Journal of Economics, Finance and Accounting Studies (JEFAS)
ISSN: 2709-0809
Website: https://al-kindipublisher.com/index.php/jefas

JEFAS

The Bayesian Kaplan Meier Model Under the Classical Nonparametric Bootstrap

Hamimes Ahmed[1]* & Benamirouche Rachid [2]

[1]*Phd student, National Higher School of statistics and applied economics, Tipaza,Algeria, Assistant Professor at the Faculty of Medicine, University of Constantine3, Algeria*

[2]*Professor, National Higher School of statistics and applied economics, Tipaza, Algeria*

Corresponding Author: Hamimes Ahmed, E-mail: ahmedhamimes@yahoo.com

ARTICLE INFO

Received: October 15, 2020
Accepted: December 01, 2020
Volume: 2
Issue: 2

KEYWORDS

Bayesian approach,
bootstrap,
unemployment durations,
the local employment agency of Ain El Benian.

ABSTRACT

This study aims to introduce and encourage the other to use bootstrap methods in statistical survival analysis. We show how to bootstrap the Kaplan-Meier Bayesian estimator and pay attention to its advantage unlike the classical Bayesian analysis of the Kaplan Meier method. In our study, we essentially try to focus on the application of Kaplan Meier Bayesian models in the estimation of unemployment durations of those registered with the local employment agency of Ain El Benian, with the aim of to determine the role of bootstrap in improving the quality of estimation. We find that the choice between the period models of the same family is likely to produce a multitude of decisions from the results of the application. A Bayesian survival method based on the Kaplan Meier model with the classical nonparametric bootstrap provides realistic solutions for a number of individuals of different nature, simple and relatively easy to exploit numerically global durations.

1. Introduction

The Bayesian concept differs from the classical concept, the meaning of which is a random variable whose behavior is assumed to be known, by associating it with a probability distribution on the space Θ called a priori distribution and noted $\pi\ (\theta)$, and with Through this design, the statistical analysis makes it possible to consider all the qualitative and quantitative information on the uncertainty in the model. Then, if we use Bayes' theorem which allows to reverse the probabilities, we can deduce the a posteriori distribution $\pi\ (\theta\ /\ x)$ which allows us to build inferential procedures in the most natural way possible, which also explains the persistence of this paradigm, against all odds for 250 years.

One of the most frequently used nonparametric methods for estimating survival function is the Kaplan-Meier method. In science, we very often have to deal with small samples. There are several reasons for this. In medical science, for example, the most common is the rarity of the disease or the difficulty of bringing together patients with the same biochemical parameters. In addition, we very often have censored data. Several works have been based on the improvement of this estimator. Khizanov and Maĭboroda (2015), proposed a modification based on a mixing model with various concentrations. Kaciroti, et al. (2012) presented Kaplan Meier's survival model with informative censorship and in a Bayesian framework. It may happen that Kaplan-Meier gives the same probabilities of survival for two groups with the same number of events and censored observations, although the duration between consecutive events (i.e. wait times) can vary considerably, Zaman et al (2011) addressed this problem. Zieliński (1999) used local smoothing of the Kaplan-Meier estimator based on an approximation by the Weibull distribution function. Zieliński (2002), introduces a Kaplan-Meier estimator based on an approximation by the Weibull distribution, Zieliński also studied a smoothing of KME such that the resulting estimator is a strictly decreasing function of time, the smoothed KME seems be more specific than the original. Shafiq Mohammad et al (2007), presented a weighting of the Kaplan Meier estimator under the sine function for heavy censorship data.

When a small sample size does not allow us to use classical statistical methods or when they are used, they can give us results that are too general and even false. With computer simulation, we can generate many samples based on the original sample data and we can more accurately evaluate the parameters determined on the bootstrap. In statistics, bootstrap techniques are statistical inference methods based on the multiple replication of data from the data set studied using resampling techniques, Hastie et al (2008) defined the word "bootstrap" as a reuse effective sample. They date from the late 1970s, when the possibility of intensive computer calculations became affordable.

This contribution, we will give a Bayesian alternative of the classical Kaplan Meier estimator based on the classical nonparametric bootstrap method. In our application we will analyze the durations of global unemployment in the National Employment Agency (ANEM) of Ain El Benian. We are working on a sample of 1064 unemployed individuals observed between 01/01/2011 and 15/07/2013. This application allows to practically demonstrate that Bayesian procedures constitute an essential element for the improvement of the quality of inference because of the difference which exists in the interpretation and the estimation of curves and durations of exiting unemployment.

2. The bootstrap method

Theorem (Glivenko-Cantelli theorem)

$$\underset{x\in IR}{\operatorname{Sup}}|F_n(x)-F(x)|\xrightarrow{n\to\infty}0\,,p.s.$$

Definition we call a plug-in estimator of a parameter of F, the estimator obtained by replacing the function F by the empirical distribution:

$$\hat{\theta}=t(\hat{F})$$

This principle of substitution (or plug-in) is asymptotically justified since we have almost certainly, according to the Glivenko-Cantelli theorem.

The idea of bootstrap is due to B. Efron is based on the following elementary principle:

If n is large F_n nest close to F, we will therefore have a good approximation of the distribution of T by using F_n instead of F.

