, Y) and the estimates from the k th iteration, is given by

$$L_A(\theta^{[k+1]} | \theta^{[k]}) = \sum_{a=0}^{\infty} \log p_A^{[k+1]}(a) p_A^{[k]}(a | x, y, \delta^X, \delta^Y), \quad (4.5)$$

where $x \wedge y = \min(x, y)$,

$$p_A(a | x, y, 1, 1) = \frac{p_A(a) p_B(x-a) p_C(y-a)}{p_{XY}(x, y)}, \quad (4.6)$$

$$p_A(a | x, y, 0, 1) = \frac{p_A(a) S_B(x-a) p_C(y-a)}{\mathbb{P}(X > x, Y = y)}, \quad (4.7)$$

$$p_A(a | x, y, 1, 0) = \frac{p_A(a) p_B(x-a) S_C(y-a)}{\mathbb{P}(X > x, Y = y)}, \quad (4.8)$$

$$p_A(a | x, y, 0, 0) = \frac{p_A(a) S_B(x-a) S_C(y-a)}{\mathbb{P}(X > x, Y > y)}, \quad (4.9)$$

and

$$p_{XY}(x, y) = \sum_{a=0}^{x \wedge y} p_A(a) p_B(x-a) p_C(y-a). \quad (4.10)$$

Remark. The lower limit of the summation in equation (4.10) is zero because we work with non-negative processes. The upper limit is also a consequence of non-negativity and the fact that A cannot be bigger than X or Y , due to equation (3.7).

The next step is to compute the derivatives of equation (4.5) with respect to $\theta^{[k+1]}$ and set them equal to zero to obtain the values of the next iteration. Given equation (4.3), it is easy to see that:

$$\frac{\partial \log p_A(a)}{\partial G_j^A} = \begin{cases} 0 & \text{if } j > a \\ \frac{-1}{N_j^A - G_j^A} & \text{if } j = a \\ \frac{1}{G_j^A} & \text{if } j < a. \end{cases} \quad (4.11)$$

Combining equations (4.11) and (4.5) yields

$$\frac{\partial L_A(\theta^{[k+1]}|\theta^{[k]})}{\partial G_j^{[k+1],A}} = \frac{S_A^{[k]}(j|x, y, \delta^X, \delta^Y)}{G_j^{[k+1],A}} - \frac{p_A^{[k]}(j|x, y, \delta^X, \delta^Y)}{N_j^{[k+1],A} - G_j^{[k+1],A}}. \quad (4.12)$$

Setting this last expression equal to zero and solving for $G_j^{[k+1],A}$ yields the value for the next iteration:

$$\frac{G_j^{[k+1],A}}{N_j^{[k+1],A}} = \frac{S_A^{[k]}(j|x, y, \delta^X, \delta^Y)}{S_A^{[k]}(j-1|x, y, \delta^X, \delta^Y)}, \quad (4.13)$$

where we have used $S_A(j) + p_A(j) = S_A(j-1)$, since A is discrete with jumps of size one.

In the case of n observations, equation (4.13) becomes:

$$\frac{G_j^{[k+1],A}}{N_j^{[k+1],A}} = \frac{\sum_{i=1}^n S_A^{[k]}(j|x_i, y_i, \delta_i^X, \delta_i^Y)}{\sum_{i=1}^n S_A^{[k]}(j-1|x_i, y_i, \delta_i^X, \delta_i^Y)}. \quad (4.14)$$

The solutions for G_j^B and G_j^C are completely analogous, and their respective formulae are provided below:

$$p_B(b|x, y, 0, 1) = \frac{p_B(b) \sum_{a=x-b+1}^y p_A(a) p_C(y-a)}{\mathbb{P}(X > x, Y = y)}, \quad (4.15)$$

$$p_B(b|x, y, 1, 0) = \frac{p_B(b) p_A(x-b) S_C(y-x+b)}{\mathbb{P}(X = x, Y > y)}, \quad (4.16)$$

$$p_B(b|x, y, 0, 0) = \frac{p_B(b) \sum_{a=x-b+1}^{\infty} p_A(a) S_C(y-a)}{\mathbb{P}(X > x, Y > y)}, \quad (4.17)$$

$$p_C(c|x, y, 1, 0) = \frac{p_C(c) \sum_{a=y-c+1}^x p_A(a) p_B(x-a)}{\mathbb{P}(X = x, Y > y)}, \quad (4.18)$$

$$p_C(c|x, y, 0, 1) = \frac{p_C(c) p_A(y-c) S_B(x-y+c)}{\mathbb{P}(X > x, Y = y)}, \quad (4.19)$$

and

$$p_C(c|x, y, 0, 0) = \frac{p_C(c) \sum_{a=y-c+1}^{\infty} p_A(a) S_B(x-a)}{\mathbb{P}(X > x, Y > y)}. \quad (4.20)$$

The final step is to consider bivariate left-truncation. In this case, we observe $(x, y, \delta^X, \delta^Y, t^X, t^Y)$ if $(x \geq t^X)$ and $(y \geq t^Y)$ and nothing otherwise. The complete data, therefore, consists of observations $(x, y, \delta^X, \delta^Y, t^X, t^Y)$ regardless of whether $(x \geq t^X)$ and $(y \geq t^Y)$ is verified. If we denote the truncation event by \mathcal{A} , the expectation of the complete log-likelihood of A for a sample of size n becomes

$$L_A(\theta^{[k+1]}|\theta^{[k]}) = \sum_{i=1}^n \sum_{a=0}^{\infty} \log p_A^{[k+1]}(a) p_A^{[k]}(a|x_i, y_i, \delta_i^X, \delta_i^Y) + (M^{[k]} - n) \sum_{a=0}^{\infty} \log p_A^{[k+1]}(a) p_A^{[k]}(a|\mathcal{A}), \quad (4.21)$$

where, in general,

$$M^{[k]} = \frac{n}{p^{[k]}(\mathcal{A}^c)} \quad (4.22)$$

is the total number of samples. That is, the number of observations (n) plus the unobserved samples due to left-truncation.

\mathcal{A} consists of three different events: $(T^X \leq X, T^Y > Y)$, $(T^X > X, T^Y \leq Y)$ and $(T^X > X, T^Y > Y)$. On the other hand, the complementary of \mathcal{A} , i.e. the observation event, is defined as

$$\mathcal{A}^c = (T^X \leq X, T^Y \leq Y). \quad (4.23)$$

Given that

$$p(\cdot) = p(\cdot|\mathcal{A})p(\mathcal{A}) + p(\cdot|\mathcal{A}^c)p(\mathcal{A}^c), \quad (4.24)$$

it is straightforward to compute $p_A(a|\mathcal{A}^c)$ and then use equation (4.24) to condition on \mathcal{A} . For that purpose, we also need the probability of the observation event, which can be computed as

$$p(\mathcal{A}^c) = \sum_{x=0}^{\infty} \sum_{y=0}^{\infty} p_{XY}(x, y) \mathbb{P}(T^X \leq x, T^Y \leq y), \quad (4.25)$$

where $p_{XY}(x, y)$ is defined as in equation (4.10) and, using the truncation assumptions of §2,

$$p(T^X \leq x, T^Y \leq y) = \sum_{t=0}^x p_{T^X}(t) p_{\epsilon}(\epsilon \leq y - t), \quad (4.26)$$

with p_{T^X} and p_{ϵ} the probability distributions of T^X and ϵ , respectively.

Remark. From equation (4.25), it is clear that we need the distribution of the truncation variables (T^X, T^Y) . This distribution can be computed from the marginal distributions of T^X and ϵ . We assume that these are also urn distributions as in equation (4.3). Therefore, as mentioned previously in this section, the total set of parameters to be estimated becomes $\theta = (G_j^A, N_j^A, G_j^B, N_j^B, G_j^C, N_j^C, G_j^{T^X}, N_j^{T^X}, G_j^{\epsilon}, N_j^{\epsilon})_{j \in \mathbb{N}_0}$ in the presence of left-truncation.

