Procesos gaussianos
(Barber, Capítulo 14)

Miguel Palomino

Regresión lineal bayesiana

Consideramos un modelo dado por una combinación lineal de funciones base fijas:

$$y = \sum_i w_i \phi_i(x) = \mathbf{w}^\mathsf{T} \boldsymbol{\phi}$$

Si lo evaluamos en N puntos $x^1, \ldots, x^N$ y apilamos los resultados y^n en un vector $\mathbf{y}$, podemos escribir:

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}$$

donde $\boldsymbol{\Phi} = [\boldsymbol{\phi}(x^1), \ldots, \boldsymbol{\phi}(x^N)]^\mathsf{T}$. Suponiendo una distribución previa $\mathbf{w} \sim \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\right)$, resulta que $p(\mathbf{y}|\mathbf{x})$ es gaussiana por ser una combinación lineal de distribuciones gaussianas con media

$$\langle \mathbf{y} \rangle = \boldsymbol{\Phi} \langle \mathbf{w} \rangle = \mathbf{0}$$

y varianza

$$\mathbf{C} = \langle \mathbf{y}\mathbf{y}^\mathsf{T} \rangle - \langle \mathbf{y} \rangle \langle \mathbf{y} \rangle^\mathsf{T} = \boldsymbol{\Phi} \langle \mathbf{w}\mathbf{w}^\mathsf{T} \rangle \boldsymbol{\Phi}^\mathsf{T} = \frac{1}{\alpha} \boldsymbol{\Phi}\boldsymbol{\Phi}^\mathsf{T} = \mathbf{K}$$

El modelo de regresión lineal bayesiano induce una distribución gaussiana en la que los pesos han sido integrados

$$p(\mathbf{y}|\mathbf{x}) \sim \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})$$

y en la que la matriz de la covarianza $\mathbf{K}$ depende solo de los datos de entrada:

$$\mathbf{K}_{n,n'} = \frac{1}{\alpha}\boldsymbol{\phi}(x^n)^\mathsf{T}\boldsymbol{\phi}(x^{n'}) \qquad n, n' = 1, \ldots, N$$

- En un proceso gaussiano se especifica directamente $\mathbf{K}$ utilizando una función de covarianza k:

$$\mathbf{K}_{n,n'} = k(x^n, x^{n'})$$

- k debe generar una matriz definida positiva, como

$$k(\mathbf{x}^n, \mathbf{x}^{n'}) = \boldsymbol{\phi}(x^n)^\mathsf{T}\boldsymbol{\phi}(x^{n'})$$

- Dada k, existe una representación correspondiente en términos de funciones base; sin embargo, en muchos casos la representación utiliza un número infinito de funciones.

Procesos gaussianos

Sea $f : \mathcal{X} \to \mathbb{R}$ una función evaluada en un conjunto de puntos $\{\mathbf{x}^i\}_{i=1}^N$. Si $[f(\mathbf{x}^1), \dots, f(\mathbf{x}^N)]$ sigue una distribución normal para cualquier conjunto de puntos se dice que $f : \mathcal{X} \to \mathbb{R}$ es un proceso gaussiano.

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

Queda determinado por su función media y su función de covarianza o *kernel*:

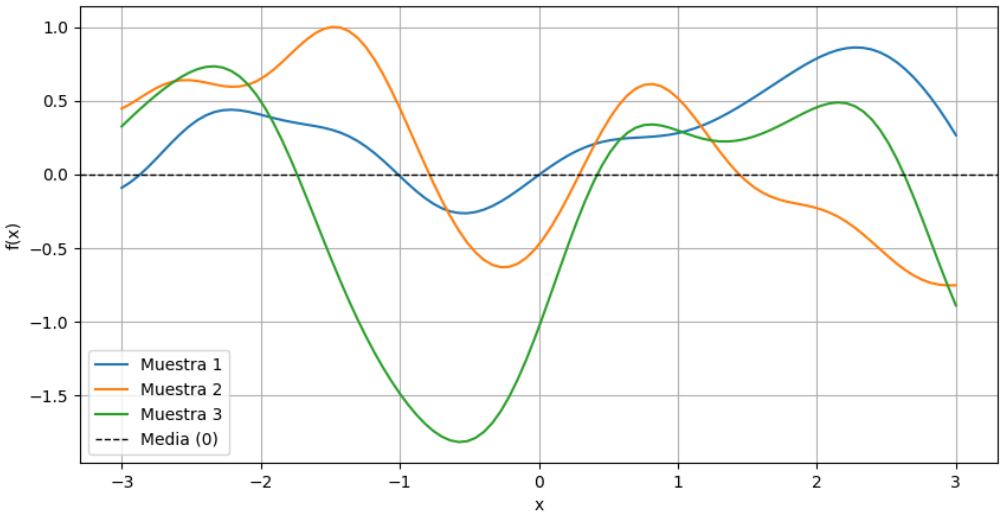
$$m(\mathbf{x}) = \langle f(\mathbf{x}) \rangle$$
$$k(\mathbf{x}, \mathbf{x}') = \langle (f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}')) \rangle$$