The following figure gives an illustrative explanation of the bootstrap principle.

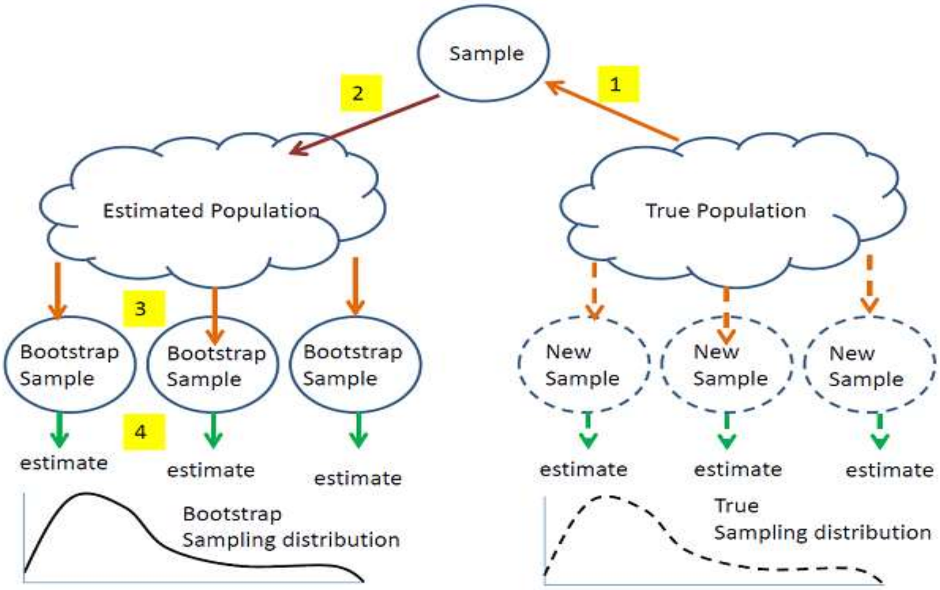

Figure 1 : the bootstrap procedure.

In figure (1) observed quantities are denoted by solid curves and unobserved quantities by dotted curves. The goal is to estimate the true distribution such that the true sampling distribution is calculated by taking new samples from the real population, calculating T and then accumulating all the values of T in the sampling distribution. This is the goal denoted by (1), but the problem we find in the first step is related to the new samples is expensive, so instead we take a single sample and we use it to estimate the population in step (3). We (3) then take samples (on the computer) of the estimated population, calculate T from each (4), and accumulate all of the values of T into an estimate of the sampling distribution. From this estimated sampling distribution, we can estimate the desired characteristics of the sampling distribution.

We are therefore led to draw samples of n values in the F_n distribution, which amounts to resampling in the sample $x_1, x_2, \dots, x_n$in other words to carrying out draws with replacement of n values among the n values observed: the values observed $x_1, x_2, \dots, x_n$ are therefore repeated according to the realizations of a multinomia1 vector $K_1, K_2, \dots, K_n$ of effective n and of probabilities p_i egal to$1/n$.

We study the properties of T_n via $T_n^* = T(X_1^*, X_2^*, \dots, X_n^*)$ where the X_n^* are i.i.d variables with distribution function F_n.
Is $X^* = (X_1^*, X_2^*, \dots, X_n^*)$ bootstrap sample. We note indifferently:

$$P(X_j^* = X_i^*/X) = 1/n, \qquad 1 \leq i, j \leq n$$

or

$$P(X_j^* = X_i^*/F_n) = 1/n, \qquad 1 \leq i, j \leq n$$

because as soon as we know, we can deduce F_n and vice versa.
The expectation of T_n:

$$E(T_n) = \int \dots \int T(x_1, x_2, \dots, x_n) dF(x_1) \dots dF(x_n)$$

And thus estimated by:

$$E(T_n^*) = \int \dots \int T(x_1, x_2, \dots, x_n) dF(x_1) \dots dF(x_n)$$

It is also written:

$$E(T_n^*) = \frac{1}{n^n} \int \dots \int T(x_1, x_2, \dots, x_n) \sum_{l_1=1}^{n} \delta_{X_{l_1}}(x_1) \dots \sum_{l_n=1}^{n} \delta_{X_{l_n}}(x_n) dx_1 \dots dx_n$$

$$\frac{1}{n^n} \sum_{l_1=1}^{n} \dots \sum_{l_n=1}^{n} T(X_{l_1}, \dots, X_{l_n})$$

For example, we estimate $E(T_n)$ by the empirical mean over all the prints:

$$E_B(T_n^*) = \frac{1}{B} \sum_{k=1}^{B} T_n^{*(k)},$$

where B is the number of draws with discount made and $T_n^{*(k)}$ n the statistic obtained in the k-th draw. Similarly, the distribution function G of T_n est estimated by:

$$G^*{}_B(x) = \frac{1}{B} \sum_{k=1}^{B} 1_{T_n^{*(k)} \leq x}$$

In the nonparametric bootstrap. The new data is simulated by resampling from the original data (with replacement), and the parameters are calculated either directly from the empirical distribution or by applying a model to these surrogate data (see Figure 2).

