Noname manuscript No.
(will be inserted by the editor)

Rejoinder on: Nonparametric estimation in mixture cure models with covariates

Ana López-Cheda · Yingwei Peng · María Amalia Jácome

Received: date / Accepted: date

We thank all discussants for their insightful comments on our paper. The comments include some suggestions on possible extensions and some potential issues and concerns in our current work. We respond to the comments as follows.

Bo Han and Xiaoguang Wang raised the issue of the efficiency of the proposed nonparametric estimator and suggested potential gains when the incidence or the latency part is replaced with a parametric model, particularly based on the goodness-of-fit tests proposed by Müller and van Keilegom (2019) and Geng et al. (2023). We agree that parametric methods may be more efficient than nonparametric methods when the parametric requirements are met. The parametric methods should be considered if the parametric assumptions can be verified, and the recent progress in the goodness-of-fit tests in the mixture cure model is useful to provide support for using more efficient parametric methods.

They also suggested future work of developing alternative methods to estimate the weights in the EM algorithm by plugging nonparametric estimates of the cure probability in the incidence part and the survival function in the latency part. We like to point out that, in the proposed work, the weights are estimated as a function of the nonparametric estimates of the survival function in the latency part and the cure probability in the incidence part. The advantage of the proposed method is that the nonparametric estimate of the cure probability will not change between the EM algorithm, and only the nonparametric survival function estimate in the latency part needs to be updated. Alternative methods to estimate the weights are possible. For example, Li et al. (2020) showed a way to estimate the cure probability in the incidence part by using a support vector machine method, which is more flexible than a typical parametric method for the incidence part. The idea may be generalized to a fully nonparametric approach to update the weights.

Further work is indeed needed on the asymptotic properties of the proposed nonparametric estimator. Although the bootstrap method can be used in practice to assess the significance of the results, the lack of the asymptotic distribution for the proposed estimator may hinder the development of potentially simple test procedures.

Address(es) of author(s) should be given

Ricardo Cao suggested a neat way to present the mixture cure model with shared covariates in the latency and incidence parts and then showed how to use the marginals to remove unwanted covariates in the estimates from the estimate of the overall survival function. We also used the idea of marginals in the paper, such as at the beginning of Section 2 to motivate the cure rate estimator and in Equation (12) to motivate the alternative estimator. We thank Ricardo for providing the useful additions to fully explore this idea and we agree that it will be interesting to investigate the differences between the proposed estimator NPSXZ in the paper and the estimates that are completely determined by marginals. Simplicity is gained at the cost of computing Beran's estimator $\hat{S}^B(t|\mathbf{z},\mathbf{x})$ on the entire set of covariates. When the number of covariates $\mathbf{x}$ or $\mathbf{z}$ is medium or large, the curse of dimensionality makes the computation of Beran's estimator a real challenge. One caveat we have (it is mentioned in the paper too) is that the ideas of marginals should work if $\mathbf{x}$ and $\mathbf{z}$ are independent. César Sánchez-Sellero and Wenceslao González-Manteiga also question the consistency of the estimates from the method of using marginals in their comments when $\mathbf{x}$ and $\mathbf{z}$ are not independent. How the performance of the estimators depends on independence between $\mathbf{x}$ and $\mathbf{z}$ remains unclear.

We primarily focus on iid observations $\{(\tilde{t}_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i),\ i = 1, \ldots, n\}$ in this work. We thank Ricardo Cao for noticing that the iid setting is not clearly stated. The asymptotic properties of the proposed estimators strongly rely on those of Beran's estimator. So far the only available asymptotic theory for Beran's estimator with independent observations is confined to univariate covariates. Liang et al. (2012) studied the strong and weak convergence for Beran's estimator with multivariate covariates for a left-truncated and right-censored data where the lifetime observations are assumed to form a stationary α-mixing sequence. The results of Liang et al. (2012) particularized in the independent setting and without left truncation have been key to proving the asymptotic properties in Section 4. Specifically, both the convergence in Theorem 2 and, consequently, the asymptotic normal distribution for the convergence in distribution in Section 4, are strongly based on Equations (18) and (19) of the paper:

$$\sup_{\mathbf{x}\in I_{\mathbf{X}}} \sup_{\tau_1\le t\le\tau_2} \left|\widehat{S}_{u,h_2}(t|\mathbf{x}) - S_u(t|\mathbf{x})\right| = O\left((\log n)^{1/2}(nh_2^p)^{-1/2} + h_2^l\right) \quad a.s.$$