Observaciones

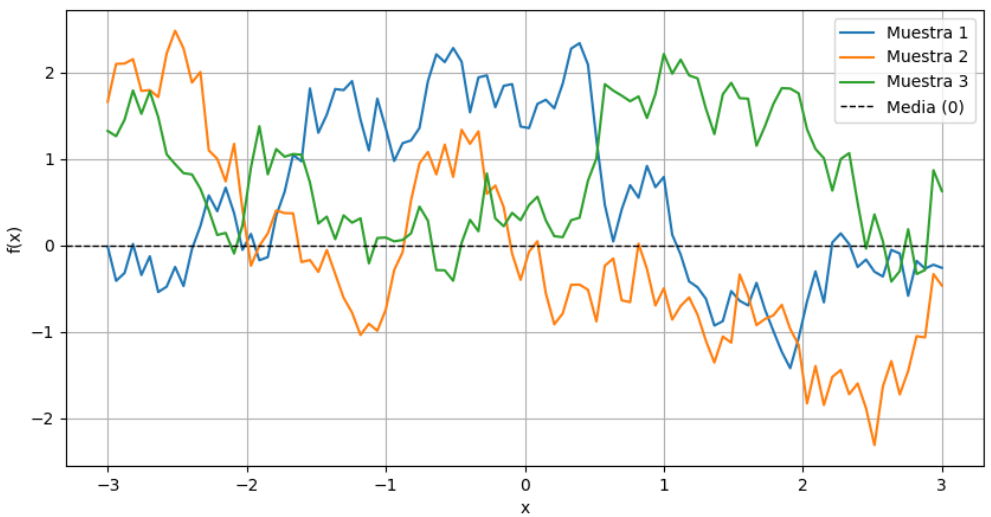
En la mayoría de situaciones no se tiene información sobre la media y por simetría se toma cero.

Dada k, el proceso gaussiano especifica una distribución sobre funciones: dados puntos $\mathbf{x}^1, \dots, \mathbf{x}^N$, se muestrean $y_1 = f(\mathbf{x}^1), \dots, y_n = f(\mathbf{x}^N)$ utilizando la correspondiente distribución gaussiana.

Muestras de procesos gaussianos



$$k(x, x') = \exp(-\frac{1}{2}(x - x')^2)$$



$$k(x, x') = \exp(-|x - x'|)$$

Inferencia en ausencia de ruido

A partir de un conjunto de datos $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$ y nuevos puntos de entrada $\mathbf{x}^*$, un PG con media cero crea un modelo gaussiano de las salidas $\{\mathbf{y}, \mathbf{y}^*\}$ dadas las entradas $\{\mathbf{x},\ \mathbf{x}^*\}$:

$$p(\mathbf{y}, \mathbf{y}^* | \mathbf{x}, \mathbf{x}^*) = \mathcal{N}\left(\begin{array}{c} \mathbf{y} \\ \mathbf{y}^* \end{array} \middle| \mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{x}, \mathbf{x}) & \mathbf{K}(\mathbf{x}, \mathbf{x}^*) \\ \mathbf{K}(\mathbf{x}^*, \mathbf{x}) & \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix} \right)$$

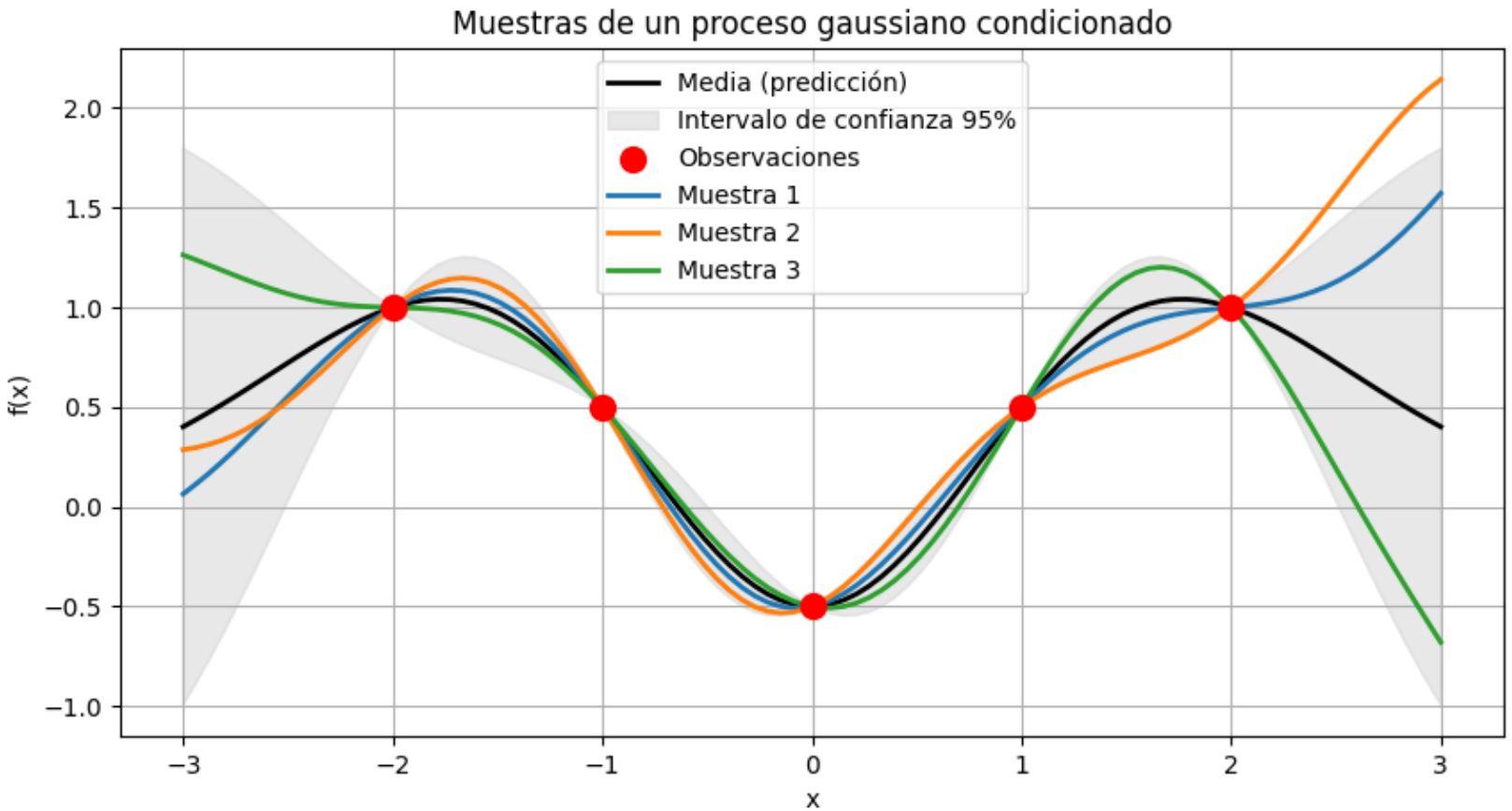
Para N puntos de entrenamiento y N' de prueba, $\mathbf{K}(\mathbf{x}, \mathbf{x}^*)$ denota la matriz de covarianza $N \times N'$ evaluada en todos los pares de puntos, y análogamente para las restantes submatrices:

