Test (2023) 32: 506-509 https://doi.org/10.1007/s11749-023-00860-3

Comments on: Nonparametric estimation in mixture cure models with covariates

Philippe Lambert[1,2]

Received: date / Accepted: date

First, I would like to thank the authors for this stimulating contribution (López-Cheda et al., 2023) to the nonparametric literature on cure survival models. The methodology is clearly laid out and naturally combines Beran's estimator with the EM algorithm to handle the unknown cure status of units for which no event has been reported.

I was puzzled by the results of the simulation in several ways. First of all, I was a bit surprised by the design of the study and the very small values chosen for the first two sample sizes ($n = 50$ and $n = 100$). Given the large proportion of cured subjects and the additional right-censoring process also acting on the reported times for the non-cured subjects, the amount of information remaining for the nonparametric estimation of $S_u(t|x)$ becomes very sparse. A parametric model using historical information for its specification would probably make much more sense in comparable practical settings, although I do not expect compelling scientific results from such modest studies.

I would like to return to the results of the simulation study for Setting 2 where the time-to-event distribution for the non-cured units is a mixture of two Weibull distributions, only one of which depends on covariate x, see Fig. 1. From its analytical form, one can see that $S_u(t|x)$ only depends on x through its square, meaning that $S_u(t|x) = S_u(t|-x)$ when $x \in (-10, 10)$. Therefore, given that z and x are generated independently, I would have expected to see rather close values for MISE(x) for opposite values of x and, thus, some symmetry in the reported graph (see the middle row of Fig. 1 for MISE in the authors' paper). One can also see from Fig. 1 in my comments that the proportion of censored event-times for non-cured units is expected to grow with x. This might explain the counter-performance (as measured by MISE) of some estimators for large values of x.

In simulation Setting 3, I am not sure that the authors can claim that the NPSXZ and NPSXZ2 estimators outperform the others. From what I understand from the graphs for MISE at the bottom of the reported figure, it seems that all

[1] Institut de Mathématique, Université de Liège, Belgium.
[2] Institut de statistique, biostatistique et sciences actuarielles (ISBA), Université catholique de Louvain, Belgium.
E-mail: p.lambert@uliege.be

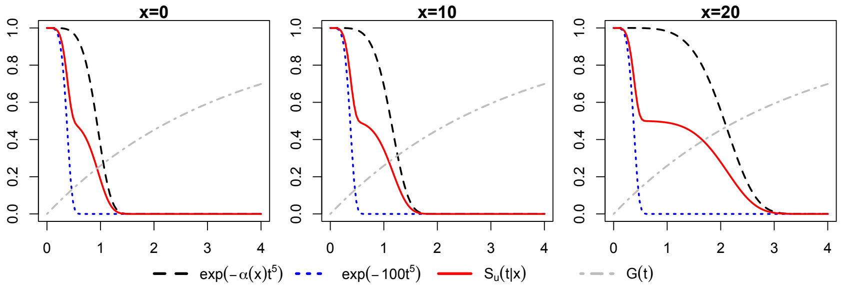

Fig. 1 Simulation Setting 2: conditional survival function for the uncured units, $S_u(t|x)$, and c.d.f. of the censoring variable, $G(t)$.

methods have similar performance, with the exception of the NPSXX estimator for small values of x and the semiparametric estimator when the sample size is (very) small.

To conclude my comments on the simulation results, I would like to stress that reporting MISE instead of RMISE tends to overemphasize differences that might not be relevant from a practical standpoint. In addition, focusing on MISE by not commenting the bias and the variance of the estimators does not enable to understand if the relative counter-performances of some estimation methods can be attributed to relatively larger biases or variances (or both).

The simulation study focused on the simple setting where a single covariate is assumed to impact the time-to-event distribution for the non-cured subjects. The same remark applies to the incidence part of the model. How does the method behave in more demanding settings where multiple covariates are available and needed to account for the complexity of the studied population? Beyond the theoretical extension of the estimation method, what would be the authors' expectations on the computational aspect?

The application section was very interesting to understand how the authors plan to use the developed tool in practice. I am a bit skeptical about the claim that banks for which no event was reported in the 2006-2017 period are not susceptible to bankruptcy and can somehow be considered 'cured'. Indeed, bankruptcy is fortunately a rare event for a bank and a much longer follow-up would be desirable to quantify this risk. Therefore, I wonder if another definition of 'cured' would not be more appropriate in this case, such as the absence of bankruptcy by the end of 2017 (despite the major financial and banking crisis of 2008). A distinction should also be made between systemic and non-systemic banks, as governments and central banks tend to intervene to support systemic banks when their activity is compromised.

