Una breve introducción a los procesos Gaussianos para el modelado de series temporales

Joaquin Cavieres
Dr. en Estadística

Overview

Sobre mi..

Joaquin Cavieres, Doctor en Estadística, Universidad de Valparaíso (Chile).

Figura 1: Day

Figura 2: Night

Sobre mi..

Joaquin Cavieres, Doctor en Estadística, Universidad de Valparaíso (Chile).

Figura 1: Day

Figura 2: Night

Tesis de Doctorado: Computational methods for a smoothing thin plate spline in spatial models.

- Efficient estimation for a smoothing thin plate spline in a two-dimensional space (In press)
- Thin plate spline model under skew-normal random errors: estimation and diagnostic analysis for spatial data ▸ Link

Áreas de investigación

1. Modelado estadístico y Estadística espacial
   - GLM, GLMM and GAM.
   - Método SPDE, Funciones de Base Radial (RBF, en inglés)

2. Inferncia Bayesiana y métodos computacionales
   - Aproximación de Laplace, diferenciación automática
   - Método MCMC (por ejemplo, Hamiltonian Monte Carlo)
   - Factorización de matrices (por ejemplo, $\mathcal{H}$-matrix)

Softwares estadísticos

- `TMB`, `Stan`, `INLA`, `Rcpp/RcppArmadillo`

Actualmente soy un investigador postdoctoral en el grupo de Geoinformática en la Universidad de Bayreuth, Alemania (▸ Link)

- Preferential sampling for spatial modelling.

Más información en mi página web: ▸ Link

Motivación

Las series temporales son conjuntos de datos secuenciales que se registran en intervalos regulares a lo largo del tiempo. Estos datos pueden representar cualquier tipo de variable, por ejemplo: valor de acciones, temperaturas, ventas mensuales, niveles de contaminación, entre otros.

El análisis de series temporales tiene una gran importancia debido a su gran capacidad para analizar patrones y tendencias en el tiempo. De esta manera, podemos descubrir la evolución de un fenómeno, identificar estacionalidades, detectar cambios abruptos y predecir futuros valores.

Por ejemplo, en el ámbito financiero, el análisis de series temporales permite pronosticar el comportamiento de los precios de los activos y ayudar en la toma de decisiones de inversión. En el campo de la economía, las series temporales se utilizan para medir el crecimiento económico, el desempleo y otros indicadores macroeconómicos.

