Suavizado con P-splines de datos censurados utilizando pesos de Kaplan-Meier

Tesis doctoral

Jorge Virto Moreno

Director:
Jesus María Orbe Lizundia

Programa de Doctorado Interuniversitario en Economía:
Instrumentos del Análisis Económico

Escuela de Doctorado de la Universidad del País Vasco

Departamento de Métodos Cuantitativos
Facultad de Economía y Empresa

Bilbao, 2 de abril de 2023

Suavizado con P-splines de datos censurados utilizando pesos de Kaplan-Meier

Tesis doctoral

Jorge Virto Moreno

Director:

Jesus María Orbe Lizundia

Programa de Doctorado Interuniversitario en Economía:
Instrumentos del Análisis Económico

Escuela de Doctorado de la Universidad del País Vasco

Departamento de Métodos Cuantitativos
Universidad del País Vasco / Euskal Herriko Unibertsitatea

Suavizado con P-splines de datos censurados utilizando pesos de Kaplan-Meier

Jorge Virto Moreno

Resumen

El objetivo de esta tesis es trabajar en el problema de ajuste de una curva no paramétrica en el contexto específico de datos censurados. Esta situación es muy habitual en el análisis de supervivencia o duración cuando se estudia la relación entre el tiempo de supervivencia, la variable de interés, y alguna covariable relevante. Frecuentemente, no se conoce la relación entre las dos variables y en lugar de asumir alguna relación paramétrica particular se asume solamente que es una función suave de los datos. Por lo tanto, se considera un enfoque no paramétrico para el ajuste de la curva que evita una especificación incorrecta de la forma funcional que conduciría a una estimación sesgada y a conclusiones equivocadas.

Cuando los datos disponibles están completos, es decir, sin datos censurados, el problema de la adaptación de las curvas no paramétricas ha sido ampliamente estudiado. Se han propuesto y analizado muchos métodos utilizando diferentes enfoques. Así, por ejemplo, existen métodos basados en los suavizados por núcleos (kernel smoothers) que obtienen la estimación en cada valor de la covariable como una función, usualmente un promedio ponderado de las observaciones locales de la variable de interés. Otro enfoque son los métodos basados en el suavizado con splines (spline smoothers). En la literatura del suavizado vía splines existen diferentes propuestas, pero se pueden distinguir dos enfoques principales: los splines de suavizado (smoothing splines) y los splines de regresión (regression splines). Además, existe un enfoque que combina lo mejor de estos dos campos, los splines con penalizaciones. Dentro de este enfoque una de las propuestas de más éxito son los splines penalizados (P-splines) de Eilers y Marx. Para el caso de datos censurados este enfoque parece muy interesante.

El objetivo de está investigación es proponer un nuevo método para adaptar los P-splines al caso de una muestra de datos censurados y analizar su funcionamiento mediante simulaciones en distintos contextos. Además, se usaran conjuntos de datos reales para ilustrar la metodología y comparar su rendimiento con alternativas parámetricas, como por ejemplo los modelos de duración acelerada (AFT) o algún modelo semi-paramétrico.

En primer lugar se propone una extensión del enfoque de splines penalizados utilizando ponderaciones de Kaplan-Meier para tener en cuenta el efecto de la censura y técnicas de validación cruzada generalizadas para elegir el parámetro de alisamiento

adaptado al caso de muestras censuradas. Además, se ha ampliado esta propuesta al marco de modelos aditivos generalizados (GAM), introduciendo una corrección del efecto de la censura, lo que permite estimar inmediatamente modelos más complejos. También se ha utilizado un conjunto de datos reales, la supervivencia de los pacientes en lista de espera del programa de trasplante de corazón de Stanford para ilustrar la metodología propuesta, que se muestra como una buena alternativa frente a otras especificaciones paramétricas o semi-paramétricas.

En segundo lugar se han desarrollado y analizado distintas alternativas para elegir el nivel óptimo de suavizado y la ubicación y el número del nodos para los estimadores propuestos: la extensión al caso de datos censurados del enfoque de P-splines y los GAM corregidos para tener en cuenta el efecto de la censura usando los pesos Kaplan-Meier. Para la elección del nivel óptimo de suavizado se han adaptado al caso censurado el criterio de validación cruzada generalizado y la modificación propuesta por Kim y Gu para evitar el habitual sobre-ajuste al utilizar el criterio de validación cruzada generalizado. Con respecto a la elección del número de nodos se ha propuesto una pequeña variación de la propuesta de Ruppert para tener en cuenta la censura. Por último, en cuanto a la elección de la ubicación de los nodos, además del caso de nodos equidistantes, se ha propuesto y analizado una adaptación al caso de datos censurados que usa vectores de nodos no uniformes con una ubicación de los mismos en función de los pesos de Kaplan-Meier. Se ha estudiado el rendimiento de las distintas propuestas mediante un amplio estudio de simulación que considera relaciones funcionales de diversos grados de complejidad entre la variable de respuesta censurada y un regresor en situaciones con diferencias en la información disponible, donde se combinan distintos tamaños de muestra y niveles de censura.

En tercer y último lugar, se ha extendido la metodología anterior al caso en que existe más de una variable explicativa, obteniendo avances significativos en diferentes aspectos. El trabajo presenta fundamentalmente dos importantes aportaciones con respecto a la primera propuesta. Por una parte, extiende la metodología a un contexto más general, permitiendo su utilización en problemas más complejos, más habituales en la práctica. Así, se propone un método de estimación que permite estimar modelos donde la forma funcional del efecto de algunas variables explicativas sobre la variable a explicar es conocida y por tanto, puede incorporarse al modelo de forma paramétrica (componente paramétrica) y variables explicativas donde la forma funcional del efecto sobre la variable de interés es desconocida (componente no paramétrica), lo que se conoce como un modelo semi-paramerico para datos censurados. En este contexto se han propuesto estimadores tanto para la parte paramétrica como para la parte no paramétrica. Por otra parte, se propone un estimador de las varianzas para ambas componentes, paramétrica y no paramétrica, y se proporcionan las herramientas necesarias para poder realizar inferencia en este tipo de modelos.

Las propuestas son analizadas en tres casos distintos y bajo diferentes escenarios de censura y tamaño muestral. Las principales conclusiones son que la nueva metodología estima de forma muy satisfactoria tanto la componente paramétrica como la no paramétrica. Además, la precisión de los estimadores mejora con el tamaño muestral y a medida que se reduce el nivel de censura en la muestra. Finalmente, se muestra que las probabilidades de cubrimiento de los intervalos de confianza propuestos, tanto para la componente paramétrica como para la no paramétrica,

coinciden o se aproximan a las probabilidades de cubrimiento nominales en cada uno de los escenarios analizados.

Por último, la aplicación presentada con datos reales sirve para ilustrar las ventajas que presenta nuestra propuesta frente a algunas de las alternativas existentes en la literatura para estimar modelos semiparamétricos de regresión censurados. Se puede apreciar que la metodología propuesta resulta una metodología flexible y robusta que no necesita realizar ciertos supuestos en la modelización del problema, que podrían resultar demasiado restrictivos y además difíciles de verificar o comprobar en la práctica. Por tanto, resulta muy útil para su aplicación práctica en una gran variedad de contextos donde la variable a explicar presenta problemas de censura.

Acerca del formato de este trabajo

Este trabajo no sigue la estructura tradicional de las tesis doctorales, sino que se acoge al formato *tesis por compendio de publicaciones*, ampliamente difundido en la Unión Europea y que ha sido aceptado recientemente por la Universidad del País Vasco/Euskal Herriko Unibertsitatea. En concreto, el capítulo IX de la Normativa de Gestión de las enseñanzas de doctorado recoge la reglamentación de las tesis por compendio de publicaciones (BOPV de 5 de marzo de 2020) y en su artículo 42 establece sus requisitos:

> *Podrán optar por la presentación de la tesis en esta modalidad aquellos doctorandos o doctorandas que tengan publicados o aceptados para su publicación al menos tres contribuciones realizadas durante el periodo de permanencia del doctorando o doctoranda en el programa de doctorado . . . En cualquier caso, todas las contribuciones deberán corresponder a la línea de investigación de su Plan de Investigación de doctorado.* Las contribuciones deberán ser artículos en revistas científicas que aparezcan en la última relación publicada por el Journal Citation Reports (SCI y/o SSCI) o SCOPUS, . . . Al menos una de ellas deberá estar en el primer o segundo cuartil de su categoría . . .
> *En los casos de coautoría y cuando el orden de los firmantes no sea el alfabético, el doctorando o doctoranda deberá ser el o la primera o segunda firmante de la contribución. El director de la tesis acreditará que la aportación del doctorando o doctoranda es relevante y se ajusta a su Plan de Investigación . . .*

Esta tesis se ha estructurado siguiendo las directrices del artículo 43 del capítulo IX de la Normativa de Gestión de las enseñanzas de doctorado que establece que las tesis doctorales presentadas por compendio de publicaciones deben incluir los elementos que se enumeran a continuación:

- Una sección inicial de síntesis, con una extensión mínima orientativa de 10.000 palabras, que contenga:
  - Introducción, en la que se realice una presentación de la tesis y se justifique la unidad temática.
  - Marco teórico en el que se inscribe el tema de la tesis y herramientas metodológicas utilizadas.
  - Hipótesis y objetivos generales y específicos a alcanzar, indicando en qué publicación o publicaciones se abordan.
  - Resumen y, en su caso, discusión de los resultados obtenidos.
  - Fuentes referenciadas.

- La segunda sección de la tesis estará formada por las conclusiones de la misma.
- La tercera sección corresponderá al Anexo, que debe contener los artículos, libros o capítulos de libro publicados o aceptados, bajo el título de "Trabajos publicados" o, si fuera el caso, "Trabajos Publicados o aceptados". Se incluirá la versión íntegra publicada o aceptada de cada contribución, ... se incluirán las referencias bibliográficas completas, y se indicará el factor de impacto de la revista en el año de la publicación, su posición relativa en la categoría a la que pertenece, y/u otros indicios de calidad.

Índice general

I Sección inicial de síntesis **1**

1. Introducción y marco teórico **3**
1.1. Caso univariante . . . 5
1.2. Selección de parámetros . . . 6
1.3. Modelo semiparamétrico . . . 7

2. Metodología propuesta **11**
2.1. Caso univariante . . . 11
2.1.1. P-splines . . . 11
2.1.2. P-splines censurados . . . 12
2.2. Selección de parámetros . . . 16
2.2.1. Parámetro de suavizado . . . 16
2.2.2. Nodos . . . 19
2.3. Modelo semiparamétrico . . . 20
2.3.1. Método de estimación . . . 21
2.3.2. Algoritmo . . . 21
2.3.3. Estimadores de las varianzas . . . 22
2.3.4. Código . . . 23

3. Resumen de resultados **25**
3.1. Resultados caso univariante . . . 25
3.1.1. Estudio de simulación . . . 25
3.1.2. Extensión a los modelos GAM . . . 29
3.1.3. Aplicación empírica: trasplantes cardíacos de Stanford . . . 31
3.2. Resultados selección de parámetros . . . 33
3.2.1. Estimaciones con la mejor elección de parámetros para los estimadores ckmPS y ckmGAM . . . 42
3.2.2. MECM según las diferentes elecciones de parámetros para las distintas funciones analizadas . . . 53
3.3. Resultados modelo semiparamétrico . . . 59
3.3.1. Estudio de simulación . . . 59
3.3.2. Estimaciones de la parte no paramétrica utilizando diferentes niveles de censura y tamaños de muestra . . . 66
3.3.3. Comparación de resultados según la forma de generar la variable censura: Uniforme versus Normal . . . 69
3.3.4. Comparación de resultados según la distribución del error: Normal versus Weibull . . . 72
3.3.5. Aplicación empírica: datos CBP . . . 75

4. Código en R **79**
4.1. Código para el caso univariante 79
4.2. Código para la selección de parámetros 85
4.3. Código para el modelo semiparamétrico 92

5. Bibliografía **99**

II Conclusiones **105**

6. Conclusiones **107**
6.1. Conclusiones caso univariante 107
6.2. Conclusiones selección de parámetros 108
6.3. Conclusiones modelo semiparamétrico 109

III Trabajos Publicados **111**

7. Penalized spline smoothing using Kaplan–Meier weights with censored data **115**

8. Selecting the smoothing parameter and knots for an extension of penalized splines to censored data **131**

9. Penalized spline smoothing using Kaplan-Meier weights in semiparametric censored regression models **165**

Índice de figuras

2.1. Estimación P-splines con datos no censurados y censurados 13
2.2. Propuesta de estimación P-splines censurada 15
2.3. Estimación con dos valores diferentes de λ 17
2.4. Criterios para la elección del parámetro λ, GCV_1 versus GCV_2 18
2.5. ECM para el parámetro λ en la red 0,000001 a 4 y para λ_{GCV_1} y λ_{GCV_2} 19

3.1. Errores Cuadráticos Medios del método P-splines censurado utilizando diferentes niveles de censura y tamaños de muestra (caso 1) 26
3.2. Función estimada utilizando el estimador P-splines censurado (caso 1) 26
3.3. Errores Cuadráticos Medios del método P-splines censurado utilizando diferentes niveles de censura y tamaños de muestra (caso 2) 28
3.4. Función estimada utilizando el estimador P-splines censurado (caso 2) 28
3.5. Media de las funciones estimadas: P-splines censurados y función *gam* 30
3.6. Extensión a modelos GAM: media de las funciones estimadas 31
3.7. Relación estimada mediante tres metodologías: AFT lognormal, enfoque de Stute y P-splines censurado 33
3.8. Función estimada utilizando el estimador ckmPS para la función cuadrática: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 43
3.9. Función estimada utilizando el estimador ckmGAM para la función cuadrática: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 44
3.10. Función estimada utilizando el estimador ckmPS para la función bump: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 45
3.11. Función estimada utilizando el estimador ckmGAM para la función bump: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 46
3.12. Función estimada utilizando el estimador ckmPS para la función logística: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 47
3.13. Función estimada utilizando el estimador ckmGAM para la función logística: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 48
3.14. Función estimada utilizando el estimador ckmPS para la función sinusoidal con dos ciclos: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 49
3.15. Función estimada utilizando el estimador ckmGAM para la función sinusoidal con dos ciclos: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 50
3.16. Función estimada utilizando el estimador ckmPS para la función sinusoidal con tres ciclos: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 51
3.17. Función estimada utilizando el estimador ckmGAM para la función sinusoidal con tres ciclos: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes 52

3.18. MECM según las diferentes elecciones de parámetros para la función cuadrática 54
3.19. MECM según las diferentes elecciones de parámetros para la función bump 55
3.20. MECM según las diferentes elecciones de parámetros para la función logística 56
3.21. MECM según las diferentes elecciones de parámetros para la función sinusoidal con dos ciclos 57
3.22. MECM según las diferentes elecciones de parámetros para la función sinusoidal con tres ciclos 58
3.23. Resultados del estudio de simulación para la función cuadrática: Errores Cuadráticos Medios para la parte no paramétrica , $\widehat{\alpha}_1$ y $\widehat{\alpha}_2$ utilizando diferentes niveles de censura y tamaños de muestra 63
3.24. Resultados del estudio de simulación para la función sinusoidal: Errores Cuadráticos Medios para la parte no paramétrica , $\widehat{\alpha}_1$ y $\widehat{\alpha}_2$ utilizando diferentes niveles de censura y tamaños de muestra 64
3.25. Resultados del estudio de simulación para la función logística: Errores Cuadráticos Medios para la parte no paramétrica , $\widehat{\alpha}_1$ y $\widehat{\alpha}_2$ utilizando diferentes niveles de censura y tamaños de muestra 65
3.26. Estimación de la parte no paramétrica utilizando diferentes niveles de censura y tamaños de muestra para la función cuadrática 66
3.27. Estimación de la parte no paramétrica utilizando diferentes niveles de censura y tamaños de muestra para la función sinusoidal 67
3.28. Estimación de la parte no paramétrica utilizando diferentes niveles de censura y tamaños de muestra para la función logística 68
3.29. Estimación componente no paramétrica utilizando tres metodologías: AFT lognormal, enfoque de Stute y estimador CPS 76

Índice de tablas

3.1. Estimación de los coeficientes de regresión y sus desviaciones típicas utilizando los métodos AFT lognormal y de Stute 32
3.2. Ejemplos de forma funcional 34
3.3. Nueve escenarios: tres tamaños de muestra por tres niveles de censura 35
3.4. Resultados del estudio de simulación para la función cuadrática . . . 36
3.5. Resultados del estudio de simulación para la función bump 37
3.6. Resultados del estudio de simulación para la función logística 39
3.7. Resultados del estudio de simulación para la función sinusoidal con dos ciclos 40
3.8. Resultados del estudio de simulación para la función sinusoidal con tres ciclos 41
3.9. Tres casos de estudio 59
3.10. Modelo semiparamétrico: resultados para la función cuadrática 62
3.11. Modelo semiparamétrico: resultados para la función sinusoidal 62
3.12. Modelo semiparamétrico: resultados para la función logística 62
3.13. Resultados del estudio de simulación de la función cuadrática: censura Uniforme versus Normal 69
3.14. Resultados del estudio de simulación de la función sinusoidal: censura Uniforme versus Normal 70
3.15. Resultados del estudio de simulación de la función logística: censura Uniforme versus Normal 71
3.16. Resultados del estudio de simulación de la función cuadrática: error Normal versus Weibull 72
3.17. Resultados del estudio de simulación de la función sinusoidal: error Normal versus Weibull 73
3.18. Resultados del estudio de simulación de la función logística: error Normal versus Weibull 74
3.19. Estimación de los coeficientes de regresión y sus desviaciones típicas para el conjunto de datos de Cirrosis Biliar Primaria de la Clínica Mayo a partir de los métodos AFT, Stute y CPS 75

Índice de funciones de R

4.1. *kmw.cp*: función para calcular los pesos Kaplan-Meier 80
4.2. *pswc*: función para el cálculo del estimador P-splines censurado 81
4.3. *gamkm*: función para el cálculo del estimador en un modelo GAM censurado con un regresor 83
4.4. *gamkm2d2l*: función para el cálculo del estimador en un modelo GAM censurado con dos regresores 84
4.5. *nodos.km*: función para calcular la ubicación de los nodos usando los pesos Kaplan-Meier 86
4.6. *pswc* modificada: función para el cálculo del estimador ckmPS 87
4.7. *gamkm* modificada: función para el cálculo del estimador ckmGAM con un regresor 90
4.8. *semipswc*: función para el cálculo del estimador P-splines censurado en un modelo semiparamétrico 93

Parte I

Sección inicial de síntesis

Capítulo 1

Introducción y marco teórico

En este trabajo se considera el problema del ajuste no paramétrico de curvas en el contexto específico de los datos censurados, es decir, cuando la muestra no se observa completamente porque algunos de los valores de los datos están censurados. Así, para algunos individuos en lugar del valor real de la variable de interés lo que se observa es un valor mínimo, y se sabe que el valor real es mayor que este valor mínimo. Esto se conoce como una muestra de datos censurada a la derecha. Esta situación es muy común en los análisis de supervivencia y duración cuando se analiza la relación entre el tiempo de supervivencia, la variable de interés, y un grupo de covariables.

Con frecuencia, la función que relaciona la variable de interés con las covariables no se conoce, y en lugar de asumir una relación paramétrica particular, como por ejemplo la especificación de regresión lineal habitual, se asume únicamente que es una función suave de los datos. Por lo tanto, consideramos un enfoque de ajuste de curvas no paramétrico que evite una especificación incorrecta de la forma funcional que conduciría a una estimación sesgada y a conclusiones erróneas.

El problema del ajuste no paramétrico de curvas cuando los datos disponibles son completos, es decir, no hay datos censurados, ha sido ampliamente analizado y existen numerosos estudios en esta área. Hay métodos basados en suavizadores de kernel (Silverman, 1986; Härdle, 1990) que obtienen la estimación de la variable de interés en cada valor del regresor como una función, normalmente una media ponderada, de los valores de las observaciones locales. Otro enfoque es el de los métodos basados en suavizadores spline (Eubank, 1988; Wahba, 1990; Green and Silverman, 1994; Wood, 2017). Los splines son trozos de funciones polinómicas encajadas en puntos conocidos como nodos, donde se fijan ciertas condiciones o restricciones relativas a la continuidad de la función y de algunas de sus derivadas. Nuestra propuesta, para el caso de una muestra de datos censurada, entra dentro del enfoque de los splines.

Los splines dependen del grado del polinomio y del número y ubicación de los nodos. La elección de estos elementos ha sido ampliamente estudiada (por ejemplo, Friedman and Silverman, 1989; Ruppert, 2002). En la literatura sobre splines suavizadores se pueden encontrar varias propuestas, pero se pueden distinguir dos enfoques principales: splines de suavización y splines de regresión.

Los splines de suavizado podrían presentarse como la solución a la introducción de la penalización por rugosidad para la estimación de curvas (ver Green and Silverman, 1994). La solución del problema de minimización es un spline cúbico natural con tantos nodos como valores diferentes tenga de la variable X (ver Reinsch, 1967;

Green and Silverman, 1994). Las muestras de datos grandes requieren un gran número de parámetros, ya que es necesario que haya tantos nodos como valores diferentes tenga la variable X, lo que puede dar lugar a problemas computacionales.

Una posible solución a este problema es reducir el problema de la dimensionalidad mediante el uso de un conjunto de q funciones de base: $B_1, \ldots, B_q$, de modo que la función objetivo a estimar, $f(\cdot)$, pueda ser reescrita como: $f(x) = \sum_{j=1}^{q} \gamma_j B_j(x)$ donde γ_j es el coeficiente asociado a la función base j-th. En la literatura sobre splines de regresión se pueden encontrar varias formas alternativas de calcular estas bases. Esta propuesta reduce el problema de la dimensionalidad pero genera uno nuevo: la elección del número y ubicación de los nodos. La idea de suavizar con splines penalizados para evitar el problema de la selección de nodos se remonta a O'Sullivan (1986, 1988), pero fueron Eilers and Marx (1996) quienes la simplificaron y generalizaron introduciendo la combinación de B-splines y una penalización basada en diferencias entre coeficientes adyacentes. Esta propuesta se conoce como el enfoque de splines penalizados, P-splines (para más detalles véase la referencia Eilers et al., 2015). Los P-splines pueden considerarse una generalización del suavizado spline con una elección más flexible de bases, penalizaciones y nodos. También pueden considerarse como los splines de regresión pero con una penalización sobre los coeficientes de los splines.

Las bases tratadas hasta ahora son muy útiles para representar el suavizado con una variable predictora. Para suavizar para más variables, los suavizadores multidimensionales parecen más adecuados, por ejemplo los splines de placa fina (Duchon, 1977; Wood, 2003) o los suavizadores de producto tensorial (De Boor, 2001; Currie et al., 2004; Wood, 2006).

Esta tesis se enmarca en el enfoque P-splines de Eilers and Marx (1996) en el contexto específico de los datos censurados. Como ya se ha mencionado, este tipo de datos es muy común en el análisis de supervivencia y duración. Se han propuesto varios métodos en este ámbito. La mayoría de ellos pueden clasificarse en dos clases principales: los que modelizan la función de riesgo, *hazard regression models*, y los modelos de regresión de duración acelerada, *accelerated failure time (AFT) regression models*.

Bajo la primera de estas propuestas los investigadores estudian el efecto de los distintos regresores en una probabilidad condicional (*i.e.*, la función de riesgo); el enfoque más popular es el modelo de regresión de función de riesgo proporcional de Cox (Cox, 1972). El uso de P-splines en modelos de esta clase no es nuevo, con referencias tan tempranas como Cai et al. (2002) donde se considera un enfoque de modelo mixto para suavizar la función de riesgo en un modelo de tipo Cox. Kauermann (2005) utiliza P-splines para ajustar modelos de riesgo no proporcionales y Kauermann and Khomski (2006) amplían esta metolología para incluir un efecto calendario no paramétrico. Hennerfeind et al. (2006) proponen un modelo de supervivencia geo-aditivo con suavizado mediante P-splines bayesianos. Kneib and Fahrmeir (2007) extienden estas propuestas al considerar una estimación máximo-verosimil penalizada en el contexto de un modelo mixto que incorpora un término no paramétrico en la función de riesgo, coeficientes que varían en el tiempo y efectos no lineales de las covariables continuas, un componente espacial y tiene en cuenta la posible heterogeneidad de distintos grupos. Los enfoques anteriores tratan únicamente con observaciones censuradas por la derecha. Kneib (2006) extiende estos modelos para su uso con datos censurados por intervalo. Otras contribuciones a la

modelización de datos censurados utilizando P-splines se pueden encontrar en la revisión de Ruppert et al. (2009).

Bajo el enfoque de los modelos de regresión de duración acelerada se estudia el efecto directo que la covariable tiene sobre la variable de respuesta o alguna transformación de la misma, de forma similar a la que se utiliza en los modelos de regresión clásicos (véase, por ejemplo, Kalbfleisch and Prentice, 2002). A diferencia del anterior enfoque, no existe un gran uso de P-splines en este tipo de modelos. Komárek et al. (2005) proponen una metodo de estimación de máxima verosimilitud para un AFT utilizando P-splines para suavizar la densidad de la distribución de error. Lambert (2013) considera un modelo aditivo no paramétrico para la localización y dispersión utilizando P-splines. Más recientemente, Konrath et al. (2015) utilizando P-splines bayesianos introducen una extensión de la regresión AFT que permite el modelado estadístico conjunto de efectos lineales, efectos no lineales suaves y una estructura de error flexible.

1.1. Caso univariante

La presente tesis se sitúa en el contexto de los modelos AFT, pero utiliza un enfoque diferente. En un primer momento, véase apartados 2.1 y 3.1, se analiza el caso más sencillo, cuando interesa analizar la relación entre una variable de interés censurada por la derecha, denotada aquí como T, y una única covariable relevante X:

$$t_i = f(x_i) + \epsilon_i \qquad i = 1, 2, \ldots, n$$

donde ϵ_i es el término de error, n es el tamaño de la muestra y $f(\cdot)$ es una función suave de los datos.

En este contexto, siguiendo la idea de Stute (1993), se propone una extensión del enfoque de los P-splines de Eilers and Marx (1996) utilizando los pesos de Kaplan-Meier para tener en cuenta el efecto de la censura y las técnicas generalizadas de validación cruzada para elegir el parámetro de suavizado adaptado para el caso de muestras censuradas. El método propuesto, además de no asumir una forma funcional para la relación entre la variable de respuesta censurada y el regresor, no necesita imponer una distribución de probabilidad específica para la variable de respuesta, que suele ser desconocida en la práctica. Además, el método es muy fácil de aplicar y reduce considerablemente la dimensionalidad del problema en comparación con el enfoque de splines de suavizado. Por último, centrándonos en los modelos de supervivencia, cabe señalar que se modeliza directamente el efecto de la covariable sobre la variable de duración, lo que facilita la interpretación de los resultados.

Utilizando diversos estudios de simulación se ha analizado la eficacia del método propuesto obteniendo un rendimiento bastante satisfactorio. Además, se ha ampliado esta propuesta al marco de modelos aditivos generalizados (GAM), introduciendo una corrección del efecto de la censura, lo que permite estimar inmediatamente modelos más complejos. También se ha utilizado un conjunto de datos reales, la supervivencia de los pacientes en lista de espera del programa de trasplante de corazón de Stanford para ilustrar la metodología propuesta, que se muestra como una buena alternativa cuando la distribución de probabilidad para la variable de respuesta y la forma funcional no se conocen en los modelos de regresión censurada.

1.2. Selección de parámetros

Como sucede también en problemas con datos no censurados, esta metodología necesita, por una parte, elegir un parámetro de suavizado y, por otra parte, elegir el número y la localización de los nodos, que no son fijos como en los splines de suavizado. En el apartado 2.2 se proponen métodos de selección óptimos para los parámetros anteriores, parámetro de suavizado, número de nodos y localización de los mismos, adecuados para su aplicación en contextos de datos censurados. En la literatura este tema ha sido estudiado por Aydin and Yilmaz (2018) pero ellos consideran las versiones transformadas de las observaciones censuradas, llamadas datos sintéticos, propuestas por Koul et al. (1981); ver Lai and Ying (1992); Zhou (1992) para más detalles. En el enfoque presentado en este trabajo no se necesita generar una variable sintética, en su lugar vamos a utilizar la variable respuesta original, censurada, que se observa en la muestra (sin modificarla) y se utilizan los pesos Kaplan-Meier para controlar el efecto de la censura. Además de tener en cuenta el efecto de la censura mediante un enfoque diferente, cabe destacar que existen dos diferencias adicionales respecto al trabajo de Aydin and Yilmaz (2018), ya que en el criterio de estimación se utilizan bases y términos de penalización diferentes.

La elección del nivel óptimo de suavizado es el factor más importante de cara al ajuste. Para el caso no censurado se suelen utilizar el criterio de información de Akaike, el criterio de validación cruzada generalizado (GCV) o la modificación propuesta por Kim and Gu (2004) para evitar el habitual sobre-ajuste al utilizar el criterio GCV. Pero la aplicación de estos criterios sin adaptarlos al caso de datos censurados lleva a elegir parámetros de suavizado que generaran importantes sesgos en la estimación. En el apartado 2.2.1 se proponen y comparan varias adaptaciones al caso censurado del criterio de validación cruzada generalizado y de la modificación propuesta por Kim and Gu (2004). Dado que hay un término de penalización que controla la suavidad de la función, la elección del número de nodos no es una cuestión crucial, siempre y cuando se elija un número de nodos que sea lo suficientemente grande como para ajustarse a los datos. Es común en la literatura de datos no censurados seleccionar el número de nodos aplicando la fórmula por defecto presentada en Ruppert (2002). En el apartado 2.2.2 se ha propuesto una pequeña variación de esta propuesta para tener en cuenta la perdida de información en la muestra debido a la existencia de datos censurados y se compara con la forma de selección en el caso no censurado. Por último, en cuanto a la elección de la ubicación de los nodos, cuestión no tratada anteriormente en la literatura para el caso censurado. además del caso de nodos equidistantes habitualmente utilizado en muestras sin censura, se ha propuesto y analizado una adaptación al caso de datos censurados que usa vectores de nodos no uniformes con una ubicación de los mismos en función de los pesos de Kaplan-Meier. En resumen, se ha desarrollado una propuesta metodológica para la elección del nivel óptimo de suavizado, el número y la posición de los nodos para los estimadores propuestos.

Todas estas propuestas son analizadas, en el apartado 3.2, en un exhaustivo y completo estudio de simulación que considera cinco relaciones funcionales de distinta complejidad para nueve escenarios que presentan diferentes niveles de censura y tamaños muestrales. Es decir, tenemos un total de 45 escenarios donde, además de analizar el comportamiento de las distintas propuestas de selección, se analiza la importancia relativa que tiene cada tipo de elección. Las conclusiones más im-

portantes que se derivan del análisis indican que los métodos de selección utilizados en el contexto no censurado no son de aplicación válida en un contexto de datos censurados. Por tanto, las propuestas presentadas para el caso censurado resultan un avance importante y son muy necesarias. La elección del parámetro de suavizado resulta ser la más importante y la propuesta que se hace para la selección del parámetro de suavizado es fundamental para una buena estimación. La propuesta de selección del número de nodos tiene una importancia relativa menor, pero mejora siempre a la forma de selección habitualmente utilizada en el caso no censurado. En cuanto a la propuesta de localización de los nodos no igualmente espaciados puede mejorar los resultados en contextos de tamaño de muestra pequeño y niveles de censura elevados. Además, hay que recalcar que la mejora que se produce con las propuestas de selección de parámetros es creciente a medida que aumentamos el nivel de censura en los datos, por tanto, una elección correcta de los parámetros es aún más importante en contextos de censura elevada.

1.3. Modelo semiparamétrico

La extensión del método P-splines al caso de respuestas censuradas utilizando ponderaciones Kaplan-Meier que se ha propuesto en los apartados 2.1 y 3.1 considera el caso de una covariable única y, además, no proporciona herramientas que permitan realizar inferencias estadísticas. Por tanto, tiene una utilidad limitada en la práctica, donde la variable de respuesta suele depender de un amplio conjunto de variables explicativas y resulta interesante realizar inferencias. En lo que sigue se trata de extender la metodología anterior a un contexto más general, permitiendo su utilización en problemas más complejos, más habituales en la práctica, proponiendo un método de estimación que permita estimar con datos censurados modelos donde la forma funcional del efecto de algunas variables explicativas sobre la variable a explicar es conocida y por tanto, puede incorporarse al modelo de forma paramétrica (componente paramétrica) y, además, incorporan un componente no paramétrico para modelar efectos en los que no se conoce la relación funcional, *i.e.* un modelo de regresión semiparamétrico para datos censurados. Dicha extensión es un problema bien estudiado para el caso de datos no censurados (véase, por ejemplo, Heckman, 1986; Schimek, 2000; Holland, 2017).

En la literatura de datos censurados se pueden encontrar varios trabajos que permiten estimar un modelo de regresión sin necesidad de elegir una distribución de probabilidad específica. Estos estudios consideran varios enfoques de mínimos cuadrados, e incluyen los trabajos de Koul et al. (1981) y Leurgans (1987), que proponen transformar la variable censurada, y los de Miller (1976), Buckley and James (1979) y Stute (1993), que presentan propuestas con un enfoque similar pero sin transformar la variable a explicar. También existe el enfoque de métodos de estimación basados en rangos, *rank-based estimation methods*, (véase por ejemplo Tsiatis, 1990; Lai and Ying, 1992; Jin et al., 2003).

Estas propuestas suponen un avance considerable en la especificación del modelo, evitando los sesgos derivados de elecciones erróneas de la distribución de probabilidad. Pero es posible ir aún más lejos en la flexibilización de estas metodologías, ya que todas estas propuestas consideran una relación paramétrica conocida para especificar el efecto de las variables explicativas sobre la variable a explicar. En la práctica, es bastante frecuente que no se conozca la relación funcional entre las va-

riables regresoras y el resultado. Una forma de evitar errores que pueden llevar a conclusiones sesgadas en la especificación de estos efectos es no imponer una relación funcional paramétrica específica entre la variable a explicar y la variable explicativa, sino suponer únicamente que esa relación es una función suave, *i. e.* considerar una estimación no paramétrica de ese efecto específico. Como ya se ha comentado, la estimación de relaciones funcionales no paramétricas con datos no censurados ha sido ampliamente estudiada y se han presentado diversas propuestas en la literatura. Se pueden agrupar en dos enfoques diferentes: métodos basados en kernel smoothers (Silverman, 1986; Härdle, 1990) y métodos basados en spline smoothers (Eubank, 1988; Wahba, 1990; Green and Silverman, 1994; Eilers and Marx, 1996; Wood, 2017).

La aplicación de estas técnicas de estimación no paramétrica no es sencilla en el caso de datos censurados, por lo que los estudios anteriores deben adaptarse para tener en cuenta el efecto de la censura en el proceso de estimación. La presente propuesta se enmarca dentro del enfoque de los spline smoothers en el contexto específico de los modelos de regresión semiparamétrica con datos censurados. Este modelo de regresión semiparamétrica ya ha sido estudiado y discutido en relación con muestras sin observaciones censuradas. Fue analizado inicialmente por Heckman (1986) y Rice (1986) utilizando un enfoque con suavizadores spline y por Speckman (1988) utilizando un enfoque con suavizadores kernel. Varios autores han investigado la inferencia en el modelo de regresión semiparamétrica cuando la variable de respuesta está sujeta a censura a la derecha. Orbe et al. (2003) utilizan un enfoque basado en splines de suavizado mientras que Zou et al. (2011) y Chen et al. (2015) utilizan splines penalizados y B-splines monótonos, respectivamente. Aydin and Yilmaz (2018) aplican las ideas propuestas por Koul et al. (1981) en el contexto de un modelo de regresión lineal parcial. De Uña Álvarez and Roca Pardiñas (2009) consideran el uso de kernel smoothers en un modelo de regresión aditiva censurada.

Es en este contexto donde se encuentra la metodología de estimación, tanto para la parte paramétrica como para la parte no paramétrica, propuesta en los apartados 2.3 y 3.3. El método propuesto, como en el caso univariante, además de no asumir una forma funcional para la relación entre la variable de respuesta censurada y algunos de los regresores, no necesita imponer una distribución de probabilidad específica para la variable de respuesta, que suele ser desconocida en la práctica. También se han propuesto distintas alternativas para estimar las varianzas para ambas componentes, paramétrica y no paramétrica, y se proporcionan las herramientas necesarias para desarrollar inferencias estadísticas en este marco general. Estas propuestas son analizadas en tres casos distintos y bajo diferentes escenarios de censura y tamaño muestral. Las principales conclusiones son que la nueva metodología estima de forma muy satisfactoria tanto la componente paramétrica como la no paramétrica. Además, la precisión de los estimadores mejora con el tamaño muestral y a medida que se reduce el nivel de censura en la muestra. Finalmente, se muestra que las probabilidades de cubrimiento de los intervalos de confianza propuestos, tanto para la componente paramétrica como para la no paramétrica, coinciden o se aproximan a las probabilidades de cubrimiento nominales en cada uno de los escenarios analizados.

Por último, la aplicación presentada con datos reales sirve para ilustrar las ventajas que presenta esta propuesta frente a algunas de las alternativas existentes en la literatura para estimar modelos semiparamétricos de regresión censurados. Se puede apreciar que la metodología propuesta resulta una metodología flexible y ro-

busta que no necesita realizar ciertos supuestos en la modelización del problema, que podrían resultar demasiado restrictivos y además difíciles de verificar o comprobar en la práctica. Por tanto, resulta muy útil para su aplicación práctica en una gran variedad de contextos donde la variable a explicar presenta problemas de censura.

Capítulo 2

Metodología propuesta

2.1. Caso univariante

En este apartado se propone una extensión del enfoque de los P-splines de Eilers and Marx (1996) utilizando los pesos de Kaplan-Meier para tener en cuenta el efecto de la censura siguiendo la idea de Stute (1993). Como paso previo se procede a describir brevemente el método de splines penalizados (P-splines) presentado en Eilers and Marx (1996) para el caso de datos no censurados.

2.1.1. P-splines

Sea una muestra de observaciones (t_i, x_i) para $i = 1, \ldots, n$ y sea la siguiente regresión simple no paramétrica:

$$T = f(X) + \epsilon$$

donde ϵ es un término de error que satisface $E(\epsilon|X) = 0$. Es decir, no se asume ninguna forma funcional de $f(\cdot)$.

Una forma de estimar $f(\cdot)$ de forma flexible es utilizar como base para la regresión las bases de B-splines (De Boor, 2001; Dierckx, 1993). Así, se utiliza un conjunto de q funciones base B-splines de grado d, $B_1(x), \ldots, B_q(x)$, y aplicando mínimos cuadrados, es posible estimar la función $f(\cdot)$, la curva ajustada, como $\hat{f}(x) = \sum_{j=1}^{q} \hat{\gamma}_j B_j(x)$ minimizando la siguiente expresión:

$$\sum_{i=1}^{n} \left[t_i - \sum_{j=1}^{q} \gamma_j B_j(x_i) \right]^2 \tag{2.1}$$

donde $B_j(x_i)$ denota el valor para x_i del B_j B-spline para una rejilla de nodos equidistantes. Eilers and Marx (1996) proporcionan una breve pero interesante revisión de las B-splines, describiendo sus características y propiedades generales y explican cómo calcularlas.

Con objeto de simplificar el problema de elegir el número y la posición de los nodos, O'Sullivan (1986, 1988) añade un término de penalización en la segunda derivada de la curva. Así, la expresión a minimizar se define como

$$\sum_{i=1}^{n} \left[t_i - \sum_{j=1}^{q} \gamma_j B_j(x_i) \right]^2 + \lambda \int_{x_{min}}^{x_{max}} \left[\sum_{j=1}^{q} \gamma_j B_j''(x_i) \right]^2 dx$$

Eilers and Marx (1996) utilizan una aproximación diferente, el estimador P-spline, que también es una solución basada en la estimación B-spline habitual junto con el enfoque de los smoothing spline.

El estimador P-splines simplifica y generaliza la propuesta de O'Sullivan introduciendo una penalización basada en la diferencias de orden k entre los coeficientes adyacentes de la bases de B-splines γ_j. De esta forma, se tiene en cuenta un nivel general de suavidad de la función estimada al imponer que los coeficientes B-splines adyacentes deban ser similares. Por tanto, la correspondiente expresión de mínimos cuadrados penalizados es:

$$\sum_{i=1}^{n}\left[t_i - \sum_{j=1}^{q}\gamma_j B_j(x_i)\right]^2 + \lambda \sum_{j=k+1}^{q}(\Delta^k\gamma_j)^2 \tag{2.2}$$

donde $\Delta\gamma_j$ denota la diferencia entre los coeficientes $(\gamma_j - \gamma_{j-1})$ y $\Delta^k\gamma_j$ es la diferencia de orden k. Cuanto más parecidos sean los coeficientes, más suave es $f(\cdot)$. Eilers and Marx (1996) muestran que esta penalización es una buena aproximación de la integral del cuadrado de la k-ésima derivada de la función. Además, este método reduce considerablemente la dimensionalidad del problema al pasar del número de valores diferentes de la variable X, con splines de suavizado, al número de B-splines. Adicionalmente, este tipo de penalización es más flexible, ya que es independiente del grado del polinomio utilizado para construir los B-splines.

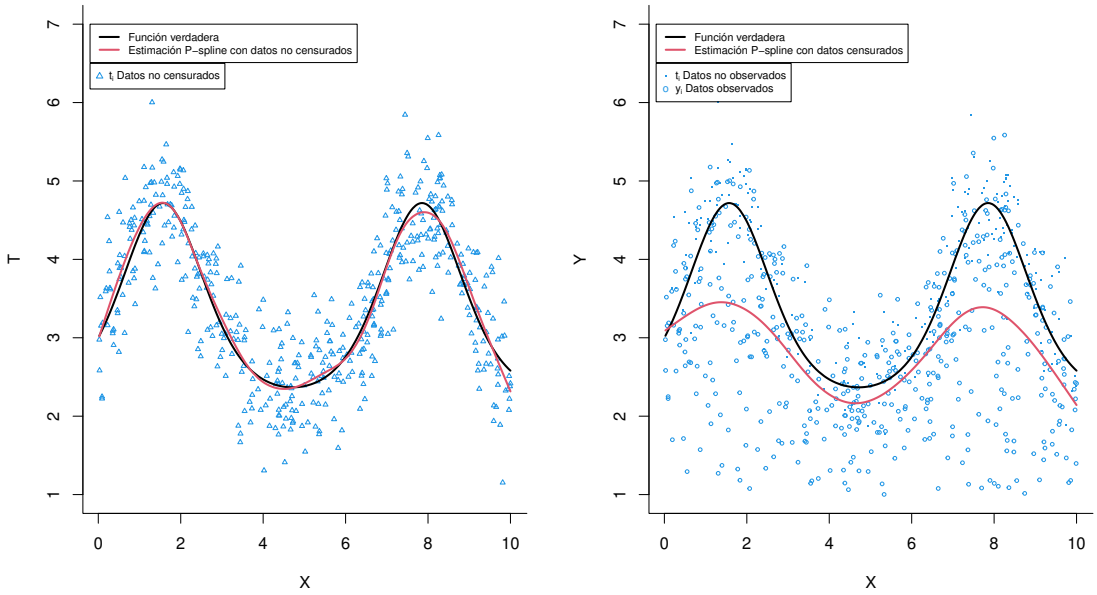
Como en otros métodos de alisado, hay que elegir el óptimo nivel de suavizado. En concreto, en este método debe elegirse el valor del parámetro λ. Estos autores sugieren utilizar validación cruzada o el criterio de información de Akaike (AIC) para elegir el valor de λ. Eilers and Marx (1996) resumen las ventajas de su propuesta de P-splines comparándola con otras metodologías de suavizado.

2.1.2. P-splines censurados

Hasta el momento se ha considerado una muestra en la que la variable de interés T es completamente conocida. A veces esta variable T no es completamente conocida porque su valor está censurado para algunos individuos, una situación común en los análisis de supervivencia, vida y duración, en los que T mide el tiempo transcurrido hasta la aparición de un suceso. Estos tiempos se conocen como tiempos de vida o supervivencia cuando el suceso de interés es la muerte de un individuo, como suele ocurrir en el ámbito biomédico. En otros ámbitos, como la economía o la ingeniería, se conoce como duración o tiempo de fallo de un individuo o elemento. Una de las características más importantes que suelen aparecer con este tipo de datos es la existencia de observaciones censuradas, y el patrón de censura más común es la censura por la derecha, *i.e.* el hecho de que para algunos individuos el suceso de interés aún no se ha producido al final del estudio. Por tanto, sabemos que su tiempo de supervivencia T es mayor que un valor observado, es decir, no conocemos su valor real. El objetivo de este estudio es extender el enfoque P-splines de Eilers and Marx (1996) a este tipo de situaciones.

Supongamos que $t_1, \ldots, t_n$ son observaciones independientes de una función de distribución de probabilidad desconocida F del tiempo de supervivencia T. Estos valores pueden no ser todos observables debido a los tiempos de censura $c_1, \ldots, c_n$ de cada uno de los individuos. Sean $x_1, \ldots, x_n$ los valores de la covariable X. Por

Figura 2.1: Estimación P-splines con datos no censurados y censurados



tanto, cuando hay datos censurados, en una muestra de tamaño n se observan $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$ donde $y_i = min(t_i, c_i)$ es el tiempo de supervivencia observado, que es el mínimo entre el tiempo de supervivencia t_i y el valor de censura c_i. Además, se sabe qué observaciones no están censuradas, mediante la variable de estado o indicador de censura $\delta_i = I(t_i \leq c_i)$.

Para ilustrar la necesidad de una propuesta que tenga en cuenta el efecto de los datos censurados se presenta una muestra simulada en la figura 2.1. Esta figura muestra dos paneles: a la izquierda, una muestra con información completa y, a la derecha, una muestra con observaciones censuradas. El panel izquierdo de la figura 2.1 muestra el diagrama de dispersión del tiempo de supervivencia frente a la covariable, las parejas de valores (t_i, x_i), junto a la función verdadera, $f(\cdot)$, y la curva ajustada utilizando el enfoque P-splines de Eilers and Marx (1996), en una situación no censurada, *i.e.* todos los tiempos de supervivencia son conocidos exactamente y se utiliza la muestra de n observaciones (t_i, x_i). En el panel derecho se ha añadido la muestra censurada, es decir, debido a la censura no todos los tiempos de supervivencia se conocen con exactitud y se dibuja la muestra de n tiempos de supervivencia observados (y_i, x_i), donde $y_i = min(t_i, c_i)$, representados en el panel como puntos circulares. En este ejemplo el 60 % de las observaciones coinciden con los puntos del gráfico de la izquierda $(y_i = t_i)$ porque no están censuradas. El otro 40 % son observaciones censuradas, con valores inferiores a los puntos correspondientes en el panel de la izquierda $(y_i = c_i)$.

Además, el panel izquierdo muestra la estimación de la curva $f(\cdot)$ por el método P-splines para la muestra sin datos censurados (línea roja) junto con la verdadera función. Como puede verse, el método de estimación capta la verdadera forma funcional. Esta estimación sólo es posible para el caso no censurado, en el que se dispone de información completa, *i.e.* la muestra (t_i, x_i) es totalmente conocida. Pero si los datos están censurados la estimación anterior no puede obtenerse. Es decir, la censura significa que no todos los tiempos de supervivencia se conocen con exactitud y

debe utilizarse la muestra de n tiempos de supervivencia observados (y_i, x_i). El panel derecho muestra la curva estimada utilizando P-splines para el caso de la muestra con observaciones censuradas (línea roja). Como puede observarse, la estimación del panel derecho tiene un sesgo importante y, por tanto, debiera corregirse teniendo en cuenta el efecto de la censura. Ese es el principal objetivo en este estudio.

En primer lugar, se adapta el enfoque de B-splines a situaciones de datos censurados, *i.e.* B-splines censurados. Partiendo del trabajo de Stute (1993) se utilizan los pesos Kaplan Meier para tener en cuenta el efecto de la censurada. Así, la minimización de la suma de cuadrados en (2.1) se modifica por la siguiente fórmula de mínimos cuadrados ponderados:

$$\sum_{i=1}^{n} w_{[i]} \left[y_{(i)} - \sum_{j=1}^{q} \gamma_j B_j(x_{[i]}) \right]^2 \tag{2.3}$$

donde $y_{(1)}, \ldots, y_{(n)}$ son los valores ordenados del tiempo de supervivencia observado $y_i = min(t_i, c_i)$, $x_{[i]}$ es el valor de la covariable asociado a la i-esima observación ordenada, $y_{(i)}$, y $w_{[i]}$ es el peso Kaplan-Meier asignado a $y_{(i)}$.

Los pesos Kaplan-Meier se pueden calcular como la contribución o incremento del estimador de Kaplan-Meier $(\hat{F}_n)$ de la función de distribución F de la variable T en cada valor $y_{(i)}$ (Kaplan and Meier, 1958), esto es:

$$w_{[i]} = \hat{F}_n(y_{(i)}) - \hat{F}_n(y_{(i-1)}) = \frac{\delta_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \left[\frac{n-j}{n-j+1} \right]^{\delta_{[j]}} \tag{2.4}$$

donde $\delta_{[i]}$ es el valor del indicador de censura asociado a la i-ésima observación ordenada no censurada $y_{(i)}$.

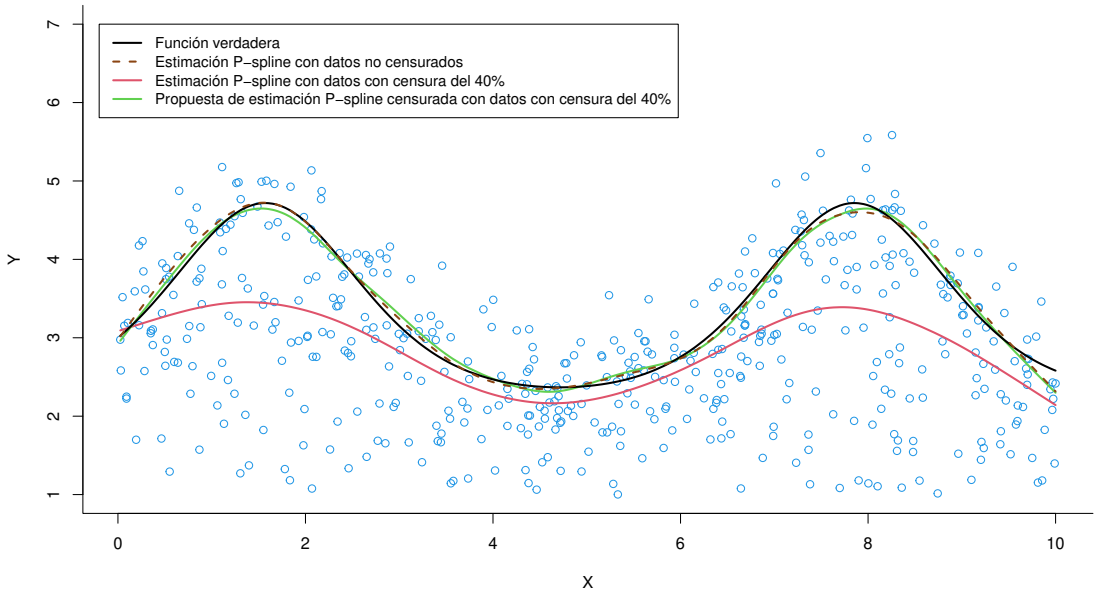
La solución por mínimos cuadrados ponderados del problema de minimización de la ecuación (2.3) es $\hat{f}(x) = B\hat{\gamma}$, donde B denota la matriz $n \times q$ con $B_{ij} = B_j(x_i)$ y $\hat{\gamma}$ es un vector de $q \times 1$ coeficientes $\hat{\gamma}_1, ...\hat{\gamma}_q$. Además, $\hat{\gamma} = (B'WB)^{-1}B'WY$, donde W es una matriz diagonal de $n \times n$ con los pesos Kaplan-Meier y Y es el vector de tiempos de supervivencia observados. Al igual que en el caso no censurado, este enfoque de B-splines censurado tiene el problema de la selección de los nodos. Como se mencionó anteriormente, una posible solución es elegir un gran número de nodos y utilizar un término de penalización para controlar la suavidad de la función. Así, basándonos en la propuesta de Eilers and Marx (1996), se modifica la expresión (2.2) para tener en cuenta la presencia de datos censurados y se propone lo que se puede denominar una aproximación P-splines censurada minimizando la siguiente expresión:

$$\sum_{i=1}^{n} w_{[i]} \left[y_{(i)} - \sum_{j=1}^{q} \gamma_j B_j(x_{[i]}) \right]^2 + \lambda \sum_{j=k+1}^{q} (\Delta^k \gamma_j)^2 \tag{2.5}$$

Por tanto, en el problema de minimización (2.5) se tienen en cuenta varias cuestiones: la bondad del ajuste mediante la suma ponderada de los residuos al cuadrado, la suavidad de la función mediante el término de penalización y la presencia de datos censurados mediante los pesos Kaplan-Meier. La expresión (2.5) se puede reescribir en forma matricial como

$$(Y - B\gamma)'W(Y - B\gamma) + \lambda\gamma' D_k' D_k \gamma \tag{2.6}$$

Figura 2.2: Propuesta de estimación P-splines censurada



donde D_k es la representación matricial del operador de diferencias Δ^k. La diferencia más habitual en la práctica es la de orden $k = 2$ En este caso, la representación matricial del operador de diferencias es

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots \\ 0 & 1 & -2 & 1 & \dots \\ 0 & 0 & 1 & -2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

La expresión (2.6) se minimiza para $\hat{\gamma} = (B'WB + \lambda D_k' D_k)^{-1} B'WY$. Por tanto, la curva estimada utilizando la propuesta metodología de P-splines censurados es $\hat{f}(x) = \sum_{j=1}^{q} \hat{\gamma}_j B_j(x)$. En la sección 4.1 puede encontrarse el código de R de la función *pswc* (función de R 4.2) que calcula el estimador P-splines censurado.

Este enfoque se ha demostrado consistente en el contexto del modelo de regresión lineal censurado paramétrico puro Stute (1993) y para la regresión no lineal paramétrica censurada Stute (1999), siempre que se cumplan las siguientes condiciones de identificabilidad: i) independencia entre los tiempos de vida y los tiempos de censura; y ii) dada la duración, la covariable no proporciona más información sobre si la observación está censurada o no, un supuesto más débil que la independencia entre los tiempos de censura y la covariable (véase Stute, 1993, 1999, para una discusión de estos supuestos).

Finalmente, se puede aplicar la metodología propuesta, los P-splines censurados, a la muestra simulada de la figura 2.1, *i.e* minimizar la ecuación (2.5) utilizando B-splines de orden 3 y un término de penalización de orden 2, los valores más comunes en la práctica. Se ha utilizado un valor óptimo para el parámetro de suavizado (véase el apartado 2.2.1) y un número de nodos equidistantes calculado aplicando la fórmula por defecto presentada en Ruppert (2002). La figura 2.2 muestra la función verdadera $f(\cdot)$ junto con los resultados de la estimación del método P-splines para

las muestras no censuradas y censuradas y los resultados del método de P-splines censurado propuesto. En la figura 2.2 se puede observar el buen funcionamiento del método de P-splines censurado, cuyos valores estimados se aproximan mucho a la función verdadera $f(\cdot)$ y a la curva ajustada en el caso no censurado. En cualquier caso, hay que tener en cuenta que esta última curva estimada (la correspondiente al caso no censurado) no puede estimarse en la práctica porque no se conoce la muestra completa debido a la censura. Además, se puede apreciar que el sesgo de la estimación del método P-splines para los datos censurados se corrige con la propuesta P-splines censurada.

2.2. Selección de parámetros

Como se ha visto en el apartado anterior, al igual que en el caso no censurado, las propuestas de estimación presentadas requieren, por un lado, la elección del parámetro de suavizado λ y, por otro, la elección del número y la ubicación de los nodos. La elección del parámetro de suavizado, es decir, el valor elegido para la ponderación del término de penalización, es la decisión más importante. La elección de los nodos es menos importante y parece que, una vez elegido un número de nodos suficientemente grande para reflejar las características de la función que se quiere estimar, el posible sobreajuste de la estimación se controla mediante el término de penalización presente en el método, que controla la suavidad de la función estimada.

2.2.1. Parámetro de suavizado

Para ilustrar la importancia de la elección del parámetro de suavizado, que ya se ha destacado como una elección fundamental para obtener una buena estimación de la curva o función $f(\cdot)$ en cualquier técnica de suavizado, se presenta la Figura 2.3. Esta figura resume las estimaciones del método P-splines censurado (ckmPS) que minimiza la expresión (2.5) para dos elecciones diferentes del parámetro de suavizado ($\lambda = 0,001$ y $\lambda = 1$).

En ambos casos se utilizan B-splines de orden 3 y un término de penalización de orden 2 (los valores más comunes en la práctica) y se consideran 1000 réplicas Monte Carlo. La figura 2.3(a) muestra la función verdadera $f(x)$ y la media y los límites superior e inferior del 95 % de los valores estimados utilizando el método P-splines censurado con $\lambda = 0{,}001$ para un escenario con un nivel de censura del 25 % y un tamaño de muestra de $n = 500$. La figura 2.3(b) muestra los resultados para el mismo caso pero con $\lambda = 1$. Como puede verse en la figura 2.3(b), una elección incorrecta del parámetro de suavizado λ puede conducir a una mala estimación de la función.

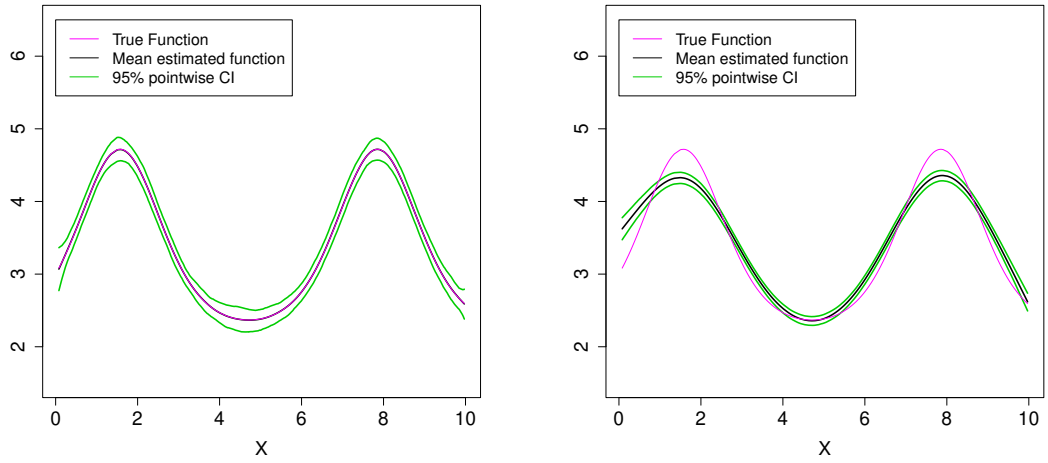
Para el caso no censurado, la utilización del criterio de validación cruzada generalizada para la elección del parámetro de suavizado (Wahba, 1990) supone que el parámetro de suavizado óptimo es el valor que minimiza la siguiente expresión:

$$GCV_{nc} = \sum_{i=1}^{n} \frac{(t_{(i)} - \hat{t}_{(i)})^2}{(n - tr(H_{nc}))^2}$$

donde $H_{nc} = B(B'B + \lambda D_k' D_k)^{-1} B'$ es la matriz de suavizado del caso no censurado.

Figura 2.3: Estimación con dos valores diferentes de λ

(a) $\lambda = 0{,}001$ (b) $\lambda = 1$



Como primera aproximación al caso censurado se podría elegir un valor de λ que minimice la expresión habitual para el criterio GCV, pero donde la matriz de suavizado H del caso no censurado se sustituye por la del caso censurado H_c:

$$GCV_1 = \sum_{i=1}^{n} \frac{(y_{(i)} - \hat{y}_{(i)})^2}{(n - tr(H_c))^2} \tag{2.7}$$

donde $H_c = B(B^{'}WB + \lambda D_k^{'} D_k)^{-1} B^{'}W$ es la matriz de suavizado del caso censurado y W es una matriz diagonal con los pesos Kaplan-Meier, $w_{[i]}$, asociados con los correspondientes valores observados de la supervivencia, $y_{(i)}$.

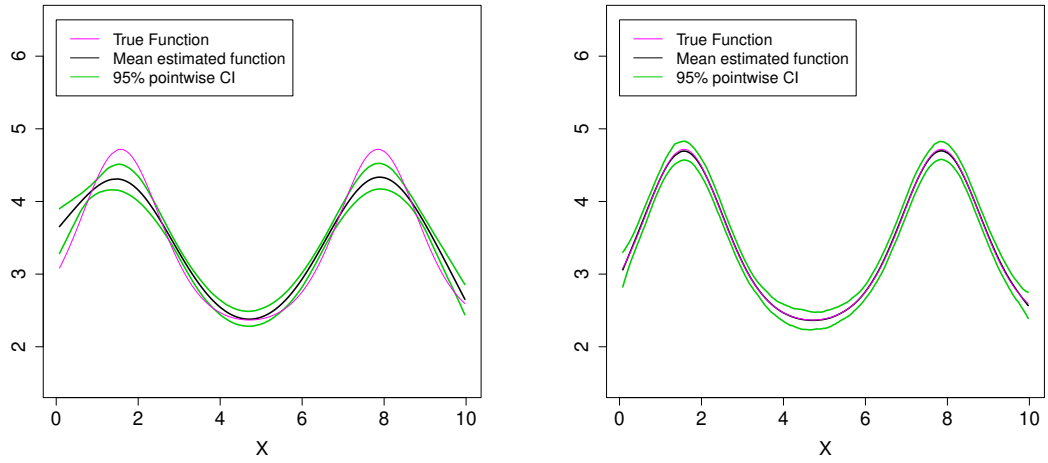
La figura 2.4(a) resume las estimaciones del método P-splines censurado que minimiza la expresión (2.5) utilizando el valor que optimiza la expresión (2.7) como parámetro de suavizado λ. En ella se muestran la media y los límites superior e inferior del 95 % de los valores estimados utilizando el método P-splines censurado para 1000 réplicas Monte Carlo. Como puede observarse, el método proporciona una estimación pobre, que empeora a medida que se incrementa el porcentaje de observaciones censuradas en la muestra. Para corregir esto, parece necesario tener en cuenta el efecto de la censura en el numerador, además de en el denominador de la expresión (2.7) con la matriz H_c. Para ello se propone la siguiente modificación de la expresión (2.7):

$$GCV_2 = \sum_{i=1}^{n} \frac{w_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{(n - tr(H_c))^2} \tag{2.8}$$

Esta nueva expresión ahora tiene en cuenta el efecto de la censura en el numerador de la expresión GCV, ponderando con sus correspondientes pesos Kaplan-Meier cada sumando del numerador, es decir, el cuadrado de las diferencias $(y_{(i)} - \hat{y}_{(i)})$. La figura 2.4(b) muestra un mejor resultado de las estimaciones realizadas cuando se utiliza la expresión GCV_2 de la fórmula (2.8) como criterio para la elección del

Figura 2.4: Criterios para la elección del parámetro λ, GCV_1 versus GCV_2

(a) GCV_1 (b) GCV_2



parámetro λ en comparación con los resultados de la figura 2.4(a), que utiliza la expresión GCV_1 de la fórmula (2.7) .

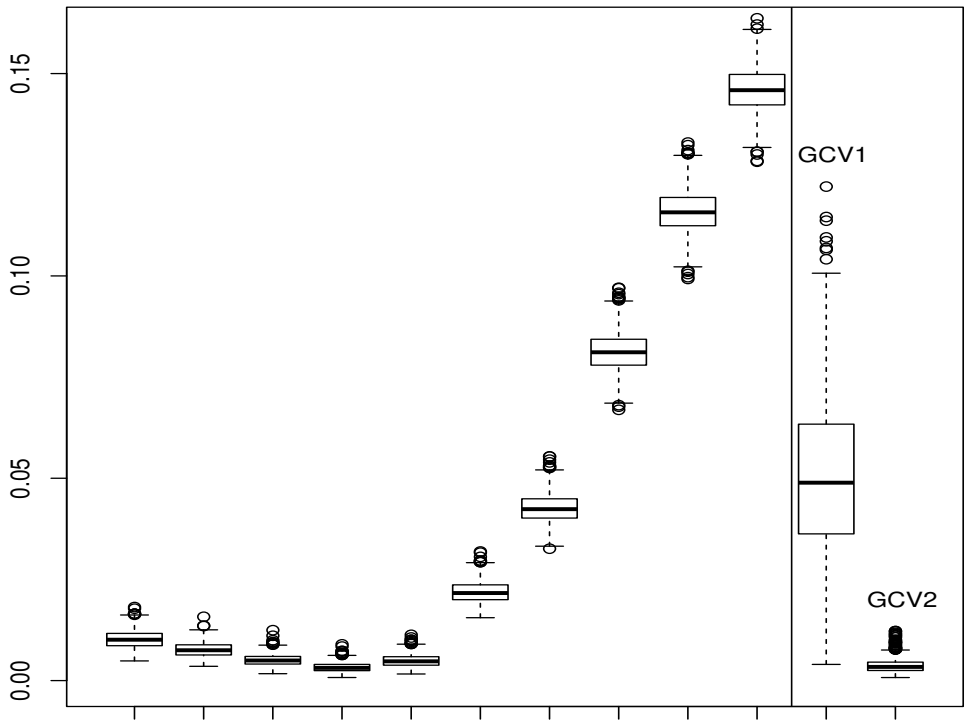
Además, se ha realizado un estudio de simulación para ver cómo funcionan las propuestas GCV_1 y GCV_2. Se han elegido una serie de valores candidatos para el parámetro λ, y para cada valor candidato se ha calculado el error cuadrático medio (ECM) de cada una de 1000 réplicas de Monte Carlo. El panel izquierdo de la figura 2.5 muestra el ECM de estimar el modelo para cada valor específico del parámetro de suavizado en la red de posibles valores para el parámetro λ que va de $0,000001$ a 4. El panel derecho de la figura 2.5 muestra el ECM de la estimación del modelo con el valor del parámetro de suavizado elegido mediante los criterios GCV_1 y GCV_2. Se puede observar que el criterio GCV_2 propuesto funciona bien.

En la literatura para el caso no censurado se ha encontrado que el GCV tiene una tendencia a sobreajustar la función estimada. Una posible solución a este problema es establecer un parámetro de ajuste $\phi \geq 1$ en la expresión GCV_2 que añada una penalización adicional. De este modo, la propuesta final sería:

$$GCV_c = \sum_{i=1}^{n} \frac{w_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{(n - \phi tr(H_c))^2} \tag{2.9}$$

Para el caso no censurado, el CGV habitual utiliza un valor $\phi = 1$, pero Wood (2017) propone corregir el problema de sobreajuste realizando lo que se denomina doble validación cruzada e introduciendo un valor de $\phi = 1,5$. Este valor se deriva de varias formas en la literatura (e.g. Kim and Gu, 2004). En la sección 3.2 se presenta un amplio estudio de simulación en el que, entre otros aspectos, se analiza el rendimiento de la propuesta (2.9) para elegir el nivel óptimo de suavizado en muestras censuradas y se comparan los resultados considerando los valores $\phi = 1$ y $\phi = 1,5$. Además, en esta propuesta de adaptación al caso censurado, también se analizan dos posibles variaciones a la hora de controlar el efecto de la censura: utilizar los pesos de Kaplan-Meier ($w_{[i]}$), tal y como se incluyen en la propuesta presentada, o sus cuadrados ($w^2_{[i]}$) para ponderar el numerador del criterio GCV_c.

Figura 2.5: ECM para el parámetro λ en la red 0,000001 a 4 y para λ_{GCV_1} y λ_{GCV_2}



2.2.2. Nodos

Como ya se ha mencionado, la elección de los nodos no es fundamental siempre que se elija un número suficientemente grande para ajustarse a las características de los datos, dado que se utiliza un término de penalización para evitar un posible sobreajuste de la estimación. En cualquier caso, en esta sección se va a analizar el efecto sobre la estimación tanto del número de nodos como de su ubicación.

Es habitual en la literatura utilizar un número de nodos elegido aplicando la fórmula presentada en Ruppert (2002) basada en la elección por defecto de Wand. Ruppert propone elegir el siguiente número de nodos:

$$K_{rp} = round\left(\text{mín}\left(\frac{m}{4}, 40\right)\right) \tag{2.10}$$

donde m es el número de valores diferentes que toma el regresor X. Ruppert (2002) estudia el rendimiento de esta regla por defecto en el enfoque de splines penalizados y concluye que elige un número efectivo de nodos en todos los casos estudiados.

Para el caso censurado, se propone a continuación una modificación que tiene en cuenta la perdida de información muestral disponible debido a la existencia de datos censurados. Así, el número de nodos (K) a utilizar puede seleccionarse automáticamente con siguiente fórmula de selección:

$$K_c = round\left(\text{mín}\left(\frac{m}{4}, 40\right) \cdot (1 - PC)\right) \tag{2.11}$$

donde PC representa el nivel de censura, en tanto por uno, existente en la muestra analizada.

Por último, para elegir la localización de los nodos (L), además de nodos equidistantes (L_{eq}) se propone y analiza una adaptación al caso de datos censurados que utiliza vectores de nodos no uniformes con un espaciado entre los mismos elegido en función de los pesos de Kaplan-Meier (L_{km}). De forma análoga a la elección de nodos no uniformes utilizando los cuantiles del regresor X (véase, por ejemplo, Ruppert et al., 2003) se propone una elección de la ubicación de los nodos que divida el rango de la variable X en intervalos continuos con igual suma de las pesos Kaplan-Meier. Esto garantiza que cada intervalo tenga el mismo peso Kaplan-Meier.

Todas estas propuestas (secciones 2.2.1 y 2.2.2) se analizan y comparan en un estudio de simulación que se presenta en la sección 3.2 para distintas situaciones. En la sección 4.2 puede encontrarse el código de R desarrollado para incorporar las modificaciones que generalizan los criterios de elección de parámetros desarrollados en esta sección: las modificaciones de las funciones de R 4.2 y 4.3, *pswc* modificada (función de R 4.6), que calcula el estimador P-splines censurado (estimador ckmPS) y *gamkm* modificada (función de R 4.7) que calcula el estimador en un modelo GAM univariante corrigiendo el efecto de la censura con los pesos Kaplan-Meier (estimador ckmGAM).

2.3. Modelo semiparamétrico

La existencia de observaciones censuradas es muy común en el análisis de supervivencia o de duración, donde se pretende analizar una variable que mide la duración de un suceso o estado o el tiempo que transcurre hasta que ocurre un determinado suceso. En esta sección se estudia un modelo que permite analizar el efecto de determinadas variables explicativas sobre una variable a explicar T, la variable duración o normalmente su transformación logarítmica, donde algunas de las observaciones están censuradas. Es relativamente habitual separar los efectos de las variables explicativas del modelo en dos componentes: un componente paramétrico, que capta la relación entre algunas variables explicativas (X) y la variable de respuesta asumiendo una forma funcional paramétrica específica y otro componente, no paramétrico, que recoge los efectos de otras variables explicativas (Z) cuya forma funcional se desconoce y que dejamos sin especificar, sin asumir una relación relación paramétrica concreta. Por tanto, se considera un modelo de regresión semiparamétrico pero en un contexto en el que la variable a explicar en el modelo está censurada por la derecha:

$$T_i = X_i'\alpha + f(Z_i) + \epsilon_i \qquad i = 1, \ldots, n \tag{2.12}$$

donde se supone que los valores de la variable T, $t_1, \ldots, t_n$, son independientes y generados con una función de distribución de probabilidad desconocida F. Es decir, no se supone una distribución de probabilidad específica para el término de error. Además, algunas observaciones de esa variable T están censuradas por la derecha. Por lo tanto, lo que realmente se observa en la muestra es la variable $y_i = \text{mín}(t_i, c_i)$, donde los valores $c_1, \ldots, c_n$ son los valores de la variable censurada, C. Para el mecanismo de censura se supone: a) que la duración y la censura son independientes entre sí y, b) dada la duración, las covariables no proporcionan más información sobre si la censura tendrá lugar o no, *i.e.*, $P[T \leq C|X, Z, T] = P[T \leq C|T]$ (véase Stute, 1993, 1999, para una discusión de estos supuestos).

El indicador $\delta_i = I(t_i \leq c_i)$ muestra si un valor t_i es observado, *i.e.*, no está cen-

surado. Además, X_i es el vector $(p \times 1)$ que recoge los valores de las p variables explicativas del componente paramétrico para el i-ésimo individuo, α es el vector $(p \times 1)$ de coeficientes del modelo asociado a estos regresores, $f(Z)$ representa el componente no paramétrico del modelo, que recoge la forma funcional desconocida del efecto de la variable regresora Z y ϵ es el término de error que satisface $E(\epsilon|X, Z) = 0$ y $Var(\epsilon|X, Z) = \sigma^2$.

A continuación se extiende la metodología presentada en la sección 2.1.2 a un entorno más general, permitiendo su utilización en problemas más complejos, más habituales en la práctica, proponiendo un método de estimación que permite estimar modelos de regresión semiparamétricos para datos censurados. En este contexto se proponen estimadores tanto para la parte paramétrica como para la parte no paramétrica. Además, se propone un estimador de las varianzas para ambas componentes, paramétrica y no paramétrica, y se proporcionan las herramientas necesarias para poder realizar inferencia en este tipo de modelos.

2.3.1. Método de estimación

Siguiendo las ideas presentadas en la sección 2.1.2 se extiende la metodología de los P-splines al contexto de muestras con observaciones censuradas en un modelo semiparamétrico. Así, para estimar el modelo (2.12) se propone minimizar la siguiente expresión:

$$\sum_{i=1}^{n} w_{[i]} \left[y_{(i)} - x'_{[i]}\alpha - \sum_{j=1}^{q} \gamma_j B_j(z_{[i]}) \right]^2 + \lambda \sum_{j=k+1}^{q} (\Delta^k \gamma_j)^2 \tag{2.13}$$

donde $y_{(1)}, \ldots, y_{(n)}$ son los valores ordenados de la variable observada $y_i = min(t_i, c_i)$, $x'_{[i]}$ es el vector $(1 \times p)$ con los valores de los regresores del componente paramétrico para el individuo asociado a la observación ordenada $y_{(i)}$ y $w_{[i]}$ es el peso Kaplan-Meier asociado a esta observación $y_{(i)}$.

Además, $\Delta\gamma_j$ denota la diferencia entre los coeficientes de B-splines adyacentes y $\Delta^k\gamma_j$ indica que esta diferencia es de orden k. Esta diferencia mide la suavidad de la función $f(z)$, cuanto mayor sea la diferencia entre los coeficientes de B-splines adyacentes menos suave será la función. Por último, el parámetro λ es el parámetro de suavizado que controla el grado de suavidad de la función estimada en el proceso de estimación.

