

APLICACIONES ACTUARIALES MEDIANTE GAUSSIAN PROCESS REGRESSION: VIDA Y NO VIDA¹

ACTUARIAL APPLICATIONS THROUGH GAUSSIAN PROCESS REGRESSION: LIFE AND NON-LIFE

David Rius Carretero^{2 3}, Salvador Torra Porras⁴

²*Máster en Ciencias Actuariales y Financieras. Facultad de Economía y Empresa.
Universidad de Barcelona. Avenida Diagonal s/n 08034 Barcelona. España.*

⁴*Dpto. Econometría, Estadística y Economía Aplicada, Riskcenter-IREA;
Universitat de Barcelona, Av. Diagonal, 690, 08034 Barcelona. España.*

Fecha de recepción: 27/09/2022

Fecha de aceptación: 22/11/2022

Resumen

En este trabajo se ha realizado una breve introducción sobre la metodología Regresión de Proceso Gaussiano (GPR) y dos aplicaciones en el ámbito Actuarial. Por un lado, se ha realizado un ejercicio de interpolación sobre las tablas de mortalidad PASEM Unisex 2020, concluyendo que el GPR es una excelente herramienta de interpolación, y que nos permite una tarificación más ajustada en el ramo de Vida. Por otro lado, se ha integrado el GPR como medida de predicción de provisiones en los ramos de No-Vida, obteniendo unos resultados prometedores. Por último, se concluye que un GPR puede ser un instrumento útil, siempre y cuando, se realice una buena selección del Kernel y un correcto período de entrenamiento del modelo.

¹ Este artículo está adscrito al trabajo final de máster que pueden encontrar al siguiente enlace: https://github.com/drius1996/TFM_GPR

² Autor para correspondencia: davidrius96@hotmail.com

³ *Agradecimientos: Quería agradecer el amor y apoyo que han tenido mis padres, hermano y pareja durante estos años de formación académica y profesional, y dedicarles este trabajo como el final de un ciclo, al igual que al resto de mi familia y amigos, por su apoyo incondicional.*

⁴storra@ub.edu

Palabras clave: proceso gaussiano, normal multivariante, covarianza, Ciencias Actariales, distribuciones.

Abstract

In this work, a brief introduction has been made on the Gaussian Process Regression (GPR) methodology and two applications in the Actuarial field. On the one hand, an interpolation exercise has been carried out on the PASEM Unisex 2020 mortality tables, concluding that the GPR is an excellent interpolation tool, and that it allows us a more adjusted pricing in the Life branch. On the other hand, the GPR has been integrated as a predictive measure for provisions in Non-Life branches, obtaining promising results. Finally, it is concluded that a GPR can be a useful instrument, as long as a good Kernel selection and a correct model training period are carried out.

Keywords: gaussian process, multivariate normal, covariance, Actuarial Science, distributions.

1. Introducción

El actual desarrollo tecnológico nos permite elaborar nuevas técnicas más complejas, como, por ejemplo, los modelos de Deep Learning o Machine Learning. En este trabajo trataremos uno de estos modelos de aprendizaje supervisado con la finalidad de ampliar las posibilidades de los modelos de regresión, a través los Procesos Gaussianos (GP) como técnica de predicción.

Los GP son procesos estocásticos no lineales ampliamente conocidos en el ámbito matemático y financiero. Esta técnica aún se encuentra en actual de desarrollo, pues, a día de hoy surgen nuevas aplicaciones y metodologías de cálculo que mejoran el GP, como por ejemplo, en la optimización de sus hiperparámetros. Su nombre viene dado por Carl Friedrich Gauss, ya que está basado en la idea de una distribución gaussiana o normal, pues en términos generales, un GP es una extensión continua de una distribución gaussiana multivariante.

El objetivo del presente artículo es introducir la aplicabilidad de los procesos Gaussianos dentro del ámbito actuarial. Para ello, tras una necesaria explicación del funcionamiento de un GP y cómo funciona la regresión de éste, se procederá a la aplicación empírica del GP a través de dos ejemplos prácticos que se sitúan en el entorno actuarial.

La primera parte de las aplicaciones, utilizaremos los GP para obtener la interpolación mensual de las tablas de mortalidad anuales PASEM 2020 Unisex de segundo orden⁵, con el objetivo de obtener una mayor precisión en la estimación de la prima en seguros de tarificación anual renovable (TAR), acorde con la edad actuarial.

Y, con respecto a la segunda propuesta, aplicaremos la metodología de los GP para la predicción de provisiones técnicas de siniestros pendientes de declaración, realizando una comparativa con el método clásico de cálculo de provisiones: el método Chain-Ladder determinista. De este modo, seremos capaces de modelar funciones no paramétricas que nos permitirán obtener las tareas de predicción e interpolación en términos probabilísticos.

2. Proceso Gaussiano

El GP es una colección de variables aleatorias, tal que, para cualquier subconjunto finito de puntos, tienen una distribución conjuntamente gaussiana, siguiendo estos, una distribución normal multivariada, indexadas por tiempo o espacio. Dicho de otro modo, un GP es una generalización de las distribuciones de probabilidades gaussianas multivariadas.

Previo a indagar en un GP, primero detallaremos brevemente que es una distribución gaussiana y como llegamos a este GP.

2.1. Distribución Gaussiana

Una distribución gaussiana es una distribución que queda definida por dos parámetros, la media y su desviación típica, y su función de densidad (la que nos asignará los valores más probables para un conjunto de datos) queda explicitada de la siguiente forma:

Partiendo de un determinado $input = x$, la ecuación para una distribución normal univariante tiene la siguiente función de densidad⁶:

Función de densidad de una Distribución Gaussiana Univariante:

⁵ PASEM 2020 Unisex de 2º orden se utilizan para los seguros de Vida-Riesgo que no contemplen seguros de decesos.

⁶ Función que nos define, para una variable aleatoria, que valor puede tomar en base a la probabilidad relativa.

$$f(x) = p(y|\mu, \sigma^2) \quad (1)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

Donde, dominio de $f = \mathbb{R}$:

$\mu = E\{x\}$, siendo este el valor esperado. $\sigma^2 = Var(x)$, siendo la varianza, y $\sigma = sd(x)$, siendo la desviación estándar. Este tipo de distribución consta de una sola dimensión: $p = 1$. $X \sim N(\mu, \sigma)$, donde los datos de entrada siguen la distribución de una normal, con dos parámetros, siendo independientes e idénticamente distribuidas (*i.i.d.*).

En el caso del input que no sea definido por una sola dimensión, es decir tenga dos o más dimensiones, obtenemos una Distribución Gaussiana Multivariante, siendo una generalización de la distribución normal unidimensional a dos o más dimensiones. Además, según el Teorema del Límite Central (TLC), la suma de variables aleatorias independientes tiende a una Gaussiana⁷.

$$y \sim N(\mu, \sigma^2); \sum_{i=1}^n y_i = N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2) \quad (3)$$

Los parámetros para una distribución multivariante se componen por un vector de medias (μ) y una matriz de varianzas (Σ), donde se expresa como: $X \sim N_p(\mu, \Sigma)$.

Vector $x = (x_1, \dots, x_p)^T$, es un vector aleatorio p-dimensional.

Vector de medias: $E(x) = \mu = (\mu_1, \dots, \mu_p)'$, donde, $E(x_i) = \mu_i$

Vector de covarianzas: $Var(x) = \Sigma$, cuya posición (i, j) es $\sigma_{i,j} = Cov(X_i, X_j)$

$$Var(x) = E[(X - \mu)(X - \mu)'] = E(XX') - \mu\mu'$$

Función de densidad de una Distribución Gaussiana Multivariante:

p-dimensional (2 o más):

$$f_X(x_1, \dots, x_p) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}, x \in \mathbb{R}^p \quad (4)$$

μ = vector real, siendo la esperanza matemática de X. $|\Sigma|$ = determinante de la matriz; si es escalar (matriz 1×1), se reduce a la ecuación de una distribución normal univariante.

⁷ Si escalamos una variable gaussiana, también es gaussiana.

La matriz $\Sigma = AA^T$ es la matriz de covarianza de los componentes X_j . Es real simétrica, semidefinida positiva de dimensión $p \times p$ (de este modo, sabemos que es invertible). En su diagonal se encuentra las varianzas de las X_j 's, mientras que alrededor de la diagonal, observamos las covarianzas entre las variables.

$$X \sim N_p(\mu, \Sigma) \rightarrow \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \quad (5)$$

Tal y como podemos observar, la definición para un GP y una distribución multivariante es relativamente similar, pero la principal diferencia es que un GP tiene un espacio y un dominio continuo (infinito-decimal), mientras que una distribución multivariante, corresponde con variables aleatorias discretas.

2.2. Regresión del Proceso Gaussiano

En un GP se asume que un proceso subyacente $f(x)$ sigue una distribución gaussiana, donde esta distribución viene definido por los parámetros de la media y covarianza.

Definidos $m(x)$ como la función de media y $k(x, x')$ como la función de covarianza. Un ejemplo matemático viene dado por Rasmussen *et al.*, (2005), el cual, toma la siguiente forma:

$$m(x) = E[f(x)], \quad (6)$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x')))] = Cov(x, x'), \quad (7)$$

$$f(x) \sim GP(m(x), k(x, x')) \quad (8)$$

Donde, para un conjunto finito de datos de entrada (inputs), la función marginal toma la siguiente expresión:

$$f(x) \sim N(m(x), k(x, x')) \quad (9)$$

La función obtenida es la distribución a priori, la cual, a medida que vayamos integrando los inputs, la regresión se adaptará a la forma y comportamiento de éste, obteniendo la distribución a posteriori.