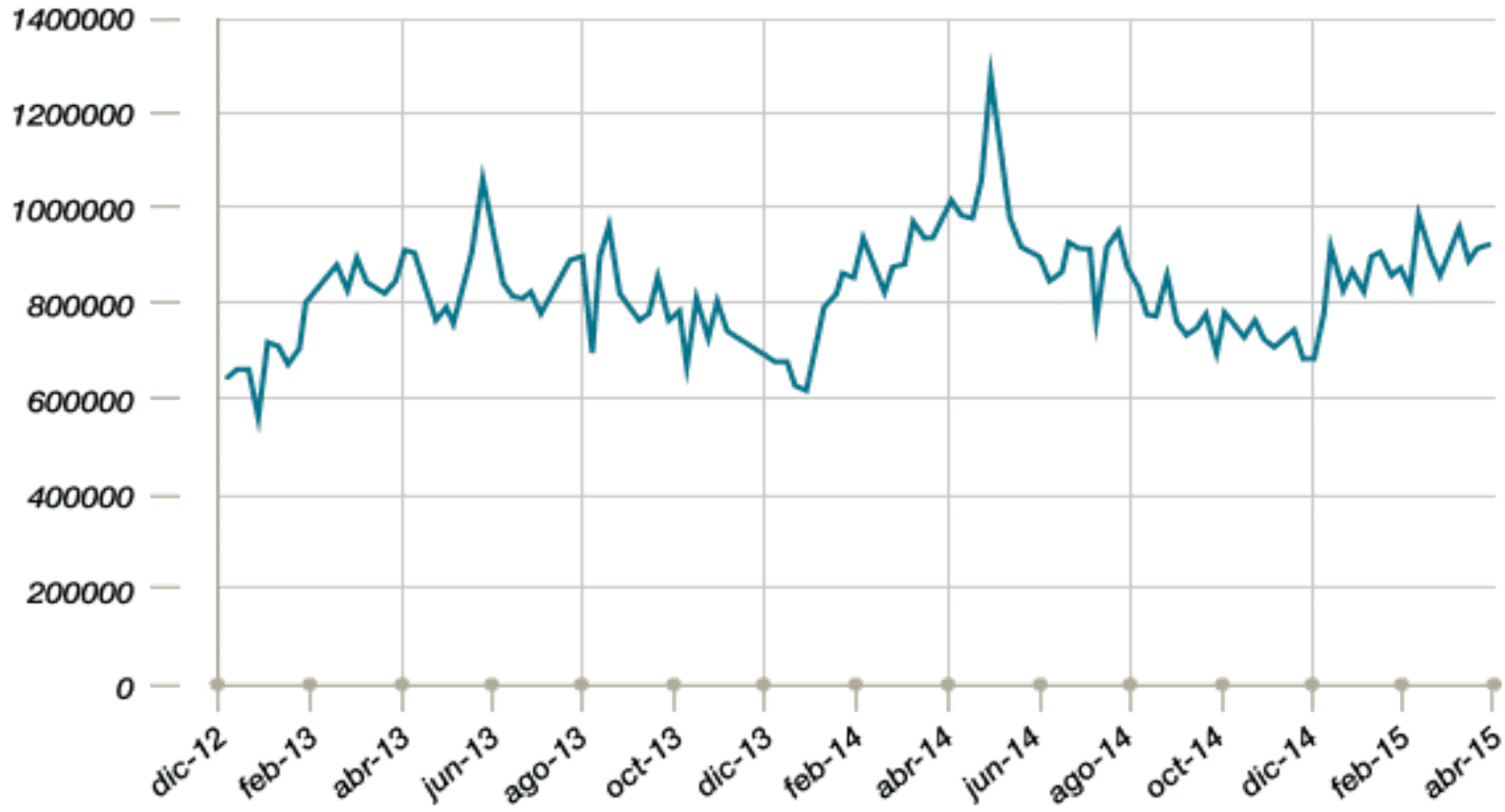

Figura 3: Serie de tiempo de las ventas de un producto en particular (Source: `https://www.pricing.cl/conocimiento/series-de-tiempo/`).

Figura 4: Precio de cierre para las acciones de la compañía AXP American Express (2003 2013) (Source: `https://blogs.sap.com/2013/11/27/brief-stock-market-time-series-data-analysis/`).

Otros campos donde las series temporales permiten realizar diversos análisis:

- Meteorología: predecir el clima y los patrones climáticos, lo que ayuda en la planificación de actividades agrícolas, el diseño de sistemas de alerta temprana y la gestión de desastres naturales.
- Medicina: monitorear y predecir la propagación de enfermedades, poder analizar la eficacia de los tratamientos y estudiar la evolución de los síntomas a lo largo del tiempo.

Para modelar series temporales se requieren modelos adecuados para comprender y explotar la información contenida en estas series. Así,

Los modelos matemáticos/estadísticos aplicados en series temporales nos permiten capturar tendencias y patrones temporales, generar pronósticos precisos, tomar decisiones informadas, identificar anomalías y evaluar políticas y estrategias respecto al fenómeno evualado.

Para modelar series temporales se requieren modelos adecuados para comprender y explotar la información contenida en estas series. Así,

Los modelos matemáticos/estadísticos aplicados en series temporales nos permiten capturar tendencias y patrones temporales, generar pronósticos precisos, tomar decisiones informadas, identificar anomalías y evaluar políticas y estrategias respecto al fenómeno evualado.

Tipos de modelos para series temporales?

- Media movil (Moving Average, MA)
- Autoregresivo (Autoregressive, AR)
- Suavizado exponencial (Exponential Smoothing)
- etc...

Tipos de modelos para series temporales?

- Media movil (Moving Average, MA)
- Autoregresivo (Autoregressive, AR)
- Suavizado exponencial (Exponential Smoothing)
- etc...

También existen variantes más avanzadas de estos modelos, así como otros modelos más complejos, como los procesos autorregresivos de media móvil (ARMA) y los GP en series temporales.

Procesos Gaussianos (GP)

Definición general

Un GP en series temporales es un tipo de modelo estadístico utilizado para describir y predecir datos secuenciales en el tiempo. Se basa en la suposición de que los valores de la serie temporal siguen una distribución gaussiana (o normal).

En un GP la distribución conjunta de cualquier número finito de observaciones en diferentes momentos de tiempo se puede describir completamente mediante su media y su matriz de covarianza.

La propiedad clave de un GP es que la distribución de cualquier valor futuro dado los valores pasados y presentes también es gaussiana.