$$\sup_{\mathbf{z}\in I_{\mathbf{Z}}} |\widehat{\pi}_{h_1}(\mathbf{z}) - \pi(\mathbf{z})| = O\left((\log n)^{1/2}(nh_1^q)^{-1/2} + h_1^l\right) \quad a.s.$$

where p and q are the dimensions of $\mathbf{x}$ and $\mathbf{z}$, respectively, and l is the order of the multivariate kernel, $\mathbf{K}(\cdot)$, as stated in conditions (A1) and (A1'). For the NPSXZ estimator $\widehat{S}_{u,h}(t|\mathbf{x})$ to converge almost surely to $S_u(t|\mathbf{x})$ as n goes to infinity as stated in Theorem 2, the asymptotic bias, which is represented by h_1^l and h_2^l in the order of the negligible terms, must be killed, as Ricardo Cao pointed out. In other words, $h_2^l/[(\log n)^{1/2}(nh_2^p)^{-1/2})] = O(1)$ and $h_1^l/[(\log n)^{1/2}(nh_1^q)^{-1/2})] = O(1)$. This is obtained when $(\log n)^{-1}nh_2^{p+2l} = O(1)$ and $(\log n)^{-1}nh_1^{q+2l} = O(1)$, respectively. In the particular case of univariate covariates X and Z ($p = q = 1$), if a usual kernel of second order ($l = 2$) is considered, these latter assumptions for the bandwidths are fulfilled if $nh_1^5 \to 0$ and $nh_2^5 \to 0$ when $n \to \infty$, as Ricardo Cao mentioned. The assumptions $(\log n)^{-1}nh_2^{p+2l} = O(1)$ and $(\log n)^{-1}nh_1^{q+2l} = O(1)$ are needed but not included in Theorem 2 of the paper, we are very grateful to Ricardo Cao for pointing this out.

The conditions needed for Theorem 2 are the same assumptions as in Liang et al. (2012). Let us denote $\mathbf{W}$ a general d-dimensional vector of covariates with density function $f_{\mathbf{W}}(\mathbf{w})$, define $D_{\mathbf{W}} = \left\{\mathbf{w} \in \mathbb{R}^d | f_{\mathbf{W}}(\mathbf{w}) > 0\right\}$, and let $I_{\mathbf{W}}$ be a compact set of $\mathbb{R}^d$ included in $D_{\mathbf{W}}$. Theorem 2 holds for $\mathbf{x} \in I_{\mathbf{X},\mathbf{e}}$ and $\mathbf{z} \in I_{\mathbf{Z},\mathbf{e}}$, with $I_{\mathbf{W},\mathbf{e}} = \{\mathbf{w} \pm \mathbf{e}, \mathbf{w} \in I_{\mathbf{W}}\}$ mentioned in assumptions (A2)-(A4), where $\mathbf{e} = (e_1, \ldots e_d)$ for small $e_i > 0$. The meaning of *small* seems not decisive as long as $\inf_{w \in I_{\mathbf{w},\mathbf{e}}} f_{\mathbf{W}}(\mathbf{w}) \geq \delta_0 > 0$. The normality of $\widehat{S}_{u,h}(t|\mathbf{x})$ is outlined for $\mathbf{x} \in D_{\mathbf{X}}$ and $\mathbf{z} \in D_{\mathbf{Z}}$ such that assumptions (A2')-(A4') are fulfilled in the neighborhoods $U(\mathbf{x})$ and $U(\mathbf{z})$. Finally, it is important to mention that the bandwidth h_{1i} denotes the smoothing parameter used in the EM algorithm to estimate $\pi(\mathbf{z}_i)$, for $i = 1, \ldots, n$, and h_{2i} is the bandwidth for the estimation of $S_u(t|\mathbf{x}_i)$. The subscript i emphasizes the local nature of the bandwidths in the sense that a different pair of bandwidths (h_{1i}, h_{2i}) is needed to compute the aforementioned functions conditioned on $\mathbf{x} = \mathbf{x}_i$ and $\mathbf{z} = \mathbf{z}_i$ using all the observations $\{(\tilde{t}_j, \delta_j, \mathbf{x}_j, \mathbf{z}_j), j = 1, \ldots, n\}$. Notwithstanding the foregoing, these bandwidths depend on n in the usual way and must fulfill the general conditions for a bandwidth h, given by $h \to 0$ and $(\log n)^{-1} nh^p \to 0$ if h is used to estimate $\pi(\mathbf{z})$, or $(\log n)^{-1} nh^q \to 0$ if h is used to estimate $S_u(t|\mathbf{x})$.

