

Estimación no paramétrica de la probabilidad de mora en riesgo de crédito

Rebeca Peláez Suárez
Departamento de Matemáticas
Universidade da Coruña
ORCID: 0000-0002-3386-5751
rebeca.pelaez@udc.es

Directores:
Ricardo Cao Abad
Juan M. Vilar Fernández

Keywords: Bootstrap, Censored data, Kernel method, Mixture cure models, Survival analysis

MSC Subject classifications: 62N01, 62N02, 62G05, 62G08, 91G40.

El riesgo de crédito se define como la posible pérdida que asume un agente económico en caso de que la contraparte incumpla sus obligaciones contractuales. Las deudas procedentes de clientes con créditos impagados tienen un importante impacto en la solvencia de los bancos y otras entidades financieras. Por esta razón, el Comité de Basilea para la Supervisión Bancaria estableció en 2004 un conjunto de mecanismos estándar para la medición del riesgo de crédito en las instituciones financieras. Según este Acuerdo de Basilea, uno de los elementos de mayor influencia en el riesgo de crédito es la probabilidad de mora, denotada por PD, por sus siglas en inglés, *probability of default*. Se considera que un préstamo está en mora si lleva más de 90 días impagado. Para un tiempo fijo, t , y un horizonte de mora, b , la probabilidad de mora puede definirse como la probabilidad de que un crédito que ha sido pagado hasta el instante t , caiga en mora no más tarde del instante $t + b$. Para estimar la PD, los bancos e instituciones financieras utilizan covariables que contienen información sobre los créditos y los correspondientes clientes. La información contenida en estas covariables suele resumirse mediante alguna combinación lineal de las mismas que mide la solvencia de clientes o futuros clientes. Esta combinación lineal recibe el nombre de puntuación crediticia o *credit scoring* en inglés. En este escenario, la variable de interés es el tiempo hasta la caída en mora, únicamente observable cuando la caída en mora tiene lugar durante el tiempo de observación de los créditos; en otro caso, el dato es censurado y el tiempo observado es el tiempo hasta la censura. La puntuación crediticia juega el papel de variable predictora, proporcionando información acerca del tiempo hasta la caída en mora. Se asume, por tanto, que la probabilidad de mora es una curva, $PD(t|x)$, a lo largo del tiempo, t , que depende de la puntuación crediticia, x .

El análisis de supervivencia es un conjunto de procedimientos estadísticos que permiten describir y estudiar datos cuando la variable de interés es el tiempo que transcurre hasta que se produce cierto evento y una proporción de estos tiempos están censurados. Por otro lado, los estimadores no paramétricos son estimadores flexibles que requieren, a lo sumo, condiciones de continuidad y diferenciabilidad para la curva subyacente. Esto hace que estas técnicas sean apropiadas para estudiar el escenario anteriormente presentado. En este trabajo se proponen modelos de supervivencia que permiten estimar de forma no paramétrica la probabilidad de mora como función de la puntuación crediticia en créditos personales. Sea T la variable que denota el tiempo hasta la caída en mora y X la puntuación crediticia. Dados x un valor fijo de la puntuación crediticia y b un horizonte de mora, la probabilidad de mora se escribe formalmente del siguiente modo:

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

donde $S(t|x)$ es la función de supervivencia condicional del tiempo hasta la mora, T . Así, empleando técnicas del análisis de supervivencia para proponer estimadores de $S(t|x)$, es posible obtener estimadores no paramétricos de la PD. En la primera fase de esta investigación se consideran cuatro estimadores para la función de supervivencia condicional: el estimador de Beran, 1981, el estimador de I. M. A. Van Keilegom, 1999, el estimador lineal local ponderado propuesto por Cai, 2003 y el estimador de Nadaraya-Watson ponderado presentado en Peláez, Cao y Vilar, 2021b. Estos estimadores de la supervivencia se transforman de acuerdo a la expresión dada en (1) para obtener los correspondientes estimadores de la PD y se demuestran resultados generales acerca de las propiedades asintóticas de los estimadores de la PD así contruidos. Para analizar el comportamiento de los estimadores de la PD basados en los estimadores de Beran, Van Keilegom-Akritas, lineal local ponderado y Nadaraya-Watson ponderado de la supervivencia, se lleva a cabo un estudio de simulación basado en varios modelos con diferentes escenarios de censura. Los resultados obtenidos muestran que el estimador de Beran es el que ofrece las mejores aproximaciones de la PD, en cuanto a que proporciona un menor error cuadrático medio integrado en la mayor parte de los escenarios analizados. Además, requiere un menor tiempo de computación.