Cuando la media toma valor igual a cero, es decir, la media de las variables aleatorias son cero, la función viene explicada enteramente por la función Kernel seleccionada a priori, de forma que cualquier función Kernel puede usarse como función de covarianza. Una función Kernel, en términos estadísticos son funciones predeterminadas que se asignan para definir la matriz de covarianzas. A continuación, entraremos en detalle sobre estas funciones Kernel⁸, pudiendo observar su relevancia para la regresión de un GP.

2.3. Funciones de Covarianza (Kernels)

Tal y como hemos comentado previamente, primero definiremos un GP a priori sobre funciones. Este vendrá definido por la selección del Kernel. Comúnmente, previo a la selección, se define que la función de la media $m(x)=0$, de este modo, solo deberíamos, por el momento, definir una función de covarianza o Kernel. Más adelante observaremos como introducir una media distinta a cero.

En términos estadísticos, un Kernel o núcleo es un método no paramétrico que nos permite estimar la función de densidad de probabilidad de una variable aleatoria, a partir de un número finito de observaciones. La estimación de densidad que obtenemos es una medida relativa de verosimilitud, de este modo, cuanto mayor sea su valor, mayor evidencia de que la observación pertenece a la distribución.

A diferencia de un histograma convencional, que agrupa por intervalos, el Kernel nos permite obtener las distribuciones continuas, una para cada observación de la muestra. La agregación de las contribuciones individuales entre el total de observaciones permite obtener una curva final que describe la distribución conjunta de la muestra. Es importante, a la hora de estimar un Kernel, la selección del ancho de banda (ℓ), es decir, que tan suavizada debe ser cada una de las distribuciones marginales. El ancho de banda (ℓ)⁹ es un parámetro que se encuentra en todas las funciones de Kernel (excepto el

⁸ En este artículo solo detallaremos los Kernels utilizados en esta actividad. En el enlace adjunto en el Anexo de este artículo se encuentra el trabajo completo que contiene una ampliación de este apartado.

⁹ Este ancho de banda controla cuanto se puede expandir la influencia de cada observación. Cuando se selecciona un ancho de banda grande, se necesitarán de un elevado número de observaciones para poder calcular la estimación en cada punto, lo que puede hacer aumentar nuestro sesgo de estimación (El sesgo de estimación es la diferencia entre el valor esperado del estimador contra el valor real de este parámetro) al definir excesiva suavidad. En el caso contrario, si se selecciona un ancho de banda estrecho, las observaciones necesarias son inferiores, lo que puede comportar en un aumento de la variabilidad y falta de suavidad.

Kernel Polinomial) y definirá la forma de este. Hay varios criterios para encontrar el valor óptimo de este parámetro, tal y como detallaremos en los siguientes apartados.

Cuando se emplea un Kernel con distribución normal, este valor ℓ es igual a la desviación típica, y los pesos asignados siguen la distribución de una normal estándar: $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$.

Es importante destacar que la función de covarianza (Kernel), define la forma y la importancia de los pesos que se asocian a cada observación para el cálculo de la estimación en un GP.

Hay algunos métodos para estimar el mejor tipo de Kernel y ancho de banda que se ajusten a nuestros datos, los comúnmente más utilizados son, en primer lugar, la validación cruzada¹⁰, aunque requiere de un mayor coste computacional, y en segundo lugar, por la máxima verosimilitud. En este trabajo, analizaremos la mejor estimación a través de la máxima verosimilitud bayesiana.

Por último, la función de densidad para un conjunto de variables aleatorias independientes e idénticamente distribuidas, su estimador de densidad de Kernel es: $\hat{f}_\ell(x) = \frac{1}{n\ell} \sum_{i=1}^n K\left(\frac{x-x_i}{\ell}\right)$, donde K es el núcleo.

Tal y como hemos observado hasta ahora, la regresión del GP nos permite obtener funciones que nos definan la función general de una regresión, pero esta última, se verá modificada en base a la selección del Kernel que utilicemos para su obtención, y más, cuando la media es igual a cero, ya que la función del input dependerá en su totalidad de la función asignada al Kernel.

2.3.1. Tipo de Kernels

Definido $x = \{x_1, \dots, x_p\}$ y el correspondiente output $f(x)$: $cov(f(x), f(x')) = k(x, x')$, siendo una matriz simétrica definida positiva. A continuación, visualizaremos algunos de los kernels más comúnmente usados:

¹⁰ La validación cruzada consiste en una técnica que evalúa los resultados de análisis estadísticos que garanticen independencia entre los datos de entrenamiento (training) y los datos de prueba (test), en base a varios escenarios planteados (múltiples iteraciones).

Kernel Exponencial Cuadrado (SE) o (RBF):

$$k_{SE}(x, x') = \sigma^2 \exp\left(-\frac{1}{2\ell^2} \|x - x'\|^2\right) \quad (10)$$

Donde σ^2 es la varianza (diagonal de la matriz¹¹).

Kernels Mattern [Matérn, 1960]

Los kernels de clase Matérn más comunes en ámbitos de Machine Learning son los siguientes:

$$k_{Matérn}^{1/2}(x, x') = \exp\left(-\frac{\|x - x'\|}{\ell}\right) \quad (11)$$

$$k_{Matérn}^{3/2}(x, x') = \left(1 + \frac{\sqrt{3}\|x - x'\|}{\ell}\right) \exp\left(-\frac{\sqrt{3}\|x - x'\|}{\ell}\right) \quad (12)$$

$$k_{Matérn}^{5/2}(x, x') = \left(1 + \frac{\sqrt{5}\|x - x'\|}{\ell} + \frac{5\|x - x'\|^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}\|x - x'\|}{\ell}\right) \quad (13)$$

Rational Quadratic (RQ):

$$k_{RQ}(x, x') = \sigma^2 \left(\frac{1 + \|x - x'\|^2}{2\alpha\ell^2}\right)^{-\alpha} \quad (14)$$

Donde α es un parámetro de variación de escala. Cuando α tiende a infinito, el Kernel RQ se asemeja a un Kernel SE a medida que alpha toma valores crecientes.

Donde, para el conjunto de Kernels, $\|x - x'\| = d$, siendo d la distancia euclidiana entre dos valores. Estos kernels nos dan como salida una matriz de covarianzas K de tamaño $[n, n]$, la cual, nos describe la covarianza entre dos inputs.

Sin añadir ninguna observación, los resultados serían nuestras distribuciones bayesianas a priori, basadas exclusivamente en el diseño del modelo. Es una suposición de cómo se distribuyen conjuntamente los valores de la función subyacente que tratamos de modelar. Como X_1 y X_2 son fijos, podemos decir

¹¹ Normalmente se emplea con un valor de $\sigma^2 = 1$, de este modo, el valor es constante.

que el GP a priori nos describe una distribución sobre funciones, siendo así una distribución gaussiana multivariada de p dimensiones.

A medida que se van añadiendo las observaciones de la muestra, iríamos obteniendo las distribuciones a posteriori condicionadas a estas observaciones.

Los resultados inferenciales de estas ecuaciones dependerán de los valores de los denominados hiperparámetros, por ejemplo, α y ℓ , ya comentados, los cuales, definirán el comportamiento del modelo. Uno de los métodos más utilizados en el ámbito estadístico para la obtención del valor óptimo de estos hiperparámetros es a través de las estimaciones máximas a posteriori (MAP) con algunas selecciones a priori. Es decir, la estimación óptima de los hiperparámetros puede obtenerse a través del enfoque de máxima verosimilitud.

Por último, cabe decir que la obtención de la matriz de covarianzas no se limita a estos kernels, pues también existe la posibilidad de realizar operaciones con kernels, es decir, se pueden combinar diferentes funciones de covarianza en base a la obtención de un mejor modelo. Por ejemplo:

La suma de las funciones de covarianza es una función de covarianza válida:

$$k_{new}(x, x') = \sum_i k_i(x, x') \quad (15)$$

De la misma forma, el producto de las funciones de covarianza es una función de covarianza válida:

$$k_{new}(x, x') = \prod_i k_i(x, x') \quad (16)$$

A continuación, visualizaremos el impacto de la definición de estos parámetros sobre cada una de las funciones Kernel que hemos detallado previamente. Para esta visualización de distribuciones a prior y sin introducción de ruido, definimos un vector de valores X comprendido entre -1 y 1, de longitud 200 observaciones, usando la función “*mvrnorm*¹²”, crearemos cinco distribuciones que siguen una normal multivariante, modificando los parámetros respectivos de cada función:

¹² En el capítulo de aplicabilidad empírica entraremos en detalle sobre los argumentos de esta función.

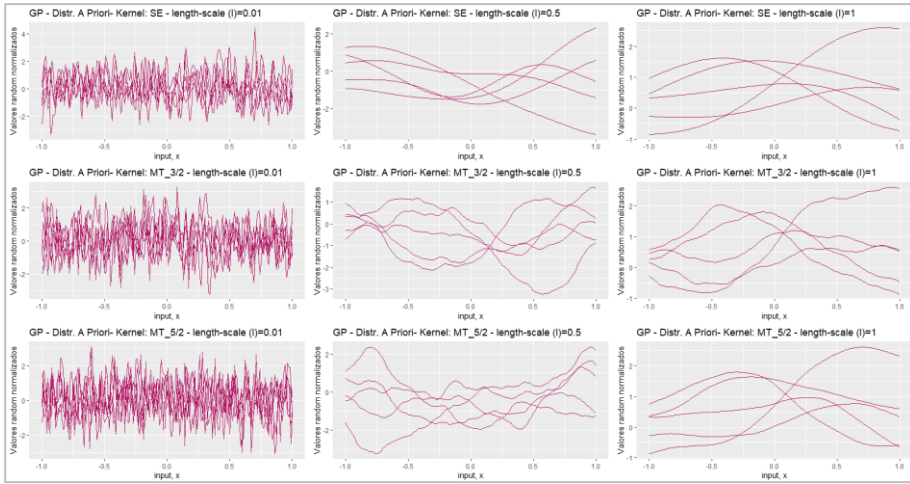


Figura 1. Distribuciones a priori de los siguientes kernels en base a la modificación del parámetro length-scale: Exponencial, Mattern 3/2 y Mattern 5/2. Elaboración propia con RStudio.