$$\mathbf{K}(\mathbf{x}, \mathbf{x}^*)_{n,n'} = k(\mathbf{x}^n, \mathbf{x}^{*n'})$$

Utilizando resultados estándar de condicionamiento sobre gaussianas, la distribución predictiva resulta una normal:

$$\begin{aligned} p(\mathbf{y}^* | \mathbf{x}^*, \mathcal{D}) = \mathcal{N}(\mathbf{y}^* |\ & \mathbf{K}(\mathbf{x}^*, \mathbf{x})\mathbf{K}(\mathbf{x}, \mathbf{x})^{-1}\mathbf{y}, \\ & \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{K}(\mathbf{x}^*, \mathbf{x})\mathbf{K}(\mathbf{x}, \mathbf{x})^{-1}\mathbf{K}(\mathbf{x}, \mathbf{x}^*)) \end{aligned} \tag{1}$$

Muestras de procesos gaussianos condicionados



Inferencia con ruido

Normalmente no tendremos acceso a los valores de la función sino a versiones con ruido (independiente y gaussiano):

$$y = f(\mathbf{x}) + \epsilon, \qquad \text{donde } \epsilon \sim \mathcal{N}\left(\epsilon|0, \sigma^2\right)$$

Queremos predecir la señal limpia $\mathbf{f}^*$ asociada a $\mathbf{x}^*$. Puesto que

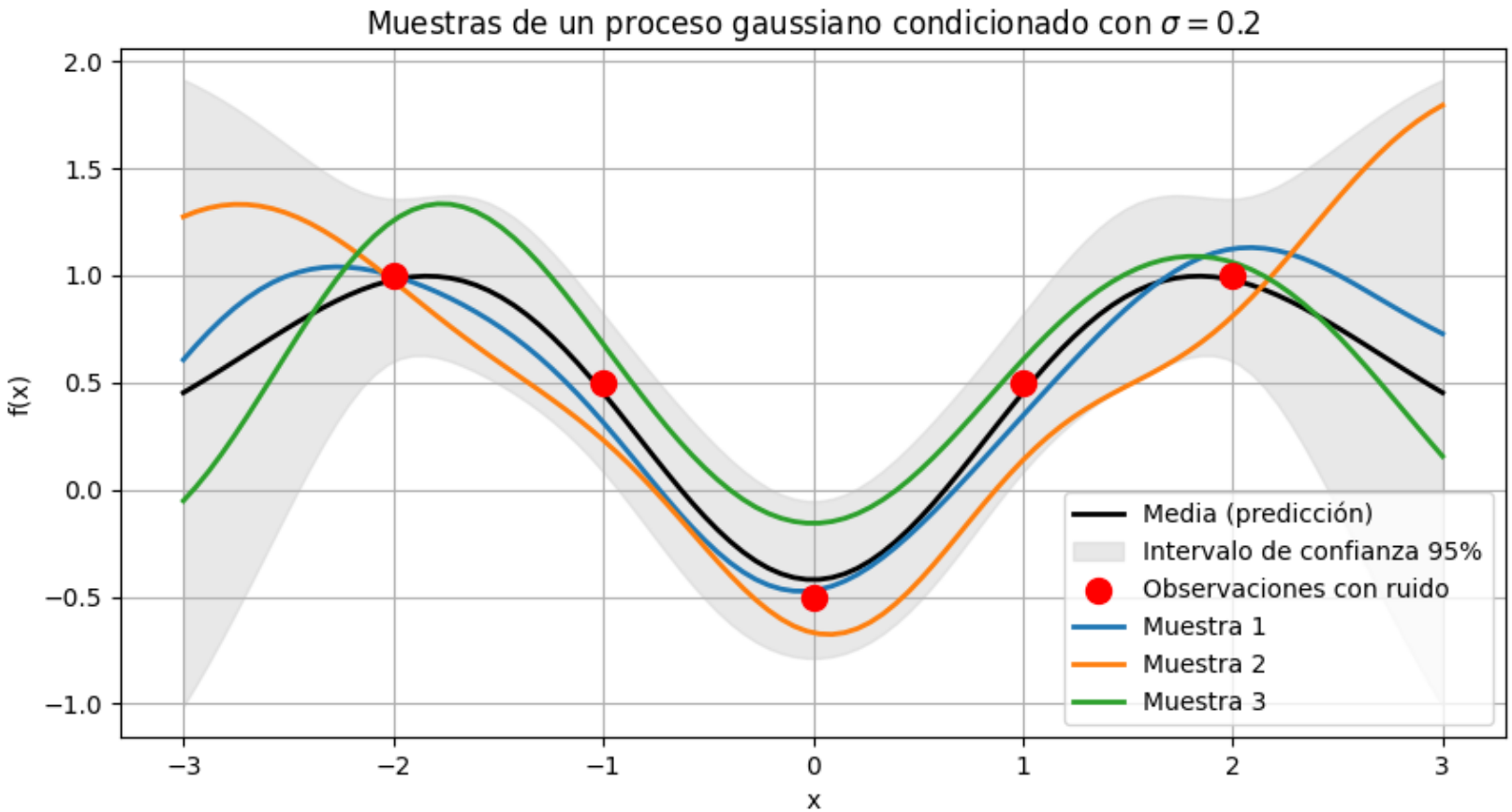
$$\begin{aligned} \langle y \rangle &= \langle f \rangle + \langle \epsilon \rangle = 0 \\ \langle y^m y^n \rangle &= \langle f^m f^n \rangle + \langle f^m \epsilon^n \rangle + \langle f^n \epsilon^m \rangle + \langle \epsilon^m \epsilon^n \rangle \\ &= k(\mathbf{x}^m, \mathbf{x}^n) + \sigma^2 \delta_{m,n} \end{aligned}$$