Las estimaciones de la probabilidad de mora obtenidas mediante estos estimadores son muy razonables, pero tienen una variabilidad excesiva y son curvas muy rugosas. El origen de esta variabilidad es el cociente de funciones supervivencias que debe hacerse para estimar la PD (véase (1)). Este cociente magnifica los saltos que caracterizan a los estimadores de la función de supervivencia, estimadores suaves en la covariable, pero funciones a saltos en la variable tiempo. Por ello, se propone un estimador no paramétrico de la función de supervivencia condicional doblemente suavizado, tanto en la covariable como en la variable temporal (véase Peláez, Cao y Vilar, 2022b) que posteriormente se utilizará para construir un estimador de la PD también suavizado en la variable tiempo. La técnica aquí presentada es general y permite suavizar en la variable tiempo numerosos estimadores ya conocidos de la función de supervivencia condicional. Sin embargo, este trabajo se centra principalmente en el estimador suavizado de la supervivencia basado en el estimador clásico de Beran. Se demuestran sus propiedades asintóticas: se obtiene una representación casi segura del estimador, expresiones asintóticas del sesgo y varianza y la distribución límite del mismo. Mediante estudios de simulación, se comprueba que este suavizado supone una considerable reducción del error de estimación.

Basándose en este estimador suavizado de Beran para la función de supervivencia, se presenta un estimador no paramétrico de la probabilidad de mora con doble suavizado que pretende solucionar el problema de variabilidad observado en las estimaciones de la PD obtenidas hasta el momento. Se obtienen las expresiones asintóticas del sesgo y la varianza del estimador de la probabilidad de mora basado en el estimador de Beran suavizado. También se demuestra la normalidad asintótica del

mismo. El estudio de simulación realizado muestra que el suavizado de la variable temporal reduce significativamente el error cometido en la estimación de la PD. Esta técnica implica un aumento considerable del tiempo de cálculo. Sin embargo, la variabilidad y la rugosidad de las estimaciones se ven claramente reducidas. Además, las simulaciones llevadas a cabo permiten concluir que otros estimadores de la PD también presentan estas mejoras cuando se suavizan en la variable temporal. También se analiza la relación entre los dos parámetros de suavizado involucrados en el estimador y su influencia en el error cuadrático medio integrado. Estos contenidos pueden consultarse en Peláez, Cao y Vilar, 2021a.

La utilidad del doble suavizado se ilustra mediante el análisis del conjunto de créditos bancarios concedidos por cierta entidad financiera española, donde se observan las diferencias entre los estimadores de Beran y Beran suavizado de la PD en un problema real. Los datos consisten en una muestra de 10000 créditos de consumo concedidos entre julio de 2004 y noviembre de 2006. Se dispone de la puntuación crediticia de cada prestatario, el tiempo de vida observado del crédito en meses y el indicador de mora. El porcentaje de censura de la muestra es de 92.8%; equivalentemente, la proporción de créditos en los que se observa el impago es de 7.2%. De la muestra original se obtuvo una submuestra intencionadamente sesgada, para no mostrar la verdadera situación de solvencia del banco y preservar así la confidencialidad. Se estima la probabilidad de mora a un horizonte de $b = 5$ meses para un cliente con puntuación crediticia $x = 0,8$ a lo largo del intervalo de tiempo $[0, 25]$. El resultado se puede ver en la Figura 1.