La expresión a minimizar (2.13) puede reescribirse en forma matricial como sigue:

$$(Y - X\alpha - B\gamma)' W (Y - X\alpha - B\gamma) + \lambda \gamma' D_k' D_k \gamma \tag{2.14}$$

donde X es la matriz de diseño $(n \times p)$ para las variables del componente paramétrico, Y es el vector de la variable observada a explicar, B es una matriz $(n \times q)$ donde $B_{ij} = B_j(z_i)$, W es una matriz diagonal $(n \times n)$ con los pesos Kaplan-Meier y D_k es la matriz utilizada para reescribir el término Δ^k en forma matricial.

2.3.2. Algoritmo

El proceso de optimización de la expresión (2.14) conduce a las siguientes ecuaciones:

$$(X'WX)\,\alpha = X'W(Y - B\gamma) \tag{2.15}$$

$$(B'WB + \lambda D_k' D_k)\,\gamma = B'W(Y - X\alpha) \tag{2.16}$$

En la práctica, las estimaciones de α y γ pueden obtenerse mediante un proceso iterativo o *backfitting algorithm* que resuelve iterativamente cada conjunto de ecuaciones (2.15) y (2.16) hasta alcanzar la convergencia de los estimadores. A continuación se describen los pasos del algoritmo:

Paso 1. En la ecuación (2.16) dar un valor inicial de $\widehat{\alpha}_{(0)} = \vec{0}$ y estimar γ como $\widehat{\gamma}_{(0)} = [B'WB + \lambda D_k' D_k]^{-1} B'WY$.

Paso 2. Substituir γ por $\widehat{\gamma}_{(0)}$ en la ecuación (2.15) y estimar α mediante $\widehat{\alpha}_{(1)} = [X'WX]^{-1} X'W(Y - B\widehat{\gamma}_{(0)}) = [X'WX]^{-1} X'W(I - H_c)Y$ donde $H_c = B\left(B'WB + \lambda D_k' D_k\right)^{-1} B'W$ es la matriz de suavizado para el caso censurado obtenida de la ecuación (2.16).

Paso 3. Substituir α por $\widehat{\alpha}_{(1)}$ en la ecuación (2.16) y estimar γ como $\widehat{\gamma}_{(1)} = [B'WB + \lambda D_k' D_k]^{-1} B'W(Y - X\widehat{\alpha}_{(1)})$.

Paso 4. Iterar los pasos 2 a 3 hasta alcanzar la convergencia.

Se considera que el algoritmo ha convergido cuando la diferencia entre los GCV_c (ver ecuación 2.9) de dos iteraciones sucesivas es menor que un umbral realmente pequeño, por ejemplo, $|GCV_c(nuevo) - GCV_c(anterior)| < 0{,}00001 \cdot GCV_c(nuevo)$.

Al igual que en el caso no censurado o en el caso univariante censurado, el método de estimación presentado requiere, por un lado, la elección del parámetro de suavizado λ y, por otro, la elección del número y la ubicación de los nodos. Estos parámetros se eligen siguiendo la metodología propuesta en la sección 2.2.

2.3.3. Estimadores de las varianzas

En esta sección se desarrollan las herramientas necesarias para realizar inferencias estadísticas para los componentes paramétricos y no paramétricos del modelo (2.12).

Para determinar la varianza del componente paramétrico, primero se resuelve la ecuación (2.16) obteniendo $\gamma = (B'WB + \lambda D_k' D_k)^{-1} B'W(Y - X\alpha)$. Sustituyendo $B\gamma = H_c(Y - X\alpha)$ en (2.15) se obtiene $(X'WX)\,\alpha = X'W\left[Y - H_c(Y - X\alpha)\right]$. Y resolviendo esto para α se obtiene $\widehat{\alpha} = [X'W(I - H_c)X]^{-1} X'W(I - H_c)Y$. En consecuencia, la matriz de varianzas y covarianzas de este estimador puede expresarse como:

$$\begin{aligned} \widehat{Var}(\widehat{\alpha}) &= \widehat{\sigma}^2 \Big\{ (X'W(I - H_c)X)^{-1} X'W(I - H_c)(I - H_c)^t WX \\ &\quad \left((X'W(I - H_c)X)^{-1}\right)^t \Big\} \end{aligned} \tag{2.17}$$

Igualmente en la ecuación (2.15) se despeja $\alpha = (X'WX)^{-1} X'W(Y - B\gamma)$. Sustituyendo $X\alpha = X\left(X'WX\right)^{-1} X'W(Y - B\gamma) = H_p(Y - B\gamma)$, donde $H_p = X\left(X'WX\right)^{-1} X'W$, en (2.16) se obtiene $\left(B'WB + \lambda D_k' D_k\right)\gamma = B'W\left[Y - H_p(Y - B\gamma)\right]$. Y despejando γ se obtiene $\widehat{\gamma} = [B'W(I - H_p)B + \lambda D_k' D_k]^{-1} B'W(I - H_p)Y$. En consecuencia, la matriz de varianzas y covarianzas de este estimador puede expresarse como:

$$\begin{aligned} \widehat{Var}(\widehat{\gamma}) &= \widehat{\sigma}^2 \Big\{ [B'W(I - H_p)B + \lambda D_k' D_k]^{-1} B'W(I - H_p)(I - H_p)^t WB \\ &\quad \left([B'W(I - H_p)B + \lambda D_k' D_k]^{-1} \right)^t \Big\} \end{aligned} \tag{2.18}$$

Para calcular estas varianzas estimadas se necesita estimar el parámetro σ^2. Para ello se propone el estimador dado por la siguiente expresión:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} nw_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{n - tr(H_c) - p}$$

2.3.4. Código

En la sección 4.3 puede encontrarse el código de R de la extensión de la metodología de los P-splines al contexto de muestras con observaciones censuradas en un modelo semiparamétrico, la función de R 4.8 *semipswc*. Esta función calcula, para un modelo semiparamétrico, las estimaciones de la componente no paramétrica y sus deviaciones típicas asociadas, las estimaciones de los parámetros de la componente paramétrica y sus deviaciones típicas asociadas, la estimación de la desviación típica del error y, además, proporciona los valores del parámetro de suavizado óptimo, el criterio GCV_c, el número de iteraciones necesarias para la estimación y los grados de libertad efectivos.

Capítulo 3

Resumen de resultados

3.1. Resultados caso univariante

3.1.1. Estudio de simulación

En esta sección se estudia el desempeño de la metodología propuesta en la sección 2.1 mediante un estudio de simulación. Para ello se consideran dos escenarios diferentes para la relación entre el logaritmo del tiempo de supervivencia, que se denota como T, y una covariable relevante X

$$t_i = f(x_i) + \epsilon_i,$$

donde los valores de la variable T no se conocen completamente porque algunas observaciones están censuradas por la derecha.

Caso 1: función sinusoidal

Para el primer escenario se considera el siguiente modelo para la función $f(\cdot)$:

$$f(x_i) = 2 + exp(sin(x_i))$$

donde la covariable X sigue una distribución uniforme en el intervalo $(0, 10)$, *i.e.* se utiliza una función $f(\cdot)$ con dos valores máximos locales y un mínimo. El término de error ϵ se genera como una variable normal $N(0, 0, 5)$. Para estudiar el efecto de la censura se considera una variable de censura C generada independientemente a partir de una distribución uniforme $U(1, b)$. El valor del parámetro b cambia para considerar tres niveles diferentes de censura: 10 %, 25 % y 40 %. Por tanto, se observa $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$ una muestra de tamaño n, donde $y_i = min(t_i, c_i)$ es la duración observada, *i.e.* el mínimo entre el tiempo de supervivencia t_i y el valor de censura c_i. Además, se sabe a través de la variable indicadora $\delta_i = I(t_i \leq c_i)$ qué observaciones no están censuradas. Se utilizan tres tamaños de muestra: $n = 200$, $n = 500$ y $n = 1000$. Para cada ejemplo se utilizan 1000 réplicas de Monte Carlo.

Se considera, como es habitual en la práctica, que la forma funcional de la relación entre la variable respuesta y la covariable es desconocida y se estima el modelo utilizando el método de P-splines censurado propuesto en la sección 2.1. Los resultados de la estimación se resumen en las figuras 3.1 y 3.2. La figura 3.1 muestra los gráficos de caja (uno para cada nivel de censura considerado) con los resultados del

Figura 3.1: Errores Cuadráticos Medios del método P-splines censurado utilizando diferentes niveles de censura y tamaños de muestra (caso 1)

(a) n=200 (b) n=500 (c) n=1000

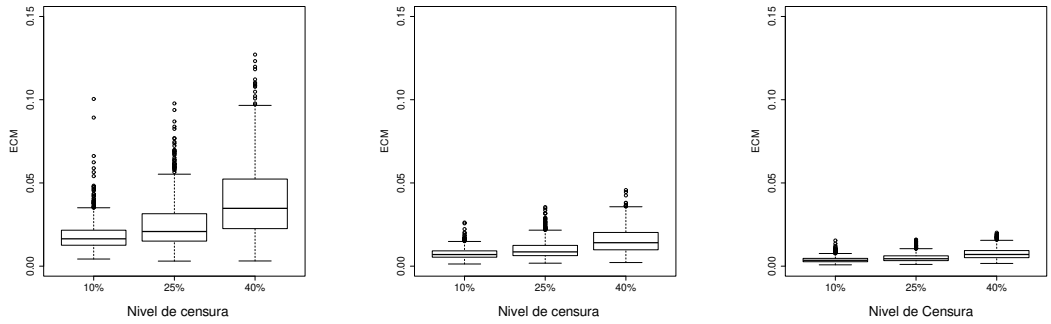
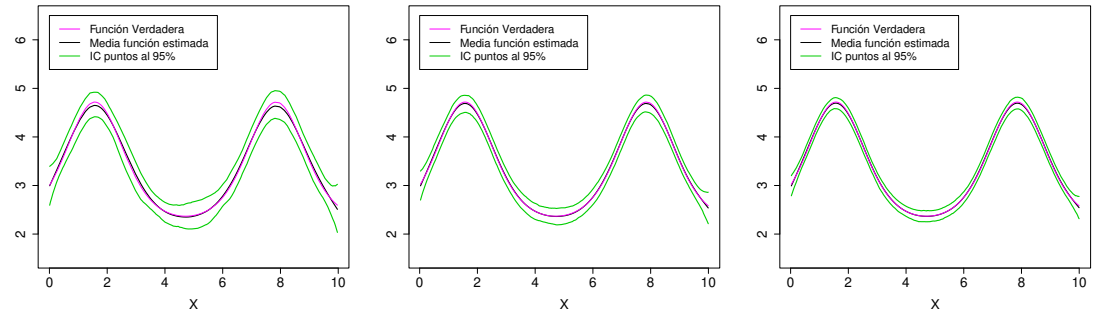
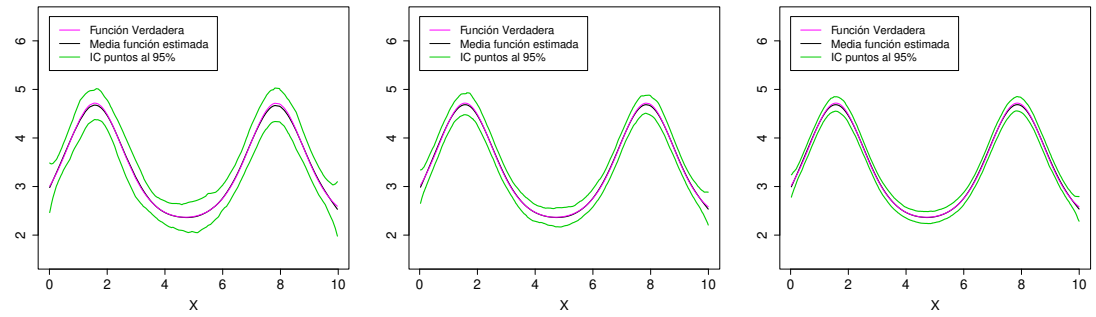


Figura 3.2: Función estimada utilizando el estimador P-splines censurado (caso 1)

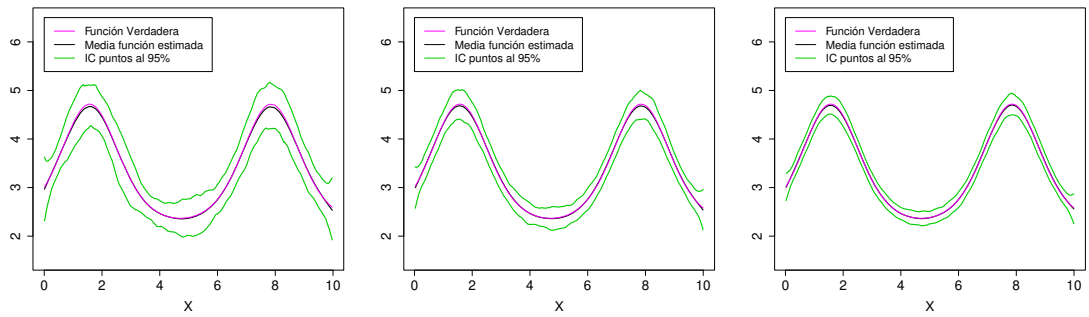
(a) C=10 % y n=200 (b) C=10 % y n=500 (c) C=10 % y n=1000



(d) C=25 % y n=200 (e) C=25 % y n=500 (f) C=25 % y n=1000



(g) C=40 % y n=200 (h) C=40 % y n=500 (i) C=40 % y n=1000



Error Cuadrático Medio (ECM) para cada réplica, donde el ECM se ha calculado como:

$$ECM = \frac{\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2}{n}$$

La figura 3.1(a) muestra el resultado de la estimación para la muestra de $n = 200$ observaciones, la figura 3.1(b) para $n = 500$, y la figura 3.1(c) para $n = 1000$. La figura 3.2 muestra la función verdadera $f(x)$ y la media y los límites superior e inferior del 95 % de los valores estimados utilizando el método P-splines censurado para cada nivel de censura y tamaño de muestra.

Del análisis de los resultados se puede concluir que el comportamiento del método P-splines censurado propuesto es bueno en todos los casos considerados: la función estimada $\hat{f}(x)$ recupera la verdadera forma funcional. La figura 3.1 muestra que el error cuadrático medio disminuye al aumentar el tamaño de la muestra para cada nivel de censura considerado. La figura 3.2 muestra que la media de las funciones estimadas está muy próxima a la función verdadera para todos los casos considerados. Además, a medida que aumenta el tamaño de la muestra, los límites superior e inferior del 95 % de los valores estimados se aproximan más a la función real. El efecto del nivel de censura es el esperado: los resultados son más precisos a niveles más bajos de censura y la variabilidad aumenta con el nivel de censura.

Caso 2: función cuadrática

En este segundo ejemplo se considera una relación cuadrática entre la variable respuesta y la covariable:

$$f(x_i) = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$$

donde $\beta_1 = 2$, $\beta_2 = 4$ y $\beta_3 = -1$. La covariable X toma valores de una distribución uniforme en el intervalo $(0, 4)$ y ϵ se genera como una variable normal $N(0, 0, 5)$. La variable de censura C se genera independientemente a partir de una distribución uniforme $U(2, b)$, cambiando el valor del parámetro b para considerar tres niveles diferentes de censura: 10 %, 25 % y 40 %. Como en el caso anterior, el tiempo de supervivencia observado y_i es el mínimo entre el tiempo de supervivencia t_i y el valor de censura c_i.

Se han utilizado los mismos tamaños de muestra ($n = 200$, $n = 500$ y $n = 1000$) y número de réplicas (1000) que en el primer escenario y se obtiene un resultado similar, el método P-splines censurado muestra un buen comportamiento como se observa en las figuras 3.3 y 3.4. La figura 3.3(a) muestra el error cuadrático medio (ECM) para cada réplica para el caso de $n = 200$, la figura 3.3(b) para $n = 500$ y la figura 3.3(c) para $n = 1000$. Como puede observarse, el error cuadrático medio disminuye a medida que aumenta el tamaño de la muestra.

Por otro lado, las figuras 3.4(a) a 3.4(c) muestran el efecto de aumentar el tamaño de la muestra en la media de las funciones estimadas y los límites superior e inferior del 95 % de los valores estimados de la función para un nivel de censura de 10 %. En las figuras 3.4(d) a 3.4(f) tenemos la misma información para una censura del 25 % y en las figuras 3.4(g) a 3.4(i) para una censura de 40 %. Estas figuras muestran el buen ajuste de la media de las funciones estimadas, y se puede observar que a medida que aumenta el tamaño de la muestra el desempeño del método mejora.

Finalmente, las estimaciones en las simulaciones de esta sección 3.1 se han obtenido minimizando la expresión (2.5) mediante B-splines de grado 3 y una penalización

Figura 3.3: Errores Cuadráticos Medios del método P-splines censurado utilizando diferentes niveles de censura y tamaños de muestra (caso 2)

(a) n=200 (b) n=500 (c) n=1000

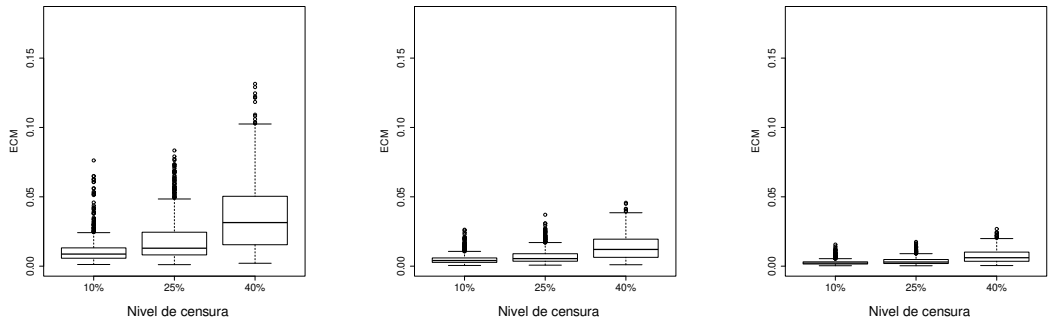
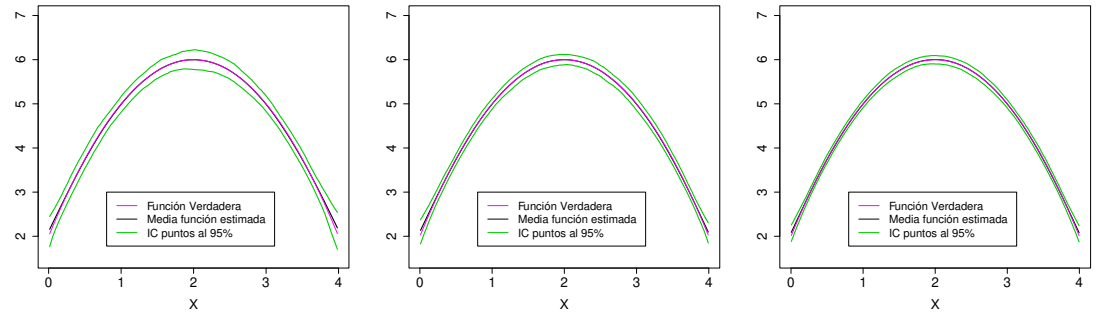
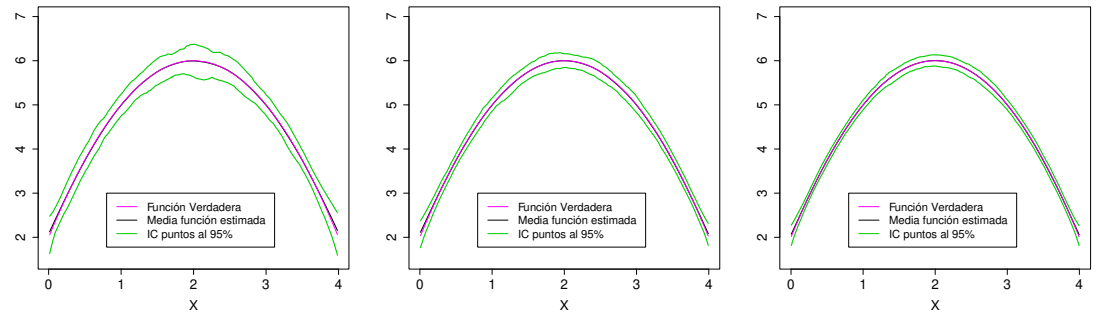


Figura 3.4: Función estimada utilizando el estimador P-splines censurado (caso 2)

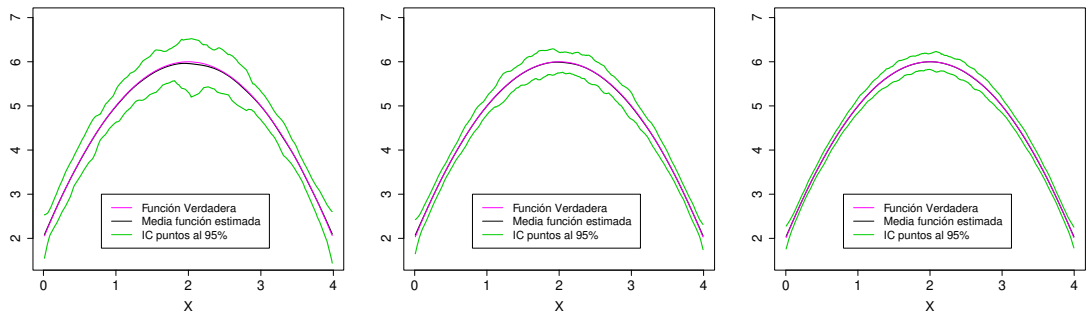
(a) C=10 % y n=200 (b) C=10 % y n=500 (c) C=10 % y n=1000



(d) C=25 % y n=200 (e) C=25 % y n=500 (f) C=25 % y n=1000



(g) C=40 % y n=200 (h) C=40 % y n=500 (i) C=40 % y n=1000



de orden 2, valores habituales en la práctica. El parámetro de suavizado λ se ha elegido mediante validación cruzada utilizando la ecuación (2.8) y el número de nodos equidistantes utilizando la fórmula (2.11). También se han realizado simulaciones adicionales para muestras de pequeño tamaño y considerando distribuciones de error no normales y asimétricas, como la distribución Weibull. Los nuevos resultados obtenidos (no mostrados) confirman el buen rendimiento del método propuesto y son coherentes con los presentados en esta sección.

3.1.2. Extensión a los modelos GAM

Las ideas que se han utilizado para ampliar la metodología P-splines, propuesta por Eilers y Marx, para el caso una variable respuesta censurada utilizando los pesos Kaplan Meier se pueden extender de manera sencilla a otros tipos de modelos. En este apartado se extiende la metodología propuesta en la sección 2.1 al marco de los modelos aditivos generalizados (GAM). En esta sección se analizará el funcionamiento de esta extensión para modelos con uno y dos regresores.

Un regresor

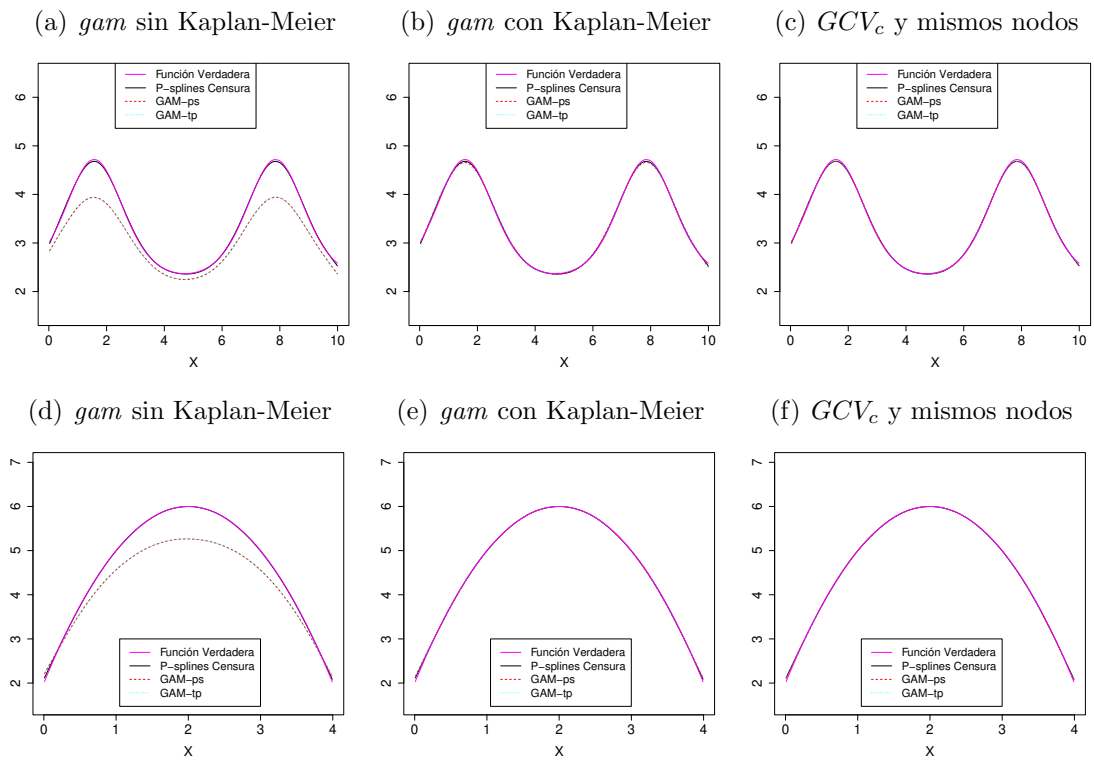
Para ilustrar el comportamiento de la extensión se utilizan los mismos escenarios que en las simulaciones del apartado anterior: una función sinusoidal (caso 1) y otra cuadrática (caso 2). La figura 3.5 muestra la media de las funciones estimadas obtenidas con el estimador propuesto en la sección 2.1 (P-splines censurado) y las basadas en la función *gam* del paquete R *mgcv* (Wood and Wood, 2015; R Core Team, 2015) para un tamaño muestral de 500 observaciones con un nivel de censura del 25 % y 1000 réplicas.

Las figuras 3.5(a) a 3.5(c) muestran los resultados para el caso 1 y las figuras 3.5(d) a 3.5(f) para el caso 2. Las figuras 3.5(a) y 3.5(d) muestran estimaciones obtenidas usando la función *gam* con bases de tipo *tp* (Wood, 2003, *thin plate regression splines*) y *ps* (P-splines estilo Eilers y Marx) sin corregir el efecto de la censura. Como era de esperar, la estimación tiene un sesgo sustancial. Las figuras 3.5(b) y 3.5(e) presentan la estimación utilizando la función *gam* pero corregida teniendo en cuenta el efecto de la censura con los pesos Kaplan-Meier. Como puede observarse, la estimación mejora considerablemente, con ligeras diferencias entre los tres estimadores analizados. En las figuras 3.5(c) y 3.5(f) no sólo se corrige el efecto de la censura con los pesos Kaplan-Meier, sino que el parámetro de suavizado y el número de nodos se eligen utilizando las propuestas de la sección 2.2: parámetro de suavizado seleccionado por GCV_c, fórmula (2.9), y número de nodos calculado utilizando la fórmula (2.11). Como se muestra en estas últimas figuras, las curvas son tan similares que resultan indistinguibles. En la sección 4.1 puede encontrarse el código de R utilizado en esta sección, la función *gamkm* (función de R 4.3) que calcula el estimador en un modelo GAM corrigiendo el efecto de la censura con los pesos Kaplan-Meier para el caso de un regresor.

Dos regresores

A continuación se presenta una ilustración en un modelo muy sencillo en el que la variable de respuesta depende linealmente de dos funciones suaves desconocidas de los regresores. Se utilizan las mismas formas funcionales con el mismo tamaño de

Figura 3.5: Media de las funciones estimadas: P-splines censurados y función *gam*



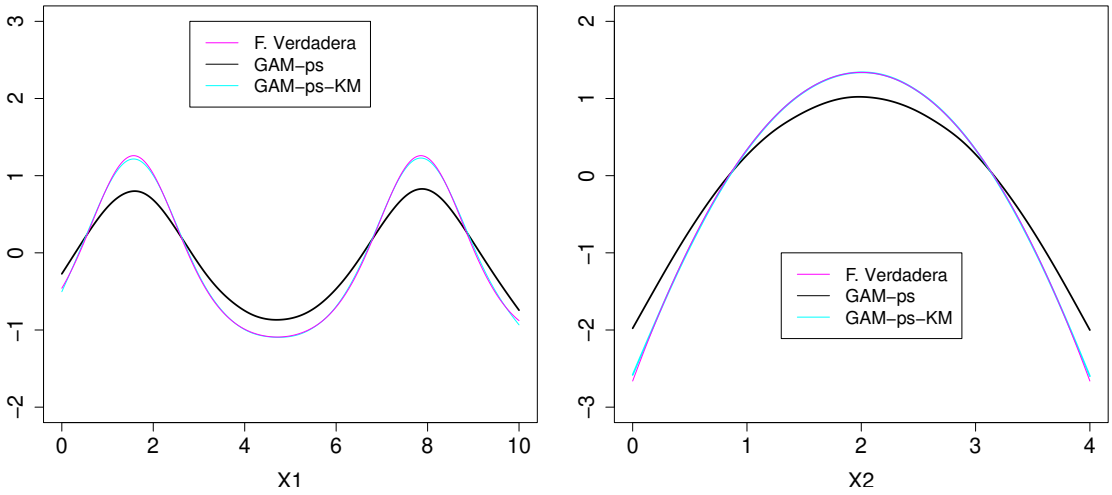
muestra y nivel de censura que en el caso anterior, pero se agregan en el modelo de forma aditiva:

$$f(x_{1i}, x_{2i}) = f(x_{1i}) + f(x_{2i}) + \epsilon_i = 2 + exp(sin(x_{1i})) + 4x_{2i} - x_{2i}^2 + \epsilon_i$$

La figura 3.6 muestra la media de las funciones estimadas obtenidas utilizando la función *gam* con base *ps* (P-splines al estilo de Eilers y Marx). Para cada regresor, esta figura muestra la media de la función estimada sin corregir el efecto de la censura (GAM-ps, línea negra), y la función estimada utilizando los pesos Kaplan-Meier para tener en cuenta el efecto de la censura (GAM-ps-KM, línea azul) donde el parámetro de suavizado y el número de nodos se eligen siguiendo la metodología presentada en la sección 2.2. Los resultados obtenidos muestran que el método funciona bien. Se han realizado ejercicios similares utilizando otras bases (*thin plate regression splines* o *penalized cubic regression splines*) con resultados indistinguibles. En la sección 4.1 puede encontrarse el código de R utilizado en esta sección, la función *gamkm2d2l* (función de R 4.4) que calcula el estimador en un modelo GAM corrigiendo el efecto de la censura con los pesos Kaplan-Meier para el caso de dos regresores.

Se han realizado simulaciones adicionales, para tres niveles de censura (10 %, 25 % y 40 %) y tres tamaños de muestra (n=200, n=500 y n=1000) en los casos de un único regresor y de dos regresores, con resultados análogos a los presentados en la sección 3.1.1, es decir, la estimación mejora a medida que aumenta el tamaño muestral y el efecto del nivel de censura es el esperado: los resultados son más precisos con niveles de censura más bajos y la variabilidad aumenta con la censura.

Figura 3.6: Extensión a modelos GAM: media de las funciones estimadas



3.1.3. Aplicación empírica: trasplantes cardíacos de Stanford

El conjunto de datos *Stanford Heart Transplant* contiene información sobre pacientes que han recibido un trasplante de corazón. Este programa de trasplantes comenzó en octubre de 1967 y se realizó el seguimiento de los pacientes hasta febrero de 1980. Este conjunto de datos se ha utilizado anteriormente, por ejemplo, en Miller and Halpern (1982) y Escobar and Meeker (1992), en modelos de regresión paramétrica censurada. El conjunto de datos proporciona información sobre el tiempo de supervivencia observado desde la fecha de trasplante, en días, para 184 pacientes, además de información sobre distintas características de los mismos: edad en el momento del primer trasplante, en años, un indicador del estado del paciente (vivo o muerto) en febrero de 1980, ... Como en Miller and Halpern (1982), se ha restringido la muestra a pacientes trasplantados de corazón que sobreviven al menos 10 días (176 pacientes). El conjunto de datos puede consultarse en Miller and Halpern (1982) y también puede descargarse del paquete R *survival* (Therneau, 2015; R Core Team, 2015).

Miller and Halpern (1982) y Escobar and Meeker (1992) analizan la asociación entre la covariable edad del paciente y el tiempo de supervivencia llegando a la conclusión de que es razonable suponer una relación cuadrática entre el tiempo de supervivencia en logaritmos, (T), y la $Edad$:

$$T = \beta_1 + \beta_2 Edad + \beta_3 Edad^2 + \epsilon \tag{3.1}$$

Asumiendo que la especificación paramétrica anterior es correcta, se pueden utilizar dos metodologías de estimación propuestas en la literatura del análisis de supervivencia para ajustar el modelo (3.1). Estos estimadores pueden utilizarse como referencia para evaluar el desempeño del método P-splines censurado propuesto en la sección 2.1.2. El primer enfoque, más restrictivo, son los modelos de duración acelerada (*AFT models,* Kalbfleisch and Prentice, 2002), modelos donde se supone una distribución de probabilidad conocida de la variable de respuesta y que permiten estimar los coeficientes β del modelo mediante máxima verosimilitud. Así,

Tabla 3.1: Estimación de los coeficientes de regresión y sus desviaciones típicas utilizando los métodos AFT lognormal y de Stute

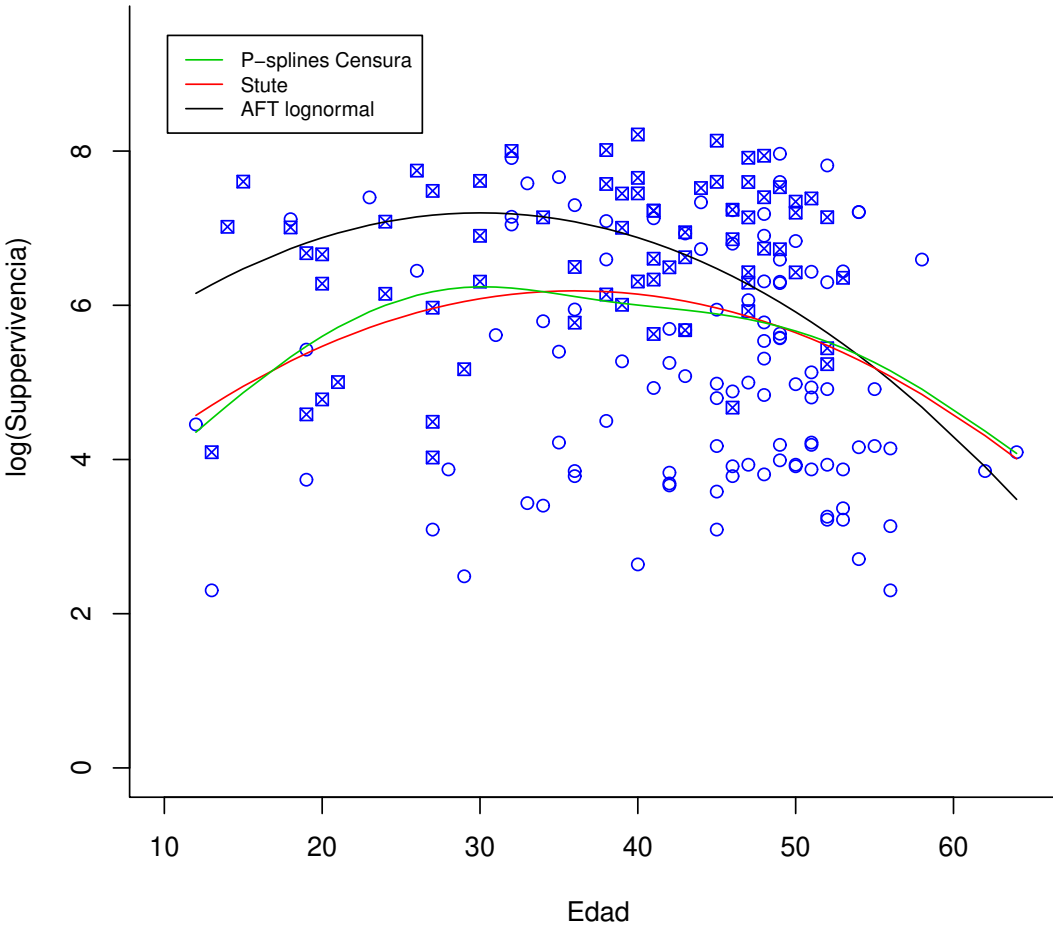
	Intercepto		*Edad*		$Edad^2$	
Método	$\hat{\beta}_1$	sdev($\hat{\beta}_1$)	$\hat{\beta}_2$	sdev($\hat{\beta}_2$)	$\hat{\beta}_3$	sdev($\hat{\beta}_3$)
AFT lognormal	4.3010	1.6529	0.1931	0.0903	-0.0032	0.0012
Stute	2.5574	1.7171	0.2013	0.0886	-0.0028	0.0011

considerando un modelo AFT lognormal, se va a proceder a estimar los coeficientes β suponiendo una distribución de probabilidad normal, lo que podríamos considerar como un método paramétrico censurado de estimación. La segunda metodología, propuesta por Stute (1993), es menos restrictiva en el sentido de que no necesita ningún supuesto sobre la distribución de probabilidad de la variable de respuesta, pero también se basa en la forma funcional presentada en la ecuación (3.1). Es decir, se necesita conocer la forma funcional que relaciona la variable respuesta y la covariable. Por tanto, puede clasificarse como un método semiparamétrico censurado de estimación. Esta última metodología estima los coeficientes mediante mínimos cuadrados ponderados utilizando como ponderaciones los pesos Kaplan-Meier (Stute, 1993). Los resultados, para la especificación cuadrática supuesta, de las estimaciones utilizando el método AFT lognormal y el método de Stute se presentan en la tabla 3.1. Como se puede observar, los coeficientes estimados de los efectos lineales y cuadráticos son muy similares para ambos métodos de estimación. La validez de los resultados de la estimación para estos dos enfoques se basa en la confianza en que la relación especificada, ecuación (3.1), es correcta. Si este no fuera el caso, ambos enfoques conducirían a conclusiones erróneas.

Como una solución robusta ante una mala especificación de la forma funcional se estima el modelo utilizando el método P-splines censurado. Así, la función estimada se obtiene al minimizar la ecuación (2.5) con B-splines cúbicos y un término de penalización de orden dos. Esta propuesta es más flexible que los mencionadas anteriormente, ya que no asume ni una distribución de probabilidad concreta para la variable respuesta, ni una forma funcional específica para la relación entre edad y supervivencia. Este último enfoque puede considerarse como un método de estimación no paramétrico censurado. También se ha estimado la relación con la función *gam*, incorporando la corrección para tener en cuenta el efecto de la censura y con el parámetro de suavizado y el número de nodos elegidos utilizando las propuestas de la sección 2.2, con resultados indistinguibles (no mostrados). La figura 3.7 muestra las estimaciones de estos tres enfoques en el diagrama de dispersión del logaritmo del tiempo de supervivencia observado frente a la edad. Los pacientes representados con ◯ han muerto y los indicados con ⊠ están vivos en febrero de 1980; es decir, estos últimos tienen tiempos de supervivencia censurados. Sesenta y nueve pacientes están vivos y ciento siete muertos en febrero de 1980.

En conclusión, la metodología AFT y la propuesta de Stute pueden aplicarse sólo cuando se conoce con exactitud la forma funcional del efecto de la covariable X sobre la variable de respuesta. En esta aplicación parece que la relación entre la supervivencia logarítmica y la edad es cuadrática, por lo que ambas metodologías funcionan razonablemente bien. Sin embargo, el enfoque no paramétrico P-splines

Figura 3.7: Relación estimada mediante tres metodologías: AFT lognormal, enfoque de Stute y P-splines censurado



censurado estima adecuadamente la relación cuadrática, obteniendo resultados muy similares a los anteriores. No obstante, si la forma funcional o la distribución de probabilidad se eligen erróneamente, estos dos métodos conducirían a una especificación incorrecta del modelo y, por tanto, a conclusiones incorrectas. Una ventaja importante del enfoque propuesto es que no necesita asumir ninguna forma funcional y, por lo tanto, evita este problema.

3.2. Resultados selección de parámetros

En esta sección se utiliza un estudio de simulación para analizar el comportamiento de las propuestas de estimación presentadas en las secciones 2.1.2 y 3.1.2, es decir, el estimador P-splines censurado (ckmPS) y el estimador en un modelo GAM corrigiendo el efecto de la censura con los pesos Kaplan-Meier (ckmGAM), para las diferentes propuestas de selección de parámetros presentadas en la sección 2.2. Para ello se consideran cinco ejemplos diferentes (tabla 3.2) para la relación entre la variable de respuesta, que se denota como T, y una covariable relevante X

$$t_i = f(x_i) + \epsilon_i,$$

Tabla 3.2: Ejemplos de forma funcional

Ejemplo	x_i	$f(x_i)$	ϵ_i
Cuadrática	$x_i \sim U[0,4]$	$2 + 4x_i - x_i^2$	$\epsilon_i \sim N(0, 0{,}50)$
Bump	$x_i \sim U[0,1]$	$2 + x_i + 2exp[-\{16(x_i - 0{,}5)\}^2]$	$\epsilon_i \sim N(0, 0{,}25)$
Logística	$x_i \sim U[0,1]$	$2 + \frac{1}{1 + exp\{-20(x_i - 0{,}5)\}}$	$\epsilon_i \sim N(0, 0{,}20)$
Sinusoidal 2	$x_i \sim U[0,10]$	$2 + exp\{sin(x_i)\}$	$\epsilon_i \sim N(0, 0{,}30)$
Sinusoidal 3	$x_i \sim U[0,10]$	$2 + exp\{sin(1{,}6x_i)\}$	$\epsilon_i \sim N(0, 0{,}30)$

donde los valores de la variable T no se conocen completamente porque algunas observaciones están censuradas.

Para estudiar el efecto de los datos censurados en la estimación se utiliza una variable de censura C generada a partir de una distribución uniforme $U(1, b)$. El valor del parámetro b cambia para considerar tres niveles distintos de censura: 10 %, 25 % y 40 %. Por tanto, se observa una muestra de tamaño n, $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$, donde $y_i = min(t_i, c_i)$ es el tiempo de supervivencia observado, es decir, el mínimo entre el tiempo de supervivencia, t_i, y el valor de censura, c_i. Además, se conoce a través de la variable indicadora $\delta_i = I(t_i \leq c_i)$ qué observaciones no están censuradas. En las simulaciones se consideran tres tamaños de muestra: $n = 200$, $n = 500$ y $n = 1000$. Para cada uno de estos nueve escenarios (véase la tabla 3.3) se utilizan 1000 réplicas de Monte Carlo.

Como es habitual en la práctica se considera que la forma funcional de la relación entre la variable respuesta y la covariable es desconocida y se estima el modelo utilizando dos métodos diferentes: el método P-splines censurado (ckmPS) y el estimador en un modelo GAM pero ponderado por los pesos Kaplan-Meier para tener en cuenta el efecto de la censura (ckmGAM). El valor óptimo del parámetro de suavizado λ se obtiene minimizando el estadístico GVC_c (véase la expresión 2.9) para dos valores diferentes del parámetro de ajuste ($\phi = 1$ o $\phi = 1{,}5$) y utilizando las ponderaciones de Kaplan-Meier (w_i) o sus cuadrados (w_i^2) para tener en cuenta el efecto de la censura. Por tanto, hay cuatro expresiones diferentes del criterio GCV_c dependiendo de los valores elegidos para ϕ y w_i. Para elegir el número de nodos (K) aplicamos la fórmula por defecto presentada en Ruppert (2002) basada en la elección por defecto de Wand (K_{rp}, ecuación 2.10) o la modificación propuesta en la subsección 2.2.2 (K_c, ecuación 2.11). Por último, con respecto a la ubicación de los nodos (L), se proponen nodos uniformemente espaciados (L_{eq}) o vectores de nodos no uniformes con el espaciado de los nodos en función de los pesos de Kaplan-Meier (L_{km}, subsección 2.2.2).

Tabla 3.3: Nueve escenarios: tres tamaños de muestra por tres niveles de censura

Tamaño muestral (n)	Nivel de censura (C)	Escenario (s)	Descripción
200	10 %	1	n=200 & C=10 %
	25 %	2	n=200 & C=25 %
	40 %	3	n=200 & C=40 %
500	10 %	4	n=500 & C=10 %
	25 %	5	n=500 & C=25 %
	40 %	6	n=500 & C=40 %
1000	10 %	7	n=1000 & C=10 %
	25 %	8	n=1000 & C=25 %
	40 %	9	n=1000 & C=40 %

Como resultado de todas estas posibles elecciones (ϕ, exponente de w_i, número y ubicación de los nodos) hay dieciséis resultados posibles para cada uno de los dos estimadores. En las simulaciones se calculan todas estas posibles estimaciones para cada una de las mil réplicas del conjunto de datos en cada uno de los nueve escenarios (véase la tabla 3.3) en cada ejemplo (véase la tabla 3.2) . Como medida de la bondad del ajuste se calcula el error cuadrático medio (ECM) para cada replica del conjunto de datos ($j = 1, 2, , \ldots, 1000$) en cada uno de los nueve escenarios ($s = 1, 2, , \ldots, 9$) para cada ejemplo. El ECM se define como:

$$ECM_{(j,s)} = \frac{\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2}{n} \qquad j = 1, 2, , \ldots, 1000 \qquad s = 1, 2, , \ldots, 9$$

Para sintetizar el rendimiento del estimador en cada escenario se calcula la media aritmética de los errores cuadráticos medios (MECM):

$$MECM_s = \frac{1}{1000}\sum_{j=1}^{1000} ECM_{(j,s)} \qquad s = 1, 2, , \ldots, 9$$

Para resumir el comportamiento de los estimadores propuestos con las diferentes elecciones de parámetros (ϕ, exponente de w_i, número y ubicación de los nodos) se propone una medida global relativa de la precisión que tiene en cuenta el comportamiento conjunto de cada estimador en los nueve escenarios. Para ello, la MECM del estimador propuesto se compara con la MECM óptima en cada escenario ($MECM_{s-opt}$) y se agregan los resultados de todos los escenarios:

$$G = \sum_{s=1}^{9} \frac{MECM_s - MECM_{s-opt}}{MECM_{s-opt}}$$

Las tablas 3.4 a 3.8 resumen los resultados para cada ejemplo presentando los diferentes estimadores ordenados de mejor a peor (de menor a mayor valor de G). Por ejemplo, la primera fila de la tabla 3.4 da los resultados de el estimador ckmPS utilizando el criterio GCV_c con un valor ϕ de 1,5 y un exponente para los pesos Kaplan-Meier de 1 junto a un número de nodos equidistantes (L_{eq}) igual a K_c. Las nueve últimas columnas presentan los valores de la media aritmética de los errores

cuadráticos medios (MECMs, multiplicadas por 1000) del estimador en los distintos escenarios. En ocho de los escenarios el valor obtenido coincide con el MECM óptimo (MECM mínimo resaltado en gris oscuro) y en el restante escenario cinco, el valor está muy próximo a ese valor mínimo (coloreado en gris claro para diferencias inferiores al 5 % con respecto al valor óptimo en ese escenario). Como medida global del comportamiento del estimador en los nueve escenarios se obtiene un valor de G de 0,25, muy próximo al $G = 0$ que se obtendría si fuera el mejor estimador en todos los escenarios.

Tabla 3.4: Resultados del estudio de simulación para la función cuadrática

						(1000 · MECM) en cada escenario								
						s=1	s=2	s=3	s=4	s=5	s=6	s=7	s=8	s=9
	GCV_c		Knots			10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
Estimador	ϕ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmPS	1.5	1	L_{eq}	K_c	0.25	8.9	10	15.3	3.9	4.6	6.9	2.1	2.5	3.9
ckmGAM	1.5	1	L_{eq}	K_c	1.22	8.9	10.4	15.7	3.9	4.5	7.2	2.1	2.5	3.9
ckmGAM	1.5	1	L_{eq}	K_{rp}	2.24	8.9	10	16.2	3.9	4.6	7.2	2.1	2.5	4.2
ckmPS	1.5	1	L_{eq}	K_{rp}	2.33	8.9	10	16.1	3.9	4.6	7.3	2.1	2.5	4.2
ckmPS	1.5	2	L_{eq}	K_c	9.47	8.9	10.2	18.6	3.9	4.7	8.9	2.1	2.5	5
ckmGAM	1.5	2	L_{eq}	K_c	10.56	8.9	10.6	19	3.9	4.6	9	2.1	2.6	5
ckmPS	1.5	2	L_{eq}	K_{rp}	16.54	8.9	10.3	20.6	3.9	4.8	10.4	2.1	2.5	6
ckmGAM	1.5	2	L_{eq}	K_{rp}	16.78	8.9	10.3	20.7	3.9	4.8	10.5	2.1	2.5	6
ckmPS	1.5	1	L_{km}	K_c	28.62	14.8	15.6	20.7	5	5.5	7.8	2.4	2.8	4.3
ckmGAM	1.5	1	L_{km}	K_c	31.10	15.4	16.4	19.9	5.1	5.4	7.9	2.6	2.9	4.2
ckmPS	1	1	L_{eq}	K_c	36.43	9.9	13.8	26.9	4.4	6.2	11.1	2.3	3.2	6
ckmGAM	1	1	L_{eq}	K_c	36.61	10	14.4	26.7	4.2	5.8	11.5	2.3	3.3	6
ckmPS	1.5	2	L_{km}	K_c	37.68	14.9	16	23.4	5	5.7	9.5	2.4	2.9	5.3
ckmGAM	1.5	2	L_{km}	K_c	39.62	15.5	16.6	22.4	5.1	5.6	9.5	2.6	3	5.2
ckmGAM	1.5	1	L_{km}	K_{rp}	39.88	16.4	17.4	22.7	5.1	5.8	8.4	2.6	3.1	4.8
ckmPS	1.5	1	L_{km}	K_{rp}	40.11	15.6	17	24.5	5.1	5.9	8.9	2.5	3	4.9
ckmPS	1	2	L_{eq}	K_c	49.85	9.9	14.3	31.6	4.4	6.6	13.5	2.3	3.4	7.3
ckmGAM	1	2	L_{eq}	K_c	51.09	10.1	15	31.2	4.2	6.2	13.9	2.4	3.6	7.3
ckmGAM	1.5	2	L_{km}	K_{rp}	52.44	16.5	17.7	26.2	5.2	6	11.2	2.6	3.1	6.3
ckmPS	1.5	2	L_{km}	K_{rp}	54.61	15.7	17.4	28.8	5.1	6	12	2.5	3.1	6.7
ckmGAM	1	1	L_{km}	K_c	56.97	15.3	18.5	28.2	5.1	6.4	11.7	2.7	3.6	6.1
ckmPS	1	1	L_{km}	K_c	62.6	15.2	19.6	31	5.4	7.1	11.8	2.6	3.6	6.2
ckmGAM	1	1	L_{eq}	K_{rp}	63.63	10	14.5	40.8	4.4	6.5	15.9	2.3	3.4	8.4
ckmPS	1	1	L_{eq}	K_{rp}	63.68	10	14.5	41.1	4.4	6.5	15.8	2.3	3.4	8.4
ckmGAM	1	2	L_{km}	K_c	68.54	15.3	19.1	31.6	5.1	6.7	13.8	2.7	3.8	7.3
ckmPS	1	2	L_{km}	K_c	73.78	15.3	20	34.1	5.4	7.5	13.7	2.6	3.8	7.4
ckmGAM	1	1	L_{km}	K_{rp}	83.82	16	19.6	39.1	5.3	7.3	15.9	2.7	3.7	8.5
ckmPS	1	1	L_{km}	K_{rp}	97.17	15.9	22.1	46.5	5.5	7.8	17	2.7	3.9	8.8
ckmPS	1	2	L_{eq}	K_{rp}	98.06	10	15.3	62	4.4	7	20.8	2.3	3.7	11.1
ckmGAM	1	2	L_{km}	K_{rp}	106.5	16.1	20.2	46.7	5.3	7.7	20.2	2.7	4	11
ckmPS	1	2	L_{km}	K_{rp}	120.93	15.9	22.9	54.5	5.5	8.3	21.4	2.7	4.2	11.4
ckmGAM	1	2	L_{eq}	K_{rp}	124.09	10	15.2	98	4.4	7	20.8	2.3	3.7	11.1

El análisis de los resultados muestra que para los dos estimadores analizados (estimadores ckmPS y ckmGAM) la mejor elección de parámetros viene dada por un GCV_c con un valor de ϕ de 1,5 y un exponente de los pesos Kaplan-Meier de 1 más un número K_c de nodos equidistantes. Las figuras 3.8 a 3.17 muestran un resumen gráfico de los resultados de estos dos estimadores con la elección óptima de parámetros. Las subfiguras (a) a (c) de las figuras 3.8 a 3.17 presentan los gráficos de caja con los resultados del error cuadrático medio (ECM) para cada escenario en cada uno de los cinco ejemplos. El error cuadrático medio disminuye cuando aumenta el tamaño de muestra para cada nivel de censura considerado. El efecto del nivel de censura es el esperado: los resultados son más más precisos con niveles de censura más bajos y la variabilidad aumenta con el nivel de censura. Las subfiguras (d) a (l) de estas figuras 3.8 a 3.17 presentan la función verdadera $f(x)$ y la media y los límites superior e inferior del 95 % de los valores estimados utilizando estos estimadores para

Tabla 3.5: Resultados del estudio de simulación para la función bump

						(1000 · MECM) en cada escenario								
						s=1	s=2	s=3	s=4	s=5	s=6	s=7	s=8	s=9
	GCV_c		Knots			10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
Estimador	ϕ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmPS	1.5	1	L_{eq}	K_c	0.17	6.5	8.1	13.4	2.8	3.4	4.9	1.5	1.8	2.5
ckmGAM	1.5	1	L_{eq}	K_c	0.17	6.5	8.1	13.4	2.8	3.4	4.9	1.5	1.8	2.5
ckmGAM	1.5	1	L_{eq}	K_{rp}	2.78	6.5	8.1	13.5	2.9	3.5	5.3	1.5	1.8	2.7
ckmPS	1.5	1	L_{eq}	K_{rp}	3.52	6.5	8.1	13.5	2.9	3.5	5.3	1.6	1.8	2.7
ckmPS	1.5	2	L_{eq}	K_c	4.78	6.4	8.3	16	2.8	3.5	5.4	1.5	1.8	2.7
ckmGAM	1.5	2	L_{eq}	K_c	4.95	6.4	8.3	16.2	2.8	3.5	5.4	1.5	1.8	2.7
ckmPS	1	1	L_{eq}	K_c	7.57	6.6	9.2	17.4	2.9	3.6	5.3	1.5	1.8	2.6
ckmGAM	1	1	L_{eq}	K_c	8.16	6.6	9.2	18.1	2.9	3.6	5.3	1.5	1.8	2.6
ckmPS	1.5	2	L_{eq}	K_{rp}	13.78	6.5	8.3	19.6	2.8	3.5	6.8	1.5	1.8	3.3
ckmGAM	1.5	2	L_{eq}	K_{rp}	14.67	6.5	8.3	20.4	2.8	3.5	6.9	1.5	1.8	3.3
ckmPS	1.5	1	L_{km}	K_c	16.10	7.5	9.2	14.5	3.5	4.2	5.6	1.8	2	2.8
ckmPS	1.5	2	L_{km}	K_c	17.43	7.4	9.1	15.4	3.5	4.2	5.8	1.8	2	2.9
ckmPS	1	2	L_{eq}	K_c	17.52	6.6	10	24.1	2.9	3.8	5.8	1.5	1.9	2.8
ckmPS	1	1	L_{km}	K_c	20.42	7.5	10	16.7	3.5	4.3	5.8	1.8	2.1	2.8
ckmGAM	1	1	L_{km}	K_c	20.62	7.6	9.7	15.6	3.7	4.4	5.9	1.8	2.1	2.8
ckmPS	1.5	1	L_{km}	K_{rp}	21.27	7.6	9.3	16.3	3.6	4.3	6.1	1.8	2.1	3
ckmGAM	1	2	L_{eq}	K_c	21.83	6.6	10	29.3	2.9	3.8	5.8	1.5	1.9	2.8
ckmGAM	1	2	L_{km}	K_c	22.33	7.5	9.8	16.9	3.7	4.4	6	1.8	2.1	2.9
ckmPS	1	2	L_{km}	K_c	24.19	7.5	10.2	18.9	3.5	4.4	6	1.8	2.1	3
ckmGAM	1.5	2	L_{km}	K_c	26.14	8	9.5	15.5	4.3	4.6	6	2	2.1	2.9
ckmPS	1.5	2	L_{km}	K_{rp}	27.23	7.5	9.3	19.3	3.5	4.3	7.1	1.8	2.1	3.4
ckmGAM	1.5	1	L_{km}	K_c	28.4	8.1	9.7	15.4	4.4	4.8	6.1	2	2.2	2.9
ckmGAM	1.5	1	L_{km}	K_{rp}	35.22	8.2	9.9	16.8	4.5	5.1	6.8	2.1	2.3	3.1
ckmGAM	1.5	2	L_{km}	K_{rp}	35.71	8.1	9.7	18.8	4.4	4.8	7.1	2	2.2	3.4
ckmGAM	1	1	L_{km}	K_{rp}	37.25	7.6	10.7	24.6	3.8	4.6	7.1	1.9	2.2	3.4
ckmPS	1	1	L_{km}	K_{rp}	37.51	7.5	10.4	27.4	3.5	4.6	7.3	1.8	2.2	3.4
ckmGAM	1	2	L_{km}	K_{rp}	48.12	7.6	10.5	33.4	3.8	4.7	8	1.8	2.2	3.9
ckmPS	1	1	L_{eq}	K_{rp}	50.95	6.7	12	54	2.9	3.9	7.2	1.5	1.9	3.3
ckmPS	1	2	L_{km}	K_{rp}	51.58	7.5	10.7	37.6	3.5	4.7	8.2	1.8	2.3	3.9
ckmGAM	1	1	L_{eq}	K_{rp}	52.77	6.7	12.6	55.2	2.9	3.9	7.2	1.5	1.9	3.3
ckmPS	1	2	L_{eq}	K_{rp}	86.39	6.7	12.6	84.9	2.9	4.1	9.1	1.6	2	3.9
ckmGAM	1	2	L_{eq}	K_{rp}	106.31	6.7	10.4	112.3	2.9	4.1	9.2	1.6	2	3.9

cada nivel de censura y tamaño de muestra. A partir de estos resultados se puede concluir que ambos enfoques estiman adecuadamente la relación en todos los casos considerados: la función estimada recupera la verdadera forma funcional.

En los cuarenta y cinco escenarios analizados (nueve escenarios para cinco ejemplos) sólo hay dos en los que la combinación anterior ($\phi = 1{,}5$, w_i, L_{eq}, K_c) no es la mejor. La primera excepción se da para la función bump en el escenario con pocos datos ($n = 200$) y un nivel de censura bajo (C=10 %), donde un GCV_c con un exponente de dos en los pesos Kaplan-Meier presenta resultados ligeramente mejores. En algunos casos, cuando el porcentaje de censura es pequeño y la muestra no es muy grande las observaciones censuradas se acumulan en zonas donde la supervivencia es elevada, lo que da lugar a un ligero sesgo a la baja en esas zonas. Un exponente de dos para los pesos Kaplan-Meier en el GCV_c puede ayudar a reducir el sesgo en estos picos de la función. La segunda excepción se da para la función sinusoidal con tres ciclos en el escenario con pocos datos ($n = 200$) y un nivel de censura alto (C=40 %), donde el número de nodos propuesto por Ruppert (K_{rp}) funciona un poco mejor que la propuesta K_c. Dado que el perfil de la función es bastante complejo, en un escenario con poca información (pocos datos y alto nivel de censura) un mayor número de nodos se adapta mejor a la situación.

No todas las elecciones de parámetros tienen la misma relevancia a la hora de obtener una buena estimación. Para analizar la importancia de cada una de las opciones en las figuras 3.18 a 3.22 se comparan la MECM de cada escenario (la

misma información que en las tablas 3.4 a 3.8) para cada uno de los cinco ejemplos, según se utiliza el estimador ckmPS frente al ckmGAM (subfigura a), un valor de ϕ de 1 frente a 1,5 (subfigura b), una ponderación de uno para los pesos Kaplan-Meier, w_i, en el numerador del criterio GGV_c frente a su cuadrado, w_i^2, (subfigura c), un número de nodos K_c frente a K_{rp} (subfigura d) y, por último, en la subfigura (e), nodos equidistantes (L_{eq}) frente a un vector de nodos no uniforme con el espaciado de los mismos en función de los pesos Kaplan-Meier (L_{km}).

La selección del método de estimación no parece demasiado importante: los dos métodos analizados se comportan de forma muy similar y dan resultados muy parecidos (véanse las subfiguras (a) de las figuras 3.18 a 3.22). Por otro lado, la elección del parámetro ϕ es de gran importancia, y está importancia crece con la censura. De forma análoga a lo encontrado en la literatura para el caso no censurado, la elección de $\phi = 1{,}5$ es mejor en casi todas las situaciones que $\phi = 1{,}0$, y la diferencia aumenta a medida que aumenta la censura (véanse las subfiguras (b) de las figuras 3.18 a 3.22). En cuanto a la elección del exponente de los pesos Kaplan-Meier en el criterio GCV_c, si la censura no es muy grande ambos exponentes funcionan de forma similar (véanse las subfiguras (c) de las figuras 3.18 a 3.22), pero cuando la censura es grande (C=40 %) el exponente 1 (w_i^1) es claramente mejor que 2 (w_i^2). Ocurre algo similar con el número de nodos: no es muy importante si la censura es pequeña, pero a medida que aumenta la censura la elección de K_c es claramente mejor que la propuesta de Ruppert, K_{rp}, (véanse las subfiguras (d) en las figuras 3.18 a 3.22). Por último, con respecto a la ubicación de los nodos, los nodos equidistantes (L_{eq}) funcionan claramente mejor, excepto en muestras de tamaño pequeño ($n = 200$) con censura grande (C=40 %), donde el espaciado de los nodos en función de los pesos Kaplan-Meier (L_{km}) en ocasiones reduce los errores cuadráticos medios más grandes.

Como conclusión, se encuentra que en los cinco ejemplos estudiados la elección de parámetros que genera la mejor estimación es la de $\phi = 1{,}5$ y exponente 1 de los pesos Kaplan-Meier en el criterio GCV_c y K_c nodos equidistantes. A medida que aumenta el nivel de censura, incluso con muestras de gran tamaño, también empiezan a encontrarse diferencias importantes entre los resultados de las distintas especificaciones. Por tanto, si la censura es grande es más importante seleccionar los parámetros de forma óptima. Con respecto al tamaño de la muestra, las mayores diferencias se producen con $n = 200$. Es interesante observar que las elecciones más importantes son $\phi = 1{,}5$ y nodos equidistantes, porque dan lugar a estimadores con un rendimiento muy bueno independientemente de la elección de los otros parámetros, que no es tan decisiva. Por otro lado, las peores combinaciones de parámetros se presentan, en general, para valores de $\phi = 1{,}0$ y número de nodos elegido siguiendo la propuesta de Ruppert, propuestas habituales en la literatura para datos no censurados, independientemente de la elección del resto de parámetros.

Estos resultados son robustos a cambios en la variabilidad o en la distribución del término de error. Se han realizado simulaciones adicionales con una varianza mayor de la perturbación aleatoria y considerando distribuciones de error no normales y asimétricas, como la distribución de Weibull. Los nuevos resultados obtenidos (no mostrados) confirman el buen comportamiento del marco propuesto y son coherentes con los presentados en esta sección. Si se aumenta la varianza (disminuyendo la relación señal/ruido en un 50 %) se mantienen los resultados anteriores. Lo más destacable es que aumenta la importancia de la elección de los parámetros, ya que las

Tabla 3.6: Resultados del estudio de simulación para la función logística

						(1000 · MECM) en cada escenario								
						s=1	s=2	s=3	s=4	s=5	s=6	s=7	s=8	s=9
	GCV_c		Knots			10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
Estimador	ϕ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmPS	1.5	1	L_{eq}	K_c	0	1.8	2.2	3.3	0.8	1	1.3	0.4	0.5	0.7
ckmGAM	1.5	1	L_{eq}	K_c	0	1.8	2.2	3.3	0.8	1	1.3	0.4	0.5	0.7
ckmPS	1.5	1	L_{eq}	K_{rp}	0.85	1.8	2.2	3.3	0.8	1	1.4	0.4	0.5	0.7
ckmGAM	1.5	1	L_{eq}	K_{rp}	0.85	1.8	2.2	3.3	0.8	1	1.4	0.4	0.5	0.7
ckmPS	1.5	2	L_{eq}	K_c	5.32	1.8	2.2	3.9	0.8	1	1.5	0.4	0.5	0.8
ckmGAM	1.5	2	L_{eq}	K_c	5.32	1.8	2.2	3.9	0.8	1	1.5	0.4	0.5	0.8
ckmPS	1.5	2	L_{eq}	K_{rp}	6.17	1.8	2.2	3.9	0.8	1	1.6	0.4	0.5	0.8
ckmGAM	1.5	2	L_{eq}	K_{rp}	6.17	1.8	2.2	3.9	0.8	1	1.6	0.4	0.5	0.8
ckmPS	1.5	1	L_{km}	K_c	17	2.1	2.5	3.4	1	1.2	1.5	0.5	0.6	0.8
ckmPS	1.5	1	L_{km}	K_{rp}	19.38	2.1	2.5	3.6	1	1.2	1.7	0.5	0.6	0.8
ckmPS	1.5	2	L_{km}	K_c	19.72	2.1	2.5	3.7	1	1.2	1.7	0.5	0.6	0.8
ckmGAM	1.5	1	L_{km}	K_c	22.9	2.2	2.5	3.4	1.1	1.3	1.5	0.6	0.6	0.8
ckmPS	1.5	2	L_{km}	K_{rp}	24.36	2.1	2.5	4.1	1	1.2	1.9	0.5	0.6	0.9
ckmGAM	1.5	2	L_{km}	K_c	25.95	2.2	2.5	3.8	1.1	1.3	1.7	0.6	0.6	0.8
ckmGAM	1.5	1	L_{km}	K_{rp}	26.84	2.2	2.6	3.5	1.2	1.3	1.7	0.6	0.6	0.8
ckmGAM	1.5	2	L_{km}	K_{rp}	31.48	2.2	2.6	3.9	1.2	1.3	1.9	0.6	0.6	0.9
ckmPS	1	1	L_{eq}	K_c	34.05	2	2.9	6.5	0.9	1.2	1.9	0.5	0.6	1
ckmGAM	1	1	L_{eq}	K_c	34.38	2	2.9	6.6	0.9	1.2	1.9	0.5	0.6	1
ckmGAM	1	1	L_{km}	K_c	40.16	2.2	3	5.4	1.1	1.4	2	0.5	0.7	1
ckmPS	1	1	L_{km}	K_c	41.14	2.2	3.1	5.7	1	1.4	2.1	0.5	0.7	1
ckmPS	1	2	L_{eq}	K_c	48.25	2	3	8.6	0.9	1.3	2.1	0.5	0.7	1.1
ckmPS	1	2	L_{km}	K_c	52.11	2.2	3.2	7.5	1	1.5	2.3	0.5	0.7	1.1
ckmGAM	1	2	L_{km}	K_c	52.71	2.2	3.1	8	1.1	1.4	2.2	0.5	0.7	1.1
ckmGAM	1	2	L_{eq}	K_c	56	2	3	10.9	0.9	1.3	2.1	0.5	0.7	1.1
ckmGAM	1	1	L_{km}	K_{rp}	57.89	2.2	3.2	7.9	1.1	1.4	2.6	0.5	0.7	1.2
ckmPS	1	1	L_{km}	K_{rp}	62.8	2.2	3.3	7.9	1.1	1.5	2.8	0.5	0.7	1.3
ckmPS	1	1	L_{eq}	K_{rp}	67.24	2	3	12.5	0.9	1.3	2.6	0.5	0.7	1.2
ckmGAM	1	1	L_{eq}	K_{rp}	69.26	2	3	13.1	0.9	1.3	2.6	0.5	0.7	1.2
ckmGAM	1	2	L_{km}	K_{rp}	74.94	2.2	3.3	9.8	1.1	1.5	3.1	0.5	0.7	1.5
ckmPS	1	2	L_{km}	K_{rp}	78.96	2.2	3.4	9.6	1.1	1.6	3.2	0.5	0.8	1.5
ckmPS	1	2	L_{eq}	K_{rp}	104.66	2	3.2	21.1	0.9	1.3	3.1	0.5	0.7	1.4
ckmGAM	1	2	L_{eq}	K_{rp}	139.67	2	3.2	31.5	0.9	1.3	3.1	0.5	0.7	1.4

diferencias entre el rendimiento de los estimadores aumentan, incluso con censuras pequeñas. Además, el número de combinaciones que dan lugar a un buen estimador disminuye, por lo que la elección correcta de los parámetros adquiere aún mayor importancia. Si se cambia la distribución del error de Normal a Weibull se mantienen los resultados generales, pero el exponente dos para los pesos Kaplan-Meier en la expresión del criterio GCV_c aparece a veces entre las mejores opciones, especialmente con porcentajes de censura bajos.