The summation over i in equation (4.21) does not depend on the truncation event, and thus it can be evaluated using equations (4.6)–(4.9). On the other hand, the truncation component in equation (4.21) can be computed using equations (4.24) and (4.25) and

$$p_A(a|\mathcal{A}^c) = \frac{p_A(a)}{p(\mathcal{A}^c)} \sum_{b=0}^{\infty} \sum_{c=0}^{\infty} p_B(b) p_C(c) \mathbb{P}(T^X \leq a + b, T^Y \leq a + c). \quad (4.27)$$

The optimal configuration $\theta^{[k+1]}$ of the next iteration is then obtained by applying the derivative operator to equation (4.21) and then using equation (4.11). This yields

$$\frac{G_j^{[k+1],A}}{N_j^{[k+1],A}} = \frac{\sum_{i=1}^n S_A^{[k]}(j|x_i, y_i, \delta_i^X, \delta_i^Y) + (M^{[k]} - n) S_A^{[k]}(j|\mathcal{A})}{\sum_{i=1}^n S_A^{[k]}(j-1|x_i, y_i, \delta_i^X, \delta_i^Y) + (M^{[k]} - n) S_A^{[k]}(j-1|\mathcal{A})}. \quad (4.28)$$

The results for B and C are completely analogous, and the necessary formulae to compute the estimates are provided in equations (4.15)–(4.20), (4.29) and (4.30). For the truncation variables, equation (4.28) is actually less involved, since these variables do not suffer from censoring. The conditional probabilities are given in equations (4.31) and (4.32), and the estimators in equations (4.33) and (4.34), respectively.

$$p_B(b|\mathcal{A}^c) = \frac{p_B(b)}{p(\mathcal{A}^c)} \sum_{a=0}^{\infty} \sum_{c=0}^{\infty} p_C(c) p_A(a) \mathbb{P}(T^X \leq a + b, T^Y \leq a + c), \quad (4.29)$$

$$p_C(c|\mathcal{A}^c) = \frac{p_C(c)}{p(\mathcal{A}^c)} \sum_{a=0}^{\infty} \sum_{b=0}^{\infty} p_B(b) p_A(a) \mathbb{P}(T^X \leq a + b, T^Y \leq a + c), \quad (4.30)$$

$$p_{T^X}(t|\mathcal{A}^c) = \frac{p_{T^X}(t)}{p(\mathcal{A}^c)} \sum_{e=-\infty}^{\infty} p_{\epsilon}(e) \mathbb{P}(X \geq t, Y \geq t + e), \quad (4.31)$$

$$p_{\epsilon}(e|\mathcal{A}^c) = \frac{p_{\epsilon}(e)}{p(\mathcal{A}^c)} \sum_{t=0}^{\infty} p_{T^X}(t) \mathbb{P}(X \geq t, Y \geq t + e), \quad (4.32)$$

$$\frac{G_j^{[k+1],T^X}}{N_j^{[k+1],T^X}} = \frac{\sum_{i=1}^n \dot{\{t_i > j\}} + (M^{[k]} - n) S_{T^X}^{[k]}(j|\mathcal{A})}{\sum_{i=1}^n \dot{\{t_i \geq j\}} + (M^{[k]} - n) S_{T^X}^{[k]}(j-1|\mathcal{A})}, \quad (4.33)$$

$$\text{and} \quad \frac{G_j^{[k+1],\epsilon}}{N_j^{[k+1],\epsilon}} = \frac{\sum_{i=1}^n \dot{\{e_i > j\}} + (M^{[k]} - n) S_{\epsilon}^{[k]}(j|\mathcal{A})}{\sum_{i=1}^n \dot{\{e_i \geq j\}} + (M^{[k]} - n) S_{\epsilon}^{[k]}(j-1|\mathcal{A})}, \quad (4.34)$$

where $T_n^X = (t_1, \dots, t_n)$ and $\epsilon_n = (e_1, \dots, e_n)$ are the observed samples of T^X and ϵ , respectively.

4.2. The expectation-reinforcement algorithm

The ER algorithm aims at combining the reinforcement mechanism of RUPs with the EM algorithm offering the possibility of embedding prior knowledge and experts' judgements into the estimates. This feature is particularly useful in the modelling of extreme events [39], epistemic uncertainty

[40,41], or when there are possible biases in the sample, as is the case with LTRC data. For example, it is well-known that the left-truncation effect produces a positive bias in the average lifetime of an individual in medical and survival studies, making the average lifetime look larger than it actually is. An expert familiar with this phenomenon can correct this bias by selecting a prior distribution with a smaller expected value than the one observed in-sample. We illustrate an example of such bias in §5.2, where we analyse the consequences of ignoring left-truncation in the estimators. In the context of extreme events that, due to their nature, are rarely present or not present at all in the data, this bias can also be corrected by choosing a prior distribution that gives a larger weight to these events. For example, due to the right-censoring effect, it is extremely rare to observe unusually large expected lifetimes in survival studies. Hence, in order to incorporate such unusual lifetimes in the posterior distribution, this information can be extracted, for example, from mortality tables and included in the prior distribution, correcting the final estimates.

Moreover, non-parametric estimators usually suffer from overfitting [42]. Such a problem occurs when the model calibrates too well to the sample data, making the procedure highly sensitive to small variations in the sample properties, and thus reducing its predictive power out-of-sample. From the bias-variance trade-off point of view, non-parametric estimators tend to have a very small bias, as they capture all the features of the dataset. On the other hand, 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, this trade-off can be controlled as follows: for high strengths of belief (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. In the opposite case, with almost zero strength of belief (or strong reinforcement), the posterior will adapt to the data as much as possible. In the first scenario, the variance of the model with respect to different datasets is zero, but the bias will be arbitrarily high depending on the choice of the prior distribution. In the second scenario, the bias should be considerably small, but the model will be very sensitive to small changes in the data, resulting in a large variance. Thus the trade-off can be balanced by choosing intermediate values of the strength of belief and reinforcement parameters. Nevertheless, non-parametric estimators also have other ways of dealing with the bias-variance trade-off, such as the stopping criteria of the optimization algorithm, or the use of smoothing kernels on the final estimates (see [20]).

The goal of the ER algorithm is to combine a prior distribution, given by the pairs $\{\beta_j, \omega_j\}_{j \in \mathbb{N}_0}$ defined in equation (3.6), with the estimates obtained from the EM algorithm, so as to obtain a posterior distribution that mixes experts' knowledge and data.

In the following, we illustrate how to include a prior distribution to the estimates in equation (4.28). The same procedure can be applied analogously to the other variables (B, C, T^X, ϵ) .

Let the pairs $\{\beta_j^A, \omega_j^A\}_{j \in \mathbb{N}_0}$ define a prior distribution through equation (3.6). Next, we follow the steps described in §4.1 to obtain the EM estimates of A . Then, we define the posterior configuration of A using the following expression

$$G_j^A = \omega_j^A + r \left[\sum_{i=1}^n S_A^{\text{EM}}(j | x_i, y_i, \delta_i^X, \delta_i^Y) + (M^{\text{EM}} - n) S_A^{\text{EM}}(j | \mathcal{A}) \right], \quad (4.35)$$

and

$$N_j^A = \beta_j^A + \omega_j^A + r \left[\sum_{i=1}^n S_A^{\text{EM}}(j - 1 | x_i, y_i, \delta_i^X, \delta_i^Y) + (M^{\text{EM}} - n) S_A^{\text{EM}}(j - 1 | \mathcal{A}) \right], \quad (4.36)$$

where r is the reinforcement parameter defined in §3.1, and the EM superscript indicates the estimates obtained with the EM algorithm.

By giving different values to the reinforcement and belief parameters,² we can control the weight of each component on the estimates. Similarly to the behaviour shown in §3.1, when the strength of belief tends to zero, we recover the estimates of the EM algorithm, while if the reinforcement tends to zero instead, the posterior distribution equals the prior distribution.

The pseudocode to implement the ER algorithm can be found in algorithm 1. It refers to the equations related to the estimates of variable A , but analogous formulae for the other variables can be found in §4.1.