Tal y como podemos observar en la Figura 1, tanto la elección del Kernel como el valor de los parámetros de este, influirán significativamente en la función de distribución. A medida que aumentamos el valor de ℓ , observamos un comportamiento común en todos los Kernels, siendo un suavizado de nuestras distribuciones a priori. También podemos apreciar como las funciones Mattern recogen una mayor irregularidad en su continuidad (curvas) respecto el resto de las funciones.

Esta visualización gráfica previa nos permite, en un inicio y a criterio del especialista, tratar de seleccionar los Kernels que mejor se puedan adaptar a los datos de entrada, y, en base a la dimensión y tipología de estos, la selección de los valores de los parámetros óptima¹³.

2.4. Regresión del GP a Priori

Una vez tengamos definido el Kernel que vamos a utilizar para definir nuestra matriz de covarianzas, el siguiente paso es ir añadiendo la información, es decir, integrar las observaciones.

¹³ En el apartado de hiperparametrización, detallaremos otras metodologías para obtener el valor del parámetro óptimo adecuado a los inputs.

A continuación, visualizamos la ecuación que nos permitirá observar cómo se estructura cada una de las funciones que debemos definir como continuas, siempre y cuando las funciones de media y auto covarianzas sean continuas también. Podemos observar en la siguiente ecuación $[X_1, f(X_1)]$ como el par de entradas y X_2 como el conjunto que queremos realizar la predicción a N puntos.

$$\begin{bmatrix} f(X_1) \\ f(X_2) \end{bmatrix} \sim N \left(\begin{bmatrix} m(X_1) \\ m(X_2) \end{bmatrix}, \begin{pmatrix} k(X_1, X_1) & k(X_1, X_2) \\ k(X_2, X_1) & k(X_2, X_2) \end{pmatrix} \right) \quad (17)$$

X_1 es una matriz de tamaño $N \times P$, siendo N el recuento de datos de entrenamiento y P la dimensión donde se ubican estos datos. Por otro lado, X_2 es una matriz de tamaño $M \times P$, siendo M el número de ejemplos a predecir.

2.4.1. Procesos gaussianos con media igual a cero (sin ruido)

Por simplicidad de cálculo definimos $m(X_1)$ y $m(X_2)$ iguales a cero, la distribución a priori de la función de densidad gaussiana queda como:

$$p \left(\begin{bmatrix} f(X_1) \\ f(X_2) \end{bmatrix} \right) = \frac{1}{(2\pi)^{\frac{n_1+n_2}{2}} \det(k)^{1/2}} \exp \left(-\frac{1}{2} \begin{bmatrix} f(X_1) \\ f(X_2) \end{bmatrix} K^{-1} \begin{bmatrix} f(X_1) \\ f(X_2) \end{bmatrix} \right) \quad (18)$$

Siendo K^{-1} la inversa de la matriz. Por otro lado, la distribución a posteriori del proceso gaussiano $f(X_2)$ es:

$$f(X_2) | f(X_1), X_1, X_2 \sim N(k(X_2, X_1)k(X_1, X_1)^{-1}f(X_1)) \quad (19)$$

$$k(X_2, X_2) - k(X_2, X_1)k(X_1, X_1)^{-1}k(X_1, X_2)) \quad (20)$$

2.4.2. Procesos Gaussianos con media distinta a cero (con ruido)

Hasta ahora hemos tratado el GP siendo la media cero, aunque esta suposición no siempre se cumple, con lo cual, realizaremos un breve desarrollo para esas situaciones donde la media realmente no es cero.

La integración de ruido en un GP no trata de modificar la ecuación, sino que, partiendo de la ecuación resultante para un modelo sin ruido, se añadirán unos parámetros bajo la suposición que la media no es igual a cero.

Al igual que en una regresión lineal, $y = f(x) + \varepsilon$, añadimos a nuestra ecuación del GP ruido independiente, con varianza σ_n^2 , nos quedaría la siguiente ecuación:

$$f(x) \sim GP(m(x), k(x, x')) + \sigma_n^2 \delta_{xx'} \quad (21)$$

Donde $\delta_{xx'}$ es un valor delta de Kronecker que toma valor cero si $x \neq x'$, y valor uno en caso de que $x = x'$. Nótese $m(x) \neq 0$, para valores de x .

Ecuación a Priori:

$$\begin{aligned} cov(y_1, y_2) &= k(X_1, X_2) + \sigma_n^2 \delta_{xx'}, \\ \text{o igualmente, } cov(y) &= k(X_1, X_2) + \sigma_n^2 I_N \end{aligned} \quad (22)$$

$$\begin{bmatrix} f(X_1) \\ f(X_2) \end{bmatrix} \sim N \left(\begin{bmatrix} m(X_1) \\ m(X_2) \end{bmatrix}, \begin{pmatrix} k(X_1, X_1) + \sigma_n^2 I_N & k(X_1, X_2) \\ k(X_2, X_1) & k(X_2, X_2) \end{pmatrix} \right) \quad (23)$$

Ecuación a Posteriori:

$$f(X_2) | f(X_1), X_1, X_2 \sim N(k(X_2, X_1)(k(X_1, X_1) + \sigma_n^2 I_N)^{-1} f(X_1)) \quad (24)$$

$$k(X_2, X_2) - k(X_2, X_1)(k(X_1, X_1) + \sigma_n^2 I_N)^{-1} k(X_1, X_2) \quad (25)$$

Observamos que al introducir solo los términos adicionales de suposición de existencia de ruido, las matrices de covarianza solo se modifican al añadir estos parámetros, pero no la estructura en sí. Por ejemplo, para el Kernel exponencial cuadrático quedaría de la siguiente forma:

$$k_{SE}(x, x') = \exp\left(\frac{-1}{2\ell^2} \|x - x'\|^2\right) + \sigma_n^2 \delta_{xx'} \quad (26)$$

Una vez definida la ecuación posterior, ya podemos obtener la media posterior y la covarianza.

2.5. Predicción de los Procesos Gaussianos

La predicción en los GP se basa en dos partes. En primer lugar, obtener el valor esperado de la media, y en segundo lugar, en base a la matriz de covarianza asignada (Kernel).

En primer lugar, la media posterior es un vector de longitud n , mientras que la matriz de covarianzas es una matriz de tamaño $n \times n$, donde, dado el GP a priori y una verosimilitud gaussiana, la función posterior también es un GP. A continuación, podemos observar la ecuación de la obtención de las funciones de predicción cuando los hiperparámetros son fijos.

$$m_{2|1} = m(X_2) + k(X_2, X_1)(k(X_1, X_1) + \sigma_n^2 I_N)^{-1}(f(X_1) - m(X_1)) \quad (27)$$

$$\Sigma_{2|1} = k(X_2, X_2) - k(X_2, X_1)(k(X_1, X_1) + \sigma_n^2 I_N)^{-1}k(X_2, X_1)^T \quad (28)$$

Las anteriores expresiones¹⁴ no tienen aplicación hasta que previamente hayamos encontrado, el valor óptimo de los hiperparámetros, sujeto a un conjunto de observaciones y dada una función de covarianza previamente asignada, ya podríamos obtener la densidad posterior.

Tomar en consideración que en el caso que la estimación sea $m(x)=0$, tan solo haría falta extraer de las ecuaciones el término $\sigma_n^2 I_N$.

Agrupando las dos últimas ecuaciones definidas previamente, la distribución de predicción final, con integración de ruido, quedaría como:

$$Y(x_*) \sim N(m_*(x_*), \Sigma_{2|1}(x_*) + \sigma^2(x_*)) \quad (29)$$

Donde $Y(x_*)$ podría definirse como nuestra función final.

Posteriormente, entraremos en el proceso de entrenamiento de nuestro GP, es decir, el procedimiento para la obtención de los hiperparámetros (parámetros del Kernel y la varianza del ruido si precede).

¹⁴ Nótese como la estimación de la media posterior es una combinación lineal de las observaciones de la función $f(x)$, mientras que la varianza es independiente de las observaciones de la función $f(x)$.

A continuación, realizaremos un pequeño ejercicio donde puede visualizarse la obtención de la distribución a posteriori a medida que vamos introduciendo nuestros inputs en la regresión del GP. Para este ejemplo usaremos un Kernel Exponencial, aplicado a los datos utilizados para la obtención de la Figura 1. Ahora, en vez de crear 5 distribuciones, crearemos 100, para obtener un intervalo de confianza más ajustado entre punto y punto y poder recoger una mayor variabilidad de la varianza. Como hemos visto hasta ahora, nuestra función a prior dependía enteramente solo de nuestros valores de X, el siguiente paso será ir introduciendo en nuestra regresión los valores de Y observados. Por ejemplo, queremos predecir los valores comprendidos entre $X = -1$ y 1 , y nuestros valores observados (*train*) son los siguientes: $x_i = (-0.81, -0.58, -0.2, 0.33, 0.51, 0.9)$, por otro lado, nuestros valores conocidos de Y son los siguientes: $y_i = (3.21, 1.8, 0.2, -1.9, 1.75, 2.1)$. Por último, hemos definido un ancho de banda de 0.1 y un valor para la integración de ruido de 0.2.