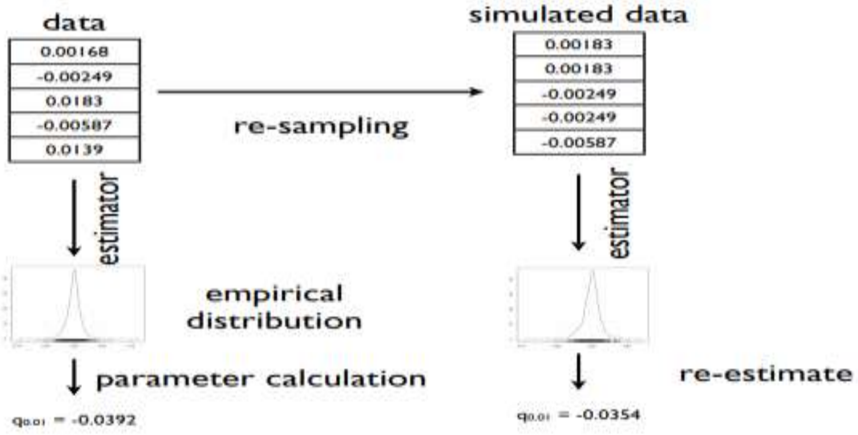

Figure 2 : Schematic of the nonparametric bootstrap.

3. Kaplan Meier's method and Bayesian statistics

3.1. Kaplan Meier estimator for the survival function

The method of Kaplan and Meier, gives us a non-parametric estimator of the survival function. Which, moreover, does not pose any preliminary hypothesis on the algebraic form of the survival function.

If we observed k distinct survival times. Arranged in ascending order, which are $t_1, t_2, \ldots t_n$. In time t_i there are n_i individuals who are at risk of dying, and who survived after this period and are not censored. Noting d_i the number of deaths on the date t_i. To simplify the notations, we put $t_0 = 0$ and $d_0 = 0$. Then, the Kaplan and Meier estimator for the survival function $S\,(t)$ is given as follows:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \qquad (1)$$

3.2. Derivation of the Kaplan and Meier estimator

The implicit idea of Kaplan and Meier's method is as follows: We have the following recursive equation:

$$\begin{aligned} P(X > t) &= P(X > t_{i-1}, X \geq t_i) \\ &= P(X > t_i / X \geq t_{i-1}) P(X > t_{i-1}) \\ &= P(X > t_i / X \geq t_{i-1}) P(X > t_{i-1} / X \geq t_{i-2}) P(X > t_{i-2}) \ldots, \end{aligned}$$

So the probability of surviving until time t_i is the probability of surviving t_i knowing that we were alive at t_{i-1}, that is,

$$S(t_i) = P(X > t_i / X \geq t_i) * S(t_{i-1})$$

and when the risk of death estimated by

$$\hat{q}_i = \frac{d_i}{n_i} \ ,$$

this equation is given by:

$$S(t_i) = P(X > t_i / X \geq t_i) * S(t_{i-1})$$
$$S(t_i) = \frac{n_i - d_i}{n_i} * S(t_{i-1})$$
$$S(t_i) = (1 - \hat{q}_i) * S(t_{i-1})$$

With $t_0 = 0$ and $d_0 = 0$

We know that the variable is a continuous and positive variable T. Therefore, theoretically and under the condition of continuity, it is rarely - if not impossible - to find observations for two identical individuals i and j. But in practice, we work on data which is measured, generally, over a well-specified period of time (days, weeks, months), therefore the event: $T_i = T_j$ is possible.

However, we model the survival time as a discrete random variable taking as values $t_1 < t_2, \ldots < t_n$. We denote by p_ the conditional probability that an individual survives until time t_i, and that he has lived time t_{i-1}:

$$p_i = P(X > t_i / X \geq t_i); i = 1, \ldots k$$

From equation (1)

$$S(t_i) = \prod_{j=1}^{i} p_j \qquad (2)$$

The maximum likelihood method is used to estimate the values of p_i.

In the date t_i there are d_i individuals for each of them, the probability of dying is 1-p_i, $i = 1,2, \ldots, k$. With the hypothesis of Independence between the survival times of individuals-. On the other hand, there are $n_i - d$ individuals who have - independently for each one - a probability of p_i.

Therefore, the likelihood function is

$$\begin{aligned} L(p_1, p_2, \ldots, p_n) &= \prod_{i=1}^{k} {q_i}^{d_i} (1 - q_i)^{n_i - d_i}, \\ &= \prod_{i=1}^{k} (1 - p_i)^{d_i} {p_i}^{n_i - d_i} \end{aligned}$$

By equivalence the log-likelihood is

$$lnL = \sum_{i=1}^{m} [d_i ln\,(1 - p_i) + (n_i - d_i) ln(p_i)],$$

Put the first derivatives of the log-likelihood with respect to p_i, $i = 1,2, \dots, k$ as linear equations equal to zero, we find

$$\frac{d_i}{1 - p_i} = \frac{n_i - d_i}{p_i}, i = 1,2, \dots, n$$

These equations will be solved, which gives us as a solution, the maximum likelihood estimators

$$p_i = 1 - \frac{d_i}{n_i}$$

Replace these estimators in equation (2)

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \qquad (3)$$

Now, if we fix a date t, and assuming that $t_i \leq t \leq t_{i+1}$ for some $i = 1,2, \dots, k$. Because, there is no death between the date t_i and t_{i+1}, the survival function s (t), is estimated by $\hat{S}(t) = \hat{S}(t_i)$. which equation (1) gives us.