Tabla 3.7: Resultados del estudio de simulación para la función sinusoidal con dos ciclos

						(1000 · MECM) en cada escenario								
						s=1	s=2	s=3	s=4	s=5	s=6	s=7	s=8	s=9
	GCV_c		Knots			10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
Estimador	ϕ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmGAM	1.5	1	L_{eq}	K_c	0	6.1	7.1	10	2.7	3.1	4.3	1.4	1.7	2.3
ckmPS	1.5	1	L_{eq}	K_c	0.18	6.2	7.1	10	2.7	3.1	4.3	1.4	1.7	2.3
ckmPS	1.5	2	L_{eq}	K_c	1.97	6.1	7.2	10.5	2.7	3.1	4.6	1.4	1.7	2.4
ckmGAM	1.5	2	L_{eq}	K_c	2.19	6.1	7.2	10.7	2.7	3.1	4.6	1.4	1.7	2.4
ckmPS	1.5	1	L_{eq}	K_{rp}	2.68	6.2	7.2	10.3	2.7	3.2	4.5	1.4	1.8	2.4
ckmGAM	1.5	1	L_{eq}	K_{rp}	2.68	6.2	7.2	10.3	2.7	3.2	4.5	1.4	1.8	2.4
ckmGAM	1.5	2	L_{eq}	K_{rp}	6.58	6.1	7.2	11.5	2.7	3.2	5	1.4	1.8	2.7
ckmPS	1.5	2	L_{eq}	K_{rp}	6.65	6.2	7.2	11.4	2.7	3.2	5	1.4	1.8	2.7
ckmPS	1.5	1	L_{km}	K_c	6.91	6.9	7.8	10.7	2.8	3.3	4.5	1.5	1.8	2.4
ckmGAM	1.5	1	L_{km}	K_c	7.61	6.7	7.7	10.5	2.8	3.3	4.5	1.6	1.9	2.4
ckmGAM	1.5	2	L_{km}	K_c	8.66	6.7	7.7	10.9	2.8	3.3	4.8	1.6	1.8	2.5
ckmPS	1.5	2	L_{km}	K_c	8.83	6.9	7.8	11.3	2.8	3.3	4.8	1.5	1.8	2.5
ckmGAM	1.5	1	L_{km}	K_{rp}	12.38	6.9	8.2	11.5	2.9	3.4	4.8	1.6	1.9	2.6
ckmPS	1.5	1	L_{km}	K_{rp}	13.38	7	8.2	12	2.9	3.4	4.9	1.6	1.9	2.6
ckmGAM	1.5	2	L_{km}	K_{rp}	15.42	6.9	8.2	12.2	2.9	3.4	5.3	1.6	1.9	2.8
ckmPS	1	1	L_{eq}	K_c	15.79	6.4	8.6	13.9	2.8	3.6	5.2	1.5	1.9	2.7
ckmGAM	1	1	L_{eq}	K_c	16.23	6.4	8.6	14.3	2.8	3.6	5.2	1.5	1.9	2.7
ckmPS	1.5	2	L_{km}	K_{rp}	16.69	7	8.2	13	2.9	3.4	5.5	1.5	1.9	2.9
ckmPS	1	2	L_{eq}	K_c	19.64	6.4	8.7	14.6	2.8	3.6	5.7	1.5	2	2.9
ckmGAM	1	1	L_{km}	K_c	19.8	6.9	9	13.9	2.9	3.7	5.3	1.6	2	2.7
ckmGAM	1	2	L_{eq}	K_c	19.97	6.4	8.7	14.9	2.8	3.6	5.7	1.5	2	2.9
ckmPS	1	1	L_{km}	K_c	21.01	7.1	9.3	14	2.9	3.7	5.4	1.6	2	2.7
ckmGAM	1	2	L_{km}	K_c	23.07	6.9	9.1	14.9	2.9	3.7	5.7	1.6	2	2.9
ckmPS	1	2	L_{km}	K_c	25.38	7.1	9.6	14.8	2.9	3.8	5.8	1.6	2.1	2.9
ckmPS	1	1	L_{eq}	K_{rp}	34.58	6.5	9.3	21.2	2.8	3.8	7	1.5	2	3.4
ckmGAM	1	1	L_{km}	K_{rp}	36.2	7.1	9.8	18.6	3	3.9	7	1.6	2.1	3.4
ckmGAM	1	1	L_{eq}	K_{rp}	38.69	6.5	9.3	24.9	2.8	3.8	7	1.5	2	3.4
ckmPS	1	1	L_{km}	K_{rp}	43.5	7.3	10.2	21.8	3	4	7.3	1.6	2.2	3.6
ckmGAM	1	2	L_{km}	K_{rp}	46	7.1	9.9	22.1	3	4	7.9	1.6	2.2	3.9
ckmPS	1	2	L_{km}	K_{rp}	51.84	7.3	10.5	24.3	3	4.1	8.2	1.6	2.2	4.1
ckmPS	1	2	L_{eq}	K_{rp}	56.38	6.5	9.4	35.5	2.8	3.9	7.9	1.5	2.1	3.9
ckmGAM	1	2	L_{eq}	K_{rp}	67.71	6.5	9.4	45.7	2.8	3.9	7.9	1.5	2.1	3.9

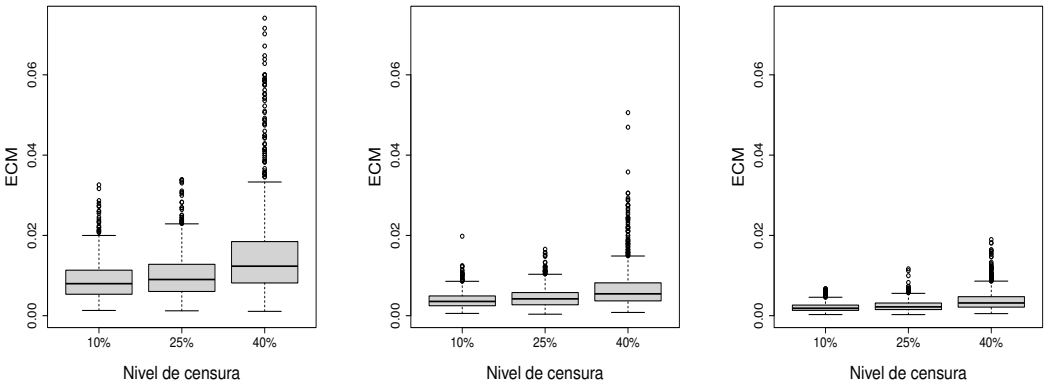
Tabla 3.8: Resultados del estudio de simulación para la función sinusoidal con tres ciclos

						(1000 · MECM) en cada escenario								
						s=1	s=2	s=3	s=4	s=5	s=6	s=7	s=8	s=9
	GCV_c		Knots			10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
Estimador	ϕ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmPS	1.5	1	L_{eq}	K_c	0.73	8.6	10.4	17.8	3.8	4.5	6.1	2.1	2.4	3.2
ckmGAM	1.5	1	L_{eq}	K_c	0.86	8.6	10.4	18	3.8	4.5	6.1	2.1	2.4	3.2
ckmPS	1.5	1	L_{eq}	K_{rp}	2.53	8.6	10.5	16.7	3.9	4.6	6.5	2.1	2.5	3.4
ckmGAM	1.5	1	L_{eq}	K_{rp}	2.53	8.6	10.5	16.7	3.9	4.6	6.5	2.1	2.5	3.4
ckmPS	1.5	2	L_{eq}	K_c	4.99	8.6	11.2	21.2	3.8	4.6	6.4	2.1	2.4	3.3
ckmGAM	1.5	2	L_{eq}	K_c	5.06	8.6	11.2	21.3	3.8	4.6	6.4	2.1	2.4	3.3
ckmPS	1	1	L_{eq}	K_c	11.39	8.8	12.1	25.4	3.9	4.9	6.7	2.1	2.5	3.4
ckmPS	1.5	1	L_{km}	K_c	11.56	9.9	12	17.8	4.4	5.2	6.8	2.3	2.6	3.4
ckmGAM	1	1	L_{eq}	K_c	11.79	8.8	12.1	26	3.9	4.9	6.7	2.1	2.5	3.4
ckmPS	1.5	2	L_{eq}	K_{rp}	12.55	8.6	11.2	26.8	3.8	4.6	7.3	2.1	2.5	3.8
ckmPS	1.5	2	L_{km}	K_c	14.88	9.8	12.9	20.2	4.4	5.2	7.1	2.3	2.6	3.5
ckmGAM	1.5	2	L_{eq}	K_{rp}	14.88	8.6	11.2	30.3	3.8	4.6	7.3	2.1	2.5	3.8
ckmGAM	1	1	L_{km}	K_c	15.77	9.9	12.8	19.9	4.5	5.4	7	2.3	2.7	3.5
ckmGAM	1.5	1	L_{km}	K_c	16.72	10.6	12.5	17.9	4.9	5.4	6.9	2.5	2.7	3.4
ckmPS	1.5	1	L_{km}	K_{rp}	17.79	10	12.5	19.6	4.5	5.4	7.5	2.3	2.8	3.8
ckmGAM	1.5	2	L_{km}	K_c	18.11	10.5	13	19.3	4.8	5.4	6.9	2.5	2.7	3.5
ckmPS	1	1	L_{km}	K_c	18.37	9.9	13.3	22	4.4	5.5	7.2	2.3	2.7	3.6
ckmPS	1	2	L_{eq}	K_c	19.93	8.8	14	31.7	3.9	5	7.2	2.1	2.6	3.6
ckmGAM	1	2	L_{km}	K_c	20.78	9.9	13.6	24	4.5	5.4	7.4	2.3	2.7	3.7
ckmGAM	1	2	L_{eq}	K_c	21.04	8.8	15.6	30.8	3.9	5	7.2	2.1	2.6	3.6
ckmPS	1.5	2	L_{km}	K_{rp}	22.67	10	13.2	22.6	4.5	5.4	8.1	2.3	2.8	4.1
ckmPS	1	2	L_{km}	K_c	23.17	9.9	14.3	25.3	4.4	5.5	7.6	2.3	2.8	3.7
ckmGAM	1.5	1	L_{km}	K_{rp}	23.96	11	13.3	19.4	5	5.8	7.6	2.5	2.9	3.8
ckmGAM	1.5	2	L_{km}	K_{rp}	27.01	10.9	13.8	21.6	5	5.7	8	2.5	2.9	4
ckmGAM	1	1	L_{km}	K_{rp}	32.12	10.2	13.8	28.7	4.6	5.9	8.9	2.3	2.9	4.4
ckmPS	1	1	L_{km}	K_{rp}	34.16	10.1	14.4	29.4	4.5	5.9	9.2	2.3	3	4.5
ckmPS	1	1	L_{eq}	K_{rp}	38.78	8.9	13	51.6	3.9	5.3	8.8	2.1	2.7	4.3
ckmGAM	1	2	L_{km}	K_{rp}	42.38	10.2	14.8	37.6	4.6	6	9.8	2.3	2.9	4.8
ckmPS	1	2	L_{km}	K_{rp}	43.64	10.1	15.6	36.7	4.5	6.1	10	2.3	3	4.9
ckmGAM	1	1	L_{eq}	K_{rp}	49.05	8.9	12.9	67.2	3.9	5.3	8.8	2.1	2.7	4.3
ckmPS	1	2	L_{eq}	K_{rp}	90.14	8.9	14.1	119.7	3.9	5.5	10	2.1	2.8	4.8
ckmGAM	1	2	L_{eq}	K_{rp}	119.09	8.9	14.1	163.2	3.9	5.5	10	2.1	2.8	4.8

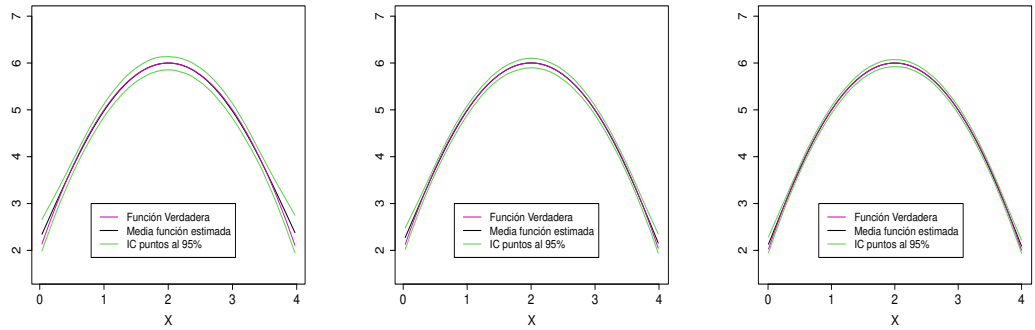
3.2.1. Estimaciones con la mejor elección de parámetros para los estimadores ckmPS y ckmGAM

Figura 3.8: Función estimada utilizando el estimador ckmPS para la función cuadrática: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes

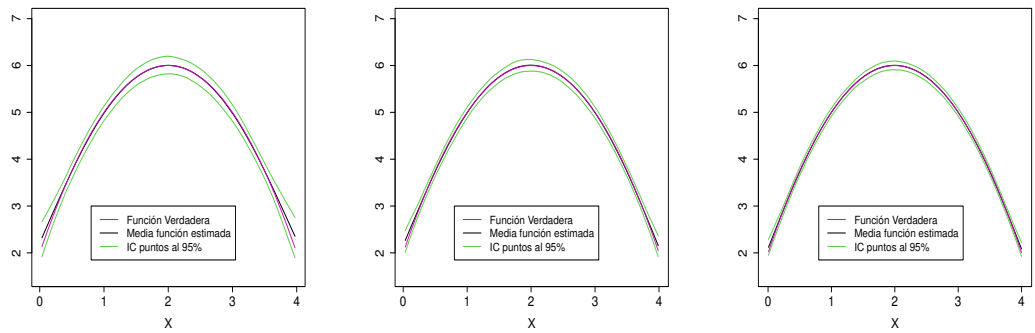
(a) n=200 (b) n=500 (c) n=1000



(d) C=10 and n=200 (e) C=10 and n=500 (f) C=10 and n=1000



(g) C=25 and n=200 (h) C=25 and n=500 (i) C=25 and n=1000



(j) C=40 and n=200 (k) C=40 and n=500 (l) C=40 and n=1000

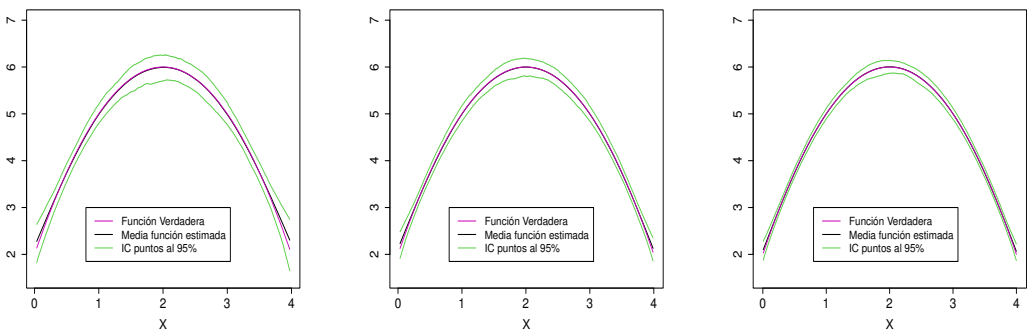


Figura 3.9: Función estimada utilizando el estimador ckmGAM para la función cuadrática: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes

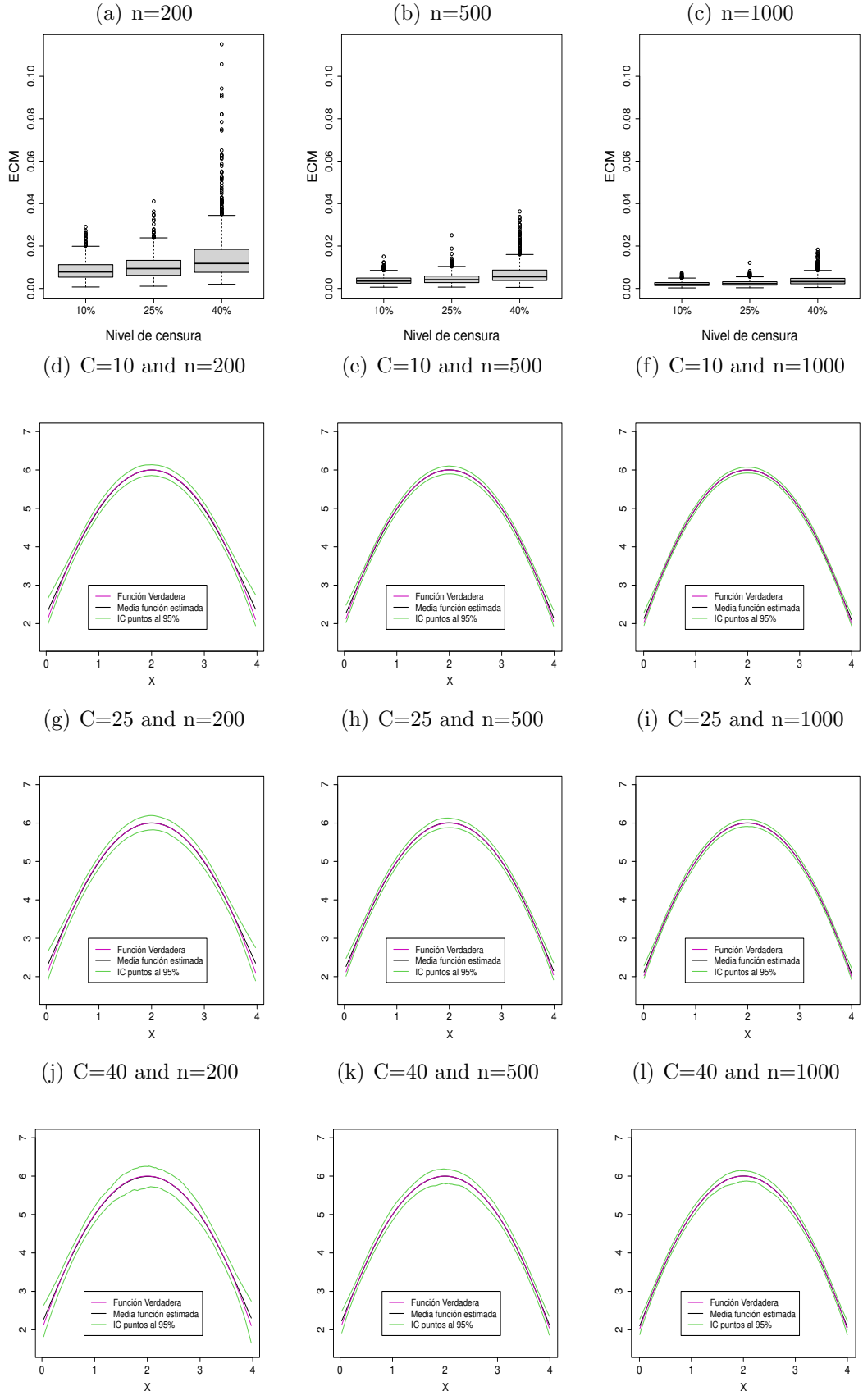
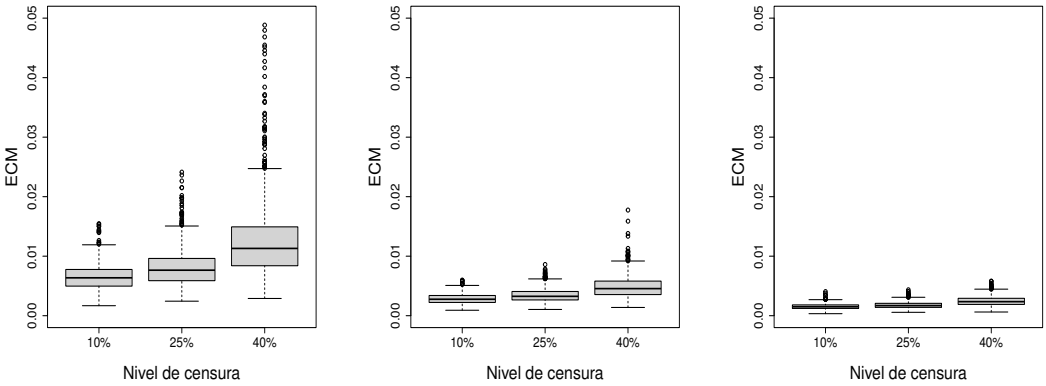
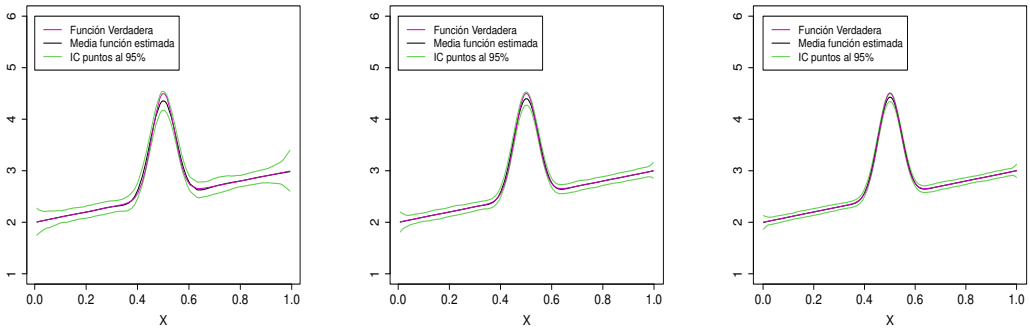


Figura 3.10: Función estimada utilizando el estimador ckmPS para la función bump: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes

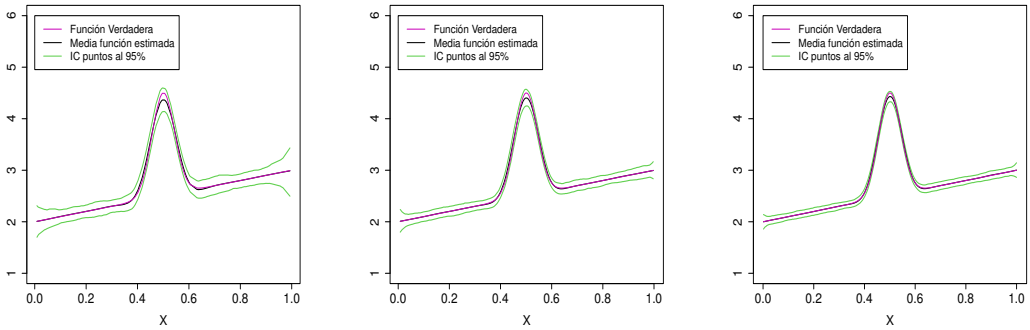
(a) n=200 (b) n=500 (c) n=1000



(d) C=10 and n=200 (e) C=10 and n=500 (f) C=10 and n=1000



(g) C=25 and n=200 (h) C=25 and n=500 (i) C=25 and n=1000



(j) C=40 and n=200 (k) C=40 and n=500 (l) C=40 and n=1000

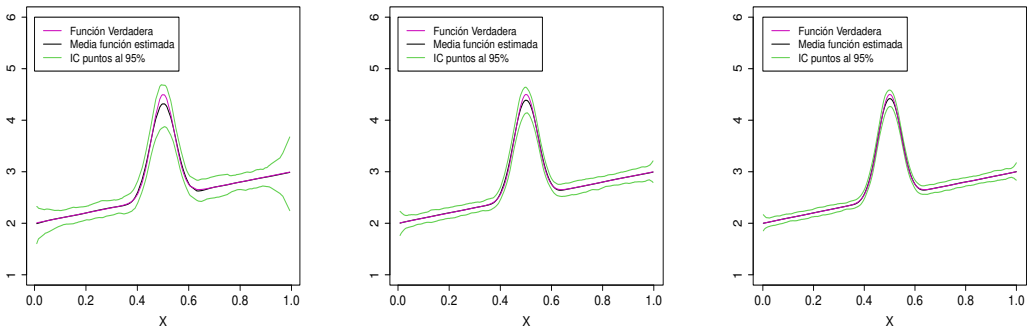


Figura 3.11: Función estimada utilizando el estimador ckmGAM para la función bump: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes

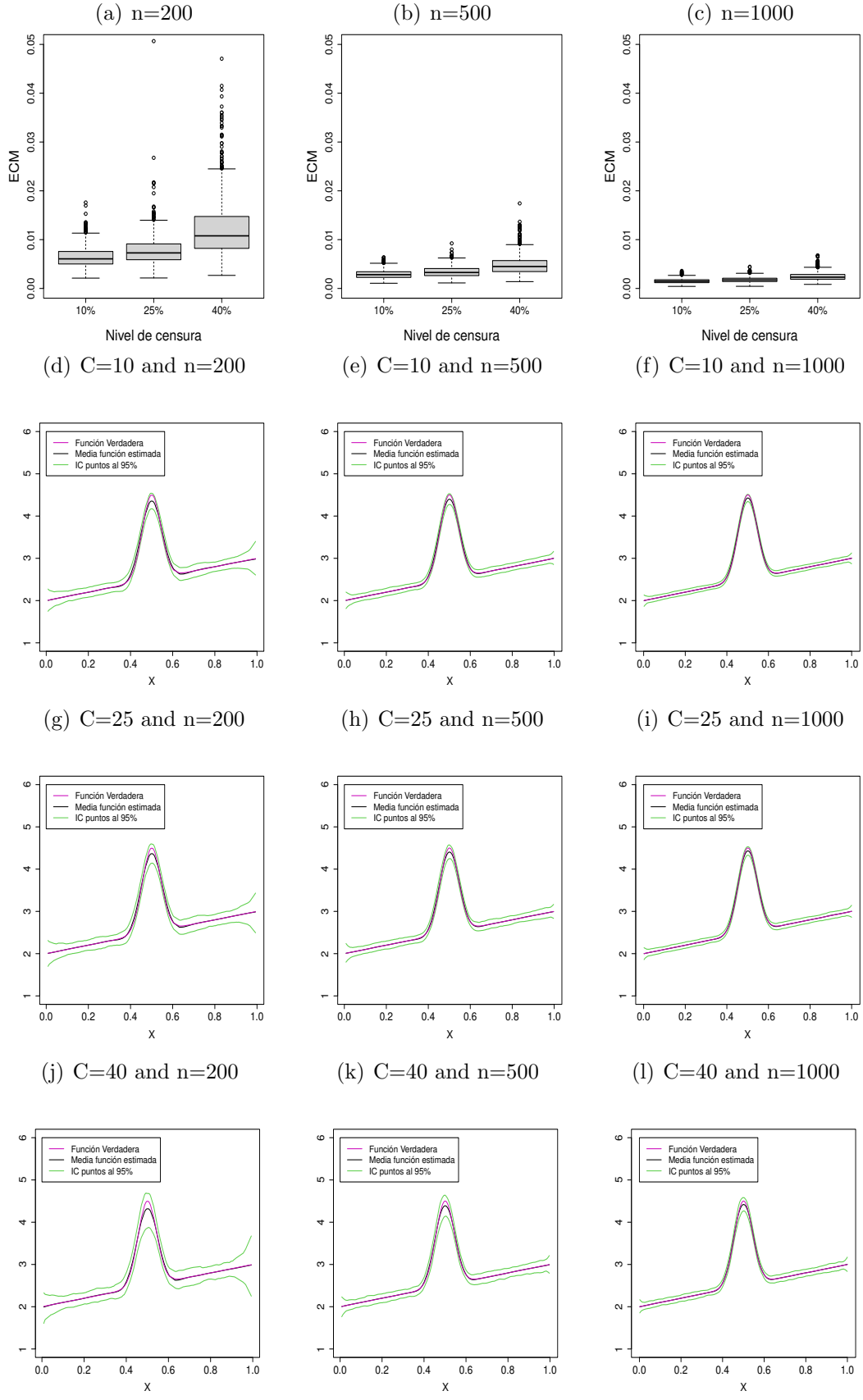
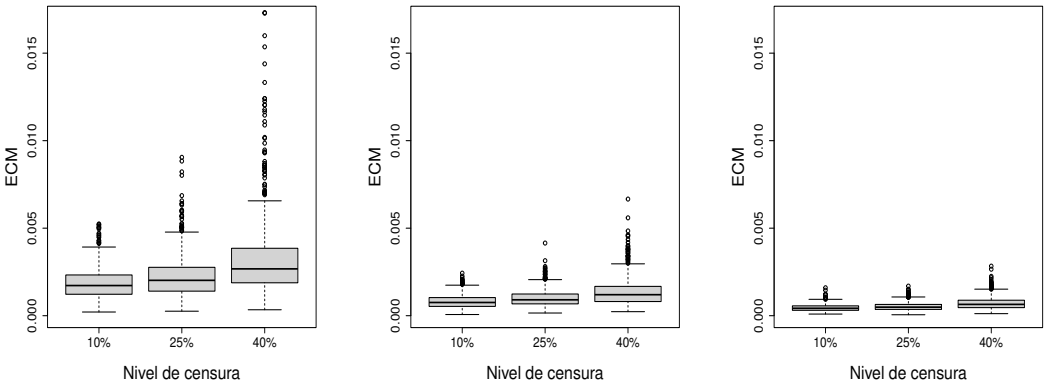
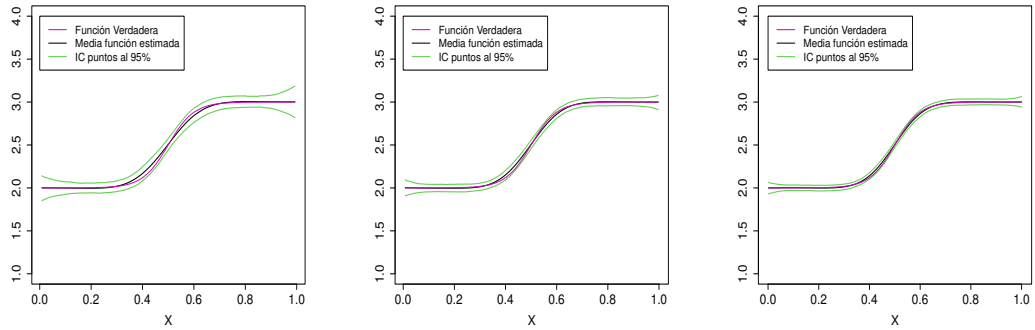


Figura 3.12: Función estimada utilizando el estimador ckmPS para la función logística: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes

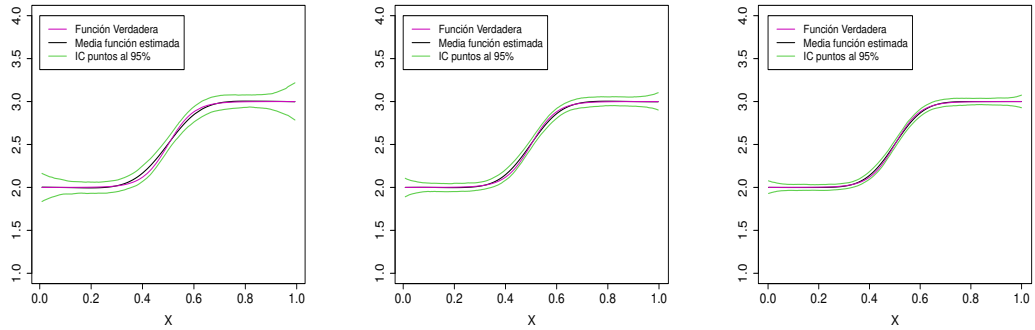
(a) n=200 (b) n=500 (c) n=1000



(d) C=10 and n=200 (e) C=10 and n=500 (f) C=10 and n=1000



(g) C=25 and n=200 (h) C=25 and n=500 (i) C=25 and n=1000



(j) C=40 and n=200 (k) C=40 and n=500 (l) C=40 and n=1000

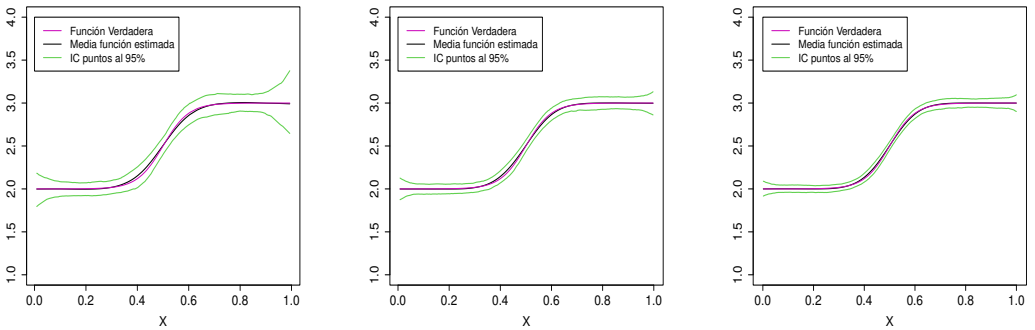


Figura 3.13: Función estimada utilizando el estimador ckmGAM para la función logística: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes

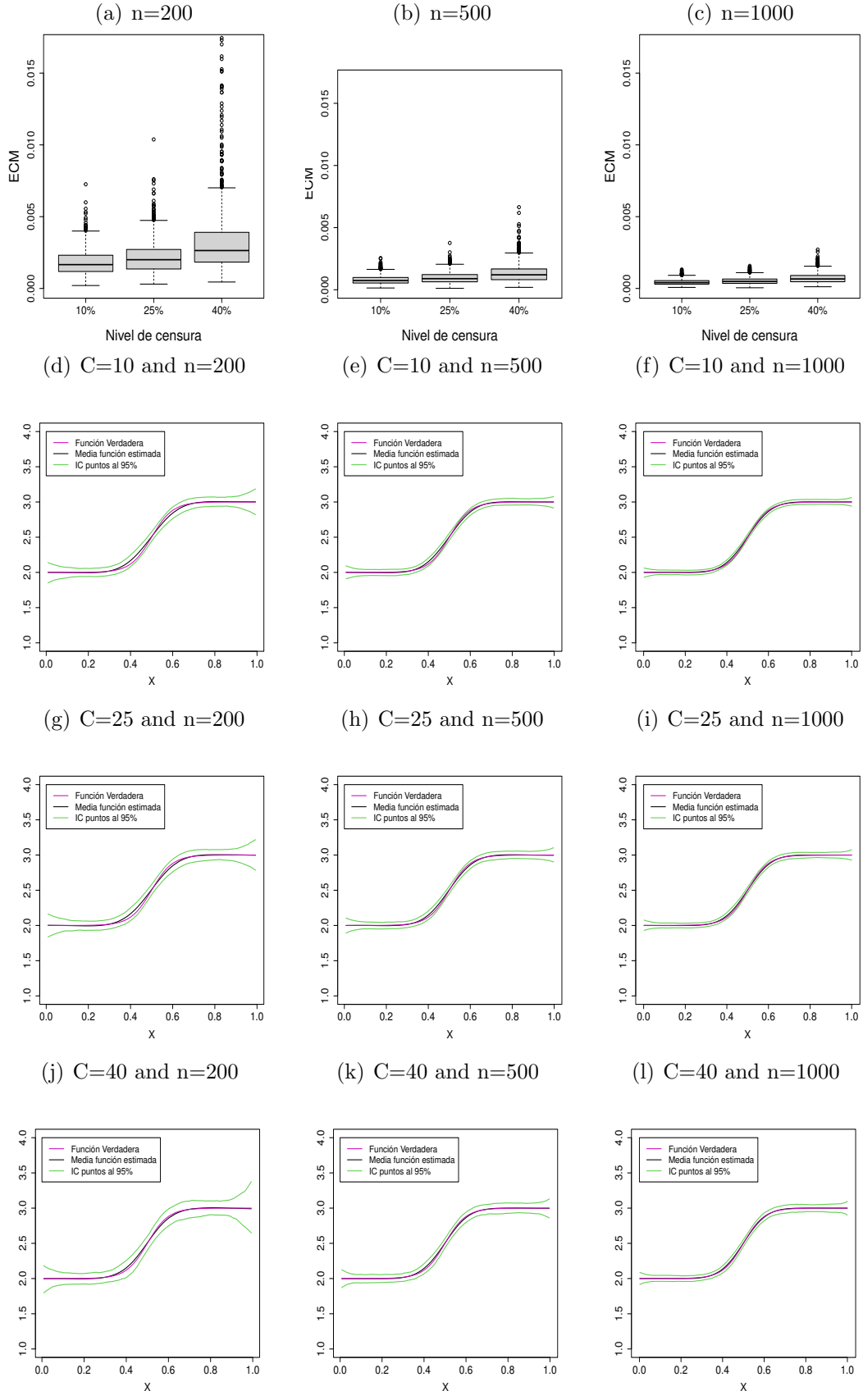
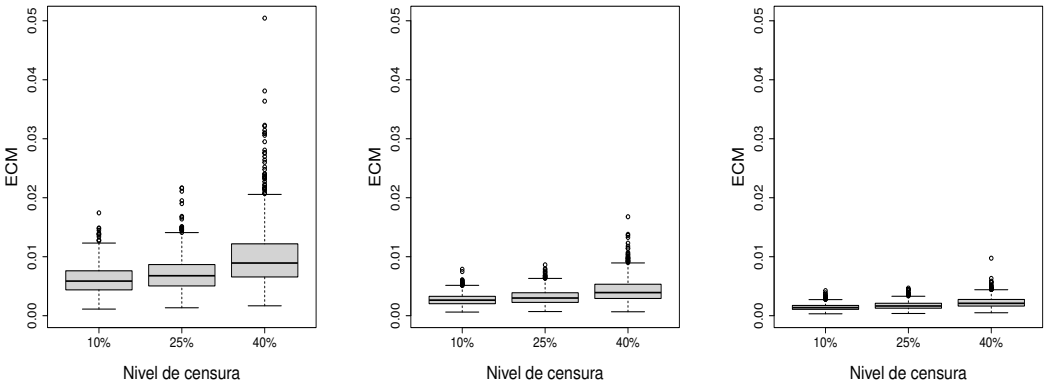
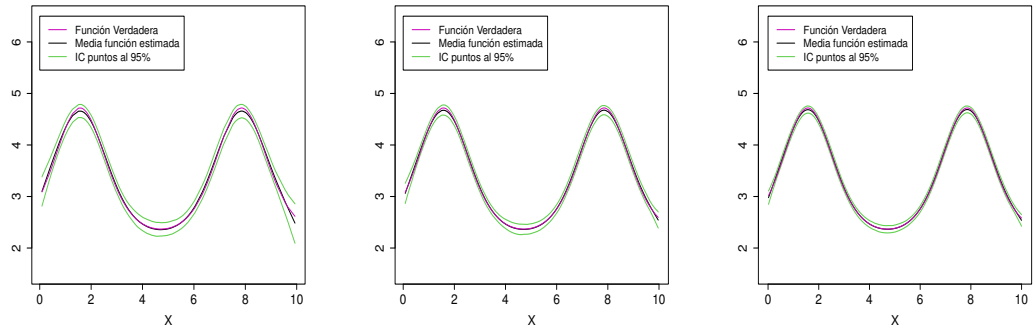


Figura 3.14: Función estimada utilizando el estimador ckmPS para la función sinusoidal con dos ciclos: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes

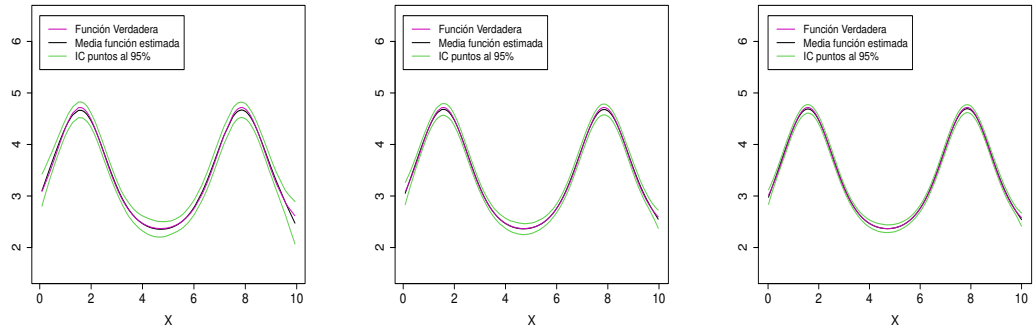
(a) n=200 (b) n=500 (c) n=1000



(d) C=10 and n=200 (e) C=10 and n=500 (f) C=10 and n=1000



(g) C=25 and n=200 (h) C=25 and n=500 (i) C=25 and n=1000



(j) C=40 and n=200 (k) C=40 and n=500 (l) C=40 and n=1000

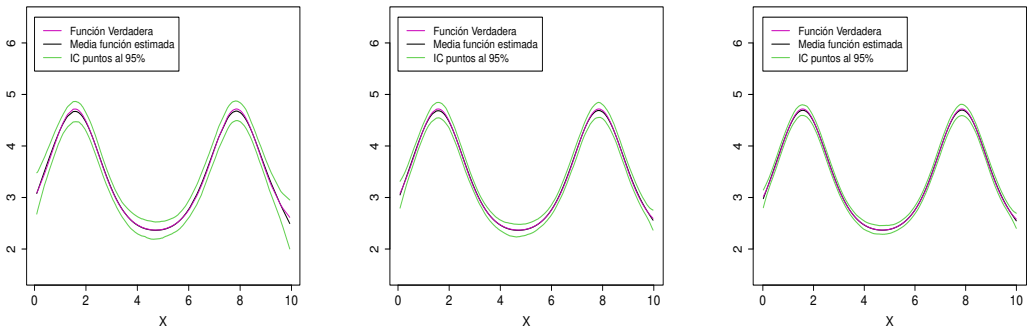


Figura 3.15: Función estimada utilizando el estimador ckmGAM para la función sinusoidal con dos ciclos: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes

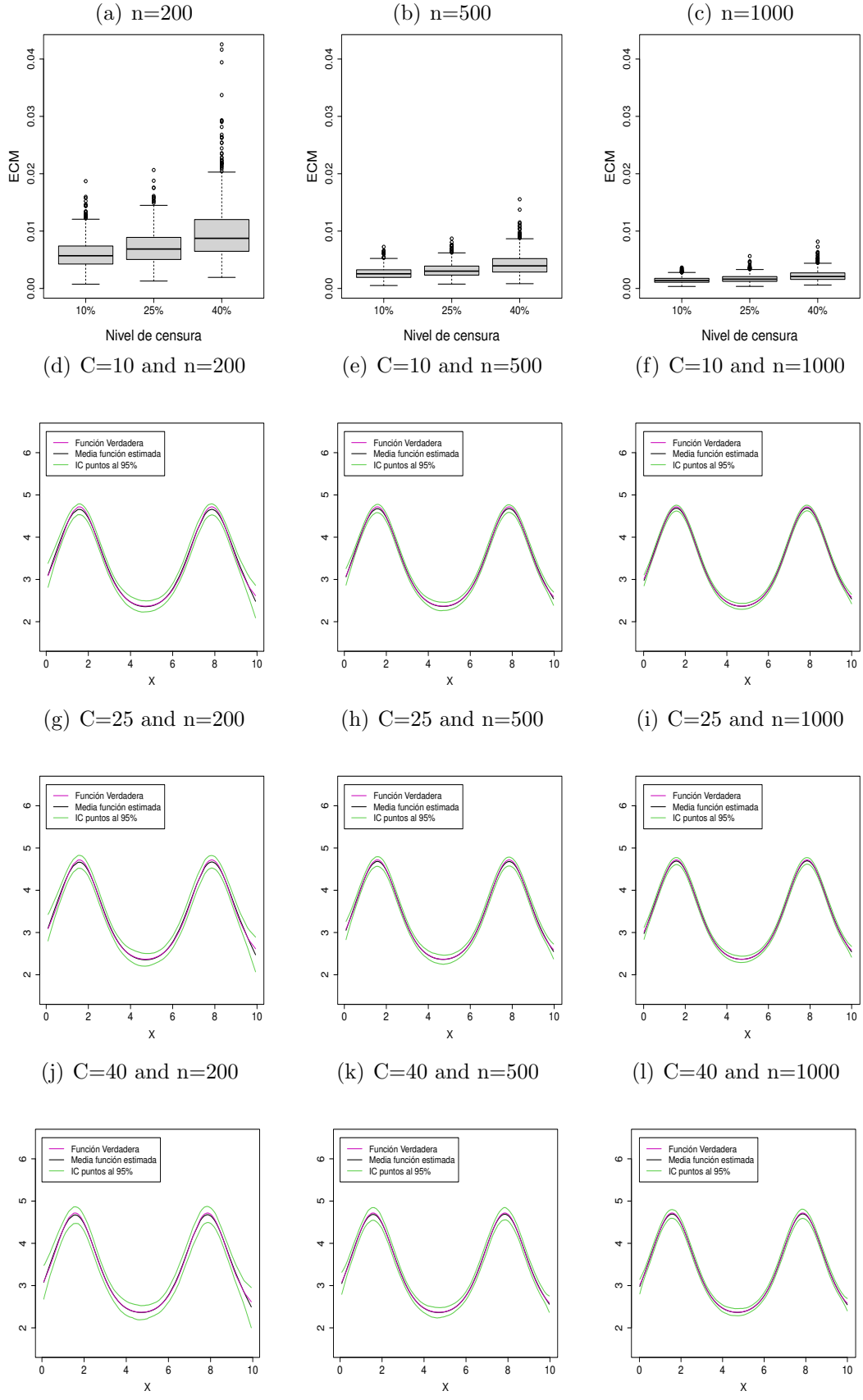
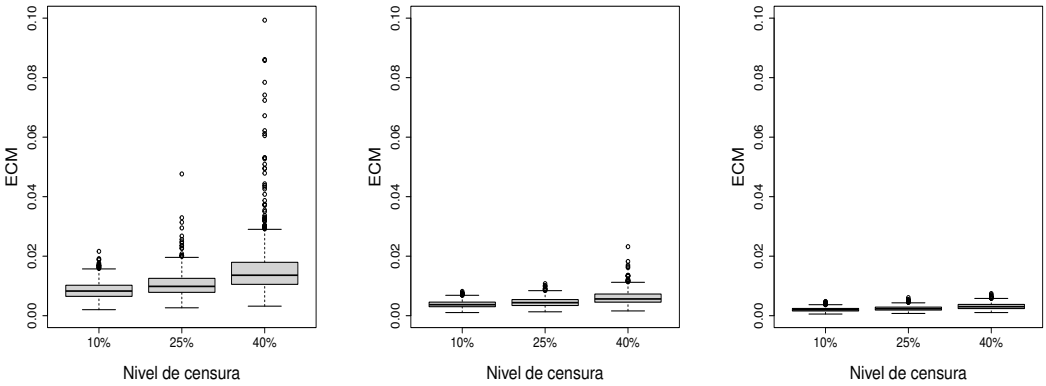
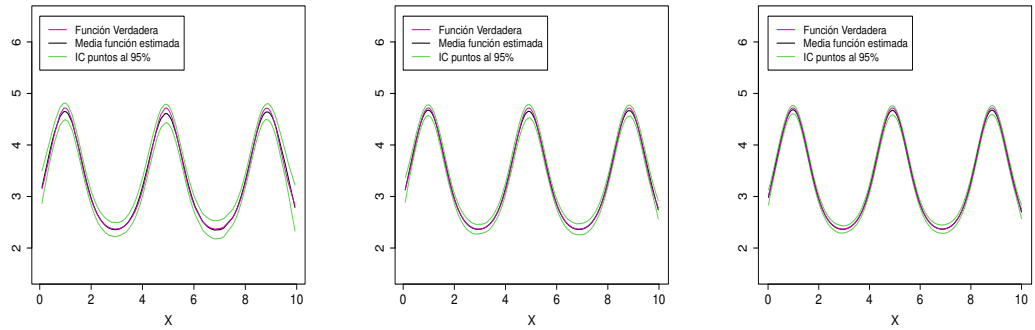


Figura 3.16: Función estimada utilizando el estimador ckmPS para la función sinusoidal con tres ciclos: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes

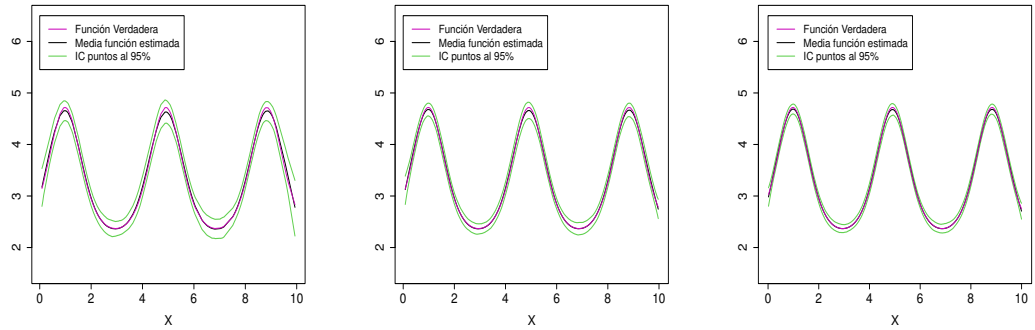
(a) n=200 (b) n=500 (c) n=1000



(d) C=10 and n=200 (e) C=10 and n=500 (f) C=10 and n=1000



(g) C=25 and n=200 (h) C=25 and n=500 (i) C=25 and n=1000



(j) C=40 and n=200 (k) C=40 and n=500 (l) C=40 and n=1000

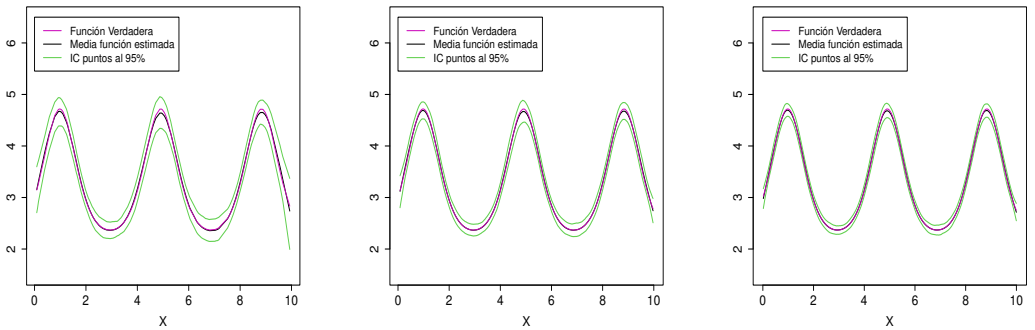
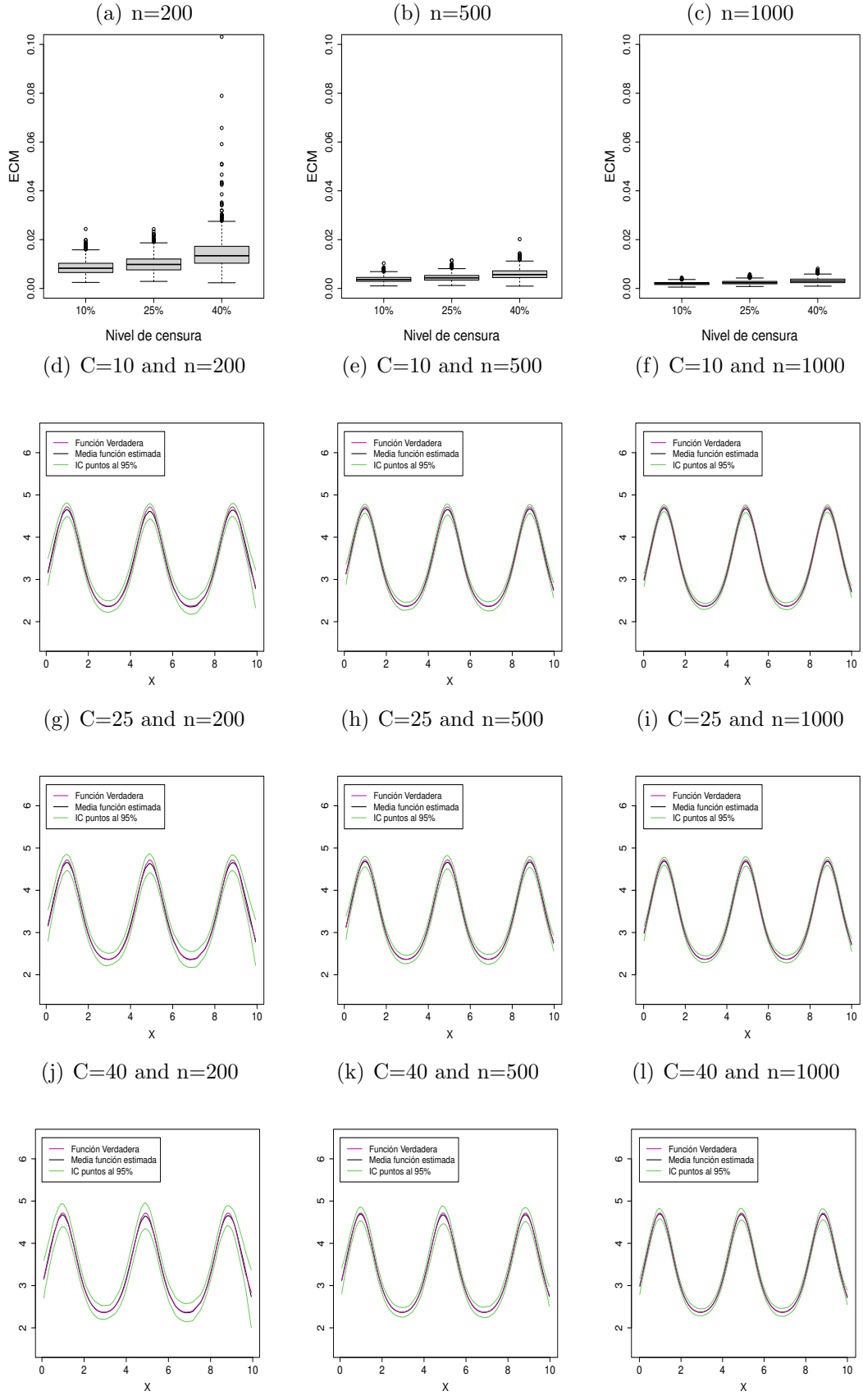


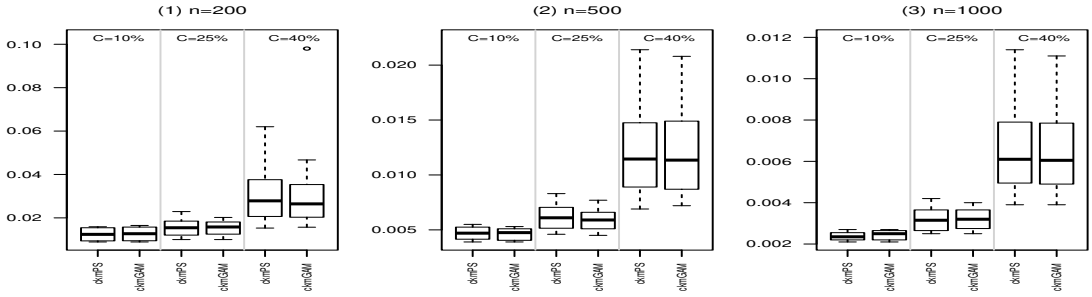
Figura 3.17: Función estimada utilizando el estimador ckmGAM para la función sinusoidal con tres ciclos: GVC_c con $\phi = 1{,}5$ y w_i^1 & K_c nodos equidistantes



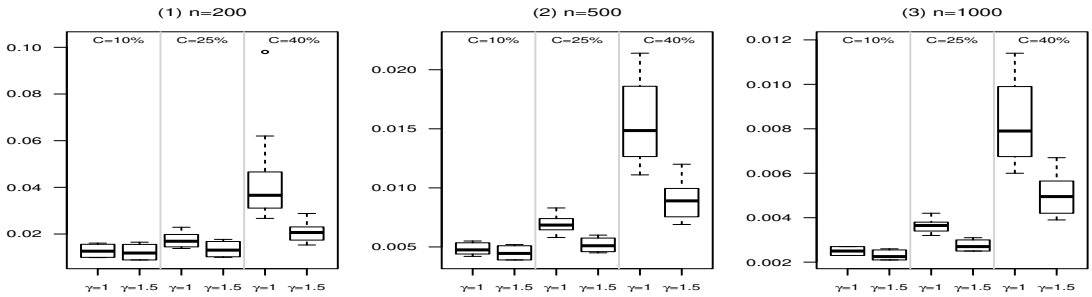
3.2.2. MECM según las diferentes elecciones de parámetros para las distintas funciones analizadas

Figura 3.18: MECM según las diferentes elecciones de parámetros para la función cuadrática

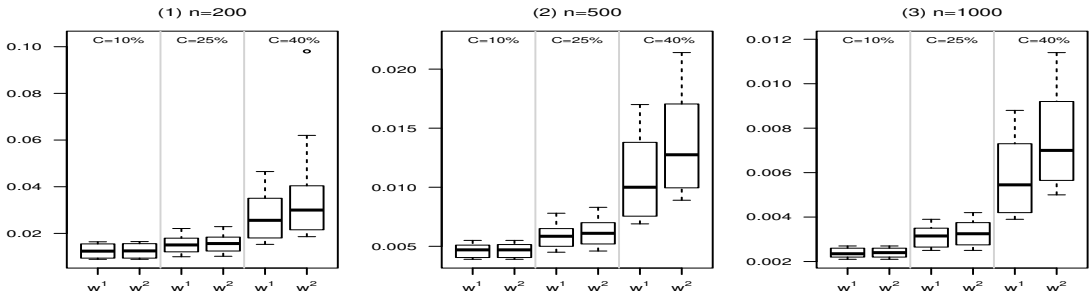
(a) ckmPS vs ckmGAM



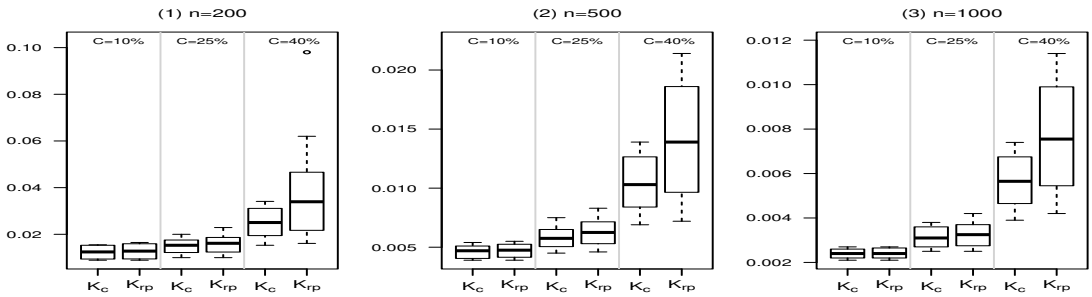
(b) $\phi = 1$ vs $\phi = 1{,}5$



(c) w^1 vs w^2



(d) K_c vs K_{rp}



(e) L_{eq} vs L_{km}

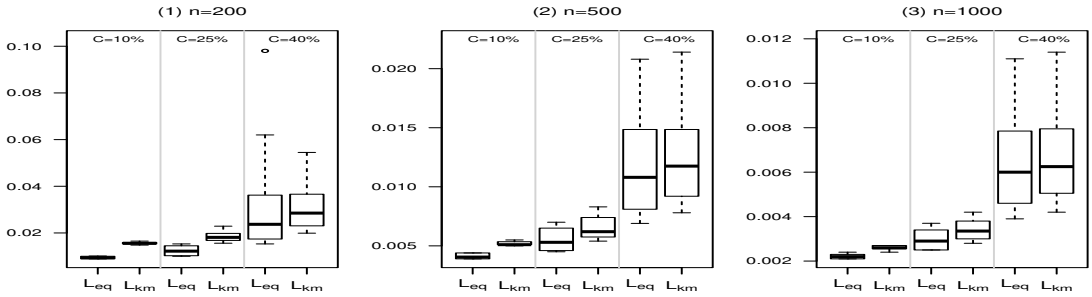
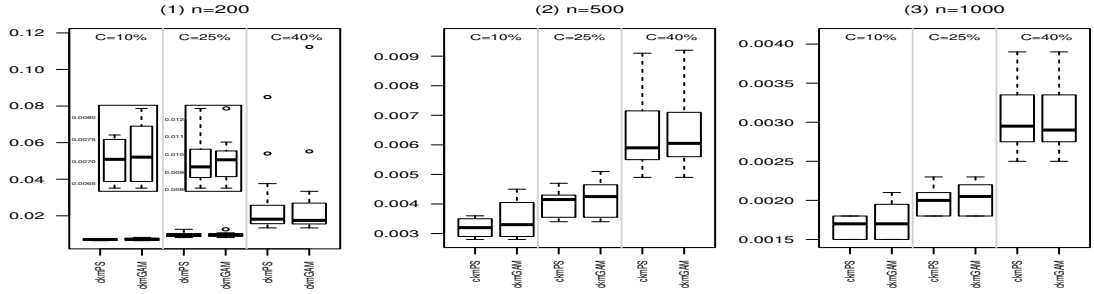
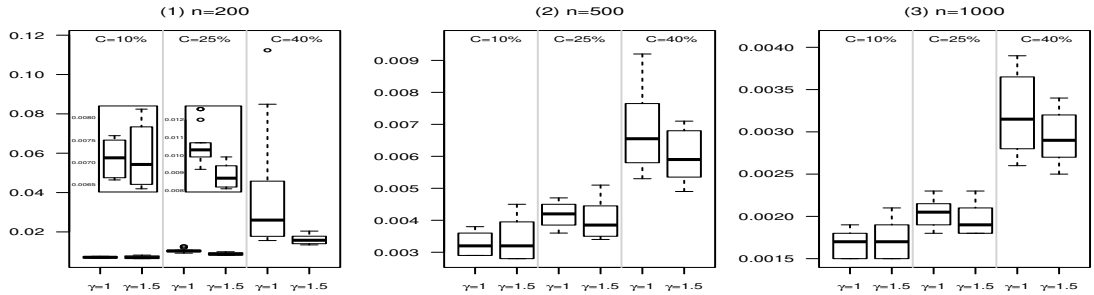


Figura 3.19: MECM según las diferentes elecciones de parámetros para la función bump

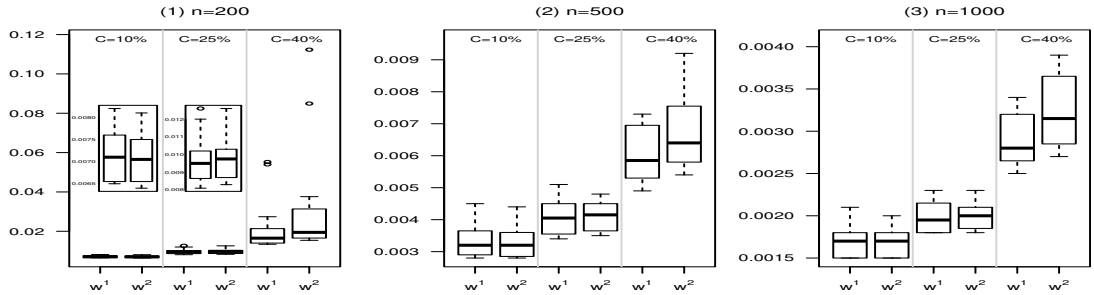
(a) ckmPS vs ckmGAM



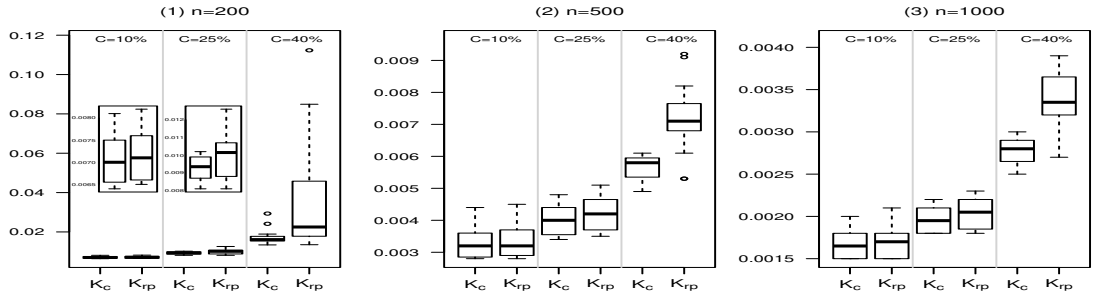
(b) $\phi = 1$ vs $\phi = 1{,}5$



(c) w^1 vs w^2



(d) K_c vs K_{rp}



(e) L_{eq} vs L_{km}

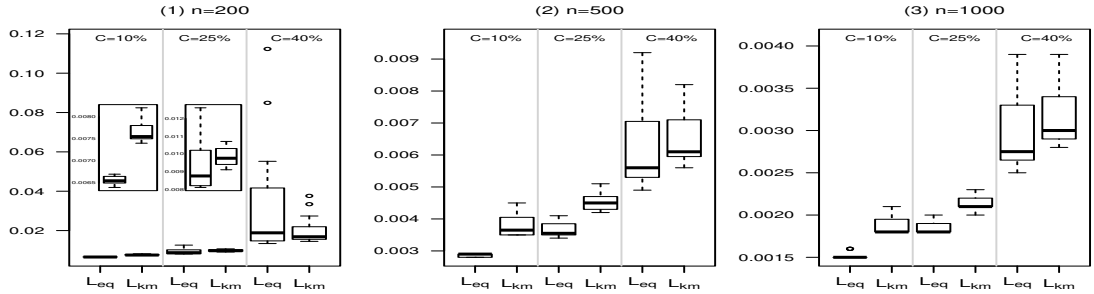
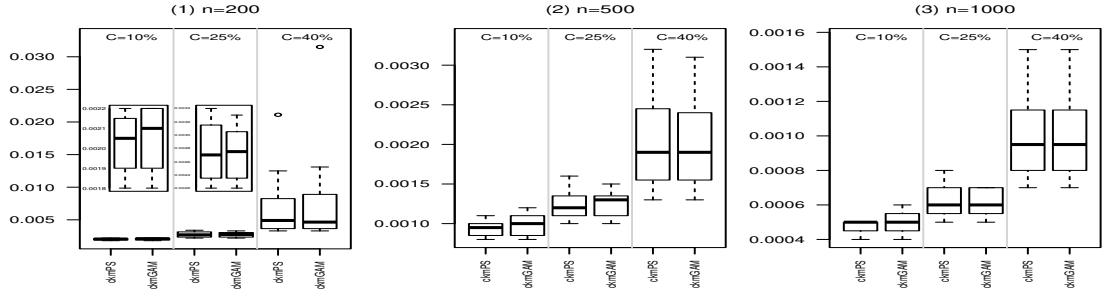
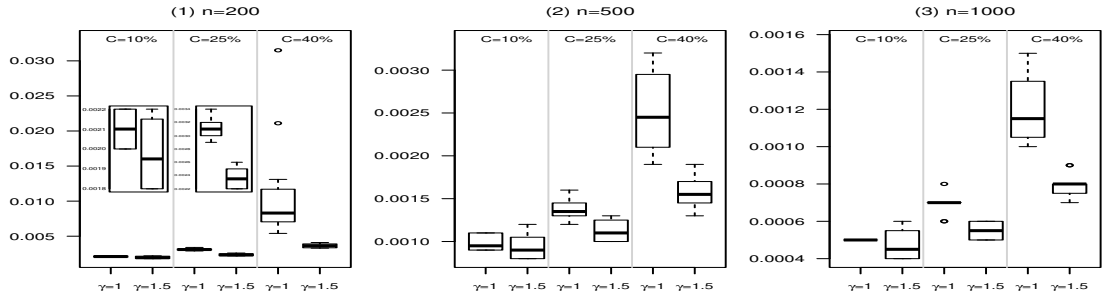


Figura 3.20: MECM según las diferentes elecciones de parámetros para la función logística

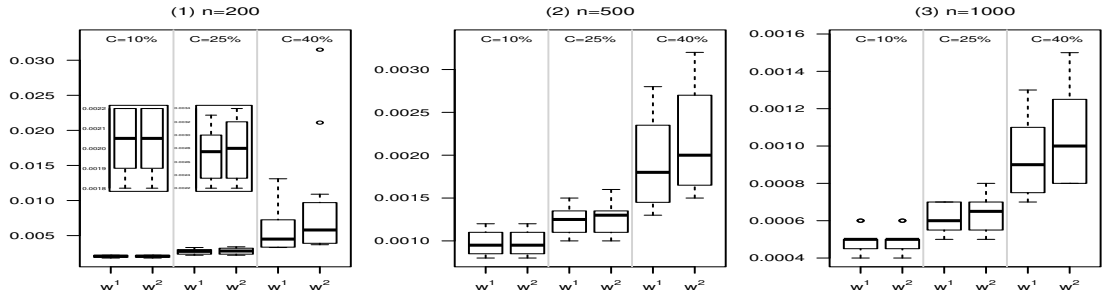
(a) ckmPS vs ckmGAM



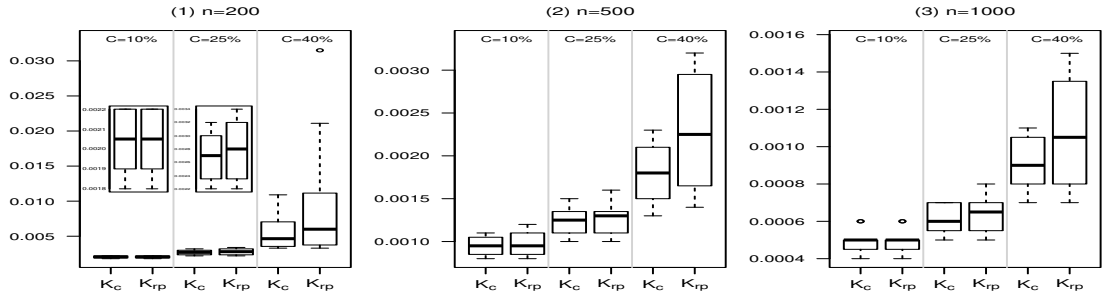
(b) $\phi = 1$ vs $\phi = 1{,}5$



(c) w^1 vs w^2



(d) K_c vs K_{rp}



(e) L_{eq} vs L_{km}

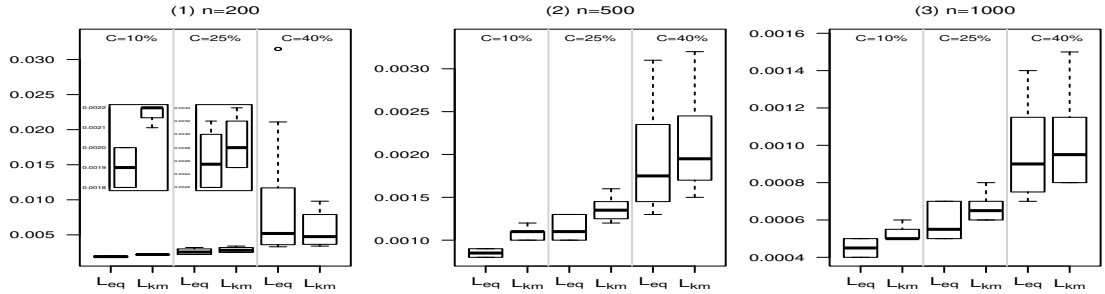
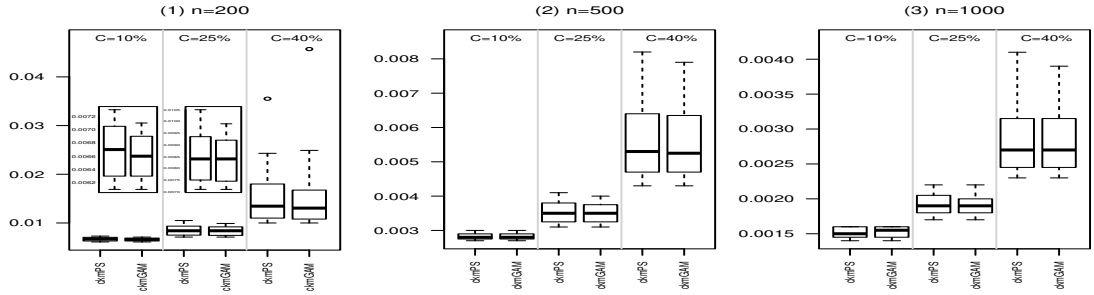
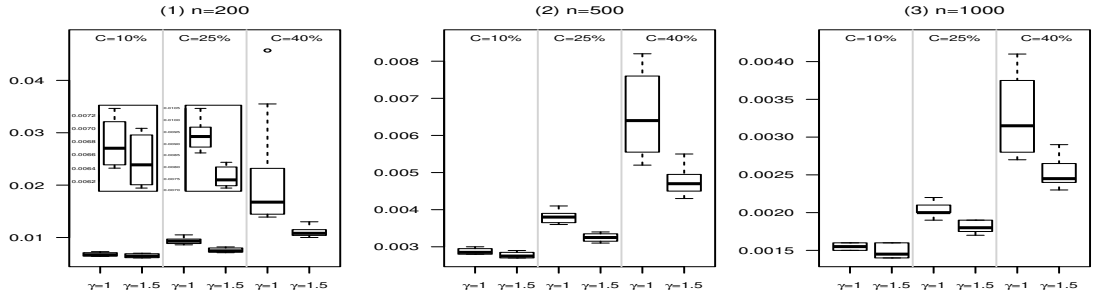


Figura 3.21: MECM según las diferentes elecciones de parámetros para la función sinusoidal con dos ciclos

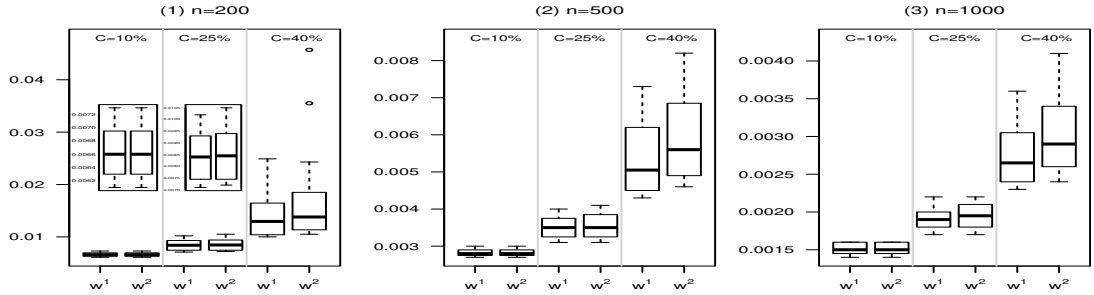
(a) ckmPS vs ckmGAM



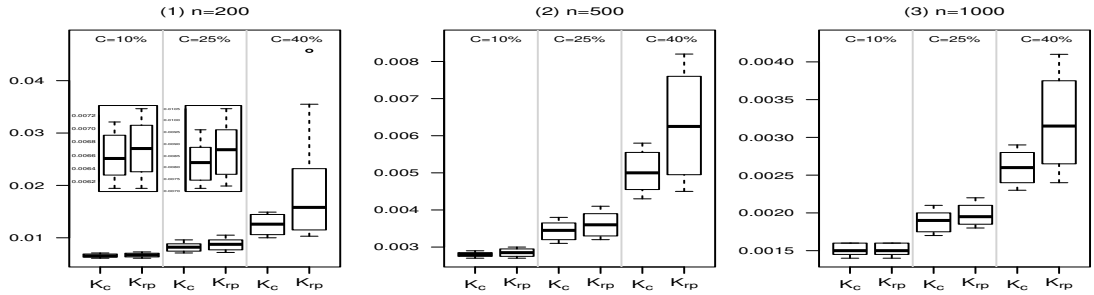
(b) $\phi = 1$ vs $\phi = 1{,}5$



(c) w^1 vs w^2



(d) K_c vs K_{rp}



(e) L_{eq} vs L_{km}

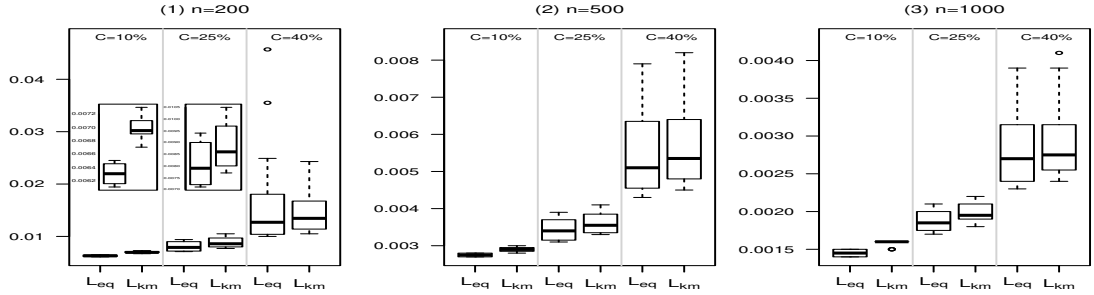
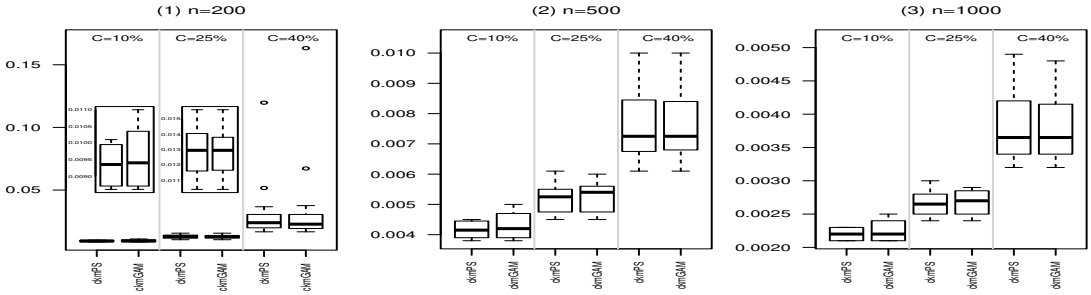
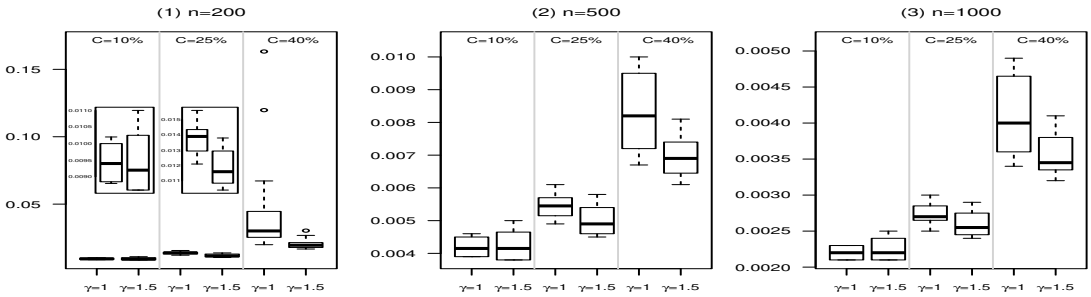


Figura 3.22: MECM según las diferentes elecciones de parámetros para la función sinusoidal con tres ciclos

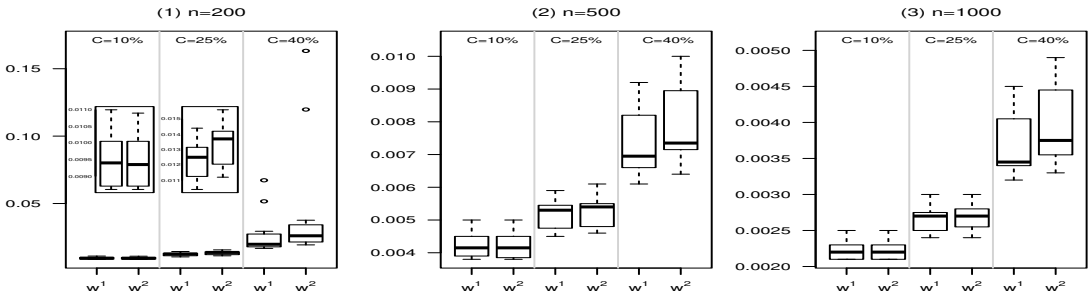
(a) ckmPS vs ckmGAM



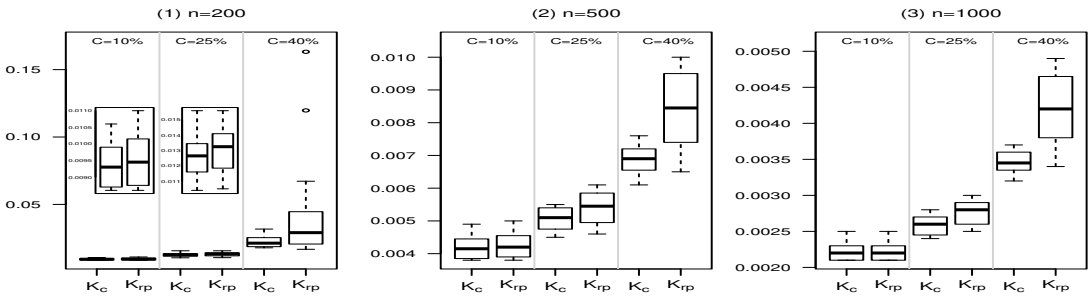
(b) $\phi = 1$ vs $\phi = 1{,}5$



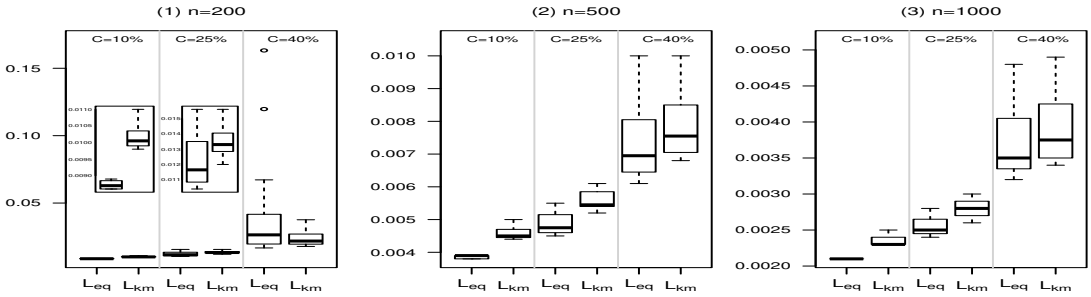
(c) w^1 vs w^2



(d) K_c vs K_{rp}



(e) L_{eq} vs L_{km}



3.3. Resultados modelo semiparamétrico

3.3.1. Estudio de simulación

Tabla 3.9: Tres casos de estudio

Nombre	z_i	$f(z_i)$	σ^2_ϵ	ratio SR
Caso (i): Cuadrática	$z_i \sim U[0,4]$	$2 + 4z_i - z_i^2$	0,40	3,5
Caso (ii): Sinusoidal	$z_i \sim U[0,10]$	$2 + exp\{sin(z_i)\}$	0,20	3,3
Caso (iii): Logística	$z_i \sim U[0,1]$	$2 + \frac{1}{1+exp\{-20(z_i-0{,}5)\}}$	0,06	3,3

En esta sección se estudia el desempeño de la metodología propuesta mediante un estudio de simulación. Para ello se considera el siguiente modelo semiparamétrico:

$$T_i = \alpha_1 X_{1i} + \alpha_2 X_{2i} + f(Z_i) + \epsilon_i \tag{3.2}$$

donde para el componente paramétrico del modelo: la variable X_1 se genera a partir de una distribución uniforme $U(0,2)$, X_2 a partir de una distribución uniforme $U(-1,3)$, siendo los valores de los coeficientes $\alpha_1 = -1$ y $\alpha_2 = 1$. Para el componente no paramétrico, se consideran tres casos diferentes para la relación $f(\cdot)$ entre T y una covariable relevante Z, véase la tabla 3.9 para las formas funcionales elegidas y la distribución de probabilidad de la variable Z. Para la distribución del término de error (ϵ) se ha utilizado la distribución normal $N(0,\sigma^2)$, donde el valor del parámetro σ^2 se ha elegido para obtener una relación señal/ruido (SR) similar en cada ejemplo (tabla 3.2). Para estudiar el efecto de los datos censurados, se considera una variable censura C generada independientemente de una distribución uniforme $U(1,b)$. El valor del parámetro b cambia para considerar tres niveles diferentes de censura: 10 %, 25 % y 40 %.