²By equation (3.6) 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 $\{\beta, \omega\}$ for (A, B, C, T^X, ϵ) according to equation (3.6).
 Choose initial estimates for the first iteration of the ER algorithm.
while *stop_criteria* = *False* **do**
 Update the distributions using equation (4.3) with the estimates from the previous iteration.
 Compute $M^{[k]}$ via equation (4.22).
 Compute the conditional probabilities from equations (4.6)–(4.9) and (4.27).
 Compute new estimates using equation (4.28).
end while
 Compute the posterior distribution using equations (4.35) and (4.36).

4.3. Error analysis

In this section, we comment on the numerical errors incurred when approximating the infinite summations appearing in the formulae of §4.1.

First, we start considering the case without left-truncation. This is the same as computing equation (4.28) (and the analogous formulae for B and C) with $M^{[k]} = n$. In that case, the formulae to be computed are equations (4.6)–(4.9). For that purpose, the distribution of (X, Y) must be known for all observed values. Therefore, there is no need to compute the distribution of (X, Y) for values outside of the observation range. Denoting by x_{\max} and y_{\max} the maximum observed values of X and Y , respectively, and taking into account equation (4.10), it is clear that the distributions of A , B and C do not need to be computed for values greater than $(x_{\max} \vee y_{\max})$, x_{\max} and y_{\max} , respectively, where \vee is the maximum operator.

In the presence of left-truncation, equations (4.31) and (4.32) require the entire survival distribution of (X, Y) , hence the distribution of (X, Y) for values greater than x_{\max} and y_{\max} is also needed. We assume that the distributions of T^X and ϵ are limited to the intervals $[0, T_{\max}^X]$ and $[\epsilon_{\min}, \epsilon_{\max}]$, respectively, since there are no observations outside of these ranges. Next, assume we compute the distribution of (X, Y) only in the domain $[0, x_M] \times [0, y_M]$. Then equation (4.31)—the analysis of equation (4.32) is analogous, and therefore omitted—can be approximated as

$$p_{T^X}(t|\mathcal{A}^c) \approx \hat{p}_{T^X}(t|\mathcal{A}^c) := \frac{p_{T^X}(t)}{p(\mathcal{A}^c)} \sum_{e=\epsilon_{\min}}^{\epsilon_{\max} \wedge (y_M - t)} p_{\epsilon}(e) \mathbb{P}(X \geq t, Y \geq t + e), \quad (4.37)$$

where, due to the nature of the left-truncation effect, $T_{\max}^X \leq x_{\max}$, and thus only the upper bound y_M affects equation (4.37).

The error incurred by this approximation is given by

$$p_{T^X}(t|\mathcal{A}^c) - \hat{p}_{T^X}(t|\mathcal{A}^c) = \frac{p_{T^X}(t)}{p(\mathcal{A}^c)} \sum_{e=y_M - t + 1}^{\epsilon_{\max}} p_{\epsilon}(e) \mathbb{P}(X \geq t, Y \geq t + e), \quad (4.38)$$

which can be upper-bounded by

$$p_{T^X}(t|\mathcal{A}^c) - \hat{p}_{T^X}(t|\mathcal{A}^c) \leq \frac{p_{T^X}(t)}{p(\mathcal{A}^c)} S_{\epsilon}(y_M - t) \mathbb{P}(X \geq t, Y \geq y_M) =: \Omega_{y_M}(t), \quad (4.39)$$

for $t \geq y_M - \epsilon_{\max}$ and zero otherwise.

For large enough values of y_M , the convergence of equation (4.39) towards zero with respect to y_M depends mostly on the right-tail of the distribution of Y . This, in turn, depends on the differentiability of the marginal distribution of Y . If the first k derivatives are nonzero, then the order of convergence is $\mathcal{O}(y_M^{-k-1})$. For infinitely differentiable distributions, the order of convergence is exponential, i.e. $\mathcal{O}(e^{-\gamma y_M})$, for some $\gamma, r > 0$.

We illustrate this using the same dataset as in §5.1, in which Y follows a Poisson distribution with parameter $\lambda_Y = 65$. In figure 2, we compute $\Omega_{y_M} := \sum_{t=0}^{T_{\max}^X} \Omega_{y_M}(t)$ for several values of y_M and compare it with the theoretical convergence of the Poisson distribution, which is $\mathcal{O}(e^{-y_M^2/\lambda_Y})$. The results confirm the exponential convergence of Ω_{y_M} in the Poisson example. For reference, in this example $y_{\max} = 85$, thus in this case it is not necessary to choose y_M much larger than y_{\max} to obtain good approximations. On the other hand, if the differentiability of the distribution of (X, Y) is expected to be small, large values of y_M should be used to guarantee convergence. In particular, choosing $y_M = T_{\max}^X + \epsilon_{\max}$ guarantees that $\Omega_{y_M} = 0$.

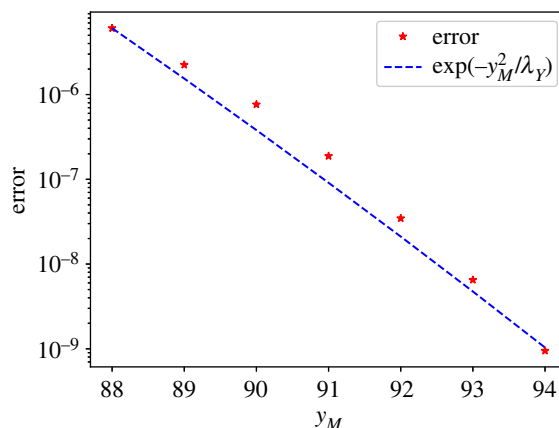


Figure 2. Empirical versus theoretical convergence of Ω_{y_M} in the Poisson simulated example.

Table 1. Proposed scenarios defined by the values of the belief and reinforcement parameters. The values are used throughout §5. The superscripts in the ER columns denote ‘low’ and ‘high’, respectively, referring to the weight of the belief parameters. Here r is again the reinforcement parameter and c is the strength of belief parameter.

	EM	ER ^l	ER ^h
r	—	10^4	1
c	0	1	10

5. Numerical results

We perform several numerical experiments to test the performance of the algorithms developed in §4. We start in §5.1 with an example that uses simulated data for which the reference distributions are known. Then, in §5.2, we analyse a Canadian dataset of coupled lifetimes widely used in the literature of actuarial sciences in the context of joint annuity evaluation [20–22]. This dataset is known for its complexity, due to the strong presence of censoring and truncation. Since the reference distribution for this problem is not known, we compare the results of the ER algorithm with the Frank copula of Frees *et al.* [21], which has offered satisfactory results. Moreover, we compare our results with those of Arias & Cirillo [20], where a B-RUP was first used, but only assuming right-censoring, that is, without left-truncation. This comparison allows us to study the impact of left-truncation on the empirical dataset, and to illustrate the advantages of the ER algorithm over the Markov chain Monte Carlo approach used in [20].

Remark. Note that, in the context of non-parametric estimators, it is not possible to obtain information about the distribution of (X, Y) for values outside the intervals $(\max(X_n^*|\delta_n^X = 1) \geq X \geq \min(X_n^*|\delta_n^X = 1))$ and $(\max(Y_n^*|\delta_n^Y = 1) \geq Y \geq \min(Y_n^*|\delta_n^Y = 1))$, where the condition $\delta_n = 1$ implies that we refer to uncensored observations. In other words, the final estimates correspond to the distribution of (X, Y) within those intervals.