Un poco de teoría..

1. Distribución gaussiana multivariada

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{1}$$

Donde:

- $\boldsymbol{x}$ es el vector de valores observados.
- $\boldsymbol{\mu}$ es el vector de medias.
- $\mathbf{\Sigma}$ es la matriz de covarianza.

El vector de medias ($\boldsymbol{\mu}$) representa las medias esperadas de las variables aleatorias en cada punto de tiempo, mientras que la matriz de covarianza ($\boldsymbol{\Sigma}$) describe las relaciones de covarianza entre las variables aleatorias en diferentes puntos de tiempo.

2 Distribución condicional gaussiana

$$p(\mathbf{x}_t|\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \tag{2}$$

Donde:

- $\boldsymbol{x}_t$ es el valor en el tiempo t.
- $\boldsymbol{\mu}_t$ es el vector de medias condicionales.
- $\boldsymbol{\Sigma}_t$ es la matriz de covarianza condicional.

La distribución condicional nos permite obtener una descripción completa de la incertidumbre en los valores futuros, actualizar la información a medida que se obtienen nuevos datos, aprovechar las propiedades analíticas y computacionales, y adaptarse a diferentes escenarios y supuestos.

3 Predicción de un valor futuro en un GP

$$p(\mathbf{x}_{t+1}|\mathbf{x}_1, \ldots, \mathbf{x}_t) = \mathcal{N}(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}) \tag{3}$$

Donde:

- $\boldsymbol{x}_{t+1}$ es el valor futuro en el tiempo $t + 1$.
- $\boldsymbol{\mu}_{t+1}$ es el vector de medias de la predicción.
- $\boldsymbol{\Sigma}_{t+1}$ es la matriz de covarianza de la predicción.

Las ecuaciones previamente mostradas forman la base del modelado de GP en series temporales, permitiendo la caracterización probabilística de los datos observados y la predicción de valores futuros.

Modelo de regresión usando GP

El modelo de regresión utilizando un GP es un enfoque no paramétrico que nos permite realizar predicciones flexibles sin asumir una forma funcional específica para la relación subyacente.

El objetivo es estimar una función desconocida a partir de puntos de datos observados. La idea principal es tratar a la función como una variable aleatoria y modelarla como un GP.

El modelo de regresión utilizando un GP es un enfoque no paramétrico que nos permite realizar predicciones flexibles sin asumir una forma funcional específica para la relación subyacente.

El objetivo es estimar una función desconocida a partir de puntos de datos observados. La idea principal es tratar a la función como una variable aleatoria y modelarla como un GP.

El proceso de regresión involucra dos pasos principales:

- Ajuste (entrenamiento)
- Predicción

Ajuste (entrenamiento)

1. Dado (x, y), donde x es la variable de entrada e y es la variable respuesta, construimos una matriz de covarianza basada en la función de covarianza elegida y los puntos de entrada.
2. La matriz de covarianza captura la estructura de correlación entre los puntos de entrada y proporciona una medida de similitud entre diferentes observaciones

Qué tipo de función de covarianza en un GP?

Existen distintas funciones de covarianza (kernels) que sirven para calcular la matriz de covarianza. Aquí sólo nos remitiremos a utilizar la función de covarianza exponencial cuadrada:

$$k(x, x') = \alpha^2 \exp\left(\frac{-||x - x'||^2}{2\rho^2}\right) \tag{4}$$

Donde:

- α^2 es la varianza de la función.
- ρ es el parámetro de longitud de escala que controla cuán rápido cae la correlación a medida que la distancia aumenta.

Para más sobre funciones de covarianza ver: Williams and Rasmussen (2006); Deisenroth et al. (2019); Murphy (2012)

Predicción

1. Dado un nuevo punto $x^\star$, calculamos la distribución predictiva, que es una distribución gaussiana que representa los posibles valores de la respuesta $y^\star$.
2. La distribución predictiva se determina mediante la distribución previa (prior), los datos observados y la función de covarianza.

Modelo de regresión utilizando GP

Modelamos una función subyacente como un GP en lugar de asumir una relación lineal entre x e y.

$$y = f(x) + \epsilon \tag{5}$$

Donde:

- y es la variable respuesta.
- x es la variable de entrada.
- $f(x)$ es una función desconocida que se modela como un proceso gaussiano.
- ϵ es el término de error ($\mathcal{N} \sim (0, \sigma^2)$)

Modelo de regresión GP para serie temporal

$$y(t) = f(t) + \epsilon(t) \tag{6}$$

Donde:

- $y(t)$ es la observación de la serie temporal en el tiempo t.
- $f(t)$ es una función desconocida que se modela como un GP en el tiempo t.
- ϵ es el término de error ($\mathcal{N} \sim (0, \sigma^2)$) en el tiempo t.