Por tanto, se observa $(y_1, x_{11}, x_{21}, z_1, \delta_1), \ldots, (y_n, x_{1n}, x_{2n}, z_n, \delta_n)$ una muestra de tamaño n, donde $y_i = min(t_i, c_i)$ es el tiempo de supervivencia observado, *i.e.*, el mínimo entre el tiempo de supervivencia t_i y el valor de censura c_i. Además, se conoce a través de la variable indicadora $\delta_i = I(t_i \leq c_i)$ qué observaciones no están censuradas. Se consideran tres tamaños de muestra: $n = 200$, $n = 500$ y $n = 1000$ y para cada uno de los nueve escenarios, tres tamaños de muestra por tres niveles de censura, se utilizan 1000 réplicas de Monte Carlo.

Para cada uno de los 27 casos analizados en este estudio de simulación se ha estimado el modelo (3.2) siguiendo la propuesta de estimación presentada en la sección 2.3, el estimador P-splines censurado (CPS), donde la elección del parámetro de suavizado λ y el número de nodos de las B-splines se han elegido mediante las fórmulas (2.9) y (2.11), respectivamente. Las tablas 3.10, 3.11 y 3.12 presentan un resumen general de los resultados obtenidos para cada combinación de nivel de censura y tamaño muestral en cada uno de las tres formas funcionales estudiadas

en el modelo (3.2). La tabla 3.10 resume la estimación del caso (i), donde $f(z)$ es una función cuadrática. Las dos primeras filas de la tabla 3.10 presentan el error cuadrático medio (ECM) de cada coeficiente estimado ($\hat{\alpha}_1$ y $\hat{\alpha}_2$) del componente paramétrico:

$$ECM(\hat{\alpha}_p) = \frac{1}{1000}\sum_{j=1}^{1000}(\alpha_p - \hat{\alpha}_{pj})^2 \qquad p = 1, 2$$

y en la tercera fila la media aritmética del error cuadrático medio (MECM) del componente no paramétrico:

$$MECM = \frac{1}{1000}\sum_{j=1}^{1000}\left(\frac{\sum_{i=1}^{n}(f(z_i) - \hat{f}_j(z_i))^2}{n}\right)$$

Las filas cuatro a seis de la tabla 3.10 presentan el sesgo empírico y las filas siete a nueve las probabilidades de cobertura de los intervalos de confianza del 95 % basados en el remuestreo. Las tablas 3.11 y 3.12 presentan la misma información para las estimaciones del caso (ii) y (iii), donde $f(z)$ es una función sinusoidal y una función logística, respectivamente. En estas tablas 3.10 a 3.12 se observa el buen comportamiento del método propuesto en términos del ECM, la MECM, el sesgo empírico y las probabilidades de cobertura.

A continuación se analiza la estimación de cada componente del modelo (3.2) por separado. Para el caso (i), función cuadrática, las subfiguras 3.23.(a) a 3.23.(c) presentan el ECM para el componente no paramétrico utilizando diferentes niveles de censura y tamaños de muestra, donde, como puede observarse, las estimaciones del componente no paramétrico mejoran a medida que aumenta el tamaño de la muestra y disminuye el nivel de censura en la misma. Las subfiguras 3.23.(d) a 3.23.(i) muestran las estimaciones de los coeficientes del componente paramétrico ($\hat{\alpha}_1$ y $\hat{\alpha}_2$) donde se puede observar el buen comportamiento de los estimadores utilizados, con una precisión que mejora a medida que aumenta el tamaño de la muestra y disminuye el nivel de censura. Adicionalmente, en la figura 3.26 se presenta el valor medio de las estimaciones de la función de forma cuadrática en comparación con la verdadera forma funcional a estimar. Como se puede observar, la estimación funciona muy bien reflejando la verdadera forma funcional de $f(\cdot)$. En esta figura 3.26, también se puede comprobar el buen comportamiento de los intervalos de confianza asintóticos generados con los estimadores de las varianzas propuestas en el apartado 2.3.3. Como puede observarse, para un nivel de confianza del 95 %, el intervalo de confianza medio propuesto (líneas azules) es coherente con el correspondiente intervalo del percentil 95 de las simulaciones (líneas verdes). Por último, las probabilidades de cobertura de los intervalos de confianza presentadas en la tabla 3.10 muestran que la probabilidad de cobertura real se aproxima bastante a la probabilidad de cobertura nominal. Resultados análogos, donde se aprecia el buen comportamiento de las propuestas presentadas, se obtienen para el caso (ii), función sinusoidal, figuras 3.24 y 3.27, y para el caso (iii), función logística, figuras 3.25 y 3.28.

Se han realizado simulaciones adicionales considerando una distribución normal para la variable de censura y también simulaciones adicionales considerando distribuciones de error no normales, como la distribución de Weibull. La forma de generar la censura que se ha utilizado en esta tesis es común en la literatura y ha sido utilizada previamente, por ejemplo, en Stute (1999), Jin et al. (2003) y De Uña Álvarez

and Roca Pardiñas (2009). En cualquier caso, se han realizado simulaciones adicionales considerando una distribución normal para la variable censura. Las tablas del apartado 3.3.3 muestran los resultados de la estimación cuando los tiempos de censura se generan a partir de una distribución uniforme (uniforme), una distribución normal (normal) y la diferencia entre ambas (dif), en azul cuando los resultados mejoran con censura normal, en rojo en caso contrario. De forma análoga, las tablas del apartado 3.3.4 muestran los resultados de la estimación cuando el término de error se genera a partir de una distribución no normal (weibull), una distribución normal (normal) y la diferencia entre ambas (dif), en rojo cuando los resultados mejoran con el término de error normal, en azul en caso contrario. Los nuevos resultados obtenidos confirman el buen funcionamiento del método propuesto, independientemente de la forma de generar la censura o la distribución del error, y son coherentes con los presentados en esta sección.

Tabla 3.10: Modelo semiparamétrico: resultados para la función cuadrática

	$n = 200$			$n = 500$			$n = 1000$		
% Censura	10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
	ECM ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and MECM ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	3.090	3.741	5.965	1.324	1.440	2.370	0.521	0.656	0.992
$\widehat{\alpha}_2$	0.722	0.906	1.581	0.275	0.302	0.541	0.121	0.181	0.259
$\widehat{f}$	9.783	12.170	21.109	4.126	5.056	8.730	2.105	2.580	4.424
	Sesgo empírico								
$\widehat{\alpha}_1$	-0.00149	0.00099	0.01130	-0.00319	-0.00290	0.00042	0.00214	0.00354	0.00575
$\widehat{\alpha}_2$	0.00289	0.00206	-0.00257	0.00049	0.00039	-0.00335	0.00036	-0.00036	-0.00195
$\widehat{f}$	-0.00033	-0.00148	-0.01630	0.00239	0.00272	0.00067	-0.00232	-0.00253	-0.00575
	Probabilidades de cobertura de los intervalos de confianza al 95 %								
$\widehat{\alpha}_1$	0.938	0.955	0.947	0.928	0.950	0.947	0.946	0.946	0.948
$\widehat{\alpha}_2$	0.945	0.946	0.934	0.941	0.960	0.943	0.960	0.926	0.957
$\widehat{f}$	0.938	0.941	0.923	0.939	0.939	0.925	0.946	0.936	0.933

Tabla 3.11: Modelo semiparamétrico: resultados para la función sinusoidal

	$n = 200$			$n = 500$			$n = 1000$		
% Censura	10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
	ECM ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and MECM ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	0.806	1.060	1.521	0.285	0.362	0.560	0.136	0.154	0.233
$\widehat{\alpha}_2$	0.189	0.266	0.376	0.062	0.087	0.132	0.035	0.044	0.064
$\widehat{f}$	4.088	5.205	7.970	1.702	2.023	3.072	0.870	1.047	1.545
	Sesgo empírico								
$\widehat{\alpha}_1$	-0.00543	-0.00202	0.00083	0.00085	0.00098	0.00209	0.00060	0.00016	0.00148
$\widehat{\alpha}_2$	-0.00060	-0.00073	-0.00145	0.00051	0.00058	-0.00093	-0.00016	-0.00017	-0.00136
$\widehat{f}$	0.00674	0.00311	-0.00152	-0.00116	-0.00167	-0.00324	-0.00021	0.00010	-0.00077
	Probabilidades de cobertura de los intervalos de confianza al 95 %								
$\widehat{\alpha}_1$	0.944	0.936	0.925	0.956	0.955	0.944	0.948	0.956	0.940
$\widehat{\alpha}_2$	0.949	0.930	0.938	0.952	0.938	0.944	0.944	0.943	0.962
$\widehat{f}$	0.930	0.927	0.918	0.932	0.941	0.932	0.942	0.938	0.941

Tabla 3.12: Modelo semiparamétrico: resultados para la función logística

	$n = 200$			$n = 500$			$n = 1000$		
% Censura	10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
	ECM ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and MECM ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	0.072	0.098	0.172	0.024	0.036	0.064	0.011	0.016	0.027
$\widehat{\alpha}_2$	0.017	0.025	0.046	0.006	0.008	0.019	0.003	0.004	0.009
$\widehat{f}$	0.309	0.397	0.710	0.128	0.164	0.311	0.065	0.085	0.169
	Sesgo empírico								
$\widehat{\alpha}_1$	0.00012	0.00029	0.00101	0.00042	0.00061	0.00035	-0.00029	-0.00022	-0.00011
$\widehat{\alpha}_2$	0.00026	0.00015	-0.00008	-0.00015	-0.00026	-0.00002	-0.00002	-0.00003	-0.00014
$\widehat{f}$	-0.00037	-0.00038	-0.00098	-0.00022	-0.00029	-0.00040	0.00035	0.00020	0.00014
	Probabilidades de cobertura de los intervalos de confianza al 95 %								
$\widehat{\alpha}_1$	0.944	0.955	0.933	0.953	0.944	0.941	0.946	0.941	0.948
$\widehat{\alpha}_2$	0.950	0.930	0.924	0.939	0.947	0.938	0.957	0.938	0.956
$\widehat{f}$	0.944	0.939	0.916	0.943	0.938	0.93 0	0.949	0.941	0.938

Figura 3.23: Resultados del estudio de simulación para la función cuadrática: Errores Cuadráticos Medios para la parte no paramétrica , $\widehat{\alpha}_1$ y $\widehat{\alpha}_2$ utilizando diferentes niveles de censura y tamaños de muestra

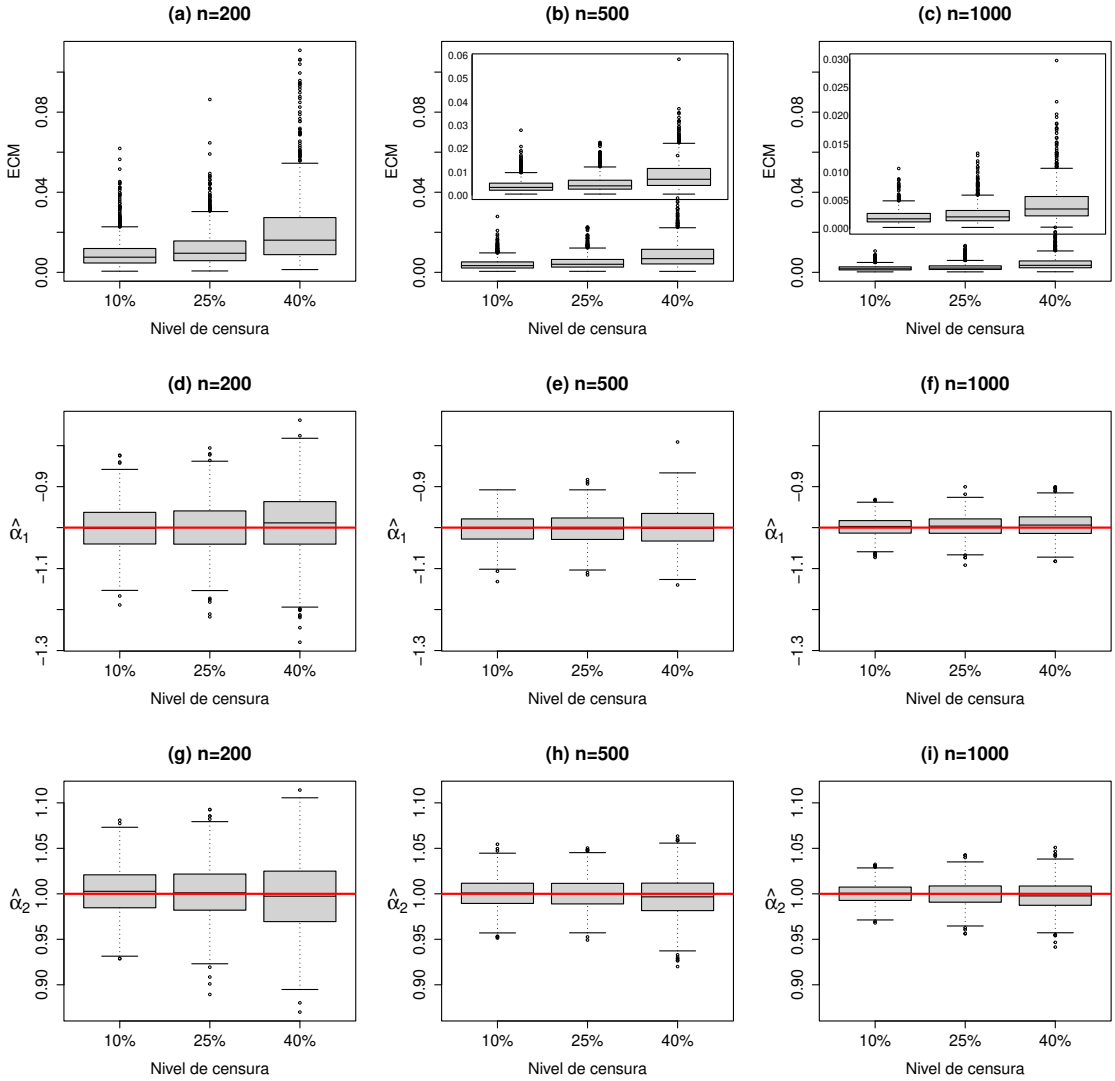


Figura 3.24: Resultados del estudio de simulación para la función sinusoidal: Errores Cuadráticos Medios para la parte no paramétrica , $\widehat{\alpha}_1$ y $\widehat{\alpha}_2$ utilizando diferentes niveles de censura y tamaños de muestra

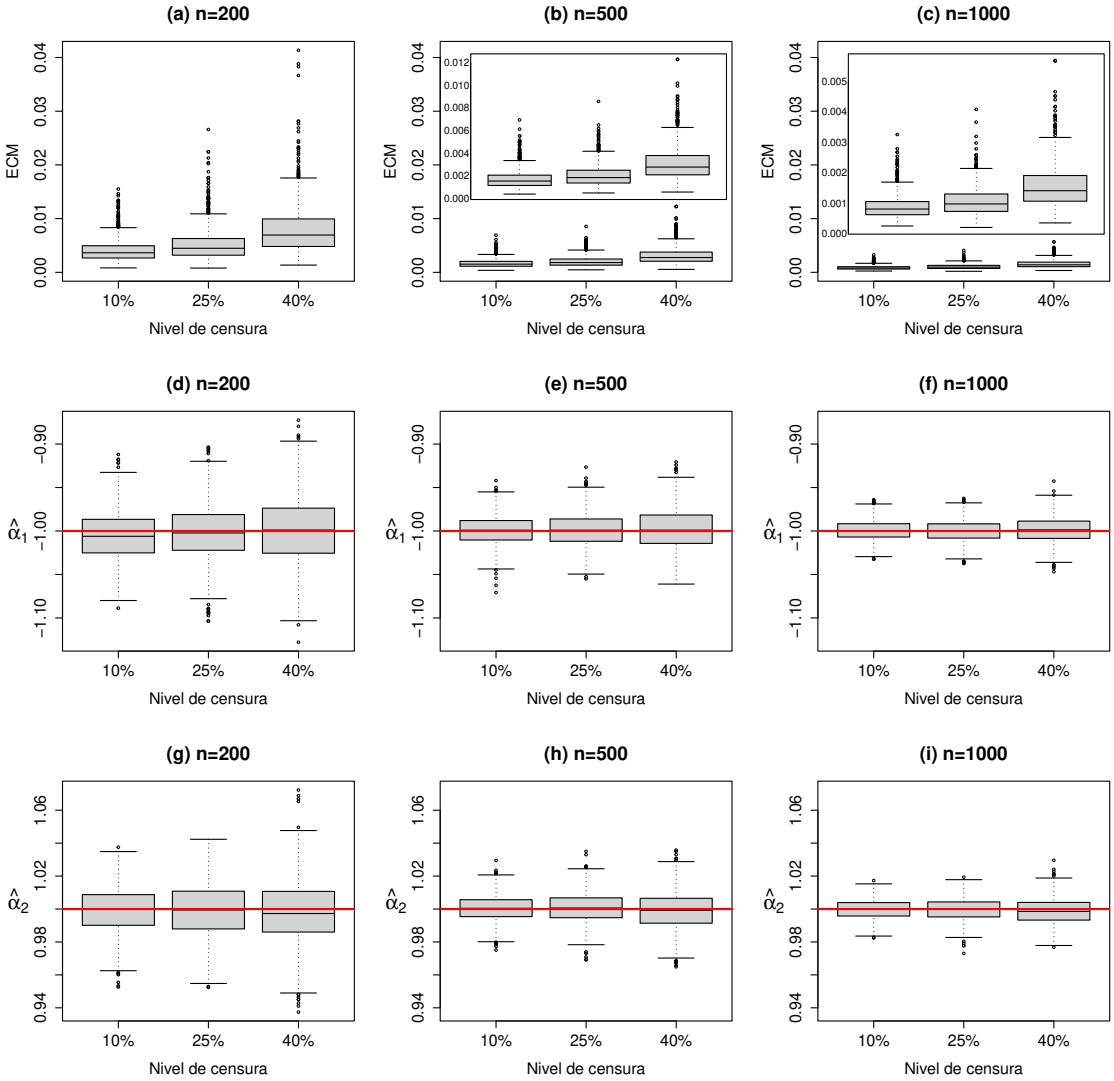
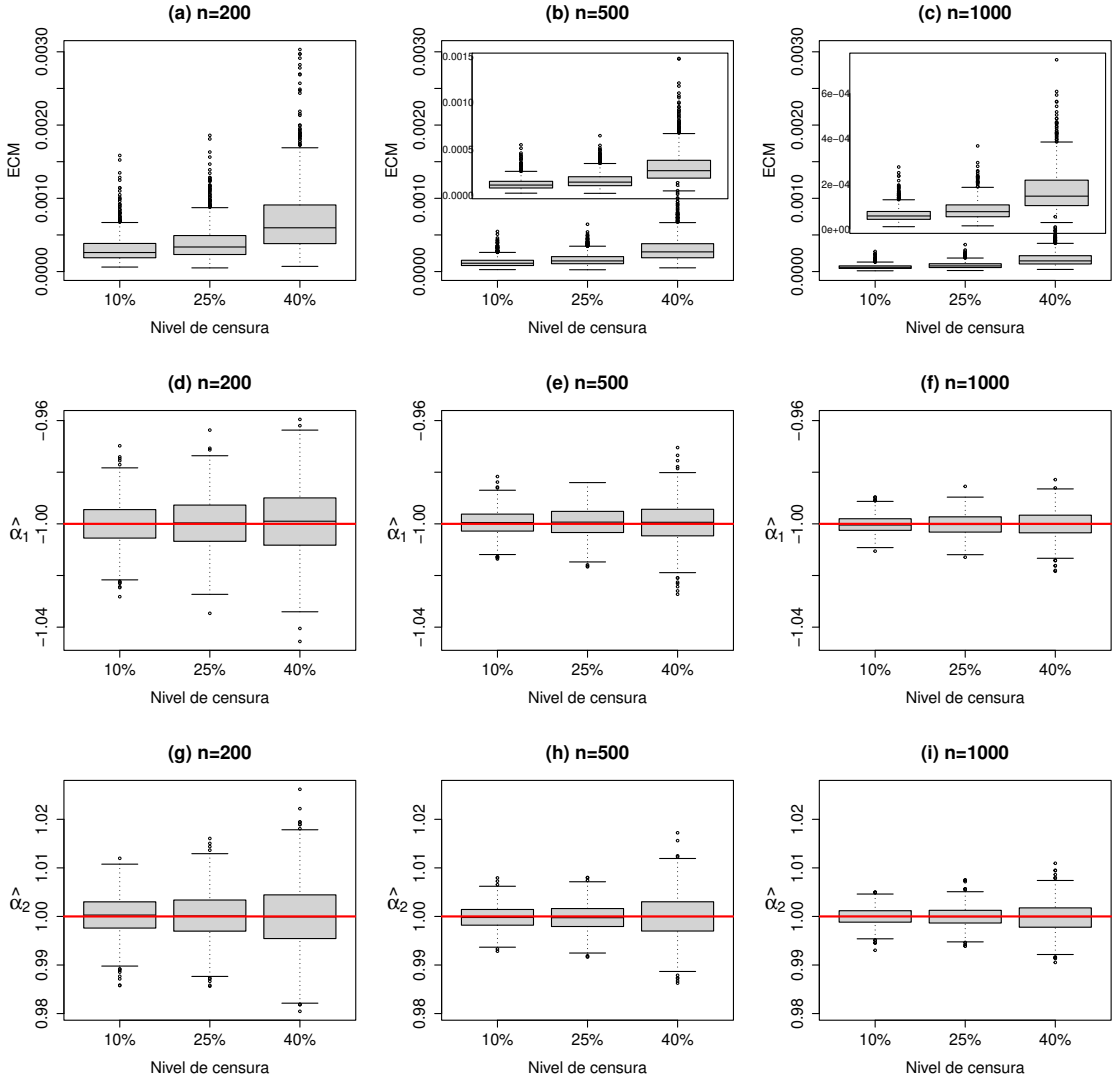


Figura 3.25: Resultados del estudio de simulación para la función logística: Errores Cuadráticos Medios para la parte no paramétrica , $\widehat{\alpha}_1$ y $\widehat{\alpha}_2$ utilizando diferentes niveles de censura y tamaños de muestra



3.3.2. Estimaciones de la parte no paramétrica utilizando diferentes niveles de censura y tamaños de muestra

Figura 3.26: Estimación de la parte no paramétrica utilizando diferentes niveles de censura y tamaños de muestra para la función cuadrática

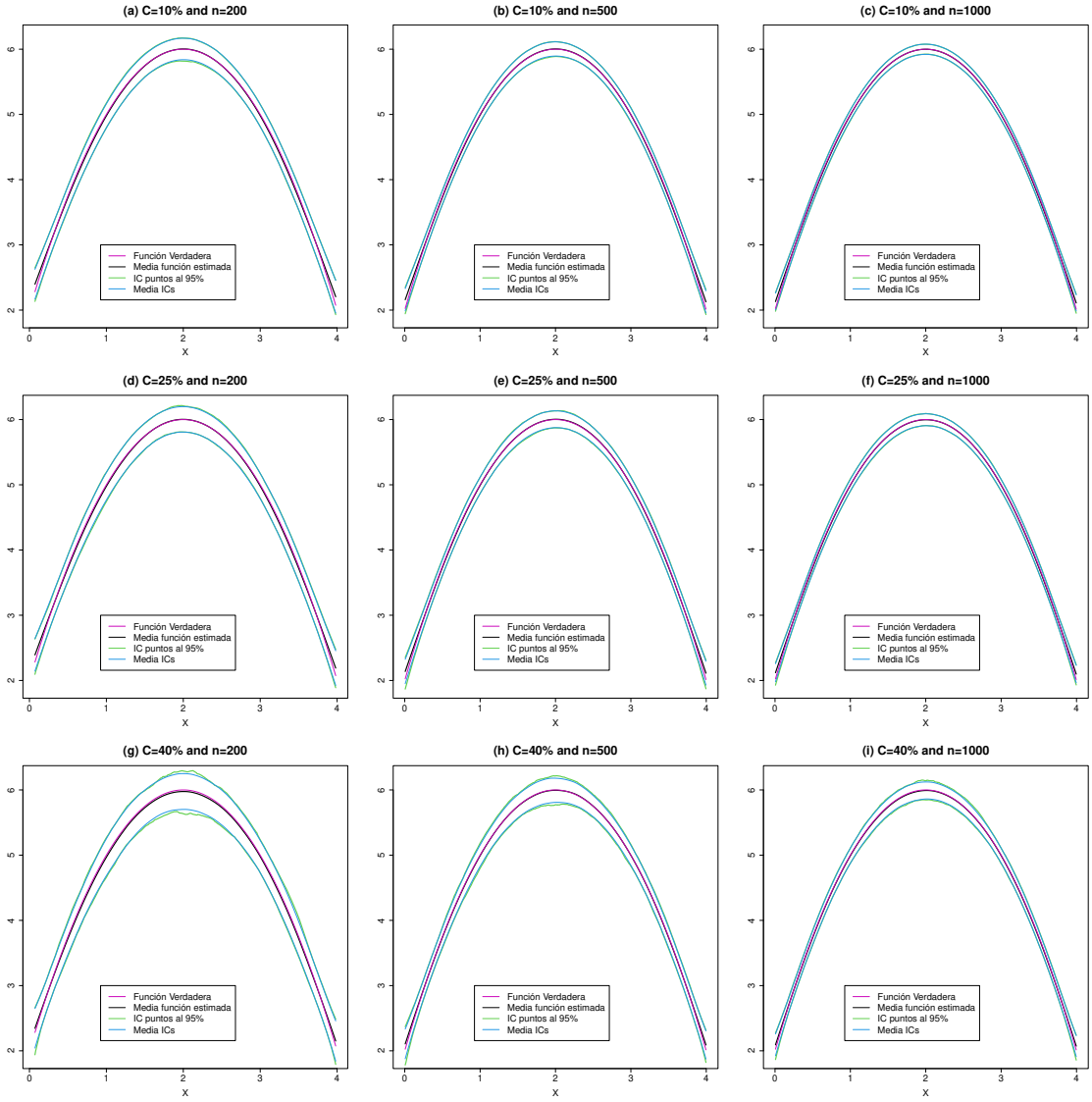


Figura 3.27: Estimación de la parte no paramétrica utilizando diferentes niveles de censura y tamaños de muestra para la función sinusoidal

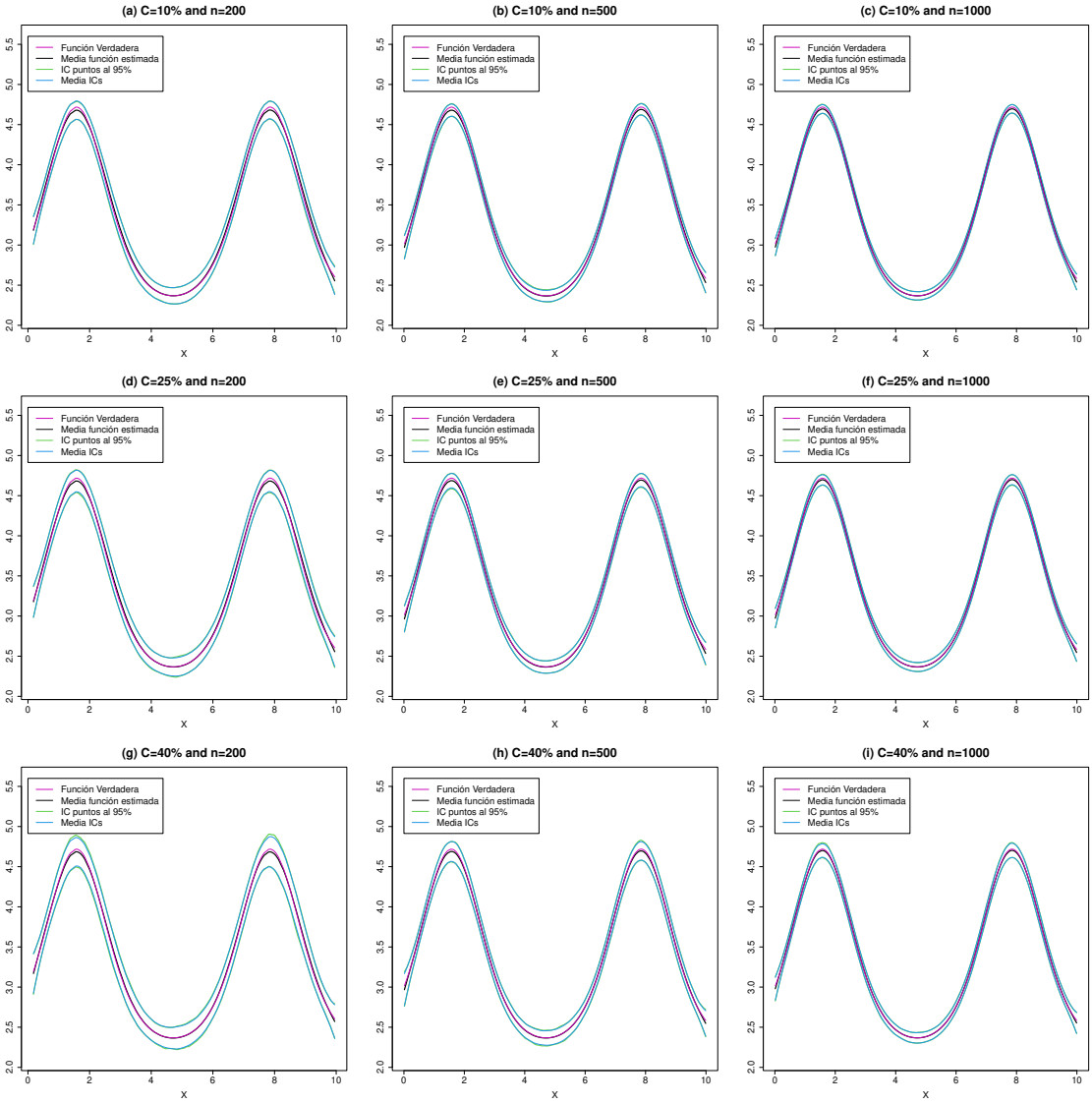
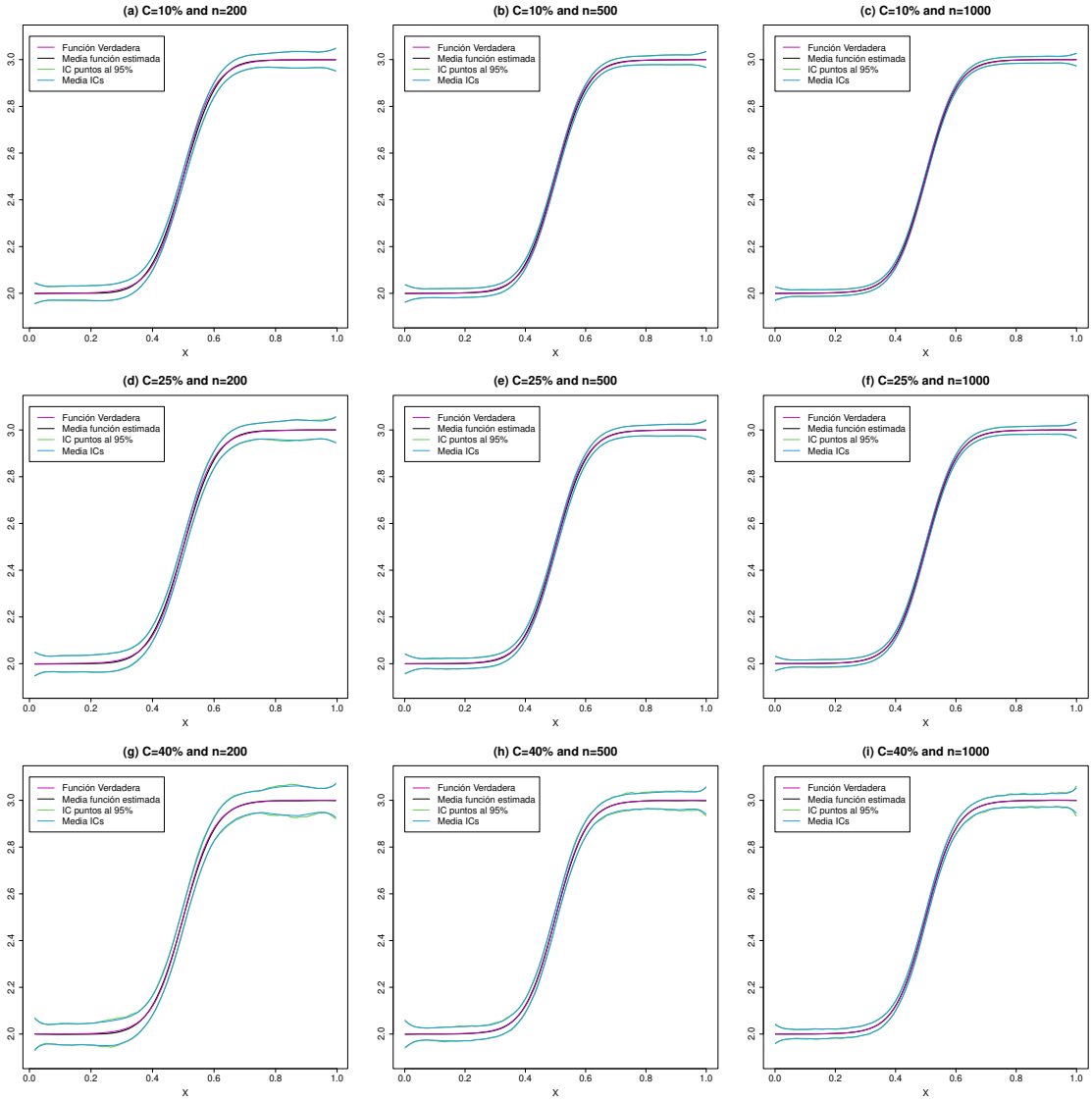


Figura 3.28: Estimación de la parte no paramétrica utilizando diferentes niveles de censura y tamaños de muestra para la función logística



3.3.3. Comparación de resultados según la forma de generar la variable censura: Uniforme versus Normal

Función cuadrática

Tabla 3.13: Resultados del estudio de simulación de la función cuadrática: censura Uniforme versus Normal

	$n = 200$			$n = 500$			$n = 1000$		
% Censura	10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
	ECM ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and MECM ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$ uniforme	3.0901	3.7413	5.9655	1.3243	1.4403	2.3698	0.5213	0.6562	0.9920
$\widehat{\alpha}_1$ normal	2.8802	4.0996	6.1351	1.1905	1.5104	2.0583	0.5262	0.6234	1.0139
$\widehat{\alpha}_1$ dif	0.2100	-0.3583	-0.1696	0.1338	-0.0702	0.3115	-0.0049	0.0328	-0.0219
$\widehat{\alpha}_2$ uniforme	0.7219	0.9065	1.5812	0.2750	0.3022	0.5407	0.1214	0.1810	0.2589
$\widehat{\alpha}_2$ normal	0.7495	0.9158	1.4396	0.2761	0.3200	0.4641	0.1287	0.1844	0.2487
$\widehat{\alpha}_2$ dif	-0.0275	-0.0093	0.1416	-0.0011	-0.0178	0.0766	-0.0073	-0.0033	0.0101
$\widehat{f}$ uniforme	9.7827	12.1700	21.1092	4.1263	5.0560	8.7302	2.1053	2.5797	4.4236
$\widehat{f}$ normal	9.8002	12.8025	20.5185	4.0547	5.0552	7.8583	2.1142	2.5215	4.1507
$\widehat{f}$ dif	-0.0176	-0.6325	0.5906	0.0716	0.0008	0.8719	-0.0088	0.0582	0.2729
	Sesgo empírico								
$\widehat{\alpha}_1$ uniforme	-0.0015	0.0010	0.0113	-0.0032	-0.0029	0.0004	0.0021	0.0035	0.0057
$\widehat{\alpha}_1$ normal	0.0026	0.0014	-0.0023	-0.0021	-0.0001	0.0010	0.0029	0.0000	0.0018
$\widehat{\alpha}_1$ dif	-0.0011	-0.0004	0.0090	0.0011	0.0028	-0.0005	-0.0008	0.0035	0.0040
$\widehat{\alpha}_2$ uniforme	0.0029	0.0021	-0.0026	0.0005	0.0004	-0.0033	0.0004	-0.0004	-0.0019
$\widehat{\alpha}_2$ normal	0.0039	0.0023	-0.0018	0.0003	-0.0001	0.0003	0.0000	-0.0003	-0.0005
$\widehat{\alpha}_2$ dif	-0.0010	-0.0003	0.0008	0.0002	0.0002	0.0030	0.0003	0.0001	0.0015
$\widehat{f}$ uniforme	-0.0003	-0.0015	-0.0163	0.0024	0.0027	0.0007	-0.0023	-0.0025	-0.0058
$\widehat{f}$ normal	-0.0070	-0.0044	0.0016	0.0023	-0.0010	-0.0037	-0.0024	-0.0001	-0.0023
$\widehat{f}$ dif	-0.0067	-0.0029	0.0147	0.0001	0.0017	-0.0030	-0.0001	0.0025	0.0034
	Probabilidades de cobertura de los intervalos de confianza al 95 %								
$\widehat{\alpha}_1$ uniforme	0.9380	0.9550	0.9470	0.9280	0.9500	0.9470	0.9460	0.9460	0.9480
$\widehat{\alpha}_1$ normal	0.9570	0.9310	0.9270	0.9480	0.9510	0.9590	0.9500	0.9620	0.9480
$\widehat{\alpha}_1$ dif	-0.0190	0.0240	0.0200	-0.0200	-0.0010	-0.0120	-0.0040	-0.0160	0.0000
$\widehat{\alpha}_2$ uniforme	0.9450	0.9460	0.9340	0.9410	0.9600	0.9430	0.9600	0.9260	0.9570
$\widehat{\alpha}_2$ normal	0.9510	0.9420	0.9440	0.9460	0.9520	0.9500	0.9560	0.9490	0.9400
$\widehat{\alpha}_2$ dif	-0.0060	0.0040	-0.0100	-0.0050	0.0080	-0.0070	0.0040	-0.0230	0.0170
$\widehat{f}$ uniforme	0.9380	0.9410	0.9230	0.9390	0.9390	0.9250	0.9460	0.9360	0.9330
$\widehat{f}$ normal	0.9420	0.9300	0.9230	0.9400	0.9360	0.9380	0.9410	0.9430	0.9290
$\widehat{f}$ dif	-0.0040	0.0110	0.0000	-0.0010	0.0030	-0.0130	0.0050	-0.0070	0.0040

Función sinusoidal

Tabla 3.14: Resultados del estudio de simulación de la función sinusoidal: censura Uniforme versus Normal

	$n = 200$			$n = 500$			$n = 1000$		
% Censura	10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
	ECM ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and MECM ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$ uniforme	0.8056	1.0597	1.5211	0.2854	0.3623	0.5600	0.1363	0.1536	0.2334
$\widehat{\alpha}_1$ normal	0.7666	1.0216	1.5935	0.3085	0.3797	0.5102	0.1289	0.1597	0.2234
$\widehat{\alpha}_1$ dif	0.0389	0.0381	-0.0724	-0.0231	-0.0174	0.0498	0.0074	-0.0060	0.0100
$\widehat{\alpha}_2$ uniforme	0.1889	0.2656	0.3760	0.0624	0.0865	0.1318	0.0348	0.0442	0.0638
$\widehat{\alpha}_2$ normal	0.1984	0.2365	0.4014	0.0659	0.0785	0.1232	0.0345	0.0384	0.0671
$\widehat{\alpha}_2$ dif	-0.0094	0.0291	-0.0254	-0.0035	0.0081	0.0086	0.0003	0.0058	-0.0033
$\widehat{f}$ uniforme	4.0876	5.2049	7.9701	1.7023	2.0233	3.0724	0.8704	1.0467	1.5450
$\widehat{f}$ normal	4.0758	5.0213	7.9228	1.7166	2.0608	2.9544	0.8653	1.0287	1.5325
$\widehat{f}$ dif	0.0118	0.1836	0.0473	-0.0143	-0.0375	0.1180	0.0051	0.0180	0.0125
	Sesgo empírico								
$\widehat{\alpha}_1$ uniforme	-0.0054	-0.0020	0.0008	0.0009	0.0010	0.0021	0.0006	0.0002	0.0015
$\widehat{\alpha}_1$ normal	-0.0030	-0.0024	-0.0015	0.0004	0.0015	-0.0002	0.0003	0.0004	0.0011
$\widehat{\alpha}_1$ dif	0.0024	-0.0004	-0.0006	0.0005	-0.0005	0.0019	0.0003	-0.0003	0.0004
$\widehat{\alpha}_2$ uniforme	-0.0006	-0.0007	-0.0014	0.0005	0.0006	-0.0009	-0.0002	-0.0002	-0.0014
$\widehat{\alpha}_2$ normal	-0.0009	-0.0005	-0.0021	0.0006	0.0006	0.0003	-0.0003	-0.0002	-0.0005
$\widehat{\alpha}_2$ dif	-0.0003	0.0002	-0.0006	-0.0001	0.0000	0.0006	-0.0001	0.0000	0.0009
$\widehat{f}$ uniforme	0.0067	0.0031	-0.0015	-0.0012	-0.0017	-0.0032	-0.0002	0.0001	-0.0008
$\widehat{f}$ normal	0.0042	0.0031	0.0020	-0.0005	-0.0026	-0.0007	0.0001	0.0000	-0.0010
$\widehat{f}$ dif	0.0025	0.0000	-0.0005	0.0007	-0.0010	0.0026	0.0001	0.0001	-0.0003
	Probabilidades de cobertura de los intervalos de confianza al 95 %								
$\widehat{\alpha}_1$ uniforme	0.9440	0.9360	0.9250	0.9560	0.9550	0.9440	0.9480	0.9560	0.9400
$\widehat{\alpha}_1$ normal	0.9510	0.9450	0.9240	0.9400	0.9500	0.9400	0.9420	0.9420	0.9460
$\widehat{\alpha}_1$ dif	-0.0070	-0.0090	0.0010	0.0160	0.0050	0.0040	0.0060	0.0140	-0.0060
$\widehat{\alpha}_2$ uniforme	0.9490	0.9300	0.9380	0.9520	0.9380	0.9440	0.9440	0.9430	0.9620
$\widehat{\alpha}_2$ normal	0.9420	0.9530	0.9310	0.9540	0.9570	0.9430	0.9490	0.9630	0.9420
$\widehat{\alpha}_2$ dif	0.0070	-0.0230	0.0070	-0.0020	-0.0190	0.0010	-0.0050	-0.0200	0.0200
$\widehat{f}$ uniforme	0.9300	0.9270	0.9180	0.9320	0.9410	0.9320	0.9420	0.9380	0.9410
$\widehat{f}$ normal	0.9340	0.9310	0.9190	0.9330	0.9360	0.9360	0.9390	0.9410	0.9400
$\widehat{f}$ dif	-0.0040	-0.0040	-0.0010	-0.0010	0.0050	-0.0040	0.0030	-0.0030	0.0010

Función logística

Tabla 3.15: Resultados del estudio de simulación de la función logística: censura Uniforme versus Normal

	$n = 200$			$n = 500$			$n = 1000$		
% Censura	10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
	ECM ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and MECM ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$ uniforme	0.0721	0.0980	0.1723	0.0242	0.0365	0.0641	0.0110	0.0164	0.0269
$\widehat{\alpha}_1$ normal	0.0732	0.0869	0.1363	0.0267	0.0326	0.0470	0.0116	0.0145	0.0219
$\widehat{\alpha}_1$ dif	-0.0011	0.0112	0.0360	-0.0025	0.0038	0.0171	-0.0006	0.0019	0.0050
$\widehat{\alpha}_2$ uniforme	0.0174	0.0248	0.0460	0.0061	0.0081	0.0192	0.0030	0.0043	0.0090
$\widehat{\alpha}_2$ normal	0.0169	0.0221	0.0373	0.0057	0.0075	0.0123	0.0031	0.0038	0.0059
$\widehat{\alpha}_2$ dif	0.0005	0.0027	0.0087	0.0004	0.0006	0.0069	-0.0001	0.0005	0.0031
$\widehat{f}$ uniforme	0.3090	0.3974	0.7105	0.1277	0.1644	0.3106	0.0650	0.0854	0.1690
$\widehat{f}$ normal	0.3125	0.3829	0.6037	0.1281	0.1570	0.2283	0.0669	0.0797	0.1252
$\widehat{f}$ dif	-0.0034	0.0146	0.1068	-0.0003	0.0073	0.0823	-0.0019	0.0057	0.0437
	Sesgo empírico								
$\widehat{\alpha}_1$ uniforme	0.0001	0.0003	0.0010	0.0004	0.0006	0.0004	-0.0003	-0.0002	-0.0001
$\widehat{\alpha}_1$ normal	0.0010	0.0006	-0.0002	0.0004	0.0008	0.0003	-0.0001	-0.0000	-0.0003
$\widehat{\alpha}_1$ dif	-0.0009	-0.0003	0.0008	0.0000	-0.0001	0.0001	0.0002	0.0002	-0.0002
$\widehat{\alpha}_2$ uniforme	0.0003	0.0002	-0.0001	-0.0002	-0.0003	-0.0000	-0.0000	-0.0000	-0.0001
$\widehat{\alpha}_2$ normal	0.0003	-0.0001	0.0001	-0.0003	-0.0002	-0.0001	-0.0001	-0.0001	-0.0001
$\widehat{\alpha}_2$ dif	0.0000	0.0000	-0.0001	-0.0001	0.0001	-0.0001	-0.0001	-0.0001	0.0000
$\widehat{f}$ uniforme	-0.0004	-0.0004	-0.0010	-0.0002	-0.0003	-0.0004	0.0003	0.0002	0.0001
$\widehat{f}$ normal	-0.0013	-0.0006	-0.0001	-0.0000	-0.0007	0.0001	0.0002	0.0002	0.0003
$\widehat{f}$ dif	-0.0009	-0.0002	0.0008	0.0002	-0.0004	0.0003	0.0001	0.0000	-0.0002
	Probabilidades de cobertura de los intervalos de confianza al 95 %								
$\widehat{\alpha}_1$ uniforme	0.9440	0.9550	0.9330	0.9530	0.9440	0.9410	0.9460	0.9410	0.9480
$\widehat{\alpha}_1$ normal	0.9450	0.9490	0.9420	0.9510	0.9460	0.9470	0.9590	0.9480	0.9410
$\widehat{\alpha}_1$ dif	-0.0010	0.0060	-0.0090	0.0020	-0.0020	-0.0060	-0.0130	-0.0070	0.0070
$\widehat{\alpha}_2$ uniforme	0.9500	0.9300	0.9240	0.9390	0.9470	0.9380	0.9570	0.9380	0.9560
$\widehat{\alpha}_2$ normal	0.9420	0.9450	0.9190	0.9570	0.9420	0.9490	0.9420	0.9490	0.9470
$\widehat{\alpha}_2$ dif	0.0080	-0.0150	0.0050	-0.0180	0.0050	-0.0110	0.0150	-0.0110	0.0090
$\widehat{f}$ uniforme	0.9440	0.9390	0.9160	0.9430	0.9380	0.9300	0.9490	0.9410	0.9380
$\widehat{f}$ normal	0.9440	0.9380	0.9270	0.9410	0.9430	0.9420	0.9440	0.9430	0.9380
$\widehat{f}$ dif	0.0000	0.0010	-0.0110	0.0020	-0.0050	-0.0120	0.0050	-0.0020	0.0000

3.3.4. Comparación de resultados según la distribución del error: Normal versus Weibull

Función cuadrática

Tabla 3.16: Resultados del estudio de simulación de la función cuadrática: error Normal versus Weibull

	$n = 200$			$n = 500$			$n = 1000$		
% Censura	10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
	ECM ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and MECM ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$ normal	3.0901	3.7413	5.9655	1.3243	1.4403	2.3698	0.5213	0.6562	0.9920
$\widehat{\alpha}_1$ weibull	2.9503	3.7218	5.8524	1.1325	1.3875	2.1032	0.5225	0.6148	0.9362
$\widehat{\alpha}_1$ dif	0.1398	0.0195	0.1131	0.1918	0.0528	0.2665	-0.0012	0.0414	0.0558
$\widehat{\alpha}_2$ normal	0.7219	0.9065	1.5812	0.2750	0.3022	0.5407	0.1214	0.1810	0.2589
$\widehat{\alpha}_2$ weibull	0.7463	0.9186	1.3748	0.2478	0.2935	0.5190	0.1297	0.1537	0.2370
$\widehat{\alpha}_2$ dif	-0.0243	-0.0121	0.2064	0.0272	0.0087	0.0217	-0.0083	0.0273	0.0219
$\widehat{f}$ normal	9.7827	12.1700	21.1092	4.1263	5.0560	8.7302	2.1053	2.5797	4.4236
$\widehat{f}$ weibull	9.0430	11.0839	18.2726	3.6671	4.3299	7.0197	1.8880	2.2861	3.5127
$\widehat{f}$ dif	0.7397	1.0861	2.8366	0.4592	0.7261	1.7105	0.2173	0.2936	0.9110
	Sesgo empírico								
$\widehat{\alpha}_1$ normal	-0.0015	0.0010	0.0113	-0.0032	-0.0029	0.0004	0.0021	0.0035	0.0057
$\widehat{\alpha}_1$ weibull	-0.0010	-0.0026	0.0051	-0.0033	-0.0031	0.0022	0.0009	0.0033	0.0050
$\widehat{\alpha}_1$ dif	0.0005	-0.0016	0.0062	-0.0001	-0.0002	-0.0018	0.0012	0.0002	0.0007
$\widehat{\alpha}_2$ normal	0.0029	0.0021	-0.0026	0.0005	0.0004	-0.0033	0.0004	-0.0004	-0.0019
$\widehat{\alpha}_2$ weibull	0.0022	0.0017	-0.0025	0.0012	0.0007	-0.0017	0.0006	0.0008	-0.0015
$\widehat{\alpha}_2$ dif	0.0007	0.0004	0.0001	-0.0007	-0.0003	0.0017	-0.0002	-0.0004	0.0005
$\widehat{f}$ normal	-0.0003	-0.0015	-0.0163	0.0024	0.0027	0.0007	-0.0023	-0.0025	-0.0058
$\widehat{f}$ weibull	-0.0014	-0.0006	-0.0092	0.0022	0.0021	-0.0037	-0.0015	-0.0044	-0.0061
$\widehat{f}$ dif	-0.0011	0.0009	0.0071	0.0002	0.0006	-0.0031	0.0008	-0.0019	-0.0004
	Probabilidades de cobertura de los intervalos de confianza al 95 %								
$\widehat{\alpha}_1$ normal	0.9380	0.9550	0.9470	0.9280	0.9500	0.9470	0.9460	0.9460	0.9480
$\widehat{\alpha}_1$ weibull	0.9470	0.9580	0.9540	0.9540	0.9570	0.9560	0.9430	0.9580	0.9560
$\widehat{\alpha}_1$ dif	0.0090	-0.0030	-0.0010	0.0180	-0.0070	-0.0030	-0.0030	-0.0040	-0.0040
$\widehat{\alpha}_2$ normal	0.9450	0.9460	0.9340	0.9410	0.9600	0.9430	0.9600	0.9260	0.9570
$\widehat{\alpha}_2$ weibull	0.9500	0.9530	0.9380	0.9510	0.9610	0.9530	0.9510	0.9590	0.9580
$\widehat{\alpha}_2$ dif	0.0050	0.0010	0.0040	0.0080	-0.0010	0.0040	0.0090	0.0150	-0.0010
$\widehat{f}$ normal	0.9380	0.9410	0.9230	0.9390	0.9390	0.9250	0.9460	0.9360	0.9330
$\widehat{f}$ weibull	0.9550	0.9530	0.9390	0.9540	0.9570	0.9530	0.9540	0.9560	0.9510
$\widehat{f}$ dif	0.0070	0.0060	0.0160	0.0070	0.0040	0.0220	0	0.0080	0.0160

Función sinusoidal

Tabla 3.17: Resultados del estudio de simulación de la función sinusoidal: error Normal versus Weibull

	$n = 200$			$n = 500$			$n = 1000$		
% Censura	10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
	ECM ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and MECM ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$ normal	0.8056	1.0597	1.5211	0.2854	0.3623	0.5600	0.1363	0.1536	0.2334
$\widehat{\alpha}_1$ weibull	0.7714	1.0138	1.4871	0.2997	0.3376	0.5465	0.1211	0.1637	0.2312
$\widehat{\alpha}_1$ dif	0.0342	0.0459	0.0340	-0.0142	0.0247	0.0135	0.0151	-0.0101	0.0021
$\widehat{\alpha}_2$ normal	0.1889	0.2656	0.3760	0.0624	0.0865	0.1318	0.0348	0.0442	0.0638
$\widehat{\alpha}_2$ weibull	0.1742	0.2507	0.4123	0.0659	0.0843	0.1353	0.0353	0.0425	0.0639
$\widehat{\alpha}_2$ dif	0.0147	0.0149	-0.0362	-0.0036	0.0023	-0.0035	-0.0005	0.0017	-0.0001
$\widehat{f}$ normal	4.0876	5.2049	7.9701	1.7023	2.0233	3.0724	0.8704	1.0467	1.5450
$\widehat{f}$ weibull	3.8214	4.8065	7.4514	1.5904	1.8785	2.8877	0.8108	1.0015	1.4402
$\widehat{f}$ dif	0.2662	0.3984	0.5187	0.1119	0.1448	0.1847	0.0596	0.0451	0.1048
	Sesgo empírico								
$\widehat{\alpha}_1$ normal	-0.0054	-0.0020	0.0008	0.0009	0.0010	0.0021	0.0006	0.0002	0.0015
$\widehat{\alpha}_1$ weibull	-0.0046	-0.0036	-0.0010	0.0009	0.0007	0.0021	-0.0000	-0.0000	0.0017
$\widehat{\alpha}_1$ dif	0.0008	-0.0016	-0.0002	0	0.0003	0	0.0006	0.0001	-0.0002
$\widehat{\alpha}_2$ normal	-0.0006	-0.0007	-0.0014	0.0005	0.0006	-0.0009	-0.0002	-0.0002	-0.0014
$\widehat{\alpha}_2$ weibull	-0.0015	-0.0000	-0.0010	0.0008	0.0006	-0.0014	0.0003	-0.0002	-0.0008
$\widehat{\alpha}_2$ dif	-0.0009	0.0007	0.0005	-0.0003	0	-0.0005	-0.0001	0	0.0005
$\widehat{f}$ normal	0.0067	0.0031	-0.0015	-0.0012	-0.0017	-0.0032	-0.0002	0.0001	-0.0008
$\widehat{f}$ weibull	0.0061	0.0036	-0.0008	-0.0018	-0.0015	-0.0032	-0.0002	0.0001	-0.0018
$\widehat{f}$ dif	0.0006	-0.0005	0.0008	-0.0007	0.0002	0	0	0	-0.0010
	Probabilidades de cobertura de los intervalos de confianza al 95 %								
$\widehat{\alpha}_1$ normal	0.9440	0.9360	0.9250	0.9560	0.9550	0.9440	0.9480	0.9560	0.9400
$\widehat{\alpha}_1$ weibull	0.9470	0.9400	0.9330	0.9510	0.9500	0.9410	0.9580	0.9510	0.9470
$\widehat{\alpha}_1$ dif	0.0030	0.0040	0.0080	0.0050	0.0050	-0.0030	-0.0060	0.0050	0.0070
$\widehat{\alpha}_2$ normal	0.9490	0.9300	0.9380	0.9520	0.9380	0.9440	0.9440	0.9430	0.9620
$\widehat{\alpha}_2$ weibull	0.9610	0.9480	0.9360	0.9420	0.9490	0.9450	0.9440	0.9460	0.9550
$\widehat{\alpha}_2$ dif	-0.0100	0.0180	-0.0020	-0.0060	0.0110	0.0010	0	0.0030	0.0070
$\widehat{f}$ normal	0.9300	0.9270	0.9180	0.9320	0.9410	0.9320	0.9420	0.9380	0.9410
$\widehat{f}$ weibull	0.9430	0.9400	0.9300	0.9420	0.9480	0.9450	0.9450	0.9460	0.9470
$\widehat{f}$ dif	0.0130	0.0130	0.0120	0.0100	0.0070	0.0130	0.0030	0.0080	0.0060

Función logística

Tabla 3.18: Resultados del estudio de simulación de la función logística: error Normal versus Weibull

	$n = 200$			$n = 500$			$n = 1000$		
% Censura	10 %	25 %	40 %	10 %	25 %	40 %	10 %	25 %	40 %
	ECM ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and MECM ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$ normal	0.0721	0.0980	0.1723	0.0242	0.0365	0.0641	0.0110	0.0164	0.0269
$\widehat{\alpha}_1$ weibull	0.0714	0.1023	0.1601	0.0259	0.0357	0.0630	0.0110	0.0143	0.0272
$\widehat{\alpha}_1$ dif	0.0006	-0.0043	0.0122	-0.0017	0.0007	0.0011	0	0.0020	-0.0003
$\widehat{\alpha}_2$ normal	0.0174	0.0248	0.0460	0.0061	0.0081	0.0192	0.0030	0.0043	0.0090
$\widehat{\alpha}_2$ weibull	0.0164	0.0252	0.0433	0.0058	0.0074	0.0181	0.0030	0.0040	0.0089
$\widehat{\alpha}_2$ dif	0.0010	-0.0004	0.0027	0.0003	0.0007	0.0011	0	0.0003	0.0001
$\widehat{f}$ normal	0.3090	0.3974	0.7105	0.1277	0.1644	0.3106	0.0650	0.0854	0.1690
$\widehat{f}$ weibull	0.2997	0.3940	0.6428	0.1187	0.1548	0.2841	0.0605	0.0784	0.1628
$\widehat{f}$ dif	0.0093	0.0034	0.0677	0.0090	0.0096	0.0266	0.0045	0.0070	0.0061
	Sesgo empírico								
$\widehat{\alpha}_1$ normal	0.0001	0.0003	0.0010	0.0004	0.0006	0.0004	-0.0003	-0.0002	-0.0001
$\widehat{\alpha}_1$ weibull	0.0006	0.0001	0.0003	0.0009	0.0007	0.0007	-0.0002	-0.0002	0.0002
$\widehat{\alpha}_1$ dif	-0.0005	0.0002	0.0007	-0.0005	-0.0001	-0.0003	0.0001	0	-0.0001
$\widehat{\alpha}_2$ normal	0.0003	0.0002	-0.0001	-0.0002	-0.0003	-0.0000	-0.0000	-0.0000	-0.0001
$\widehat{\alpha}_2$ weibull	0.0001	0.0003	-0.0001	-0.0003	-0.0003	-0.0002	-0.0000	0.0001	-0.0001
$\widehat{\alpha}_2$ dif	0.0002	-0.0001	0	-0.0001	0	-0.0002	0	-0.0001	0
$\widehat{f}$ normal	-0.0004	-0.0004	-0.0010	-0.0002	-0.0003	-0.0004	0.0003	0.0002	0.0001
$\widehat{f}$ weibull	-0.0009	-0.0005	-0.0006	-0.0007	-0.0005	-0.0007	0.0003	0.0001	-0.0001
$\widehat{f}$ dif	-0.0005	-0.0001	0.0004	-0.0005	-0.0002	-0.0003	0	0.0001	0
	Probabilidades de cobertura de los intervalos de confianza al 95 %								
$\widehat{\alpha}_1$ normal	0.9440	0.9550	0.9330	0.9530	0.9440	0.9410	0.9460	0.9410	0.9480
$\widehat{\alpha}_1$ weibull	0.9470	0.9470	0.9350	0.9520	0.9460	0.9370	0.9630	0.9510	0.9430
$\widehat{\alpha}_1$ dif	0.0030	0.0020	0.0020	0.0010	0.0020	-0.0040	-0.0090	0.0080	-0.0050
$\widehat{\alpha}_2$ normal	0.9500	0.9300	0.9240	0.9390	0.9470	0.9380	0.9570	0.9380	0.9560
$\widehat{\alpha}_2$ weibull	0.9550	0.9290	0.9240	0.9540	0.9570	0.9450	0.9530	0.9550	0.9450
$\widehat{\alpha}_2$ dif	-0.0050	-0.0010	0	0.0070	-0.0040	0.0070	0.0040	0.0070	0.0010
$\widehat{f}$ normal	0.9440	0.9390	0.9160	0.9430	0.9380	0.9300	0.9490	0.9410	0.9380
$\widehat{f}$ weibull	0.9480	0.9430	0.9300	0.9530	0.9500	0.9410	0.9550	0.9510	0.9400
$\widehat{f}$ dif	0.0040	0.0040	0.0140	0.0040	0.0120	0.0110	-0.0040	0.0080	0.0020

3.3.5. Aplicación empírica: datos CBP

Tabla 3.19: Estimación de los coeficientes de regresión y sus desviaciones típicas para el conjunto de datos de Cirrosis Biliar Primaria de la Clínica Mayo a partir de los métodos AFT, Stute y CPS

	edad	edema	trt	log(albumina)	log(bili)
AFT	-0.0246	-0.7692	-0.0627	1.4880	-0.5356
	(0.0065)	(0.2303)	(0.1273)	(0.5268)	(0.0694)
Stute	-0.0166	-0.9249	-0.0950	1.6161	-0.3028
	(0.0076)	(0.3489)	(0.1371)	(0.6015)	(0.0732)
CPS	-0.0168	-0.9163	-0.0991	1.6197	-0.3061
	(0.0064)	(0.1900)	(0.1291)	(0.4578)	(0.0633)

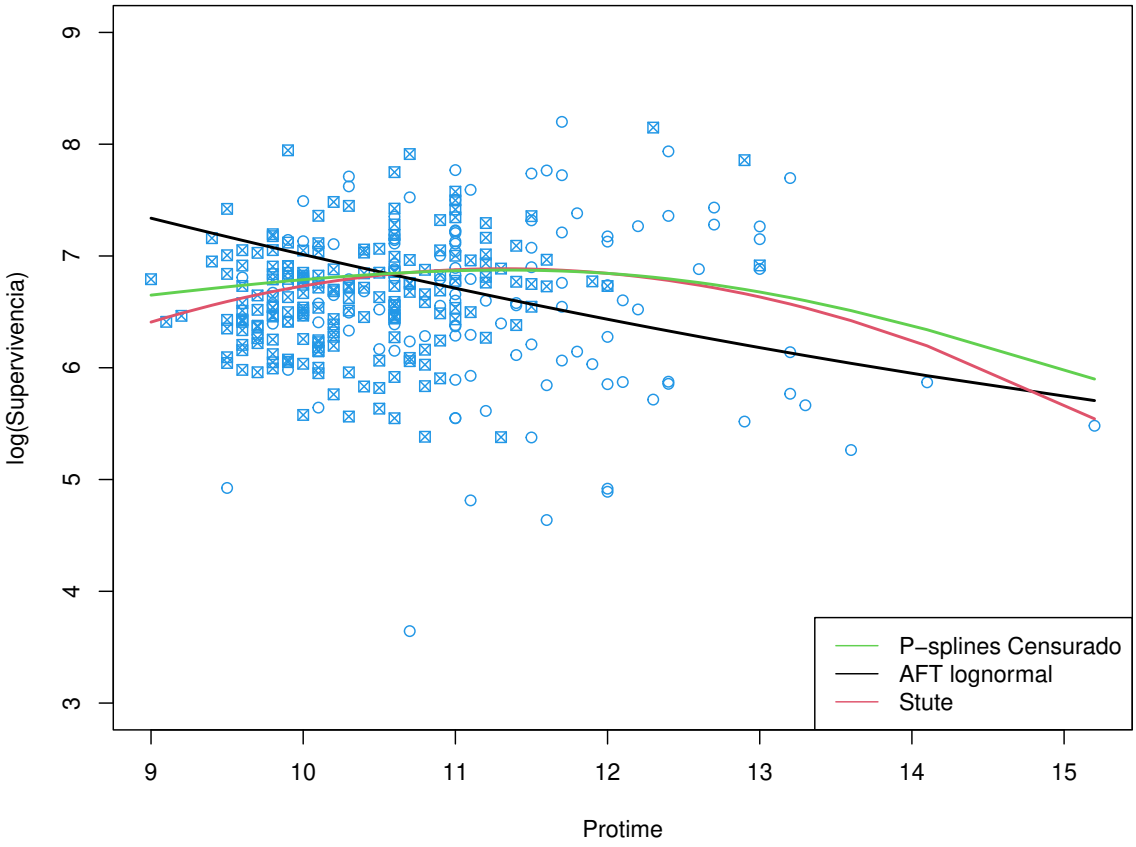
El conjunto de datos de Cirrosis Biliar Primaria contiene información de 418 pacientes de la Clínica Mayo con colangitis biliar primaria (CBP), anteriormente denominada cirrosis biliar primaria, una enfermedad autoinmune del hígado. Los primeros 312 casos del conjunto de datos participaron en un ensayo de la Clínica Mayo sobre CBP realizado entre 1974 y 1984 en el que se comparó el fármaco D-penicilamina (tratamiento) con un placebo. El conjunto de datos proporciona información sobre el tiempo de supervivencia observado desde la fecha de registro en el ensayo , además de un gran número de variables clínicas, bioquímicas, serológicas e histológicas como la edad del paciente en el momento del primer diagnóstico, la gravedad del edema (0 sin edema, 0,5 moderado y 1 para edema grave), valores sanguíneos relacionados con la función hepática como bilirrubina, albúmina, fosfotasa alcalina y tiempo de protrombina junto a otras variables explicativas, y un indicador del estado del paciente (vivo o muerto) en julio de 1986. El conjunto de datos puede descargarse del paquete R *survival* (Therneau, 2021; R Core Team, 2018). Además existe un conjunto independiente de casos adicionales, 106 pacientes con colangitis biliar primaria de la Clínica Mayo que eran elegibles para el ensayo pero se negaron a participar. Este conjunto de datos se ha utilizado anteriormente en la literatura, por ejemplo, en los estudios de Dickson et al. (1989), Therneau and Grambsch (2000) y Fleming and Harrington (2005), en modelos de regresión censurada.

Los estudios de Therneau and Grambsch (2000) y Fleming and Harrington (2005) analizan la relación entre las distintas covariables y la supervivencia del paciente. En ellos se llega a la conclusión de que la edad, el edema, los logaritmos de bilirrubina y albúmina y el tiempo de protrombina son las variables que mejor explican la supervivencia de los pacientes. Asimismo, estos estudios analizan la necesidad de transformar alguna de las variables continuas en el modelo propuesto, concluyendo que es probable que la relación entre el tiempo de protrombina (protime) y la supervivencia de los pacientes no sea lineal.

En esta aplicación se incorpora la variable protime en el modelo de forma flexible, suponiendo únicamente que el tiempo de protrombina entra en el modelo a través de una función suave desconocida $f(\cdot)$:

$$log(T) = \alpha_1 + \alpha_2 edad + \alpha_3 edema + \alpha_4 trt + \alpha_5 log(albumina) + \alpha_6 log(bili) + f(protime) + \epsilon \tag{3.3}$$

Figura 3.29: Estimación componente no paramétrica utilizando tres metodologías: AFT lognormal, enfoque de Stute y estimador CPS



Se ha estimado el modelo (3.3) utilizando el método P-splines censurado propuesto en la sección 2.3. Para evaluar el comportamiento de este estimador se ha propuesto como alternativa una relación cuadrática entre el logaritmo de la supervivencia y la variable protime, *i.e.*, $f(protime) = \alpha_7 protime + \alpha_8 protime^2$ en la ecuación (3.3). Asumiendo que esta especificación paramétrica es correcta, se pueden utilizar dos metodologías propuestas en la literatura sobre análisis de supervivencia para ajustar el modelo (3.3). Estos estimadores pueden utilizarse como referencia para evaluar el rendimiento del método P-splines censurado.

El primer enfoque, el más restrictivo, es la metodología paramétrica de los modelos de duración acelerada (*Accelerated Failure Time*, AFT, Kalbfleisch and Prentice, 2002), que suponen conocidas la distribución de probabilidad de la variable respuesta y la forma funcional que relaciona la variable protime y la supervivencia del paciente, y que estima los coeficientes α del modelo utilizando el estimador de máxima verosimilitud. Así, en este caso se propone un modelo AFT lognormal y se estiman los coeficientes α asumiendo una distribución de probabilidad normal.

La segunda metodología, propuesta por Stute (1993), es menos restrictiva en el sentido de que no necesita conocer la distribución de probabilidad de la variable de respuesta, pero también supone una forma funcional conocida, en el caso que nos ocupa la cuadrática. Es decir, necesita conocer la forma de la relación entre la variable respuesta y la covariable. Esta metodología estima los coeficientes mediante mínimos cuadrados ponderados utilizando los pesos Kaplan-Meier (Stute, 1993).

La tabla 3.19 presenta las estimaciones de los componentes paramétricos del modelo (3.3) utilizando estos tres métodos. Puede observarse que los tres métodos generan estimaciones similares y dan como resultado una estimación del modelo biológicamente razonable. Como ya se ha descrito en la bibliografía, los tres métodos coinciden en que el tratamiento con el fármaco D-penicilamina (trt) no tiene un efecto significativo en la supervivencia de los pacientes. La figura 3.29 muestra las estimaciones de la función $f(protime)$ para los tres enfoques, con el diagrama de dispersión del logaritmo del tiempo de supervivencia observado frente al tiempo de protrombina. Los pacientes indicados con ◯ han fallecido y los indicados con ⊠ están vivos en julio de 1986; es decir, los pacientes fallecidos tienen tiempos de supervivencia no censurados y los pacientes vivos tienen tiempos de supervivencia censurados.

En conclusión, el buen funcionamiento de la metodología AFT y de la propuesta de Stute depende de la correcta especificación de la relación entre la duración y la variable protime. En esta aplicación parece que la relación entre el logaritmo de la supervivencia y el tiempo de protrombina es cuadrática, por lo que ambas metodologías funcionan razonablemente bien. La propuesta de estimación que se desarrolla en esta tesis no necesita asumir una forma funcional paramétrica específica y, sin embargo, estima adecuadamente la relación obteniendo resultados muy similares a los anteriores. Sin embargo, si la forma funcional se hubiera elegido erróneamente estos métodos paramétricos tendrían un grave problema de especificación incorrecta y, por tanto, llevarían a conclusiones erróneas. En este sentido, se puede considerar los P-splines censurados como una solución robusta a la especificación errónea del modelo.

Capítulo 4

Código en R

4.1. Código para el caso univariante

En este apartado se muestra el código de R utilizado para calcular el estimador P-splines censurado (*pswc*, sección 2.1.2) y el estimador en un modelo GAM corrigiendo el efecto de la censura con los pesos Kaplan-Meier (*gamkm*: un regresor, *gamkm2d2l*: dos regresores, sección 3.1.2), además de una función necesaria para ambas que permite calcular los pesos Kaplan-Meier (*kmw.cp*, sección 2.1.2, ecuación 2.4).

El código ha sido escrito usando R-3.4.4 (plataforma: x86_64-pc-linux-gnu, 64-bit) con versiones de los paquetes mgcv_1.8-23, nlme_3.1-131.1 y survival_2.41-3.