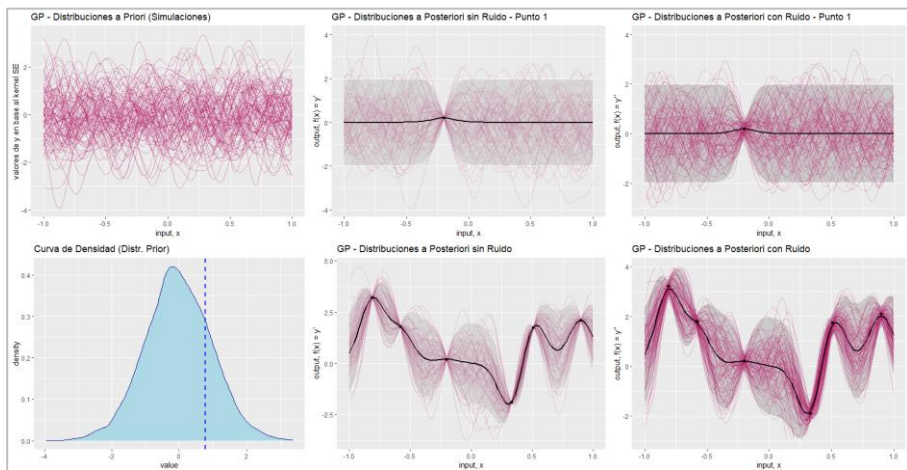


Figura 2. Visualización de una regresión por proceso gaussiano, con Kernel exponencial, ancho de banda definido igual a 0.1 y ruido 0.2. El intervalo de confianza asignado tiene un nivel de confianza asignado del 95% (bandas grises). Elaboración propia con RStudio.

En la Figura 2 podemos observar, en primer lugar, la distribución a priori, preparada a través de la realización de cien distribuciones que siguen una normal multivariante que depende completamente del Kernel exponencial. Las dos últimas figuras de la primera fila se visualizan como se adapta la regresión a medida que integramos puntos observados de nuestro input, en este caso, perteneciente al $x_3 = -0.2$ y a $y_3 = 0.2$, con y sin ruido en la

regresión. A medida que vayamos integrando los puntos, nuestra distribución a posteriori se irá adaptando paulatinamente a las variables observadas, definiendo también los intervalos de confianza. La primera figura de la segunda fila es un histograma de nuestra distribución a priori, pudiendo observar el punto medio y cuáles son esos valores que más se repiten, reflejándose que, con los parámetros y variables definidas, nuestra regresión tiende a cero. La segunda figura de la fila dos es la regresión del GP sin ruido con toda la muestra introducida, por otro lado, la última figura es la regresión del GP con ruido.

La primera diferencia destacable que podemos observar es que, cuando la media no es cero, la regresión no estrangula tanto los puntos, (efecto cuello de botella) y se ensancha el margen de actuación de la media¹⁵. Por otro lado, podemos observar que, como la media tiende a cero en los tramos con pocas observaciones, la regresión puede tomar valores con una tendencia que no sea correcta entre los puntos de la muestra. Por este motivo, es importante realizar un proceso de entrenamiento del modelo. Una de las metodologías para evitar que suceda esta casuística es definir una función para nuestros parámetros, evitando que tomen valores constantes a lo largo del período de muestra.¹⁶ Otra posible solución es con la integración de ruido, que permitirá al modelo una mayor flexibilidad de adaptación. Cabe decir que este comportamiento vendrá condicionado a la distancia entre los puntos de la muestra, es decir, si se observa una alta varianza entre dos puntos muy cercanos, cuando previamente no se había observado este comportamiento a lo largo del camino, en los siguientes puntos de la muestra el modelo estima con esta alta volatilidad, forzando la regresión a fuertes caídas y subidas.

2.6. Hiperparametrización

La correcta elección de los valores de los parámetros nos permitirá explicar mejor los datos de entrenamiento. Para la obtención de estos hiperparámetros óptimos entramos en un proceso de aprendizaje de parámetros, en el cual, se irán modificando sus parámetros hasta alcanzar un modelo que se adapte lo suficientemente bien para explicar los datos de entrenamiento.

¹⁵ Podemos observar que la integración con ruido permite un mayor margen de maniobra en el valor esperado de la media, siendo un método indicado para esas regresiones que requieren una mayor flexibilidad tanto en observaciones conocidas como para predecir.

¹⁶ En el capítulo de aplicabilidad empírica, observaremos como llevar a cabo este proceso.

Algunos de los hiperparámetros observados hasta ahora son σ_n^2 o ℓ , los cuales agruparemos como θ para facilitar la metodología explicativa de estos hiperparámetros.

La obtención de esta optimización puede obtenerse por varios criterios en los GP, el método que trataremos en esta sección es por selección bayesiana.

2.6.1. Verosimilitud de la selección bayesiana

En términos generales, el criterio de máxima verosimilitud (MV) trata de encontrar el modelo que, con una mayor probabilidad, haya generado la máxima similitud sobre los datos observados. Es un modelo generativo para series de tiempo.

Para nuestro caso, un modelo generativo es cualquiera que sea capaz de modelar la variable output como una variable aleatoria Y a través de una distribución condicionada a los inputs de entrada X y a los parámetros θ .

Tal y como hemos observado previamente, la regla genérica de bayes para realizar inferencia bayesiana es:

$$posterior = \frac{verosimilitud * prior}{verosimilitud marginal} \quad (30)$$

Como la obtención de la verosimilitud marginal puede implicar una integral de difícil resolución, la metodología tiene a derivarse por aproximación de máxima verosimilitud. Para los GP de media cero, la verosimilitud de los parámetros toman la siguiente forma.

En primer lugar, si observamos la probabilidad marginal $p(y|X, \theta, \sigma)$:

$$p(y|X, \theta, \sigma) = \int p(Y|f, X, \theta, \sigma)p(f|X, \theta, \sigma)df \quad (31)$$

$$p(y|X, \theta, \sigma) = \frac{1}{(2\pi|K_y|)^{n/2}} \exp\left(-\frac{1}{2}(y - m)^T K_y^{-1}(y - m)\right) \quad (32)$$

Donde $m = m(X)$ y $K_y = K_\theta(X, X) + \sigma_n^2 I_N$.

Esta ecuación nos permitirá adaptar a una Gaussiana, aunque sea continua, no nos afectará a la hora de la resolución, ya que los datos que disponemos son

variables discretas que siguen una distribución multivariante. Cada distribución marginal de cada predicción obtenida, (*output*) es una gaussiana univariante.

Como la finalidad de este proceso es la obtención de un modelo con probabilidad máxima, maximizaremos la probabilidad marginal con respecto a los parámetros del modelo. En este caso, como la probabilidad marginal es una función exponencial, podemos tomar logaritmos para contrarrestar la exponencial y aprovecharemos que la función logarítmica aumenta estrictamente, con lo cual, el *output* obtenido por $\log(p(y|X))$ son iguales a $p(y|X)$, obteniendo el mismo valor para los parámetros en los dos casos.

A continuación, extraemos el Negative Log-Likelihood (NLL), de este modo, quedará definida nuestra función objetivo al minimizar la log-verosimilitud que equivale a maximizar la exponencial.

$$NLL = -\log P(Y|X, \theta) \quad (33)$$

$$NLL = -\frac{1}{2} \log(|K_y|) - \frac{1}{2} (y - m)^T K_y^{-1} (y - m) + \frac{n}{2} \log(2\pi) \quad (34)$$

En orden lineal, la primera parte de la ecuación (34) nos definirá en qué sentido está orientada la función, la cual, penaliza la complejidad¹⁷. Por otro lado, la parte intermedia de la ecuación nos indica la distancia orientada de que la media, es similar a la distancia euclidiana para contemplar el ajuste de los datos. Es otro término de penalización que aumentará proporcionalmente ante la complejidad de un modelo y es la única que depende de las observaciones de Y. En último lugar, es un término constante de normalización, que no modifica el valor del resultado. De este modo, a través de esta ecuación, ya podríamos tratar de encontrar cuál es la función que tiene un mayor comportamiento similar a la media de todos los puntos. Podremos observar que a mayor ancho de banda (ℓ), menos complejas son las funciones, son más suaves, con lo cual, también se vuelven más rígidas, provocando que no se ajuste correctamente a los datos si estos son muy dispersos. En resumen, la ecuación da un mayor peso a las funciones que se ajusten correctamente a los datos sin ser tan complejas.

Para el desarrollo empírico de nuestra regresión según el GP, seguiremos la misma metodología que (Flaxman *et al.*, 2015) es decir, la inferencia

¹⁷ Es un término de penalización por el error de un modelo al ajustar los valores observados que depende solamente del Kernel y las entradas.

bayesiana, el cual, nos permite definir una distribución conjunta de probabilidad $p(\cdot)$ para los hiperparámetros de la función de covarianza:

$$\theta \sim p(\theta) \quad (35)$$

$$f|X, \theta \sim N_n(M(X), K_\theta(X, X)) \quad (36)$$

Donde $K_\theta(X, X)$ es una función del input de nuestros datos y de los hiperparámetros. En vez de tomar un valor fijo o constante, los hiperparámetros de la función de covarianza vendrán definidos por una distribución, de la familia gaussiana, por ejemplo: exponencial, gamma, log-normal, etc.¹⁸

A continuación, realizaremos una breve introducción metodológica sobre las dos aplicaciones prácticas realizadas. En primer lugar, los seguros de tarificación anual renovable (TAR) en Vida, para el escenario que nos es de interés, y, en segundo lugar, de la estructura y algunas técnicas para el cálculo de provisiones técnicas (PT) en seguro No-vida.