Ejemplo 1

Simulamos los siguientes datos:

- n $= 200$
- Variable de entrada "Day"$\rightarrow$ t
- Simulamos un GP con dist.multivariante Normal y mat.cov

$$\Sigma = \alpha^2 \exp\left(-\phi^2 d^2\right), \quad \text{con} \quad \alpha = 2, \;\; \rho = 5$$

- $\mu = GP \sim MVN(0, \Sigma)$
- Variable respuesta $y \sim N(\mu, \sigma^2)$, con $\sigma^2 = 0{,}3$

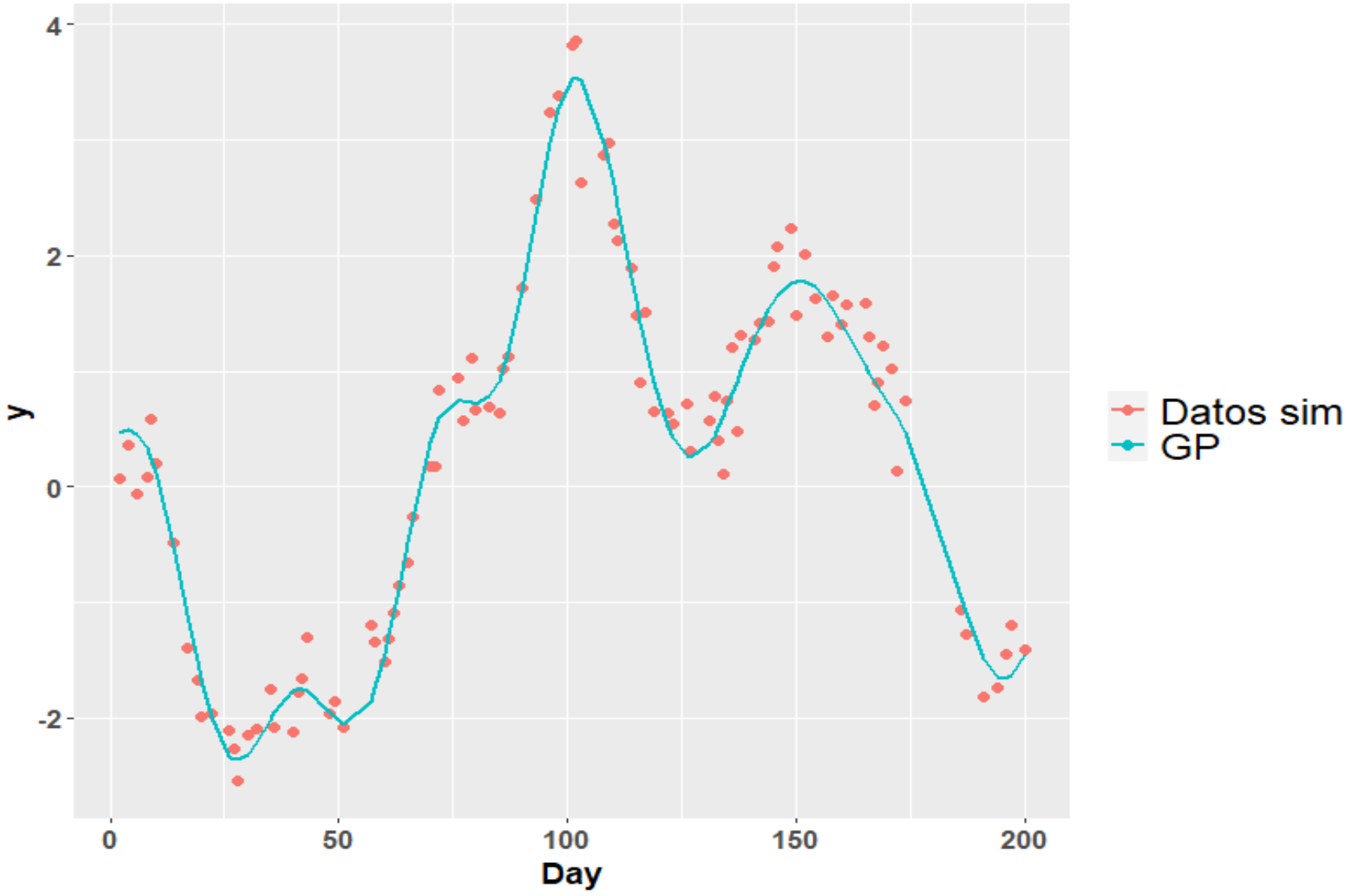

Figura 5: Datos simulados utilizando un GP

Utilizando el software `Stan` (Gelman et al. (2015); Carpenter et al. (2017)) podemos hacer inferencia mediante el método MCMC y obtener las distribuciones posteriores de los parámetros utilizados en la simulación.