la distribución $p(\mathbf{y}, \mathbf{f}^*|\mathbf{x}, \mathbf{x}^*)$ es una normal de media cero con covarianza

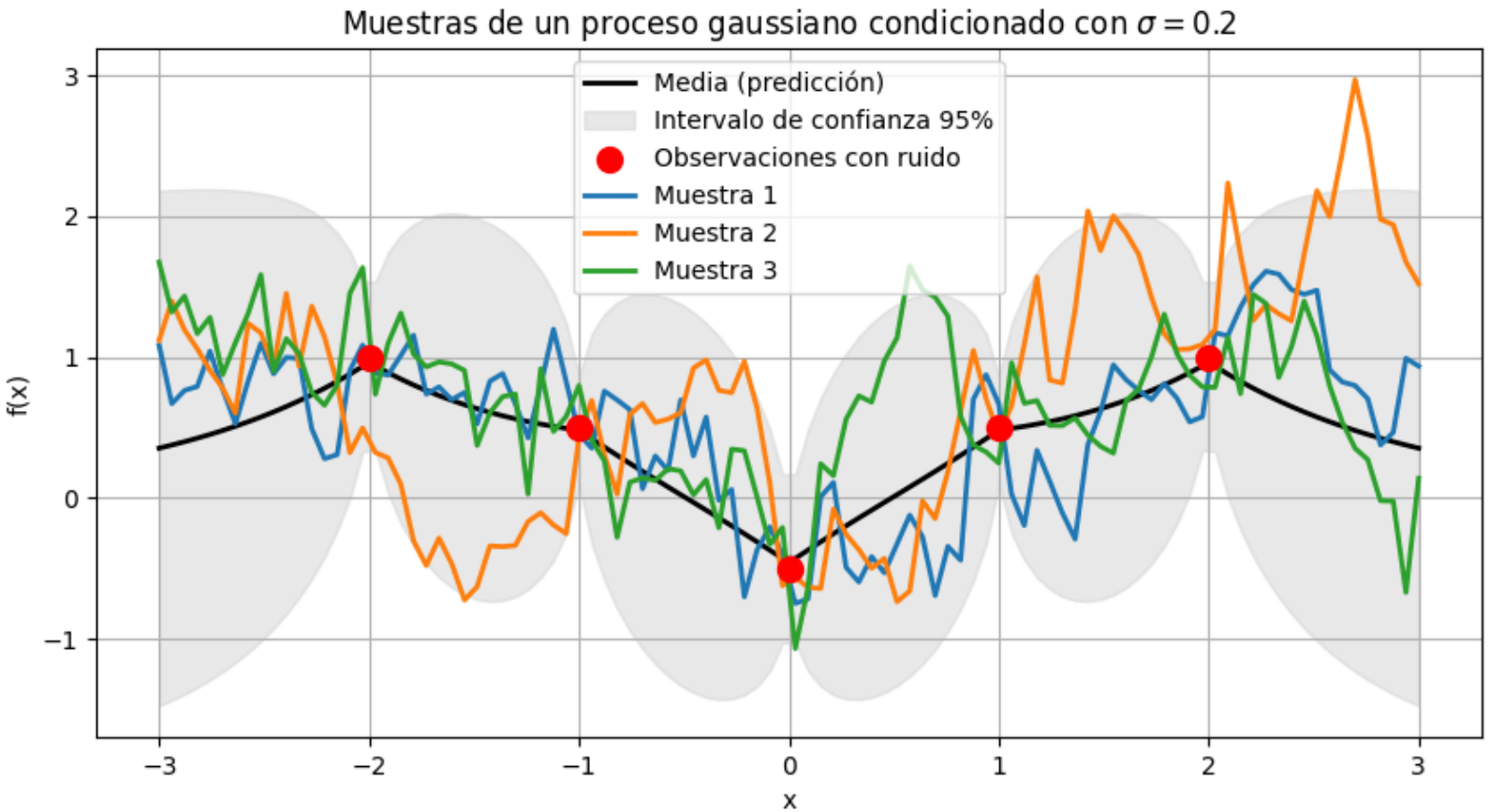
$$\begin{pmatrix} \mathbf{K}(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I} & \mathbf{K}(\mathbf{x}, \mathbf{x}^*) \\ \mathbf{K}(\mathbf{x}^*, \mathbf{x}) & \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) \end{pmatrix}$$