3.3. Context of using the Kaplan and Meier method

This estimator makes it possible to approach the empirical form taken by the risk of leaving the state, without adopting any specification of distribution (see Planchet and Thérond, 2006).

• This estimator requires knowing exactly all the study entry and exit dates of all individuals; which is a drawback if the registers are incorrectly entered.

• this method is relevant if each time interval considered is small relative to the speed of the variation of the survival function. This is to ensure that the discretization does not generate a significant bias on the estimate.

The Kaplan and Meier estimator cannot take into account the effect of individual characteristics by breaking down the study population into subpopulations (e.g. sex, PSC, ...)[2].

3.4. The Bayesian conception of the Kaplan and Meier method

We assume that the number of deaths in the interval of time is a Binomial distribution given by

$$d_i \sim \beta in(n_i, q_i)$$

when the outputs in the intervals $[t_i, t_{i+1}[$ [being independent of each other, we write

$$f(d/q_i) = \prod_{i=1}^{m} C_{n_i}^{d_i} {q_i}^{d_i} (1 - q_i)^{n_i - d_i},$$

using the maximum likelihood method, the risk of death estimated by

$$\hat{q}_i = \frac{d_i}{n_i}\ ,$$

From a Bayesian perspective, we assume an a priori for q_ (i,), and when the distribution used in the case of proportions is that of Beta, we set:

$$q_{i \sim} beta(\alpha, \beta)$$

3.4.1. Jeffreys' distribution

In the non-informative case the most used method is that of Jeffreys, the a priori measurement of Jeffreys (possibly improper) defined by:

$$\pi^*(q_i) \propto\ I^{1/2}(q_i) \Rightarrow \pi^*(q_i) \propto n^{1/2} \big(q_i(1 - q_i)\big)^{-1/2}$$

If we set $\pi(q_i) = A * \pi^*(q_i)$, and we integrate the two parts of the equation with respect to $q_{i,}$, we find:

$$1 = \int_0^1 A * n^{1/2} \big(q_i(1 - q_i)\big)^{-1/2} dq_i = A * n^{1/2} \int_0^1 \big(q_i(1 - q_i)\big)^{-1/2} dq_i$$
$$= A \times n^{1/2} \times \beta(1/2\,; 1/2)$$

[2] Les méthodes appropriées pour ça sont les modèles proportionnels.

So we have :

$$\mathrm{A} = 1/\left(n^{1/2} \times \beta(1/2\,;1/2)\right)$$

From A, the a priori function is written as follows:

$$\pi(q_i) = A * \pi^*(q_i) = \frac{1}{\sqrt{n}\beta(1/2\,;1/2)} n^{1/2}\big(q_i(1-q_i)\big)^{1/2-1}$$

$$= \frac{1}{\beta(1/2\,;1/2)} q_i^{-1/2}(1-q_i)^{-1/2}; 0 < q_i < 1$$

so

$$q_i \qquad \sim \mathcal{B}e\left(\frac{1}{2},\frac{1}{2}\right) \qquad (4)$$

3.4.2. Uniform distribution

In the case of a uniform measure with respect to the Lebesgue measure, the a priori distribution is not invariant by reparametrization, we generally set

$$q_i \sim \mathcal{B}e(1,1),$$

This distribution poses a crucial problem because the support is vague with a uniform probability, but in the majority of duration models the risk function takes low values.

3.4.3. The vague a priori distribution

A solution for the previous distribution is vague prior distribution, it is a proper distribution with a very large variance, according to this distribution, the prior distribution is considered to be weak informative, it provides solutions in the use of algorithms. We ask:

$$q_i \sim \beta(0{,}01{,}0{,}01) \qquad (5)$$

3.4.4. La loi a priori hiérarchique

One of the methods to introduce the relation between the probability q_i_and the independent and unknown causal series is the logistic transformation given by:

$$logit\ (q_i) \overset{\text{def}}{=} ln\frac{q_i}{1-q_i}, \qquad q_i \in]0{,}1[$$

and

$$\mu_i = logit\ (q_i), \text{ i.e.}$$

$$q_i = \frac{exp(\mu_i)}{1+exp(\mu_i)},$$

Our problem remains nonparametric, we pose a Gaussian model for μ_i with unknown hyperparameters as follows:

$$\mu_i \sim \mathcal{N}(\vartheta;\tau)$$

$$\vartheta \sim \mathcal{N}(0;\,0{,}001)$$

In our case the hyperparameter of the variance τ is only made up of positive values, so we set the demi-Cauchy distribution. τ follows a half-Cauchy distribution if that density is:

$$\pi(\tau) = 2/\pi\rho\,[1 + (\tau/\rho)^{-1}], \qquad \tau > 0$$

we notice

$$\tau \sim HC(\rho)$$

ρ is the median of demi-Cauchy, which means that a priori beliefs are easily expressed.