3. Aplicación a los Seguros de Vida

3.1. Tarificación Anual Renovable

Un seguro de tipo TAR en vida-riesgo¹⁹ son los seguros que ofrecen cobertura para una sola anualidad y que pueden modificarse a fecha de renovación. Al ser anuales, la prima también tiende a ser de pago anual, la cual queda sujeta a incrementarse en base a la edad del asegurado, a la siniestralidad histórica, a la revalorización de los capitales contratados y a otros índices económicos como el tipo de interés máximo o la actualización a nuevas tablas de mortalidad designados por la Dirección General de Seguros y Fondos de Pensiones (DGSFP) y el Boletín Oficial de Estado (BOE), respectivamente. Hay otros factores internos que pueden modificar la cuantía de dicha prima como los gastos internos o de administración (GGI) y los gastos externos o comerciales (GGE), así como otros factores que componen las bases técnicas (BT) de una póliza que queden sujetos a modificarse.

En nuestra propuesta empírica, nos ceñiremos al siguiente escenario. La obtención de las tasas de mortalidad se extraerá de las tablas de mortalidad

¹⁸ En el capítulo de aplicación empírica entraremos en detalle sobre las funciones de parámetros.

¹⁹ Incluyen esos seguros que dan cobertura para vida y accidentes.

PASEM2020_General_2ndo_orden²⁰, Unisex, destinada a los seguros de vida-riesgo, excluyendo cualquier cobertura por seguro de decesos. La propuesta empírica solo contemplará la garantía de fallecimiento por cualquier causa, el cual, incluye cobertura para todo tipo de fallecimiento, aunque se excluyen algunos sucesos como las casuísticas que ya recoge el Consorcio de Compensación de Seguros o el fallecimiento bajo efectos de embriaguez. Cabe decir que cada compañía difiere en algunos de los artículos del condicionado general, y adaptará las coberturas en base al perfil del asegurado en las condiciones particulares y/o especiales.

El cálculo que nos permitirá obtener las tasas de mortalidad y así la prima neta, parte de los valores de la cohorte l_x . Estos valores son reales positivos $l: [0, \omega] \rightarrow \mathbb{R}$, y representan el monto de personas supervivientes nacidos en el mismo instante, desde el nacimiento hasta la extinción de este grupo de personas Ayuso *et al.*, (2007), $l(0)$ es el tamaño inicial, donde las tablas que usaremos equivalen a 1.000.000 personas y cada l_x es el número de supervivientes en cada tramo de edad igual a x . Por otro lado, el cálculo de probabilidad de fallecimiento, q_x , para este escenario, donde $h=1$, siendo h el período de interés, el tanto anual de mortalidad se calcula a través de la siguiente ecuación:²¹

$$q_x = \frac{(l_{x-1} - l_x)}{l_{x-1}} \quad (37)$$

Esta ecuación equivale a la variación del valor de cohorte por tramo de edad. La aplicación consistirá en obtener la interpolación mensual de cada q_x anual, y donde el pago, en caso de que el suceso se dé a cabo, es a mes vencido²².

Una vez obtenidas las probabilidades de fallecimiento, a continuación, construiremos las ecuaciones para obtener la prima neta por asegurado. Definidas unas bases técnicas y un tipo de interés (I_1) anual, y para un intervalo de interpolación definido cada 12 períodos²³:

$$\text{Tasas puras} = q_x * (1 + I_1)^{1/12} \quad (38)$$

Donde $(1 + i)^{1/12}$ representa la capitalización financiera mensual.

²⁰ https://www.boe.es/diario_boe/txt.php?id=BOE-A-2020-17154

²¹ Para una mayor extensión de información, consultar Ayuso *et al.*, (2007).

²² Bajo la perspectiva de una compañía aseguradora, solo nos focalizaremos en el intervalo de edad comprendido entre 15 y 70 años, simulando la situación de que el producto viene ofrecido de una aseguradora con destinación hacia una compañía o colectivo laboral, y no a particulares.

²³ Interpolación para 12 períodos dentro de cada anualidad, obteniendo así la actualización financiera mensual. Se ha realizado la interpolación a nivel mensual por criterio y por limitación de coste computacional.

$$Tasas\ Tarifa = \frac{Tasas\ puras}{(1 - GGI - GGE)} \quad (39)$$

Donde GGI y GGE son los gastos internos y externos expresados en términos porcentuales sobre la prima.

$$Prima\ Neta = Capital * Tasas\ Tarifa \quad (40)$$

El capital es el monto acumulado que deberá abonar el asegurador al asegurado en caso de que este último fallezca.

Como podemos observar, el cálculo es sencillo²⁴ y no se han incluido cobertura o factores adicionales como un recargo estándar o por contratación de mayores coberturas, participación por beneficios, descuentos sobre garantías o grupos de riesgo, ya que tan solo nos interesa el comportamiento de q_x y no una simulación real en sí. Por este motivo, nos focalizaremos sobre la prima neta, sin entrar en detalle sobre el cálculo de la prima total, una vez se han integrado los impuestos necesarios como el del consorcio o el I.P.S.

3.2. Interpolación tablas de mortalidad

Con la finalidad de obtener nuestras nuevas tasas de mortalidad, nuestro input se comprende de dos variables. En primer lugar, tenemos la edad anual (x) y en segundo lugar, los valores de la cohorte²⁵ a cada tramo $l_x = y$, el cual, deberemos estandarizar para integrarlo correctamente en la regresión del GP. Tal y como se observa en la ecuación (41).

$$l'_i = \frac{l_i - \bar{l}}{sd(l)} \quad (41)$$

La edad anual la mantendremos en su formato original ya que no afecta en esta aplicación. Para su elaboración de esta parte de la actividad se ha creado una tabla con los datos de cinco asegurados con datos aleatorios que contiene tres variables: el identificador del asegurado, el monto de capital asegurado y la fecha de nacimiento de cada uno de estos. En este escenario, hemos definido

²⁴ En términos continuos, se tiende a tomar logaritmos sobre el tipo de interés y se multiplica por la ecuación con variables discretas: $\frac{t_h}{Ln(1+t_1)} * m/V^{(h)} A_{x:\overline{n}}$.

²⁵ En el Anexo de este trabajo se encuentra el detalle con los valores de cohorte y una tabla adicional con el detalle de los datos de los asegurados aleatorios.

las siguientes BT's: GGI=10%, GGE=10%, la fecha de efecto es el 01/05/2022 y $I_1 = 0.46\%$ ²⁶

El siguiente paso es extraer la edad actuarial de cada asegurado, calculado como la diferencia entre la fecha de efecto común y la fecha de nacimiento individual y dividido entre 365.25, de este modo, recogemos el impacto de los años bisiestos²⁷.

Procedemos a iniciar la interpolación por la regresión de un GP de nuestros datos. Previo, cabe destacar los Kernels utilizados: Kernel Exponencial, Mattern 3/2, Mattern 5/2, Racional Cuadrático. Se han descartado el Kernel periódico y el Kernel polinomial, ya que debido a la tipología de nuestros datos, estos no se ajustan.

Para la obtención de nuestras simulaciones se ha utilizado la función “*mvrnorm*”²⁸ del paquete “MASS”. Esta función nos permite crear una muestra de datos con distribución normal multivariante con parámetros μ y σ definidos, definiendo 50 caminos aleatorios que siguen una normal multivariante.²⁹ A continuación, visualizaremos las distribuciones a prior en base al tipo de Kernel seleccionado, la definición de los valores de los hiperparámetros y de las distribuciones en base si integramos o no ruido.

Cabe comentar que no se ha estimado el valor óptimo de los parámetros, pues se han seleccionado a criterio del investigador en base a la tipología de los datos de entrada y a un análisis de prueba-error. El parámetro ancho de banda seleccionado es un valor constante igual a 1, al igual que el parámetro alpha, mientras que por otro lado, el parámetro de ruido se ha definido con un valor de 0.15, también constante.

Nuestros inputs X son una secuencia de valores (edad) que van desde 0 a 120 años, los valores Y son los valores de cohorte para cada tramo de edad y la predicción será los valores de Y' que se encuentran a nivel mensual, es decir, X' es la interpolación de X mensual ($X' = X * 1/12$).

²⁶<http://www.dgsfp.mineco.es/es/Regulacion/DocumentosRegulacion/Resolución%20tipo%20de%20inter%20de%20seguros%202022.pdf>

²⁷ Normalmente, una vez obtenida la edad actuarial (para la aplicación, mantendremos los dos decimales, ya que será el indicativo para saber a qué mensualidad pertenecen) se redondean los decimales a la unidad en base a si el decimal es inferior o superior a $x.5$.

²⁸ Esta función realiza la descomposición de la matriz de covarianzas a través de los valores propios (*eigen values*), es más estable que la descomposición de Choleski, pero también más lenta computacionalmente.

²⁹ Parte del código utilizado para esta aplicación empírica se ha basado en el blog online: Boettiger, (2012).

Previo a realizar la predicción, si visualizamos los valores de cohorte podemos sacar las siguientes conclusiones: observamos que la reducción del monto de población decrece muy levemente hasta, aproximadamente 50 años. A partir de los 50, cae progresivamente hasta los 100 aproximadamente, llegando a 0 supervivientes alrededor de los 110 años. Con estos datos, podemos extraer ya la conclusión que las probabilidades de fallecimiento crecerán exponencialmente a partir de los 50 años, penalizando el valor de las primas notablemente.