Resultados

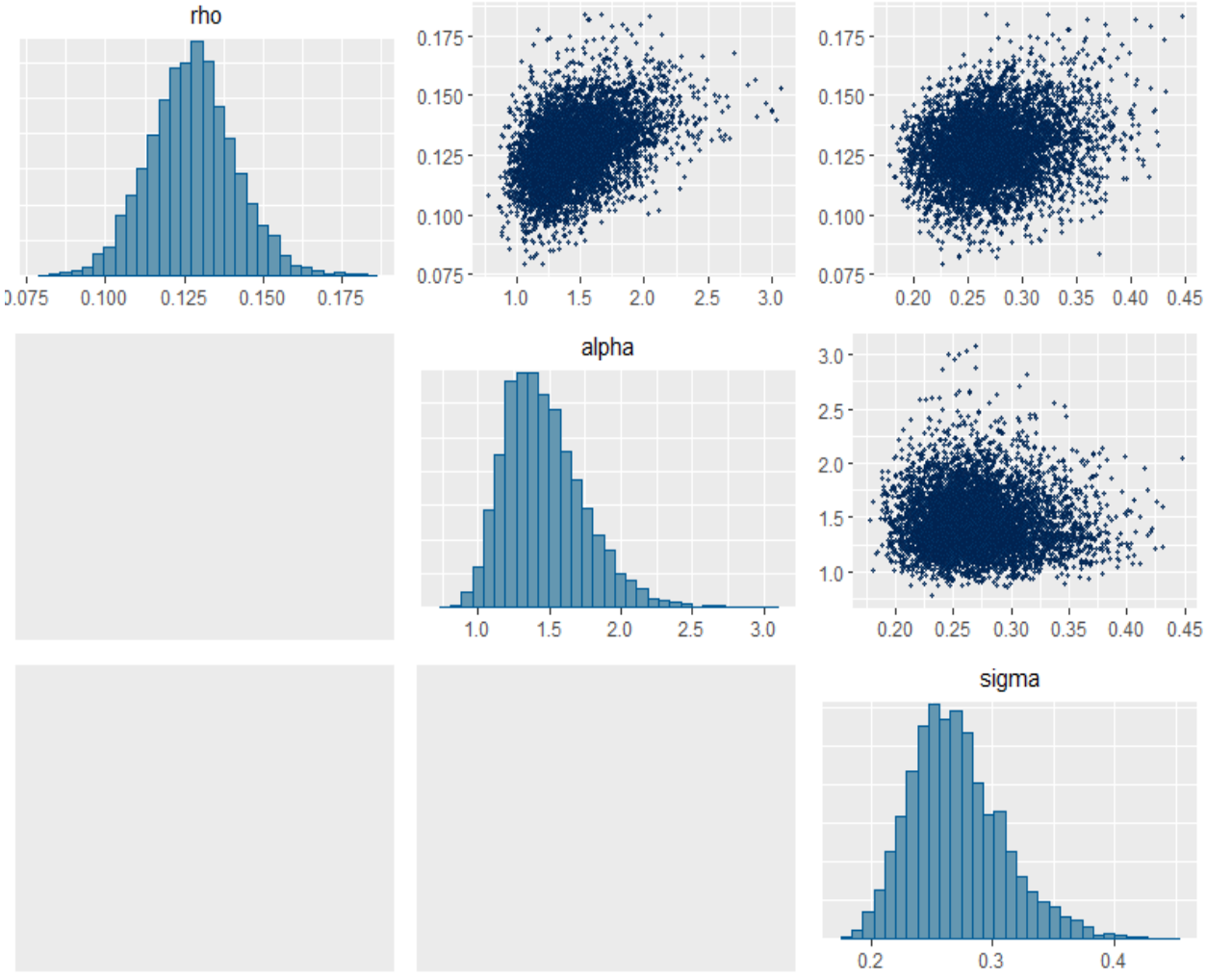

Figura 6: Distribución posterior para los parámetros del modelo

Resultados

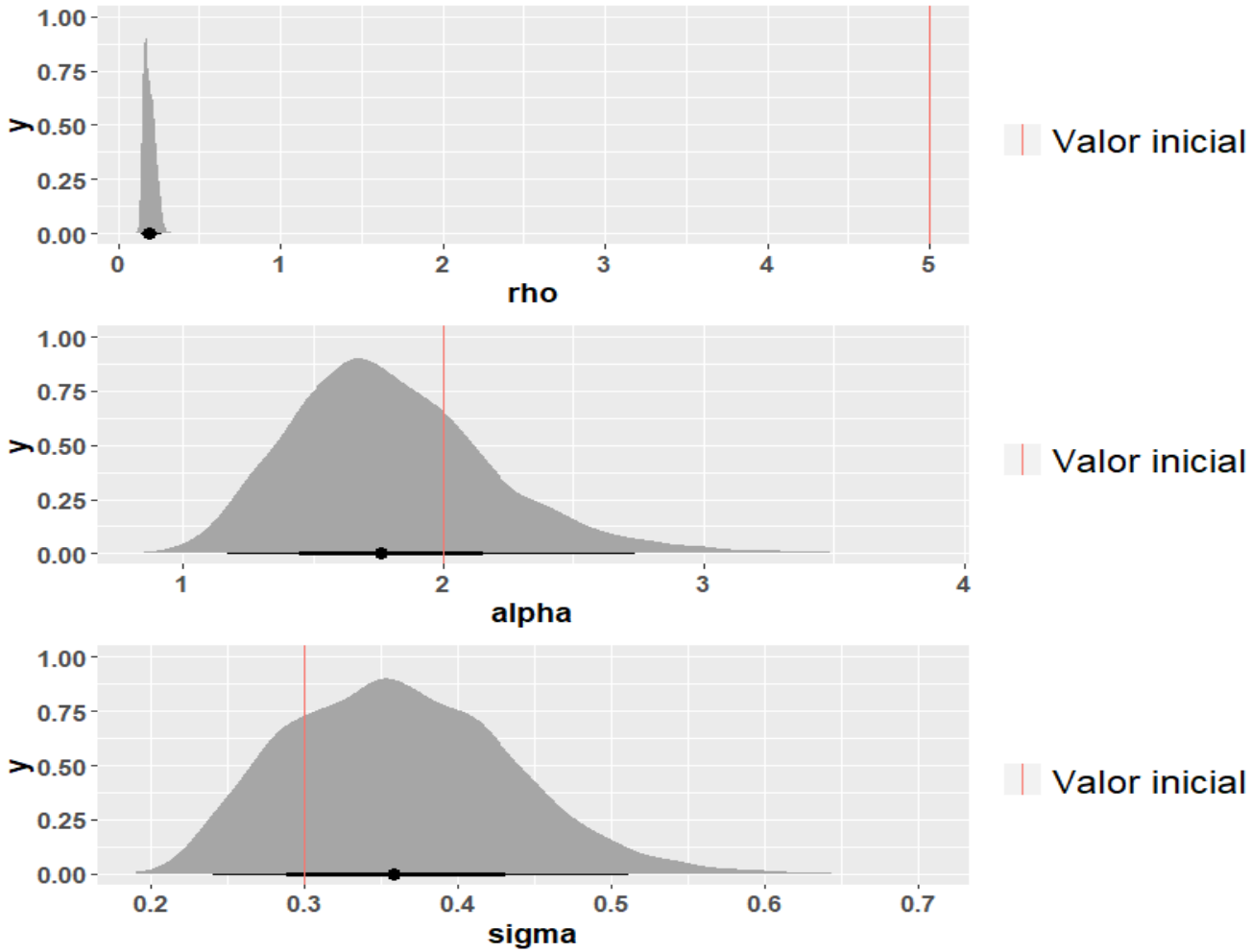

Figura 7: Distribución posterior de los parámetros del modelo