Función de R 4.1: *kmw.cp*: función para calcular los pesos Kaplan-Meier

```
## ------------------------------------------------------
## Nombre de la función: kmw.cp
## ------------------------------------------------------
## Descripción: Calcula los pesos Kaplan-Meier
## ------------------------------------------------------
## Librerías requeridas: survival
## ------------------------------------------------------
## Uso: kmw.cp(y, cen)
##
##      y:    tiempo de seguimiento
##      cen:  indicador de estado,
##            usualmente 0/FALSE=vivo, 1/TRUE=fallecido
## ------------------------------------------------------
## Valor: pesos Kaplan-Meier
## ------------------------------------------------------


kmw.cp <- function(y, cen) {
    require(survival)
    y.un <- unique(sort(y))
    kme <- survfit(Surv(y, cen) ~ 1)
    FKM <- 1 - kme$surv
    yjs <- kme$time
    yjs.fr <- kme$n.event
    yjs.fr[yjs.fr == 0] <- 1
    w.i <- c(FKM[1], diff(FKM))/yjs.fr
    Wkm <- w.i[match(y, y.un)]
    # Observaciones censuradas: pesos igual a 0
    Wkm[cen == 0] <- 0
    return(Wkm)
}
## fin de la función kmw.cp
```

Función de R 4.2: *pswc*: función para el cálculo del estimador P-splines censurado

```
## ------------------------------------------------------
## Nombre de la función: pswc
## ------------------------------------------------------
## Descripción: extensión del método P-splines
##              de Eilers & Marx 1996 al caso de datos
##              censurados utilizando pesos Kaplan-Meier
## ------------------------------------------------------
## Librerías requeridas: survival, splines
## ------------------------------------------------------
## Uso: pswc(x, y, cen, grado.b, ndx, lambda)
##
##      x:    regresor
##      y:    tiempo de seguimiento
##      cen:  indicador de estado,
##            usualmente 0/FALSE=vivo, 1/TRUE=fallecido
##      grado.b:  grado del polinomio a trozos
##                (3= splines cúbicos)
##      ndx+1:    numero de nodos
##      lambda:   parámetro de suavizado
## ------------------------------------------------------
## Valor: estimación noparamétrica de la relación entre
##        la variable respuesta (y) y el regresor (x)
## ------------------------------------------------------


pswc <- function(x, y, cen, grado.b, ndx, lambda) {
    require(survival)
    require(splines)
    # Calcula pesos Kaplan-Meier
    Wkm <- kmw.cp(y, cen)

    ## Ordenado por X
    x.ox <- x[order(x)]
    y.ox <- y[order(x)]
    cen.ox <- cen[order(x)]
    ind.ox <- c(1:length(x))[order(x)]

    ## Ordenado por Y y por censura
    y.oy <- y[order(y, !cen)]
    x.oy <- x[order(y, !cen)]
    W.oy <- Wkm[order(y, !cen)]
    ind.oy <- c(1:length(y))[order(y, !cen)]

    ## Base B-spline y Nodos
    if (missing(grado.b)) {
        grado.b <- 3
    }
```

```
    unix <- min(abs(x[x != 0]))/1000
    xl <- min(x) - unix
    xr <- max(x) + unix
    dx <- (xr - xl)/ndx
    knots <- seq(xl - grado.b * dx, xr + grado.b * dx,
                 by = dx)
    BBB.ox  <- spline.des(knots, x.ox,
                          grado.b + 1, 0 * x.ox)$design
    # Orden inicial.
    # Si las X's están ordfenadas, no hace nada
    BBB.ini <- BBB.ox[order(ind.ox),]
    BBB.oy  <- BBB.ini[order(y, !cen), ]

    ## Matriz representación del operador de diferencias
    ## (diferencias de orden 2)
    D = diff(diff(diag(ncol(BBB.oy))))
    DtD = t(D) %*% D

    ## Matriz con los pesos Kaplan-Meier
    matW.oy <- diag(W.oy)
    ## Elección del nivel de suavizado óptimo via GCVc
    if (missing(lambda)) {
        pswcGCVc <- function(lambda, BBB.oy, matW.oy, DtD,
                             y.oy, W.oy) {
            pena <- solve(t(BBB.oy) %*% matW.oy %*%
                          BBB.oy + (lambda * DtD)) %*%
                          t(BBB.oy) %*% matW.oy %*% y.oy
            s <- sum(W.oy * (y.oy - BBB.oy %*% pena)^2)
            hasterisco <- solve(t(BBB.oy) %*% matW.oy %*%
                                BBB.oy + (lambda * DtD)) %*%
                                t(BBB.oy) %*% matW.oy %*% BBB.oy
            trazah <- sum(diag(hasterisco))
            gcv <- s/(nrow(BBB.oy) - trazah)^2
            return(gcv)
        }
        # Valor óptimo del parámetro de suavizado lambda
        lambda <- optimize(pswcGCVc, c(0, 10000),
                           tol = 1e-06, BBB.oy = BBB.oy,
                           matW.oy = matW.oy, DtD = DtD,
                           y.oy = y.oy, W.oy = W.oy)[[1]]
    }

    ## Estimación
    pena <- solve(t(BBB.oy) %*% matW.oy %*%
                  BBB.oy + (lambda * DtD)) %*%
                  t(BBB.oy) %*% matW.oy %*% y.oy
    return(BBB.ini %*% pena)
} ## fin de la función pswc
```

Función de R 4.3: *gamkm*: función para el cálculo del estimador en un modelo GAM censurado con un regresor

```
## ------------------------------------------------------------
## Nombre de la función: gamkm
## ------------------------------------------------------------
## Descripción: función gam con pesos Kaplan-Meier y
##              selección del parámetro de suavizado
##              via GCVc: un regresor
## ------------------------------------------------------------
## Librerías requeridas: mgcv
## ------------------------------------------------------------
## Uso: gamkm(y, x, kn, kmw, base)
##
##      y:      tiempo de seguimiento
##      x:      regresor
##      kn:     dimensión de la base que representa el
##              término de suavizado
##      kmw:    pesos Kaplan-Meier
##      base:   cadena de dos letras que indica la base
##              de suavizado (penalizada) a utilizar
##
## ------------------------------------------------------------
## Valor: objecto de la clase "gam"
## ------------------------------------------------------------

gamkm <- function(y, x, kn, kmw, base) {
    require(mgcv)
    gamkmGCVc <- function(lambda, y, x, kn, kmw, base) {
        nobs <- length(y)
        mod.gam <- gam(y ~ s(x, bs = base, k = kn,
                        m = c(2, 2)), sp = lambda,
                        weights = kmw)
        rss <- sum(kmw * (y - fitted(mod.gam))^2)
        trF <- sum(mod.gam$hat)
        # GCVc
        gcv <- nobs * rss/(nobs - trF)^2
        return(gcv)
    }

    # Selección del parámetro de suavizado
    lambdaopt <- optimize(gamkmGCVc, c(0, 10000),
                          tol = 1e-06, y = y, x = x,
                          kn = kn, kmw = kmw,
                          base = base)[[1]]
    return(gam(y ~ s(x, bs = base, k = kn, m = c(2, 2)),
               sp = lambdaopt, weights = kmw, gamma = 1))
} ## fin de la función gamkm
```

Función de R 4.4: *gamkm2d2l*: función para el cálculo del estimador en un modelo GAM censurado con dos regresores

```
## ------------------------------------------------------
## Nombre de la función: gamkm2d2l
## ------------------------------------------------------
## Descripción: función gam con pesos Kaplan-Meier y
##              selección del parámetro de suavizado
##              via GCVc: dos regresores
## ------------------------------------------------------
## Librerías requeridas: mgcv
## ------------------------------------------------------
## Uso: gamkm2d2l(y, x1, x2, kn, kmw, base)
##
##      y:      tiempo de seguimiento
##      x1:     regresor 1
##      x2:     regresor 2
##      kn:     dimensión de la base que representa el
##              término de suavizado
##      kmw:    pesos Kaplan-Meier
##      base:   cadena de dos letras que indica la base
##              de suavizado (penalizada) a utilizar
## ------------------------------------------------------
## Valor: objecto de la clase "gam"
## ------------------------------------------------------

gamkm2d2l <- function(y, x1, x2, kn, kmw, base) {
    gamkmGCVc <- function(lsp, y, x1, x2, kn, kmw, base) {
        nobs <- length(y)
        mod.gam <- gam(y ~ s(x1, bs = base, k = kn,
                        m = c(2, 2)) + s(x2, bs = base,
                        k = kn, m = c(2, 2)),
                        sp = exp(lsp), weights = kmw)
        rss <- sum(kmw * (y - fitted(mod.gam))^2)
        trF <- sum(mod.gam$hat)
        # GCVc
        gcv <- nobs * rss/(nobs - trF)^2
        return(gcv)
    }
    # Selección del parámetro de suavizado
    lambdaopt <- optim(c(0, 0), gamkmGCVc, y = y, x1 = x1,
                        x2 = x2, kn = kn, kmw = kmw,
                        base = base)
    return(gam(y ~ s(x1, bs = base, k = kn,
                      m = c(2, 2)) + s(x2, bs = base,
                      k = kn, m = c(2, 2)),
                      sp = exp(lambdaopt$par),
                      weights = kmw, gamma = 1))
} ## fin de la función gamkm2d2l
```

4.2. Código para la selección de parámetros

En este apartado se muestra el código de R desarrollado para incorporar las modificaciones que generalizan los criterios de elección de parámetros desarrollados en la sección 2.2 para el estimador P-splines censurado (*pswc* modificada, función para calcular el estimador ckmPS) y el estimador en un modelo GAM corrigiendo el efecto de la censura con los pesos Kaplan-Meier (*gamkm* modificada, función para calcular el estimador ckmGAM), además de una función necesaria para ambas que permite calcular vectores de nodos no uniformes con un espaciado entre los mismos elegido en función de los pesos de Kaplan-Meier (*nodos.km*, ubicación de los nodos L_{km}, sección 2.2.2).

El código ha sido escrito usando R-3.4.4 (plataforma: x86_64-pc-linux-gnu, 64-bit) con versiones de los paquetes mgcv_1.8-23, nlme_3.1-131.1 y survival_2.41-3.

Función de R 4.5: *nodos.km*: función para calcular la ubicación de los nodos usando los pesos Kaplan-Meier

```
## ------------------------------------------------------
## Nombre de la función: nodos.km
## ------------------------------------------------------
## Descripción: Calcula un vector de nodos no uniformes
##              con un espaciado entre los mismos elegido
##              en función de los pesos Kaplan-Meier
## ------------------------------------------------------
## Librerías requeridas: survival
## ------------------------------------------------------
## Uso: nodos.km(X, km, ndx)
##
##      X:       regresor
##      km:      pesos Kaplan-Meier
##      ndx+1:   numero de nodos
## ------------------------------------------------------
## Valor: vector de nodos
## ------------------------------------------------------

nodos.km <- function(X,km,ndx) {
      xl <- min(X)
      xr <- max(X)
      dx <- (xr-xl)/ndx
      unix <- min(km[km!=0])/1000
      Xord <- sort(X)
      kmord <- km[order(X)]
      kmord[length(kmord)] <- kmord[length(kmord)]+unix
      Xsc  <- Xord[kmord!=0]
      kmsc <- kmord[kmord!=0]
      kma <- cumsum(kmsc)
      c(seq(xl-3*dx, xl, length=4),
        Xsc[cumsum(table(factor(findInterval(kma,
        seq(min(kma), (max(kma)+unix),length=ndx+1)),
        levels=1:ndx)))], seq(xr+1*dx, xr+3*dx, length=3))
} ## end of function nodos.km
```

Función de R 4.6: *pswc* modificada: función para el cálculo del estimador ckmPS

```
## ------------------------------------------------------
## Nombre de la función: pswc
## ------------------------------------------------------
## Descripción: extensión del método P-splines
##              de Eilers & Marx 1996 al caso de datos
##              censurados utilizando pesos Kaplan-Meier
## ------------------------------------------------------
## Librerías requeridas: survival, splines
## ------------------------------------------------------
## Uso: pswc(x, y, cen, grado.b, ndx, knots=F, lambda,
##           gamma, pexp)
##
##      x:     regresor
##      y:     tiempo de seguimiento
##      cen:   indicador de estado,
##             usualmente 0/FALSE=vivo, 1/TRUE=fallecido
##      grado.b:  grado del polinomio a trozos
##                (3= splines cúbicos)
##      ndx+1:    numero de nodos
##      knots:    knots=T, nodos no equidistantes creados
##                usando los pesos Kaplan-Meier
##                knots=F, nodos equidistantes
##      lambda:   parámetro de suavizado
##      gamma:    multiplica los grados de libertad del
##                modelo en el criterio GCVc
##      pexp:     exponente de los pesos Kaplan-Meier en
##                el criterio GCVc
## ------------------------------------------------------
## Valor: estimación noparamétrica de la relación entre
##        la variable respuesta (y) y el regresor (x)
##        y valor del parámetro óptimo de suavizado
## ------------------------------------------------------

pswc <- function(x, y, cen, grado.b, ndx, knots=F, lambda,
                 gamma, pexp) {
    require(survival)
    require(splines)
    # Calcula pesos Kaplan-Meier
    Wkm <- kmw.cp(y, cen)

    ## Ordenado por X
    x.ox <- x[order(x)]
    y.ox <- y[order(x)]
    cen.ox <- cen[order(x)]
    ind.ox <- c(1:length(x))[order(x)]

    ## Ordenado por Y y por censura
```

```
y.oy <- y[order(y, !cen)]
x.oy <- x[order(y, !cen)]
W.oy <- Wkm[order(y, !cen)]
ind.oy <- c(1:length(y))[order(y, !cen)]

## Base B-spline y Nodos
if (missing(grado.b)) {
    grado.b <- 3
}
ifelse(knots==TRUE,
{
knots <- nodos.km(x,Wkm,ndx)
},
{
unix <- min(abs(x[x != 0]))/1000
xl <- min(x) - unix
xr <- max(x) + unix
dx <- (xr - xl)/ndx
knots <- seq(xl - grado.b * dx, xr + grado.b * dx,
             by = dx)
}
)

BBB.ox  <- spline.des(knots, x.ox,
                      grado.b + 1, 0 * x.ox)$design
# Orden inicial.
# Si las X's están ordfenadas, no hace nada
BBB.ini <- BBB.ox[order(ind.ox),]
BBB.oy  <- BBB.ini[order(y, !cen), ]

## Matriz representación del operador de diferencias
## (diferencias de orden 2)
D = diff(diff(diag(ncol(BBB.oy))))
DtD = t(D) %*% D

## Matriz con los pesos Kaplan-Meier
matW.oy <- diag(W.oy)
## Elección del nivel de suavizado óptimo via GCVc
if (missing(lambda)) {
    if (missing(gamma)) {
    gamma <- 1
    }
    if (missing(pexp)) {
    pexp <- 1
    }
    pswcGCVc <- function(lambda, BBB.oy, matW.oy, DtD,
                         y.oy, W.oy) {
```

```
        pena <- solve(t(BBB.oy) %*% matW.oy %*%
                      BBB.oy + (lambda * DtD)) %*%
                      t(BBB.oy) %*% matW.oy %*% y.oy

        s <- sum(W.oy^pexp * (y.oy - BBB.oy %*% pena)^2)

        hasterisco <- solve(t(BBB.oy) %*% matW.oy %*%
                      BBB.oy + (lambda * DtD)) %*%
                      t(BBB.oy) %*% matW.oy %*% BBB.oy

        trazah <- sum(diag(hasterisco))
        gcv <- s/(nrow(BBB.oy) - gamma * trazah)^2
        return(gcv)
      }

      # Valor óptimo del parámetro de suavizado lambda
      lambda <- optimize(pswcGCVc, c(0, 10000),
                         tol = 1e-06, BBB.oy = BBB.oy,
                         matW.oy = matW.oy, DtD = DtD,
                         y.oy = y.oy, W.oy = W.oy)[[1]]
  }

  ## Estimación
  pena <- solve(t(BBB.oy) %*% matW.oy %*%
                BBB.oy + (lambda * DtD)) %*%
                t(BBB.oy) %*% matW.oy %*% y.oy
  est.cnpl <- BBB.ini %*% pena
  return(list(estimacion= est.cnpl, landa= lambda))
} ## fin de la función pswc
```

Función de R 4.7: *gamkm* modificada: función para el cálculo del estimador ckmGAM con un regresor

```
## ------------------------------------------------------
## Nombre de la función: gamkm
## ------------------------------------------------------
## Descripción: función gam con pesos Kaplan-Meier y
##              selección del parámetro de suavizado
##              via GCVc: un regresor
## ------------------------------------------------------
## Librerías requeridas: mgcv
## ------------------------------------------------------
## Uso: gamkm(y, x, cen, kn, knots, kmw, base, gamma, pexp)
##
##      y:      tiempo de seguimiento
##      x:      regresor
##      cen:  indicador de estado,
##            usualmente 0/FALSE=vivo, 1/TRUE=fallecido
##      kn:     dimensión de la base que representa el
##              término de suavizado (kn=ndx+3)
##      knots:    knots=T, nodos no equidistantes creados
##                usando los pesos Kaplan-Meier, que debe
##                coincidir con el valor kn suministrado
##                knots=F, nodos equidistantes
##      kmw:    pesos Kaplan-Meier
##      base:   cadena de dos letras que indica la base
##              de suavizado (penalizada) a utilizar
##      gamma:    multiplica los grados de libertad del
##                modelo en el criterio GCVc
##      pexp:     exponente de los pesos Kaplan-Meier en
##                el criterio GCVc
## ------------------------------------------------------
## Valor: objecto de la clase "gam"
## ------------------------------------------------------

gamkm <- function(y, x, cen, kn, knots=F, kmw, base,
                  gamma, pexp) {
    ndx <- kn-3
    require(mgcv)
    if (missing(kmw)) {
    require(survival)
    kmw <- kmw.cp(y, cen)
    }
    if (missing(gamma)) {
        gamma <- 1
    }
    if (missing(pexp)) {
        pexp <- 1
    }
```

```
    ifelse(knots==TRUE,
        {
        knots <- nodos.km(x,kmw,ndx)
        },
        {
        unix <- min(abs(x[x != 0]))/1000
        xl <- min(x) - unix
        xr <- max(x) + unix
        dx <- (xr - xl)/ndx
        knots <- seq(xl - 3 * dx, xr + 3 * dx, by = dx)
    })
    gamkmGCVc <- function(lambda, y, x, kn, kmw, base,
                          knots) {
        nobs <- length(y)
        mod.gam <- gam(y ~ s(x, bs = base, k = kn,
                       m = c(2, 2)), sp = lambda,
                       weights = kmw, knots=list(x=knots))
        rss <- sum(kmw^pexp * (y - fitted(mod.gam))^2)
        trF <- sum(mod.gam$hat)
        # GCVc
        gcv <- nobs * rss/(nobs - gamma * trF)^2
        return(gcv)
    }

    # Selección del parámetro de suavizado
    lambdaopt <- optimize(gamkmGCVc, c(0, 10000),
                          tol = 1e-06, y = y, x = x,
                          kn = kn, kmw = kmw, base = base,
                          knots=list(x=knots))[[1]]
    return(gam(y ~ s(x, bs = base, k = kn, m = c(2, 2)),
                    sp = lambdaopt, weights = kmw,
                    knots=list(x=knots), gamma = 1))
} ## fin de la función gamkm
```

4.3. Código para el modelo semiparamétrico

En este apartado se muestra el código de R de la extensión de la metodología de los P-splines al contexto de muestras con observaciones censuradas en un modelo semiparamétrico (*semipswc*, sección 2.3).

El código ha sido escrito usando R-3.6.3 (plataforma: x86_64-pc-linux-gnu, 64-bit) con versiones de los paquetes mgcv_1.8-23, nlme_3.1-131.1 y survival_2.41-3.

Función de R 4.8: *semipswc*: función para el cálculo del estimador P-splines censurado en un modelo semiparamétrico

```
## ------------------------------------------------------
## Nombre de la función: semipswc
## ------------------------------------------------------
## Descripción: extensión del método P-splines
##              de Eilers & Marx 1996 al caso de datos
##              censurados utilizando pesos Kaplan-Meier
##              en un modelo semiparamétrico
## ------------------------------------------------------
## Librerías requeridas: survival, splines, MASS
## ------------------------------------------------------
## Uso: semipswc(Z, x, y, cen, grado.b, ndx, knots=F,
##               lambda, gamma, pexp, toler=1e-05)
##
##      Z:    matriz de regresores parte paramétrica
##      x:    regresor componente no paramétrica
##      y:    tiempo de seguimiento
##      cen:  indicador de estado,
##            usualmente 0/FALSE=vivo, 1/TRUE=fallecido
##      grado.b:  grado del polinomio a trozos
##                (3= splines cúbicos)
##      ndx+1:    numero de nodos
##      knots:    knots=T, nodos no equidistantes creados
##                usando los pesos Kaplan-Meier
##                knots=F, nodos equidistantes
##      lambda:   parámetro de suavizado
##      gamma:    multiplica los grados de libertad del
##                modelo en el criterio GCVc
##      pexp:     exponente de los pesos Kaplan-Meier en
##                el criterio GCVc
##      toler:    precisión deseada
## ------------------------------------------------------
## Valor:
##       estnp:    estimación componente no paramétrica
##       alfas:    estimación parámetros componente
##                 paramétrica
##       sdnp:     estimación desviación típica componente
##                 no  paramétrica
##       sdalfas:  estimación desviación típica componente
##                 paramétrica
##       landa:    parámetro de suavizado óptimo
##       sd:       desviación típica estimada del error
##       iter:     número de iteraciones
##       gvc:      valor del criterio GVCc
##       rdf:      grados de libertad efectivos
## ------------------------------------------------------
```

```
semipswc <- function(Z, x, y, cen, grado.b, ndx, knots=F,
                     lambda, gamma, pexp, toler=1e-05) {

    require(survival)
    require(splines)
    require(MASS)
    # Calcula pesos Kaplan-Meier
    Wkm <- kmw.cp(y, cen)

    ## Ordenado por X
    x.ox <- x[order(x)]
    y.ox <- y[order(x)]
    cen.ox <- cen[order(x)]
    ind.ox <- c(1:length(x))[order(x)]

    ## Ordenado por Y y por censura
    y.oy <- y[order(y, !cen)]
    x.oy <- x[order(y, !cen)]
    W.oy <- Wkm[order(y, !cen)]
    # Regresores parte Paramétrica
    Z.oy <- Z[order(y, !cen),]
    matZ.oy <- cbind(rep(1,length(y)),Z.oy)
    ind.oy <- c(1:length(y))[order(y, !cen)]

    ## Base B-spline y Nodos
    if (missing(grado.b)) {
        grado.b <- 3
    }
    ifelse(knots==TRUE,
    {
    knots <- nodos.km(x,Wkm,ndx)
    },
    {
        unix <- min(abs(x[x != 0]))/1000
    xl <- min(x) - unix
    xr <- max(x) + unix
    dx <- (xr - xl)/ndx
    knots <- seq(xl - grado.b * dx, xr + grado.b * dx,
                 by = dx)
    }
    )

    BBB.ox  <- spline.des(knots, x.ox,
                          grado.b + 1, 0 * x.ox)$design
    # Orden inicial.
    # Si las X's están ordfenadas, no hace nada
    BBB.ini <- BBB.ox[order(ind.ox),]
    BBB.oy  <- BBB.ini[order(y, !cen), ]
```

```
## Matriz representación del operador de diferencias
## (diferencias de orden 2)
D = diff(diff(diag(ncol(BBB.oy))))
DtD = t(D) %*% D

## Matriz con los pesos Kaplan-Meier
matW.oy <- diag(W.oy)


## Matriz de proyección: parte paramétrica
Hp.oy <- matZ.oy[,-1] %*% solve(t(matZ.oy[,-1]) %*%
         matW.oy %*% matZ.oy[,-1]) %*%
         t(matZ.oy[,-1]) %*% matW.oy

## Matriz identidad
id.mat <- diag(rep(1,nrow(BBB.oy)))


## Función para una iteración: semipswc.iter
## Paso 1
## Estimación parte no paramétrica tras restar la
## estimación parte paramétrica
## Parámetros iniciales iguales a cero (alfas=0)

semipswc.iter <- function(matZ.oy, BBB.oy, BBB.ini,
                          matW.oy, DtD, W.oy,
                          alphaold, y.oy, lambda,
                          gamma, pexp) {

   # Inicio con alfas=0
   ymenosalphaold.oy <- y.oy-(matZ.oy %*% alphaold)

   # Elección del nivel de suavizado óptimo via GCVc
    if (missing(lambda)) {
        if (missing(gamma)) {
        gamma <- 1
        }
        if (missing(pexp)) {
        pexp <- 1
        }
        pswcGCVc <- function(lambda, BBB.oy, matW.oy,
                             DtD, y.oy, W.oy) {

          pena <- solve(t(BBB.oy) %*% matW.oy %*%
                        BBB.oy + (lambda * DtD)) %*%
                        t(BBB.oy) %*% matW.oy %*% y.oy
```

```
            s <- sum(W.oy^pexp * (y.oy - BBB.oy %*%
                                  pena)^2)

            hasterisco <- solve(t(BBB.oy) %*%
                          matW.oy %*% BBB.oy +
                          (lambda * DtD)) %*%
                          t(BBB.oy) %*%
                          matW.oy %*% BBB.oy

            trazah <- sum(diag(hasterisco))
            gcv <- s/(nrow(BBB.oy) - gamma * trazah)^2
            return(gcv)
          }

    # Valor óptimo del parámetro de suavizado lambda
    lambda <- optimize(pswcGCVc, c(0, 10000),
                       tol = 1e-06, BBB.oy = BBB.oy,
                       matW.oy = matW.oy, DtD = DtD,
                       y.oy = ymenosalphaold.oy,
                       W.oy = W.oy)[[1]]
    }

    # Estimación
    pena <- solve(t(BBB.oy) %*% matW.oy %*% BBB.oy +
                  (lambda * DtD)) %*% t(BBB.oy) %*%
                  matW.oy %*% ymenosalphaold.oy
     est.cnpl <- BBB.ini %*% pena

    # Paso 2: estimar los alfas
    alphanew<- solve(t(matZ.oy) %*% matW.oy %*%
                     matZ.oy) %*% t(matZ.oy) %*%
                     matW.oy %*%
                     (y.oy - (BBB.oy %*% pena))

    # GCV
    s <- sum(W.oy^pexp * (y.oy - BBB.oy %*% pena)^2)
    hasterisco <- solve(t(BBB.oy) %*% matW.oy %*%
                        BBB.oy + (lambda * DtD)) %*%
                        t(BBB.oy) %*% matW.oy %*%
                        BBB.oy
    trazah <- sum(diag(hasterisco))
    gcv <- s/(nrow(BBB.oy) - gamma * trazah)^2

    return(list(solu = est.cnpl, landa = lambda,
                solupar = alphanew, gcv = gcv))
    }
## Fin función para una iteración: semipswc.iter
```

```
 ## Iteraciones
 ## Criterio de parada: diferencia entre GCVs

 i <- 0
 not.converged <- TRUE
 # Para evitar convergencia immediatamente
 old.gcv <- -100

 # Incorpora término independiente para mejorar
 # la velocidad de convergencia (Marra & Wood 2012)
 alphaold <- c(rep(0,ncol(matZ.oy)))
 while (not.converged & i<=100) {
        if(i == 100){warning("No converge en 100 it.")}
        i <- i+1
        print(i)
        mod.iter <- semipswc.iter(matZ.oy=matZ.oy,
                    BBB.oy=BBB.oy, BBB.ini=BBB.ini,
                    matW.oy=matW.oy, DtD=DtD,
                    W.oy=W.oy, alphaold=alphaold,
                    y.oy=y.oy, lambda=lambda,
                    gamma=gamma, pexp=pexp)

       alphaold <- mod.iter[[3]]
       new.gcv  <- mod.iter[[4]]
       if (abs(new.gcv-old.gcv)<toler*new.gcv) {
           not.converged <- FALSE
           }
       old.gcv <- new.gcv
}

# Resultado última iteración
mod.fin <- semipswc.iter(matZ.oy=matZ.oy[,-1],
                         BBB.oy=BBB.oy, BBB.ini=BBB.ini,
                         matW.oy=matW.oy, DtD=DtD,
                         W.oy=W.oy, alphaold=
                         alphaold[2:ncol(matZ.oy)],
                         y.oy=y.oy, lambda=lambda,
                         gamma=gamma, pexp=pexp)

## Parámetro de suavizado óptimo
lambda <- mod.fin[[2]]

## Estimaciones parte paramétrica
alpha.fin <- matrix(data=mod.fin[[3]],ncol=1)

## Residuos
res <- (y - Z%*%alpha.fin - mod.fin[[1]])
```

```
    ## Hc: matriz suavizado
    hasterisco <- solve(t(BBB.oy) %*% matW.oy %*%
                        BBB.oy + (lambda * DtD)) %*%
                        t(BBB.oy) %*% matW.oy %*% BBB.oy
    trazah <- sum(diag(hasterisco))

    hc.oy   <- BBB.oy %*% solve(t(BBB.oy) %*% matW.oy %*%
    BBB.oy + (lambda * DtD)) %*% t(BBB.oy) %*% matW.oy

    ## Estimación sigma
    s.v1 <- sum((nrow(BBB.oy)*Wkm) * res^2)
    Sest <- sqrt(s.v1/(nrow(BBB.oy) - trazah-2))
    ## Fin Estimación sigma

    ## Estimación Desviación Parte No Paramétrica
    vargammainv <- solve(t(BBB.oy) %*% matW.oy %*%
                         (id.mat-Hp.oy) %*%
                         BBB.oy + (lambda * DtD))
    vargamma.p1 <- vargammainv %*% t(BBB.oy) %*%
                   matW.oy %*% (id.mat-Hp.oy)
    vargamma.p2 <- t(vargamma.p1)
    vargamma <- vargamma.p1%*%vargamma.p2
    varnp.r.oy <- BBB.oy %*% vargamma %*% t(BBB.oy)
    sdnpv2.r.oy <- Sest*sqrt(diag(varnp.r.oy))
    sdnpv2.r <- sdnpv2.r.oy[order(ind.oy)]

    ## Estimación Desviación Parte Paramétrica
    varalfasinv <- ginv(t(matZ.oy) %*% matW.oy %*%
                        (id.mat-hc.oy) %*% matZ.oy)
    varalfas.2.p1 <- varalfasinv%*%t(matZ.oy) %*%
                     matW.oy %*% (id.mat-hc.oy)
    varalfas.2.p2 <- t(id.mat-hc.oy) %*% matW.oy %*%
                     matZ.oy %*% t(varalfasinv)
    varalfas.2 <- varalfas.2.p1%*%varalfas.2.p2

    sdalfas2 <- Sest*sqrt(diag(
             varalfas.2[2:ncol(matZ.oy),2:ncol(matZ.oy)]))

    return(list(estnp = mod.fin[[1]], alfas = alpha.fin,
                sdnp = sdnpv2.r, sdalfas = sdalfas2,
                landa = mod.fin[[2]], sd = Sest, iter = i,
                gcv = mod.fin[[4]],
                rdf = (nrow(BBB.oy) - trazah-2)
                )
          )

} ## fin de la función semipswc
```

Capítulo 5

Bibliografía

Aydin, D. and E. Yilmaz (2018). Modified estimators in semiparametric regression models with right-censored data. *Journal of Statistical Computation and Simulation 88, 1470–1498.*

Buckley, J. J. and I. R. James (1979). Linear regression with censored data. Biometrika 66, 429–436.

Cai, T., R. J. Hyndman, and M. Wand (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics 11, 784–798.*

Chen, W., X. Li, D. Wang, and G. Shi (2015). Parameter estimation of partial linear model under monotonicity constraints with censored data. Journal of the Korean Statistical Society 44, 410–418.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34, 187–202.*

Currie, I. D., M. Durbán, and P. H. Eilers (2004). Smoothing and forecasting mortality rates. Statistical Modelling 4, 279–298.

De Boor, C. (2001). *A Practical Guide to Splines, revised version*, Volume 27 of *Applied Mathematical Sciences*. New York: Springer-Verlag.

De Uña Álvarez, J. and J. Roca Pardiñas (2009). Additive models in censored regression. *Computational Statistics and Data Analysis 53, 3490–3501.*

Dickson, E. R., P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy (1989). Prognosis in primary biliary cirrhosis: Model for decision making. Hepatology 10, 1–7.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Numerical Mathematics and Scientific Computation. Oxford: Oxford University Press.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In W. Schemp and K. Zeller (Eds.), *Constructive theory of functions of several variables*, pp. 85–100. Springer.

Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science 11, 89–121.*

Eilers, P. H., B. D. Marx, and M. Durbán (2015). Twenty years of p-splines. SORT-Statistics and Operations Research Transactions 39, 149–186.

Escobar, L. A. and W. Q. Meeker (1992). Assessing influence in regression analysis with censored data. *Biometrics 48, 507–528.*

Eubank, R. L. (1988). Spline Smoothing and Nonparametric Regression. *New York: Marcel Dekker.*

Fleming, T. R. and D. P. Harrington (2005). Counting Processes and Survival Analysis. *Hoboken: New Jersey: John Wiley & Sons.*

Friedman, J. H. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling (with discussion). Technometrics 31, 3–39.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models*, Volume 58 of *Monographs on Statistics and Applied Probability.* London: Chapman and Hall.

Härdle, W. (1990). *Applied Nonparametric Regression*, Volume 19 of *Econometric Society Monographs.* Cambridge: Cambridge University Press.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society: Series B (Methodological) 48, 244–248.*

Hennerfeind, A., A. Brezger, and L. Fahrmeir (2006). Geoadditive survival models. Journal of the American Statistical Association 101, 1065–1075.

Holland, A. D. (2017). Penalized spline estimation in the partially linear model. *Journal of Multivariate Analysis 153, 211–235.*

Jin, Z., D. Y. Lin, L. J. Wei, and Z. Ying (2003). Rank-based inference for the accelerated failure time model. Biometrika 90, 341–353.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data* (Second ed.). Wiley Series in Probability and Statistics. Hoboken, New York: John Wiley & Sons, Inc.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53, 457–481.*

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. Computational Statistics and Data Analysis 49, 169–186.

Kauermann, G. and P. Khomski (2006). Additive two-way hazards model with varying coefficients. *Computational Statistics and Data Analysis 51, 1944–1956.*

Kim, Y. J. and C. Gu (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66, 337–356.

Kneib, T. (2006). Mixed model-based inference in geoadditive hazard regression for interval-censored survival times. *Computational Statistics and Data Analysis 51, 777–792.*

Kneib, T. and L. Fahrmeir (2007). A mixed model approach for geoadditive hazard regression. Scandinavian Journal of Statistics 34, 207–228.

Komárek, A., E. Lesaffre, and J. F. Hilton (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics 14, 726–745.*

Konrath, S., L. Fahrmeir, and T. Kneib (2015). Bayesian accelerated failure time models based on penalized mixtures of Gaussians: regularization and variable selection. AStA Advances in Statistical Analysis 99, 259–280.

Koul, H., V. Susarla, and J. Van-Ryzin (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics 9, 1276 – 1288.*

Lai, T. L. and Z. Ying (1992). Linear rank statistics in regression analysis with censored or truncated data. Journal of Multivariate Analysis 40, 13–45.

Lambert, P. (2013). Nonparametric additive location-scale models for interval censored data. *Statistics and Computing 23, 75–90.*

Leurgans, S. (1987). Linear models, random censoring and synthetic data. Biometrika 74, 301–309.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika 63, 449–464.*

Miller, R. G. and J. Halpern (1982). Regression with censored data. Biometrika 69, 521–531.

Orbe, J., E. Ferreira, and V. Núñez Antón (2003). Censored partial regression. *Biostatistics 4, 109–121.*

Orbe, J. and J. Virto (2018). Penalized spline smoothing using Kaplan-Meier weights with censored data. Biometrical Journal 60, 947–961.

Orbe, J. and J. Virto (2021). Selecting the smoothing parameter and knots for an extension of penalized splines to censored data. *Journal of Statistical Computation and Simulation 91, 2953–2985.*

Orbe, J. and J. Virto (2022). Penalized spline smoothing using kaplan-meier weights in semiparametric censored regression models. SORT-Statistics and Operations Research Transactions 46, 95–114.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science 1, 502–527.*

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. SIAM Journal on Scientific and Statistical Computing 9, 363–379.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik 10, 177–183.*

Rice, J. (1986). Convergence rates for partially splined models. Statistics and Probability Letter 4, 203–208.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11, 735–757.*

Ruppert, D., M. Wand, and R. Carroll (2009). Semiparametric regression during 2003–2007. Electronic Journal of Statistics 3, 1193–1256.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics 12. Cambridge University Press.

Schimek, M. G. (2000). Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference 91, 525–540.*

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis*, Volume 26 of* Monographs on Statistics and Applied Probability*. London: Chapman & Hall.*

Speckman, P. (1988). Kernel smoothing in partial linear models. Journal of the Royal Statistical Society: Series B (Methodological) 50, 413–436.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis 45, 89–103.*

Stute, W. (1999). Nonlinear censored regression. Statistica Sinica 9, 1089–1102.

Therneau, T. M. (2015). *A Package for Survival Analysis in R.* R package version 2.38.

Therneau, T. M. (2021). *A Package for Survival Analysis in R.* R package version 3.2-11.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics 18, 354–372.*

Wahba, G. (1990). Spline Models for Observational Data*, Volume 59 of* CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia: Society for Industrial and Applied Mathematics.*

Wood, S. and M. S. Wood (2015). Package 'mgcv'. R package version*, 1–7.*

Wood, S. N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65, 95–114.

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics 62, 1025–1036.*

Wood, S. N. (2017). Generalized Additive Models: An Introduction with R. *Texts in Statistical Science Series. Boca Raton, Florida: CRC press.*

Zhou, M. (1992). Asymptotic normality of the 'synthetic data' regression estimator for censored survival data. The Annals of Statistics 20, 1002–1021.

Zou, Y., J. Zhang, and G. Qin (2011). A semiparametric accelerated failure time partial linear model and its application to breast cancer. *Computational Statistics and Data Analysis 55, 1479–1487.*

Parte II

Conclusiones

Capítulo 6

Conclusiones

6.1. Conclusiones caso univariante

En la primera parte de esta tesis, secciones 2.1 y 3.1, se ha presentado un método sencillo y flexible de modelización no paramétrica en el contexto de datos censurados. En concreto, se ha desarrollado una extensión del enfoque de splines penalizados (Eilers and Marx, 1996) que constituye una novedad en este contexto, utilizando los pesos Kaplan-Meier para tener en cuenta el efecto de la censura. Asimismo, para elegir el valor óptimo del parámetro de suavizado se ha adaptado el criterio de validación cruzada generalizada para el caso de muestras censuradas.

Esta propuesta no necesita asumir una distribución de probabilidad específica para la variable respuesta, que en la práctica es aún más difícil de comprobar con datos censurados, y permite estimar la relación entre la variable de interés y el regresor sin asumir una forma funcional paramétrica específica. Por lo tanto, se evitan problemas de mala especificación del modelo que conducen a estimaciones sesgadas y conclusiones erróneas. Su aplicación en muestras con datos censurados es útil en contextos de análisis de supervivencia o duración en los que las observaciones censuradas son habituales.

Entre otras buenas propiedades (véase Eilers et al., 2015), este enfoque no paramétrico tiene la importante ventaja de que es más apropiado que el enfoque de splines de suavizado, *smoothing splines*, para muestras grandes ya que, a diferencia de los *smoothing splines*, la dimensión de la base no crece con el tamaño de la muestra. El uso de P-splines reduce considerablemente el problema de la dimensionalidad, pasando del número de observaciones en la muestra al número de B-splines. Además, dado que el uso de técnicas de validación cruzada implica estimar las curvas varias veces, es importante señalar que el enfoque de splines penalizados estima las curvas en un tiempo significativamente menor que otros enfoques de suavizado.

Por último, es muy fácil de aplicar y de interpretar porque se modela directamente el efecto de las variables explicativas sobre la supervivencia, por lo que la interpretación de los resultados es más clara y sencilla (en términos de efectos sobre el tiempo medio de supervivencia, como en los modelos estadísticos clásicos) que en los modelos de riesgo, en los que el efecto de la covariable se modela sobre una probabilidad condicional.

Los estudios de simulación realizados ilustran el buen funcionamiento del método propuesto y se puede comprobar que estima satisfactoriamente la verdadera forma funcional de la relación entre covariable y respuesta. Además, como era de esperar,

la precisión de las estimaciones mejora a medida que aumenta el tamaño de muestra y se reduce el porcentaje de censura. La aplicación a un caso real sirve para ilustrar las ventajas potenciales de su uso cuando no se conoce la verdadera forma funcional de la relación.

Además, se ha ampliado la propuesta introduciendo el uso de los pesos Kaplan-Meier para corregir el efecto de la censura en el marco de los modelos aditivos generalizados (GAM) de una forma sencilla, lo que permite estimar inmediatamente modelos más complejos. Los estudios de simulación ilustran el buen rendimiento de la extensión al marco de los modelos aditivos generalizados (GAM).

6.2. Conclusiones selección de parámetros

En las secciones 2.2 y 3.2 se indican algunas pautas a tener en cuenta a la hora de aplicar una metodología de análisis no paramétrico para el caso de datos censurados, por ejemplo, al aplicar la propuesta presentada en la sección 2.1.2. Para la aplicación de esta metodología es necesario, al igual que en el caso no censurado, elegir un parámetro de suavizado y el número y localización de los nodos. Este tema sólo ha sido estudiado previamente en la literatura para datos censurados en Aydin and Yilmaz (2018), pero bajo un enfoque metodológico diferente al de este trabajo (utilizando datos sintéticos). En esta tesis se han propuesto diferentes alternativas para elegir el nivel óptimo de suavizado y la ubicación y el número de nodos en presencia de datos censurados. Mediante un amplio estudio de simulación que considera relaciones funcionales de diversos grados de complejidad entre la variable respuesta censurada y un regresor, se ha analizado el comportamiento de las distintas propuestas en situaciones con diferencias en la información disponible, en las que se combinan tamaños de muestra y niveles de censura variados.

Para elegir el nivel óptimo de suavizado, esencial para obtener un buen ajuste, es habitual en la práctica en el contexto de datos no censurados utilizar el criterio de validación cruzada generalizada o la modificación propuesta por Kim and Gu (2004) para evitar el sobreajuste. En la sección 2.2.1 se muestra que una aplicación directa de estos criterios al caso censurado conduce a estimaciones muy sesgadas. Para corregir este problema se han adaptado los criterios anteriores utilizando los pesos Kaplan-Meier para tener en cuenta el efecto de la censura, tanto en el denominador del GCV, utilizando la matriz de proyección H_c, como en el numerador, que se pondera directamente utilizando los pesos Kaplan Meier (ecuación 2.9). Junto a la anterior propuesta, también se considera un criterio alternativo en el que los pesos Kaplan-Meier aparecen al cuadrado. Del análisis realizado en las simulaciones se concluye que en lo que respecta a la elección del exponente de los pesos Kaplan-Meier a la hora de corregir el GCV, si el nivel de censura es bajo ambos exponentes funcionan de forma similar, pero con censura alta el exponente 1 (w_i) funciona claramente mejor que el exponente 2 (w_i^2). Además, el criterio propuesto (GCV_c) presenta distintas versiones en función de los valores de ϕ. En el estudio de simulación los posibles valores de este parámetro que se utilizan son los que habitualmente se encuentran en la literatura; un valor de 1 (como en el caso del GCV ordinario) o un valor de 1,5. La conclusión obtenida es que un ϕ con valor 1,5 es mejor en casi todas las situaciones, ya que evita el sobreajuste. Este resultado es análogo al encontrado en la literatura para el caso no censurado (Kim and Gu, 2004; Wood, 2017). Es más, la mejora obtenida al utilizar el valor de $\phi = 1{,}5$ es mayor a medida que aumenta el

nivel de censura.

En cuanto a la elección del número y la ubicación de los nodos, por un lado se estudian las propuestas habituales en la literatura para el caso no censurado, es decir nodos igualmente espaciados con un número de nodos calculado aplicando la fórmula presentada en Ruppert (2002). Por otro lado, también se analiza el comportamiento de dos nuevas propuestas adaptadas al caso de datos censurados. Para el número de nodos, una modificación de la fórmula de Ruppert para tener en cuenta el porcentaje de observaciones censuradas (K_c, ecuación 2.11). De los resultados de las simulaciones se concluye que el número de nodos elegido con la expresión K_c siempre funciona mejor que el obtenido utilizando la propuesta de Ruppert. Con un nivel bajo de censura la diferencia es pequeña, pero a medida que aumenta el nivel de censura la nueva propuesta funciona claramente mejor que la de Ruppert. Para la localización de los nodos, se analizan no sólo nodos igualmente espaciados, sino también vectores de nodos no uniformes con nodos espaciados en función de los pesos de Kaplan-Meier (L_{km}). En general, los nodos igualmente espaciados obtienen mejores resultados.

Una de las principales conclusiones que se obtiene del análisis de simulación es que la adaptación del criterio GCV propuesta para el caso censurado, GCV_c, elige parámetros de suavizado que conducen a buenas estimaciones en los distintos escenarios de cada ejemplo analizado. En general, se obtienen mejores resultados cuando el numerador del GCV_c se pondera con los pesos Kaplan-Meier y con un valor ϕ de 1,5. Si se utilizan K_c nodos igualmente espaciados, donde K_c es el número de nodos resultantes de la fórmula (2.11), la combinación de parámetros obtenida produce generalmente los mejores resultados. Cuando el nivel de censura es pequeño no hay grandes diferencias en las estimaciones con distintas combinaciones de parámetros. Cuando el nivel de censura aumenta, estas diferencias son mucho mayores, por lo que la elección de los parámetros adecuados adquiere una importancia aún mayor.

Estos resultados no dependen del método de estimación utilizado en las simulaciones. En general, los dos métodos de estimación que utilizamos, P-splines censurados (ckmPS) y GAMs corregidos (ckmGAM), se comportan de forma similar, con muy buenos resultados para la combinación de parámetros propuesta. Como era de esperar, a mayor tamaño de muestra y menor nivel de censura se obtienen estimaciones más precisas. Por último, los resultados son robustos a las diferencias en la variabilidad o distribución del término de error.

6.3. Conclusiones modelo semiparamétrico

En las secciones 2.3 y 3.3 se ha propuesto y probado un método de estimación en el contexto de los modelos semiparamétricos censurados basado en el enfoque P-spline de Eilers and Marx (1996) utilizando pesos Kaplan-Meier para tener en cuenta el efecto de la censura. La metodología presentada extiende el método de estimación propuesto en la sección 2.1 a un contexto con más de una variable explicativa, muy útil desde un punto de vista práctico. Además, se han desarrollado las herramientas necesarias para realizar inferencias estadísticas en este marco general, proporcionando, por ejemplo, intervalos de confianza tanto para el componente no paramétrico como para los coeficientes asociados a los regresores del componente paramétrico. Los estudios de simulación realizados ilustran el buen funcionamiento del método de estimación, que estima satisfactoriamente tanto la componente no paramétrica como

los coeficientes asociados a la parte paramétrica en los distintos ejemplos estudiados. Como en el caso univariante, la precisión de las estimaciones mejora a medida que se reduce el nivel censurado y aumenta el tamaño de la muestra. Se han calculado en varios estudios de simulación las probabilidades de cobertura de los intervalos de confianza propuestos y se ha comprobado que la probabilidad de cobertura real se aproxima bastante a la probabilidad de cobertura nominal en todos los escenarios analizados.

Una aplicación a datos reales sirve para ilustrar las ventajas potenciales de la propuesta, que es comparable al método paramétrico AFT y al enfoque de Stute cuando la forma funcional elegida es correcta. En caso contrario, hay que mencionar que si la forma funcional o la distribución de probabilidad se eligen de forma errónea, se produciría un grave problema de especificación incorrecta del modelo en el AFT o el enfoque de Stute y, por tanto, de conclusiones incorrectas. Sin embargo, el método propuesto sería más flexible y robusto, ya que no necesita imponer una distribución de probabilidad específica para la variable de respuesta, ni asumir una forma funcional para la relación entre la variable de respuesta censurada y la covariable, que suelen ser desconocidas en la práctica. Por tanto, su aplicación en muestras con datos censurados es especialmente útil en contextos de análisis de supervivencia o duración en los que las observaciones censuradas son habituales.

Parte III

Trabajos Publicados

A continuación se incluye la versión íntegra de las tres publicaciones que, hasta el momento, ha dado lugar esta tesis:

1. En el capítulo 7 se incluye Orbe and Virto (2018),
   DOI: https://doi.org/10.1002/bimj.201700213

   Orbe, J. and J. Virto (2018). Penalized spline smoothing using Kaplan-Meier weights with censored data. *Biometrical Journal 60*, 947–961.

   La revista *Biometrical Journal* está recogida en los listados del Journal Citation Reports (JCR) del Web of Science (WoS), Science Edition. Atendiendo al índice Journal Impact Factor del JCR (JCR-JIF) tiene un valor JCR-JIF=1,255 en el año de publicación (2018), ocupando la posición 51/123, segundo cuartil (Q2) en la categoría Statistics and Probability. Atendiendo al índice Journal Citation Indicator del JCR (JCR-JCI) tiene un valor JCR-JCI=0,65 en el año de publicación (2018), ocupando la posición 54/161, segundo cuartil (Q2), primer tercil (T1) en la categoría Statistics and Probability.

2. En el capítulo 8 se incluye Orbe and Virto (2021),
   DOI: https://doi.org/10.1080/00949655.2021.1913737

   Orbe, J. and J. Virto (2021). Selecting the smoothing parameter and knots for an extension of penalized splines to censored data. *Journal of Statistical Computation and Simulation 91*, 2953–2985.

   La revista *Journal of Statistical Computation and Simulation* está recogida en los listados del Journal Citation Reports (JCR) del Web of Science (WoS), Science Edition. Atendiendo al índice Journal Impact Factor del JCR (JCR-JIF) tiene un valor JCR-JIF=1,225 en el año de publicación (2021), ocupando la posición 84/125, cuartil Q3 en la categoría Statistics and Probability. Atendiendo al índice Journal Citation Indicator del JCR (JCR-JCI) tiene un valor JCR-JCI=0,41 en el año de publicación (2021), ocupando la posición 105/163, cuartil Q3 en la categoría Statistics and Probability.

3. En el capítulo 9 se incluye Orbe and Virto (2022),
   DOI: https://doi.org/10.2436/20.8080.02.119

   Orbe, J. and J. Virto (2022). Penalized spline smoothing using Kaplan-Meier weights in semiparametric censored regression models. *SORT-Statistics and Operations Research Transactions 46*, 95–114.

   Como se trata de una aportación publicada en el año 2022, presento los valores de los últimos índices publicados, los del año 2021. La revista *SORT-Statistics and Operations Research Transactions* está recogida en los listados del Journal Citation Reports (JCR) del Web of Science (WoS), Science Edition. Atendiendo al índice Journal Impact Factor del JCR (JCR-JIF) tiene un valor JCR-JIF=1,759 en el último año publicado (2021), ocupando la posición 54/125, cuartil Q2 en la categoría Statistics and Probability. Atendiendo al índice Journal Citation Indicator del JCR (JCR-JCI) tiene un valor JCR-JCI=0,51 en el último año publicado (2021), ocupando la posición 81/163, cuartil Q2 en la categoría Statistics and Probability.

Capítulo 7

Penalized spline smoothing using Kaplan–Meier weights with censored data

Received: 27 October 2017 | Revised: 29 May 2018 | Accepted: 30 May 2018

DOI: 10.1002/bimj.201700213

RESEARCH PAPER

Biometrical Journal

Penalized spline smoothing using Kaplan–Meier weights with censored data

Jesus Orbe | Jorge Virto

Department of Econometrics and Statistics, University of the Basque Country UPV/EHU, Bilbao, Spain

Correspondence
Jesus Orbe, Department of Econometrics and Statistics, University of the Basque Country UPV/EHU, Bilbao, Spain.
Email: jesus.orbe@ehu.eus

Funding information
Eusko Jaurlaritza, Grant/Award Number: IT-642-13; Euskal Herriko Unibertsitatea, Grant/Award Number: UFI 11/03; Ministerio de Ciencia e Innovación, Grant/Award Numbers: ECO2013-40935-P, ECO2016-76884-P

Abstract

In this paper, we consider the problem of nonparametric curve fitting in the specific context of censored data. We propose an extension of the penalized splines approach using Kaplan–Meier weights to take into account the effect of censorship and generalized cross-validation techniques to choose the smoothing parameter adapted to the case of censored samples. Using various simulation studies, we analyze the effectiveness of the censored penalized splines method proposed and show that the performance is quite satisfactory. We have extended this proposal to a generalized additive models (GAM) framework introducing a correction of the censorship effect, thus enabling more complex models to be estimated immediately. A real dataset from Stanford Heart Transplant data is also used to illustrate the methodology proposed, which is shown to be a good alternative when the probability distribution for the response variable and the functional form are not known in censored regression models.

KEYWORDS

censored data, Kaplan–Meier weights, nonparametric estimation, penalized splines, survival analysis

1 | INTRODUCTION

In this paper, we consider the problem of nonparametric curve fitting in the specific context of censored data, that is when the sample is not observed completely because some of the data values are censored. Thus, for some individuals instead of the actual value of the variable of interest what is observed is a minimum value, and it is known that the actual value is larger than this minimum value. This is known as a right censored data sample. This situation is very common in survival and duration analysis when the relationship between the logarithm of survival time, denoted here as T, the variable of interest, and a relevant covariate X is studied

$$t_i = f(x_i) + \epsilon_i \qquad i = 1, 2, \ldots, n$$

where ϵ_i is the error term and n is the sample size.

Frequently, the relationship between T and X is not known, and instead of assuming a particular parametric relationship $f(\cdot)$, such as for example the usual linear regression specification, it is assumed only that $f(\cdot)$ is a smooth function of data. Thus, we consider a nonparametric curve fitting approach avoiding an incorrect specification of functional form that would lead to a biased estimate and to the wrong conclusions.

The problem of nonparametric curve fitting when the data available is complete, that is there is no censored data, has been extensively analyzed and there are numerous studies in this area. Many methods using different approaches have been proposed. There are methods based on kernel smoothers (Härdle, 1990; Silverman, 1986) that obtain the estimate in each value x_i as a function, usually a weighted average of local observations values t_i. Another approach is that of methods based on spline

smoothers (Eubank, 1988; Green & Silverman, 1994; Wahba, 1990; Wood, 2017). Splines are polynomial function pieces fitted together at points known as knots, where certain conditions or constraints concerning the continuity of the function and some of its derivatives are fixed. Our proposal, for the case of a censored data sample, falls under the splines approach.

Splines depend on the degree of the polynomial and the number and location of the knots. The choice of these elements has been widely studied (e.g. Friedman & Silverman, 1989; Ruppert, 2002). In the literature on spline smoothers various proposals can be found, but two main approaches can be distinguished: smoothing splines and regression splines.

Smoothing splines could be presented as the solution to the introduction of the roughness penalty approach to curve estimation (see Green & Silverman, 1994). The solution of the minimization problem is a natural cubic spline with as many knots as there are different values of the variable X (see Green & Silverman, 1994; Reinsch, 1967). Large data samples require a large number of parameters since there need to be as many knots as there are different values of the variable X. This approach may not therefore be satisfactory because of the size of the data.

One possible solution to this issue is to reduce the dimensionality problem by using a set of q basis functions: $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q$, so that $f(\cdot)$ can be rewritten as the basis expansion: $f(x) = \sum_{j=1}^{q} \gamma_j \boldsymbol{B}_j(x)$ where γ_j is the coefficient associated with the j-th basis function. In the literature on regression splines various alternative ways can be found of calculating these bases. This proposal reduces the dimensionality problem but generates a new one: choosing the number and location of the knots. The idea of penalized spline smoothing to avoid the problem of knot selection dates back to O'Sullivan (1986, 1988), but it was Eilers and Marx (1996) who simplified and generalized it by introducing the combination of B-splines and difference penalties. This proposal is known as the penalized splines (P-splines) approach (for further details see reference Eilers, Marx, & Durbán, 2015).

The bases covered so far are very useful for representing smoothing with one predictor variable. To smooth for more variables, multidimensional smoothers seem more suitable, for example thin-plate splines (Duchon, 1977; Wood, 2003) or tensor product smoothers (Currie, Durban, & Eilers, 2004; De Boor, 2001; Wood, 2006).

Our proposal falls under the P-splines approach of Eilers and Marx (1996) in the specific context of censored data. As already mentioned, this type of data is very common in survival and duration analysis. Several methods have been proposed in this area. Most of them can be classified into two main classes: hazard regression models and accelerated failure time (AFT) regression models.

Under the first of these approaches researchers study the effect of the covariate on a conditional probability (i.e., the hazard function); the most popular approach is Cox's proportional hazard regression model (Cox, 1972). P-spline smoothing approaches in the context of hazard regression models have been suggested before, with early references given by Cai, Hyndman, & Wand (2002) where a mixed model approach for baseline hazard smoothing in a Cox-type model is considered. Kauermann (2005) uses P-splines for fitting nonproportional hazard models and Kauermann and Khomski (2006) extend this model to include a nonparametric calendar time effect. Hennerfeind, Brezger, and Fahrmeir (2006) propose a geoadditive survival model with Bayesian P-spline smoothing. These suggestions are extended by Kneib and Fahrmeir (2007) considering mixed model-based penalized likelihood estimation incorporating nonparametric terms for the baseline hazard rate, time-varying coefficients and nonlinear effects of continuous covariates, a spatial component, and additional cluster-specific frailties. All the aforementioned approaches deal only with right censored survival times. Kneib (2006) extends these models to handle interval censored data. Other contributions on modeling censored data using P-splines are reviewed in Ruppert, Wand, and Carroll (2009).

Under the accelerated failure time regression models approach, the direct effect that the covariate has on the response variable T or some transformation of it is studied, in a way similar to that used in classical regression models (see, e.g. Kalbfleisch & Prentice, 2002). Interestingly, there is little use of P-splines in the context of AFT-type models. Komárek, Lesaffre, and Hilton (2005) propose a methodology that implementes maximum likelihood approach for an AFT using P-splines to smooth the density of the error distribution. Lambert (2013) considers a nonparametric additive model for the location and dispersion of a continuous response using P-splines. More recently, Konrath, Fahrmeir, and Kneib (2015) introduce an extension of AFT regression allowing for joint statistical modeling of regularized linear effects, smooth nonlinear effects, and a flexible error structure. They use a penalized Gaussian mixture for the error density specification that permits imputation of unobserved durations and smoothing as in a complete data situation.

Our proposal is set in the context of AFT models, but it uses a different approach. We combine penalized splines with a weighted least squares estimation method to directly smooth observed data. This proposal consists of an extension of the P-splines method of Eilers and Marx (1996) to handle censored responses using Kaplan–Meier weights following the idea in Stute (1993). The proposed method, in addition to not assuming a functional form for the relation between the censored response variable T and X, does not need to impose a specific probability distribution for the response variable, which is usually unknown in practice. Moreover, the method is very easy to implement and considerably reduces the dimensionality of the problem compared to the smoothing splines approach. Finally, focusing on survival models, note that we directly model the effect of the covariate on the duration variable, which makes the results easier to interpret.

The rest of the paper is organized as follows. The P-spline approach from Eilers and Marx (1996) is described in Section 2. Section 3 shows how to extend the P-splines method when the sample has censored observations and proposes a censored version of penalized splines. In Section 4, the methodology proposed is studied using simulation studies and is shown to perform well. In Section 5, we extend our proposal to generalize additive models (GAM) framework. Section 6 presents an application of the method to a real dataset and the paper concludes with a discussion in Section 7.

2 | PENALIZED SPLINES

We now briefly describe the penalized splines (P-splines) method presented in Eilers and Marx (1996) for the case of noncensored data. Assume a sample of observations (t_i, x_i) for $i = 1, \dots, n$ and consider the nonparametric simple regression

$$T = f(X) + \epsilon$$

where ϵ is the error term that satisfies $E(\epsilon|X) = 0$.

That is, no functional form of $f(\cdot)$ is assumed. A basis approach based on B-splines (De Boor, 2001; Dierckx, 1993) could be a flexible proposal for modeling functions. Thus, using a set of q B-splines basis functions of degree d, $B_1(x), \dots, B_q(x)$, and applying least squares, it is possible to estimate the function $f(\cdot)$, the fitted curve, as $\hat{f}(x) = \sum_{j=1}^{q} \hat{\gamma}_j B_j(x)$ by minimizing this expression:

$$\sum_{i=1}^{n} \left[t_i - \sum_{j=1}^{q} \gamma_j B_j(x_i) \right]^2 \tag{1}$$

where $B_j(x_i)$ denotes the value at x_i of the B_j B-spline for a grid of equidistant knots. Eilers and Marx (1996) provide a brief but interesting review of B-splines, describing their characteristics and general properties and explaining how to compute them.

In order to reduce the problem of choosing the number and the position of the knots, O'Sullivan (1986, 1988) adds a penalty term equivalent to the penalty term for smoothing spline literature (the integrated squared second derivative of the fitted function). Thus, the expression to be minimized is defined as

$$\sum_{i=1}^{n} \left[t_i - \sum_{j=1}^{q} \gamma_j B_j(x_i) \right]^2 + \lambda \int_{x_{min}}^{x_{max}} \left[\sum_{j=1}^{q} \gamma_j B_j''(x_i) \right]^2 dx$$

Eilers and Marx (1996) use a different penalized term and propose the P-spline estimator which, like the previous method, can also be presented as a solution based on the usual B-spline estimation and the smoothing spline approach. The P-spline estimator simplifies and generalizes the proposal of O'Sullivan by introducing a roughness penalty on the difference in adjacent B-splines coefficients γ_j. Thus, an overall level of smoothness of the estimated function is taken into account by imposing that adjacent B-splines coefficients must be similar. Therefore, the corresponding expression of penalized least squares is:

$$\sum_{i=1}^{n} \left[t_i - \sum_{j=1}^{q} \gamma_j B_j(x_i) \right]^2 + \lambda \sum_{j=k+1}^{q} (\Delta^k \gamma_j)^2 \tag{2}$$

where $\Delta\gamma_j$ denotes the difference between coefficients $(\gamma_j - \gamma_{j-1})$ and $\Delta^k\gamma_j$ is the difference of degree k. Thus, the more similar the coefficients are, the less wiggly $f(\cdot)$ is. Eilers and Marx (1996) show that this penalty is a good approximation of the integral of the square of the k-th derivative of the function. In addition, this method considerably reduces the dimensionality of the problem by decreasing it from the number of different values of variable X with smoothing splines to the number of B-splines. In addition, this type of penalty is more flexible since it is independent of the degree of the polynomial used to construct the B splines.

As in other smoothing methods, the optimal amount of smoothing must be chosen. Specifically, in this approach the value of parameter λ must be chosen. These authors suggest using cross-validation or the Akaike information criterion (AIC) to choose the value of λ. Eilers and Marx (1996) give a summary of advantages of their P-splines proposal by comparing it with other methodologies for smoothing.

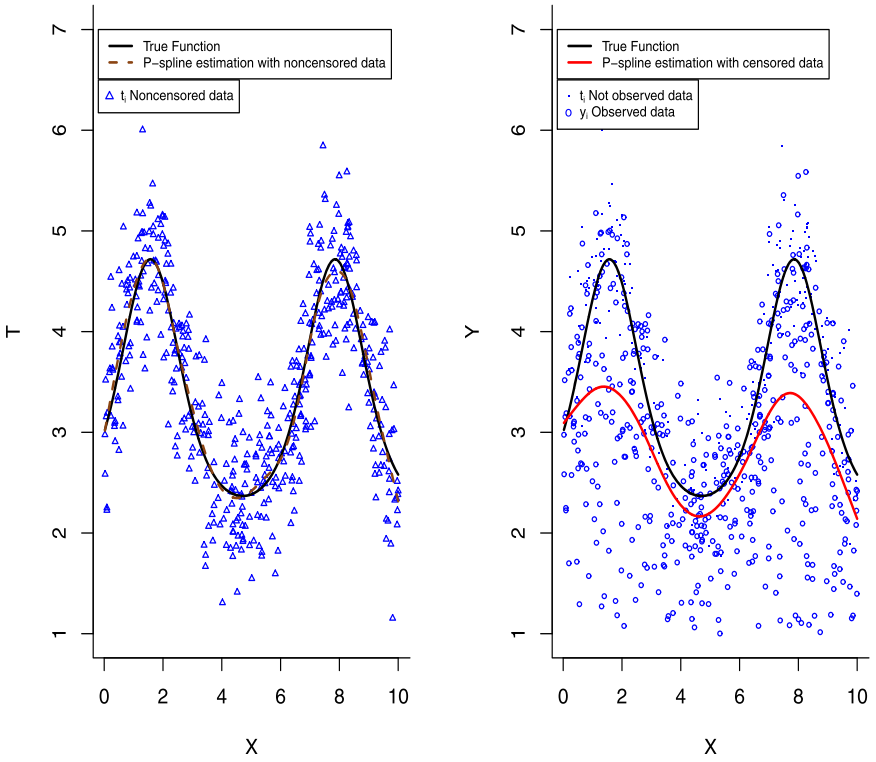


FIGURE 1 P-spline estimation with noncensored and censored data

3 | CENSORED PENALIZED SPLINES

Up to this point we have considered a sample where the interest variable T is completely known. Sometimes this variable T is not completely known because its value would be censored for some individuals. This is a common situation in survival, lifetime, and duration analyses, where T measures the time until the occurrence of an event. These times are known as lifetimes or survival times when the event of interest is the death of an individual, as is usually the case in biomedical areas. In other areas, for example economics or engineering, it is known as the duration or failure time of an individual or element. One of the most important characteristics that generally appear with this kind of data is the existence of censored observations, and the most common censorship pattern is right censoring, that is the fact that for some individuals the event of interest has not yet occurred at the end of the study, so there are right censored observations. We know that their survival time T is greater than a given value, but do not know its real value.

The aim of this study is to extend the P-spline approach of Eilers and Marx (1996) to handle censored responses following the idea of Stute (1993), using Kaplan–Meier weights. To present the proposal, assume that $t_1, \ldots, t_n$ are independent observations from an unknown probability distribution function F of the survival time T. These values may not all be observable due to censoring times $c_1, \ldots, c_n$ for each of the individuals. Moreover, $x_1, \ldots, x_n$ are the values of the covariate X. Therefore, when there is censored data, one observes $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$ a sample of size n where $y_i = min(t_i, c_i)$ is the observed survival time, which is the minimum between the survival time t_i and the censoring value c_i. In addition, it is known which observations are not censored, via the indicator variable $\delta_i = I(t_i \leq c_i)$.

To illustrate the need for a proposal to take into account the effect of censored data we present a simulated sample in Figure 1. This figure shows two panels: on the left a sample with complete information and on the right a sample with censored observations.

The left panel of Figure 1 shows the scatterplot of survival time versus the covariate (t_i, x_i), the true function $f(\cdot)$, and the fitted curve using the P-spline approach by Eilers and Marx (1996), where the noncensored situation is considered, that is all the survival times are known exactly and a sample of n observations (t_i, x_i) is used.

The right panel is constructed by adding the case of the censored sample to the left panel. That is, due to censoring not all the survival times are known exactly and the sample of n observed survival times (y_i, x_i) must be used where $y_i = min(t_i, c_i)$, represented in the panel as circle points. In this example 60% of the observations that coincide with the points in the chart on the left $(y_i = t_i)$ are uncensored. The other 40% are censored observations with values lower than their corresponding points in the left panel $(y_i = c_i)$.

The left panel presents the estimation of the curve $f(\cdot)$ by the P-splines method for the sample without censored data (brown dashed line) along with the true function. As can be seen, the estimation method captures the true functional form. This estimate is only possible for the uncensored case, where complete information is available, that is the sample (t_i, x_i) is fully known.

But if the data is censored the above estimate cannot be obtained. That is, censoring means that not all the survival times are known exactly and the sample of n observed survival times (y_i, x_i) must be used. The right panel shows the estimated curve using P-splines for the case of the sample with censored observations (solid red line). As can be seen, the estimation in the right panel has a major bias and therefore needs to be corrected by taking into account the effect of censoring. That is our main goal in this study.We start by adapting the B-splines approach to censored data situations, that is censored B-splines. Based on Stute's work (Stute, 1993), we use Kaplan–Meier weights to handle censored responses. Thus, the least squares minimizing expression (1) is modified by the following next weighted least squares formula:

$$\sum_{i=1}^{n} w_{[i]} \left[y_{(i)} - \sum_{j=1}^{q} \gamma_j B_j(x_{[i]}) \right]^2 \tag{3}$$

where $y_{(1)}, \ldots, y_{(n)}$ are the ordered values of the observed survival time $y_i = min(t_i, c_i)$, $x_{[i]}$ is the value of the covariate associated to the i-th ordered observation $y_{(i)}$, and $w_{[i]}$ is the Kaplan–Meier weight assigned to $y_{(i)}$. The Kaplan–Meier weights can be calculated as the contribution or jump of the Kaplan–Meier estimator $(\hat{F}_n)$ of the distribution function F of the variable T at each value $y_{(i)}$ (Kaplan & Meier, 1958), that is:

$$w_{[i]} = \hat{F}_n(y_{(i)}) - \hat{F}_n(y_{(i-1)}) = \frac{\delta_{[i]}}{n-i+1} \prod_{j=1}^{i-1} \left[\frac{n-j}{n-j+1} \right]^{\delta_{[j]}} \tag{4}$$

where $\delta_{[i]}$ is the value of the noncensored indicator variable associated with the i-th ordered observation $y_{(i)}$.

The weighted least squares solution of the minimization problem in Equation (3) is $\hat{f}(x) = B\hat{\gamma}$, where B denotes the $n \times q$ matrix with $B_{ij} = B_j(x_i)$ and $\hat{\gamma}$ is a $q \times 1$ vector of coefficients $\hat{\gamma}_1, ..., \hat{\gamma}_q$. In addition, $\hat{\gamma} = (B^{'} W B)^{-1} B^{'} W Y$, where W is a $n \times n$ diagonal matrix with the Kaplan–Meier weights and Y is the vector of observed survival times. As in the noncensored case this censored B-splines approach has the problem of the selection of the knots. As mentioned above, one possible solution is to choose a large number of knots and use a penalty term to control the smoothness. Thus, based on the proposal of Eilers and Marx (1996), expression (2) is modified to take into account the presence of censored data and we propose what can be called a censored P-spline approach by minimizing the following expression:

$$\sum_{i=1}^{n} w_{[i]} \left[y_{(i)} - \sum_{j=1}^{q} \gamma_j B_j(x_{[i]}) \right]^2 + \lambda \sum_{j=k+1}^{q} (\Delta^k \gamma_j)^2 \tag{5}$$

Therefore, in the minimization problem (5) we consider several issues: the goodness of the fit with the weighted sum of squared residuals, the smoothness of the function with the penalty term, and the presence of the censored data with Kaplan–Meier weights. Expression (5) can be rewritten in matrix form as

$$(Y - B\gamma)^{'} W (Y - B\gamma) + \lambda \gamma^{'} D_k^{'} D_k \gamma \tag{6}$$

where D_k is the matrix representation of the difference operator Δ^k. The most common order of the difference in practice is $k = 2$. In this case, the matrix representation of the difference operator is

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \ldots \\ 0 & 1 & -2 & 1 & \ldots \\ 0 & 0 & 1 & -2 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Expression (6) is minimized by $\hat{\gamma} = (B^{'} W B + \lambda D_k^{'} D_k)^{-1} B^{'} W Y$. Therefore the fitted curve using the censored P-splines methodology proposed is $\hat{f}(x) = \sum_{j=1}^{q} \hat{\gamma}_j B_j(x)$. This approach was shown to be consistent in the pure parametric censored linear regression model context (Stute, 1993) and for parametric censored nonlinear regression (Stute, 1999), provided that the following identifiability conditions are met: (i) independence between the lifetimes and the censoring times; and (ii) given the duration, the covariate does not provide any further information as to whether the observation is censored or not, a weaker

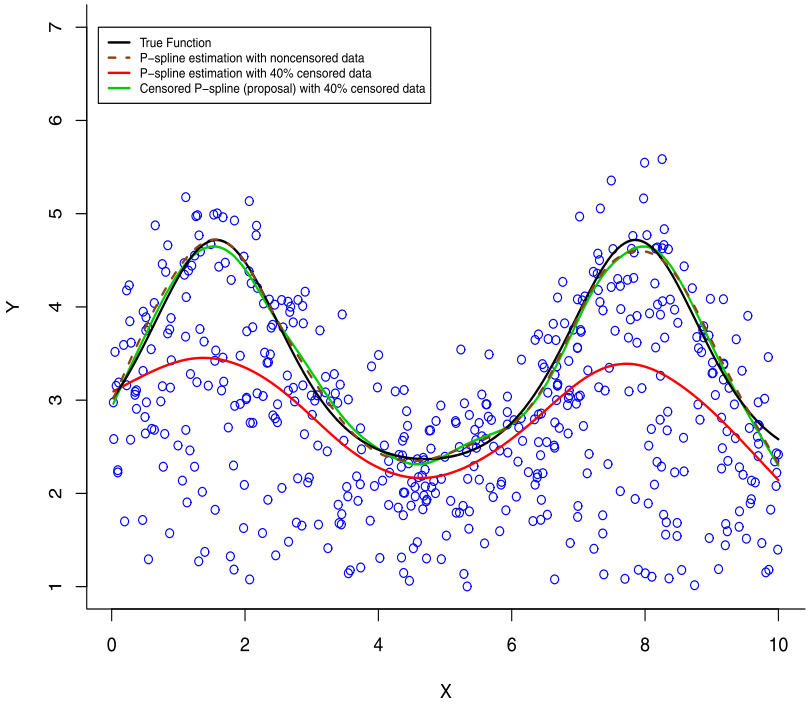


FIGURE 2 Proposed censored P-spline estimation

assumption than the independence between the censoring times and the covariate (see Stute, 1993, 1999, for a discussion of these assumptions). Under this approach in the censored context, as in the noncensored one, the location and number of the knots are not fixed as they are with the smoothing splines approach. Therefore, the number and location of the knots need to be chosen. There is a penalty term that controls the smoothness of the function, so the number of knots is not a crucial parameter; the idea is to use a sufficiently large number of knots to fit the data. In this study, we consider equidistant knots whose number can be chosen by applying the default formula presented in Ruppert (2002) based on Wand's default choice. That is, Ruppert proposes choosing the following number of knots: $K = \min([m/4], 40)$, with $[r]$ being the floor of r, that is the greatest integer value equal to or less than r, and m is the number of different values of the variable X. Ruppert (2002) studies the performance of this default rule in the penalized splines approach and concludes that the default chooses an effective number of knots in all the cases studied. Therefore this recommendation seems reasonable.

