proceedings

MDPI

Proceedings

A Doubly Smoothed PD Estimator in Credit Risk†

Rebeca Peláez Suárez [1,*], Ricardo Cao Abad [2] and Juan M. Vilar Fernández [2]

[1] Research Group MODES, Department of Mathematics, CITIC, University of A Coruña, 15001 A Coruña, Spain

[2] Research Group MODES, Department of Mathematics, CITIC, University of A Coruña and ITMATI, 15782 A Coruña, Spain; ricardo.cao@udc.es (R.C.A.); juan.vilar@udc.es (J.M.V.-F.)

* Correspondence: rebeca.pelaez@udc.es

† Presented at the 3rd XoveTIC Conference, A Coruña, Spain, 8–9 October 2020.

Published: 1 September 2020

check for updates

Abstract: In this work a doubly smoothed probability of default (PD) estimator is proposed based on a smoothed version of the survival Beran's estimator. The asymptotic properties of both the smoothed survival and PD estimators are proved and their behaviour is analyzed by simulation. The results allow us to conclude that the time variable smoothing reduce the error committed in the PD estimation.

Keywords: probability of default; risk analysis; censored data; survival analysis; nonparametric estimation; kernel estimation

1. Introduction

The debts coming from clients with unpaid credits have a important impact in the solvency of banks and other credit institutions. Therefore, one of the most crucial elements that influences the risk in credits is the probability of default (PD). For a fixed time, t, and a horizon time, b, the PD can be defined as the probability that a credit that has been paid until time t, becomes unpaid not later than time $t+b$.

The probability of default conditional on the credit scoring can be written as a transformation of the conditional survival function. Therefore, in Section 2.1 Beran's survival estimator is used to obtain a PD estimator. A time variable smoothing for this estimator is proposed in Section 2.2. In Section 3, both estimators are applied to a real data set. Section 4 contains some concluding remarks.

2. Nonparametric PD Estimators

Let $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ be a simple random sample of (X, Z, δ) where X is the credit scoring, $Z = \min\{T, C\}$ is the observed maturity, T is the time to default, C is the time until the end of the study or the time until the anticipated cancellation of the credit and $\delta = I_{\{T \leq C\}}$ is the uncensoring indicator. Let x be a fixed value of the covariate X, b a horizon time and $S(t|x)$ the conditional survival function of T. Then, the probability of default in a time horizon $t+b$ from a maturity time t is defined as follows

$$PD(t|x) = P(T \leq t+b | T > t, X = x) = 1 - \frac{S(t+b|x)}{S(t|x)}. \quad (1)$$

Replacing $S(t|x)$ with a nonparametric estimator, $\widehat{S}(t|x)$, in (1), the following estimator for $PD(t|x)$ is obtained:

$$\widehat{PD}(t|x) = 1 - \frac{\widehat{S}(t+b|x)}{\widehat{S}(t|x)}. \quad (2)$$

In [5] the theoretical results that allow to obtain, under general conditions, asymptotic properties for a PD estimator are proved. They are based on these properties for the corresponding estimator of the conditional survival function.

2.1. Beran's Estimator

Beran's survival estimator proposed in [1] is given by

$$\widehat{S}_h^B(t|x) \quad = \quad \prod_{i=1}^{n}\left(1-\frac{I_{\{Z_i\leq t,\,\delta_i=1\}}w_{i,n}(x)}{1-\sum_{j=1}^{n}I_{\{Z_j<Z_i\}}w_{n,j}(x)}\right), \tag{3}$$

where the weights are $w_{i,n}(x) = \dfrac{K\big((x-X_i)/h\big)}{\sum_{j=1}^{n}K\big((x-X_j)/h\big)}$, with $i=1,\ldots,n$, K is a kernel function and $h=h_n>0$ is a smoothing parameter. Now, replacing $\widehat{S}(t|x)$ with $\widehat{S}_h^B(t|x)$ in (2), Beran's estimator of the probability of default, $\widehat{PD}_h^B(t|x)$, is available. It was firstly used in [2].

The asymptotic properties of Beran's estimator for the conditional survival function were proven in both [3,4] under certain assumptions. From them, the expressions of the bias and the variance of the estimator $\widehat{PD}_h^B(t|x)$ can be found by using Theorem 1 in [5].

A simulation study was conducted in order to analyse the performance of Beran's estimator. Its behavior was compared with other estimators of the probability of default obtained from estimators of the survival function, including a benchmark method based on proportional hazards models. For more details about the simulation study, see [5].

The results show that the probability of default estimations obtained by means of the estimators built according to (2) are very reasonable, but they have excessive variability and they are very rough curves.

2.2. Smoothed Beran's Estimator

Beran's estimator is smoothed with respect to the covariate, but not with respect to the time variable. This fact along with the survival ratio structure of the PD estimator could be the cause of the instability of the estimations. Therefore, a time variable smoothing of the survival estimator is proposed.