Ricardo Cao was concerned about how the time-dependent covariates were handled in the real data analysis and suggested that using the baseline values of the covariates instead of their average values over the follow-up period as the time-independent variables will be more useful for prediction. Philippe Lambert also raised this concern as well. We think this is a good suggestion and we revised the analysis accordingly. The new results are presented in Figure 1 in this rejoinder. The estimates of the cure rate as functions of baseline retail deposits (COREDEP) and the baseline number of total loans (LOANS) are very similar to the ones as functions of the averaged-over-time values (see Figure 4 in the paper). However, the results for the return on assets (ROA) deserve some comments. Large values of ROA are usually associated with stronger and safer banks. The difference in the estimated density functions between the baseline values of ROA (blue color in Figure 1) and the averaged values of ROA (grey color in Figure 1) indicates that ROA values have decreased during the follow-up period, possibly during the banking crisis of 2008. The estimated probability of bankruptcy does not seem to depend on the baseline values of ROA, and the effect of ROA on the probability of bankruptcy is not significant at a significance level of 5% (p_{CvM}=0.0695, p_{KS}=0.0775).

We agree that double-index instead of single-index may be the right term for an extension of the two-parts mixture cure model to deal with the curse of dimensionality issue when the number of covariates is large, and it is what we meant in this context. In the paper, we did not discuss in detail how to decide whether model (1) or model (7) should be considered in practice. This can be determined on a priori grounds or empirical grounds, the latter requires statistical tests such as the covariate significance tests. Determining what statistical model to use based on statistical tests is, however, not deemed a good practice in statistical analysis, and thus it should be used with caution.

César Sánchez-Sellero and Wenceslao González-Manteiga speculate that the nonparametric cure rate estimate based on the estimated nonparametric survival function evaluated at the largest uncensored time may not perform as well as a

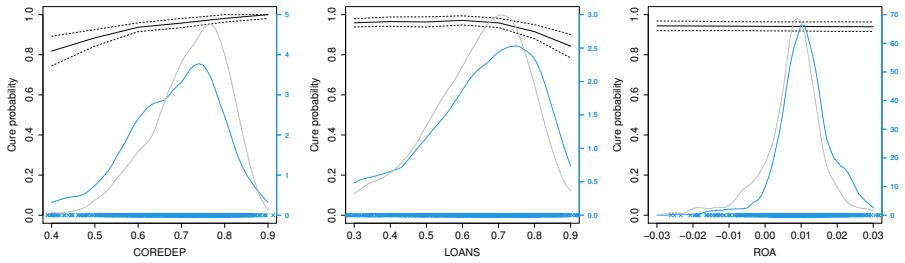

Fig. 1 Nonparametric estimation of the probability of immunity to bankruptcy (solid black line) as a function of the baseline values of COREDEP (left), LOANS (center), and ROA (right). The 95% confidence intervals (dashed black lines) are computed using the percentile bootstrap method. The blue (grey) line represents the Parzen–Rosenblatt density estimations of the baseline covariates (mean values of the covariates), using Sheather and Jones' plug-in bandwidth.

parametric model such as the logistic model because the nonparametric survival function estimate tends to be unstable due to heavy censoring at the right tail of the distribution. We agree that both the Kaplan-Meier survival function estimate and Beran's generalized Kaplan-Meier survival estimate can be unstable at the right tail if censoring is heavy at the right tail, and indeed the survival data that are suitable for cure models usually have heavy censoring at the right tail. However, there is a subtle difference in the heavy censoring for the data that are not suitable for cure models and the data that are suitable for cure models. Applying cure models to survival data requires that the data come from a study with a sufficient follow-up, which guarantees that there are uncensored times observed in the right tail of the survival function of the survival time of uncured subjects and that the cure probability is identifiable. Because of this, the nonparametric and semiparametric cure models often assume that the survival function of uncured subjects is zero for any time greater than or equal to the largest uncensored time, which means that subjects with censored times greater than the largest uncensored time are treated as cured and the censoring times do not contribute to the estimation of the survival function of uncured subjects. Thus when considering estimating the survival function of uncured subjects using the proposed nonparametric method and the existing semiparametric method, the effective censoring rate in data that contribute to the estimation is much smaller than the observed censoring rate. This can be seen in the estimator given in (9) in the paper. The weight w_j in the denominator is very small for a censored time closer to the largest uncensored time and it is 0 for the censoring times larger than the largest uncensored time. Therefore, w_j essentially reduces or removes a substantial number of the censored times in the estimation and it reduces the effective censoring rate. This explains why the nonparametric cure rate estimate, whether it is based on the Kaplan-Meier survival function estimate suggested by Maller and Zhou (1992) or based on Beran's generalized Kaplan-Meier survival estimate suggested in Xu and Peng (2014) and López-Cheda et al. (2017), tends to work well as long as sufficient follow-up is guaranteed.