The proposed model is written as follows

$$q_i = \frac{exp(\mu_i)}{1+exp(\mu_i)}$$

$$\mu_i \sim \mathcal{N}(\vartheta;\tau)$$

$$\vartheta \sim \mathcal{N}(0;\,0{,}001),\, \tau \sim HC(B)$$

$$B \sim Uniforme(0;\mathcal{T}), \text{on pose} : \mathcal{T} = 100$$

3.4.5. The parametric empirical prior distribution

Another way to solve the problem of uncertainty in the a priori measure of beta is the marginal distribution $f_\pi(d_i/\alpha,\beta)$, this method is called the empirical Bayesian approach. Morris (1983) introduces the parametric concept of this approach. For a binomial distribution and a conjugate prior distribution we set

$$f_{\pi}(d_i/\alpha,\beta)=\int_0^1 f(d_i/q_i)\,\pi(q_i/\alpha,\beta)dq_i=C_{n_i}^{d_i}\frac{B(\alpha+d_i,n_i+\beta-d_i)}{B(\alpha,\beta)}$$

This provides a beta – binomial distribution for estimating $\hat{\alpha},\hat{\beta}$ in order to calculate $\pi\left(q_i/d_i,\hat{\alpha},\hat{\beta}\right)$. The binomial beta parameters can be estimated, the estimator based on the first two moments is written by

$$\hat{\alpha}=\frac{\hat{\xi}_0(m-1-\hat{\xi}_1)}{\hat{\xi}_0+m(\hat{\xi}_1-\hat{\xi}_0)},\ \ \hat{\beta}=\frac{\left(m-\hat{\xi}_0\right)(m-1-\hat{\xi}_1)}{\hat{\xi}_0+m(\hat{\xi}_1-\hat{\xi}_0)}$$

or

$$\hat{\xi}_j=\frac{\hat{\mu}_{(j+1)}}{\hat{\mu}_{(j)}},\qquad j=1{,}2$$

and

$\hat{\mu}_{(j)}$ is the generalized moment of order j, calculated by

$$\mu_{(j)}=\frac{(-m)_j(\alpha)_j}{(\alpha+\beta)_j}(-1)^j,j=1{,}2,\ldots$$

3.4.6. la loi a priori de mélange

The prior distribution of mixture is written as follows

$$L(q_i/w,\alpha,\beta)=\sum_{j=1}^{k}w_jf(q_i/d_i,\alpha_j,\beta_j)$$

and $0\le w_j\le 1,\sum_{j=1}^{k}w_j=\ 1$

we consider that the probabilities $q=(q_1,\ldots q_m)$ are i.i.d in the likelihood is

$$L(q,w/d,\alpha,\beta)=\prod_{i=1}^{m}\sum_{j=1}^{k}w_jf(q_i/d_i,\alpha_j,\beta_j)$$

$$\propto\prod_{i=1}^{m}\sum_{j=1}^{k}w_j{q_i}^{d_i}(1-q_i)^{n_i-d_i}$$

if we set the vector of independent parameters $\theta=\left(\alpha_j,\beta_j\right)$, in Bayesian analysis where the inference is based on the posterior distribution given by

$$\pi(\theta,w/q)\propto L(q,w/d,\alpha,\beta)\ \pi(\alpha_j/\xi_j,\nu_j)\pi(\beta_j/\xi_j^*,\nu_j^*)\pi(w/\varphi_1,\ldots,\varphi_k)$$

$$\propto\prod_{i=1}^{m}\sum_{j=1}^{k}w_j{q_i}^{d_i}(1-q_i)^{n_i-d_i}\pi(\alpha_j/\xi_j,\nu_j)\pi(\beta_j/\xi_j^*,\nu_j^*)\pi(w/\varphi_1,\ldots,\varphi_k)$$

3.5 The Bayesian estimator of the Kaplan and Meier method based on the nonparametric bootstrap method

For the bootstrap estimator of the survival function, we take the bootstrap replica of a sample statistic that can be expressed for $S(t_i)=\prod_{j=1}^{i}p_j$ is the weighted multiplication:

$$\hat{S}_B(t_i)=\prod_{j=1}^{i}\hat{p}_j=\prod_{t_j\le t}\left(1-\hat{q}_j\right)$$

such as

$$\hat{q}_j=\hat{q}_{j(bootstrap)}$$

choose a random number between 1 and n through the categorial density:

$$j(bootstrap)\sim cat(\theta_i);\ \sum_{i=1}^{n}\theta_i=1;\ \theta_1=\theta_2=\cdots=\theta_n=1/n$$

therefore:

$$\hat{S}_B(t_i)=\prod_{t_j\le t}\left(1-\hat{q}_{j(bootstrap)}\right)\qquad(6)$$

4. Application

4.1. Unemployment durations and the model used

One of the most effective tools for analyzing lifespans is certainly the survival or stay function estimator. For the case of this study, this function relates to the duration of unemployment. Typically, the most widely used estimator for this estimate is the

Kaplan-Meier estimator, which allows for right-censored data. This estimator calculates the probability of knowing the event in each time interval, and we thus obtain a curve which is interpreted simply as the proportion of "survivors" for each length of stay in a given state. In other words, the proportions of individuals leaving unemployment for each period of unemployment.