In order to complete our new proposal, as in any smoothing technique, we must choose the optimal level of smoothing, that is the value of parameter λ in Equation (5). We use the generalized cross validation criterion to estimate this parameter. For the noncensored case this is the value that minimizes the following expression:

$$GCV_{nc} = \sum_{i=1}^{n} \frac{(t_{(i)} - \hat{t}_{(i)})^2}{(n - tr(H_{nc}))^2}$$

where $H_{nc} = B(B^{'}B + \lambda D_k^{'} D_k)^{-1} B^{'}$ is the smoother matrix for the noncensored case. But for the censored data case this expression must be modified to take into account the effect of censoring, so we propose minimizing the following expression:

$$GCV_c = \sum_{i=1}^{n} \frac{w_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{(n - tr(H_c))^2} \tag{7}$$

where $H_c = B(B^{'}WB + \lambda D_k^{'} D_k)^{-1} B^{'}W$ is the smoother matrix for the censored case. W is a diagonal matrix with the $w_{[i]}$ Kaplan–Meier weights associated with the corresponding $y_{(i)}$ observed survival values.

Finally, we apply the methodology of censored P-splines proposed to the sample with censored data, that is Equation (5) is minimized using B-splines of order 3 and a penalty term of order 2, the most common values in practice. The optimal value of smoothing parameter λ is obtained by minimizing the expression (7) and the number of knots using Ruppert's default. Figure 2 shows the true function $f(\cdot)$ together with the estimation results of the P-spline method for the noncensored and censored samples and the results of the censored P-spline method proposed. From Figure 2 the good performance of the censored P-spline method can be shown, as the estimated values are very close to the true function $f(\cdot)$ and the fitted curve of the noncensored case.

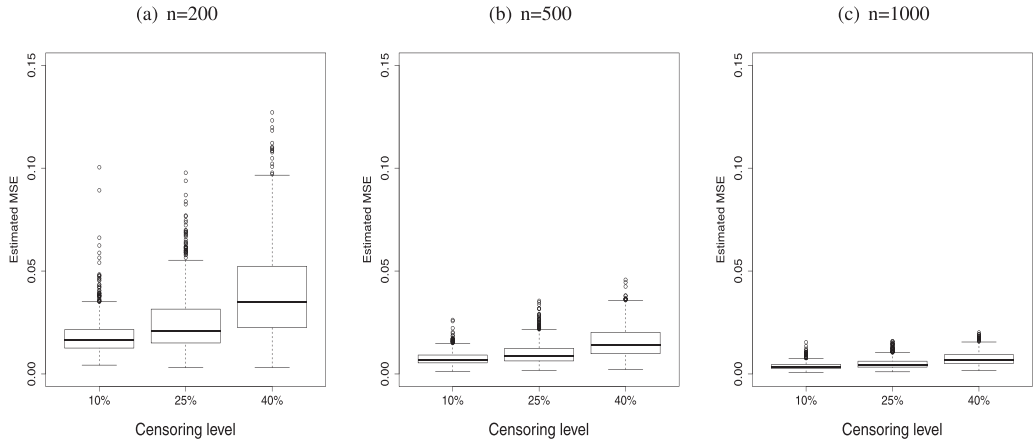


FIGURE 3 Mean square errors in the censored P-splines method using different censoring levels and sample sizes (case 1)

In any event it must be remembered that this last fitted curve (corresponding to the noncensored case) cannot be estimated in practice because the full sample is not known due to censoring. In addition, it can be appreciated that the bias of the estimation of P-spline method for the censored data is corrected with the censored P-spline proposal.

4 | SIMULATION STUDY

In this section, the performance of our proposal is studied using a simulation study. To that end we consider two different scenarios for the relationship between the logarithm of survival time, which we denote as T, and a relevant covariate X

$$t_i = f(x_i) + \epsilon_i,$$

where the values of the variable T are not completely known because some observations are censored.

4.1 | Case 1: Sinusoidal function

For the first scenario, we consider the following model for the function $f(\cdot)$:

$$f(x_i) = 2 + exp(sin(x_i))$$

where the covariate X takes values from a uniform variable in the interval (0,10), that is a $f(\cdot)$ function is used with two local maximum values and one minimum. The error term ϵ is generated as a normal variable $N(0, 0.5)$. In order to study the effect of censoring, we consider a censoring variable C generated independently from a uniform distribution $U(1, b)$. The value of parameter b changes to consider three different levels of censored data: 10%, 25%, and 40%. Therefore, we observe $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$ a sample of size n, where $y_i = min(t_i, c_i)$ is the observed survival time, that is the minimum between the survival time t_i and the censoring value c_i. In addition, it is known through the indicator variable $\delta_i = I(t_i \leq c_i)$ which observations are not censored. We use three sample sizes: $n = 200$, $n = 500$, and $n = 1,000$. For each example we consider 1000 Monte Carlo replications.

We consider, as usual in practice, that the functional form of the relationship between the response variable and the covariate is unknown and estimate the model using the censored P-spline method proposed in the previous section. The results are summarized in Figures 3 and 4. Figure 3 gives the box plots (one for each censored level considered) with the results of the estimated mean squared error (MSE) for each replication. The MSE is defined as follows:

$$MSE = \frac{\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2}{n}$$

Figure 3a shows the result for the $n = 200$ sample, Figure 3b for $n = 500$, and Figure 3c for $n = 1,000$. Figure 4 shows the mean, the pointwise 95% upper and lower oscillation limits of the values estimated using the censored P-spline method, and the true function $f(x)$ for each level of censoring and sample size.

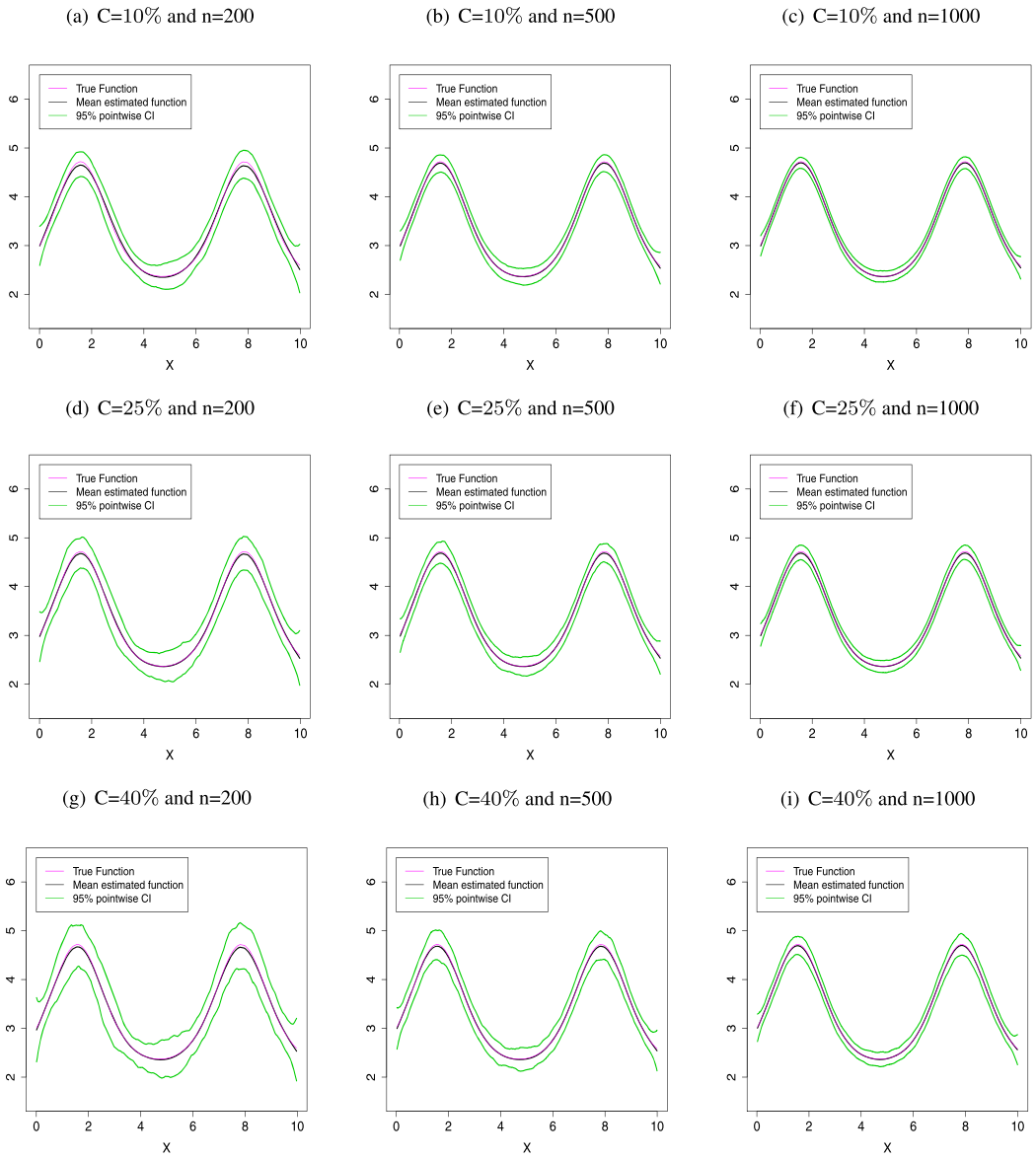


FIGURE 4 Estimated function using censored P-spline proposal (case 1)

From these results the conclusion can be drawn that the performance of the censored P-splines method proposed is good in all the cases considered: the estimated function $\hat{f}(x)$ recovers the true functional form. Figure 3 shows that the estimated mean squared error decreases when the sample size increases for each censoring level considered. Figure 4 shows that the mean of the estimated functions is very close to the true function for all the cases considered. In addition, as sample size increases the upper and lower pointwise 95% limits become closer to the true function. The effect of the censoring level is as expected: the results are more accurate with lower levels of censoring and the variability increases with the censoring level.

4.2 | Case 2: Quadratic function

Now we analyze a second example, where we consider a response variable generated following a quadratic relationship between the response variable and the covariate. That is, we consider the following model for function $f(\cdot)$:

$$f(x_i) = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$$

where $\beta_1 = 2$, $\beta_2 = 4$, and $\beta_3 = -1$. The covariate X takes values from a uniform variable in the interval of (0,4) and ϵ is generated as a normal variable $N(0, 0.5)$. The censoring variable C is generated independently from a uniform distribution

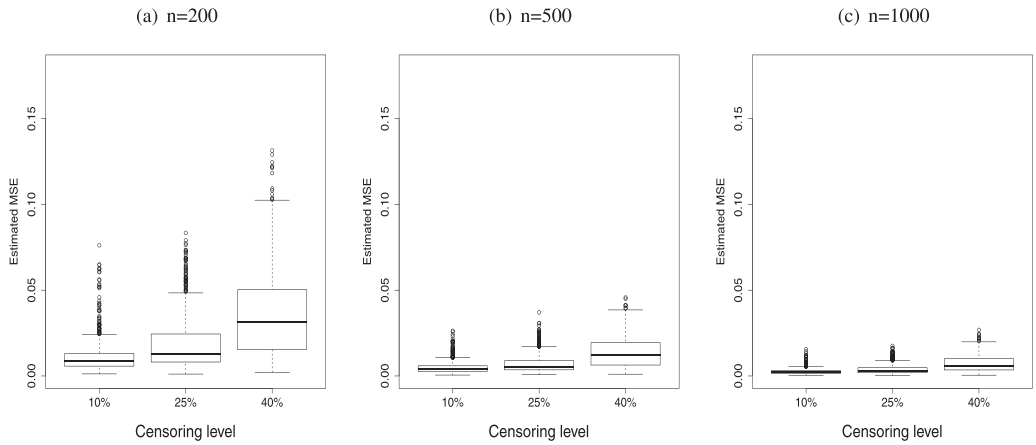


FIGURE 5 Mean square errors in the censored P-splines method using different censoring levels and sample sizes (case 2)

$U(2, b)$, changing the value of parameter b to consider three different levels of censored data: 10%, 25%, and 40%. As in the previous study, the observed survival time y_i is the minimum between the survival time t_i and the censoring value c_i.

Using the same sample sizes ($n = 200$, $n = 500$, and $n = 1,000$) and number of replications (1,000) as in the first scenario, the good performance of the censored P-splines method is shown in Figures 5 and 6. Figure 5a shows the estimated mean squared error (MSE) for each replication for the case of $n = 200$, Figure 5b for $n = 500$, and Figure 5c for $n = 1,000$. As can be seen, the estimated mean squared error decreases as sample size increases. On the other hand, Figure 6a–c shows the effect of increasing the sample size on the mean of the estimated functions and on the pointwise 95% upper and lower oscillation limits of the estimated values of the true function for a 10% censoring level, Figure 6d–f does likewise for the 25% censoring level, and Figure 6g–i for the 40% censoring level. These figures show the good fit of the mean of the estimated functions, and show that as sample size increases the performance get better.

Finally, the simulations in this subsection and in Subsection 4.1 are obtained by minimizing expression (5) using B-splines of degree 3 and a penalty of order 2, which are common values in practice. The smoothing parameter λ is chosen by cross validation using Equation (7). Moreover, the number of knots could be chosen by using the default in Ruppert (2002), but smoothing is controlled by a penalty term so the choice of the number of knots is not crucial. In some cases of small samples and high censoring levels this choice of the number of knots could cause problems in the estimation due to the lack of information in the sample. We have modified this formula for choosing knots by multiplying by one minus the proportion of censored observations in the sample (C). Therefore, the number of knots is equal to $round(\min(m/4, 40) \cdot (1 - C))$. With low censoring levels or, more importantly, with large sample sizes the results obtained would be similar to those obtained with the default in Ruppert (2002).

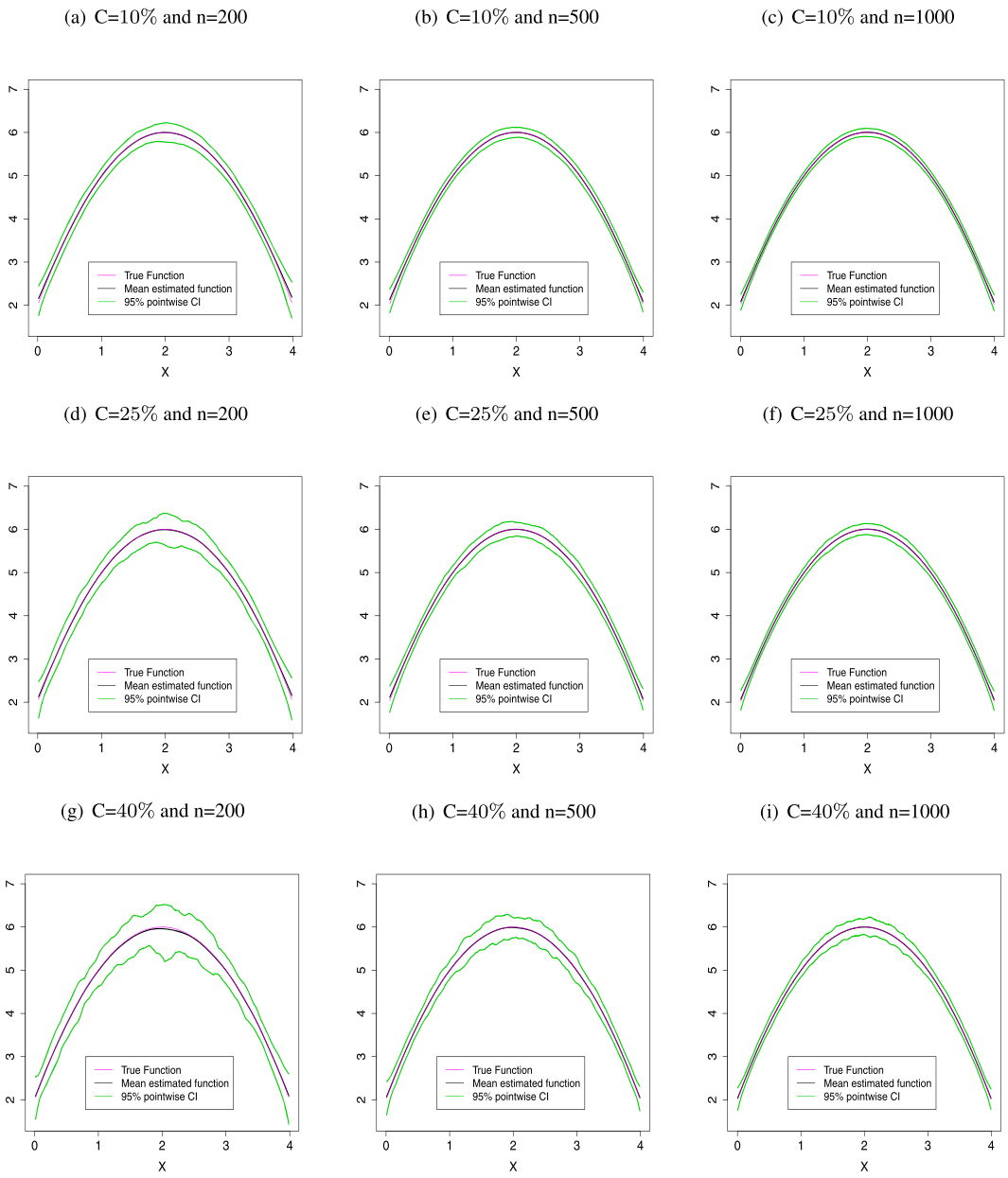
As suggested by the referees, we also conduct additional simulations for small sample size and consider nonnormal, asymmetric error distributions such as the Weibull distribution. The new results obtained (not shown) confirm the good performance of the proposed method and are consistent with those presented in this section.

5 | EXTENSION TO GAM FRAMEWORK

The ideas that we have used to extend the P-splines methodology proposed by Eilers and Marx for the case of censored variable response using the Kaplan– Meier weights could be extended to other types of model. Thus, following the valuable suggestion of an anonymous reviewer, we have considered extending our proposal to a generalized additive models (GAM) framework. In this section, we analyze the performance of this proposal for models with one and two regressors.

5.1 | One regressor

To illustrate the performance of the extension we use the same scenarios as in the simulations in the previous section: a sinusoidal function (case 1) and a quadratic one (case 2). Figure 7 shows the mean of the estimated functions obtained following our



(a) C=10% and n=200 (b) C=10% and n=500 (c) C=10% and n=1000

(d) C=25% and n=200 (e) C=25% and n=500 (f) C=25% and n=1000

(g) C=40% and n=200 (h) C=40% and n=500 (i) C=40% and n=1000

FIGURE 6 Estimated function using censored P-spline proposal (case 2)

proposal (censored P-spline) and those based on the function *gam* from R package *mgcv* (R Core Team, 2015; Wood & Wood, 2015) for a sample size of 500, a censorship level of 25% and 1,000 replications. Figure 7a–c displays the results for case 1 and Figure 7d–f for case 2. Figure 7a and d shows estimates obtained using the *gam* function with *tp* (Wood, 2003, optimal low rank approximation to thin plate spline) and *ps* (Eilers and Marx style P-splines) basis-penalty smoothers without correcting the effect of censorship. As expected, the estimation has an substantial bias. Figure 7b and e presents the estimation using the *gam* function but corrected by taking into account the effect of censoring with Kaplan–Meier weights. As can be seen, the estimation is considerably improved, with slight differences between the three estimators analyzed. In Figure 7c and f not only is the censorship effect corrected with Kaplan–Meier weights but the smoothing parameter and the number of knots are chosen using our proposal (smoothing parameter selected by GCV_c and number of knots using a modified Ruppert's formula). As shown in these last figures the curves are so similar as to be indistinguishable.

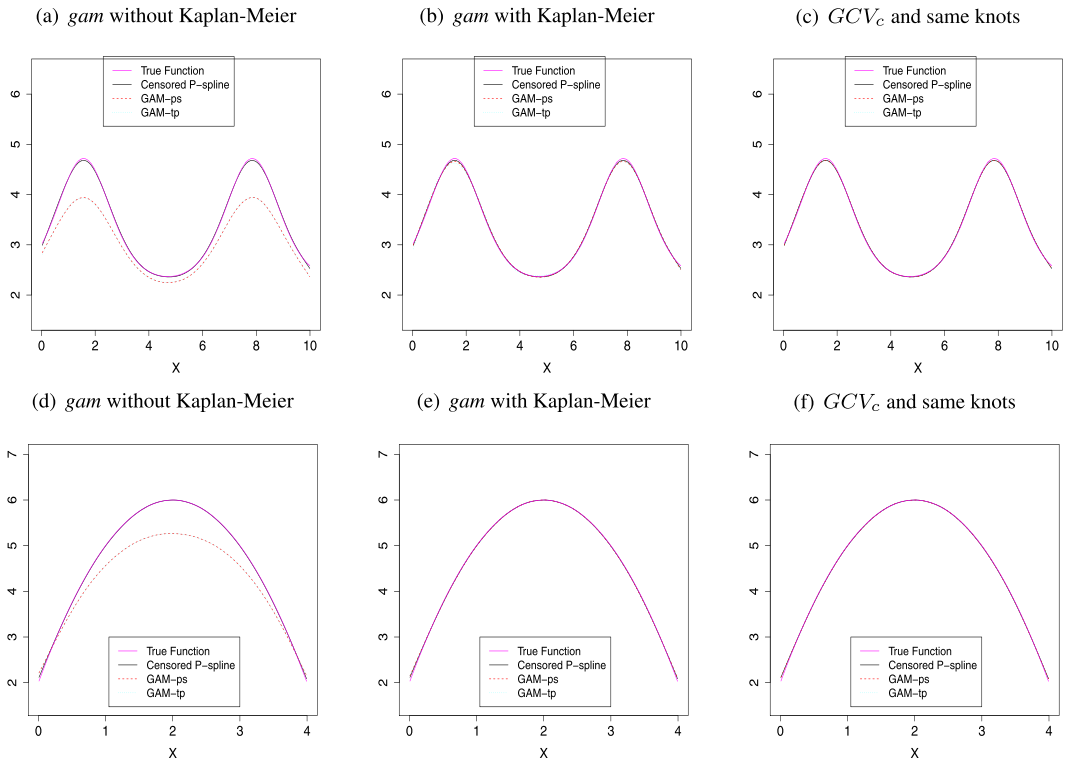


FIGURE 7 Mean of the estimated functions: censored P-spline and *gam* function

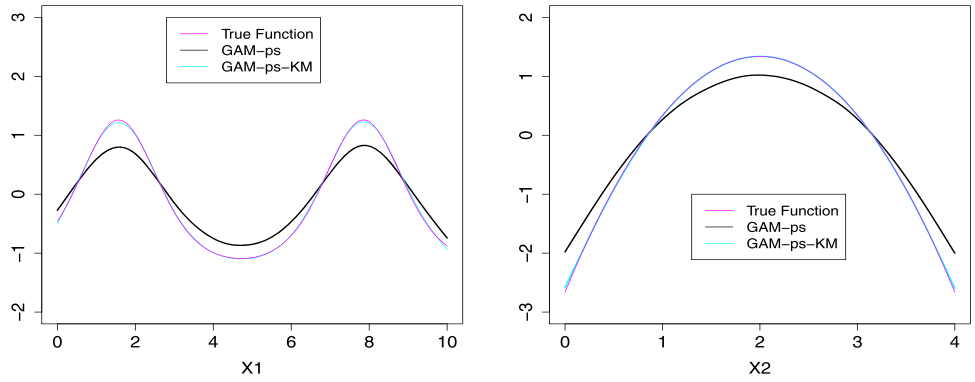


FIGURE 8 Mean of the estimated functions

5.2 | Two regressors

Now we present an illustration for a very simple model where the response variable depends linearly on two unknown smooth functions of the regressors. We use the same functional forms with the same sample size and censorship level as the previous case, but they are aggregated in the model in an additive manner:

$$f(x_{1i}, x_{2i}) = f(x_{1i}) + f(x_{2i}) + \epsilon_i = 2 + exp(sin(x_{1i})) + 4x_{2i} - x_{2i}^2 + \epsilon_i$$

Figure 8 shows the mean of the estimated functions obtained using the *gam* function with *ps* (Eilers and Marx style P-splines) basis-penalty smoother. For each regressor, this figure displays the mean of the estimated curve without correcting the effect of censorship (GAM-ps, black line), and the estimated curve using the Kaplan–Meier weights to take into account the effect of censorship (GAM-ps-KM, blue line) where the smoothing parameter and the number of knots are chosen following our proposal. The results obtained show that the method performs well. Similar exercises have been carried out using other basis-penalty smoothers (thin plate regression splines or penalized cubic regression splines) with indistinguishable results (not shown).

TABLE 1 Regression estimates and standard deviation for log survival time versus age and age squared at the time of transplant

	Intercept		Age		Age^2	
Method	$\hat{\beta}_1$	sdev($\hat{\beta}_1$)	$\hat{\beta}_2$	sdev($\hat{\beta}_2$)	$\hat{\beta}_3$	sdev($\hat{\beta}_3$)
AFT lognormal	4.3010	1.6529	0.1931	0.0903	−0.0032	0.0012
Stute	2.5574	1.7171	0.2013	0.0886	−0.0028	0.0011

To further check the performance of the method, additional simulations (not shown) have been conducted for three levels of censoring (10%, 25%, and 40%) and three sample sizes ($n = 200$, $n = 500$, and $n = 1{,}000$) in the cases of both a single regressor and two regressors. The results obtained are consistent with those presented in Section 4, that is, the estimation improves as sample size increases and the effect of censoring level is as expected: the results are more accurate with lower levels of censoring and the variability increases with the censoring level. The simulation studies illustrate the good performance of the extension of our proposal to a generalized additive models (GAM) framework. We have incorporated a correction of the censorship effect based on Kaplan–Meier weights into the GAM framework in a simple way, thus enabling more complex models to be estimated immediately. However, our approach is extended to the multiple covariate case only for a very simple example. In general it could be necessary to consider other types of basis that are more suitable for multivariate smoothing than the one considered in this illustration.

6 | EMPIRICAL APPLICATION: STANFORD HEART TRANSPLANT DATA

The Stanford Heart Transplant dataset contains information about patients who have received a heart transplant. This transplantation program began in October 1967 and the patients were followed until February 1980. These dataset have been previously used, for example, in Miller and Halpern (1982) and Escobar and Meeker (1992), in parametric censored regression models. The dataset provides information about the observed survival time from the date of transplant in days for 184 transplant cases, their ages at first transplant in years and an indicator of patient status (dead or alive) in February 1980. As in Miller and Halpern (1982) we restrict the sample to heart transplant patients who survive at least 10 days (176 patients). The dataset can be found in Miller and Halpern (1982) and can also be downloaded from the R package *survival* (R Core Team, 2015; Therneau, 2015).

The studies by Miller and Halpern (1982) and Escobar and Meeker (1992) deal with the relationship between the age covariate and the survival response variable. They conclude that a quadratic is a reasonable relationship between log survival time (T) and Age:

$$T = \beta_1 + \beta_2 Age + \beta_3 Age^2 + \epsilon \tag{8}$$

Assuming that the above parametric specification is correct, two methodologies known and proposed in the literature on survival analysis can be used to fit the model (8). These estimators can be used as a benchmark to evaluate the performance of the censored P-spline method proposed. The first and more restrictive approach is the parametric Accelerated Failure Time (AFT) methodology (Kalbfleisch & Prentice, 2002), based on the restricted assumption of knowing the probability distribution of the response variable, that estimates the β coefficients of the model using the maximum likelihood estimator. Thus, considering an AFT lognormal model, we estimate the β coefficients assuming a normal probability distribution. This can be considered as a censored parametric method of estimation. The second methodology, proposed by Stute (1993), is less restrictive in that it does not need the assumption of the probability distribution of the response variable, but it also trusts the functional form presented in Equation (8). That is, it needs to know the form of the relationship between the response variable and the covariate. This can be classed as a censored semiparametric method of estimation. The latter methodology estimates coefficients using weighted least squares via Kaplan–Meier weights (Stute, 1993). The results for the AFT lognormal method and Stute's method applied with the quadratic specification assumed above are presented in Table 1. The estimated coefficients of the linear and quadratic effects are very similar for the two afore-mentioned estimation methods.

The validity for the estimation results for these two approaches is based on the confidence of the specified relationship (8). If this relationship is not correct both approaches lead to wrong conclusions. As a robust solution to avoid this possibility of incorrect specification we use our proposal of a censored P-spline method to estimate the relationship between survival and the age covariate. Thus, the fitted curve is obtained from minimizing Equation (5) with cubic B-splines and a penalty term of order two. This approach is more flexible than those mentioned above, as it does not assume any functional form for the true relationship. It can be considered as a censored nonparametric method of estimation. We also estimate the relationship with the

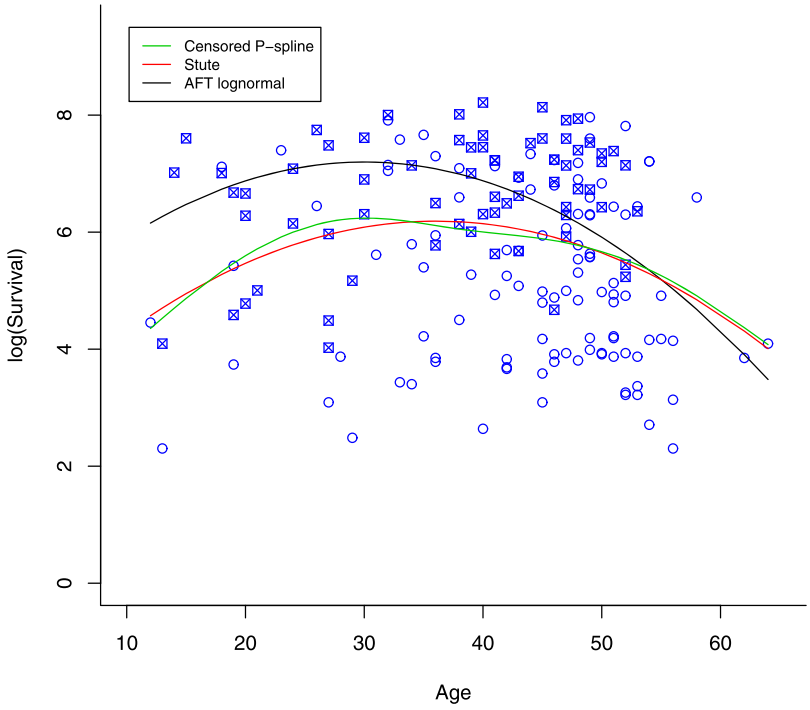


FIGURE 9 Estimated relationship using three methodologies: AFT lognormal, Stute's approach and censored P-spline

gam function incorporating the censorship correction with the smoothing parameter and the number of knots chosen using our proposal with indistinguishable results (not shown). Figure 9 shows the estimations of these three approaches with the scatterplot of observed log survival time versus age. Patients indicated by ○ are dead and those indicated by ⊠ are alive in February 1980; that is, the dead patients have uncensored survival times and the live patients have censored survival times. Sixty-nine patients are alive and one hundred seven dead in February 1980.

In conclusion, the AFT methodology and Stute's proposals can be applied only when the functional form of the effect of the covariate X on the response variable is known exactly. In this application, it seems that the relationship between log survival and age is quadratic, so both these methodologies perform reasonably well. However, the nonparametric censored P-spline approach adequately estimates the quadratic relationship, obtaining very similar results to the previous ones. Nevertheless, it must be mentioned that if the functional form or the probability distribution are wrongly chosen, these two methods lead to a serious problem of incorrect specification of the model and therefore to incorrect conclusions. An important advantage of our approach is that it does not need to assume any functional form and therefore it avoids this problem.

7 | CONCLUSION

In this paper, we introduce a simple, flexible method of nonparametric modeling in the context of censored data. Specifically, we propose an extension of the penalized splines approach (Eilers & Marx, 1996) that constitutes a novelty in this context, using Kaplan–Meier weights to take into account the effect of censorship. Moreover, for choosing the best value of the smoothing parameter we have adapted the generalized cross validation criterion for the case of censored samples. Our proposal does not need to assume a specific probability distribution for the response variable which in practice is more difficult to check in the presence of censoring and enables the true functional form between the variable of interest and the regressor to be estimated without assuming a specific parametric functional form. Therefore we avoid problems of model specification that lead to biased estimations and erroneous conclusions, as may occur when a parametric censored regression model is specified wrongly. Its application in samples with censored data is useful in contexts of survival or duration analysis where censored observations are common. Among other good properties (see reference Eilers et al., 2015), this nonparametric approach has the important advantage that it is more appropriate than the smoothing splines approach for large samples. That is, for spline smoothing the dimension of the corresponding spline basis grows with the sample size. Using P-splines reduces the dimensionality problem considerably, from the number of observations in the sample to the number of B-splines. Moreover, since the use of

cross-validation techniques entails estimating the curves several times it is important to note that the penalized splines approach estimates the curves in a significantly shorter time than other smoothing approaches. Finally, it is very easy to implement and is easy to interpret because it directly models the effect of explanatory variables on survival, so the interpretation of the results is clearer and easier (in terms of effects on mean survival time, as in the classical statistical models) than in hazard models, where the effect of the covariate is modeled on a conditional probability.

The simulation studies conducted illustrate the good performance of the method proposed and it can be verified that it estimates the true functional form satisfactorily. Furthermore, as expected, the accuracy of estimates improves as the sample size increases and the percentage of censorship is reduced. The application to real data serves to illustrate the potential advantages of its use when the true functional form of the relationship is not known.

Furthermore, we have extended our proposal introducing the correction of the censorship effect based on Kaplan–Meier weights into the generalized additive models (GAM) framework in a simple way, thus enabling more complex models to be estimated immediately. However, our approach is extended to the multiple covariate case only for a very simple example. Another possibility would be to extend the proposal to the context of partial linear models with two additive components: one linear that normally covers discrete covariates or variables that are known to influence the response variable in a linear way and the other component that is to be modeled in a nonparametric way using our methodology. To generalize our approach to the multivariate smoothing case without assuming a probability distribution is no trivial matter. Moreover, it could be more appropriate to consider another type of bases better suited to multivariate smoothing. Therefore, we believe that this interesting generalization could be the subject of a new work for study in the future (see, e.g. Wood, 2006).

Finally, in this paper we have presented the estimation method and analyzed its performance in different scenarios. For implementation of the method in a practical settings it would be an interesting topic for future study to consider the possibility of making inferences, for example using bootstrap re-sampling techniques following the idea in Orbe and Nuñez (2013).

ACKNOWLEDGMENTS

We thank an associate editor and two anonymous referees for their careful reading of the paper and suggestions, which have improved its presentation. This study was supported by the Spanish Ministry of Science and Innovation, the Basque Government and the UPV/EHU under research grants ECO2013-40935-P, ECO2016-76884-P, UPV/EHU Econometrics Research Group IT-642-13, and UFI 11/03 Sustainable Economics & Welfare. The authors have no conflict of interest to declare.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Jesus Orbe http://orcid.org/0000-0001-5543-7443

REFERENCES

Cai, T., Hyndman, R. J., & Wand, M. (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics*, *11*(4), 784–798.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B*, *34*, 187–220.

Currie, I. D., Durban, M., & Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, *4*(4), 279–298.

De Boor, C. (2001). *A practical guide to splines, revised version*. Applied mathematical sciences 27. New York, NY: Springer-Verlag.

Dierckx, P. (1993). *Curve and surface fitting with splines*. Numerical mathematics and scientific computation. Oxford, UK: Oxford University Press.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In: W. Schemp, & K. Zeller (Eds.), *Constructive theory of functions of several variables* (pp. 85–100). Berlin, DE: Springer.

Eilers, P. H. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, *11*, 89–121.

Eilers, P. H., Marx, B. D., & Durbán, M. (2015). Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*, *39*(2), 149–186.

Escobar, L. A. & Meeker, W. Q. (1992). Assessing influence in regression analysis with censored data. *Biometrics*, *48*, 507–528.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*. New York, NY: Marcel Dekker.

Friedman, J. H., & Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics*, *31*(1), 3–39.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. Monographs on statistics and applied probability 58. London, UK: Chapman and Hall.

Härdle, W. (1990). *Applied nonparametric regression*. Econometric society monographs 19. Cambridge, UK: Cambridge University Press.

Hennerfeind, A., Brezger, A., & Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, *101*(475), 1065–1075.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The statistical analysis of failure time data (2nd ed.)*. Wiley series in probability and statistics. Hoboken, NJ: John Wiley & Sons, Inc.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, *49*(1), 169–186.

Kauermann, G. & Khomski, P. (2006). Additive two-way hazards model with varying coefficients. *Computational Statistics & Data Analysis*, *51*(3), 1944–1956.

Kneib, T. (2006). Mixed model-based inference in geoadditive hazard regression for interval-censored survival times. *Computational Statistics & Data Analysis*, *51*(2), 777–792.

Kneib, T. & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, *34*(1), 207–228.

Komárek, A., Lesaffre, E., & Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, *14*(3), 726–745.

Konrath, S., Fahrmeir, L., & Kneib, T. (2015). Bayesian accelerated failure time models based on penalized mixtures of Gaussians: regularization and variable selection. *AStA Advances in Statistical Analysis*, *99*(3), 259–280.

Lambert, P. (2013). Nonparametric additive location-scale models for interval censored data. *Statistics and Computing*, *23*(1), 75–90.

Miller, R. & Halpern, J. (1982). Regression with censored data. *Biometrika*, *69*(3), 521–531.

Orbe, J. & Nuñez-Anton, V. (2013). Confidence Intervals on Regression Models with Censored Data. *Communications in Statistics-Simulation and Computation*, *42*, 2140–2159.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, *1*, 502–527.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, *9*(2), 363–379.

R Core Team (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, AT. URL https://www.R-project.org/.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik*, *10*(3), 177–183.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, *11*(4), 735–757.

Ruppert, D., Wand, M., & Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, *3*, 1193–1256.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability 26. London, UK: Chapman and Hall.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, *45*(1), 89–103.

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica*, *9*(4), 1089–1102.

Therneau, T. M. (2015). A Package for Survival Analysis in S. URL https://CRAN.R-project.org/package=survival. R package version 2.38.

Wahba, G. (1990). *Spline models for observational data*. CBMS-NSF regional conference series in applied mathematics 59. SIAM: Society for Industrial and Applied Mathematics, Philadelphia.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114.

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, *62*(4), 1025–1036.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Texts in statistical science series. Boca Raton, FL: CRC press.

Wood, S. & Wood, M. S. (2015). Package mgcv. *R package version*, pp. 1–7.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Orbe J, Virto J. Penalized spline smoothing using Kaplan–Meier weights with censored data. *Biometrical Journal*. 2018;60:947–961. https://doi.org/10.1002/bimj.201700213

Capítulo 8

Selecting the smoothing parameter and knots for an extension of penalized splines to censored data

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2021, VOL. 91, NO. 14, 2953–2985
https://doi.org/10.1080/00949655.2021.1913737

Taylor & Francis
Taylor & Francis Group

Selecting the smoothing parameter and knots for an extension of penalized splines to censored data

Jesus Orbe and Jorge Virto

Department of Econometrics and Statistics, University of the Basque Country UPV/EHU, Bilbao, Spain

ABSTRACT

The combination of P-splines and Kaplan–Meier weights provide a flexible approach to nonparametric modelling in the context of censored data. To apply this methodology, it is necessary to choose the smoothing parameter and the number and location of the knots. In this paper, we propose a new criterion for choosing the smoothing parameter adapted to the case of uncensored data. In addition, alternatives to the methods used in the literature on uncensored data are proposed for choosing the location and number of knots. Using a simulation study we analyse the effectiveness of the various alternatives proposed in situations with differences in the information available and show that their performance is quite satisfactory. A real dataset from Mayo Clinic Primary Biliary Cirrhosis data is also used to illustrate the methodology proposed. Finally, we offer some guidelines to help the user choose the parameters in the practical application of the methodology.

ARTICLE HISTORY
Received 30 June 2020
Accepted 3 April 2021

KEYWORDS
Censored data; Kaplan–Meier weights; nonparametric estimation; penalized splines; survival analysis

1. Introduction

This paper analyses the problem of fitting a nonparametric curve in the specific context of censored data. In this context, the sample is not fully observed because some of the data values are only partially known, i.e. for some individuals, it is not the actual value of the variable of interest but a lower value that is observed. This is known as a right-censored data sample. This situation is very common in survival or duration analysis when studying the relationship between survival time or the time until a certain event occurs T, the variable of interest and a relevant covariable X:

$$T = f(X) + \epsilon$$

Frequently, the functional form between the two variables, $f(\cdot)$ is not known and instead of a particular parametric relationship being assumed, it is assumed only that it is a smooth function of the data. This is, therefore, considered a nonparametric approach to fit $f(\cdot)$ that avoids any incorrect specification of the relationship that would lead to a biased estimation and wrong conclusions. The problem of estimating nonparametric curves has been widely

CONTACT Jesus Orbe jesus.orbe@ehu.eus

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

studied in cases where the data available are complete, i.e. not censored and there are many papers in this area. Numerous methods have been proposed and analysed using different approaches. For example, there are methods based on kernel smoothers [1,2] that obtain the estimate at each covariable value as a weighted average of the local observations of the variable of interest. We focus here on spline smoothers, a different approach based on smoothing with splines [3–5].

Splines are pieces of polynomial functions joined at points known as knots, where certain conditions or constraints are set on the continuity of the function and some of its derivatives. Splines depend on the degree of the polynomial and the number and location of the knots. The choice of these elements has been widely studied (for example, [6–8]). There are various proposals in the literature on spline smoothers, but two main approaches can be distinguished: smoothing splines and regression splines. There is also a methodology that combines the best of these two approaches: the splines with penalties. Under this approach, O'Sullivan [9,10], seeking to avoid the problem of knot selection, proposes the use of a large number of knots and introduces a penalty of the second derivative of the function to be estimated, analogous to smoothing splines. Eilers and Marx [11] simplify and generalize O'Sullivan's proposal by introducing a different penalty that considers the difference of the adjacent B-spline coefficients, the bases used in this regression. This proposal is known as the penalized spline (P-splines) approach (for further details see reference [12]). It reduces the size of the problem and the complexity of the calculations.

Orbe and Virto [13] propose an extension of the penalized spline approach using Kaplan–Meier weights [14] to take into account censored response variables. They also extend this proposal to the framework of generalized additive models, introducing some modifications that take into account the effect of censoring in an analogous way, which allows the possibility of estimating more complex models in the presence of censored observations.

As in the uncensored framework, to apply the above proposals the optimal level of smoothing must be determined. For the uncensored case, the generalized cross-validation criterion (GCV) or the Akaike information criterion is usually used. It is also necessary to choose the location and number of knots, which are not fixed as in smoothing splines. Given that there is a penalty term that controls the smoothness of the function, the choice of the number of knots is not a crucial issue as long as a number of knots that is large enough to fit the data is chosen. It is common in the literature to select the number of knots by applying the default formula presented in Ruppert [8].

When extending this methodology to the case of censored data, it is necessary to adapt and study the proposals for choosing the smoothing parameter and the location and number of knots. In the literature, this topic has been studied by Aydin and Yilmaz [15] but they consider the transformed versions of the censored observations, called synthetic data, proposed by Koul et al. [16]; see [17,18] for more details. In our approach in this work, we do not need to generate a synthetic variable, instead we are going to use the same censored variable that we observe in the sample without modifying it and using Kaplan–Meier weights to control the effect of the censorship. In addition to taking into account the effect of censorship through a different approach, it should be noted that we have two additional differences compared to the work of Aydin and Yilmaz [15], since in the estimation criterion we use different bases and penalty terms.

In this paper, we propose various alternatives for choosing the optimal level of smoothing and the location and number of knots for this context. These alternatives are analysed and compared in five examples of functional forms of different complexity with different sample information scenarios, where different sample sizes and levels of censoring are combined. For the proposal to choose the optimal level of smoothing, we start from the generalized cross-validation criterion (GCV) and the modification proposed by Kim and Gu [19] to avoid the overfitting that usually arise when using that criterion. But if we apply the GCV criterion normally used in the case of uncensored data it will lead us to choose a smoothing parameter that would generate important biases in the estimation. Therefore, we have formulated new criteria adapted to the censored case. Thus, in Section 3 different alternatives to the GCV criterion are proposed using Kaplan-Meier weights to take into account the effect of censoring.

For choosing the number of knots, a modification of Ruppert's proposal Ruppert [8] is adapted to the censored case. Finally, for locating the knots we propose and analyse an adaptation, in addition to the case of equally spaced knots, to the case of censored data that uses vectors of nonuniform knots with a location according to the Kaplan–Meier weights. This issue of knot location has not been addressed previously in the literature for the censored case. Therefore, in Section 3 we propose not only criteria for the choice of the number of knots but also for their location.

The performance of all these proposals is analysed by means of an extensive simulation study and as a result, a number of guidelines are proposed for their implementation in practice.

The next section describes the extensions of the estimation methodology for the censored case. Section 3 presents the proposals for choosing the smoothing parameter and the number and location of the knots. Section 4 analyses and compares the different proposals in a wide range of scenarios and ends with a summary of the main results. Section 5 presents an application of the method to a real dataset and the paper concludes with a discussion in Section 6.

2. Methodology

The penalized spline (P-splines) method presented in Eilers and Marx [11] for the case of noncensored data assumes a sample of observations (t_i, x_i) for $i = 1, \ldots, n$ and consider the nonparametric simple regression

$$T = f(X) + \epsilon$$

where T is the variable of interest, X a relevant covariate and ϵ is the error term that satisfies $E(\epsilon \mid X) = 0$.

Eilers and Marx [11] use a basis approach based on B-splines [20,21] for modelling the function $f(\cdot)$ as $\hat{f}(x) = \sum_{j=1}^{q} \hat{\gamma}_j B_j(x)$. Thus, they use a set of q B-spline basis functions of degree d, $B_1(x), \ldots, B_q(x)$, and add a penalty term to reduce the problem of choosing the number and position of the knots. The penalty term used by Eilers and Marx [11] is not the usual integrated squared second derivative of the fitted function proposed by O'Sullivan [9,10]. They introduce a different roughness penalty based on the difference in adjacent B-spline coefficients γ_j. Thus, Eilers and Marx [11] propose the P-spline estimator, which

consists of applying penalized least squares to estimate the function $f(\cdot)$, the fitted curve, by minimizing this expression:

$$\sum_{i=1}^{n}[t_i - \sum_{j=1}^{q}\gamma_j B_j(x_i)]^2 + \lambda \sum_{j=k+1}^{q}(\Delta^k \gamma_j)^2 \tag{1}$$

where $\Delta\gamma_j$ denotes the difference between B-spline coefficients $(\gamma_j - \gamma_{j-1})$, $\Delta^k\gamma_j$ is the difference of degree k and λ is the smoothing parameter.

Orbe and Virto [13] extend the P-spline approach Eilers and Marx [11] to handle censored responses. Thus, this approach combines penalized splines with a weighted least squares estimation method using Kaplan–Meier weights to smooth censored responses following the idea of Stute [22]. To present the proposal, assume that $t_1, \ldots, t_n$ are independent observations from an unknown probability distribution function F of the response variable T. These values may not all be observable due to censoring times $c_1, \ldots, c_n$ for each of the individuals. Moreover, $x_1, \ldots, x_n$ are the values of the covariate X. Therefore, when there is censored data, one observes $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$ a sample of size n where $y_i = \min(t_i, c_i)$ is the observed response variable, which is the minimum between the response variable t_i and the censoring value c_i. In addition, it is known which observations are not censored, via the indicator variable $\delta_i = I(t_i \leq c_i)$. Remember that this situation is different from the uncensored case, where complete information is available, i.e. the sample (t_i, x_i) for $i = 1, \ldots, n$ is fully known.

Thus, based on the proposal of Eilers and Marx [11], expression (1) is modified to take into account the presence of censored data and we propose what can be called a 'censored P-spline approach' by minimizing the following expression:

$$\sum_{i=1}^{n} w_{[i]}[y_{(i)} - \sum_{j=1}^{q}\gamma_j B_j(x_{[i]})]^2 + \lambda \sum_{j=k+1}^{q}(\Delta^k \gamma_j)^2 \tag{2}$$

where $y_{(1)}, \ldots, y_{(n)}$ are the ordered values of the observed response variable $y_i = \min(t_i, c_i)$, $x_{[i]}$ is the value of the covariate associated with the ith ordered observation $y_{(i)}$, and $w_{[i]}$ is the Kaplan–Meier weight assigned to $y_{(i)}$. The Kaplan–Meier weights can be calculated as the contribution or jump of the Kaplan–Meier estimator $(\hat{F}_n)$ of the distribution function F of the variable T at each value $y_{(i)}$ [14], that is,

$$w_{[i]} = \hat{F}_n(y_{(i)}) - \hat{F}_n(y_{(i-1)}) = \frac{\delta_{[i]}}{n-i+1}\prod_{j=1}^{i-1}\left[\frac{n-j}{n-j+1}\right]^{\delta_{[j]}} \tag{3}$$

where $\delta_{[i]}$ is the value of the noncensored indicator variable associated with the ith ordered observation $y_{(i)}$.

Expression (2) can be rewritten in matrix form as

$$(Y - B\gamma)W(Y - B\gamma) + \lambda\gamma' D_k' D_k \gamma \tag{4}$$

where B denotes the $n \times q$ matrix with $B_{ij} = B_j(x_i)$ and γ is a $q \times 1$ vector of coefficients $\gamma_1, \ldots, \gamma_q$. W is an $n \times n$ diagonal matrix with the Kaplan–Meier weights and Y is the

ordered vector of the observed response variable. In addition, D_k is the matrix representation of the difference operator Δ^k. The most common order of the difference in practice is $k = 2$. In this case, the matrix representation of the difference operator is

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots \\ 0 & 1 & -2 & 1 & \dots \\ 0 & 0 & 1 & -2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Expression (4) is minimized by $\hat{\gamma} = (B'WB + \lambda D_k' D_k)^{-1} B'WY$. Therefore, the fitted curve using the censored P-spline methodology (ckmPS estimator) proposed is $\hat{f}(x) = \sum_{j=1}^{q} \hat{\gamma}_j B_j(x)$.

The idea used to extend the P-spline methodology proposed by Eilers and Marx to the case of censored variable response using the Kaplan–Meier weights could be extended to other types of model. Orbe and Virto [13] extend this proposal to a generalized additive model (GAM) framework. To take into account the presence of censored observations, the same idea is taken as in the censored P-spline proposal and Kaplan–Meier weights are used to propose the estimator ckmGAM, where Kaplan–Meier weights are used to weight the contribution of each observation to the log-likelihood.

3. Choosing the smoothing parameter and knots

As in the uncensored case, the estimation proposals presented in the previous section require, on the one hand, the choice of the λ smoothing parameter and, on the other hand, the choice of the number and location of the knots. The choice of the smoothing parameter, i.e. the value chosen for the weighting of the penalty term, is the most important decision. The choice of knots is less important and it seems that once a number of knots large enough to reflect the characteristics of the function to be estimated has been chosen, the possible overfitting of the estimate is controlled by the penalty term present in the method, which controls the smoothness of the function estimated.

3.1. Smoothing parameter

To illustrate the importance of the choice of the smoothing parameter, which has already been highlighted as a fundamental choice for obtaining a good estimate of the curve or $f(\cdot)$ function, as in any smoothing technique, we present Figure 1. This figure summarizes the estimates of the censored P-spline method (ckmPS) that minimizes expression (2) for two different choices for the smoothing parameter ($\lambda = 0.001$ and $\lambda = 1$).

For both cases, B-splines of order 3 and a penalty term of order 2 (the most common values in practice) are used and 1000 Monte Carlo replications are considered. Figure 1(a) shows the results for the $\lambda = 0.001$ case, the mean and the pointwise 95% upper and lower oscillation limits of the values estimated using the censored P-spline method, and the true function $f(x)$ for a scenario that considers a 25% censoring level and a sample size of $n = 500$. Figure 1(b) shows the results for the same case but with $\lambda = 1$.

As can be seen in Figure 1(b), choosing the wrong smoothing parameter λ can lead to an incorrect estimation of the function. Therefore, the optimal level of smoothing must be

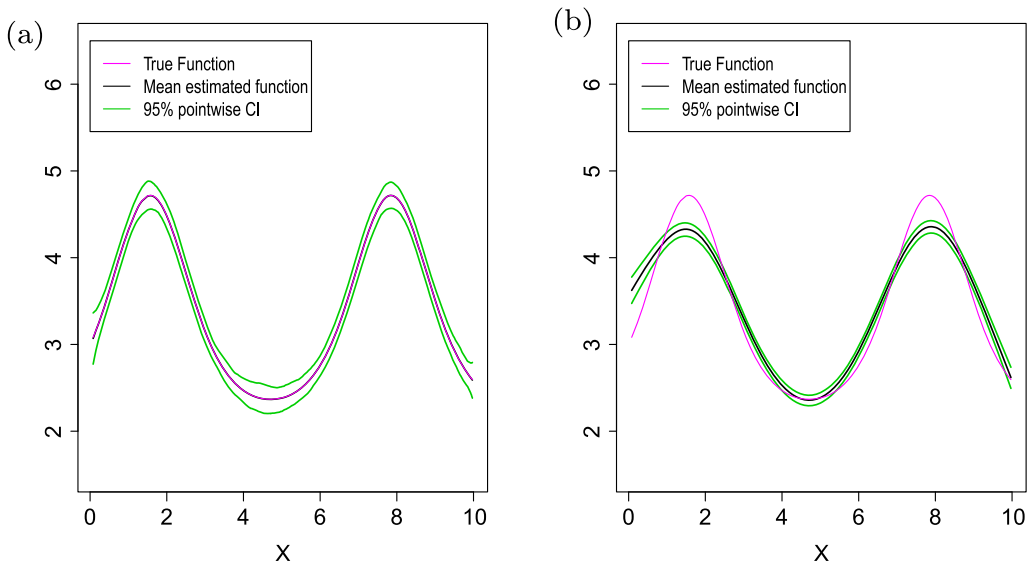


Figure 1. Estimation with two different λ values. (a) $\lambda = 0.001$, (b) $\lambda = 1$.

chosen, i.e. the value of parameter λ in equation (2). Note that there is no need to assume a specific probability distribution for the response variable, so instead of using criteria such as AIC or BIC based on deviance or likelihood we consider it more appropriate to build on the idea of generalized cross-validation and adapt it to the censored case. Thus, the GCV criterion [5] can be used to choose this parameter. As a first approximation, a value of λ could be chosen that minimizes the usual expression for the GCV criterion, where the smoothing matrix H of the uncensored case is replaced by that for the censored case H_c:

$$\mathrm{GCV}_1 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(n - tr(H_c))^2} \tag{5}$$

where $H_c = B(B'WB + \lambda D_k' D_k)^{-1} B'W$ is the smoother matrix for the censored sample case. W is a diagonal matrix with the Kaplan–Meier weights.

Figure 2(a) summarizes the estimates of the censored P-spline method which minimises the expression (2) using the value that optimizes expression (5) as a smoothing parameter λ. Thus, Figure 2(a) shows the results for the mean and the pointwise 95% upper and lower oscillation limits of the values estimated considering 1000 Monte Carlo replications. As can be seen, the method provides a poor estimation, which becomes worse as the percentage of censored observations in the sample is increased (not shown). To correct this, it seems necessary to take into account the effect of censoring on the numerator in addition to the denominator of the expression (5) with the matrix H_c. For this purpose, we propose a modification of expression (5):

$$\mathrm{GCV}_2 = \sum_{i=1}^{n} \frac{w_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{(n - tr(H_c))^2} \tag{6}$$

This new expression now also takes into account the effect of censoring on the numerator of the GCV expression, squaring the differences $(y_{(i)} - \hat{y}_{(i)})^2$ with the Kaplan–Meier weights $w_{[i]}$. Figure 2(b) shows a better performance of the estimates when the expression GCV_2

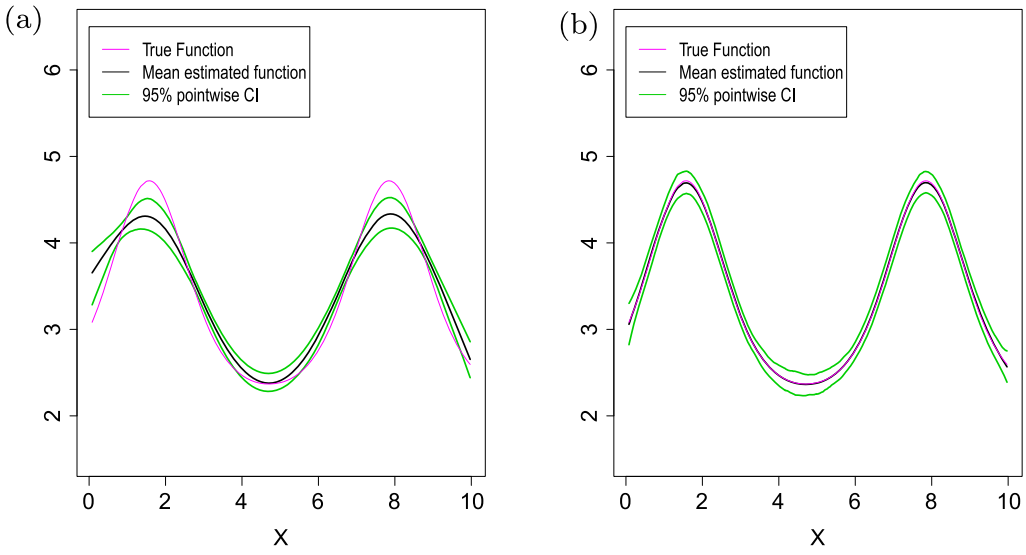


Figure 2. Criteria for the choice of the λ parameter, GCV_1 versus GCV_2. (a) GCV_1, (b) GCV_2.

from formula (6) is used as a criterion for choosing the parameter λ, compared to Figure 2(a) using the expression GCV_1 from formula (5).

In addition, we have run a simulation study to see how GCV_1 and GCV_2 proposals work. We selected a number of candidate values for parameter λ, and for each candidate value we computed the estimated mean square error (MSE) of each of the 1000 Monte Carlo replications. The left panel of Figure 3 shows the MSE of estimating the model for each specific value of the smoothing parameter in the grid of possible values for the parameter λ from 0.000001 to 4. The right panel of Figure 3 shows the MSE of estimating the model with the value of the smoothing parameter chosen by means of the GCV_1 and GCV_2 criteria. It can be seen that the proposed criterion GCV_2 works well.

As can be seen in the literature, GCV has a tendency to overfit, so it tends to do less well than it could at actual prediction. One possible solution to this problem is to set a tuning parameter $\gamma \geq 1$ in the expression GCV_2 that adds an extra penalty for model searching. In this way, our final proposal would be

$$\mathrm{GCV}_c = \sum_{i=1}^{n} \frac{w_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{(n - \gamma\ \mathrm{tr}(H_c))^2} \tag{7}$$

For the uncensored case the ordinary CGV uses $\gamma = 1$, but Wood [23] proposes correcting the overfit problem by doing what is called double cross-validation and introducing a value of $\gamma = 1.5$. This value is derived in several ways in the literature (e.g. [19]). Section 4 presents an extensive simulation study where, among other aspects, we analyse the performance of proposal (7) for choosing the optimal level of smoothing in censored samples, and compare the results considering the values $\gamma = 1$ and $\gamma = 1.5$. In addition, in this proposal for adaptation to the censored case, two possible variations are also analysed when controlling the effect of censoring: using the Kaplan–Meier weights ($w_{[i]}$), as included in the proposal presented, and their squares ($w^2_{[i]}$) to weight the numerator of the criterion.

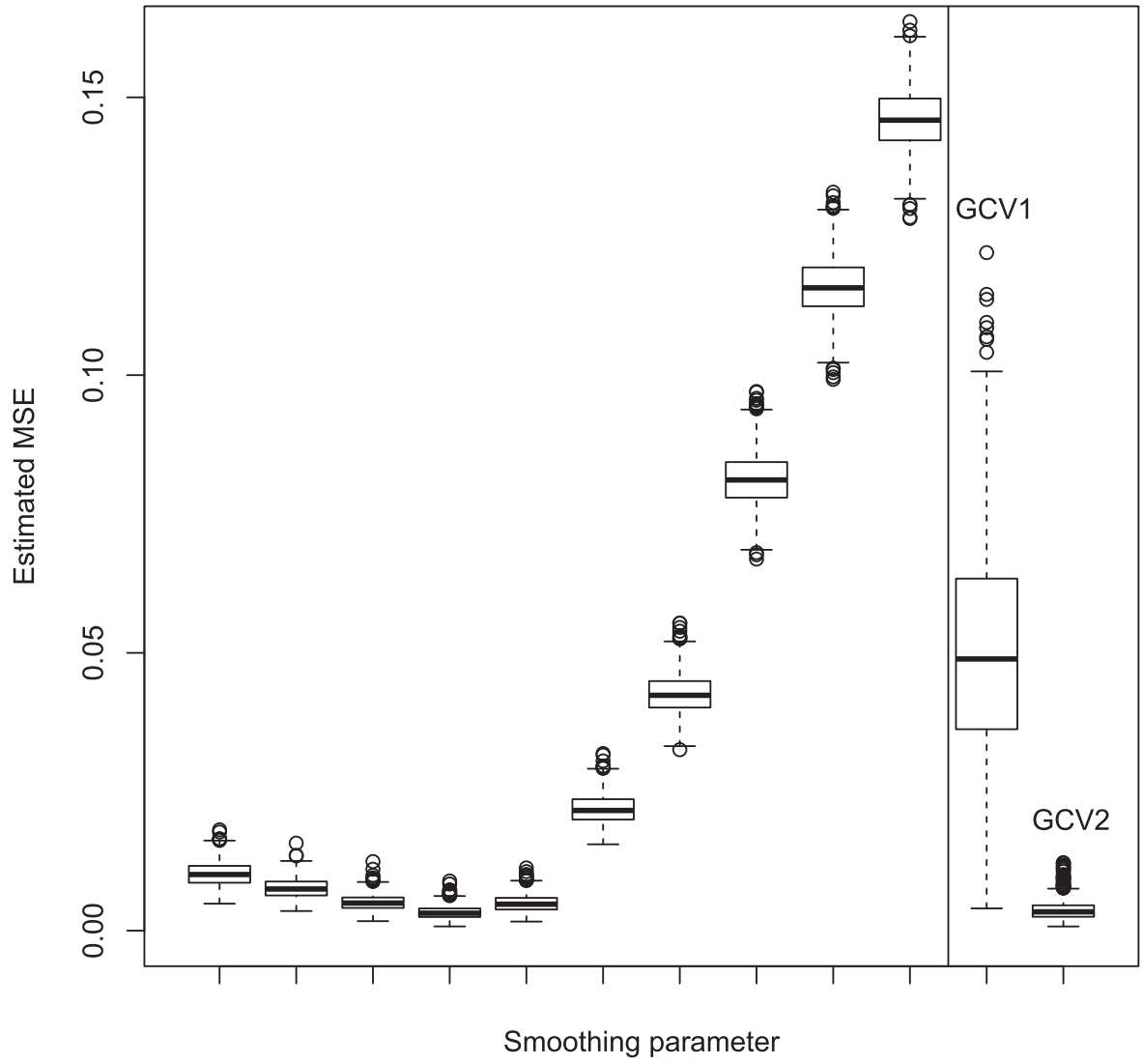


Figure 3. Estimated MSE for λ parameter in the grid 0.000001 to 4 and for λ_{GCV_1} and λ_{GCV_2}.

3.2. Knots

As already mentioned, the choice of knots is not fundamental, provided that at least enough are chosen to fit features in the data, given that a penalty term is used to prevent possible overfitting of the estimate. In any case, we also analyse here the effect on the estimation of both the number of knots and their location. It is usual in the literature to use a large enough number of knots, which can be chosen by applying the formula presented in Ruppert [8] based on Wand's default choice. That is, Ruppert proposes choosing the following number of knots:

$$K_{rp} = \text{round}\left(\min\left(\frac{m}{4}, 40\right)\right) \tag{8}$$

where m is the number of different values of the regressor X.

For the censored case, we propose a modification that takes into account the sample information available due to the existence of censored data. Thus, the number of knots (K)

to be used can be selected automatically with the selection formula that we propose:

$$K_c = \text{round}\left(\min\left(\frac{m}{4}, 40\right) \cdot (1 - C)\right) \tag{9}$$

where C is the level of censoring.

Finally, to locate knots (L), in addition to equidistant knots (L_{eq}) we propose and analyse an adaptation to the case of censored data that uses nonuniform knot vectors with the spacing of the knots as a function of the Kaplan–Meier weights (L_{km}). In a similar way to the choice of nonuniform knots using the quantiles of variable X (see, for example, [24]) it is proposed that knots be chosen in such a way that they divide the range of the variable X into continuous intervals with equal sum of the Kaplan–Meier weights. This ensures that each interval has the same Kaplan–Meier weight.

All these proposals (Sections 3.1 and 3.2) are analysed and compared in the simulation study presented in the following section for different situations.

4. Simulations

In this section, a simulation study is used to analyse the performance of the estimation with different parameter choices for the estimation proposals presented in Section 2, i.e. the ckmPS and ckmGAM estimators. To that end we consider five different examples (Table 1) for the relationship between the response variable, which we denote as T, and a relevant covariate X

$$t_i = f(x_i) + \epsilon_i,$$

where the values of the variable T are not completely known because some observations are censored. This type of data is very common in survival and duration analyses, where T measures the time (or usually its logarithm) until the occurrence of an event.

Table 1. Examples of functional form.

Example	x_i	$f(x_i)$	ϵ_i
Quadratic	$x_i \sim U[0,4]$	$2 + 4x_i - x_i^2$	$\epsilon_i \sim N(0, 0.50)$
Bump	$x_i \sim U[0,1]$	$2 + x_i + 2\exp[-\{16(x_i - 0.5)\}^2]$	$\epsilon_i \sim N(0, 0.25)$
Logit	$x_i \sim U[0,1]$	$2 + \frac{1}{1+\exp\{-20(x_i - 0.5)\}}$	$\epsilon_i \sim N(0, 0.20)$
Sine2	$x_i \sim U[0,10]$	$2 + \exp\{\sin(x_i)\}$	$\epsilon_i \sim N(0, 0.30)$
Sine3	$x_i \sim U[0,10]$	$2 + \exp\{\sin(1.6x_i)\}$	$\epsilon_i \sim N(0, 0.30)$

Table 2. Nine scenarios: three sample sizes for three levels of censoring.

Sample size (n)	Censoring level (C)	Scenario (s)	Description
	10%	1	$n = 200$ & $C = 10\%$
200	25%	2	$n = 200$ & $C = 25\%$
	40%	3	$n = 200$ & $C = 40\%$
	10%	4	$n = 500$ & $C = 10\%$
500	25%	5	$n = 500$ & $C = 25\%$
	40%	6	$n = 500$ & $C = 40\%$
	10%	7	$n = 1000$ & $C = 10\%$
1000	25%	8	$n = 1000$ & $C = 25\%$
	40%	9	$n = 1000$ & $C = 40\%$

Table 3. Results of the simulation study for the quadratic function.

						(1000 · AMSE) in each scenario								
						$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$	$s = 9$
	GCV_c		Knots			10%	25%	40%	10%	25%	40%	10%	25%	40%
Estimator	γ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmPS	1.5	1	L_{eq}	K_c	0.25	8.9	10	15.3	3.9	4.6	6.9	2.1	2.5	3.9
ckmGAM	1.5	1	L_{eq}	K_c	1.22	8.9	10.4	15.7	3.9	4.5	7.2	2.1	2.5	3.9
ckmGAM	1.5	1	L_{eq}	K_{rp}	2.24	8.9	10	16.2	3.9	4.6	7.2	2.1	2.5	4.2
ckmPS	1.5	1	L_{eq}	K_{rp}	2.33	8.9	10	16.1	3.9	4.6	7.3	2.1	2.5	4.2
ckmPS	1.5	2	L_{eq}	K_c	9.47	8.9	10.2	18.6	3.9	4.7	8.9	2.1	2.5	5
ckmGAM	1.5	2	L_{eq}	K_c	10.56	8.9	10.6	19	3.9	4.6	9	2.1	2.6	5
ckmPS	1.5	2	L_{eq}	K_{rp}	16.54	8.9	10.3	20.6	3.9	4.8	10.4	2.1	2.5	6
ckmGAM	1.5	2	L_{eq}	K_{rp}	16.78	8.9	10.3	20.7	3.9	4.8	10.5	2.1	2.5	6
ckmPS	1.5	1	L_{km}	K_c	28.62	14.8	15.6	20.7	5	5.5	7.8	2.4	2.8	4.3
ckmGAM	1.5	1	L_{km}	K_c	31.10	15.4	16.4	19.9	5.1	5.4	7.9	2.6	2.9	4.2
ckmPS	1	1	L_{eq}	K_c	36.43	9.9	13.8	26.9	4.4	6.2	11.1	2.3	3.2	6
ckmGAM	1	1	L_{eq}	K_c	36.61	10	14.4	26.7	4.2	5.8	11.5	2.3	3.3	6
ckmPS	1.5	2	L_{km}	K_c	37.68	14.9	16	23.4	5	5.7	9.5	2.4	2.9	5.3
ckmGAM	1.5	2	L_{km}	K_c	39.62	15.5	16.6	22.4	5.1	5.6	9.5	2.6	3	5.2
ckmGAM	1.5	1	L_{km}	K_{rp}	39.88	16.4	17.4	22.7	5.1	5.8	8.4	2.6	3.1	4.8
ckmPS	1.5	1	L_{km}	K_{rp}	40.11	15.6	17	24.5	5.1	5.9	8.9	2.5	3	4.9
ckmPS	1	2	L_{eq}	K_c	49.85	9.9	14.3	31.6	4.4	6.6	13.5	2.3	3.4	7.3
ckmGAM	1	2	L_{eq}	K_c	51.09	10.1	15	31.2	4.2	6.2	13.9	2.4	3.6	7.3
ckmGAM	1.5	2	L_{km}	K_{rp}	52.44	16.5	17.7	26.2	5.2	6	11.2	2.6	3.1	6.3
ckmPS	1.5	2	L_{km}	K_{rp}	54.61	15.7	17.4	28.8	5.1	6	12	2.5	3.1	6.7
ckmGAM	1	1	L_{km}	K_c	56.97	15.3	18.5	28.2	5.1	6.4	11.7	2.7	3.6	6.1
ckmPS	1	1	L_{km}	K_c	62.6	15.2	19.6	31	5.4	7.1	11.8	2.6	3.6	6.2
ckmGAM	1	1	L_{eq}	K_{rp}	63.63	10	14.5	40.8	4.4	6.5	15.9	2.3	3.4	8.4
ckmPS	1	1	L_{eq}	K_{rp}	63.68	10	14.5	41.1	4.4	6.5	15.8	2.3	3.4	8.4
ckmGAM	1	2	L_{km}	K_c	68.54	15.3	19.1	31.6	5.1	6.7	13.8	2.7	3.8	7.3
ckmPS	1	2	L_{km}	K_c	73.78	15.3	20	34.1	5.4	7.5	13.7	2.6	3.8	7.4
ckmGAM	1	1	L_{km}	K_{rp}	83.82	16	19.6	39.1	5.3	7.3	15.9	2.7	3.7	8.5
ckmPS	1	1	L_{km}	K_{rp}	97.17	15.9	22.1	46.5	5.5	7.8	17	2.7	3.9	8.8
ckmPS	1	2	L_{eq}	K_{rp}	98.06	10	15.3	62	4.4	7	20.8	2.3	3.7	11.1
ckmGAM	1	2	L_{km}	K_{rp}	106.5	16.1	20.2	46.7	5.3	7.7	20.2	2.7	4	11
ckmPS	1	2	L_{km}	K_{rp}	120.93	15.9	22.9	54.5	5.5	8.3	21.4	2.7	4.2	11.4
ckmGAM	1	2	L_{eq}	K_{rp}	124.09	10	15.2	98	4.4	7	20.8	2.3	3.7	11.1

In order to study the effect of censoring on estimation performance we consider a censoring variable C generated independently from a uniform distribution $U(1, b)$. The value of parameter b changes to consider three different levels of censored data: 10%, 25% and 40%. Therefore, we observe $(y_1, x_1, \delta_1), \ldots, (y_n, x_n, \delta_n)$ a sample of size n, where $y_i = \min(t_i, c_i)$ is the observed survival time, i.e. the minimum between the survival time t_i and the censoring value c_i. In addition, it is known through the indicator variable $\delta_i = I(t_i \leq c_i)$ which observations are not censored. We use three sample sizes: $n = 200$, $n = 500$ and $n = 1000$. For each of these nine scenarios (see Table 2), we consider 1000 Monte Carlo replications.

As usual in practice, we consider that the functional form of the relationship between the response variable and the covariate is unknown and estimate the model using two different methods: the censored P-spline method (ckmPS) and generalized additive models, but corrected by taking into account the effect of censoring with Kaplan–Meier weights (ckmGAM). The optimal value of smoothing parameter λ is obtained by minimizing

Table 4. Results of the simulation study for the bump function.