The smoothed Beran's survival estimator is given by

$$\widetilde{S}_{h,g}^B(t|x) \quad = \quad 1-\sum_{i=1}^{n}s_{(i)}\mathbb{K}\left(\frac{t-Z_{(i)}}{g}\right) \tag{4}$$

where $s_{(i)}=\widehat{S}_h^B(Z_{(i-1)}|x)-\widehat{S}_h^B(Z_{(i)}|x)$ with $Z_{(i)}$ the i-th element of the sorted sample of Z, $\mathbb{K}(t)$ the distribution function of a kernel K and $g=g_n$ is the smoothing parameter for the time variable. Finally, the smoothed Beran's PD estimator, $\widetilde{PD}_{h,g}^B(t|x)$, is obtained by replacing $\widehat{S}(t|x)$ with $\widetilde{S}_{h,g}^B(t|x)$ in Equation (2).

The asymptotic expressions for the bias and the variance of the smoothed Beran's estimator of the survival function have been recently found [6]. The results are too extensive to be shown here. By applying Theorem 1 of [5], the corresponding asymptotic properties of the smoothed Beran's estimator of the PD are obtained.

The simulation study carried out shows that the time variable smoothing significantly reduces the error committed in the PD estimation. This technique implies a considerable increase in the computation time and the improvement is not very noticeable in the estimation of the survival function. However, in the case of the PD, the variability and roughness of the estimations is clearly reduced.

3. Application to Real Data

To illustrate the differences between the estimator based on Beran's and its smoothed version, we obtain the estimation of the conditional survival function and the PD in a real data set. The data consists of a sample of 10,000 consumer credits from a Spanish bank registered between July 2004 and November 2006. The sample contains the credit scoring of each borrower, the observed lifetime and the uncensoring indicator. The sample censoring percentage is 92.8%. The probability of default is estimated using Beran's and smoothed Beran's estimators with $h = 0.05$ and $g = 3$. Figure 1 shows the result.

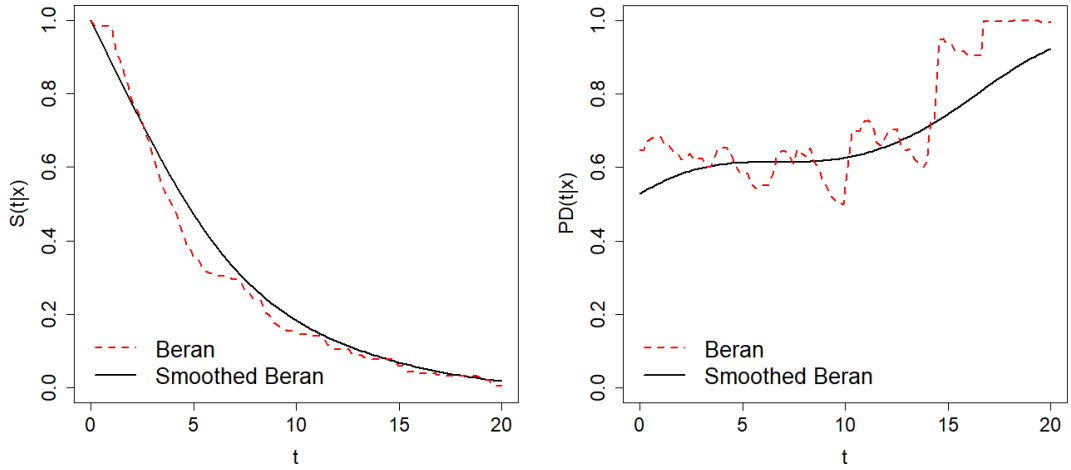

Figure 1. Estimation of $S(t|x)$ (**left**) and $PD(t|x)$ (**right**) at horizon $b = 5$ for $x = 0.8$ by means of Beran's (dashed line) and smoothed Beran's (solid line) estimators on the consumer credits dataset.

4. Conclusions

This work proposes a time variable smoothing for Beran's estimator of the conditional survival function. General asymptotic expressions for the bias and the variance of this estimator are proven. It is used to build a doubly-smoothed PD estimator whose asymptotic properties are also proved. In view of the simulation study carried out, it can be concluded that the smoothed Beran's estimator seems to reduce the estimation error committed when estimating the probability of default.

Work is currently underway to develop a method for choosing the smoothing parameters involved in the above-mentioned estimators. In addition, since the censoring probability is heavy in this context, nonparametric cure models are going to be considered in the study.

Acknowledgments: This research has been supported by MINECO Grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

1. Beran, R. *Nonparametric Regression with Randomly Censored Survival Data*; Technical Report; University of California: Oakland, CA, USA, 1981.
2. Cao, R.; Vilar, J.M.; Devia, A. Modelling consumer credit risk via survival analysis (with discussion). *Stat. Oper. Res. Trans.* **2009**, *33*, 3–30.
3. Dabrowska, D.M. Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Ann. Stat.* **1989**, *17*, 1157–1167.
4. Iglesias-Pérez, M.C.; González-Manteiga, W. Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *J. Nonparametr. Stat.* **1999**, *10*, 213–244.

5. Peláez Suárez, R.; Cao Abad, R.; Vilar Fernández, J.M. Probability of default estimation in credit risk using a nonparametric approach. *TEST* **2020**, 1–23, doi:10.1007/s11749-020-00723-1.
6. Peláez Suárez, R.; Cao Abad, R.; Vilar Fernández, J.M. Nonparametric estimation of the probability of default with double smoothing. **2020**, under preparation.