In order to study the impact of the prior distribution on the final estimates, we distinguish between two scenarios: *low* strength of belief and *high* strength of belief. In the remainder of this work, we denote 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 assume the same strengths of belief for all urns. In practice, it could be 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 where observations are sparse. Regarding the stopping criterion, we have chosen the absolute value of the relative difference between the incomplete log-likelihood (see equation (2.3)) of the current and previous iterations. When this quantity is below a certain threshold, or the number of iterations reaches a prespecified limit, the

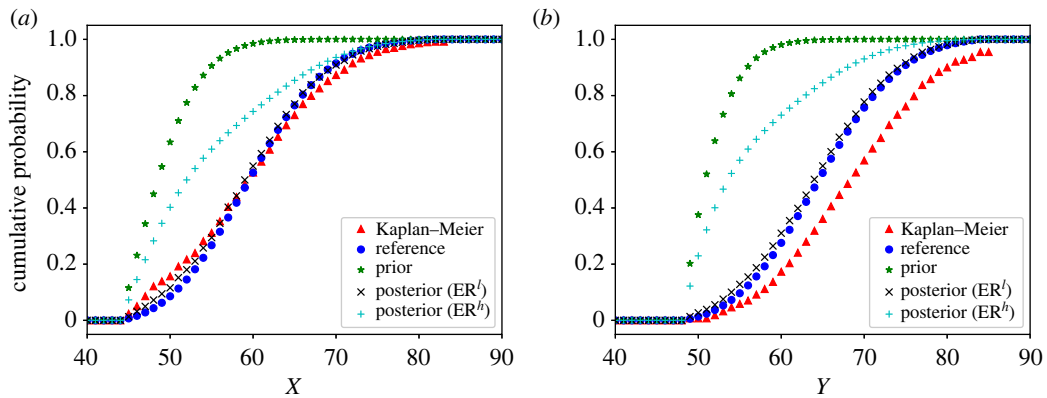


Figure 3. Fitting of the marginal distributions of X (a) and Y (b) in the Poisson simulated example.

algorithm stops. For all experiments considered in this paper, we have chosen the threshold to be 10^{-9} and the maximum number of iterations as 10^4 .

Remark. The algorithm was implemented in C++ using the g++ compiler (v. 9.4.0) and is freely available in the GitHub repository: <https://github.com/LuisSouto/Expectation-Reinforcement>. Experiments were run using an Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz processor.

5.1. Example with simulated data

The first experiment we consider consists of a simulated dataset where the observations come from a one-factor model as in equation (3.7). Hence, the model assumptions are verified and it is expected that the ER algorithm yields reasonable results. We refer to this experiment as the Poisson simulated example, since we assume Poisson distributions for all variables involved. Further experiments using a combination of discrete uniform distributions, Beta-Binomial distributions and Negative Hypergeometric distributions have also yielded positive results (not shown here) and are available upon request.

The data is generated using the following distributions: $A \sim \text{Poi}(40)$, $B \sim \text{Poi}(20)$, $C \sim \text{Poi}(25)$, $T \sim \text{Poi}(70)$, $\epsilon \sim \text{Poi}(7) - 5$ and $\Delta \sim \text{Poi}(2)$, where $\text{Poi}(\lambda)$ denotes a Poisson distribution with parameter λ . Therefore, in this case $X \sim \text{Poi}(60)$ and $Y \sim \text{Poi}(65)$, respectively, and the correlation between them is approximately 0.64. The reader is referred to §2.1 for a review of the censoring and truncation assumptions. The sample consists of 10^4 pairs of LTRC triplets $((X_n^*, T_n^X, \delta_n^X), (Y_n^*, T_n^Y, \delta_n^Y))$. The probability of truncation is approximately 82%, and the number of unobserved samples due to truncation is 53 096. Within the observed sample, the percentage of double censored observations is 65.73% and the percentage of observations with at least X or Y censored is 92.23%. Therefore, this is a dataset with both a high probability of censoring and truncation.

Following algorithm 1, we start by choosing a prior distribution for each variable. In this example, we have chosen $F_0^A = \text{Poi}(20)$, $F_0^B = \text{Poi}(20)$, $F_0^C = \text{Poi}(20)$, $F_0^T = \text{Poi}(50)$ and $F_0^{\epsilon+10} = \text{Poi}(10)$, where $F_0^{\epsilon+5} = \text{Poi}(10)$ implies that $\epsilon \sim \text{Poi}(10) - 5$ according to the prior distribution. The value -5 in the distribution of ϵ was inferred from its minimum observed value. The next step is the choice of the initial estimates. Both the ER and EM algorithms are local optimization algorithms, and whether a global or local minimum is reached, therefore, depends on the choice of the initial estimate. In practice, it is common to use several initial estimates to overcome this problem (see [43] and references therein), but we use the prior distributions as initial estimates for simplicity.

Remark. All results regarding this dataset are conditioned on $45 \leq X \leq 82$ and $49 \leq Y \leq 84$, since these are the maximum and minimum uncensored values observed. Moreover, the computational cost of the ER algorithm under the specified stopping criterion and hardware was 24 s.

In figure 3, we present the ER estimates of the marginal distributions of X and Y . For comparison purposes, the KM estimator and the prior distribution are also shown. Note that the KM estimator is clearly biased with respect to the reference distribution due to the presence of bivariate left-truncation. The plots also illustrate the impact of the strength of belief parameters on the final estimates. For low strengths of belief (ER^l), the distribution only takes the data into account, and matches well with the

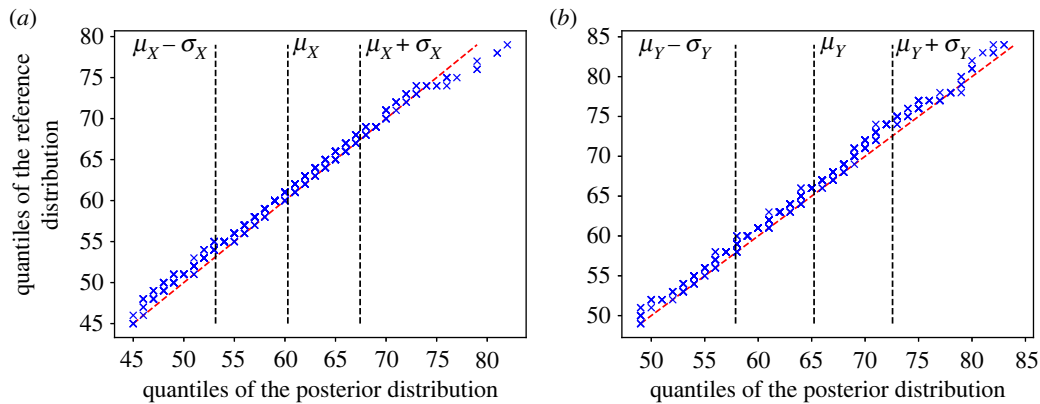


Figure 4. QQ-plot comparing the ER^l posterior and reference distributions. The dashed lines correspond to the mean (μ) and one standard deviation (σ) from the mean in each direction. (a) QQ-plot: X and (b) QQ-plot: Y .

Table 2. Permutation test results for samples from the ER^l marginal estimators. The first column gives the value of the test statistic, the second is the p -value and the last column checks whether the null hypothesis is rejected or not. The number of permutations is 10^5 in all tests.

	value	p (%)	H_0
Mean(X)	0.304	53.73	do not reject
Mean(Y)	0.722	14.49	do not reject
Var(X)	−8.891	7.81	do not reject
Var(Y)	1.195	82.71	do not reject

Table 3. Table with the mean values and variances of X and Y , the correlation between X and Y and the probability of truncation in the Poisson simulated example.

	mean(X)	mean(Y)	var(X)	var(Y)	corr(X , Y)	$p(\mathcal{A})$
reference	60.2875	65.2269	51.0608	53.8907	0.5939	0.8254
prior	49.8009	52.1530	15.5717	9.4091	0.1905	0.5551
ER^l	59.8433	64.7188	57.5964	56.2892	0.6402	0.8382
ER^h	55.1751	56.8845	72.3250	56.6427	0.7776	0.4266
KM	59.9198	68.1797	71.9409	61.5444	—	—

reference distribution. For high strengths of belief (ER^h), the distribution is more influenced by the prior knowledge, specially in areas with fewer observations.

A further analysis of this comparison can be found in figure 4, where we show the QQ-plots for the ER^l marginal distributions against the reference ones, and table 2, which contains the results of permutation tests for the mean values and the variances of X and Y also obtained with the ER^l estimates. The QQ-plots were generated using two samples of size 500 from each marginal of the ER^l posterior distribution, and comparing them with a sample of the same size from the reference solution. The same samples are also used to perform the permutation tests in table 2. We use the difference in mean values as test statistics for the mean values themselves, and the difference in variances as test statistics for the variances themselves. The null hypothesis assumes that the ER^l estimates match with the reference ones with a 5% confidence threshold.