						(1000 · AMSE) in each scenario								
						$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$	$s = 9$
	GCV_c		Knots			10%	25%	40%	10%	25%	40%	10%	25%	40%
Estimator	γ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmPS	1.5	1	L_{eq}	K_c	0.17	6.5	8.1	13.4	2.8	3.4	4.9	1.5	1.8	2.5
ckmGAM	1.5	1	L_{eq}	K_c	0.17	6.5	8.1	13.4	2.8	3.4	4.9	1.5	1.8	2.5
ckmGAM	1.5	1	L_{eq}	K_{rp}	2.78	6.5	8.1	13.5	2.9	3.5	5.3	1.5	1.8	2.7
ckmPS	1.5	1	L_{eq}	K_{rp}	3.52	6.5	8.1	13.5	2.9	3.5	5.3	1.6	1.8	2.7
ckmPS	1.5	2	L_{eq}	K_c	4.78	6.4	8.3	16	2.8	3.5	5.4	1.5	1.8	2.7
ckmGAM	1.5	2	L_{eq}	K_c	4.95	6.4	8.3	16.2	2.8	3.5	5.4	1.5	1.8	2.7
ckmPS	1	1	L_{eq}	K_c	7.57	6.6	9.2	17.4	2.9	3.6	5.3	1.5	1.8	2.6
ckmGAM	1	1	L_{eq}	K_c	8.16	6.6	9.2	18.1	2.9	3.6	5.3	1.5	1.8	2.6
ckmPS	1.5	2	L_{eq}	K_{rp}	13.78	6.5	8.3	19.6	2.8	3.5	6.8	1.5	1.8	3.3
ckmGAM	1.5	2	L_{eq}	K_{rp}	14.67	6.5	8.3	20.4	2.8	3.5	6.9	1.5	1.8	3.3
ckmPS	1.5	1	L_{km}	K_c	16.10	7.5	9.2	14.5	3.5	4.2	5.6	1.8	2	2.8
ckmPS	1.5	2	L_{km}	K_c	17.43	7.4	9.1	15.4	3.5	4.2	5.8	1.8	2	2.9
ckmPS	1	2	L_{eq}	K_c	17.52	6.6	10	24.1	2.9	3.8	5.8	1.5	1.9	2.8
ckmPS	1	1	L_{km}	K_c	20.42	7.5	10	16.7	3.5	4.3	5.8	1.8	2.1	2.8
ckmGAM	1	1	L_{km}	K_c	20.62	7.6	9.7	15.6	3.7	4.4	5.9	1.8	2.1	2.8
ckmPS	1.5	1	L_{km}	K_{rp}	21.27	7.6	9.3	16.3	3.6	4.3	6.1	1.8	2.1	3
ckmGAM	1	2	L_{eq}	K_c	21.83	6.6	10	29.3	2.9	3.8	5.8	1.5	1.9	2.8
ckmGAM	1	2	L_{km}	K_c	22.33	7.5	9.8	16.9	3.7	4.4	6	1.8	2.1	2.9
ckmPS	1	2	L_{km}	K_c	24.19	7.5	10.2	18.9	3.5	4.4	6	1.8	2.1	3
ckmGAM	1.5	2	L_{km}	K_c	26.14	8	9.5	15.5	4.3	4.6	6	2	2.1	2.9
ckmPS	1.5	2	L_{km}	K_{rp}	27.23	7.5	9.3	19.3	3.5	4.3	7.1	1.8	2.1	3.4
ckmGAM	1.5	1	L_{km}	K_c	28.4	8.1	9.7	15.4	4.4	4.8	6.1	2	2.2	2.9
ckmGAM	1.5	1	L_{km}	K_{rp}	35.22	8.2	9.9	16.8	4.5	5.1	6.8	2.1	2.3	3.1
ckmGAM	1.5	2	L_{km}	K_{rp}	35.71	8.1	9.7	18.8	4.4	4.8	7.1	2	2.2	3.4
ckmGAM	1	1	L_{km}	K_{rp}	37.25	7.6	10.7	24.6	3.8	4.6	7.1	1.9	2.2	3.4
ckmPS	1	1	L_{km}	K_{rp}	37.51	7.5	10.4	27.4	3.5	4.6	7.3	1.8	2.2	3.4
ckmGAM	1	2	L_{km}	K_{rp}	48.12	7.6	10.5	33.4	3.8	4.7	8	1.8	2.2	3.9
ckmPS	1	1	L_{eq}	K_{rp}	50.95	6.7	12	54	2.9	3.9	7.2	1.5	1.9	3.3
ckmPS	1	2	L_{km}	K_{rp}	51.58	7.5	10.7	37.6	3.5	4.7	8.2	1.8	2.3	3.9
ckmGAM	1	1	L_{eq}	K_{rp}	52.77	6.7	12.6	55.2	2.9	3.9	7.2	1.5	1.9	3.3
ckmPS	1	2	L_{eq}	K_{rp}	86.39	6.7	12.6	84.9	2.9	4.1	9.1	1.6	2	3.9
ckmGAM	1	2	L_{eq}	K_{rp}	106.31	6.7	10.4	112.3	2.9	4.1	9.2	1.6	2	3.9

the GVC_c statistic (see expression (7)) for two different tuning parameter values ($\gamma = 1$ or $\gamma = 1.5$) and using the Kaplan–Meier weights (w_i) or their squares (w_i^2) to take into account the effect of censoring. Thus there are four different expressions of the GCV_c depending on the values chosen for γ and w_i. To choose the number of knots (K), we apply the default formula presented in Ruppert [8] based on Wands default choice (K_{rp}, equation (8)) or the modification propose in Section 3.2 (K_c, equation (9)). Finally, with regard to the location of the knots (L), we propose evenly spaced knots (L_{eq}) or nonuniform knot vectors with the spacing of the knots as a function of the Kaplan–Meier weights (L_{km}), Section 3.2.

As a result of all these possible choices (γ, exponent of w_i, number and location of knots), there are sixteen possible outcomes for each of the two estimators. They are used to calculate the corresponding estimation in each of the thousand dataset replications on each of the nine scenarios (see Table 2) in each example (see Table 1).

Table 5. Results of the simulation study for the logit function.

						(1000 · AMSE) in each scenario								
						$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$	$s = 9$
	GCV_c		Knots			10%	25%	40%	10%	25%	40%	10%	25%	40%
Estimator	γ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmPS	1.5	1	L_{eq}	K_c	0	1.8	2.2	3.3	0.8	1	1.3	0.4	0.5	0.7
ckmGAM	1.5	1	L_{eq}	K_c	0	1.8	2.2	3.3	0.8	1	1.3	0.4	0.5	0.7
ckmPS	1.5	1	L_{eq}	K_{rp}	0.85	1.8	2.2	3.3	0.8	1	1.4	0.4	0.5	0.7
ckmGAM	1.5	1	L_{eq}	K_{rp}	0.85	1.8	2.2	3.3	0.8	1	1.4	0.4	0.5	0.7
ckmPS	1.5	2	L_{eq}	K_c	5.32	1.8	2.2	3.9	0.8	1	1.5	0.4	0.5	0.8
ckmGAM	1.5	2	L_{eq}	K_c	5.32	1.8	2.2	3.9	0.8	1	1.5	0.4	0.5	0.8
ckmPS	1.5	2	L_{eq}	K_{rp}	6.17	1.8	2.2	3.9	0.8	1	1.6	0.4	0.5	0.8
ckmGAM	1.5	2	L_{eq}	K_{rp}	6.17	1.8	2.2	3.9	0.8	1	1.6	0.4	0.5	0.8
ckmPS	1.5	1	L_{km}	K_c	17	2.1	2.5	3.4	1	1.2	1.5	0.5	0.6	0.8
ckmPS	1.5	1	L_{km}	K_{rp}	19.38	2.1	2.5	3.6	1	1.2	1.7	0.5	0.6	0.8
ckmPS	1.5	2	L_{km}	K_c	19.72	2.1	2.5	3.7	1	1.2	1.7	0.5	0.6	0.8
ckmGAM	1.5	1	L_{km}	K_c	22.9	2.2	2.5	3.4	1.1	1.3	1.5	0.6	0.6	0.8
ckmPS	1.5	2	L_{km}	K_{rp}	24.36	2.1	2.5	4.1	1	1.2	1.9	0.5	0.6	0.9
ckmGAM	1.5	2	L_{km}	K_c	25.95	2.2	2.5	3.8	1.1	1.3	1.7	0.6	0.6	0.8
ckmGAM	1.5	1	L_{km}	K_{rp}	26.84	2.2	2.6	3.5	1.2	1.3	1.7	0.6	0.6	0.8
ckmGAM	1.5	2	L_{km}	K_{rp}	31.48	2.2	2.6	3.9	1.2	1.3	1.9	0.6	0.6	0.9
ckmPS	1	1	L_{eq}	K_c	34.05	2	2.9	6.5	0.9	1.2	1.9	0.5	0.6	1
ckmGAM	1	1	L_{eq}	K_c	34.38	2	2.9	6.6	0.9	1.2	1.9	0.5	0.6	1
ckmGAM	1	1	L_{km}	K_c	40.16	2.2	3	5.4	1.1	1.4	2	0.5	0.7	1
ckmPS	1	1	L_{km}	K_c	41.14	2.2	3.1	5.7	1	1.4	2.1	0.5	0.7	1
ckmPS	1	2	L_{eq}	K_c	48.25	2	3	8.6	0.9	1.3	2.1	0.5	0.7	1.1
ckmPS	1	2	L_{km}	K_c	52.11	2.2	3.2	7.5	1	1.5	2.3	0.5	0.7	1.1
ckmGAM	1	2	L_{km}	K_c	52.71	2.2	3.1	8	1.1	1.4	2.2	0.5	0.7	1.1
ckmGAM	1	2	L_{eq}	K_c	56	2	3	10.9	0.9	1.3	2.1	0.5	0.7	1.1
ckmGAM	1	1	L_{km}	K_{rp}	57.89	2.2	3.2	7.9	1.1	1.4	2.6	0.5	0.7	1.2
ckmPS	1	1	L_{km}	K_{rp}	62.8	2.2	3.3	7.9	1.1	1.5	2.8	0.5	0.7	1.3
ckmPS	1	1	L_{eq}	K_{rp}	67.24	2	3	12.5	0.9	1.3	2.6	0.5	0.7	1.2
ckmGAM	1	1	L_{eq}	K_{rp}	69.26	2	3	13.1	0.9	1.3	2.6	0.5	0.7	1.2
ckmGAM	1	2	L_{km}	K_{rp}	74.94	2.2	3.3	9.8	1.1	1.5	3.1	0.5	0.7	1.5
ckmPS	1	2	L_{km}	K_{rp}	78.96	2.2	3.4	9.6	1.1	1.6	3.2	0.5	0.8	1.5
ckmPS	1	2	L_{eq}	K_{rp}	104.66	2	3.2	21.1	0.9	1.3	3.1	0.5	0.7	1.4
ckmGAM	1	2	L_{eq}	K_{rp}	139.67	2	3.2	31.5	0.9	1.3	3.1	0.5	0.7	1.4

As a measure of goodness-of-fit, the estimated mean squared error (MSE) is calculated for each dataset replication ($j = 1, 2, \ldots, 1000$) for each of the nine scenarios ($s = 1, 2, \ldots, 9$) in each example. The MSE is defined as follows:

$$\mathrm{MSE}_{(j,s)} = \frac{\sum_{i=1}^{n} (f(x_i) - \hat{f}(x_i))^2}{n} \quad j = 1, 2, \ldots, 1000 \; s = 1, 2, \ldots, 9$$

To summarize the performance of the estimator in each scenario the average of the mean squared errors (AMSE) is calculated:

$$\mathrm{AMSE}_s = \frac{1}{1000} \sum_{j=1}^{1000} \mathrm{MSE}_{(j,s)} \quad s = 1, 2, \ldots, 9$$

To summarize the performance of the proposed estimators with the different parameter choices (γ, exponent of w_i, number and location of knots), a relative overall measure of

Table 6. Results of the simulation study for the sinusoidal function with two cycles.

						(1000 · AMSE) in each scenario								
						$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$	$s = 9$
	GCV_c		Knots			10%	25%	40%	10%	25%	40%	10%	25%	40%
Estimator	γ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmGAM	1.5	1	L_{eq}	K_c	0	6.1	7.1	10	2.7	3.1	4.3	1.4	1.7	2.3
ckmPS	1.5	1	L_{eq}	K_c	0.18	6.2	7.1	10	2.7	3.1	4.3	1.4	1.7	2.3
ckmPS	1.5	2	L_{eq}	K_c	1.97	6.1	7.2	10.5	2.7	3.1	4.6	1.4	1.7	2.4
ckmGAM	1.5	2	L_{eq}	K_c	2.19	6.1	7.2	10.7	2.7	3.1	4.6	1.4	1.7	2.4
ckmPS	1.5	1	L_{eq}	K_{rp}	2.68	6.2	7.2	10.3	2.7	3.2	4.5	1.4	1.8	2.4
ckmGAM	1.5	1	L_{eq}	K_{rp}	2.68	6.2	7.2	10.3	2.7	3.2	4.5	1.4	1.8	2.4
ckmGAM	1.5	2	L_{eq}	K_{rp}	6.58	6.1	7.2	11.5	2.7	3.2	5	1.4	1.8	2.7
ckmPS	1.5	2	L_{eq}	K_{rp}	6.65	6.2	7.2	11.4	2.7	3.2	5	1.4	1.8	2.7
ckmPS	1.5	1	L_{km}	K_c	6.91	6.9	7.8	10.7	2.8	3.3	4.5	1.5	1.8	2.4
ckmGAM	1.5	1	L_{km}	K_c	7.61	6.7	7.7	10.5	2.8	3.3	4.5	1.6	1.9	2.4
ckmGAM	1.5	2	L_{km}	K_c	8.66	6.7	7.7	10.9	2.8	3.3	4.8	1.6	1.8	2.5
ckmPS	1.5	2	L_{km}	K_c	8.83	6.9	7.8	11.3	2.8	3.3	4.8	1.5	1.8	2.5
ckmGAM	1.5	1	L_{km}	K_{rp}	12.38	6.9	8.2	11.5	2.9	3.4	4.8	1.6	1.9	2.6
ckmPS	1.5	1	L_{km}	K_{rp}	13.38	7	8.2	12	2.9	3.4	4.9	1.6	1.9	2.6
ckmGAM	1.5	2	L_{km}	K_{rp}	15.42	6.9	8.2	12.2	2.9	3.4	5.3	1.6	1.9	2.8
ckmPS	1	1	L_{eq}	K_c	15.79	6.4	8.6	13.9	2.8	3.6	5.2	1.5	1.9	2.7
ckmGAM	1	1	L_{eq}	K_c	16.23	6.4	8.6	14.3	2.8	3.6	5.2	1.5	1.9	2.7
ckmPS	1.5	2	L_{km}	K_{rp}	16.69	7	8.2	13	2.9	3.4	5.5	1.5	1.9	2.9
ckmPS	1	2	L_{eq}	K_c	19.64	6.4	8.7	14.6	2.8	3.6	5.7	1.5	2	2.9
ckmGAM	1	1	L_{km}	K_c	19.8	6.9	9	13.9	2.9	3.7	5.3	1.6	2	2.7
ckmGAM	1	2	L_{eq}	K_c	19.97	6.4	8.7	14.9	2.8	3.6	5.7	1.5	2	2.9
ckmPS	1	1	L_{km}	K_c	21.01	7.1	9.3	14	2.9	3.7	5.4	1.6	2	2.7
ckmGAM	1	2	L_{km}	K_c	23.07	6.9	9.1	14.9	2.9	3.7	5.7	1.6	2	2.9
ckmPS	1	2	L_{km}	K_c	25.38	7.1	9.6	14.8	2.9	3.8	5.8	1.6	2.1	2.9
ckmPS	1	1	L_{eq}	K_{rp}	34.58	6.5	9.3	21.2	2.8	3.8	7	1.5	2	3.4
ckmGAM	1	1	L_{km}	K_{rp}	36.2	7.1	9.8	18.6	3	3.9	7	1.6	2.1	3.4
ckmGAM	1	1	L_{eq}	K_{rp}	38.69	6.5	9.3	24.9	2.8	3.8	7	1.5	2	3.4
ckmPS	1	1	L_{km}	K_{rp}	43.5	7.3	10.2	21.8	3	4	7.3	1.6	2.2	3.6
ckmGAM	1	2	L_{km}	K_{rp}	46	7.1	9.9	22.1	3	4	7.9	1.6	2.2	3.9
ckmPS	1	2	L_{km}	K_{rp}	51.84	7.3	10.5	24.3	3	4.1	8.2	1.6	2.2	4.1
ckmPS	1	2	L_{eq}	K_{rp}	56.38	6.5	9.4	35.5	2.8	3.9	7.9	1.5	2.1	3.9
ckmGAM	1	2	L_{eq}	K_{rp}	67.71	6.5	9.4	45.7	2.8	3.9	7.9	1.5	2.1	3.9

accuracy is proposed that takes into account the joint behaviour of each estimator in the nine scenarios. To that end, the AMSE of the proposed estimator is compared with the optimal AMSE in each scenario ($AMSE_{s-opt}$) and the results of all scenarios are aggregated:

$$G = \sum_{s=1}^{9} \frac{AMSE_s - AMSE_{s\text{-opt}}}{AMSE_{s\text{-opt}}}$$

Tables 3–7 summarize the results for each example by presenting the different estimators ranked from best to worst (from lowest to highest value of G). For example, the first row of Table 3 gives the results for the ckmPS estimator with a GCV_c criterion with a γ value of 1.5 and an exponent of Kaplan–Meier weights of 1 plus a number of knots equidistantly distributed (L_{eq}) adding up to K_c. The last nine columns present the AMSE values (multiplied by 1000) of the estimator in the different scenarios. In eight of the scenarios, the value obtained coincides with the optimum AMSE and in the remaining scenario, scenario five, the value is very close to that minimum value. As an overall measure of the behaviour of

Table 7. Results of the simulation study for the sinusoidal function with three cycles.

						(1000 · AMSE) in each scenario								
						$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$	$s = 9$
	GCV_c		Knots			10%	25%	40%	10%	25%	40%	10%	25%	40%
Estimator	γ	w^{exp}	L	K	G	$n = 200$			$n = 500$			$n = 1000$		
ckmPS	1.5	1	L_{eq}	K_c	0.73	8.6	10.4	17.8	3.8	4.5	6.1	2.1	2.4	3.2
ckmGAM	1.5	1	L_{eq}	K_c	0.86	8.6	10.4	18	3.8	4.5	6.1	2.1	2.4	3.2
ckmPS	1.5	1	L_{eq}	K_{rp}	2.53	8.6	10.5	16.7	3.9	4.6	6.5	2.1	2.5	3.4
ckmGAM	1.5	1	L_{eq}	K_{rp}	2.53	8.6	10.5	16.7	3.9	4.6	6.5	2.1	2.5	3.4
ckmPS	1.5	2	L_{eq}	K_c	4.99	8.6	11.2	21.2	3.8	4.6	6.4	2.1	2.4	3.3
ckmGAM	1.5	2	L_{eq}	K_c	5.06	8.6	11.2	21.3	3.8	4.6	6.4	2.1	2.4	3.3
ckmPS	1	1	L_{eq}	K_c	11.39	8.8	12.1	25.4	3.9	4.9	6.7	2.1	2.5	3.4
ckmPS	1.5	1	L_{km}	K_c	11.56	9.9	12	17.8	4.4	5.2	6.8	2.3	2.6	3.4
ckmGAM	1	1	L_{eq}	K_c	11.79	8.8	12.1	26	3.9	4.9	6.7	2.1	2.5	3.4
ckmPS	1.5	2	L_{eq}	K_{rp}	12.55	8.6	11.2	26.8	3.8	4.6	7.3	2.1	2.5	3.8
ckmPS	1.5	2	L_{km}	K_c	14.88	9.8	12.9	20.2	4.4	5.2	7.1	2.3	2.6	3.5
ckmGAM	1.5	2	L_{eq}	K_{rp}	14.88	8.6	11.2	30.3	3.8	4.6	7.3	2.1	2.5	3.8
ckmGAM	1	1	L_{km}	K_c	15.77	9.9	12.8	19.9	4.5	5.4	7	2.3	2.7	3.5
ckmGAM	1.5	1	L_{km}	K_c	16.72	10.6	12.5	17.9	4.9	5.4	6.9	2.5	2.7	3.4
ckmPS	1.5	1	L_{km}	K_{rp}	17.79	10	12.5	19.6	4.5	5.4	7.5	2.3	2.8	3.8
ckmGAM	1.5	2	L_{km}	K_c	18.11	10.5	13	19.3	4.8	5.4	6.9	2.5	2.7	3.5
ckmPS	1	1	L_{km}	K_c	18.37	9.9	13.3	22	4.4	5.5	7.2	2.3	2.7	3.6
ckmPS	1	2	L_{eq}	K_c	19.93	8.8	14	31.7	3.9	5	7.2	2.1	2.6	3.6
ckmGAM	1	2	L_{km}	K_c	20.78	9.9	13.6	24	4.5	5.4	7.4	2.3	2.7	3.7
ckmGAM	1	2	L_{eq}	K_c	21.04	8.8	15.6	30.8	3.9	5	7.2	2.1	2.6	3.6
ckmPS	1.5	2	L_{km}	K_{rp}	22.67	10	13.2	22.6	4.5	5.4	8.1	2.3	2.8	4.1
ckmPS	1	2	L_{km}	K_c	23.17	9.9	14.3	25.3	4.4	5.5	7.6	2.3	2.8	3.7
ckmGAM	1.5	1	L_{km}	K_{rp}	23.96	11	13.3	19.4	5	5.8	7.6	2.5	2.9	3.8
ckmGAM	1.5	2	L_{km}	K_{rp}	27.01	10.9	13.8	21.6	5	5.7	8	2.5	2.9	4
ckmGAM	1	1	L_{km}	K_{rp}	32.12	10.2	13.8	28.7	4.6	5.9	8.9	2.3	2.9	4.4
ckmPS	1	1	L_{km}	K_{rp}	34.16	10.1	14.4	29.4	4.5	5.9	9.2	2.3	3	4.5
ckmPS	1	1	L_{eq}	K_{rp}	38.78	8.9	13	51.6	3.9	5.3	8.8	2.1	2.7	4.3
ckmGAM	1	2	L_{km}	K_{rp}	42.38	10.2	14.8	37.6	4.6	6	9.8	2.3	2.9	4.8
ckmPS	1	2	L_{km}	K_{rp}	43.64	10.1	15.6	36.7	4.5	6.1	10	2.3	3	4.9
ckmGAM	1	1	L_{eq}	K_{rp}	49.05	8.9	12.9	67.2	3.9	5.3	8.8	2.1	2.7	4.3
ckmPS	1	2	L_{eq}	K_{rp}	90.14	8.9	14.1	119.7	3.9	5.5	10	2.1	2.8	4.8
ckmGAM	1	2	L_{eq}	K_{rp}	119.09	8.9	14.1	163.2	3.9	5.5	10	2.1	2.8	4.8

the estimator in the nine scenarios a value of G of 0.25 is obtained, very close to the $G = 0$ that would be obtained if it were the best estimator in all the scenarios.

Our analysis of the results shows that for the two estimators analysed (ckmPS and ckmGAM estimators), the best choice of parameters is given by a GCV_c with a value of γ of 1.5 and an exponent of the Kaplan–Meier weights of 1 plus a number of knots equidistantly distributed adding up to K_c. The results for the ckmPS estimator with this choice of parameters are summarized in Figures 4–8 (the ckmGAM estimator obtains similar results, not shown). Subfigure (a) in Figures 4–8 shows the mean, the pointwise 95% upper and lower oscillation limits of the values estimated and the true function $f(x)$ for a level of censoring of 25% and a sample size of 500 observations (scenario 5) for each example.

It is clear that the proposed approach adequately estimates the relationship in all cases and the associated pointwise 95% limits achieve a coverage which is close to the true function. This result is maintained in the rest of the scenarios (Figures 9–13). From these results, it can be concluded that the performance of the method proposed is good in all the cases

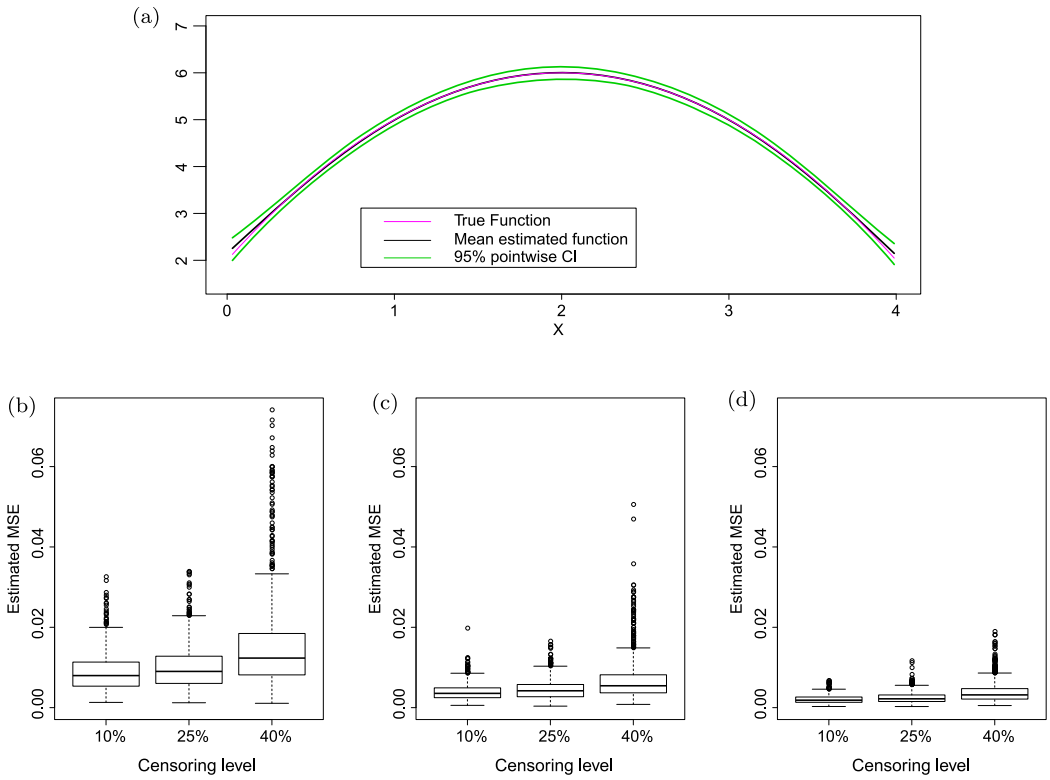


Figure 4. Results of simulation study for the ckmPS estimator for the quadratic function. (a) Scenario 5 ($n = 500$ & $C = 25$): GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots, (b) $n = 200$, (c) $n = 500$, and (d) $n = 1000$.

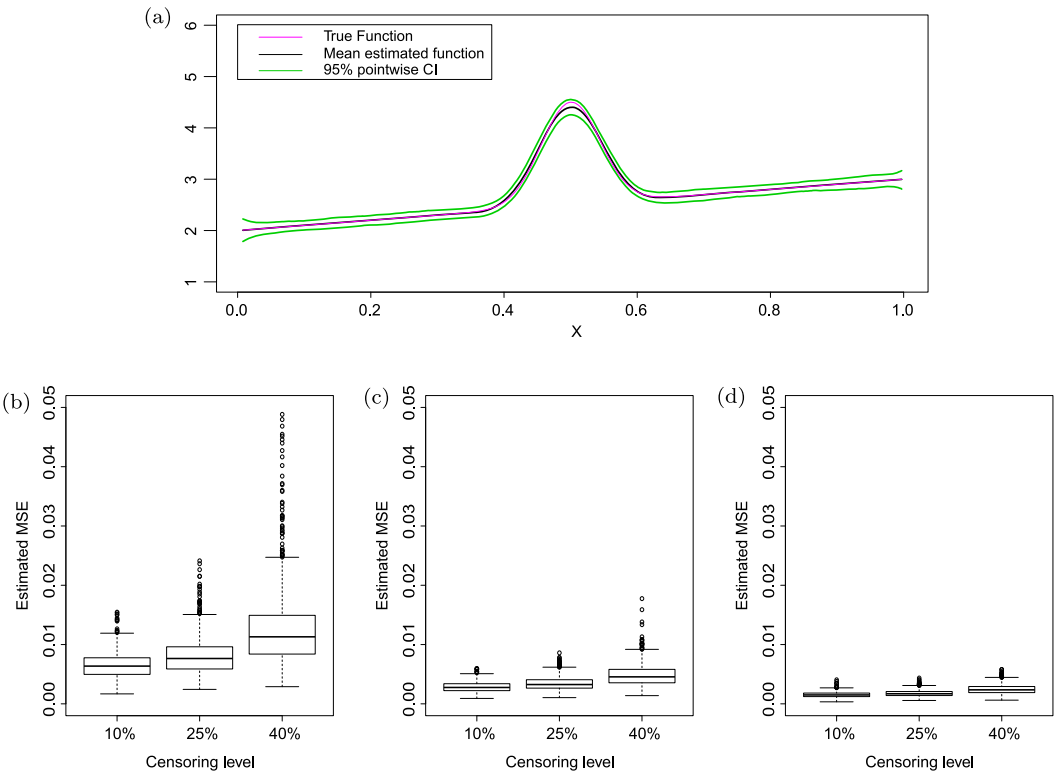


Figure 5. Results of simulation study for the ckmPS estimator for the bump function. (a) Scenario 5 ($n = 500$ & $C = 25$): GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots, (b) $n = 200$, (c) $n = 500$, and (d) $n = 1000$.

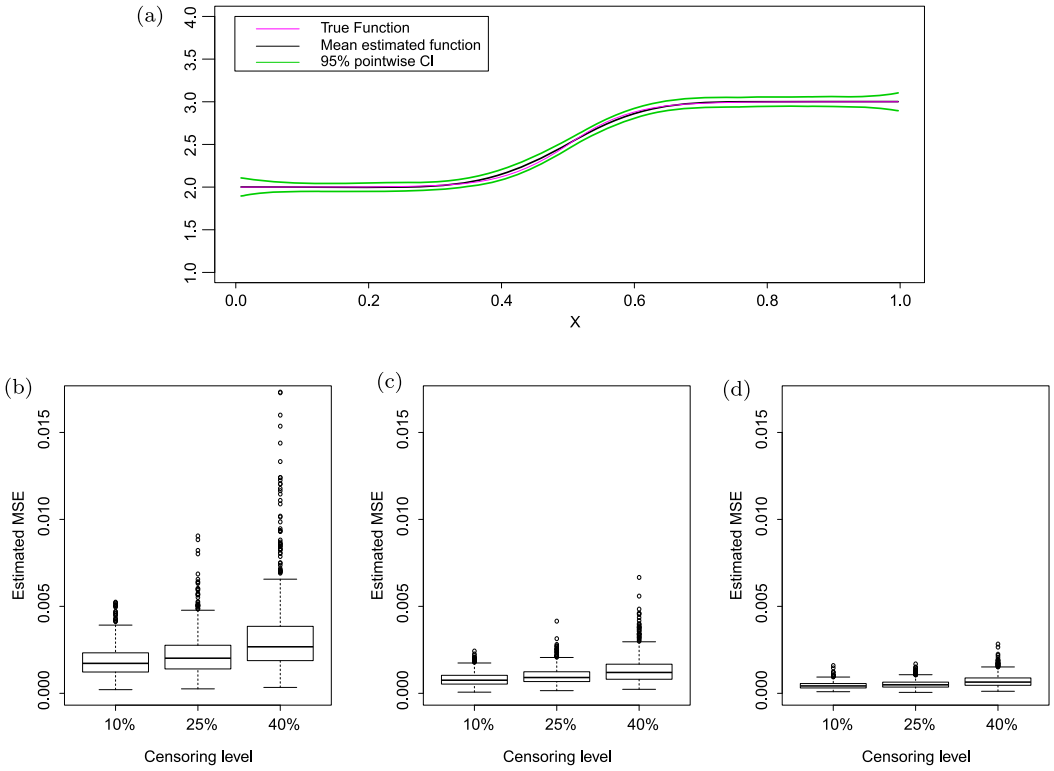


Figure 6. Results of simulation study for the ckmPS estimator for the logit function. (a) Scenario 5 ($n = 500$ & $C = 25$): GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots, (b) $n = 200$, (c) $n = 500$, and (d) $n = 1000$.

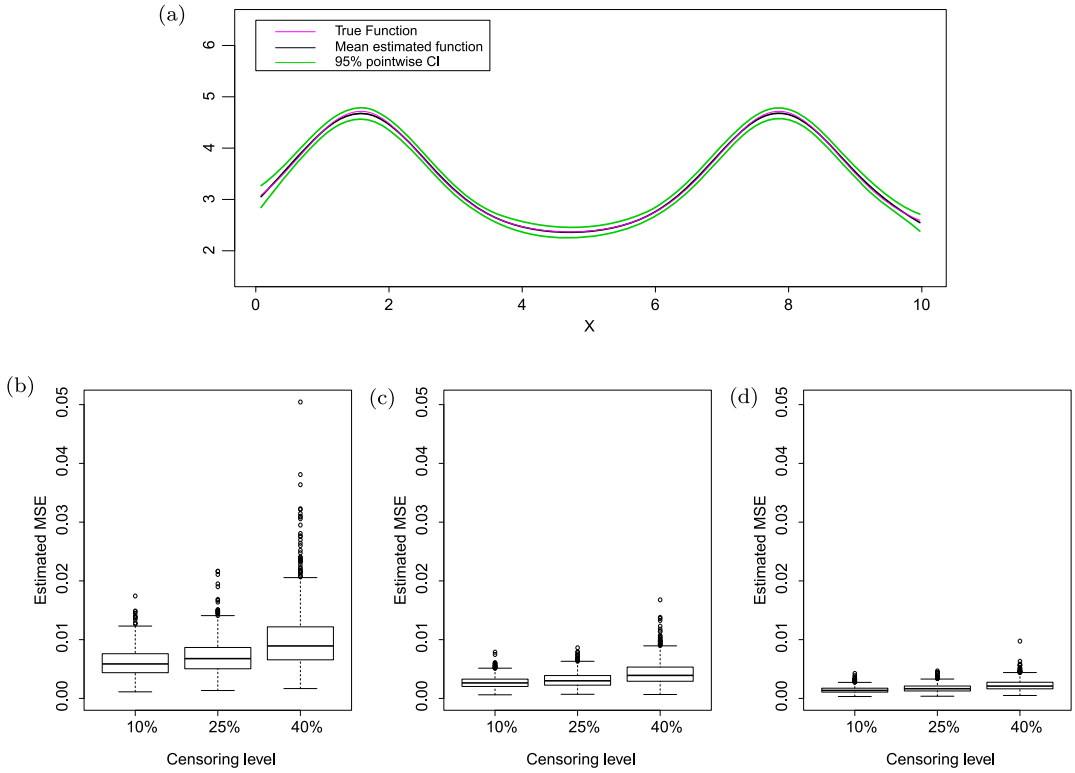


Figure 7. Results of simulation study for the ckmPS estimator for the sinusoidal function with two cycles. (a) Scenario 5 ($n = 500$ & $C = 25$): GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots, (b) $n = 200$, (c) $n = 500$, and (d) $n = 1000$.

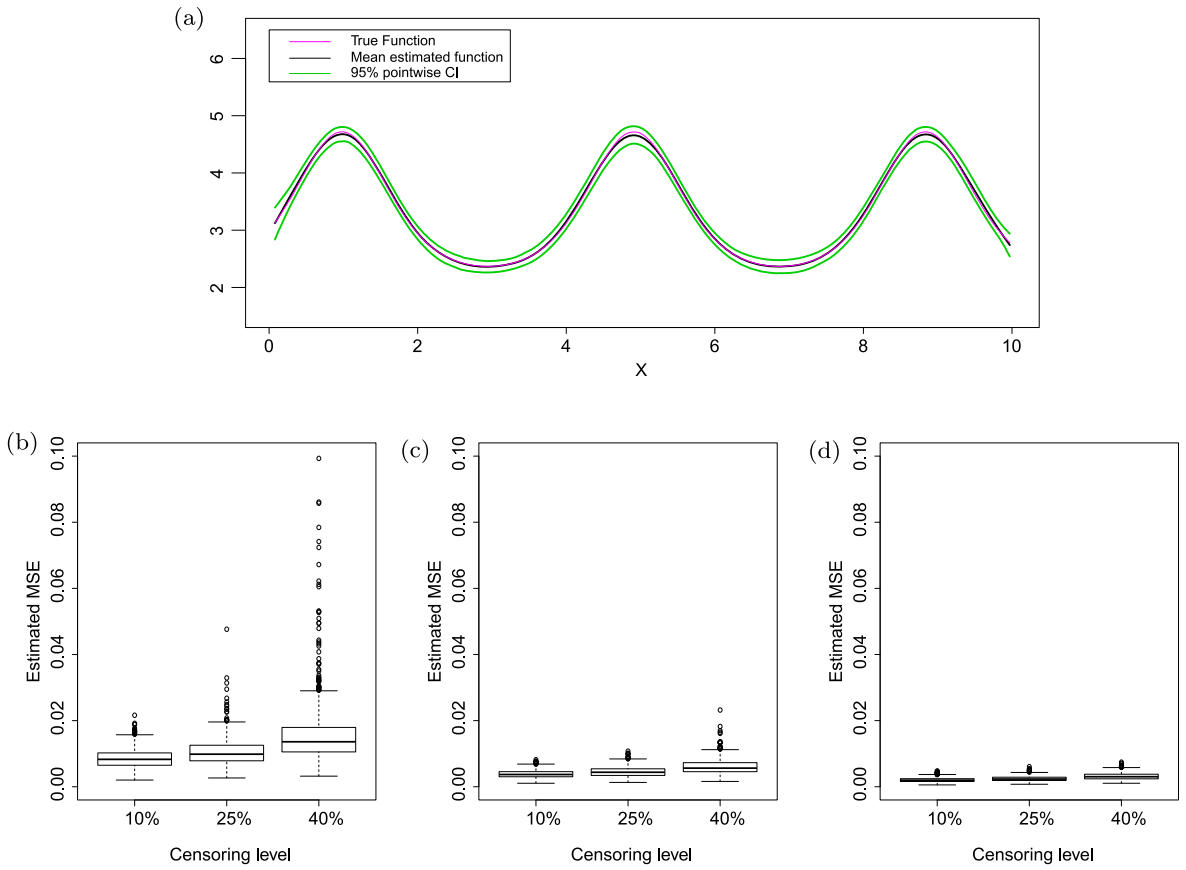


Figure 8. Results of simulation study for the ckmPS estimator for the sinusoidal function with three cycles. (a) Scenario 5 ($n = 500$ & $C = 25$): GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots, (b) $n = 200$, (c) $n = 500$, and (d) $n = 1000$.

considered: the estimated function recovers the true functional form. Subfigures (b)–(d) in Figures 4–8 gives the box plots with the results of the estimated mean squared error (MSE) for each scenario in each of the five examples. The estimated mean squared error decreases when the sample size increases for each censoring level considered. The effect of the censoring level is as expected: the results are more accurate with lower levels of censoring and the variability increases with the censoring level.

In the 45 scenarios analysed (nine scenarios for five examples) there are only two where the above combination ($\gamma = 1.5$, w_i, L_{eq}, K_c) is not the best. The first exception is for the bump function in the scenario with few data ($n = 200$) and a low censoring level ($C = 10\%$), where a GCV_c with an exponent of two in the Kaplan–Meier weights presents slightly better results. In some cases, when the percentage of censoring is small and the sample is not very large, the censored observations accumulate in areas where the duration is high, resulting in a slight downward bias in those areas. An exponent of two for Kaplan–Meier weights in the GCV_c can help reduce the bias in these function peaks. The second exception is for the sinusoidal function with three cycles in the scenario with few data ($n = 200$) and a high censoring level ($C = 40\%$), where the number of knots proposed by Ruppert (K_{rp}) performs a little better than the K_c proposal. Given that the profile of the function is quite complex, in a scenario with little information (few data and high censoring) a larger number of knots is better suited to the situation.

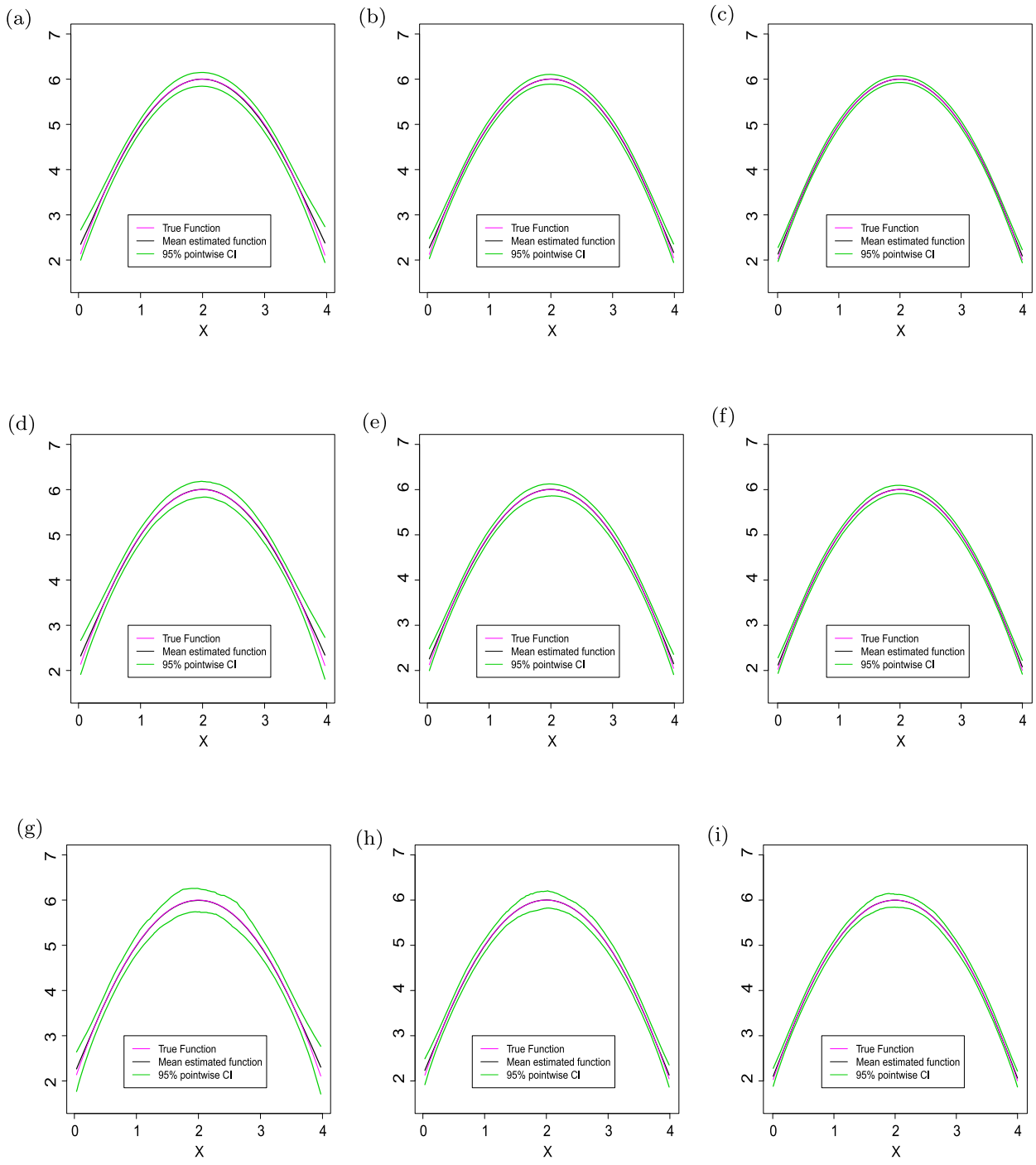


Figure 9. Estimated function using ckmPS estimator for the bump function: GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots. (a) $C = 10$ and $n = 200$, (b) $C = 10$ and $n = 500$,(c) $C = 10$ and $n = 1000$, (d) $C = 25$ and $n = 200$, (e) $C = 25$ and $n = 500$, (f) $C = 25$ and $n = 1000$, (g) $C = 40$ and $n = 200$, (h) $C = 40$ and $n = 500$, and (i) $C = 40$ and $n = 1000$.

To analyse the importance of each of the choices, Figures 14–18 compare the AMSE of each scenario (same information as in Tables 3–7) for each of the five examples as they use the estimator ckmPS versus the ckmGAM (subfigure (a)), a value of γ of 1 versus 1.5 (subfigure (b)), a weighting of the numerator of the criterion GGV_c with the Kaplan–Meier weights (w_i) versus theirs squares (subfigure (c)), a number of knots K_c versus K_{rp} (subfigure (d)) and, finally, in subfigure (e), equidistant knots (L_{eq}) versus a nonuniform knots

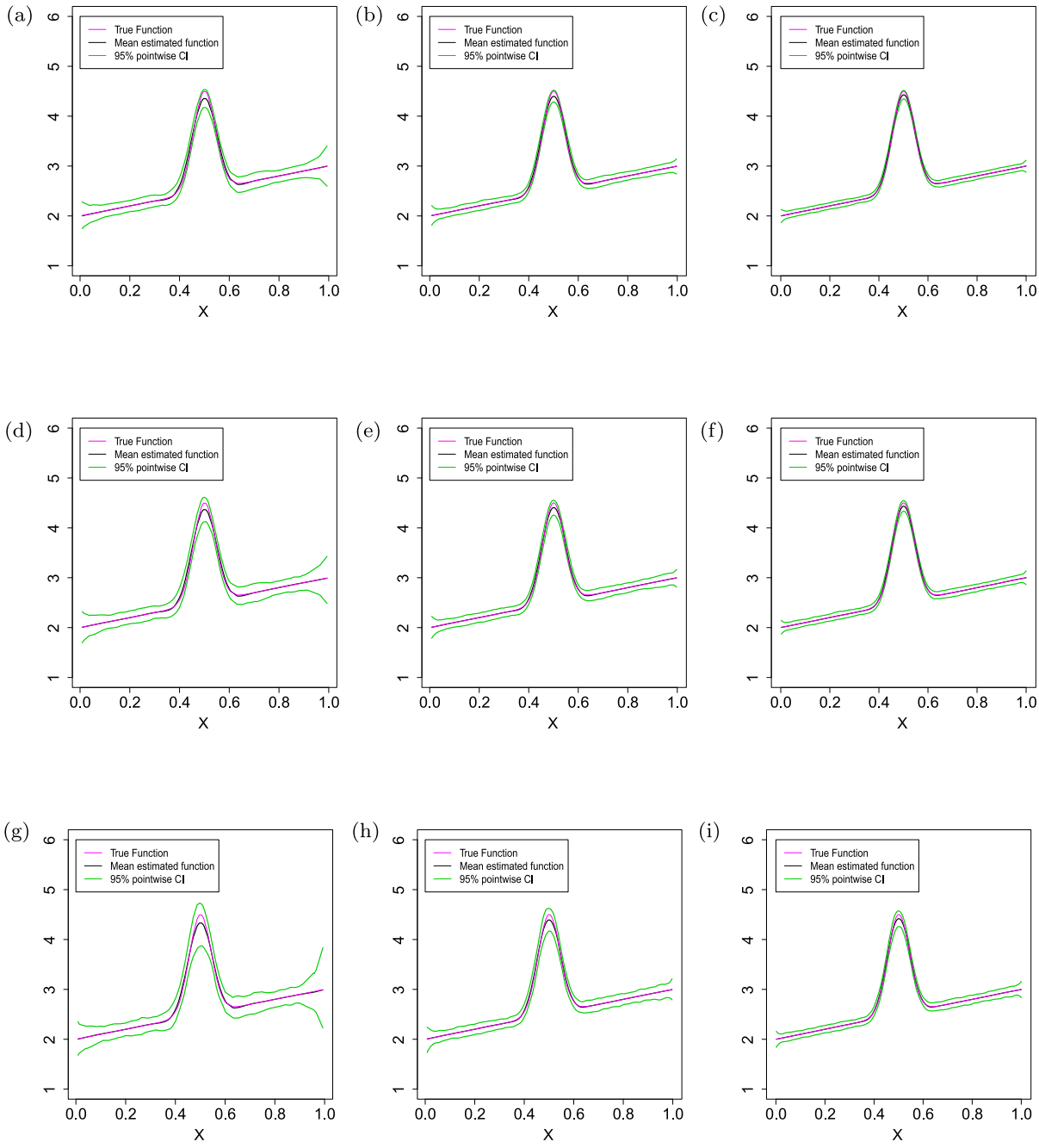


Figure 10. Estimated function using ckmPS estimator for the logit function: GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots. (a) $C = 10$ and $n = 200$, (b) $C = 10$ and $n = 500$, (c) $C = 10$ and $n = 1000$, (d) $C = 25$ and $n = 200$, (e) $C = 25$ and $n = 500$, (f) $C = 25$ and $n = 1000$, (g) $C = 40$ and $n = 200$, (h) $C = 40$ and $n = 500$, and (i) $C = 40$ and $n = 1000$.

vector with the spacing of the knots as a function of the Kaplan–Meier weights (L_{km}). It can be observed that not all choices have the same relevance for obtaining a good estimate. The selection of the estimation method does not seem too important: the two methods analysed perform very similarly and give very similar results (see subfigures (a) in Figures 14–18). On the other hand, the choice of the γ parameter is of great importance, and its importance grows with censoring. Analogous to what has been found in the literature for the uncensored case, the choice of $\gamma = 1.5$ is better in almost all situations than $\gamma = 1.0$, with the difference increasing as the censoring increases (see subfigures (b) in Figures 14–18).

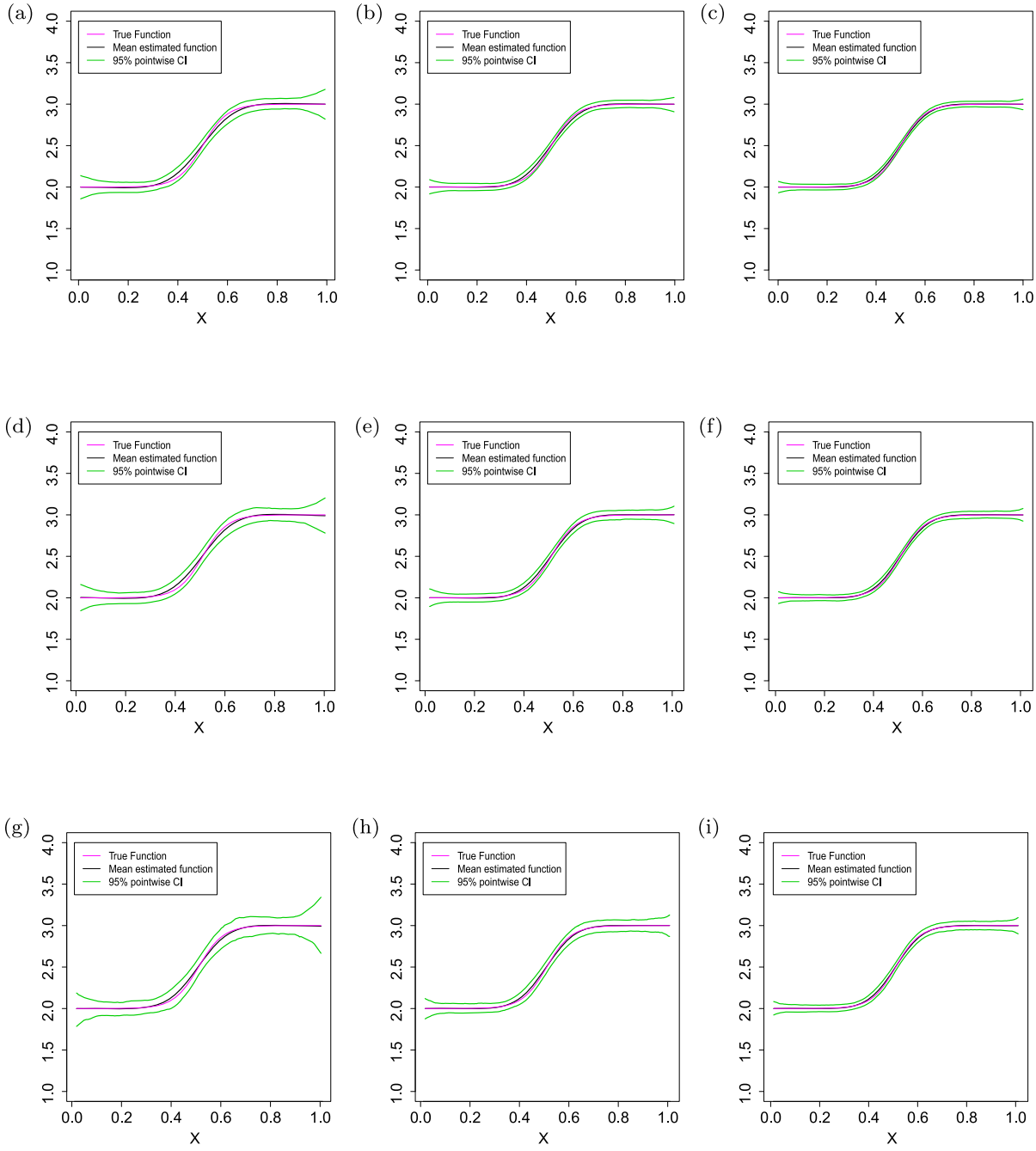


Figure 11. Estimated function using ckmPS estimator for the logit function: GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots. (a) $C = 10$ and $n = 200$, (*b*) C = 10 *and* n = 500, (c) $C = 10$ and $n = 1000$, (*d*) C = 25 *and* n = 200, (e) $C = 25$ and $n = 500$, (*f*) C = 25 *and* n = 1000, (g) $C = 40$ and $n = 200$, (*h*) C = 40 *and* n = 500, and (i) $C = 40$ and $n = 1000$.

As for the choice of the exponent in the Kaplan–Meier weights of the GCV_c if the censoring is not very large both exponents work similarly (see sub-figures (c) in Figures 14–18), but when the censoring is large ($C = 40\%$) exponent 1 (w_i^1) is clearly better than 2 (w_i^2). The situation is similar with respect to the number of knots: it is not very important if the censoring is small, but as the censoring increases the choice of K_c is clearly better than K_{rp}, which is Ruppert's proposal (see subfigures (c) in Figures 14–18). Finally, with regard to knot location, equidistant knots (L_{eq}) clearly perform better, except in samples of small size

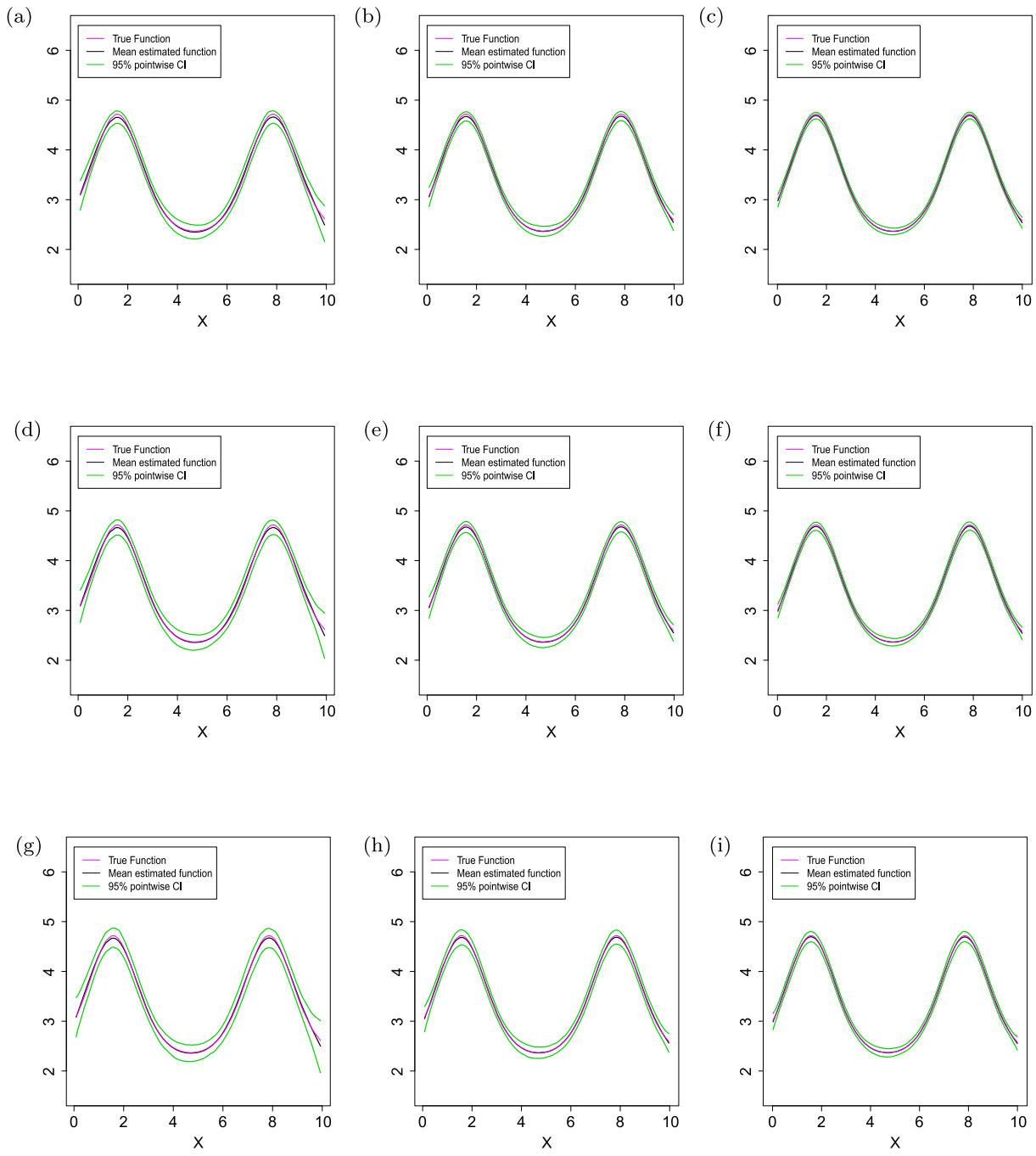


Figure 12. Estimated function using ckmPS estimator for the sinusoidal function with three cycles: GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots. (a) $C = 10$ and $n = 200$, (b) $C = 10$ and $n = 500$, (c) $C = 10$ and $n = 1000$, (d) $C = 25$ and $n = 200$, (e) $C = 25$ and $n = 500$, (f) $C = 25$ and $n = 1000$, (g) $C = 40$ and $n = 200$, (h) $C = 40$ and $n = 500$, and (i) $C = 40$ and $n = 1000$.

($n = 200$) with large censoring ($C = 40\%$) where the spacing of the knots as a function of the Kaplan–Meier weights (L_{km}) sometimes reduces the largest MSEs.

As a conclusion, we find that in the five examples studied, the choice of parameters that generates the best estimate is that of $\gamma = 1.5$, exponent 1 in the Kaplan–Meier weights of the GCV_c and K_c equidistant knots. As the level of censoring increases, even with large sample sizes, major differences also begin to be found between the results of the different specifications. Therefore, if the censoring is large it is important to make good choices.

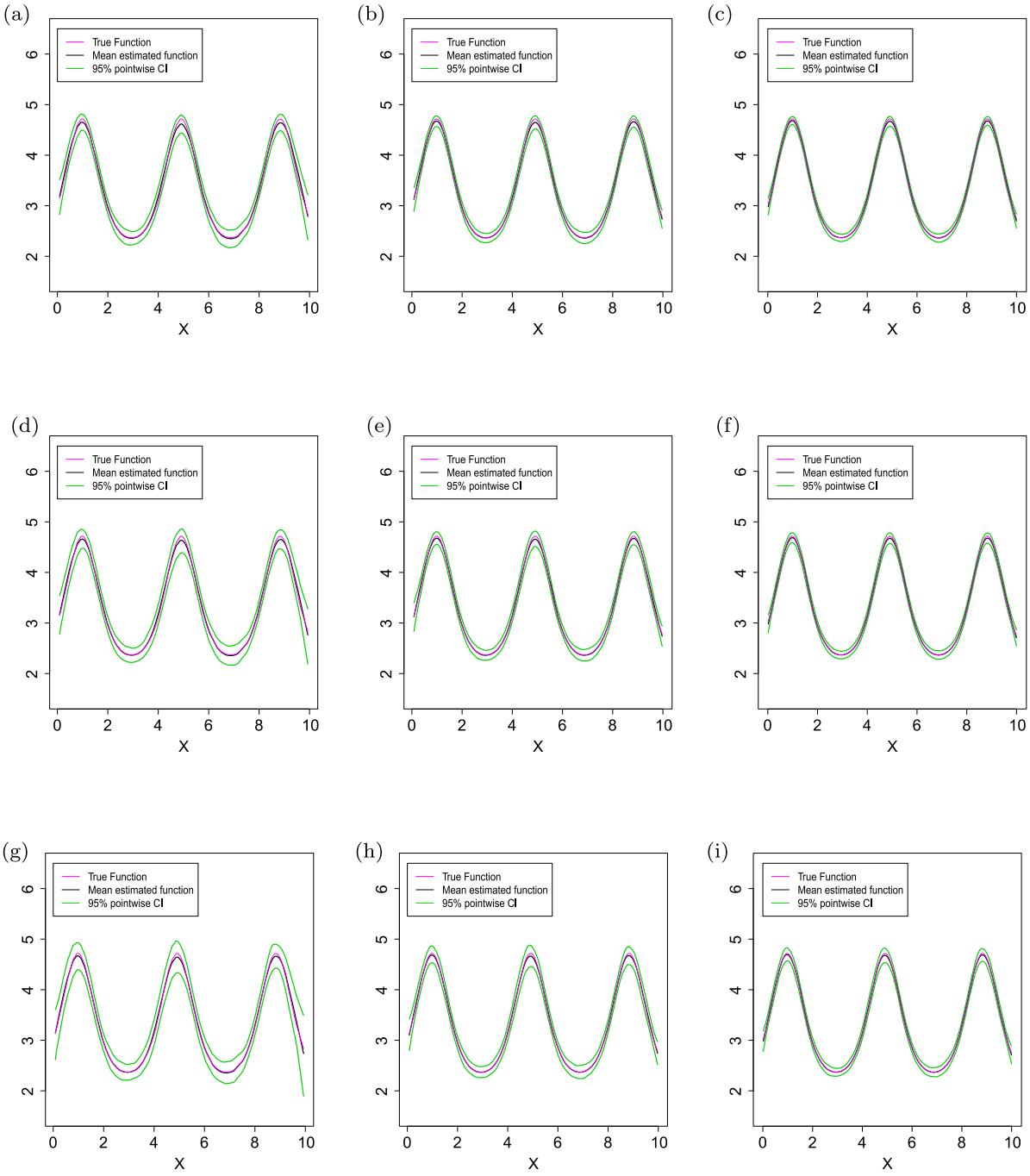


Figure 13. Estimated function using ckmPS estimator for the sinusoidal function with three cycles: GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots. (a) $C = 10$ and $n = 200$, (b) $C = 10$ and $n = 500$, (c) $C = 10$ and $n = 1000$, (d) $C = 25$ and $n = 200$, (e) $C = 25$ and $n = 500$, (f) $C = 25$ and $n = 1000$, (g) $C = 40$ and $n = 200$, (h) $C = 40$ and $n = 500$, and (i) $C = 40$ and $n = 1000$.

With respect to the sample size, the greatest differences occur with $n = 200$. It is interesting to note that the most important choices are $\gamma = 1.5$ and equidistant knots, because they give rise to estimators with a very good performance regardless of the choice of other parameters, which is not so decisive. On the other hand, the worst combinations of parameters arise in general for values of $\gamma = 1.0$ and the number of knots using Ruppert's default, usually proposed in the literature for uncensored data, independently of the choice of the rest of the parameters.

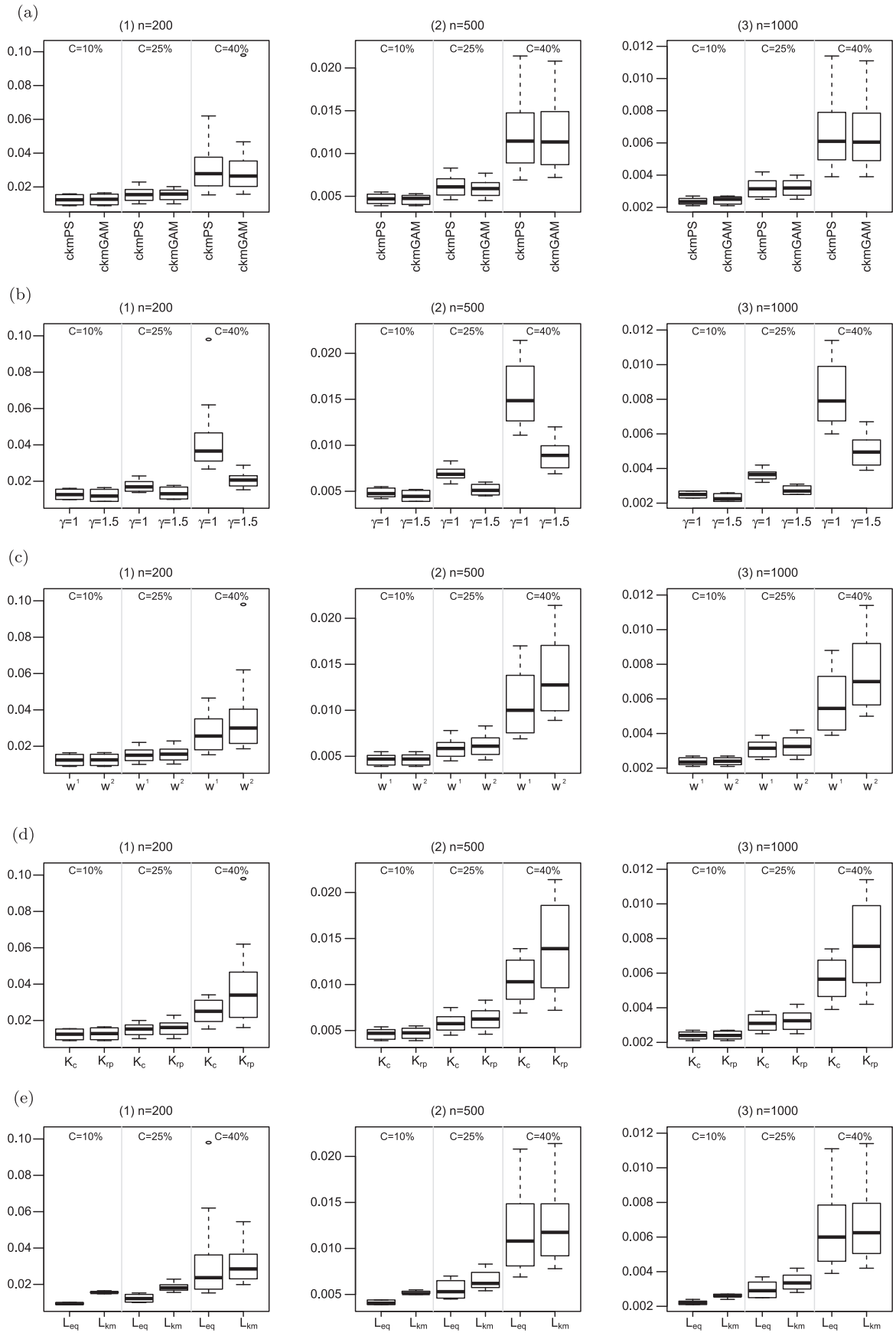


Figure 14. AMSE as a function of the different options for the quadratic function. (a) ckmPS vs ckmGAM, (b) $\gamma = 1$ vs $\gamma = 1.5$, (c) w^1 vs w^2, (d) K_c vs K_{rp}, and (e) L_{eq} vs L_{km}.

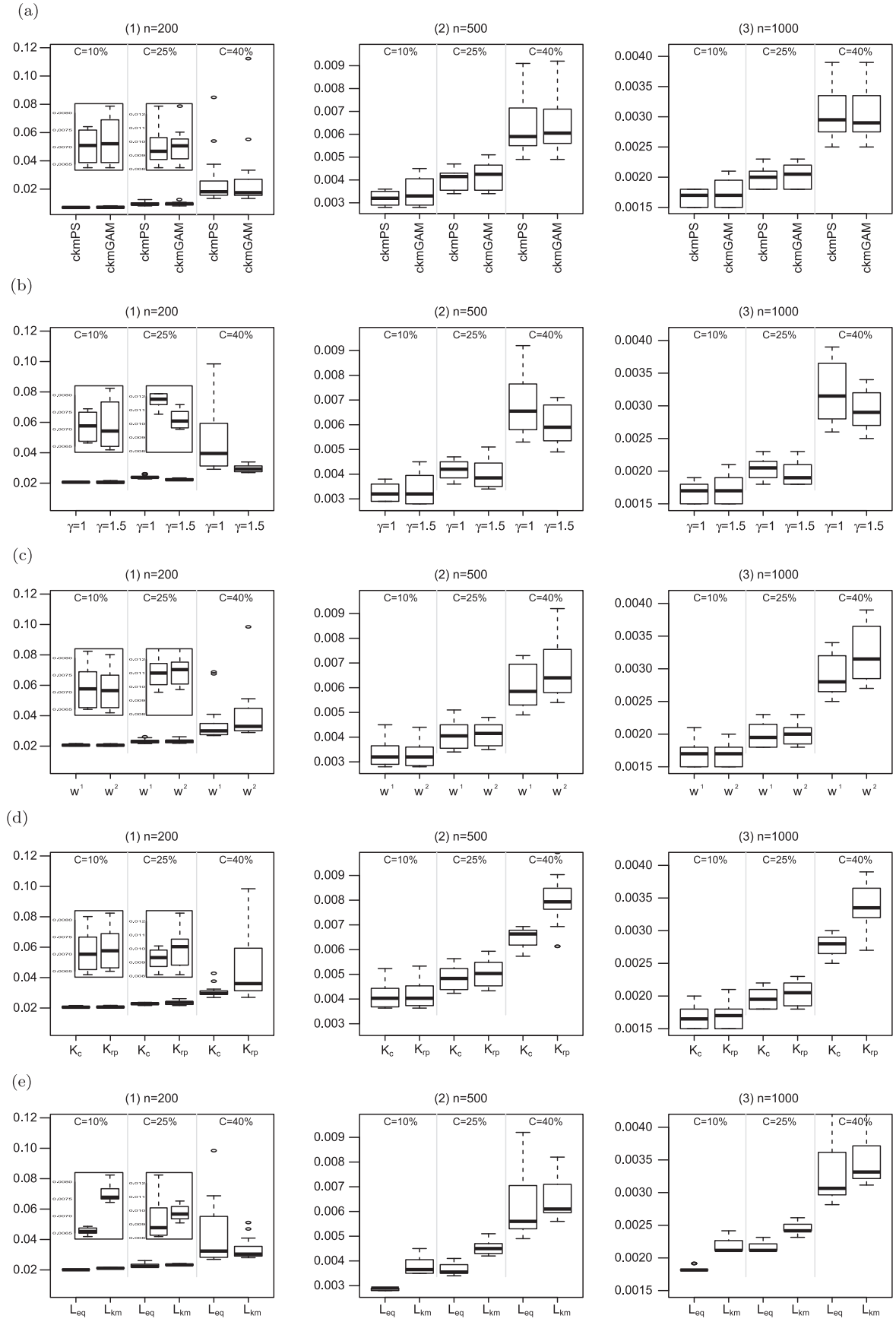


Figure 15. AMSE as a function of the different options for the bump function. (a) ckmPS vs ckmGAM, (b) $\gamma = 1$ vs $\gamma = 1.5$, (c) w^1 vs w^2, (d) K_c vs K_{rp}, and (e) L_{eq} vs L_{km}.

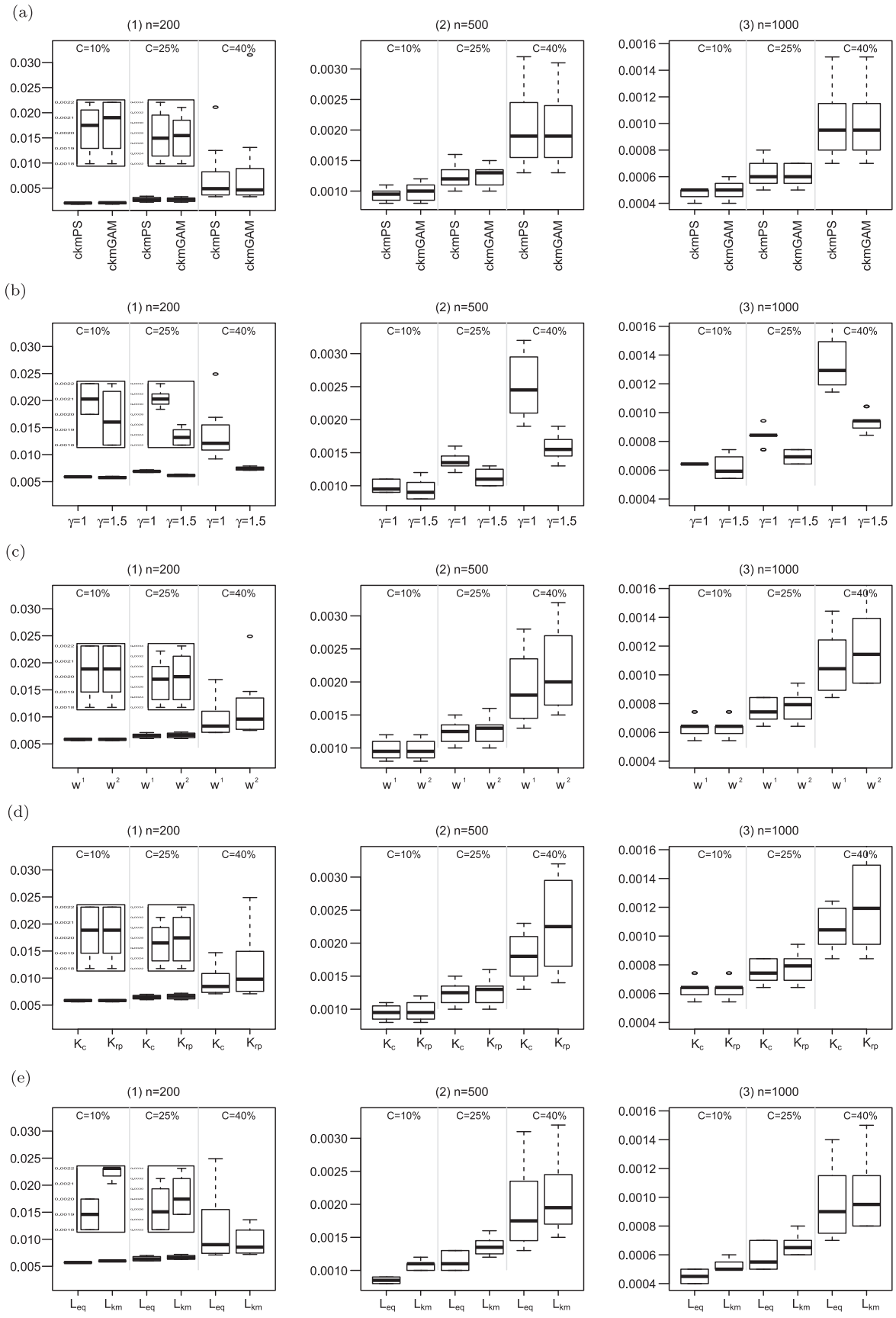


Figure 16. AMSE as a function of the different options for the logit function. (a) ckmPS vs ckmGAM, (b) $\gamma = 1$ vs $\gamma = 1.5$, (c) w^1 vs w^2, (d) K_c vs K_{rp}, and (e) L_{eq} vs L_{km}.