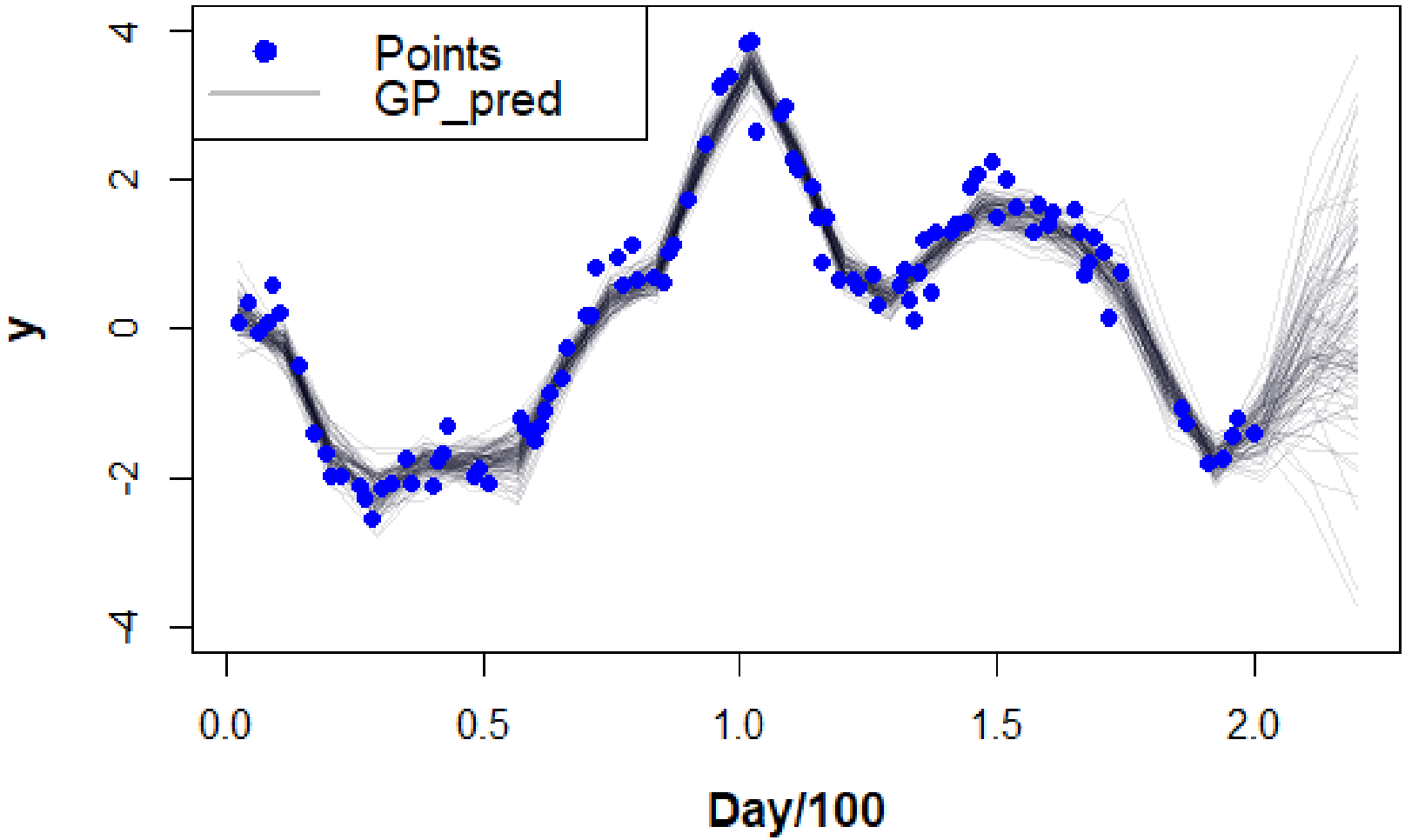

Figura 8: Predicción temporal basada en un GP

Ejemplo 2

Para este ejemplo vamos a utilizar los datos de acciones disponibles de Apple. Este conjunto de datos consta de una serie de tiempo entre el 1 de Enero 2017 hasta el 31 de Diciembre 2018.

- n = 504 (t)
- Variable respuesta = Valor de las acciones al cerrar la bolsa (`AAPL.Close`)

El objetivo es predecir dos valores en el futuro (2 días).