Table 3 contains a comparison of the mean values, variances, correlation and probability of truncation obtained with each model. Compared to the reference distribution, the ER^l estimator generates the most similar values. The ER^h estimator aims for a compromise between the data and the prior distribution,

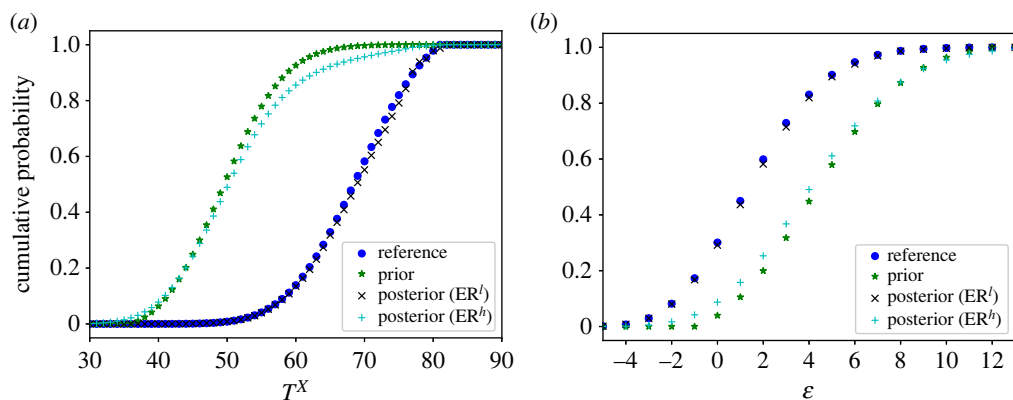


Figure 5. Fitting of the marginal distributions of T^X (a) and ϵ (b) in the Poisson simulated example.

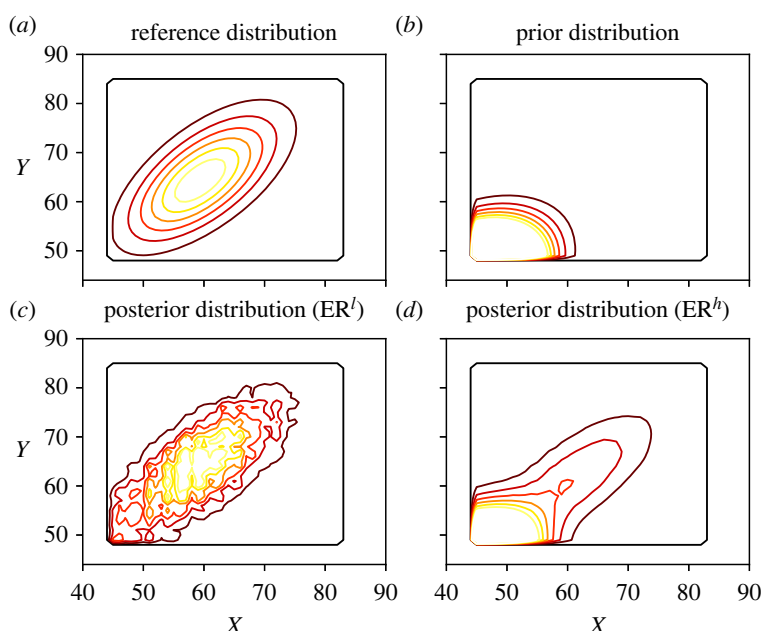


Figure 6. Contour plots of the reference, prior and posterior distributions in the Poisson simulated example. The levels for each curve are the same in all plots. On the upper row the reference and prior distributions are presented, while the lower row contains the ER posterior distributions with low (a,c) and high (b,d) strengths of belief.

which is reflected in the smaller mean values and larger variances. The KM estimator also gives reasonable results, although it is clearly biased, for example, in the mean of Y and the variance of X . Regarding the probability of truncation, the ER^l value is very close to the reference one, while the prior distribution underestimates the amount of truncation. A further comparison of the truncation variables can be seen in figure 5, where we show the marginal distributions of T^X and ϵ under the different models.

Finally, in figure 6, a contour plot of the reference, prior and posterior distributions is presented. Consistently with the previous results, the ER^l estimator is the closest to the reference distribution, while the ER^h estimator still presents features of both the prior distribution and the observed data. Note that the curves of the ER^l estimator are not smooth compared to the reference distribution, showing signs of overfitting. As mentioned in §4.2, this could be tackled using a prior distribution, modifying the stopping criteria or by using smoothing kernels.

5.2. Empirical data: coupled lifetimes

In the next experiment, we consider an empirical dataset of coupled lifetimes, which is widely used in the field of joint annuity modelling [21,22]. The data consists of almost 15 000 couples of clients of a

Table 4. Calibration of the Frank copula model using MLE. The subscripts X , Y refer to the marginals of the male and female annuitants, respectively.

μ_X	σ_X	μ_Y	σ_Y	α
84.809	9.926	87.575	7.792	−4.081

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), age at the end of the observation window, incomes, etc.

Following Luciano *et al.* [22], we remove same-sex contracts due to their scarcity in the dataset and define X and Y as the lifetime of males and females in the couple, respectively. In the same paper [22], the authors also mention that a couple may have entered into more contracts, and thus they may appear several times in the dataset. Therefore, we remove all repeated entries so that each couple is considered only once. We also remove entries where the annuitants were older than 100 years old at the beginning of the observation period. Finally, as in [21], we condition on couples that are at least 40 years old. This leaves us with a total of 11 420 male–female couples, of which only 197 are completely uncensored. Since the period of observation is, at most, 5 years, the effects of left-truncation and right-censoring are expected to be large when determining the underlying distribution. For an extended analysis of this dataset, we refer to Frees *et al.* [21] and Luciano *et al.* [22].

We analyse the effect of two different prior distributions: one that clearly underestimates the mean values and variances observed in the data (ER_1), and another that takes this information into account to generate a good starting point in the calibration (ER_2). For each choice of a prior distribution, we consider the same two scenarios introduced in table 1. Moreover, these results are compared with the Frank copula model defined in [21] in order to analyse the performance of the proposed estimators and with the results of Arias & Cirillo [20] to assess the impact of left-truncation on the posterior distribution.

Similar to the procedure followed in §5.1, we define first the prior and initial distributions for each variable. For simplicity, we assume that the prior and initial distributions coincide. For the first example we choose: $F_0^A = \text{Poi}(20)$, $F_0^B = \text{Poi}(20)$, $F_0^C = \text{Poi}(20)$, $F_0^{T^X} = \text{Poi}(50)$ and $F_0^{\epsilon-40} = \text{Poi}(10)$, where the constant in $\epsilon - 40$ was inferred from the maximum age difference observed in the dataset. According to this distribution, males and females have the same average lifetime of 40 years, with a standard deviation of almost 6.5 years. Moreover, the average age difference is 30 years, with a standard deviation of approximately 3 years. For the second example we choose: $F_0^A = \text{Poi}(40)$, $F_0^B = \text{Poi}(40)$, $F_0^C = \text{Poi}(45)$, $F_0^{T^X} = \text{Poi}(70)$ and $F_0^{\epsilon-40} = \text{Poi}(40)$. Under this choice of prior distribution, males and females have an average lifetime of 80 and 85 years, respectively, with a standard deviation of approximately 9 years. Also, the average age difference is zero, with a standard deviation of almost 6.5 years.

Regarding the copula model, Frees *et al.* [21] use Gompertz distributions for the individual lifetimes, and the Frank copula to model the dependence. The Gompertz distribution is given by

$$\text{Gomp}(x; \mu, \sigma) = 1 - \exp(e^{-(\mu/\sigma)}(1 - e^{(x/\sigma)})), \quad (5.1)$$

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

The Frank copula is defined as

$$C(u, v; \alpha) = \frac{1}{\alpha} \log \left(1 + \frac{(e^{\alpha u} - 1)(e^{\alpha v} - 1)}{e^{\alpha} - 1} \right), \quad (5.2)$$

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 $\alpha = 0$ indicates independence [44].