The choice of taking the average value of time-varying covariates as regressors instead of their full history might not be appropriate, as banks might, for example, have changed their lending policies to cope with the crisis in 2008. Would a joint model for the event time and the three time-varying covariates in the model not be more appropriate? More simply (and with a non-negligible loss of information), using the baseline value of each time-varying covariate (at the beginning of the follow-up) would make more sense than conditioning on future values as is implicit in using the mean value of time-varying covariates over the observation period.

Although the authors indicate in Sections 2 and 3 that, without loss of generality, one single continuous covariate (Z in Section 2 and X in Section 3) is considered, one might have expected to have all three covariates considered simultaneously in the regression models in the Application section. Instead, the effects of covariates on the incidence and latency parts were studied one by one. Does it reflect the difficulty of the proposed estimation strategy to handle more than one covariate simultaneously? Beyond the theoretical extension or adaptation of the presented estimation method, what would be the authors' expectations on the computational side? Related to this, no indication of the time required to produce the different estimators was provided, including the bandwidth selection part. In particular, I would be curious to have a comparison of the competing methods on this issue. I was also wondering whether confidence regions or pointwise confidence intervals for the estimations reported in Figures 5 to 6 could be obtained by relying on asymptotic results and without making use of an additional bootstrap.

To conclude, I take this opportunity to remind that the analysis of survival data with an unknown cured fraction can also be made in the framework of the promotion time model (also named the bounded hazard model) (Yakovlev and Tsodikov, 1996). It can be extended to have separate or shared covariates to model the long-term survival probability (or 'cured' status) and the event dynamics for non-cured units (Bremhorst and Lambert, 2016; Bremhorst et al., 2016; Gressani and Lambert, 2018) provided that the follow-up is sufficiently long (Lambert and Bremhorst, 2019). Then, the population hazard takes the following form, $S_p(t|z,x) = \exp(-\theta(z)F(t|x))$, with conditional cure probability $\pi(z) = \exp(-\theta(z))$, where $\theta(z)$ is the maximal value for the cumulative hazard, and where $F(t|x)$ is a distribution function modelling the dynamics in the cumulative hazard. Flexible forms based on P-splines can be taken for $\theta(z)$ and $F(t|x)$ with additive models proposed to account for nonlinear effects of covariates (Bremhorst et al., 2019). It can also be extended to include multiple time-varying covariates (Lambert and Bremhorst, 2020) in a flexible way using additive models (Lambert and Kreyenfeld, 2023).

Funding ARC project IMAL (grant 20/25-107) financed by the Wallonia-Brussels Federation and granted by the Academie universitaire Louvain.

References

1. Bremhorst, V., M. Kreyenfeld, and P. Lambert (2016). Fertility progression in Germany: An analysis using flexible nonparametric cure survival models. *Demographic Research 35*, 505–534.
2. Bremhorst, V., M. Kreyenfeld, and P. Lambert (2019). Nonparametric double additive cure survival models : an application to the estimation of the nonlinear effect of age at first parenthood on fertility progression. *Statistical Modelling 19*, 275–279.
3. Bremhorst, V. and P. Lambert (2016). Flexible estimation in cure survival models using Bayesian P-splines. *Computational Statistics and Data Analysis 93*, 270–284.
4. Gressani, O. and P. Lambert (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. *Computational Statistics and Data Analysis 124*, 151–167.

5. Lambert, P. and V. Bremhorst (2019). Estimation and identification issues in the promotion time cure model when the same covariates influence long- and short-term survival. *Biometrical Journal 61*(2), 275–289.
6. Lambert, P. and V. Bremhorst (2020). Inclusion of time-varying covariates in cure survival models with an application in fertility studies. *J. R. Statist. Soc. A 183*, 333–354.
7. Lambert, P. and M. Kreyenfeld (2023). Exogenous time-varying covariates in double additive cure survival model with application to fertility. *arXiv.2302.00331*.
8. López-Cheda, A., Y. Peng, and M. Jácome (2023). Nonparametric estimation in mixture cure models with covariates. *Test*.
9. Yakovlev, A. and A. Tsodikov (1996). *Stochastic Models for Tumor of Latency and Their Biostatistical Applications.* World Scientific Publishing Singapore.