# Clase 20 - Procesos Gaussianos I

## Aprendizaje de Máquinas - MA5204

Felipe Tobar

Department of Mathematical Engineering &
Center for Mathematical Modelling
Universidad de Chile

20 de marzo de 2021

UNIVERSIDAD
DE CHILE

# Modelos paramétricos y no paramétricos

- **Modelos paramétricos:** Son los modelos que hemos considerado hasta ahora y que se caracterizan por su cantidad fija de parámetros al momento de entrenar.
  Ejemplo: Regresión lineal/no lineal
- **Modelos no paramétricos:** Son los modelos que no tienen un número fijo de parámetros, pudiendo llegar incluso a ser infinitos.
  Ejemplo: Máquinas de soporte vectorial.

**Observación**

*Es importante hacer la distinción entre parámetros que se aprenden y los parámetros del modelo (hiperparámetros), donde estos últimos pueden ser fijos independiente de si el método es paramétrico o no paramétrico.*

En este capítulo introduciremos un método no paramétrico probabilístico de regresión no lineal, llamado procesos gaussianos ($\mathcal{GP}$). Este modelo en vez de encontrar un candidato único de la función a estimar, define una distribución sobre funciones $\mathbb{P}(f)$, donde $f$ es una función de un espacio de entrada $\mathcal{X}$ a los reales, $f : \mathcal{X} \to \mathbb{R}$. Esto tiene la virtud de permitir cuantificar la incertidumbre puntual que existe en la predicción de nuestro modelo, la cual servirá en forma de intervalos de confianza para la distribución gaussiana.

# Modelos paramétricos y no paramétricos

- **Modelos paramétricos:** Son los modelos que hemos considerado hasta ahora y que se caracterizan por su cantidad fija de parámetros al momento de entrenar.
  Ejemplo: Regresión lineal/no lineal
- **Modelos no paramétricos:** Son los modelos que no tienen un número fijo de parámetros, pudiendo llegar incluso a ser infinitos.
  Ejemplo: Máquinas de soporte vectorial.

**Observación**

*Es importante hacer la distinción entre parámetros que se aprenden y los parámetros del modelo (hiperparámetros), donde estos últimos pueden ser fijos independiente de si el método es paramétrico o no paramétrico.*

En este capítulo introduciremos un método no paramétrico probabilístico de regresión no lineal, llamado procesos gaussianos ($\mathcal{GP}$). Este modelo en vez de encontrar un candidato único de la función a estimar, define una distribución sobre funciones $\mathbb{P}(f)$, donde $f$ es una función de un espacio de entrada $\mathcal{X}$ a los reales, $f : \mathcal{X} \to \mathbb{R}$. Esto tiene la virtud de permitir cuantificar la incertidumbre puntual que existe en la predicción de nuestro modelo, la cual servirá en forma de intervalos de confianza para la distribución gaussiana.

# Modelos paramétricos y no paramétricos

- **Modelos paramétricos:** Son los modelos que hemos considerado hasta ahora y que se caracterizan por su cantidad fija de parámetros al momento de entrenar.
  Ejemplo: Regresión lineal/no lineal
- **Modelos no paramétricos:** Son los modelos que no tienen un número fijo de parámetros, pudiendo llegar incluso a ser infinitos.
  Ejemplo: Máquinas de soporte vectorial.

**Observación**

*Es importante hacer la distinción entre parámetros que se aprenden y los parámetros del modelo (hiperparámetros), donde estos últimos pueden ser fijos independiente de si el método es paramétrico o no paramétrico.*

En este capítulo introduciremos un método no paramétrico probabilístico de regresión no lineal, llamado procesos gaussianos ($\mathcal{GP}$). Este modelo en vez de encontrar un candidato único de la función a estimar, define una distribución sobre funciones $\mathbb{P}(f)$, donde $f$ es una función de un espacio de entrada $\mathcal{X}$ a los reales, $f : \mathcal{X} \to \mathbb{R}$. Esto tiene la virtud de permitir cuantificar la incertidumbre puntual que existe en la predicción de nuestro modelo, la cual servirá en forma de intervalos de confianza para la distribución gaussiana.

# Modelos paramétricos y no paramétricos

- **Modelos paramétricos:** Son los modelos que hemos considerado hasta ahora y que se caracterizan por su cantidad fija de parámetros al momento de entrenar.
  Ejemplo: Regresión lineal/no lineal
- **Modelos no paramétricos:** Son los modelos que no tienen un número fijo de parámetros, pudiendo llegar incluso a ser infinitos.
  Ejemplo: Máquinas de soporte vectorial.