We follow the same procedure of Frees *et al.* [21] to estimate the model parameters, to which we refer for all details. In table 4, the optimal parameters obtained via maximum likelihood estimation (MLE) are presented, where (μ_X, σ_X) are the 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, which justifies the use of the B-RUP model.

Remark. All results regarding this dataset are conditioned on $51 \leq X \leq 99$ and $46 \leq Y \leq 98$, since these are the maximum and minimum uncensored values observed. Moreover, the computational cost of the

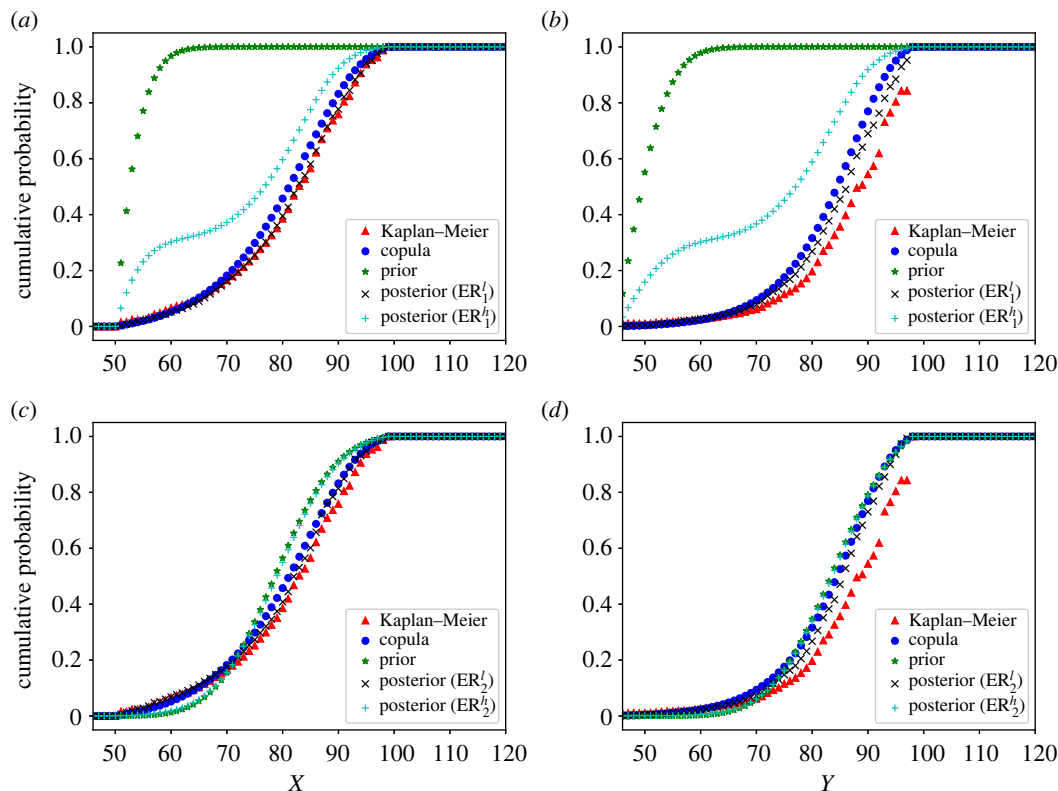


Figure 7. Marginal distributions of X and Y using the Canadian dataset. The upper row contains the ER_1 scenarios and the lower row the ER_2 scenarios. The copula and KM distributions are the same in both rows. (a) Marginal of X , (b) marginal of Y , (c) marginal of X and (d) marginal of Y .

Table 5. Table with the mean values and variances of X and Y , the correlation between X and Y and the probability of truncation using the Canadian dataset.

	mean(X)	mean(Y)	var(X)	var(Y)	corr(X , Y)	$p(\mathcal{A})$
prior ₁	53.8337	50.7572	8.1668	15.3958	0.1753	0.2919
prior ₂	79.0723	83.4871	69.1529	62.5827	0.4235	0.2668
ER ₁ ^l	81.5889	84.8201	113.6054	93.0486	0.5233	0.1840
ER ₁ ^h	72.9873	72.3154	204.6247	244.5786	0.9128	0.1161
ER ₂ ^l	80.6768	84.4110	117.8567	81.6205	0.3706	0.2070
ER ₂ ^h	79.2383	83.5579	73.0210	64.1661	0.4194	0.2594
copula	80.2418	83.3962	109.3485	83.5461	0.5145	—
KM	81.5813	87.0393	124.3929	100.5714	—	—

ER algorithm under the specified stopping criteria and hardware was 21 s for scenario ER_1 and 27 s for scenario ER_2 .

In figure 7, we show the marginal distributions of X and Y for all scenarios and models considered. As expected, the first prior distribution is considerably different from the observations, which affects significantly the posterior distribution in scenario ER_1^h . On the other hand, the posterior distribution in ER_1^l is affected only by the data, due to the choices of the reinforcement and strength of belief parameters. Note that the ER_1^l and ER_2^l estimators are almost equal, since they only depend on the observations and the choice of the initial distribution in the ER algorithm, but not on the prior

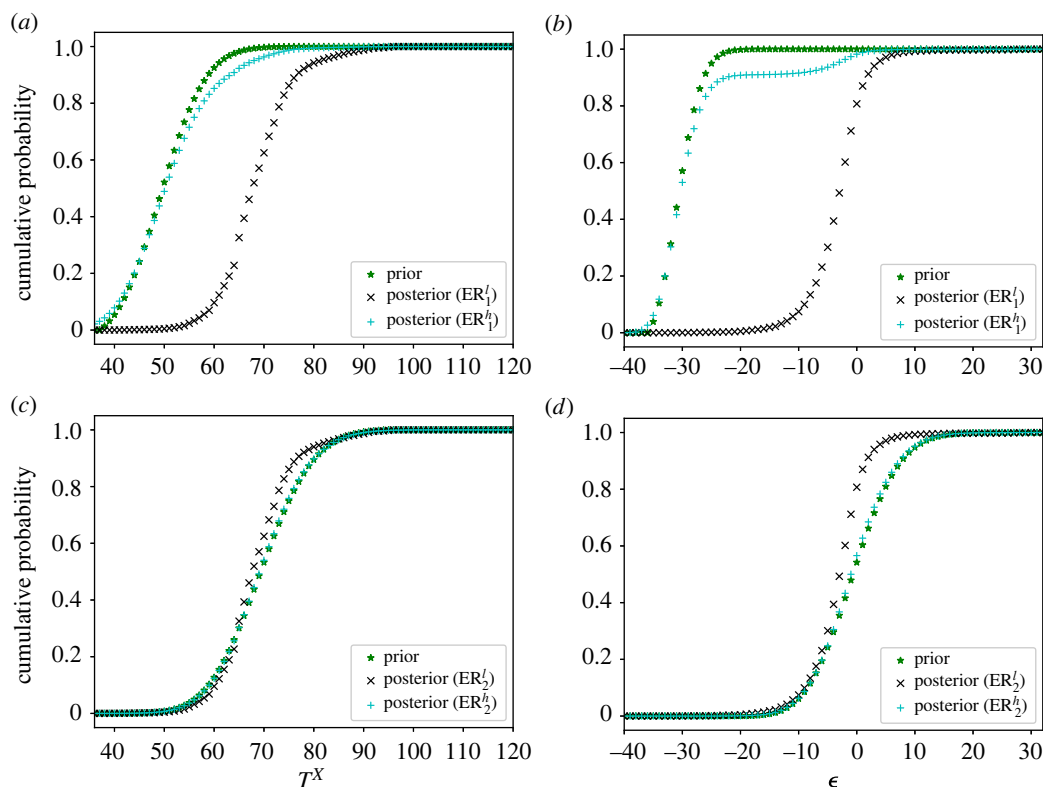


Figure 8. Fitting of the marginal distributions of T^X (a,c) and ϵ (b,d) in the Canadian dataset. The upper row contains the ER_1 scenarios and the lower row the ER_2 scenarios.

distribution. However, the ER_2^h estimator is significantly different from the ER_1^h one. Since the second prior distribution was chosen based on the data, it cannot be stated that the prior distribution brings in new information in this case. Nevertheless, it can be used as a reasonable starting point in the ER algorithm and to smoothen the posterior distribution. Regarding the copula model, the distributions obtained with the Frank copula and the ER algorithm are quite similar, signalling the robustness of the results obtained with both methods.