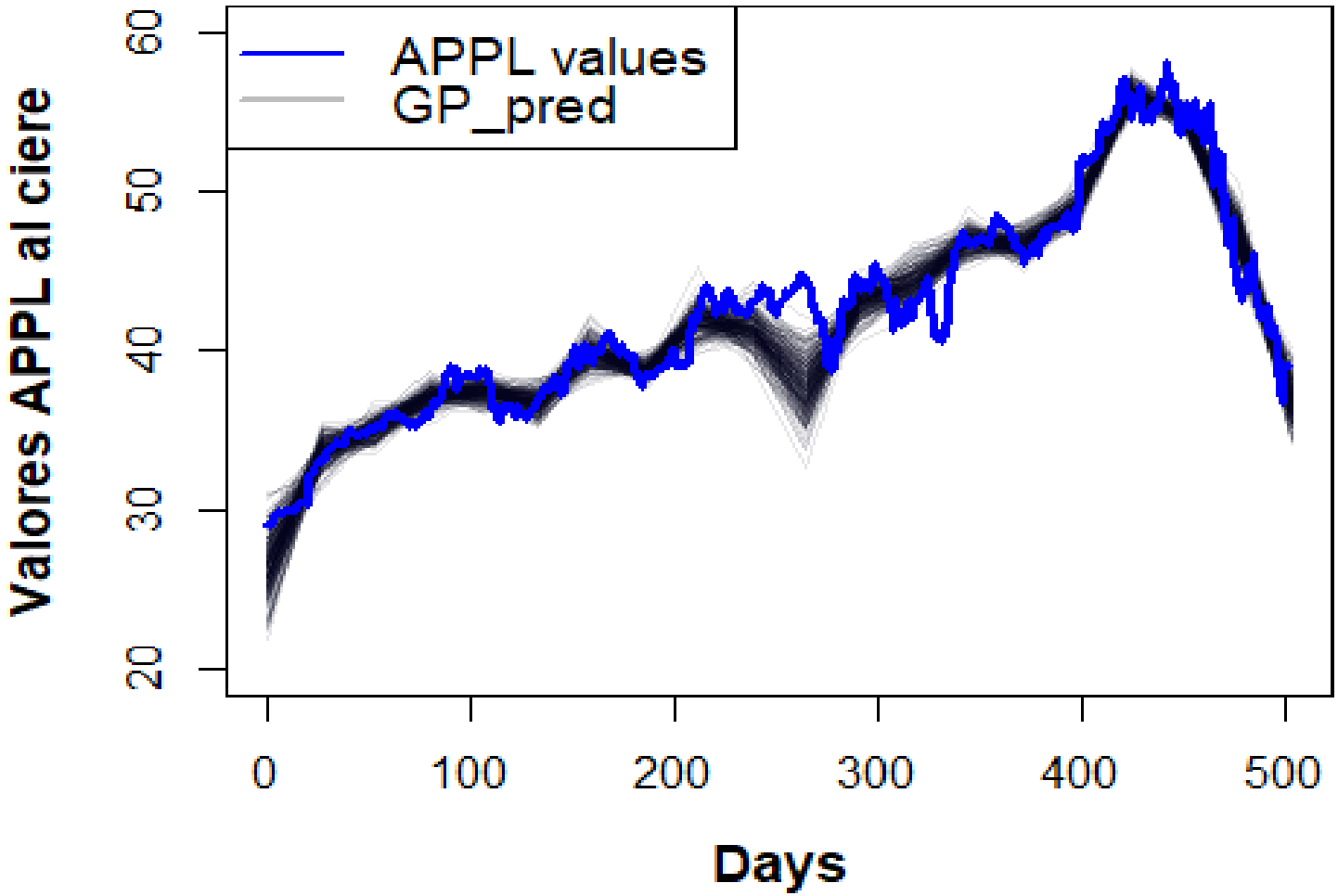

Figura 9: Predicción temporal basada en un GP para los valores de AAPL en el cierre

Conclusiones

Ventajas de usar GP en series temporales?

- Nos permite realizar pronósticos probabilísticos.
- Dada la flexibilidad de los procesos gaussianos permite modelar diferentes patrones en los datos, como tendencias, estacionalidad y correlaciones.

Desventajas de usar GP en series temporales?

- Es costoso computacionalmente.
- Supuestos de linealidad.
- Supuestos de estacionariedad.
- Especificar una función de covarianza.

GPs ofrecen flexibilidad, cuantificación de la incertidumbre , capacidad de manejo de irregularidades y adaptabilidad, lo que los convierte en una herramienta poderosa para el modelado de series temporales tales como:

- Pronósticos
- Análisis de señales
- Supuestos de linealidad.
- Detección de anomalías
- Clasificación de objetos
- .. y muchos más!

Gracias!

References I

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32.

Deisenroth, M. P., Luo, Y., and van der Wilk, M. (2019). A practical guide to gaussian processes. *Distill (cit. on pp. 79, 100)*.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.