Tras realizar un primer análisis para la selección del Kernel óptimo, hemos podido extraer las siguientes conclusiones: en primer lugar, seleccionamos el Kernel Exponencial como distribución a priori en base a la tipología de los datos. Se descartan los Kernel tipo Mattern debido a que presentan demasiada volatilidad entre cada punto (edad actuarial), y nos interesa obtener una media constante que enlace cada punto dividiendo entre 12 subperíodos. Por último, se descarta el Kernel racional cuadrático ya que por la tipología de los datos, este presenta, la curva cuadrática entre punto y punto, y si queremos suavizar este comportamiento debemos aumentar el valor de alpha hasta alcanzar una curva continua, pero al realizar este proceso, cada vez tomara una mayor similitud con el Kernel exponencial.³⁰

En siguiente paso es desestandarizar los valores de cohorte interpolados por GP y extraer las probabilidades de fallecimiento (q_x) para poder calcular la prima neta en base al desarrollo ya comentado en el capítulo anterior. Una vez obtenidas las nuevas q_x , se ha propuesto el siguiente tratamiento. El cálculo habitual de la obtención de las tasas de mortalidad corresponde a la variación de los valores de cohorte (ecuación 37), si esta variación es ahora mensual, obtendremos las primas mensuales, pero como el objetivo es calcular la variación de la prima anual, se propone el siguiente cálculo para obtener la prima anual en base a la edad actuarial con corte mensual:

$$q_x'' = q_x' + q_x * \left(1 - \frac{1}{12}\right) \quad (42)$$

Siendo, q_x'' la probabilidad de fallecimiento que aplicaremos sobre cada asegurado; q_x es la probabilidad de fallecimiento anual (calculada sobre los valores de cohorte originales); q_x' es la probabilidad de fallecimiento obtenida por interpolación cada doce períodos. De este modo, recogemos el peso que tiene la probabilidad de interpolación sobre la probabilidad real. Dicho de otro

³⁰ Por criterio de sostenibilidad, no se adjunta el detalle gráfico, pero puede encontrarse en el script de este trabajo.

modo, podemos introducir correctamente el cálculo para un seguro de vida, con cobertura anual (prima se abona a inicio del contrato) pero con pago vencido mensual (abono de la aseguradora al asegurador del capital asegurado en caso de fallecimiento). A continuación realizaremos un análisis gráfico para entender mejor este último concepto y observar los beneficios de esta interpolación. Véase Figura 3:

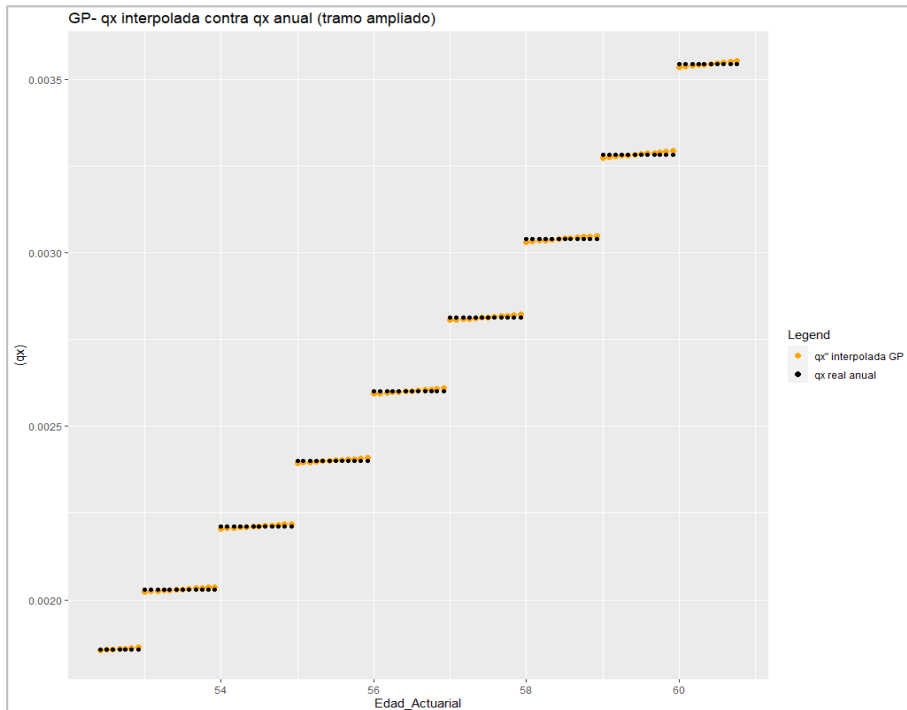


Figura 3. Comparativa entre las probabilidades de fallecimiento anuales contra las probabilidades de fallecimiento interpoladas y tratadas para el tramo de 54 a 60 años. Elaboración propia con RStudio.

En la Figura 3 observamos la diferencia entre cómo afecta, la probabilidad de fallecimiento cuando el pago es anual vencido, con actualización mensual, que tomaría un valor constante para todo el año, contra la probabilidad de fallecimiento con pago mensual vencido y actualización mensual, donde observamos que la probabilidad es creciente, encontrándose por debajo de la q_x hasta la mitad del período aproximadamente, y un valor superior al final de este, ajustando correctamente y más apropiadamente en base a la edad actuarial del asegurado.

Tabla 1

Comparativa entre las primas netas calculadas por la metodología tradicional contra la nueva metodología estimada por GP. Edad y PN son las columnas calculadas por la metodología tradicional y Edad'' y PN''(INT) por la nueva metodología

<i>id_aseg</i>	<i>Nacimiento</i>	<i>Edad Act</i>	<i>Edad ''</i>	<i>Edad</i>	<i>Capital</i>	<i>PN'' (INT)</i>	<i>PN (Real)</i>	<i>DIF PRIMAS</i>
<i>aseg_1</i>	07/09/1981	40,65	40,67	41	25.000	11,33	12,46	-1,13
<i>aseg_2</i>	01/04/1959	63,08	63,08	63	100.000	557,56	558,83	-1,27
<i>aseg_3</i>	30/01/1971	51,25	51,25	51	50.000	105,48	105,64	-0,16
<i>aseg_4</i>	15/07/1992	29,79	29,83	30	35.000	7,70	7,61	0,09
<i>aseg_5</i>	14/12/1969	52,38	52,42	52	150.000	347,99	348,08	-0,09
Total =						1.030,06	1.032,61	-2,55

Elaboración propia.

A continuación visualizaremos una tabla con los datos de los asegurados para observar esta diferencia en términos de prima neta.

En la Tabla 1 disponemos de varios contenidos: la Edad Act es la edad actuarial calculada según la fecha de efecto tomando en cuenta el año bisiesto. Por otro lado, Edad'' es el tramo de edad por interpolación al que se le aplicará la tasa, mientras que Edad es el tramo de edad por anualidad al que se le aplicará la tasa a cada asegurado, respectivamente. Observamos que ahora podemos encontrar unas tasas más personalizadas respecto la edad actuarial, donde se puede reflejar en la columna Edad''. Realmente, si observamos la columna PN(real) y PN''(INT), podemos ver que no hay una gran diferencia sobre la prima total en el acumulado final, pero si cabe decir que esta metodología se adapta levemente mejor a las necesidades del asegurado en el período de cobertura. Observamos que en los tramos de edad que se encuentran por debajo de la mitad decimal de su edad actuarial, la nueva prima neta es inferior respecto la prima total común. Por otro lado, tampoco es un patrón común para todos los tramos de edad, pues, por ejemplo, el primer asegurado, al tener una edad actuarial de 40.65, en el cálculo convencional se le aplicarían las tasas pertenecientes a la edad de 41 años, mientras que, con la nueva metodología, le pertenece el tramo de 40.67 años, el cual, supone un ahorro para el asegurado.

4. Aplicación a los Seguros de No-Vida

4.1. Triangulo de Provisiones Técnicas

Las provisiones técnicas (PT) son el importe que deben mantener las entidades aseguradoras y de reaseguro que garantizan todas las obligaciones derivadas de los contratos de seguros y que se reflejan en el balance. Estas PT's evidencian la estabilidad de la entidad frente a oscilaciones aleatorias o cíclicas de siniestralidad y posibles riesgos especiales.

En esta aplicación solo introduciremos la metodología de cálculo de las PT para los seguros de No-Vida³¹.

El método seleccionado es Chain-Ladder determinista, a través de la predicción de triángulos de provisiones.

Limitaciones: No se contemplará el cálculo de las PT con la integración de los tipos de interés ni de la inflación. Tampoco tendremos en cuenta otros elementos que la legislación de Solvencia II contempla, como importes mínimos, cálculos simplificados u opciones transitorias, por ejemplo.

Las PT, según la normativa de Solvencia, están formadas por la suma de la mejor estimación (BEL: el valor actual esperado de los flujos de caja futuros) y el margen de riesgo (margen que garantiza que el valor de las PT sea igual al importe que las entidades previamente exigirían para poder asumir y garantizar las obligaciones de los contratos). Pueden calcularse en su conjunto o por separado (defecto), en este caso, solo nos focalizaremos en el cálculo por separado de los siniestros pendientes de declarar en términos brutos.

En este apartado, analizaremos las reservas por pérdidas como un triángulo de provisiones, para el caso de los siniestros IBNR (*Incurring But Not yet Reported*)³².

³¹ No entraremos en detalle sobre la normativa estipulada el Solvencia II para simplificar los cálculos, ya que la finalidad de la actividad es observar cómo se adapta una regresión GP (con varios kernels) contra los métodos convencionales de cálculo de PT, utilizando algunos de los métodos que si requieren esta normativa.