In this application we will analyze the durations of global unemployment in the local employment agency of Ain El Benian. We are working on a sample of 1064 unemployed individuals observed between 01/01/2011 and 07/15/2013. This application allows to practically demonstrate that the Bayesian beta mixing procedures are an efficient element to solve label-swetching problem and to find estimates of good quality and good precision. By distinguishing those who found a job, the placement of the unemployed during this period gives rise to 875 right-censored observations. In this case, the variable i represents the indication that the ith unemployed person found a job after his daily period of unemployment t_i.

4.2. Analysis of unemployment durations

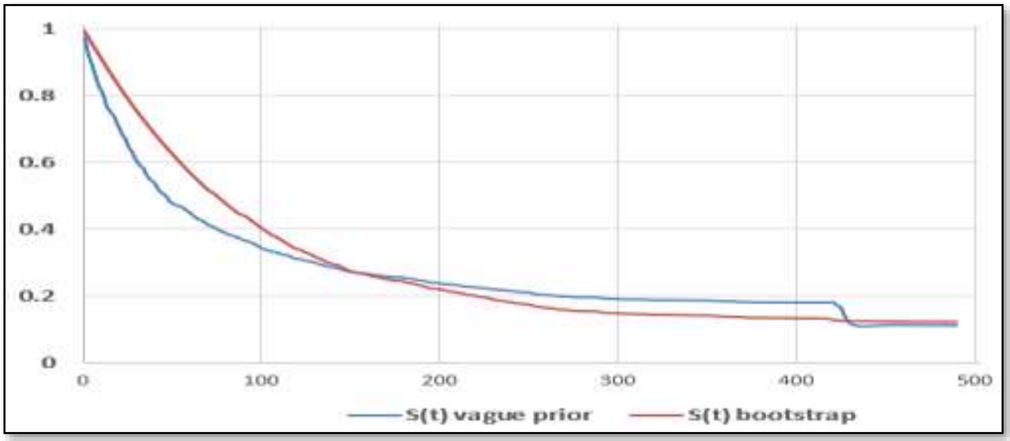

Figure 3 : Bayesian survival functions according to the classical nonparametric bootstrap method and the Bayesian method with a vague prior distribution Beta (0.01; 0.01).

According to this figure (3) a difference between the Bayesian Kaplan Meier estimator based on the classical nonparametric bootstrap method and the Bayesian method with a vague prior distribution Beta (0.0.1; 0.01) .. this difference is clear in terms of median survival, 50% of individuals were placed in the labor market after 74 days in the modified model i.e. the Bayesian Kaplan Meier estimator based on the classical 50-day non-parametric bootstrap method for the Bayesian model with a vague Beta prior distribution (0.0.1; 0.01). But, the exit from unemployment for the rest of the individuals in the sample is spread over a long period in both models, for some it even exceeds one year. We also note that a convergence between the proposed estimation models occurs gradually after the 400-day duration of leaving unemployment.

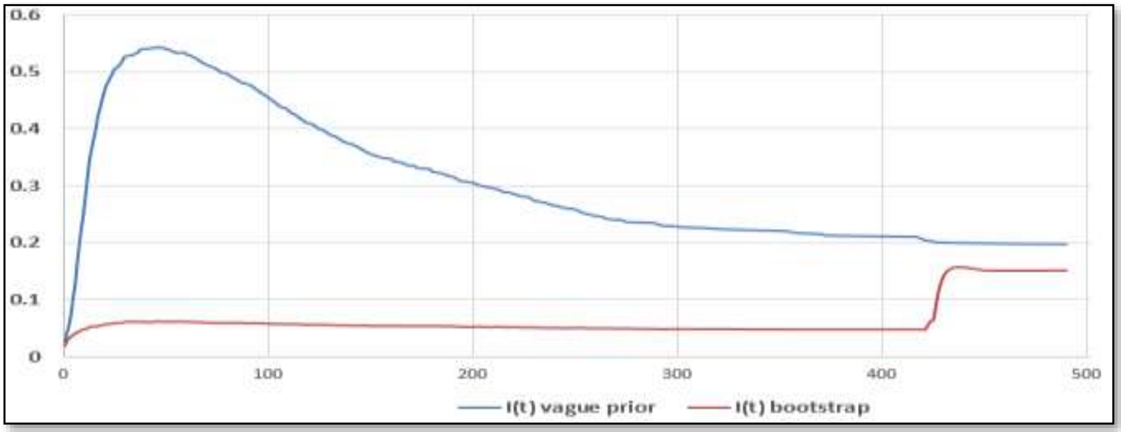

Figure 4: The credibility intervals according to the classical nonparametric bootstrap method and the Bayesian method with a vague prior distribution Beta (0.01; 0.01).

Figure (4) shows the difference between the amplitude of the credibility intervals of the two methods, it is clear that the Bayesian Kaplan Meier estimator based on the classical nonparametric bootstrap method offers practical, simple and relatively easy to use solutions. numerically exploit overall survival times in comparison with the model of the vague prior distribution Beta (0.0.1; 0.01).