Nevertheless, censoring does have a larger impact on the survival function of uncured subjects even though it may be at a less degree than in the situation

without cured subjects. César Sánchez-Sellero and Wenceslao González-Manteiga suggested a possible way to further reduce the bias and the variance of the survival estimation by considering the work of Stute (1994) and Stute (1996). They also suggested quantile methods in the estimation as an alternative way to the weighting method to lower the impact of the tail estimation on the mean square error. The quantile methods for cure models have been considered in the work of Wu and Yin (2013), Wu and Yin (2017a), and Wu and Yin (2017b). The existing methods usually assume a parametric logistic model for the incidence part and a semiparametric quantile regression model for the latency part. However, there is no reason to believe that the cure rate is always logistic, not even monotone in $\mathbf{z}$. Exploring a nonparametric method for assessing covariate impact on the quantiles of the survival distribution of uncured subjects based on the current work with the nonparametric method for the incidence part is certainly a very interesting idea and will be examined in future work.

Philippe Lambert raised a few practical issues in our numerical studies. We agree that larger sample sizes could be considered in the simulation study and better results can be expected for the nonparametric methods. Nonetheless, due to the consistency of compared methods, differences among them tend to fade away with large sample sizes. A simulation study with very small samples can be viewed as a stress test to determine their efficiency under this not-unusual situation.

Regarding the results of the simulation study for Setting 2, we understand the concern. The latency function $S_u(t|x)$ is symmetric in x, since (x, z) are generated independently hence it is reasonable to expect MISE(x) inherit the symmetry, and Figure 1 (middle row) shows similar values for MISE(x) for opposite values of x. We think that should be the case if $X \in U(-10, 10)$. However, covariate x was generated as a uniform in $(-10, 20)$. When using kernel methods near the edges of the support, kernel estimates often overspill the boundaries and are consequently biased. This boundary effect might be only noticeable in the estimation of $S_u(t|x)$ for negative values of x especially close to -10, but not for the corresponding positive values close to 10, as the upper limit of the support of X is not 10 but 20 and this boundary effect does not appear anymore. This results in different local bandwidths for opposite values of x, and consequently different MISE(x) values.

We also agree with Philippe Lambert that, in Setting 3, all methods have similar performance for large sample sizes ($n = 200$), with a slightly worse behavior of the NPSXX estimator, especially for small values of x. For smaller sample sizes, the differences become slightly more apparent; the best performance of NPSXZ, NPSXZ2, and PVK estimators is quite comparable, the first two being preferable for $n = 50$ especially for large values of x.

The key differences between RMISE and MISE are their interpretation and their behavior on large differences. A benefit of using RMISE is that the metric it produces is in terms of the unit being predicted. MISE squares the error, leading to a result more difficult to interpret, and large errors being punished or highlighted. The choice of the metric is dependent upon each use. Comparing estimating methods using either RMISE or MISE makes no difference in terms of ranking their efficiencies, and MISE is often the go-to metric for estimators of the survival function. Providing only MISE results hinders understanding the trade-off between bias and variance of the estimators, not allowing to attribute the hypothetical bad behavior of the estimators to large bias, large variance, or both. MISE does enable

a fair sound comparison of the five estimation methods in all the settings and all sample sizes for all the covariate values from a graphical viewpoint.

Regarding the comment about the effects of the covariates on the incidence and latency parts studied individually, we agree that it would be more suitable to study them jointly in the same model. However, when applying the method to multiple covariates, we faced the challenges of computing kernel-based estimation methods, such as Beran's estimator, and the curse of dimensionality. We also used single covariates in the latency and cure rate parts to avoid the risk of possible dependence between covariates in $\mathbf{x}$ and $\mathbf{z}$. Besides, although the NPSXX estimator can deal with more than one covariate, for its computation we used the `npcure` package in R that can only work with unidimensional covariates.

Bankruptcy is fortunately a rare event for a bank. Phillipe Lambert wonders whether banks not bankrupt within the 2006 - 2017 period are insusceptible to bankruptcy or not. Indeed, a much longer follow-up would be desirable to consider a bank as 'cured' from bankruptcy, and a different term, such as 'long-term survivors', would be more appropriate than 'cured' to describe the banks. We also agree that systemic banks may have a distinct behavior from non-systemic banks in terms of bankruptcy. Systemic banks are classified as Systemically Important Financial Institution (SIFI). A set of stricter requirements would apply to SIFI banks, and they tend to be supported by governments and central banks when their financial viability is compromised since they are 'too big to fail'. Including an indicator of SIFI banks in the model could yield different and interesting conclusions.