A quantitative comparison can be seen in table 5, which contains the mean values, variances, correlation and probability of truncation computed with each model. The results confirm the findings from figure 7. The quantities that are most affected due to the choice of a particular model are the variance of Y and the correlation between X and Y . The reason why variable Y appears to be more sensitive to each model, compared to X , may be the lack of information in the dataset. There are only 448 uncensored female lifetimes in the dataset, as compared to 1242 uncensored male observations. With respect to the dependence between X and Y , all models that are not strongly influenced by the choice of a prior distribution present levels of correlation around 0.5. This is in agreement with previous studies on this dataset. In particular, in comparison with the results of Arias & Cirillo [20], correlation does not appear to be highly affected by left-truncation. On the other hand, the expectation of X is around four years smaller if we take truncation into account, while the expectation of Y is reduced by 2 years. The variances are also significantly larger if we account for left-truncation, as it gives a larger weight to the left tail of the distribution. If we look at the probability of truncation, both the ER_1^l and ER_2^l estimators give a probability of almost 20%. Although this is notably smaller than the probability of truncation considered in §5.1, it is large enough to be non-negligible, in light of the comparison with the results of Arias & Cirillo [20]. The distributions of the truncation variables can be seen in figure 8, showing that the stationary assumption is not verified in this particular dataset.

Finally, in figures 9 and 10, the distributions obtained with the ER and copula methods are shown. The prior distributions are also shown for comparison. The conclusions are similar to those of the previous experiments: the ER^l estimates are very similar in both scenarios and barely affected by the choice of a prior distribution. On the other hand, the ER^h estimates are highly influenced by the prior distribution. In fact, since the second prior is based on the observations, the ER_2^h estimate shows

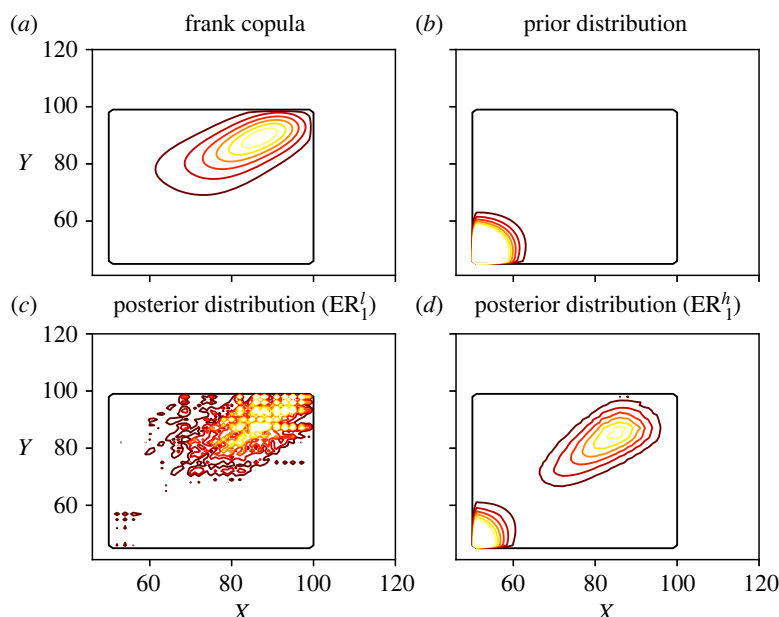


Figure 9. Contour plots of the reference, prior and posterior distributions using the Canadian dataset, under scenario ER_1 . The levels for each curve are the same in all plots. On the upper row the copula and prior distributions are presented, while the lower row contains the ER_1 posterior distributions with low (*a,c*) and high (*b,d*) strengths of belief.

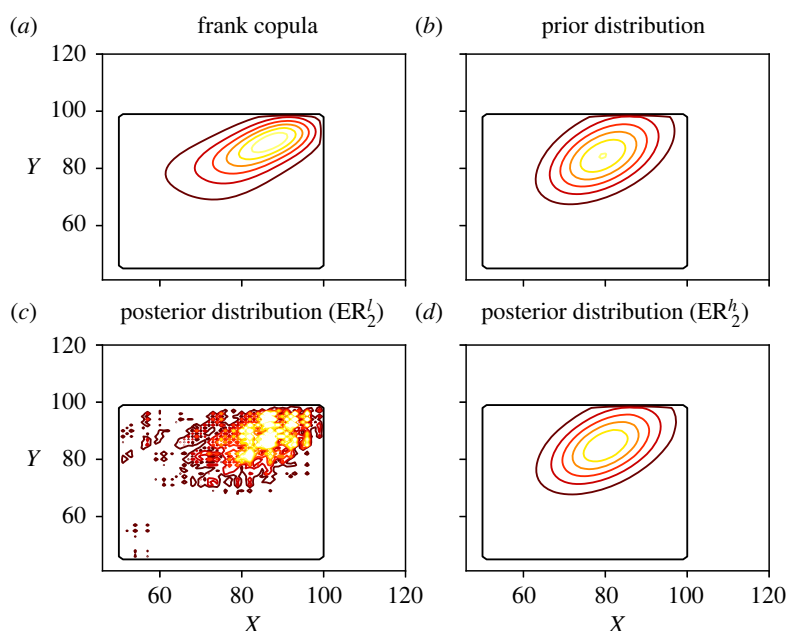


Figure 10. Contour plots of the reference, prior and posterior distributions using the Canadian dataset, under scenario ER_2 . The levels for each curve are the same in all plots. On the upper row the copula and prior distributions are presented, while the lower row contains the ER_2 posterior distributions with low (*a,c*) and high (*b,d*) strengths of belief.

hardly any difference with respect to this prior. Thus, the contrast between the ER_1^h and ER_2^h estimates illustrates how different prior distributions can be used to introduce different features in the posterior distribution. Compared to the copula model, the ER_1^l and ER_2^l estimates present some non-negligible contours in areas with a large age difference, due to their presence in the dataset. These observations may correspond to parent-child couples instead of co-living partners, which would be relevant to maintain in the distribution if these observations are of interest. Otherwise, smoothing techniques can be applied to reduce the effect of overfitting.

6. Conclusion

We have proposed a novel estimation approach for bivariate LTRC observations using RUPs and the EM algorithm. The algorithm returns not only the distribution of the observed pair (X, Y) , but also the distribution of the truncation variables. This can be of interest to analyse the mechanisms that generate biases in the observations, as well as to check whether the stationary condition can be applied. Analogously to the reference B-RUP model of Bulla *et al.* [2], the proposed ER algorithm benefits from the inclusion of experts' knowledge in the form of a prior distribution, following the Bayesian paradigm.

Performances have been tested using simulated and empirical LTRC data, showing that the algorithms are able to recover the reference distribution even under substantial amounts of left-truncation and right-censoring. The proposed methodology has also been compared to the Frank copula employed in [21], producing similar results, and thus confirming the reliability of the new approach. Lastly, the effect of left-truncation has been analysed by comparing with the results obtained in [20], which only takes right-censoring into account.

Future lines of work involve extending the one-factor model of Bulla *et al.* [2], to cope with multivariate situations. This extension is straightforward in the absence of censoring and truncation, but it is more involved in a realistic and general setting. Another relevant research direction should aim at generalizing the B-RUP to deal with other forms of dependence, in particular including negative correlation and nonlinear effects.

Data accessibility. This article has no additional data.

Authors' contributions. L.A.S.A.: formal analysis, investigation, methodology, software, validation, writing—original draft; P.C.: conceptualization, funding acquisition, investigation, project administration, resources, supervision, writing—review and editing; C.W.O.: conceptualization, funding acquisition, project administration, resources, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This research has been financed by the European Union, under the H2020-EU.1.3.1. MSCA-ITN-2018 scheme, grant no. 813261.