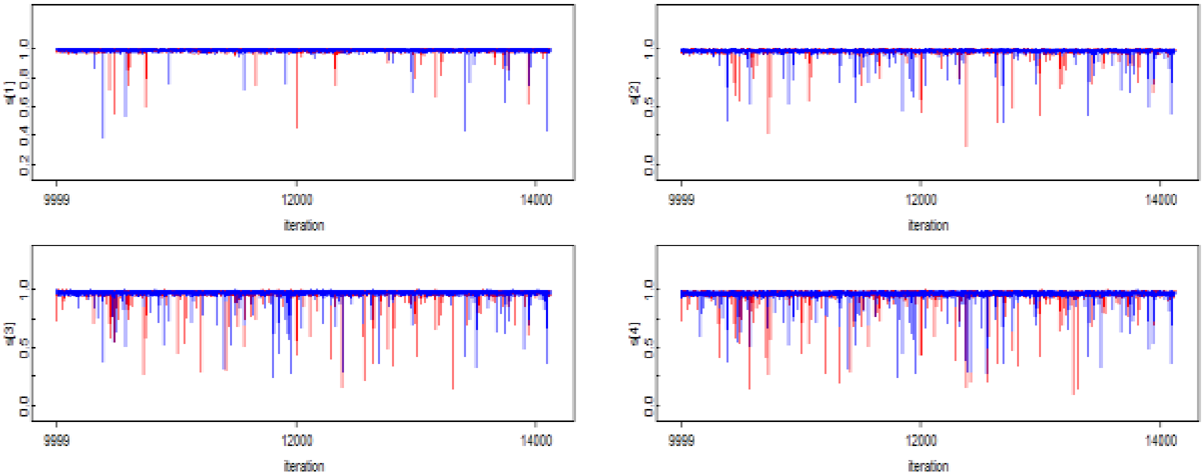

Figure 5: The trace of the a posteriori distribution for certain estimators of the survival function of in the proposed model.

In figure (5), we can see graphically a certain stationarity of the a posteriori values throughout the 20,000 iterations for the Bayesian Kaplan meier model with a classical nonparametric bootstrap. Also the two chains mix well: convergence is reached. Brooks and Gelman in 1998 proposed a generalization of the method of Gelman and Rubin which was introduced in the year 1992, it is a method of validation of the ergodic sequences of MCMC algorithms.

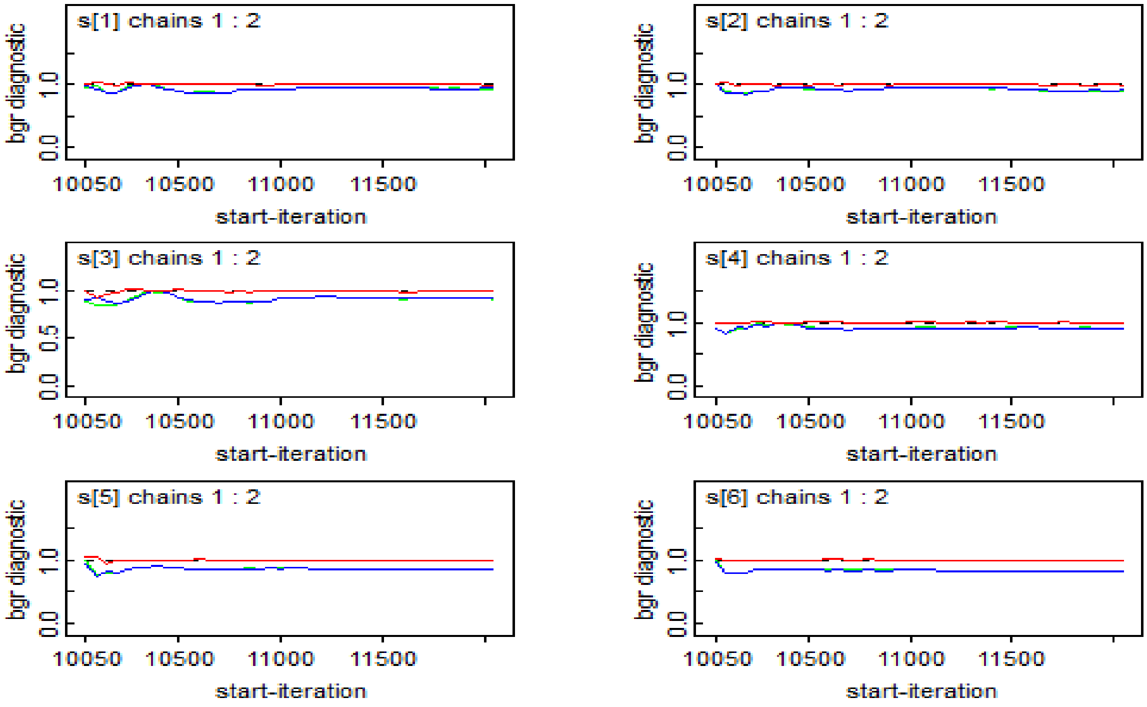

Figure 6 : The Brooks and Gelman graph "convergence - diagnosis - graph" after 20,000 iterations in the proposed beta mixing model.