Para predecir hay que sustituir en (1) la matrix $\mathbf{K}(\mathbf{x}, \mathbf{x})$ por $\mathbf{K}(\mathbf{x}, \mathbf{x}) + \sigma^2$.

Muestras de procesos gaussianos condicionados con ruido



Muestras de procesos gaussianos condicionados con ruido



Funciones de covarianza o de núcleo (*kernels*)

Dada una colección de puntos $\mathbf{x}^1, \ldots, \mathbf{x}^M$, una función de covarianza $k(\mathbf{x}, \mathbf{x}')$ define los elementos de una matriz $\mathbf{C}$ semidefinida positiva:

$$\mathbf{C}_{i,j} = k(\mathbf{x}^i, \mathbf{x}^j)$$

Reglas de construcción.

- Suma. $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$
- Producto. $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') \cdot k_2(\mathbf{x}, \mathbf{x}')$
- Espacios producto. Para $\mathbf{z} = (\mathbf{x}, \mathbf{y})^\mathsf{T}$,

$$k(\mathbf{z}, \mathbf{z}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{y}, \mathbf{y}')$$
$$k(\mathbf{z}, \mathbf{z}') = k_1(\mathbf{x}, \mathbf{x}') \cdot k_2(\mathbf{y}, \mathbf{y}')$$

- Variación de escala. $k(\mathbf{x}, \mathbf{x}') = a(\mathbf{x}) k_1(\mathbf{x}, \mathbf{x}') a(\mathbf{x}')$

Funciones de covarianza

Una función de covarianza $k(\mathbf{x}, \mathbf{x}')$ es estacionaria si depende solo de la separación $\mathbf{x} - \mathbf{x}'$, esto es,

$$k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}')$$

Escribiremos $k(\mathbf{d})$, donde $\mathbf{d} = \mathbf{x} - \mathbf{x}'$.

Las funciones del GP correspondiente solo dependen de la distancia entre las entradas: son invariantes por traslaciones, en promedio.

Funciones de covarianza: ejemplos

Estacionarias

- Exponencial cuadrática. Es infinitamente diferenciable.

$$k(\mathbf{d}) = \exp(-\frac{|\mathbf{d}|^2}{l})$$

- Matérn. Es diferenciable k veces si $\nu > k$.

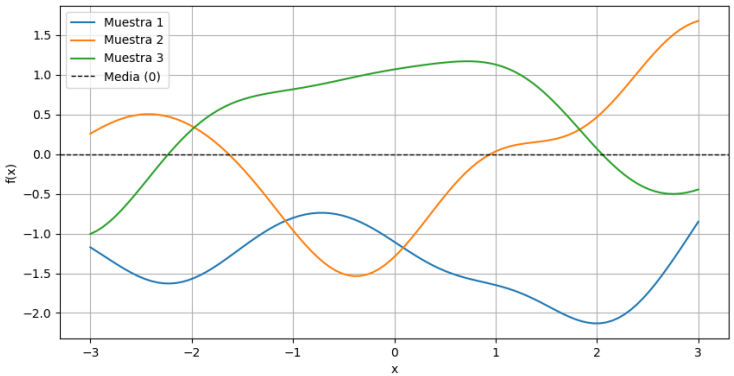
$$k(\mathbf{d}) = \frac{|\mathbf{d}|^{\nu}}{l} K_\nu(\frac{|\mathbf{d}|}{l}) \qquad K_\nu \text{ función de Bessel modificada}$$

- Periódica. En una dimensión. $k(x, x') = \exp(-\lambda \operatorname{sen}^2(\omega(x - x')))$, $\lambda > 0$.
- Cuadrática racional. $k(\mathbf{d}) = (1 + |\mathbf{d}|^2)^{-\alpha}$, $\alpha > 0$.