³² Los siniestros IBNR son esos siniestros que ya han ocurrido pero la reclamación aún no se ha tramitado y su comunicación y pago se realizará en un escenario futuro, en varios periodos, con la finalidad de compensar las pérdidas sufridas con pagos por pérdidas en varios momentos. El triángulo de provisiones está compuesto por el importe agregado total de la cartera de una aseguradora de los siniestros IBNR.

El triángulo³³ se estructura por el eje horizontal, llamado año de desarrollo, el cual, equivale al retraso de los pagos de todos los siniestros de la cartera. Por otro lado, el eje vertical pertenece al año de ocurrencia de los siniestros.

Ejemplo de un triángulo de PT (Triangulo 1) para siniestros pendientes de declaración en términos brutos de un producto No-Vida, no acumulados:

Año Origen	Año Desarrollo									
	j_1	j_2	j_3	j_4	j_5	j_6	j_7	j_8	j_9	j_{10}
i_1	$c_{1,1}$	$c_{1,2}$	$c_{1,10}$
i_2	$c_{2,1}$	∴								
$i_{...}$	∴		∴							
$i_{...}$	∴								∴	
i_{10}	$c_{10,1}$									$c_{10,10}$

Figura 2. Triángulo de provisiones técnicas. Elaboración propia.

Donde:

Eje horizontal (j): Año desarrollo respecto siniestros ocurridos en año origen.

Eje vertical (i): Año origen del siniestro.

$\mathcal{D} = \{i, j | i \in I, j \in J\}$, dominio con esquema triangular.

La Diagonal (en negrita) representa: La parte izquierda, los datos reales actuales de los que disponemos; La parte derecha de la diagonal, los valores a predecir. Respecto a los valores del cuerpo del triángulo, c_{ij} : cuantía abonada en el año de desarrollo (j) respecto a los siniestros ocurridos en el año de origen (i). C_{ij} : cuantía acumulada abonada en el año de desarrollo (j) respecto a los siniestros ocurridos en el año de origen (i).³⁴

4.1.1. Chain Ladder Determinista

La metodología Chain-Ladder (CL) es uno de los métodos más usados en el ambiente actuarial para el cálculo de PT no Vida debido a su sencillez y eficacia para la predicción.

³³ Estos triángulos de provisiones también reciben la nomenclatura de triángulos *run-off* o de desarrollo.

³⁴ Sumatorio acumulado por filas.

Este método se estructura de la siguiente forma:

En primer lugar, sobre un triángulo de provisiones con valores acumulados por filas ($C_{i,j}$), se extraen los factores de desarrollo a través de la siguiente formulación:

$$F_j = \frac{\sum_{i=1}^{k+1-j} C_{i,j}}{\sum_{i=1}^{k+1-j} C_{i,j-1}} ; \forall j = 2, \dots, k. \quad (43)$$

Donde F_j será cada factor de desarrollo que usaremos a continuación. Estos factores nos describen cuál será la evolución en su conjunto de los pagos por cada par de columnas. En términos generales, la proyección se obtendrá del producto de estos factores por el valor previo por fila acumulada:

$$\tilde{C}_{i,k-i+1+h} = C_{i,k-i+1} \prod_{m=1}^h F_{k-i+1+m} \quad (44)$$

Una vez extraídas las nuevas cuantías acumuladas (última columna del triángulo), tan solo haría falta desacumular de nuevo los valores del triángulo por filas (tomar diferencias) y extraer el sumatorio por diagonales, obteniendo el valor total de provisiones a reservar para poder hacer frente las pérdidas esperadas.

4.2. Cálculo de provisiones No-Vida

En esta segunda aplicación empírica, nuestro input consta de dos variables exógenas y una variable endógena. Las variables exógenas pertenecen a los años de inicio del importe de los siniestros y el año de desarrollo de estos, mientras que la variable endógena pertenece al cúmulo del importe de estos propios siniestros. De este modo, nos quedaría un plano espacial de tres dimensiones, donde las dos variables exógenas serían la latitud y la longitud (ejes Figura 4) y las pérdidas los valores en cada momento del tiempo: Tanto los años de origen como de desarrollo son una matriz de tamaño de 10x10, dando un total de 100 observaciones. 55 observaciones de muestra y 45 observaciones a predecir.

4.2.1. Tratamiento Input (warping input beta)

Para nuestra aplicación, tal y como hemos observado hasta ahora, deberemos darle tratamiento a nuestros inputs con la finalidad que tengan una media de

cero y una desviación estándar igual a la unidad. Seguiremos la metodología empleada en Lally y Hartman, (2018), donde el año de origen y el año de desarrollo serán tratados de la siguiente forma.

Tras observar nuestros datos, podemos observar una falta de estacionariedad, con lo cual, en línea con la transformación aplicada en Snoek et al., (2014) modificamos nuestros datos de entrada con la finalidad de obtener una mayor flexibilidad en la definición de certeza previa sobre formas particulares de no estacionariedad. De este modo, el input debe ser normalizado deformado a través de una función de distribución acumulativa beta Snoek et al., (2014), obteniendo una función que si presenta estacionariedad, modificando la función de covarianza de $k(x_p, x_q)$ a $k(w(x_p), w(x_q))$: Lally y Hartman, (2018) y Ruitenberg, (2019).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (45)$$

$$\omega_j(x_{ij}^*) = \int_0^{x_{ij}^*} \frac{\mu^{\alpha_j-1} (1-\mu)^{\beta_j-1}}{B(\alpha_j, \beta_j)} d\mu \quad (46)$$

Donde x_{ij}^* es igual a x' , los inputs normalizados de los años de origen y de desarrollo, y cada ω_j es una función de distribución acumulada (CDF) comprendido entre cero y uno.

El principal beneficio de una función beta es que puede tomar diferentes formas fácilmente solo modificando los parámetros de alpha y beta, y con el input normalizado, se puede aplicar al encontrarnos en un baremo entre cero y uno, dándonos la flexibilidad necesaria a la hora de adaptar la función.

Por otro lado, los valores (volumen de pérdidas acumuladas en nuestro caso) de la variable endógena, se estandarizarán (normalizan) con la siguiente formulación:

$$Z'_i = \frac{Z_i - \bar{Z}}{sd(Z)} \quad (47)$$

Siendo \bar{Z} el promedio del vector (media) y $sd(Z)$ la desviación estándar de la muestra. Como nuestro input estará normalizado, Lally y Hartman, (2018) propone la integración del parámetro Señal del GP (η^2)³⁵.

Al igual que con la metodología de Chain-Ladder, la predicción se obtendrá a partir de las pérdidas acumuladas por año de desarrollo.

El input (X) contiene dos variables, año desarrollo y año origen, que puede definirse como un plano de dos dimensiones, donde:

$D = (\omega_{i1}, \omega_{i2}, z_i)_{i=1}^n$, ω_1 pertenece al año de origen y ω_2 al año de desarrollo.

Así pues, nos quedaría la siguiente forma para nuestro input:

$$\omega_1(x) = \text{BetaCDF}(x, \alpha_1, \beta_1) \quad (48)$$

$$\omega_2(x) = \text{BetaCDF}(x, \alpha_2, \beta_2) \quad (49)$$

El siguiente y último paso, previo a la revisión de los datos obtenidos, es la definición de la distribución que van a seguir en un inicio todos los parámetros a estimar, una vez ya hemos definido a su vez, la distribución que seguirán nuestros datos input.

4.2.2. Distribuciones Prior e Hiperparámetros

Una vez tenemos los inputs tratados, el siguiente paso sería escoger/seleccionar cuál es la distribución que deben seguir estos inputs. Lally y Hartman, (2018) y Ruitenberg, (2019) aplican algunas modelizaciones como la distribución de t_Student (con 4 grados de libertad, media cero y varianza 1), gamma (4,4; parametrizado con forma y ratio con valor 4) y log-normal (con media cero y desviación estándar 0.5). Estas distribuciones están supeditadas a ser igual o superior a cero y permiten la flexibilidad necesaria para esos valores más atípicos (outliers).

4.2.3. Funciones de Covarianza Empleadas

Siguiendo con la metodología empleada por Lally y Hartman, (2018), a continuación, visualizaremos una tabla con las funciones de covarianza seleccionadas finalmente para realizar la predicción.

³⁵ Este parámetro multiplica cada uno de los Kernels con la finalidad de evitar que la diagonal de la matriz, cuando la distancia euclídea (input de nuestro modelo) entre x_p y x_q sea igual a cero, la diagonal de la matriz tomará valor 1 y con la integración de este parámetro, da una mayor flexibilidad al modelo.

Tabla 2

Funciones de Covarianza (Kernels) empleados para la predicción de provisiones, extrayendo los hiperparámetros a estimar. (Fuente: Ruitenberg, (2019))

Función Covarianza	Ecuación
Exponencial Cuadrático (SE)	$\eta^2(1 + \sqrt{3}d)\exp(-\sqrt{3}d)$
Matérn 3/2	$\eta^2(1 + \sqrt{5}d + 5/3d^2)\exp(-\sqrt{5}d)$
Matérn 5/2	$\eta^2\exp(-d^2)$

Elaboración propia.

En la Tabla 2 observamos las funciones de covarianza, pero no de la misma estructura que hemos definido inicialmente para los Kernels en este trabajo. Esto es debido a que la Tabla 2 refleja las ecuaciones pero sin la integración de los hiperparámetros que debemos estimar. En esta Tabla 2, d es igual a la distancia euclídea y $d^2 = (x - x')^T \psi(x - x')$, donde $\psi =$ diagonal de (ψ_1, ψ_2) (reflejan las distribuciones a priori que seguirán en un inicio).