In figure (6) the green curve shows the width of the inter-chain credibility interval at 80%. The blue curve displays the average width of the within-chain 80% credibility intervals. The red curve indicates the Brooks and Gelman statistic (i.e., the ratio of the green / blue curves). The Brooks and Gelman statistic tend towards 1, which means that there is convergence.

5. Conclusion

The objective sought in our contribution is to improve the inferential phase in the estimation of nonparametric survival times and in the presence of censorship. In the results of this work, we find:

- A Bayesian approach to survival based on the Kaplan Meier model with the classical nonparametric bootstrap offers practical solutions, simple and relatively easy to exploit numerically global durations for a set of individuals of different nature.
- The use of the MCMC method is relatively easy to implement, it provides a set of techniques very suitable for the estimation of complex models with several parameters or with a hierarchical structure.
- From the results of the application we find that the choice between the duration models of the same family is likely to generate a multitude of decisions.

The limit of this method is the use of Bayesian bootstrappe and the latter represents the future work.

References

[1] Frédéric Planchet et Pierre Thérond. (2006). Modèles de durée, Applications actuarielles. P250.

[2] Hastie,T. , Tibshirani, R. and Fried-man, J. (2008) The Elements of Sta-tistical Learning, second edition, Springer, Stanford, California.

[3] Kaciroti, N. A., Raghunathan, T. E., Taylor, J. M., & Julius, S. (2012). A Bayesian model for time-to-event data with informative censoring. *Biostatistics, 13*(2), 341-354.

[4] Khizanov, V. G., Maĭboroda, R. (2015). A modified Kaplan-Meier estimator for a model of mixtures with varying concentrations. *Theor. Probability and Math. Statist. 92* (2016), 109-116.

[5] Robbins, H. (1951). Asymptotically subminimax solutions to compound statistical decision problems. In Proc. Second Berkeley Symp. Math. Statist. Probab., volume 1. University of California Press.

[6] Robbins, H. (1955). An empirical Bayes approach to statistics. InProc. Third Berkeley Symp. Math. Statist. Probab. , volume 1. University of California Press.

[7] Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. Ann. Mathemat. Statist.,35, 1–20.

[8] Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Sta-tist, 11*, 713–723.

[9] Robert, C.P. (2006). Le choix Bayésien : principes et pratiques. Springer.

[10] Rossa, A and Zieliński, R. (2006). A simple improvement of the kaplan-meier estimator. *Communications in statistics - theory and methods, 31*(1), 147-158, doi: 10.1081/sta-120002440.

[11] Rossa, A., & Zieliński, R. (1999). Locally Weibull-smoothed Kaplan-Meier Estimator. Inst. of Mathematics.

[12] Shafiq, M., Shah, S., & Alamgir, M. (2007). Modified Weighted Kaplan Meier Estimator. *Pakistan Journal of Statistics and Operation Research, 3*(1), 39-44.

[13] Zaman, Q., Strasak, A. M., & Pfeiffer, K. P. (2011). Exact Waiting Time Survival Function. *J Biomet Biostat, 2*(117), 2.

Appendices (the code on the OpenBUGS program)

```
model
{
for (i in 1:m) {
d[i]~dbin(q[i],n[i])
q[i]~dbeta(0.01,0.01)}
for (i in 1:m) {
pr[i] <- 1/m
}
for (i in 1:m){
pick[i] ~ dcat(pr[])
qboot[i] <- q[pick[i]]
}
for (i in 1:m){
ce[i]~dbin(0.01,0.01)
}
for (i in 1:m){
qc[i]~dbeta(0.01,0.01)
}
for (i in 1:m){
```

```
p[i]<-1-qboot[i]
}
n[1]<- 1064
for(i in 2:m){
n[i]<-n[i-1]-d[i-1]-ce[i-1]
}
for (i in 2:m){
s[i]<-s[i-1]*p[i]
}
s[1]<-p[1]
}
list(m=227,
ce=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,36,39,29,42,30,6,5,0,1,1),
d=c(28,31,23,18,14,20,18,18,16,8,14,15,20,12,7,9,5,11,15,12,11,13,7,10,16,11,6,13,11,9,9,7,5,4,14,9,9,4,5,6,7,12,7,5,3,5,5,11
,5,5,3,1,3,1,2,1,8,2,5,4,5,4,4,5,3,1,3,3,6,3,3,4,3,4,1,4,4,1,4,2,1,3,2,1,1,5,4,4,2,2,1,2,3,4,3,4,2,3,2,2,1,3,1,1,2,4,1,1,1,1,1,3,3,2,3,
1,2,2,1,2,2,1,1,2,4,2,1,2,1,3,1,2,3,1,2,2,1,1,1,2,2,1,2,1,1,3,1,1,2,1,1,2,1,2,2,1,2,1,2,1,2,1,3,1,1,1,2,1,2,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,2,1,1,1,2,2,1,1,1,1,1,1,1,1,1,1,1,1,2,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,1,0,1,0,0))
list(q=c(0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.
5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,
0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.
5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,
0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.
5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,
0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5))
list(q=c( 0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,
0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9,0.9))
```