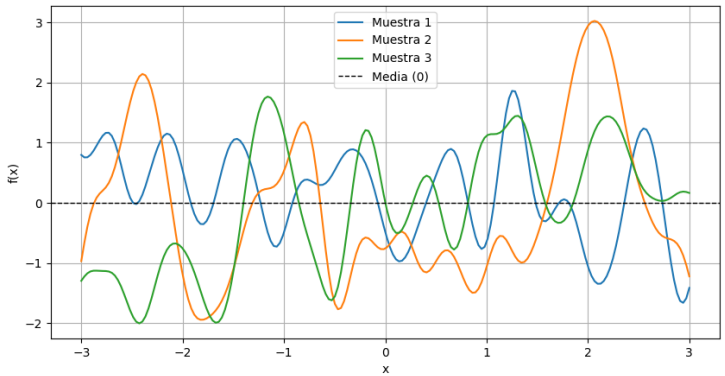
No estacionarias

- Lineal. $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\mathsf{T}\mathbf{x}'$.
- Red neuronal. $k(\mathbf{x}, \mathbf{x}') = \operatorname{sen}^{-1}(\frac{2\mathbf{x}^\mathsf{T}\Sigma\mathbf{x}'}{\sqrt{(1+2\mathbf{x}^\mathsf{T}\Sigma\mathbf{x})(1+2\mathbf{x}'^\mathsf{T}\Sigma\mathbf{x}')}})$.

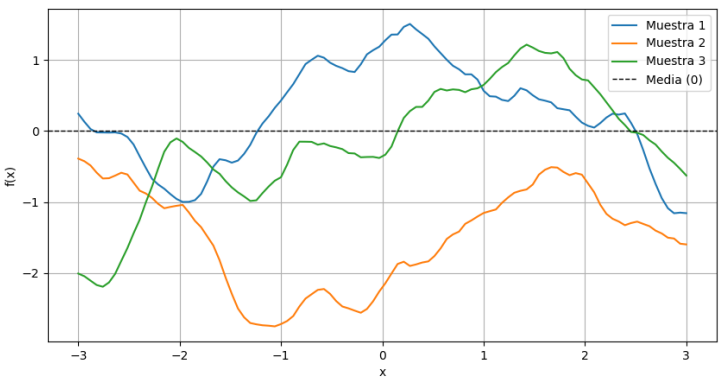
Muestras para funciones de covarianza diversas



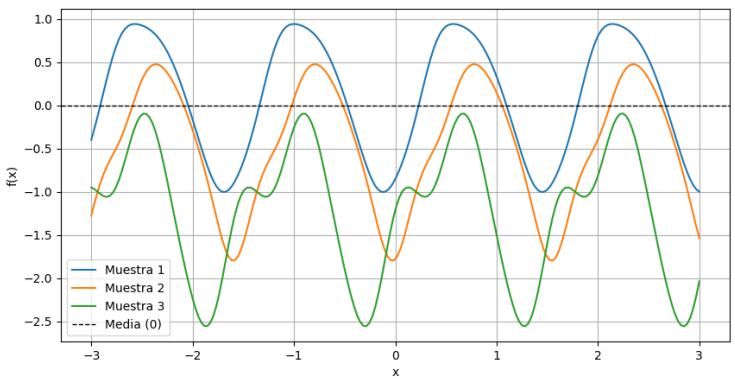
EC ($l = 1$)



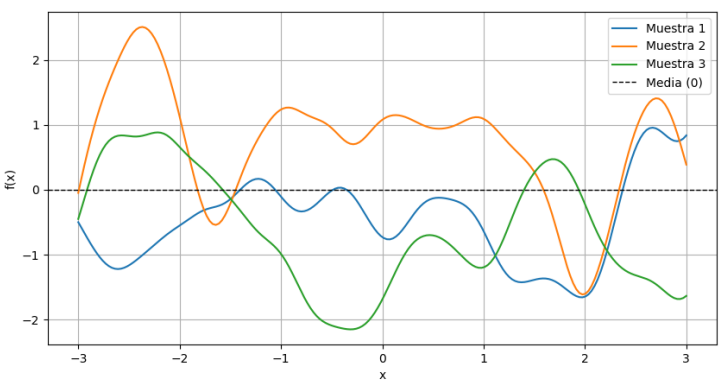
EC ($l = 0.2$)



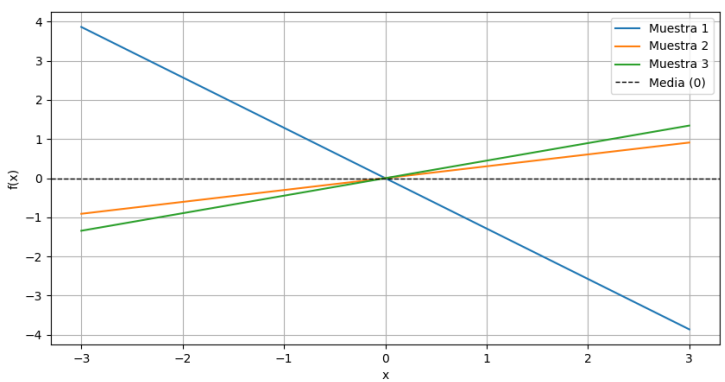
Mattern32



Periódica



CR ($\alpha = 3$)



Lineal

Selección de modelos

Para que el modelo sea útil en una aplicación hay que tomar decisiones sobre su especificación.

- Existe una multitud de familias de funciones de covarianza.
- Cada familia, a su vez, cuenta con hiperparámetros cuyos valores hay que determinar.

La selección de modelos es un proceso esencialmente abierto.

Selección bayesiana

Se trabaja con un conjunto (discreto) de posibles estructuras $\mathcal{H}_i$ y con hiperparámetros $\boldsymbol{\theta}$ que las controlan.