Philippe Lambert also pointed out the importance of including the computing time required to produce the different estimators, considering the bandwidth selection part. The proposed method is computationally intensive and the implementation time deserves some comments. Besides the sample size, the computational time also depends on the number of bootstrap resamples, B, and the length of the grids of bandwidths from which the optimal bandwidths for the cure rate and the latency are obtained. Using an AMD Ryzen 9 5950X 16-Core Processor, 3.40 GHz, 128 GB RAM, in the real data analysis of $n = 500$ commercial banks, the computational time to obtain the proposed NPSXZ latency estimator with $B = 100$ bootstrap resamples and a grid of 10 bandwidths is 2.38 hours. If the search grid is increased to 50 (100) bandwidths, then the computational time increases to 5.84 (9.96) hours. For $B = 200$ bootstrap resamples and a grid of 10 (50, 100) bandwidths, the computational time is around 4.01 (10.55, 17.90) hours, respectively.

Providing confidence regions and pointwise confidence intervals for estimates is of great interest to moving into an inferential framework. However, obtaining confidence regions and intervals based on asymptotic results requires estimating unknown functions in the expressions of the asymptotic normal distribution. In addition, the normal approximation does not usually work well in practice, the convergence is usually too slow to get good results for finite samples. A bootstrap procedure is usually considered instead to approximate confidence regions and intervals.

Time-dependent covariates are currently not considered in this work. Including time-dependent covariates in survival analysis is always a challenging task and it is currently only possible in a handful of models. There are a few recent works on including time-dependent covariates in cure models (Dirick et al., 2019; Lambert

and Bremhorst, 2019; Dong et al., 2022). The data analysis in this work clearly shows the importance and significance of extending the existing work in the future to allow time-dependent continuous covariates in nonparametric mixture cure models.

As pointed out by Philippe Lambert that this work only focuses on the mixture cure model. We did not consider the promotion time cure model. The mixture cure model arises naturally in the context of a mixed population of cured and uncured subjects and it provides a simple framework with few assumptions needed for the proposed nonparametric method. The promotion time cure model involves latent dynamics that are not easy to interpret in practice. It also involves the proportional hazards assumption, which does not meet our goal to have a nonparametric estimation method that does not make any unnecessary assumptions beyond assuming the presence of cured subjects.

References

Dirick L, Bellotti T, Claeskens G, Baesens B (2019) Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. Journal of Business & Economic Statistics 37(1):40–53

Dong Q, Peng Y, Li P (2022) Time to delisted for listed firms in Chinese stock markets: An analysis using a mixture cure model with time-varying covariates. Journal of the Operational Research 73(10):2358–2369

Geng Z, Li J, Niu Y, Wang X (2023) Goodness-of-fit test for a parametric mixture cure model with partly interval-censored data. Statistics in Medicine 42(4):407–421

Lambert P, Bremhorst V (2019) Inclusion of Time-Varying Covariates in Cure Survival Models with an Application in Fertility Studies. Journal of the Royal Statistical Society Series A: Statistics in Society 183(1):333–354

Li P, Peng Y, Jiang P, Dong Q (2020) A support vector machine based semiparametric mixture cure model. Computational Statistics 35:931–945

Liang HY, de Uña-Álvarez J, del Carmen Iglesias-Pérez M (2012) Asymptotic properties of conditional distribution estimator with truncated, censored and dependent data. TEST 21(4):790–810

López-Cheda A, Cao R, Jácome MA, Van Keilegom I (2017) Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. Computational Statistics & Data Analysis 105:144–165

Maller RA, Zhou S (1992) Estimating the proportion of immunes in a censored sample. Biometrika 79(4):731–739

Müller UU, van Keilegom I (2019) Goodness-of-fit tests for the cure rate in a mixture cure model. Biometrika 106(1):211–227

Stute W (1994) Improved estimation under random censorship. Communications in Statistics-Theory and Methods 23(9):2671–2682

Stute W (1996) Distributional convergence under random censorship when covariables are present. Scandinavian Journal of Statistics 23(4):461–471

Wu Y, Yin G (2013) Cure rate quantile regression for censored data with a survival fraction. Journal of the American Statistical Association 108:1517 – 1531

Wu Y, Yin G (2017a) Cure rate quantile regression accommodating both finite and infinite survival times. Canadian Journal of Statistics 45(1):29–43

Wu Y, Yin G (2017b) Multiple imputation for cure rate quantile regression with censored data. Biometrics 73(1):94–103

Xu J, Peng Y (2014) Nonparametric cure rate estimation with covariates. Canadian Journal of Statistics 42:1–17