4.2.4. Metodología

La aplicación propuesta consistirá en realizar diferentes predicciones sobre un mismo triangulo de provisiones. Por un lado, las obtenidas por el método de Chain-Ladder y por el otro lado, las obtenidas por los GP's empleados, diferenciando los outputs en base a los kernels seleccionados para cada predicción. La valoración de las mejores estimaciones se realizará a través de del error cuadrático medio de predicción (RMSE) acumulado,

$$RMSE \text{ agregado} = \sqrt{\frac{(\sum_{t=1}^T \hat{Z}_t - \sum_{t=1}^T Z_t)^2}{T}} \quad (50)$$

Donde T es el número total de observaciones de la muestra, Z_t el valor agregado que desearíamos obtener y \hat{Z}_t el valor agregado de las predicciones.

La estimación se realizará con un intervalo de confianza del 95% para la distribución posterior de las muestras obtenidas a través de caminos aleatorios multivariantes, en este caso, estimados por la metodología de Cadenas de Markov de Monte Carlo (MCMC)³⁶.

³⁶ El método MCMC consiste en crear muestras autocorrelacionadas a partir de una variable aleatoria, con una densidad conocida, lo que nos permite obtener aproximaciones para la distribución posterior. Dicho de

4.2.5. Bases de Datos

La base de datos seleccionada pertenece a la librería pública de R “*casdata*”³⁷, donde refleja triángulos enteros, desde el 1988 al 1997, obteniendo una matriz de 10x10, pudiendo obtener predicciones hasta el año 2006. En esta librería hay datos de múltiples empresas aseguradoras americanas para realizar ejercicios de predicciones con datos reales observados. Para esta aplicación se ha seleccionado “*comauto*”, contiene datos de reclamaciones incurridas del ramo de automóviles, seleccionando la compañía del grupo Celina Mut (Id: 353)³⁸ y el grupo Federal Ins Co (id: 388), de este modo, podemos observar cómo se adaptan estos modelos en base a dos triángulos de provisiones diferentes.

4.2.6. Resultados Predicciones Provisiones

Tabla 3

Cálculo del IBNR y del RMSE para los modelos en cuestión, en primer lugar Chain Ladder y seguido de los tres modelos estimados por GP, para la compañía Celina Mut.

Cod.Id: 353	Obs	ChainLadder	GP-SE	GP-M.3/2	GP-M.5/2
IBNR	4.272	3.125	3.420	3.493	3.743
RMSE	-	115	85	78	53

Elaboración propia.

Tabla 4

Cálculo del IBNR y del RMSE para los modelos en cuestión, en primer lugar Chain Ladder y seguido de los tres modelos estimados por GP, para la compañía Federal Ins Co.

Cod.Id: 388	Obs	ChainLadder	GP-SE	GP-M.3/2	GP-M.5/2
IBNR	85.841	32.445	-6.693	39.291	53.706
RMSE	-	5.340	9.253	4.655	3.276

Elaboración propia.

otro modo, en base a la distribución a priori ($p(\theta)$) y el valor de la verosimilitud ($p(x|\theta)$), el método devuelve la aproximación de una distribución posterior ($p(\theta|x)$).

³⁷ En el script de este trabajo se encuentra el desarrollo para instalar esta librería y filtrar la base de datos deseada.

³⁸ Por motivos de sostenibilidad, solo visualizaremos el triángulo perteneciente a Federal Ins Co, la gráfica de Celina Mut puede encontrarse en el script de esta actividad.

En la Tabla 3 y 4 podemos observar los datos de salida estimados del cálculo del IBNR y del RMSE para los dos triángulos de provisiones estimados. En primer lugar, podemos destacar que aunque inicialmente los dos triángulos tienen comportamientos asimétricos, el modelo que mejor se ha ajustado a los datos reales es el GP estimado con un Kernel Matérn 5/2.

Por otro lado, el método Chain Ladder es el peor modelo que estima las predicciones para Celina Mut, aunque cabe decir, que para la compañía Federal Ins Co, su estimación es mejor que el GP Exponencial, el cual, ofrece un IBNR negativo, descartándolo directamente como método válido para estas dos casuísticas. Añadir que, aunque en Lally y Hartman, (2018) y Ruitenberg, (2019), el Kernel exponencial tuviese una muy buena adaptación, para nuestros datos no se ajusta debido a la tendencia individual de los años de desarrollo, pues siempre fuerza las predicciones a futuro hacia la media cero al no integrar ruido y con las distribuciones definidas a priori actuales. Aun así, podemos observar que todos los modelos retornan valores bastante alejados respecto al real observado.

La predicción sigue un comportamiento bastante constante, al igual que los datos originales, lo cual facilita la predicción sin ruido en este Kernel, por otro lado, también podemos observar que la falta de importe total se debe en su mayoría a la predicción de los dos últimos años de origen, donde la predicción se queda corta en media, aunque sí que lo recoge en sus intervalos de confianza.

Tabla 5
Cálculo del IBNR por año de origen para el grupo Federal Ins Co.

<i>Cod.Id: 388</i>	1	2	3	4	5	6	7	8	9	10
<i>Obs</i>	0	-486	-437	-110	871	1.043	870	10.834	22.058	51.198
<i>GP-M.5/2</i>	0	-842	192	126	-644	1.101	1.314	5.431	8.749	37.649

Elaboración propia.

Por último, si entramos en detalle, en la Tabla 5 podemos observar el IBNR por año de predicción, observado grandes diferencias en los primeros próximos cuatro años y a partir del octavo año en adelante, donde el GP Matérn 5/2 se queda significativamente por debajo de los valores reales observados.

5. Conclusiones

Por un lado, en seguros TAR como un ejercicio de interpolación, hemos observado parte del potencial de este GP, alcanzando el objetivo planteado

inicialmente. La interpolación ha sido exitosa aun sin la realización de la optimización de los hiperparámetros. Por otro lado, también hemos integrado el GP para el cálculo de un monto de provisiones en No Vida, para el ramo automovilístico y para el conjunto de ramos de una aseguradora, siguiendo la misma metodología aplicada en Lally y Hartman, (2018). Se ha alcanzado a observar que los diferentes tipos de GP son buenos predictores, en estas dos casuísticas y se adaptan mejor que el método convencional de Chain Ladder, reflejando su potencial. Aun así, cabe destacar la importancia de la selección del Kernel y las distribuciones a priori de los parámetros, pues se observan grandes diferencias en base a la selección de éste, destacando el Kernel Exponencial como el que mejor se ajusta en el ejercicio de interpolación y el Kernel Matérn 5/2 en el ejercicio de cálculo del IBNR.

En base a la experiencia obtenida, concluimos en que el GP tiene un gran potencial, siempre y cuando se haya realizado un correcto proceso de entrenamiento, destacando su fuerza en términos de interpolación pero observando debilidades en predicciones a futuro, al menos, con las distribuciones a priori que hemos definido en este ejercicio.

Por último, concluir que la integración de un GP, sin el uso de librerías específicas³⁹ tiene un alto coste computacional y nos hemos visto frente algunas limitaciones. Por ejemplo, el ejercicio de interpolación mensual, en un inicio se buscaba la actualización diaria, pero debido a estos problemas, se ha reducido al tratamiento mensual. Para solucionar este problema, puede realizarse un GP para cada par de puntos de la muestra, es decir, en vez de extraer la continuidad de toda la regresión, puede realizarse por tramos, permitiendo optimizar este coste computacional.

6. Bibliografía

- Ayuso. M., Corrales. H, Guillén. M, Pérez-Marín. A. M, Rojo. J. L. (2007). *Estadística actuarial vida*. Universitat de Barcelona.
- Boettiger, C. (2012). *Basic regression in Gaussian processes*.
<https://www.carlboettiger.info/2012/10/17/basic-regression-in-gaussian-processes.html>
- Flaxman, S., Gelman, A., Neill, D. B., Smola, A., & Vehtari, A. (2015). *Fast hierarchical Gaussian processes*. (Unpublished manuscript).

³⁹ En este ejercicio hemos tratado de usar el mínimo de librerías posibles y tratar de desarrollar el GP por pasos, pero en el entorno de R, existen varias librerías que nos permitirían obtener una resolución más rápida y/o eficiente. Algunas de las más relevantes: “GauPro”, “GPfit”, “varycoef” y “GPFDA”. También comentar que en Python hay todo un desarrollo avanzado sobre GP’s.

- Lally, N. y Hartman, B. (2018). Estimating loss reserves using hierarchical Bayesian Gaussian process regression with input warping. *Insurance: Mathematics and Economics*, 82, 124-140.
- Rasmussen, C. E., Williams, C. K. I. y Bach, F. (2005). *Gaussian Processes for Machine Learning*. MIT Press.
- Ruitenbergh, P. (2019). *Adapting a Hierarchical Gaussian Process model to predict the loss reserve of a non-life insurer*. University of Twente.
- Snoek, J., Swersky, K., Zemel, R. S., & Adams, R. P. (2014). *Input Warping for Bayesian Optimization of Non-Stationary Functions*. ICML.

Anexos

El código utilizado para la realización de esta actividad se encuentra en el enlace adjunto a continuación. En el enlace también encontrarán el trabajo completo de este artículo, con apartados adicionales, un mayor detalle del análisis y bibliografías que no han sido incluidas en el artículo, pero si han sido relevantes para la comprensión del trabajo. También incluye algunos comentarios relevantes para esos datos que no se han incluido en este artículo, pero que sí pueden considerarse de interés.

- Script: https://github.com/drius1996/TFM_GPR