Selección de modelos: probabilidades posteriores

La probabilidad posterior de los hiperparámetros conocidos los datos de entrenamiento $\mathbf{X}, \mathbf{y}$ es:

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathcal{H}_i) = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{H}_i)p(\boldsymbol{\theta}|\mathcal{H}_i)}{p(\mathbf{y}|\mathbf{X}, \mathcal{H}_i)}$$

donde la constante normalizadora, la probabilidad de los datos para un modelo concreto, es:

$$p(\mathbf{y}|\mathbf{X}, \mathcal{H}_i) = \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{H}_i)p(\boldsymbol{\theta}|\mathcal{H}_i)d\boldsymbol{\theta} \tag{2}$$

La probabilidad posterior de un modelo es

$$p(\mathcal{H}_i|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathcal{H}_i)p(\mathcal{H}_i)}{p(\mathbf{y}|\mathbf{X})}$$

Verosimilitud marginal

Las integrales anteriores son en general difíciles de calcular; en particular, la integral (2), y en su lugar se suele maximizar la verosimilitud $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{H}_i)$. Se puede entonces aproximar la integral (2) usando una expansión local alrededor de dicho máximo.

Procesos gaussianos

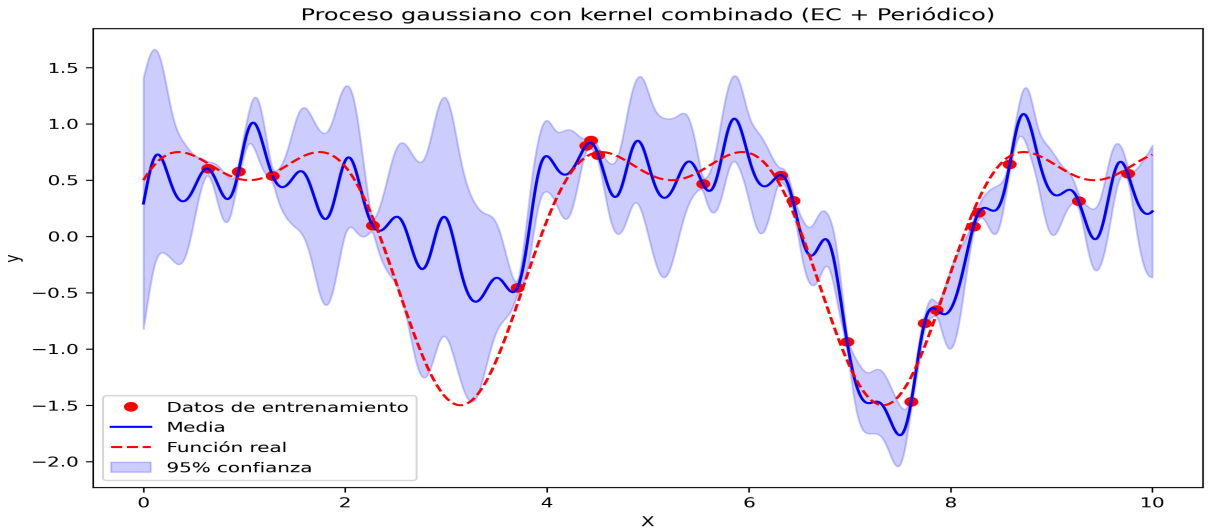
Para la mayoría de modelos, los cómputos requeridos para el enfoque bayesiano no son analíticamente tratables y no es sencillo derivar buenas aproximaciones. Los procesos gaussianos constituyen una excepción.

En el enfoque simplificado, se maximiza la verosimilitud marginal para un modelo fijo que viene dada por:

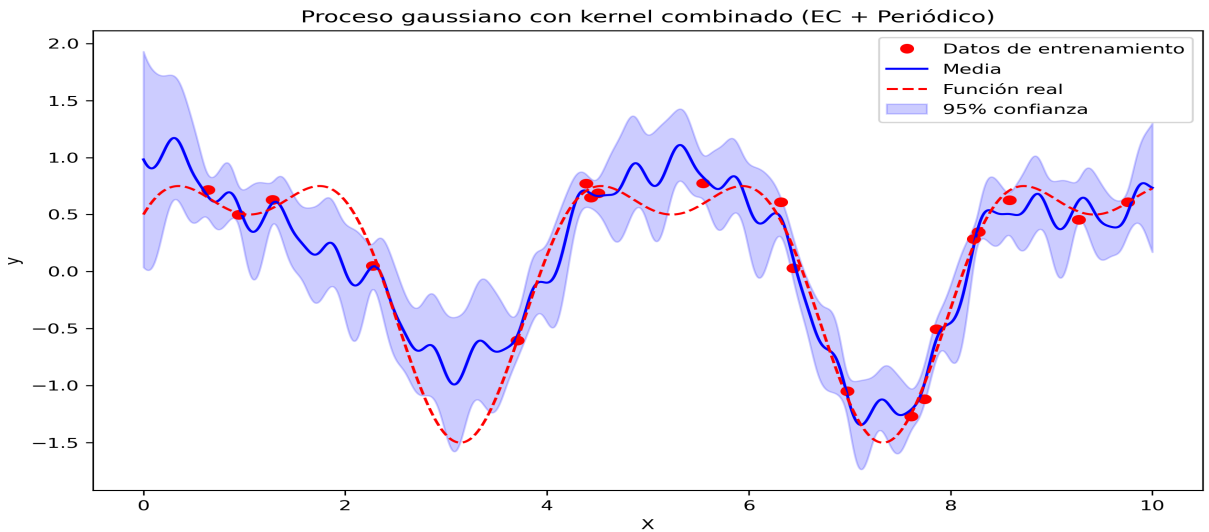
$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\mathsf{T}\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi$$

donde $\mathbf{K} = \mathbf{K}_f + \sigma^2\mathbf{I}$ es la matriz de covarianza con ruido y $\mathbf{K}_f$ la matriz de covarianza del modelo.