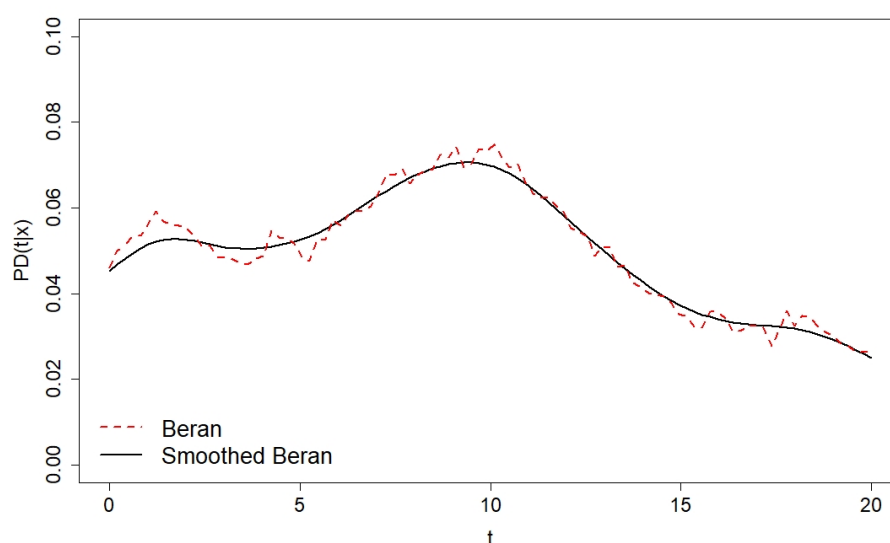


Fig. 1: Estimación de $PD(t|x)$ en el horizonte $b = 5$ para $x = 0.8$ mediante el estimador de Beran (línea discontinua) y el estimador suavizado de Beran (línea continua) en el conjunto de datos de créditos de consumo.

Aunque las estimaciones de la PD son muy similares con ambos estimadores, puede observarse cómo se reduce la rugosidad y se eliminan los saltos artificiales, no esperados en este tipo de curvas, cuando se utiliza el estimador de Beran suavizado. Según la estimación suavizada de Beran, la probabilidad

de mora tiene una tendencia decreciente. De ello se deduce que, para un individuo con esta puntuación crediticia, cuanto mayor es la madurez del crédito, menor es la probabilidad de caer en impago.

Se podría pensar en la estimación no paramétrica, y, en particular, los métodos kernel propuestos en esta tesis, como un procedimiento de estimación sin parámetros. Pero lo cierto es que es necesario estimar parámetros de suavizado o ventanas involucrados en los estimadores. La diferencia con la estimación paramétrica es que no es necesario asumir ninguna forma específica de la curva a estimar, en este caso la probabilidad de mora. Si bien los parámetros de suavizado óptimos en este escenario son aquellos que minimizan el error cuadrático medio integrado, la complejidad de las expresiones asintóticas obtenidas para el sesgo y varianza de los estimadores propuestos dificulta la búsqueda de dichos parámetros mediante este abordaje. Por este motivo, se proponen métodos basados en técnicas bootstrap para la selección automática del parámetro de suavizado en la covariable para el estimador de Beran y de los parámetros de suavizado en la covariable y en la variable tiempo para el estimador de Beran suavizado tanto de la función de supervivencia condicional, como de la probabilidad de mora. Las técnicas de remuestreo propuestas se basan en combinar un bootstrap suavizado con un bootstrap con covariables. Las ventanas bootstrap se obtienen mediante la minimización de la aproximación bootstrap del error cuadrático medio integrado. El rendimiento de las técnicas propuestas se analiza mediante simulaciones en diferentes escenarios, obteniendo resultados prometedores (véase Peláez, Cao y Vilar, 2022a). También se aborda la obtención de regiones de confianza mediante procedimientos bootstrap para la curva de probabilidad de mora, $PD(t|x)$, para un valor fijo de $x \in I \subseteq \mathbb{R}$ y t variando en el intervalo $I_T \subseteq \mathbb{R}^+$, utilizando los estimadores de Beran y de Beran suavizado.

Finalmente, se discuten técnicas para estimar la PD basadas en modelos de curación. El tiempo hasta la caída en mora podría enfrentarse no sólo a un problema de censura por la derecha, sino también a la presencia de curación. Podrían existir clientes que nunca caen en mora, es decir, no importa cuánto tiempo se observe a tales individuos, nunca experimentarán el evento de interés. Se considera el estimador no paramétrico de la función de supervivencia propuesto por López-Cheda, Cao y Jácome, 2017 y López-Cheda, Cao, Jácome e I. Van Keilegom, 2017 basado en modelos de curación de tipo mixtura y se transforma para obtener el estimador no paramétrico de modelos de curación (NPCM) de la PD. Se analizan las propiedades asintóticas de dicho estimador: se obtiene una representación casi segura y expresiones asintóticas del sesgo y la varianza, así como su normalidad asintótica. Su comportamiento se compara mediante simulación con el estimador de Beran de la PD, el estimador de Beran doblemente suavizado y con ciertos métodos paramétricos. Los resultados obtenidos muestran que el estimador NPCM proporciona buenas estimaciones de la PD, reduciendo el error cometido por las alternativas semiparamétricas. El estimador de Beran de la PD es competitivo con el estimador NPCM en la mayoría de los escenarios. El estimador de Beran doblemente suavizado es, de nuevo, la alternativa que proporciona un menor error de estimación en todos los escenarios analizados.