References

- Walker SG, Muliere P. 1997 Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Ann. Stat.* **25**, 1762–1780. (doi:10.1214/aos/1031594741)
- Bulla P, Muliere P, Walker SG. 2007 Bayesian nonparametric estimation of a bivariate survival function. *Stat. Sin.* **17**, 427–444.
- Angrist J, Bettinger E, Kremer M. 2006 Long-term educational consequences of secondary school vouchers: evidence from administrative records in Colombia. *Am. Econ. Rev.* **96**, 847–862. (doi:10.1257/aer.96.3.847)
- Antonio K, Badescu A, Gong L, Lin S, Verbelen R. 2015 Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bull.* **45**, 729–758. (doi:10.1017/asb.2015.15)
- Shen PS, Yan YF. 2008 Nonparametric estimation of the bivariate survival function with left-truncated and right-censored data. *J. Stat. Plan. Inference* **138**, 4041–4054. (doi:10.1016/j.jspi.2008.02.007)
- Klein JP, Moeschberger ML. 2003 *Survival analysis techniques for censored and truncated data*, 2nd edn. New York, NY: Springer.
- Campbell G, Földes A. 1982 Large sample properties of nonparametric statistical inference. In *Nonparametric statistical inference* (eds BV Gnedenko, ML Puri, I Vineze), pp. 103–122. Amsterdam, The Netherlands: North-Holland.
- Dabrowska DM. 1988 Kaplan–Meier estimate on the plane. *Ann. Stat.* **16**, 1475–1489. (doi:10.1214/aos/1176351049)
- Pruitt RC. 1991 *Strong consistency of self-consistent estimators: general theory and an application to bivariate survival analysis*, vol. 543. Minneapolis, MN: University of Minnesota.
- Pruitt RC. 1993 Small sample comparisons of six bivariate survival curve estimators. *J. Multivar. Anal.* **45**, 147–167. (doi:10.1080/00949659308811478)
- Pruitt RC. 1991 On negative mass assigned by the bivariate Kaplan–Meier estimator. *Ann. Stat.* **19**, 443–453. (doi:10.1214/aos/1176347992)
- Gribkova S, Lopez O. 2015 Non-parametric copula estimation under bivariate censoring. *Scand. J. Stat.* **42**, 925–946. (doi:10.1111/sjos.12144)
- Shen PS. 2014 Simple nonparametric estimators of the bivariate survival function under random left truncation and right censoring. *Comput. Stat.* **29**, 641–659. (doi:10.1007/s00180-013-0455-0)
- Lopez O. 2012 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. *Insur. Math. Econ.* **51**, 505–516. (doi:10.1016/j.insmatheco.2012.07.009)
- Kaplan EL, Meier P. 1958 Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481. (doi:10.1080/01621459.1958.10501452)
- Tsai WY, Jewell NP, Wang MC. 1987 A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **74**, 883–886. (doi:10.1093/biomet/74.4.883)
- Dempster AP, Laird NM, Rubin DB. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **39**, 1–38. (doi:10.1111/j.2517-6161.1977.tb01600.x)
- van der Laan MJ. 1994 Modified EM-estimator of the bivariate survival function. *Math. Methods Stat.* **3**, 213–243.
- Nandi S, Dewan I. 2010 An EM algorithm for estimating the parameters of bivariate Weibull distribution under random censoring. *Comput. Stat. Data Anal.* **54**, 1559–1569. (doi:10.1016/j.csda.2010.01.004)
- Arias LAS, Cirillo P. 2021 Joint and survivor annuity valuation with a bivariate reinforced urn process. *Insur. Math. Econ.* **99**, 174–189. (doi:10.1016/j.insmatheco.2021.04.004)
- Frees EW, Carriere J, Valdez E. 1996 Annuity valuation with dependent mortality. *J. Risk Insur.* **63**, 229–261. (doi:10.2307/253744)
- Luciano E, Spreeuw J, Vigna E. 2008 Modelling stochastic mortality for dependent lives. *Insur. Math. Econ.* **43**, 234–244. (doi:10.1016/j.insmatheco.2008.06.005)

23. Wang MC. 1991 Nonparametric estimation from cross-sectional survival data. *J. Am. Stat. Assoc.* **86**, 130–143. (doi:10.1080/01621459.1991.10475011)
24. Johnson NL, Kotz S. 1977 *Urn models and their application*. New York-London-Sidney: John Wiley & Sons.
25. Mahmoud HM. 2008 *Pólya urn models*. New York, NY: Chapman & Hall/CRC.
26. Feng Y, Chen X, Jia L, Song X, Mahmoud HM. 2017 Estimating the Pólya process. *Commun. Stat. Theory Methods* **46**, 9397–9406. (doi:10.1080/03610926.2016.1208242)
27. Line CLG, Philippe S. 2017 Parameter estimation of a two-colored urn model class. *Int. J. Biostat.* **13**, 20160029.
28. Marcaccioli R, Liva G. 2019 A pólya urn approach to information filtering in complex networks. *Nat. Commun.* **10**, 745. (doi:10.1038/s41467-019-08667-3)
29. Cheng D, Cirillo P. 2018 A reinforced urn process modeling of recovery rates and recovery times. *J. Bank. Finance* **96**, 1–17. (doi:10.1016/j.jbankfin.2018.08.014)
30. Wang MC. 1987 Product limit estimates: a generalized maximum likelihood study. *Commun. Stat. Theory Methods* **16**, 3117–3132. (doi:10.1080/03610928708829561)
31. Cox DR, Oakes D. 1984 *Analysis of survival data*. New York, NY: Chapman & Hall.
32. Lynden-Bell D. 1971 A method of allowing for known observational selection in small samples applied to 3CR quasars. *Mon. Not. R. Astron. Soc.* **155**, 95–118. (doi:10.1093/mnras/155.1.95)
33. Muliere P, Secchi P, Walker SG. 2000 Urn schemes and reinforced random walks. *Stoch. Process. Appl.* **88**, 59–78. (doi:10.1016/S0304-4149(99)00119-2)
34. Cirillo P, Hüslér J, Muliere P. 2010 A nonparametric urn-based approach to interacting failing systems with an application to credit risk modeling. *Int. J. Theor. Appl. Finance* **41**, 1–18. (doi:10.1142/S0219024910006170)
35. Peluso S, Mira A, Muliere P. 2015 Reinforced urn processes for credit risk models. *J. Econom.* **184**, 1–12. (doi:10.1016/j.jeconom.2014.08.003)
36. Fortini S, Petrone S. 2012 Hierarchical reinforced urn processes. *Stat. Probab. Lett.* **82**, 1521–1529. (doi:10.1016/j.spl.2012.04.012)
37. Connor RJ, Mosimann JE. 1969 Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Stat. Assoc.* **64**, 194–206. (doi:10.1080/01621459.1969.10500963)
38. Turnbull BW. 1976 The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Stat. Soc. Ser. B (Methodological)* **38**, 290–295. (doi:10.1111/j.2517-6161.1976.tb01597.x)
39. Embrechts P, Klüppelberg C, Mikosch T. 2003 *Modelling extremal events for insurance and finance*, vol. 33, 2nd edn. Heidelberg, Germany: Springer Science & Business Media.
40. Shackle GLS. 1955 *Uncertainty in economics and other reflections*. Cambridge, UK: Cambridge University Press.
41. Taleb NN. 2007 *The black swan: the impact of the highly improbable*. New York, NY: Random House.
42. James G, Witten D, Hastie T, Tibshirani R. 2013 *An introduction to statistical learning with application in R*. New York, NY: Springer.
43. O'Hagan A, Murphy TB, Gormley IC. 2012 Computational aspects of fitting mixture models via the expectation–maximization algorithm. *Comput. Stat. Data Anal.* **56**, 3843–3864. (doi:10.1016/j.csda.2012.05.011)
44. Nelsen RB. 2006 *An introduction to copulas*. New York, NY: Springer.