Ejemplo: máximos locales

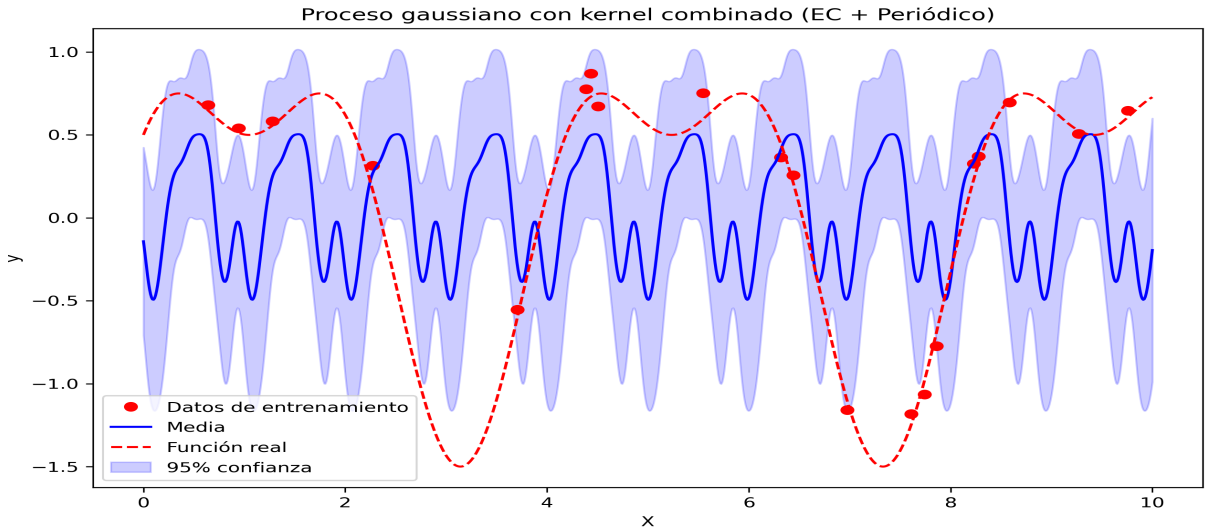


un máximo local:
$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_1, \mathcal{H}_i)$
(con `Nelder-Mead`)

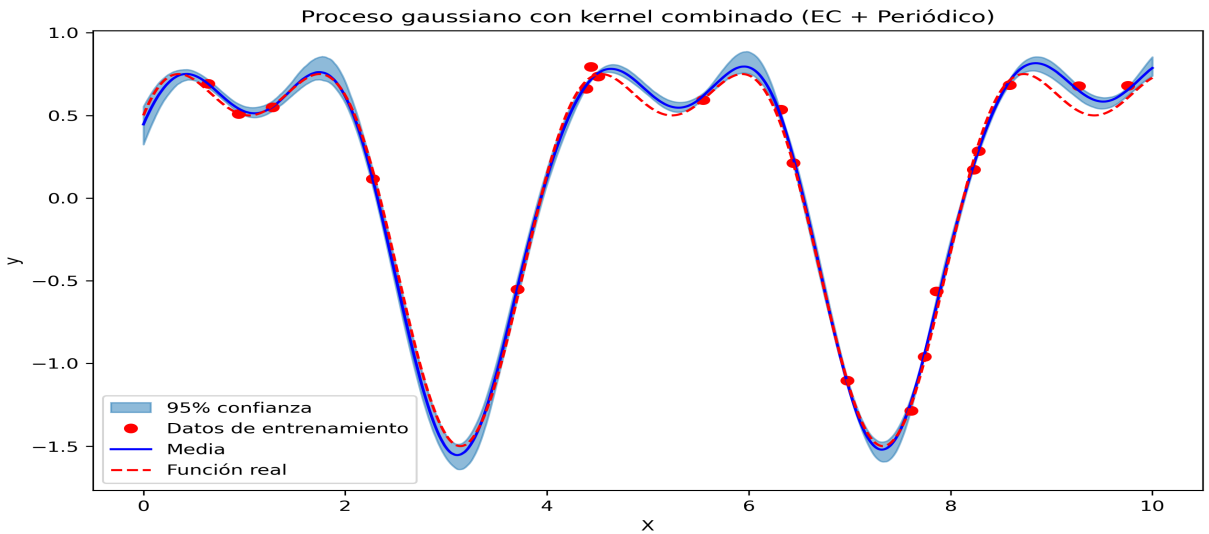


otro máximo local:
$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_2, \mathcal{H}_i)$
(con `L-BFGS-B`)

Ejemplo: otras posibilidades



(solo con covarianza periódica)



integrando todos los parámetros:
$p(\mathbf{y}|\mathbf{X}, \mathcal{H}_i)$
(Montecarlo)