La principal área de aplicación de los procedimientos propuestos y estudiados a lo largo de esta tesis es el riesgo de crédito. Sin embargo, estas técnicas pueden ser de utilidad en contextos biomédicos, donde predecir el tiempo hasta la ocurrencia de un evento condicionado a una covariable relacionada con dicho tiempo es un problema de interés. Por ejemplo, el tiempo de supervivencia de un individuo que participa en un estudio clínico y sus características fisiológicas (sexo, colesterol, edad...). El doble suavizado es una herramienta novedosa que puede aportar mejoras en los resultados. Además, el selector bootstrap propuesto resuelve el problema de elección de los parámetros de suavizado de los que depende el estimador de Beran suavizado. Para ilustrar esta aplicación se llevó a cabo un análisis de tiempos de hospitalización de pacientes de COVID-19 en Galicia, España. Las técnicas desarrolladas se emplean para estudiar en profundidad la relación entre la edad, el sexo y ciertas patologías previas, como obesidad o EPOC, con el tiempo de recuperación de pacientes infectados con

SARS-CoV-2. Los datos fueron proporcionados por el Servicio Gallego de Salud y se corresponden con ingresos en hospitales gallegos durante las primeras semanas de la pandemia.

Consideramos que la principal contribución de esta tesis es la propuesta del estimador suavizado en la covariable y en la variable tiempo basado en el estimador de Beran para la función de supervivencia que permite estimar la curva de la PD reduciendo el error de estimación de forma notable. El análisis en profundidad de las propiedades asintóticas de este estimador, así como la propuesta de selectores automáticos para los parámetros de suavizado constituyen elementos importantes de esta disertación. Sin embargo, quedan importantes cuestiones abiertas en las que pretendemos continuar trabajando en el futuro.

Agradecimientos

Esta tesis se realizó bajo la inestimable supervisión de Juan Vilar y Ricardo Cao.

Rebeca Peláez Suárez fue beneficiaria de una Ayuda para contratos predoctorales para la formación de doctores 2018 y de una beca para estancias predoctorales del Programa inMOTION Inditex-UDC 2021. La investigación fue parcialmente financiada por los proyectos MTM2017-82724-R (MINECO, España), PID2020-113578RB-100 (MICINN, España), ED431C-2016-015 y ED431C-2020-14 (Grupos de Referencia Competitiva, Xunta de Galicia), ED431G/01 y ED431G 2019/01 (Centro Singular de Investigación, Xunta de Galicia).

Acerca de la autora



Rebeca Peláez Suárez es investigadora postdoctoral en el Grupo MODES (Universidade da Coruña). Graduada en Matemáticas por la Universidad de Oviedo (2017), obtuvo el Máster en Técnicas Estadísticas por la Universidade da Coruña (2019) por el que logró el Premio Extraordinario Fin de Máster. Posteriormente, se doctoró en Matemáticas por la Universidade da Coruña (2022) con Mención Internacional y Mención Cum Laude. Sus principales intereses son la estadística no paramétrica, los datos censurados y el riesgo financiero.

Referencias

- Beran, R. (1981). «Nonparametric regression with randomly censored survival data». En: *Technical report, University of California*.
- Cai, Z. (2003). «Weighted local linear approach to censored nonparametric regression». En: *Recent Advances and Trends in Nonparametric Statistics*. Ed. por Michael G. Akritas y Dimitris N. Politis. Elsevier, Amsterdam, págs. 217-231.
- López-Cheda, A., R. Cao y M. A. Jácome (2017). «Nonparametric latency estimation for mixture cure models». En: *TEST* 26.2, págs. 353-376.

- López-Cheda, A., R. Cao, M. A. Jácome e I. Van Keilegom (2017). «Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models». En: *Computational Statistics and Data Analysis* 105.12, págs. 144-165.
- Peláez, R., R. Cao y J. M. Vilar (2021a). «Nonparametric estimation of probability of default with double smoothing». En: *SORT* 45.2, págs. 93-120.
- (2021b). «Probability of default estimation in credit risk using a nonparametric approach». En: *TEST* 30.2, págs. 383-405.
- (2022a). «Bootstrap bandwidth selection and confidence regions for double smoothed default probability estimation». En: *Mathematics* 10.9, pág. 1523.
- (2022b). «Nonparametric estimation of the conditional survival function with double smoothing». En: *Journal of Nonparametric Statistics* 34.4, págs. 1063-1090.
- Van Keilegom, I. and M.G. Akritas (1999). «Transfer of tail information in censored regression models». En: *The Annals of Statistics* 27.5, págs. 1745-1784.