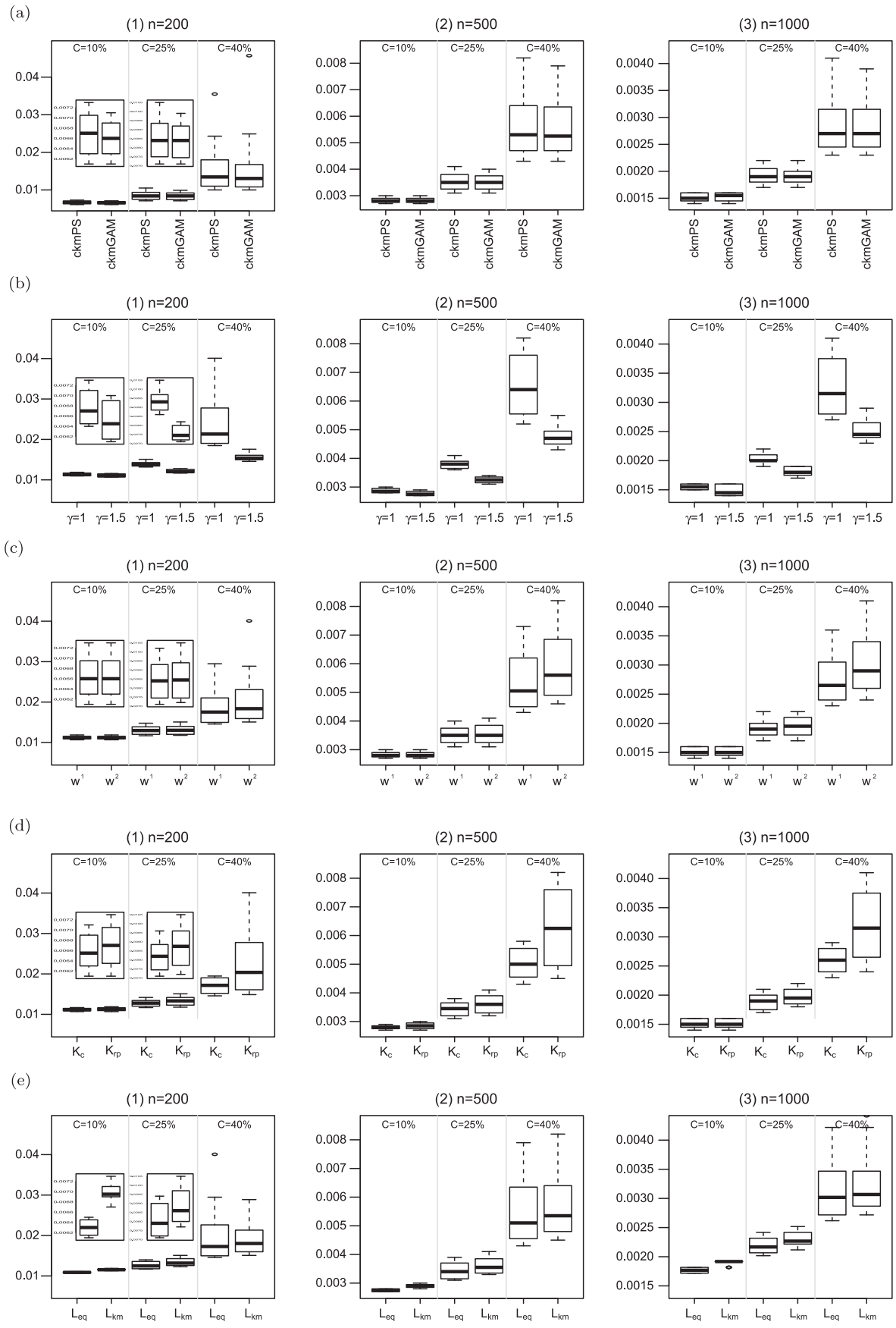


Figure 17. AMSE as a function of the different options for the sinusoidal function with two cycles. (a) ckmPS vs ckmGAM, (b) $\gamma = 1$ vs $\gamma = 1.5$, (c) w^1 vs w^2, (d) K_c vs K_{rp}, and (d) L_{eq} vs L_{km}.

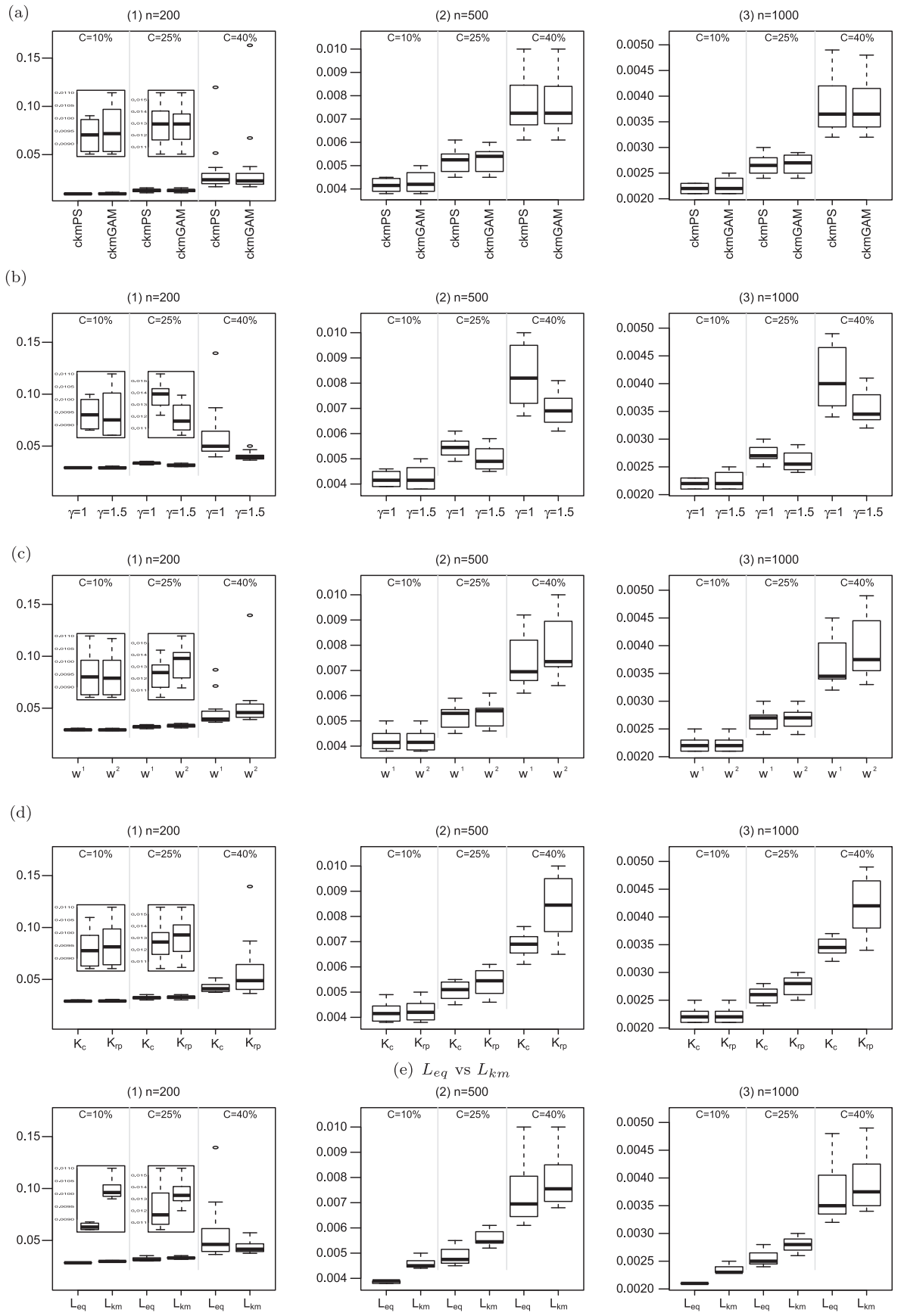


Figure 18. AMSE as a function of the different options for the sinusoidal function with three cycles. (a) ckmPS vs ckmGAM, (b) $\gamma = 1$ vs $\gamma = 1.5$, (c) w^1 vs w^2, (d) K_c vs K_{rp}, and (d) L_{eq} vs L_{km}.

These results are robust to differences in the variability or distribution of the error term. We conduct additional simulations with a greater variance of the random disturbance and consider nonnormal, asymmetric error distributions such as the Weibull distribution. The new results obtained (not shown) confirm the good performance of the proposed framework and are consistent with those presented in this section. If the variance is increased (by lowering the signal-to-noise ratio by 50%) the previous results are maintained. The most remarkable thing is that this increases the importance of the choice of parameters, as the differences between the performance of the estimators increase, even with small censoring. In addition, the number of combinations that give rise to a good estimator decreases, so the right choice of parameters becomes even more important. If the distribution of the error is changed from Normal to Weibull the general results are maintained, but exponent two for Kaplan-Meier weights in the GCV_c sometimes appears among the best options, especially with low censoring percentages.

5. Empirical application: Mayo Clinic primary biliary cirrhosis data

The Mayo Clinic Primary Biliary Cirrhosis dataset contains information from 418 Mayo Clinic patients with primary biliary cholangitis (PBC), previously called primary biliary cirrhosis, an autoimmune disease of the liver. The first 312 cases in the dataset participated in a Mayo Clinic trial in PBC conducted between 1974 and 1984. The additional cases are from an independent set of 106 Mayo Clinic primary biliary cholangitis patients who were eligible for the trial but declined to participate. This dataset has been previously used, for example, in Dickson et al. [25], Therneau and Grambsch [26] and Fleming and Harrington [27], in censored regression models. The dataset provides information about the observed survival time from the date of registration in the trial, albumin values (a low albumin level, a protein made by the liver, can be a sign of advanced liver disease) amid other explanatory variables, and an indicator of patient status (dead or alive) in July 1986. The dataset can be downloaded from the R package *survival* [28,29].

The studies by Therneau and Grambsch [26] and Fleming and Harrington [27] deal with the relationship between the albumin covariate and the survival response variable. They conclude that the relationship between patient survival (T) and albumin is likely to be nonlinear. One option would be a quadratic relationship between the logarithm of survival and albumin:

$$\log(T) = \beta_1 + \beta_2 \,\text{Albumin} + \beta_3 \,\text{Albumin}^2 + \epsilon \tag{10}$$

Assuming that the above parametric specification is correct, two methodologies known and proposed in the literature on survival analysis can be used to fit the model (10). These estimators can be used as a benchmark to evaluate the performance of the censored P-spline method proposed. The first and more restrictive approach is the parametric Accelerated Failure Time (AFT) methodology [30], based on the restricted assumption of knowing the probability distribution of the response variable, that estimates the β coefficients of the model using the maximum likelihood estimator. Thus, considering an AFT lognormal model, we estimate the β coefficients assuming a normal probability distribution. This can be considered as a censored parametric method of estimation. The second methodology, proposed by Stute [22], is less restrictive in that it does not need the assumption of the

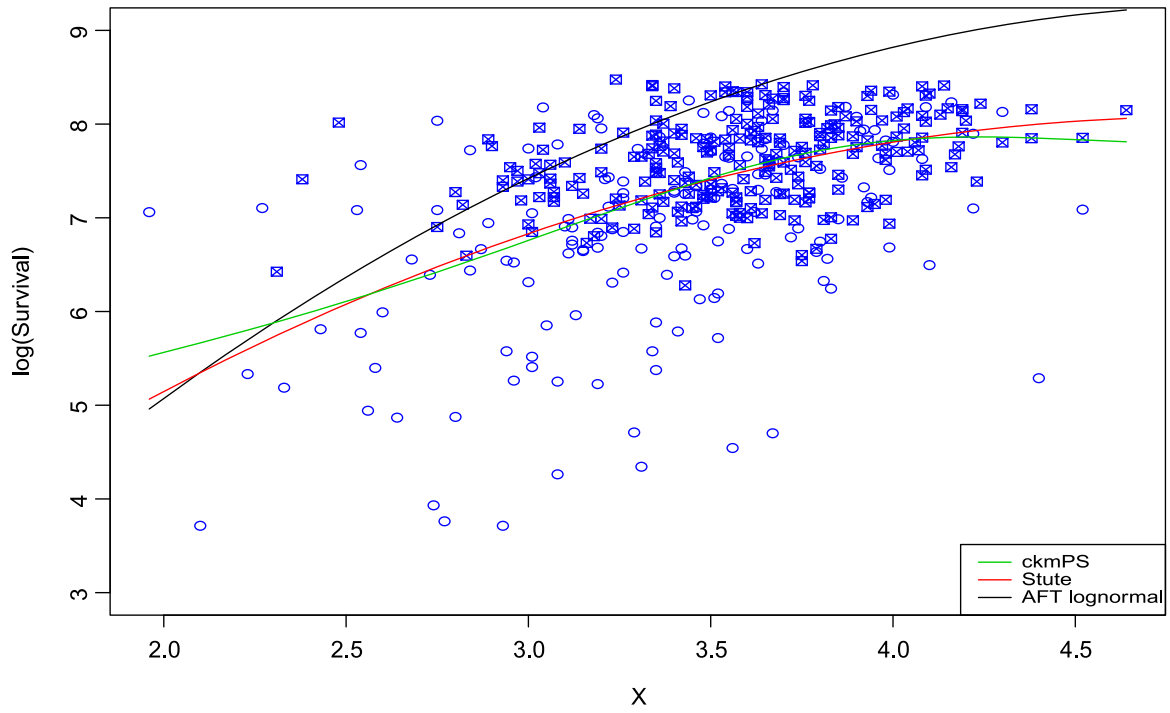


Figure 19. Estimated function using ckmPS estimator for the quadratic function: GVC_c with $\gamma = 1.5$ and w_i^1 & K_c equally spaced knots.

probability distribution of the response variable, but it also trusts the functional form presented in equation (10). That is, it needs to know the form of the relationship between the response variable and the covariate. This can be classed as a censored semiparametric method of estimation. The latter methodology estimates coefficients using weighted least squares via Kaplan–Meier weights [22].

The validity for the estimation results for these two approaches is based on the confidence of the specified relationship (10). If this relationship is not correct both approaches lead to wrong conclusions. As a robust solution to avoid this possibility of incorrect specification, we use our proposal of a censored P-spline method to estimate the relationship between survival and the albumin covariate. Thus, the fitted curve is obtained from minimizing Equation (2) with cubic B-splines and a penalty term of order two. This approach is more flexible than those mentioned above, as it does not assume any functional form for the true relationship. It can be considered as a censored nonparametric method of estimation. We estimate the relationship with the choice of parameters that generates the best estimate in the simulations, i.e. that of $\gamma = 1.5$, exponent 1 in the Kaplan-Meier weights of the GCV_c and K_c equidistant knots.

Figure 19 shows the estimations of these three approaches with the scatterplot of observed *log* survival time versus albumin. Patients indicated by ◯ are dead and those indicated by ☒ are alive in July 1986; that is, the dead patients have uncensored survival times and the live patients have censored survival times.

In conclusion, the AFT methodology and Stute's proposals can be applied only when the functional form of the effect of the covariate X on the response variable is known exactly. In this application, it seems that the relationship between log survival and albumin

is quadratic, so both these methodologies perform reasonably well. However, the non-parametric censored P-spline approach adequately estimates the quadratic relationship, obtaining very similar results to the previous ones. Nevertheless, it must be mentioned that if the functional form or the probability distribution are wrongly chosen, these two methods lead to a serious problem of incorrect specification of the model and therefore to incorrect conclusions. An important advantage of our approach is that it does not need to assume any functional form and therefore it avoids this problem. We also apply the ckmPS estimator with different parameter choices of γ, exponent of w_i and number and location of knots (not shown). There are no major differences in the estimation when varying the number and location of knots, but a value of γ equal to one and/or an exponent equal to two in the Kaplan-Meier weights of the GCV_c produces estimates that are too wiggly, in line with what would be expected from the results of the simulations.

6. Conclusions

In this paper, we present some guidelines to be taken into account when applying a non-parametric analysis methodology for the case of censored data, specifically the proposal of Orbe and Virto [13], which adapts P-splines using Kaplan–Meier weights to the censored case. For the application of this methodology, it is necessary, as in the uncensored case, to choose a smoothing parameter and the number and location of the knots. To the best of our knowledge this topic has only been previously studied in the literature in Aydin and Yilmaz [15] but under a different methodological approach than this paper (using synthetic data). We propose different alternatives here for choosing the optimal level of smoothing and the location and number of knots in the presence of censored data. Using a large simulation study that considers functional relationships of various degrees of complexity between the censored response variable and a regressor, we analyse the performance of the different proposals in situations with differences in the information available, where varied sample sizes and levels of censorship are combined.

To choose the optimal level of smoothing, which is essential to obtain a good fit, it is usual in practice in the context of uncensored data to use the generalized cross-validation criterion or the modification proposed by Kim and Gu [19] to avoid overfitting. In Section 3, we report that a direct application of these criteria to the censored case leads to highly biased estimates. To correct this problem, we adapt the previous criteria using Kaplan–Meier weights to take into account the effect of censorship, both in the denominator of the formula, using the H_c projection matrix, and in the numerator, which is directly weighted using Kaplan–Meier weights. In addition to the above proposal, in the simulations an alternative criterion using square Kaplan–Meier weights is considered. From the analysis, we conclude that in regard to choosing the exponent of the Kaplan–Meier weights to correct the GCV, if the level of censoring is low both exponents perform similarly, but with high censorship exponent 1 (w_i) clearly performs better than exponent 2 (w_i^2). In addition, the proposed criterion (GCV_c) presents different versions according to the values of γ. In the simulation study, the possible values of this parameter that we use are those usually found in the literature; a value of 1 (as in the case of the ordinary GCV) or a value of 1.5. We conclude that a γ with a value of 1.5 is better in almost all situations since it avoids overfitting. This result is similar to that found in the literature for the uncensored case

[19,23]. Furthermore, the improvement obtained by using the value of $\gamma = 1.5$ becomes greater as the level of censorship increases.

With regard to the choice of number and location of knots, on the one hand we study the usual proposals in the literature for the uncensored case, i.e. equally spaced knots and a number of knots calculated by applying the formula presented in Ruppert [8]. On the other hand, we also analyse the performance of two new proposals adapted to the case of censored data. For the number of knots, we modify Ruppert's formula to take into account the percentage of censored observations (K_c, equation 9). From the results of the simulations, we conclude that the number of knots chosen with the K_c expression always performs better than that obtained using Ruppert's proposal. With a low level of censorship the difference is small, but as the censorship level increases our proposal clearly performs better than Ruppert's. For knot location, we analyse not just equally spaced knots but also nonuniform knot vectors with knots spaced as a function of the Kaplan–Meier weights (L_{km}). In general, equally spaced knots perform better.

One of the main conclusions that we obtain from the simulation analysis is that the adaptation of the *GCV* criterion proposed for the censored case, GCV_c , chooses smoothing parameters that lead to good estimates in the different scenarios for each example analysed. In general, GCV_c obtains the best results when its numerator is weighted with the Kaplan–Meier weights and with a γ value of 1.5. If K_c equally spaced knots are used, where K_c is the number of knots resulting from our proposal, the combination of parameters obtained generally produces the best results. When the level of censorship is small, there are no major differences in the estimates with different combinations of parameters. When the level of censorship increases these differences become much larger, so choosing the right parameters becomes even more important.

These results do not depend on what estimation method is used in the simulations. In general, the two estimation methods that we use, censored P-splines (ckmPS) and corrected GAMs (ckmGAM), perform similarly, with very good results for the proposed combination of parameters. As expected, the larger the sample size and the lower the level of censorship, the more precise the estimates are. Finally, the results are robust to the differences in the variability or distribution of the error term. The application to real data serves to illustrate the potential advantages of its use when the true functional form of the relationship is not known.

Acknowledgments

We thank an associate editor and an anonymous referee for their careful reading of the paper and suggestions, which have led to significant improvements in the contents and presentation of the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Eusko Jaurlaritza [grant number IT1359-19], Ministerio de Economía, Industria y Competitividad, Gobierno de España [grant number PECO2016-76884-P] and Spanish Research Agency of the Ministry of Science and Innovation [grant number PID2019-105183GB-I00].

References

[1] Härdle W. Applied nonparametric regression. Cambridge: Cambridge University Press; 1990. (Econometric Society Monographs; 19.

[2] Silverman BW. Density estimation for statistics and data analysis. London: Chapman and Hall; 1986. (Monographs on Statistics and Applied Probability; 26).

[3] Eubank RL. Spline smoothing and nonparametric regression. New York (NY): Marcel Dekker; 1988.

[4] Green PJ, Silverman BW. Nonparametric regression and generalized linear models. London: Chapman and Hall; 1994. (Monographs on Statistics and Applied Probability; 58).

[5] Wahba G. Spline models for observational data. Philadelphia: SIAM; 1990. (CBMS-NSF Regional Conference Series in Applied Mathematics; 59).

[6] Currie ID, Durban M. Flexible smoothing with P-splines: a unified approach. Stat Model. 2002;2(4):333–349.

[7] Friedman JH, Silverman BW. Flexible parsimonious smoothing and additive modeling (with discussion). Technometrics. 1989;31(1):3–39.

[8] Ruppert D. Selecting the number of knots for penalized splines. J Comput Graph Stat. 2002;11(4):735–757.

[9] O'Sullivan F. A statistical perspective on ill-posed inverse problems (with discussion). Stat Sci. 1986;1:502–527.

[10] O'Sullivan F. Fast computation of fully automated log-density and log-hazard estimators. SIAM J Sci Stat Comput. 1988;9(2):363–379.

[11] Eilers PH, Marx BD. Flexible smoothing with B-splines and penalties (with discussion). Stat Sci. 1996;11:89–121.

[12] Eilers PH, Marx BD, Durbán M. Twenty years of P-splines. Stat Oper Res Trans. 2015;39(2):149–186

[13] Orbe J, Virto J. Penalized spline smoothing using Kaplan–Meier weights with censored data. Biom J. 2018;60:947–961.

[14] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–481.

[15] Aydin D, Yilmaz E. Modified spline regression based on randomly right-censored data: a comparative study. Commun Stat Simul Comput. 2018;47(9):2587–2611.

[16] Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data. Ann Stat. 1981;9(6):1276–1285.

[17] Lai TL, Ying Z. Asymptotically efficient estimation in censored and truncated regression models. Stat Sin. 1992;2:17–46.

[18] Zhou M. Asymptotic normality of the synthetic data regression estimation for censored survival data. Ann Stat. 1992;20(2):1002–1021.

[19] Kim YJ, Gu C. Smoothing spline Gaussian regression: more scalable computation via efficient approximation. J R Stat Soc Ser B. 2004;66:337–356.

[20] De Boor C. A practical guide to splines, revised version. New York (NY): Springer-Verlag; 2001. (Applied Mathematical Sciences; 27).

[21] Dierckx P. Curve and surface fitting with splines. Oxford: Oxford University Press; 1993. (Numerical Mathematics and Scientific Computation).

[22] Stute W. Consistent estimation under random censorship when covariables are present. J Multivar Anal. 1993;45(1):89–103.

[23] Wood SN. Generalized additive models: an introduction with R. Boca Raton, FL: CRC Press; 2017. (Texts in Statistical Science Series.

[24] Ruppert D, Wand MP, Carroll RJ. Semiparametric regression. Cambridge: Cambridge University Press; 2003.

[25] Dickson ER, Grambsch PM, Fleming TR, et al. Prognosis in primary biliary cirrhosis: model for decision making. Hepatology. 1989;10(1):1–7.

[26] Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York (NY): Springer-Verlag; 2000.

[27] Fleming TR, Harrington DP. Counting processes and survival analysis. Hoboken, NJ: John Wiley & Sons; 2005.
[28] R Core Team. A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/
[29] Therneau TM. A package for survival analysis in S. R package version 2.38; 2015. Available from: https://CRAN.R-project.org/package = survival
[30] Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. Hoboken, NJ: John Wiley & Sons Inc; 2002. (Wiley Series in Probability and Statistics.

Capítulo 9

Penalized spline smoothing using Kaplan-Meier weights in semiparametric censored regression models

SORT 46 (1) January-June 2022, 95-114 DOI: 10.2436/20.8080.02.119

Penalized spline smoothing using Kaplan-Meier weights in semiparametric censored regression models

Jesus Orbe* and Jorge Virto*

Abstract

In this article we consider an extension of the penalized splines approach in the context of censored semiparametric modelling using Kaplan-Meier weights to take into account the effect of censorship. We proposed an estimation method and develop statistical inferences in the model. Using various simulation studies we show that the performance of the method is quite satisfactory. A real data set is used to illustrate that the proposed method is comparable to parametric approaches when assuming a probability distribution of the response variable and/or the functional form. However, our proposal does not need these assumptions since it avoids model specification problems.

MSC: *62N02.*

Keywords: *Censored data, Kaplan-Meier weights, P-splines, semiparametric models, survival analysis.*

1. Introduction

In this paper we present a proposal for estimating regression models where the variable to be explained is censored. That is, our research context is a scenario where the values of the explanatory variables are fully known but some observations of the variable to be explained are not known because there is censored data. This problem is very common in survival or duration analyses, where the sample individuals analysed are tracked over time until the specific event studied occurs (death, failure, breakdown, etc) or the study ends. In practice, there are various types of censoring, but the most common is right censoring. There is an a large body of literature on censored data, much of which can be grouped into two main approaches: one comprising models that directly specify

* Department of Quantitative Methods, University of the Basque Country UPV/EHU, Bilbao, Spain
Received: October 2021
Accepted: April 2022

the effect of the explanatory variables on the variable to be explained (the most widely used of which are those known as Accelerated Failure Times (AFT) see for example Kalbfleisch and Prentice, 2002) and the other comprising hazard models, the best known and most widely applied of which is Proportional Hazard (PH), proposed by Cox (1972). In the former a regression model is specified between the logarithmic transformation of the variable to be explained and the explanatory variables. The latter specifies a relationship between the hazard function of the variable to be explained and the explanatory variables.

PH models have the advantage that the effects of the explanatory variables can be estimated without having to assume a probability distribution for the variable to be explained which is usually unknown. However they also have the drawback that the assumption of proportional hazard functions must be imposed. Another drawback of the hazard functions approach is that the effect of the explanatory variables on the variable to be explained is hard to interpret: the results obtained from Cox model fits are harder to explain to non-statisticians and provide less information than AFT-type models, which are more attractive because they can be interpreted simply and straightforwardly (Wei, 1992; Reid, 1994; Stare, Heinzl and Harrel, 2000; Swindell, 2009). Therefore, in terms of interpretability of results the linear regression model is an attractive alternative to models for hazard functions or hazard ratios. However, its main disadvantage is that the usual estimation procedure for AFT-type models requires a probability distribution to be assumed.

The proposal presented here seeks to make the modelling of this type of data more flexible without imposing restrictions or assumptions that may prove restrictive or false in practice. We also propose an approach for making inferences in this flexible model. Our proposal can be classed as an AFT type model. Several papers using this particular approach can be found in the literature which enable the regression model to be estimated with no need to choose a specific probability distribution. They consider various least squares approaches, and include the papers by Koul, Susarla and Van-Ryzin (1981) and Leurgans (1987), who propose transforming the censored variable, and those by Miller (1976), Buckley and James (1979) and Stute (1993), which present proposals with a similar approach but without transforming the variable to be explained. There is also the rank-based estimation methods approach (see for example Tsiatis, 1990; Lai and Ying, 1992; Jin et al., 2003).

These proposals represent considerable progress in the specification of the model, avoiding the biases derived from wrong choices of probability distribution. But it is possible to go even further in making these methodologies more flexible, since all these proposals consider a known parametric relationship to specify the effect of the explanatory variables on the variable to be explained. In practice, it is quite common for the functional relationship between regressor variables and outcome not to be known. One way of avoiding errors likely to lead to biased conclusions in specifying these effects is not to impose a specific parametric functional relationship between the variable to be explained and the explanatory variable, but to assume only that that relationship is a

smooth function, *i.e.* to consider a nonparametric estimation of that specific effect. The estimation of nonparametric functional relationships involving non-censored data has been widely studied and various proposals have been presented in the literature. They can be grouped into two different approaches: methods based on kernel smoothers (Silverman, 1986; Härdle, 1990) and methods based on spline smoothers (Eubank, 1988; Wahba, 1990; Green and Silverman, 1994; Eilers and Marx, 1996; Wood, 2017).

Applying these nonparametric estimation techniques is not straightforward in the case of censored data, so the earlier studies must be adapted to take into account the effect of censoring in the estimation process. Our proposal falls under the spline smoothers approach in the specific context of semiparametric regression models with censored data. This semiparametric regression model has already been studied and discussed in regard to samples without censored observations. It was initially analysed by Heckman (1986) and Rice (1986) using an approach with spline smoothers and by Speckman (1988) using an approach with kernel smoothers. Several authors have investigated inference in the semiparametric regression model when the response variable is subject to right censoring. Orbe, Ferreira and Núñez Antón (2003) use an approach based on smoothing splines while Zou, Zhang and Qin (2011) and Chen et al. (2015) use penalized splines and monotone B-splines, respectively. Aydin and Yilmaz (2018) apply the ideas proposed by Koul et al. (1981) in the context of a partial linear regression model and De Uña Álvarez and Roca Pardiñas (2009) consider the use of kernel smoothers in an additive censored regression model.

A previous paper by Orbe and Virto (2018) proposes an extension of the P-splines method of Eilers and Marx (1996), which has become very popular in applications and in theoretical work and is an active area of research (Eilers, Marx and Durbán, 2015), to handle censored responses using Kaplan-Meier weights (Kaplan and Meier, 1958). But the proposal by Orbe and Virto provides no tools to allow statistical inferences to be made, and considers the case of a unique covariate. It is therefore of limited use in practice, where the response variable usually depends on a large set of explanatory variables and it is of interest to draw inferences. Here we propose an extension of that previous paper that enables the technique to be applied to more general problems where the effect of other covariates is incorporated parametrically (parametric component) in addition to the nonparametric component for modelling effects where the functional relationship is not known, that is, a semiparametric regression model. Such extension is a well-studied problem for case of uncensored data (see, for example, Heckman, 1986; Schimek, 2000; Holland, 2017). We also develop variance estimators for both the parametric and nonparametric components and provide the tools needed to develop statistical inferences in this general framework and study performance by calculating coverage probabilities of the confidence intervals for the true values of interest in several simulation studies.

The rest of the paper is organized as follows. Section 2 shows how to extend the P-splines method when the sample has censored observations and proposes a censored data version of penalized splines. Section 3 examines the methodology proposed using simulation studies. Section 4 presents an application of the method to a real data set and Section 5 concludes.

2. Methodology

The existence of censored observations is very common in survival analysis or duration analysis, where the aim is to analyse a variable that measures the duration of an event or state or the time that elapses until a specific event occurs. In other words, we consider a model that allows us to analyse the effect of certain explanatory variables on a variable to be explained T, the duration variable or usually its logarithmic transformation, where some of its observations are censored. Furthermore, we separate the effects of the explanatory variables of the model into two components: a component captures the relationship between some explanatory variables (X) and the response variable assuming a specific parametric functional form (parametric component) and the other component captures the effects of other explanatory variables (Z) whose functional form is unknown (nonparametric component) and which we leave unspecified, without assuming a particular parametric relationship. Therefore, we are considering a semiparametric regression model but in a context where the variable to be explained in the model is right-censored:

$$T_i = X_i^{\top}\alpha + f(Z_i) + \varepsilon_i \qquad i = 1,\ldots,n \tag{1}$$

where we assume that the values of the variable T: $t_1,\ldots,t_n$ are independent and generated with an unknown probability distribution function F. That is, we are not assuming any probability distribution for the error term. In addition some observations of that variable T are not known due to the problem known as right censoring. Therefore, what we actually observe in the sample is the variable $y_i = \min(t_i, c_i)$, where the values $c_1,\ldots,c_n$ are the values of the censoring variable C. For the censoring mechanism it will be assume: a) the lifetimes and the censoring times are independent and, b) given the lifetime, the covariates do not provide any further information as to whether censoring will take place or not, *i.e.*, $P[T \leq C|X,Z,T] = P[T \leq C|T]$ (see Stute, 1993, 1999, for a discussion of these assumptions).

We use the indicator $\delta_i = I(t_i \leq c_i)$ to show whether in particular the value t_i is observed, *i.e.*, it is not censored. In addition, X_i is the $(p \times 1)$ vector that collects the values of the p explanatory variables of the parametric component for the i-th individual, α is the $(p \times 1)$ coefficients vector of the model associated with those regressor variables, $f(Z)$ represents the nonparametric component of the model, which captures the unknown functional form of the effect of the regressor variable Z and ε is the error term satisfying $E(\varepsilon|X,Z) = 0$ and $Var(\varepsilon|X,Z) = \sigma^2$.

2.1. Estimation method

Our proposal is based on the nonparametric estimation approach proposed by Eilers and Marx (1996) together with the idea of using Kaplan-Meier weights, proposed by Stute (1993), to control the effect of censoring in the estimation of the model. Thus following this particular approach, if we want to estimate the nonparametric component of the model without assuming a particular functional form $f(\cdot)$ to the unknown effect

of the regressor variable Z, we will use an approximation that rewrites or represents that effect by using a set of q B-splines type basis functions: $B_1(z),\ldots,B_q(z)$ (see, for example, Dierckx, 1993; De Boor, 2001). Thus we rewrite the unknown function as $f(z)=\sum_{j=1}^{q}\gamma_j B_j(z)$.

In order to solve the problem of choosing the number of the knots of the bases, we use the proposal of Eilers and Marx (1996) which introduces a penalty term in the estimation process of the model. This penalty term is based on the idea of previous works by O'Sullivan (1986, 1988) that propose to use a penalty term that measures the smoothness of the function through the integrated squared second derivative of the fitted function. Eilers and Marx (1996) in their proposal of the P-splines methodology suggest using, with the same idea, a different penalty term, which generalizes and simplifies the previous proposal, introducing a penalty but on the difference of the γ_j coefficients of the adjacent B-splines.

In order to account for the effect of censoring we follow the ideas of Orbe and Virto (2018) who extend the possibility of applying the P-spline methodology to the context of samples with censored observations in a simple model. Thus, to estimate the model (1) we propose to minimize the following expression:

$$\sum_{i=1}^{n} w_{[i]}\left[y_{(i)}-x_{[i]}^{\mathsf{T}}\alpha-\sum_{j=1}^{q}\gamma_j B_j(z_{[i]})\right]^2+\lambda\sum_{j=k+1}^{q}(\Delta^k\gamma_j)^2 \tag{2}$$

where $y_{(1)},\ldots,y_{(n)}$ are the ordered values of the observed variable $y_i=min(t_i,c_i)$, $x_{[i]}^{\mathsf{T}}$ is the $(1\times p)$ vector with the values of the regressors of the parametric component for the individual corresponding to the ordered observation $y_{(i)}$, $w_{[i]}$ is the Kaplan-Meier weight associated with that observation $y_{(i)}$ and this weight is calculated using the estimator $(\hat{F}_n)$ (Kaplan and Meier, 1958) of the probability distribution function F of the variable to be explained T:

$$w_{[i]}=\hat{F}_n(y_{(i)})-\hat{F}_n(y_{(i-1)})=\frac{\delta_{[i]}}{n-i+1}\prod_{j=1}^{i-1}\left[\frac{n-j}{n-j+1}\right]^{\delta_{[j]}} \tag{3}$$

without the need to assume a probability distribution for the error term, therefore a flexible methodology is used regarding to parametric assumption of the error. Furthermore $\Delta\gamma_j$ denotes the difference between the coefficients of adjacent B-splines $(\gamma_j-\gamma_{j-1})$ and $\Delta^k\gamma_j$ indicates that this difference is of order k. This difference measures the smoothness of the function $f(z)$, the larger the difference between the coefficients of adjacent B-splines the less smooth the function. Finally the parameter λ is the smoothing parameter that controls the degree of the smoothness of the estimated function in the estimation process.

The expression to minimize (2) can be rewritten in matrix form as follows:

$$(Y-X\alpha-B\gamma)^{\mathsf{T}}W(Y-X\alpha-B\gamma)+\lambda\gamma^{\mathsf{T}}D_k^{\mathsf{T}}D_k\gamma \tag{4}$$

where X is the $(n \times p)$ design matrix for the variables of the parametric component. Y is the vector of the observed variable to be explained. B is a $(n \times q)$ matrix where $B_{ij} = B_j(z_i)$. W is a $(n \times n)$ diagonal matrix with Kaplan-Meier weights. D_k is the matrix used to rewrite the Δ^k term in matrix form.

2.2. Algorithm

The optimization process of the expression (4) leads to the following equations:

$$\left(X^{\intercal}WX\right)\alpha = X^{\intercal}W(Y - B\gamma) \tag{5}$$

$$\left(B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right)\gamma = B^{\intercal}W(Y - X\alpha) \tag{6}$$

In practice, the estimations of α and γ can be obtained by means of an iterative process or backfitting algorithm that iteratively solves each set of equations (5) and (6) until the convergence of the estimators is reached. We describe the algorithm process as follows:

- Step 1. In equation (6) give initial value of $\widehat{\alpha}_{(0)} = \vec{0}$ and estimate γ by $\widehat{\gamma}_{(0)} = \left[B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right]^{-1} B^{\intercal}WY$.
- Step 2. Substitute γ by $\widehat{\gamma}_{(0)}$ in equation (5) and estimate α by $\widehat{\alpha}_{(1)} = [X^{\intercal}WX]^{-1}X^{\intercal}W(Y - B\widehat{\gamma}_{(0)}) = [X^{\intercal}WX]^{-1}X^{\intercal}W(I - H_c)Y$ where $H_c = B\left(B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right)^{-1}B^{\intercal}W$ is the smoothing matrix for the censored case obtained from equation (6).
- Step 3. Substitute α by $\widehat{\alpha}_{(1)}$ in equation (6) and estimate γ by $\widehat{\gamma}_{(1)} = \left[B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right]^{-1} B^{\intercal}W(Y - X\widehat{\alpha}_{(1)})$.
- Step 4. Iterate step 2 and step 3 until convergence is achieved.

The algorithm is considered to have converged when the difference between the GCV_c (see equation 8) of two successive iterations is less than a really small threshold: $|GCV_c(new) - GCV_c(old)| < 0.00001 \cdot GCV_c(new)$.

2.3. Choice of smoothing parameter and knots

It should be noted that in this iterative process we need to make a number of choices, such as the number of knots (K_c) and the choice of the smoothing parameter λ, in order to estimate the components of the model. The use of a penalty term in the optimization criterion makes the determination of the number of knots not a crucial decision as long as a sufficient number of knots is chosen. To choose this number of knots in samples with censored data we propose the following automatic choice criterion that takes into account the sample information available due to the existence of censored data by multiplying by one minus the proportion of censored observations:

$$K_c = round\left(\min\left(\frac{m}{4}, 40\right) \cdot (1 - PC)\right) \tag{7}$$

where m is the number of distinct values of the Z variable of the nonparametric component and PC represents the level of censoring, measured as a percentage, existing in the analysed sample. The expression (7) is a modification to the one proposed for the choice of the number of knots in Ruppert (2002) that we propose for application in contexts with censored data.

The choice of the smoothing parameter is a more relevant choice. To choose an optimal smoothing level we propose to use the following version of the generalized cross validation (GCV) criterion adapted for application in contexts with censored data:

$$GCV_c = \sum_{i=1}^{n} \frac{w_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{(n - \phi tr(H_c))^2} \tag{8}$$

where ϕ is a parameter that tries to correct for the overfitting problem that occurs when using the ordinary GCV criterion. Wood (2017) proposes to use what he refers to as the double cross validation and suggests using a value of $\phi = 1.5$. This value is justified in different ways in the literature, see for example Kim and Gu (2004) for the uncensored case and Orbe and Virto (2021) for the censored case. The performance of proposal (8) has been analysed using a simulation study and, as in the uncensored case, the choice of $\phi = 1.5$ is better in almost all situations than $\phi = 1$, with the difference increasing as the censoring increases.

2.4. Variances estimation

In this section we develop the necessary tools to perform statistical inferences for the parametric and nonparametric components.

In order to determine the variance of the parametric component, we first solve equation (6) getting $\gamma = \left(B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right)^{-1} B^{\intercal}W(Y - X\alpha)$. Therefore, substituting $B\gamma = H_c(Y - X\alpha)$ in equation (5) we get $(X^{\intercal}WX)\,\alpha = X^{\intercal}W\,[Y - H_c(Y - X\alpha)]$. Solving for α we obtain $\widehat{\alpha} = [X^{\intercal}W(I - H_c)X]^{-1} X^{\intercal}W(I - H_c)Y$. Accordingly, the variance-covariance matrix of this estimator can be expressed as:

$$\widehat{Var}(\widehat{\alpha}) = \widehat{\sigma}^2 \left\{ \left(X^{\intercal}W(I - H_c)X\right)^{-1} X^{\intercal}W(I - H_c)(I - H_c)^t WX \right.$$
$$\left. \left((X^{\intercal}W(I - H_c)X)^{-1}\right)^t \right\} \tag{9}$$

In a similar way, we solve equation (5) getting $\alpha = (X^{\intercal}WX)^{-1} X^{\intercal}W(Y - B\gamma)$. Plugging $X\alpha = X\,(X^{\intercal}WX)^{-1} X^{\intercal}W(Y - B\gamma) = H_p(Y - B\gamma)$, where $H_p = X\,(X^{\intercal}WX)^{-1} X^{\intercal}W$, in equation (6) we get $\left(B^{\intercal}WB + \lambda D_k^{\intercal}D_k\right)\gamma = B^{\intercal}W\,[Y - H_p(Y - B\gamma)]$. Solving for γ we get $\widehat{\gamma} = \left[B^{\intercal}W(I - H_p)B + \lambda D_k^{\intercal}D_k\right]^{-1} B^{\intercal}W(I - H_p)Y$. Accordingly, the variance-covariance matrix of this estimator can be expressed as:

$$\widehat{Var}(\widehat{\gamma}) = \widehat{\sigma}^2 \left\{ \left[B^{\intercal}W(I - H_p)B + \lambda D_k^{\intercal}D_k\right]^{-1} B^{\intercal}W(I - H_p)(I - H_p)^t WB \right.$$
$$\left. \left(\left[B^{\intercal}W(I - H_p)B + \lambda D_k^{\intercal}D_k\right]^{-1}\right)^t \right\} \tag{10}$$

In order to calculate these estimated variances we need to estimate the σ^2 parameter. We propose the estimator given by the following expression:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} nw_{[i]}(y_{(i)} - \hat{y}_{(i)})^2}{n - tr(H_c) - p}$$

3. Simulation study

In this section the performance of the proposed methodology is studied using a simulation study. In order to do that we consider the next semiparametric model:

$$T_i = \alpha_1 X_{1i} + \alpha_2 X_{2i} + f(Z_i) + \varepsilon_i \tag{11}$$

where for the parametric component of the model: the variable X_1 is generated from a uniform distribution $U(0,2)$, X_2 from a uniform distribution $U(-1,3)$, being $\alpha_1 = -1$ and $\alpha_2 = 1$ the values of the coefficients. For the nonparametric component, we consider three different cases for the relationship $f(\cdot)$ between T and a relevant covariate Z, see Table 1 for the chosen functional forms and the probability distribution of the variable Z. For the distribution of the error term (ε) has been used the normal distribution $N(0,\sigma^2)$, where the value of σ^2 parameter has been chosen to obtain a similar signal/noise (SN) ratio in each example (see Table 1). In order to study the effect of censoring, we consider a censoring variable C generated independently from a uniform distribution $U(1,b)$. The value of parameter b changes to consider three different levels of censored data: 10%, 25% and 40%. Therefore, we observe $(y_1,x_{11},x_{21},z_1,\delta_1),\ldots,(y_n,x_{1n},x_{2n},z_n,\delta_n)$ a sample of size n, where $y_i = min(t_i,c_i)$ is the observed survival time, *i.e.*, the minimum between the survival time t_i and the censoring value c_i. In addition, it is known through the indicator variable $\delta_i = I(t_i \leq c_i)$ which observations are not censored. We use three sample sizes: $n = 200$, $n = 500$ and $n = 1000$. For each of the nine scenarios, three sample sizes for three levels of censorship, we consider 1000 Monte Carlo replications.

Table 1. *Three Case Studies.*

Name	z_i	$f(z_i)$	σ_ε^2	SN ratio
Case (i): Quadratic	$z_i \sim U[0,4]$	$2+4z_i - z_i^2$	0.40	3.5
Case (ii): Sinusoidal	$z_i \sim U[0,10]$	$2+exp\{sin(z_i)\}$	0.20	3.3
Case (iii): Logit	$z_i \sim U[0,1]$	$2+\frac{1}{1+exp\{-20(z_i-0.5)\}}$	0.06	3.3

For each of the 27 cases analysed in this simulation study we have estimated model (11) following the estimation proposal presented in the previous section, the censored P-

spline estimator (CPS), where the choice of the smoothing parameter λ and the number of knots of B-splines have been chosen using formulas (8) and (7), respectively.

Tables 2, 3 and 4 present a general summary of the results obtained for each combination of censoring level and sample size in each of the three cases of functional forms studied for model (11). That is, Table 2 summarizes the estimation of case (i), where $f(z)$ is a quadratic function. The first two rows of Table 2 present the estimated Mean Square Error (MSE) of each coefficient (α_1 and α_2) of the parametric component:

$$MSE(\hat{\alpha}_p) = \frac{1}{1000}\sum_{j=1}^{1000}(\alpha_p - \hat{\alpha}_{pj})^2 \qquad p = 1,2$$

and the third row the Averaged Mean Square Error (AMSE) of the nonparametric component:

$$AMSE = \frac{1}{1000}\sum_{j=1}^{1000}\left(\frac{\sum_{i=1}^{n}(f(z_i) - \hat{f}_j(z_i))^2}{n}\right)$$

Rows four to six of Table 2 present the empirical bias and rows seven to nine the coverage probabilities of the 95% confidence intervals based on the resampling.

Tables 3 and 4 present the same information for the estimates of case (ii) and (iii), where $f(z)$ is a sinusoidal function and a logit function, respectively. Tables 2 to 4 show the good performance of the proposed method in terms of MSE and AMSE, empirical bias and coverage probabilities.

Furthermore, if we focus on the estimation of each component of model (11), we have that for case (i), quadratic function: Figure 1(a) presents the MSE estimates for the nonparametric component using different censoring levels and sample sizes, where, as can be seen, the estimates of the nonparametric component improve as the sample size increases and the level of censoring in the sample decreases. Figures 1(b) and (c) show the estimates of the coefficients of the parametric component (α_1 and α_2), where it can be seen that the coefficient estimates are good and that their accuracy also improves as the sample size increases and the level of censoring in the sample decreases. In addition, Figure 1(d) presents the mean value of the estimates of the quadratic form function compared to the true functional form to be estimated. As can be seen, the proposal we made works very well reflecting the true functional form of $f(\cdot)$. In this Figure 1(d), we can also verify the good performance of the asymptotic confidence intervals generated with the estimates of the variances proposed in the previous section. As can be seen, for a confidence level of 95%, the proposed mean confidence interval (blue lines) is consistent with the corresponding 95th percentile interval of the simulations (green lines). Finally, the coverage probabilities of the confidence intervals presented in Table 2 show that the actual coverage probability is quite close to the nominal coverage probability.

Similar results, where the good performance of our proposals can be appreciated, are obtained for case (ii), sinusoidal function, see Figures 2(a)-(d), and for case (iii), logit function, see Figures 3(a)-(d).

As suggested by the referees, we conduct additional simulations considering a normal distribution for the censoring variable and also additional simulations considering

non-normal error distributions such as the Weibull distribution. The new results obtained (not shown) confirm the good performance of the proposed method and are consistent with those presented in this section.

Table 2. *Results of simulation study for the quadratic function.*

	$n=200$			$n=500$			$n=1000$		
Censored %	10%	25%	40%	10%	25%	40%	10%	25%	40%
	MSE ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and AMSE ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	3.090	3.741	5.965	1.324	1.440	2.370	0.521	0.656	0.992
$\widehat{\alpha}_2$	0.722	0.906	1.581	0.275	0.302	0.541	0.121	0.181	0.259
$\widehat{f}$	9.783	12.170	21.109	4.126	5.056	8.730	2.105	2.580	4.424
	Empirical Bias								
$\widehat{\alpha}_1$	-0.00149	0.00099	0.01130	-0.00319	-0.00290	0.00042	0.00214	0.00354	0.00575
$\widehat{\alpha}_2$	0.00289	0.00206	-0.00257	0.00049	0.00039	-0.00335	0.00036	-0.00036	-0.00195
$\widehat{f}$	-0.00033	-0.00148	-0.01630	0.00239	0.00272	0.00067	-0.00232	-0.00253	-0.00575
	Coverage probabilities of the 95% confidence intervals								
$\widehat{\alpha}_1$	0.938	0.955	0.947	0.928	0.950	0.947	0.946	0.946	0.948
$\widehat{\alpha}_2$	0.945	0.946	0.934	0.941	0.960	0.943	0.960	0.926	0.957
$\widehat{f}$	0.938	0.941	0.923	0.939	0.939	0.925	0.946	0.936	0.933

Table 3. *Results of simulation study for the sinusoidal function.*

	$n=200$			$n=500$			$n=1000$		
Censored %	10%	25%	40%	10%	25%	40%	10%	25%	40%
	MSE ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and AMSE ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	0.806	1.060	1.521	0.285	0.362	0.560	0.136	0.154	0.233
$\widehat{\alpha}_2$	0.189	0.266	0.376	0.062	0.087	0.132	0.035	0.044	0.064
$\widehat{f}$	4.088	5.205	7.970	1.702	2.023	3.072	0.870	1.047	1.545
	Empirical Bias								
$\widehat{\alpha}_1$	-0.00543	-0.00202	0.00083	0.00085	0.00098	0.00209	0.00060	0.00016	0.00148
$\widehat{\alpha}_2$	-0.00060	-0.00073	-0.00145	0.00051	0.00058	-0.00093	-0.00016	-0.00017	-0.00136
$\widehat{f}$	0.00674	0.00311	-0.00152	-0.00116	-0.00167	-0.00324	-0.00021	0.00010	-0.00077
	Coverage probabilities of the 95% confidence intervals								
$\widehat{\alpha}_1$	0.944	0.936	0.925	0.956	0.955	0.944	0.948	0.956	0.940
$\widehat{\alpha}_2$	0.949	0.930	0.938	0.952	0.938	0.944	0.944	0.943	0.962
$\widehat{f}$	0.930	0.927	0.918	0.932	0.941	0.932	0.942	0.938	0.941

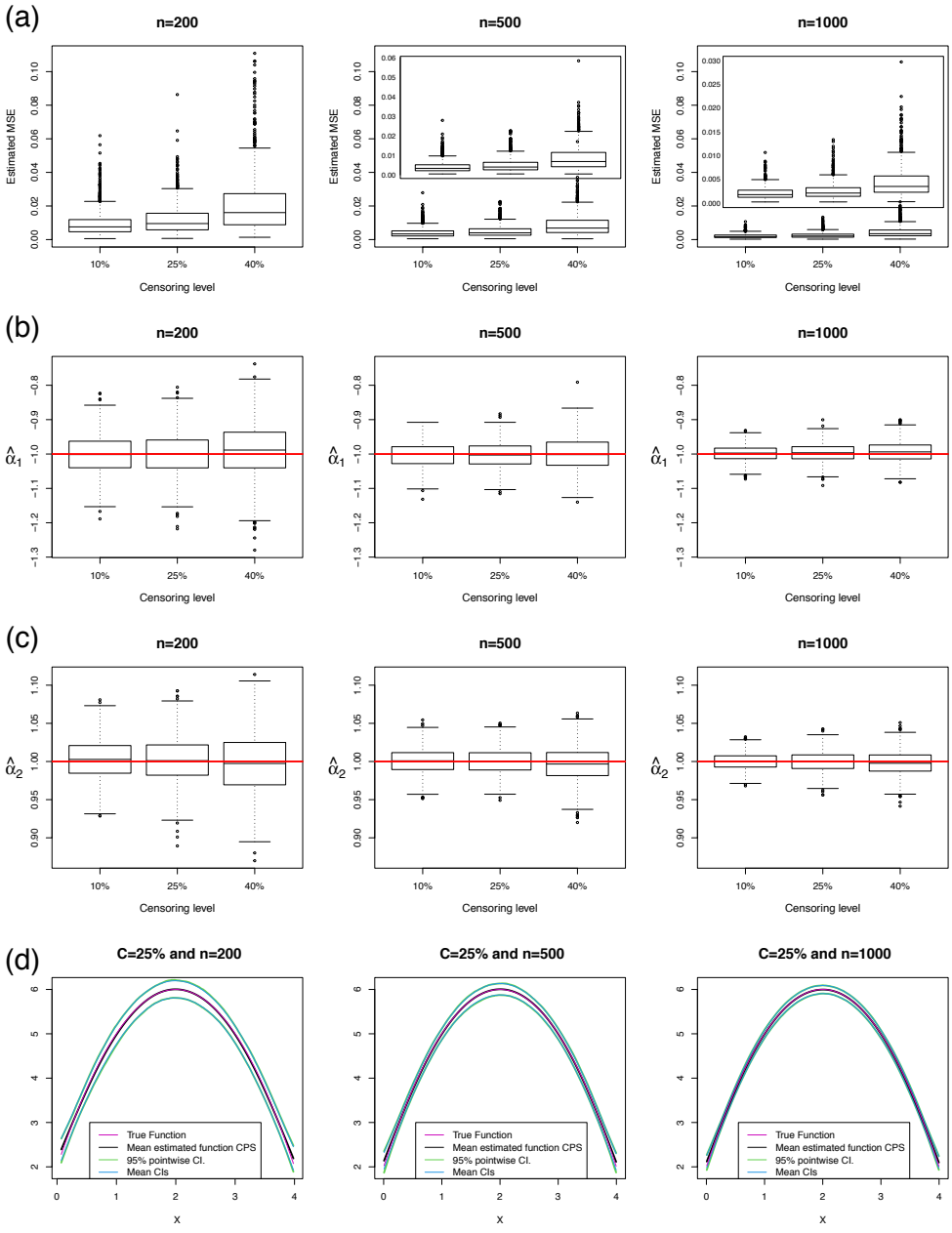


Figure 1. *Results of simulation study for the quadratic function. (a) Mean square errors for the nonparametric part using different censoring levels and sample sizes. (b)* $\hat{\alpha}_1$*. (c)* $\hat{\alpha}_2$*. (d) Mean value of the estimates of the quadratic form function compared to the true functional form to be estimated.*

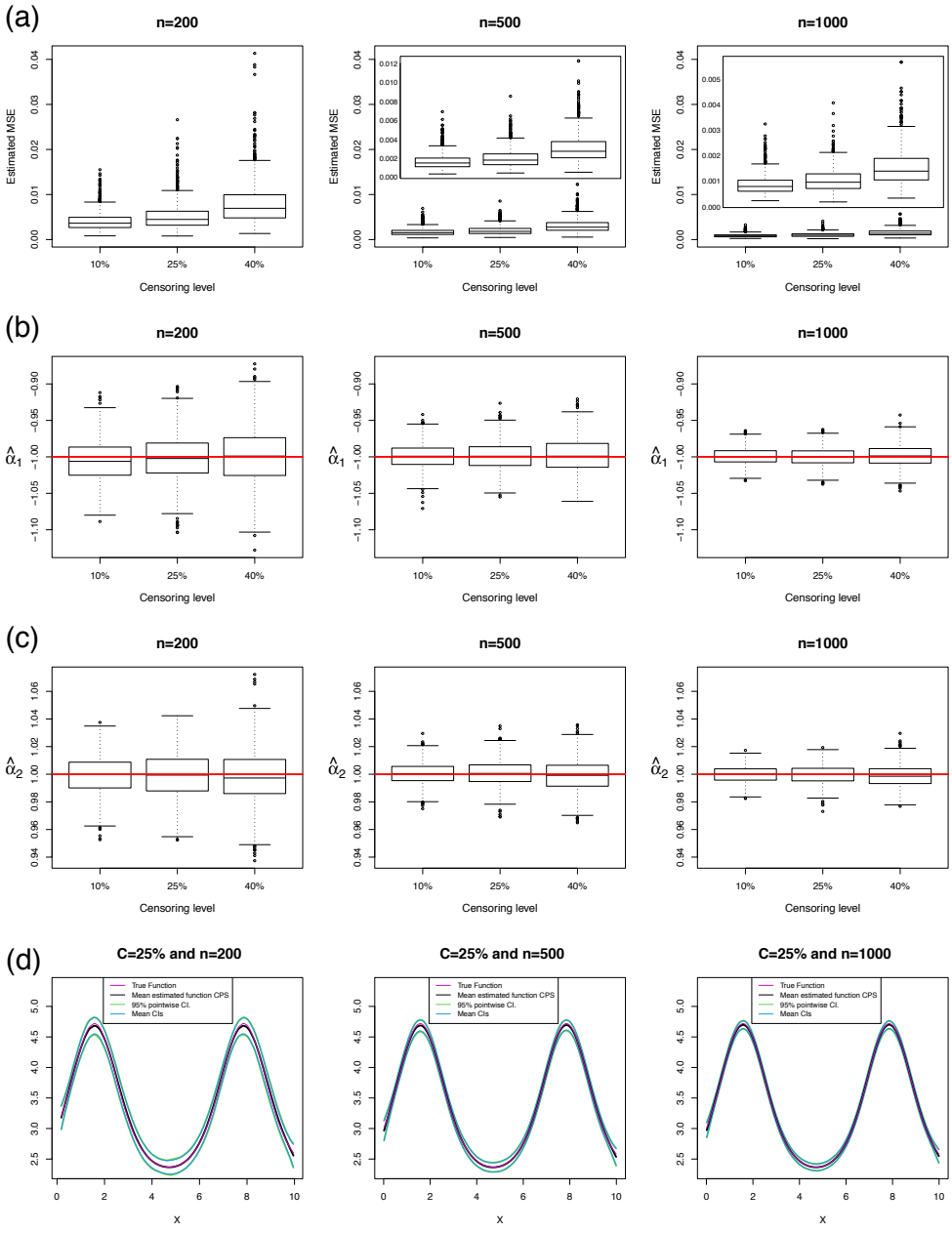


Figure 2. *Results of simulation study for the sinusoidal function. (a) Mean square errors for the nonparametric part using different censoring levels and sample sizes. (b) $\hat{\alpha}_1$. (c) $\hat{\alpha}_2$. (d) Mean value of the estimates of the sinusoidal form function compared to the true functional form to be estimated.*

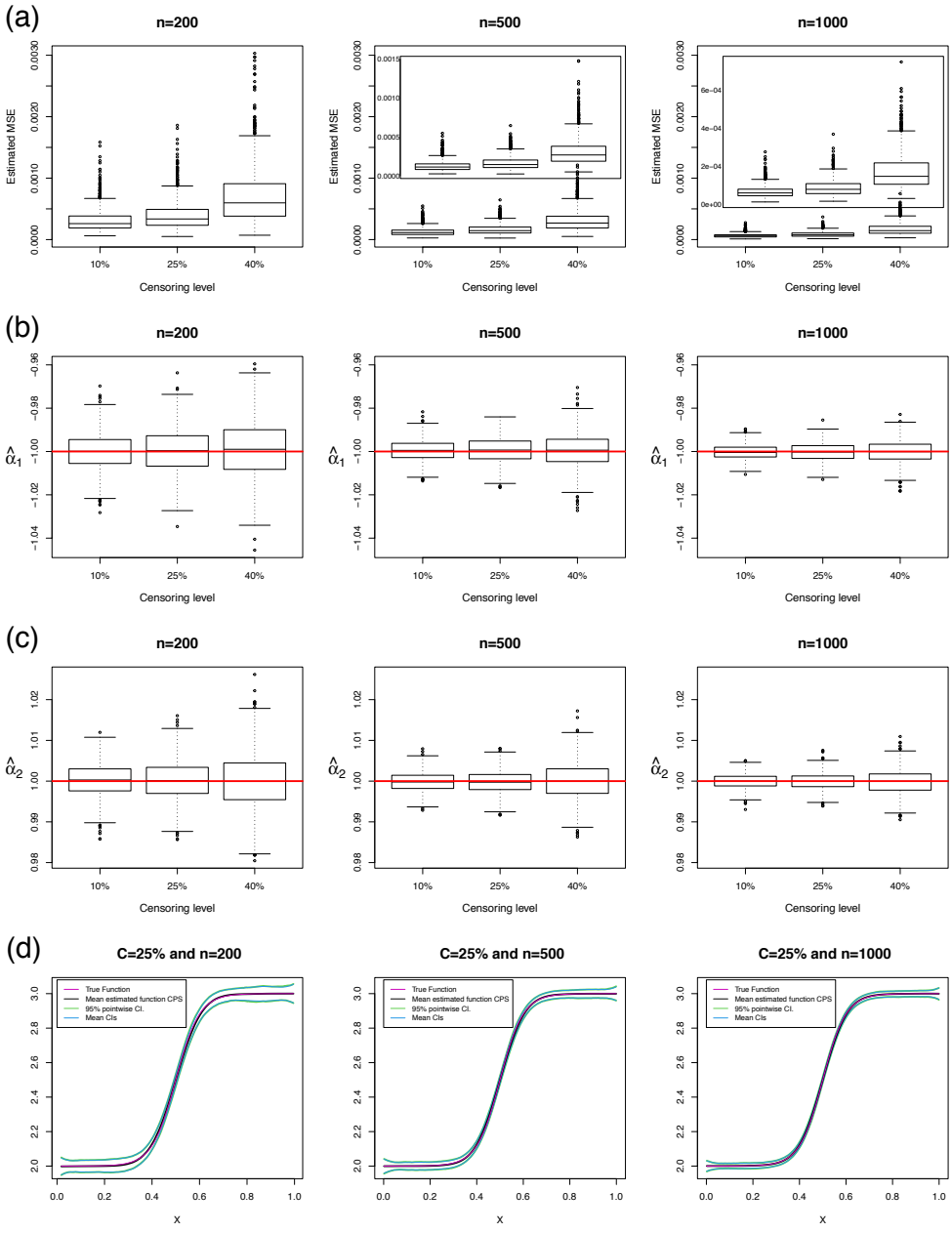


Figure 3. *Results of simulation study for the logit function. (a) Mean square errors for the nonparametric part using different censoring levels and sample sizes. (b) $\hat{\alpha}_1$. (c) $\hat{\alpha}_2$. (d) Mean value of the estimates of the logit form function compared to the true functional form to be estimated.*

Table 4. *Results of simulation study for the logit function.*

	$n = 200$			$n = 500$			$n = 1000$		
Censored %	10%	25%	40%	10%	25%	40%	10%	25%	40%
	MSE ($\widehat{\alpha}_1$ and $\widehat{\alpha}_2$) and AMSE ($\widehat{f}$) $\times 10^3$								
$\widehat{\alpha}_1$	0.072	0.098	0.172	0.024	0.036	0.064	0.011	0.016	0.027
$\widehat{\alpha}_2$	0.017	0.025	0.046	0.006	0.008	0.019	0.003	0.004	0.009
$\widehat{f}$	0.309	0.397	0.710	0.128	0.164	0.311	0.065	0.085	0.169
	Empirical Bias								
$\widehat{\alpha}_1$	0.00012	0.00029	0.00101	0.00042	0.00061	0.00035	-0.00029	-0.00022	-0.00011
$\widehat{\alpha}_2$	0.00026	0.00015	-0.00008	-0.00015	-0.00026	-0.00002	-0.00002	-0.00003	-0.00014
$\widehat{f}$	-0.00037	-0.00038	-0.00098	-0.00022	-0.00029	-0.00040	0.00035	0.00020	0.00014
	Coverage probabilities of the 95% confidence intervals								
$\widehat{\alpha}_1$	0.944	0.955	0.933	0.953	0.944	0.941	0.946	0.941	0.948
$\widehat{\alpha}_2$	0.950	0.930	0.924	0.939	0.947	0.938	0.957	0.938	0.956
$\widehat{f}$	0.944	0.939	0.916	0.943	0.938	0.93 0	0.949	0.941	0.938

4. Empirical application: PBC data

The Mayo Clinic Primary Biliary Cirrhosis dataset contains information from 418 Mayo Clinic patients with primary biliary cholangitis (PBC), previously called primary biliary cirrhosis, an autoimmune disease of the liver. The first 312 cases in the dataset participated in a Mayo Clinic trial in PBC conducted between 1974 and 1984 comparing the drug D-penicillamine (treatment) with a placebo. The dataset provides information about the observed survival time from the date of registration in the trial and a large number of clinical, biochemical, serologic and histologic variables such as patient's age at first diagnosis, severity of edema (0 no edema, 0.5 moderate and 1 for severe edema), blood values related to liver function such as bilirubin, albumin, alkaline phosphotase and prothrombin time amid other explanatory variables, and an indicator of patient status (dead or alive) in July 1986. The dataset can be downloaded from the R package *survival* (Therneau, 2021; R Core Team, 2018). The additional cases are from an independent set of 106 Mayo Clinic primary biliary cholangitis patients who were elegible for the trial but declined to participate. This dataset has been previously used, for example, in Dickson et al. (1989), Therneau and Grambsch (2000) and Fleming and Harrington (2005), in censored regression models.

The studies by Therneau and Grambsch (2000) and Fleming and Harrington (2005) deal with the relationship between the covariates and the survival response variable. They conclude that age, edema score, bilirubin and albumin logarithms and prothrombin time are the variables that best explain patient survival. In addition, these studies analyse the need for transformations of the continuous variables in the proposed model

Table 5. *Estimate and standard deviation (SD) of estimated parameters for the Mayo Clinic Primary Biliary Cirrhosis dataset from AFT, Stute and CPS methods.*

	age	edema	trt	log(albumin)	log(bili)
AFT	-0.0246	-0.7692	-0.0627	1.4880	-0.5356
	(0.0065)	(0.2303)	(0.1273)	(0.5268)	(0.0694)
Stute	-0.0166	-0.9249	-0.0950	1.6161	-0.3028
	(0.0076)	(0.3489)	(0.1371)	(0.6015)	(0.0732)
CPS	-0.0168	-0.9163	-0.0991	1.6197	-0.3061
	(0.0064)	(0.1900)	(0.1291)	(0.4578)	(0.0633)

concluding that the relationship between prothrombin time (protime) and patient survival is likely to be non-linear.

In this application we incorporate the protime variable into the model in a flexible way only assuming that prothrombin time enters in the model via some unknown smooth function $f(\cdot)$:

$$log(T) = \alpha_1 + \alpha_2 age + \alpha_3 edema + \alpha_4 trt + \alpha_5 log(albumin) + \alpha_6 log(bili) + f(protime) + \varepsilon \tag{12}$$

We estimated model (12) using the censored P-spline method proposed in section 2. To evaluate the performance of the censored P-spline estimator, a quadratic relationship between the logarithm of survival and the protime variable has been proposed as an alternative, *i.e.*, $f(protime) = \alpha_7 protime + \alpha_8 protime^2$ in equation (12). Assuming that this parametric specification is correct, two methodologies known and proposed in the literature on survival analysis can be used to fit the model (12). These estimators can be used as a benchmark to evaluate the performance of the censored P-spline method proposed. The first and more restrictive approach is the parametric Accelerated Failure Time (AFT) methodology (Kalbfleisch and Prentice, 2002), based on the restricted assumptions of knowing the probability distribution of the response variable and the functional form relating the protime variable and patient survival, that estimates the α coefficients of the model using the maximum likelihood estimator. Thus, considering an AFT lognormal model, we estimate the α coefficients assuming a normal probability distribution. The second methodology, proposed by Stute (1993), is less restrictive in that it does not need the assumption of the probability distribution of the response variable, but it also trusts the quadratic functional form. That is, it needs to know the form of the relationship between the response variable and the covariate. This methodology estimates coefficients using weighted least squares via Kaplan-Meier weights (Stute, 1993).

Table 5 presents the estimates of the parametric components of the model (12) using these three methods. It can be seen that all three methods generate similar estimates and result in a biologically reasonable model estimate. As previously reported in the literature, all three methods agree that treatment with the drug D-penicillamine (treatment) has no significant effect on patient survival.

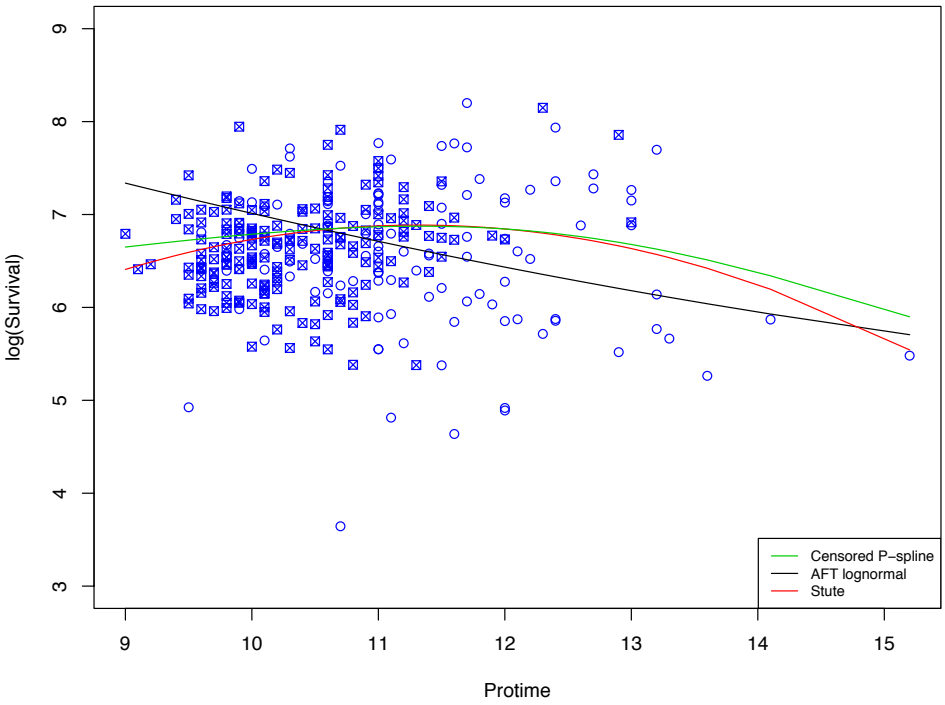


Figure 4. *Estimated relationship using three methodologies: AFT lognormal, Stute's approach and CPS estimator*

Figure 4 shows the estimates of the unknown function $f(protime)$ for the three approaches with the scatterplot of observed log survival time versus prothrombin time. Patients indicated by ◯ are dead and those indicated by ☒ are alive in July 1986; that is, the dead patients have uncensored survival times and the live patients have censored survival times.

In conclusion, the AFT methodology and Stute's proposals performance depends on the correct specification of the relationship between the duration and the protime variable. In this application it seems that the relationship between log survival and prothrombin time is quadratic, so both these methodologies perform reasonably well. Our proposal does not need to assume a specific parametric functional form and, however, it adequately estimates the relationship obtaining very similar results to the previous ones. However, if the functional form had been wrongly chosen these parametric methods would have led to a serious problem of incorrect specification and therefore to wrong conclusions. Therefore, we can see our approach as a robust solution to misspecification of the model.

5. Discussion and conclusion

In this paper, we have proposed an estimation method in the context of censored semiparametric models based on the P-spline approach of Eilers and Marx (1996) using Kaplan-Meier weights to take into account the effect of censorship. We present an extension of the estimation methodology proposed by Orbe and Virto (2018) to a context with more than one explanatory variable, which is very useful from a practical point of view. Furthermore, we develop the necessary tools to perform statistical inferences in this general framework, providing, for example, confidence intervals for both the nonparametric component and the coefficients associated with the regressors of the parametric component. The simulation studies conducted illustrate the good performance of the estimation method which satisfactorily estimates both the nonparametric component and the coefficients associated with the parametric part in the various examples studied. Furthermore, the accuracy of estimates improves as the censored level reduces and the sample size is increased. The coverage probabilities of the confidence intervals proposed have been calculated in several simulation studies and it has been found that the actual coverage probability is quite close to the nominal coverage probability in all the scenarios analysed.

The application to real data serves to illustrate the potential advantages of our proposal which is comparable with the parametric method AFT and Stute's approach when the functional form chosen is correct. Otherwise, it must be mentioned that if the functional form or the probability distribution are wrongly chosen this would lead to a serious problem of incorrect specification of the model and therefore to incorrect conclusions. The proposed method would be more flexible and robust as it does not need to impose a specific probability distribution for the response variable, nor assume a functional form for the relationship between the censored response variable and the covariate, which are usually unknown in practice. Therefore, its application in samples with censored data is particularly useful in contexts of survival or duration analysis where censored observations are common.

Funding

This study was supported by the Basque Government and the Spanish Research Agency of the Ministry of Science and Innovation under research grants UPV/EHU Econometrics Research Group IT1359-19 and PID2019-105183GB-I00.

References

Aydin, D. and Yilmaz, E. (2018). Modified estimators in semiparametric regression models with right-censored data. *Journal of Statistical Computation and Simulation*, 88:1470–1498.

Buckley, J. J. and James, I. R. (1979). Linear regression with censored data. *Biometrika*, 66:429–436.

Chen, W., Li, X., Wang, D., and Shi, G. (2015). Parameter estimation of partial linear model under monotonicity constraints with censored data. *Journal of the Korean Statistical Society*, 44:410–418.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34:187–202.

De Boor, C. (2001). *A Practical Guide to Splines, revised version*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

De Uña Álvarez, J. and Roca Pardiñas, J. (2009). Additive models in censored regression. *Computational Statistics and Data Analysis*, 53:3490–3501.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 10:1–7.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11:89–121.

Eilers, P. H., Marx, B. D., and Durbán, M. (2015). Twenty years of p-splines. *SORT-Statistics and Operations Research Transactions*, 39(2):149–186.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Fleming, T. R. and Harrington, D. P. (2005). *Counting Processes and Survival Analysis*. John Wiley & Sons, Hoboken: New Jersey.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, volume 58 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Härdle, W. (1990). *Applied Nonparametric Regression*, volume 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48:244–248.

Holland, A. D. (2017). Penalized spline estimation in the partially linear model. *Journal of Multivariate Analysis*, 153:211–235.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90:341–353.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kim, Y. J. and Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:337–356.

Koul, H., Susarla, V., and Van-Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, 9:1276 – 1288.

Lai, T. L. and Ying, Z. (1992). Linear rank statistics in regression analysis with censored or truncated data. *Journal of Multivariate Analysis*, 40:13–45.

Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika*, 74:301–309.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, 63:449–464.

Orbe, J., Ferreira, E., and Núñez Antón, V. (2003). Censored partial regression. *Biostatistics*, 4:109–121.

Orbe, J. and Virto, J. (2018). Penalized spline smoothing using Kaplan-Meier weights with censored data. *Biometrical Journal*, 60:947–961.

Orbe, J. and Virto, J. (2021). Selecting the smoothing parameter and knots for an extension of penalized splines to censored data. *Journal of Statistical Computation and Simulation*, 91:1–33.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, 1:502–527.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9:363–379.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, N. (1994). A conversation with sir david cox. *Statistical Science*, 9:439–455.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics and Probability Letter*, 4:203–208.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757.

Schimek, M. G. (2000). Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference*, 91:525–540.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50:413–436.

Stare, J., Heinzl, H., and Harrel, F. (2000). On the use of buckley and james least squares regression for survival data. In Ferligoj, A. and Mrvar, A., editors, *New Approaches in Applied Statistics*, volume 16, pages 125–134. Metodološki zvezki, Ljubljana: Eslovenia.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, 45:89–103.

Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica*, 9:1089–1102.

Swindell, W. R. (2009). Accelerated failure time models provide a useful statistical framework for aging research. *Experimental Gerontology*, 44:190–200.

Therneau, T. M. (2021). *A Package for Survival Analysis in R*. R package version 3.2-11.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, 18:354–372.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11:1871–1879.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Texts in Statistical Science Series. CRC press, Boca Raton: Florida.

Zou, Y., Zhang, J., and Qin, G. (2011). A semiparametric accelerated failure time partial linear model and its application to breast cancer. *Computational Statistics and Data Analysis*, 55:1479–1487.