**Observación**

*Es importante hacer la distinción entre parámetros que se aprenden y los parámetros del modelo (hiperparámetros), donde estos últimos pueden ser fijos independiente de si el método es paramétrico o no paramétrico.*

En este capítulo introduciremos un método no paramétrico probabilístico de regresión no lineal, llamado procesos gaussianos ($\mathcal{GP}$). Este modelo en vez de encontrar un candidato único de la función a estimar, define una distribución sobre funciones $\mathbb{P}(f)$, donde $f$ es una función de un espacio de entrada $\mathcal{X}$ a los reales, $f : \mathcal{X} \to \mathbb{R}$. Esto tiene la virtud de permitir cuantificar la incertidumbre puntual que existe en la predicción de nuestro modelo, la cual servirá en forma de intervalos de confianza para la distribución gaussiana.

# Modelos paramétricos y no paramétricos

- **Modelos paramétricos:** Son los modelos que hemos considerado hasta ahora y que se caracterizan por su cantidad fija de parámetros al momento de entrenar.
  Ejemplo: Regresión lineal/no lineal
- **Modelos no paramétricos:** Son los modelos que no tienen un número fijo de parámetros, pudiendo llegar incluso a ser infinitos.
  Ejemplo: Máquinas de soporte vectorial.

**Observación**

*Es importante hacer la distinción entre parámetros que se aprenden y los parámetros del modelo (hiperparámetros), donde estos últimos pueden ser fijos independiente de si el método es paramétrico o no paramétrico.*

En este capítulo introduciremos un método no paramétrico probabilístico de regresión no lineal, llamado procesos gaussianos ($\mathcal{GP}$). Este modelo en vez de encontrar un candidato único de la función a estimar, define una distribución sobre funciones $\mathbb{P}(f)$, donde $f$ es una función de un espacio de entrada $\mathcal{X}$ a los reales, $f : \mathcal{X} \to \mathbb{R}$. Esto tiene la virtud de permitir cuantificar la incertidumbre puntual que existe en la predicción de nuestro modelo, la cual servirá en forma de intervalos de confianza para la distribución gaussiana.

# Modelos paramétricos y no paramétricos

- **Modelos paramétricos:** Son los modelos que hemos considerado hasta ahora y que se caracterizan por su cantidad fija de parámetros al momento de entrenar.
  Ejemplo: Regresión lineal/no lineal
- **Modelos no paramétricos:** Son los modelos que no tienen un número fijo de parámetros, pudiendo llegar incluso a ser infinitos.
  Ejemplo: Máquinas de soporte vectorial.

**Observación**

*Es importante hacer la distinción entre parámetros que se aprenden y los parámetros del modelo (hiperparámetros), donde estos últimos pueden ser fijos independiente de si el método es paramétrico o no paramétrico.*

En este capítulo introduciremos un método no paramétrico probabilístico de regresión no lineal, llamado procesos gaussianos ($\mathcal{GP}$). Este modelo en vez de encontrar un candidato único de la función a estimar, define una distribución sobre funciones $\mathbb{P}(f)$, donde $f$ es una función de un espacio de entrada $\mathcal{X}$ a los reales, $f : \mathcal{X} \to \mathbb{R}$. Esto tiene la virtud de permitir cuantificar la incertidumbre puntual que existe en la predicción de nuestro modelo, la cual servirá en forma de intervalos de confianza para la distribución gaussiana.

# Proceso Gaussiano

**Definition (proceso gaussiano)**

Un proceso gaussiano ($\mathcal{GP}$) es una colección de variables aleatorias, tal que para cualquier subconjunto finito de puntos, estos tienen una distribución conjuntamente gaussiana.

Al aplicar esta definición a nuestro caso anterior, $\mathbb{P}(f)$ será un $\mathcal{GP}$ y para cualquier conjunto finito $\{x_i\}_{i=1}^n \subset \mathcal{X}$, la distribución de $\mathbb{P}(f(\mathbf{x}))$ es Gaussiana multivariada $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))^\top$). En este caso las variables aleatorias representan el valor de la función $f(x_i)$ en la posición $x_i$.

Un $\mathcal{GP}$ queda completamente caracterizado por su función de media $m(\cdot)$ y función de covarianza $K(\cdot, \cdot)$, de esta forma para cualquier conjunto finito podemos encontrar la distribución. Definimos estas funciones como

$$m(x) = \mathbb{E}\{f(x)\}$$
$$K(x, x') = \mathbb{E}\left\{(f(x) - m(x))\left(f(x') - m(x')\right)\right\}$$

# Proceso Gaussiano

**Definition (proceso gaussiano)**

Un proceso gaussiano ($\mathcal{GP}$) es una colección de variables aleatorias, tal que para cualquier subconjunto finito de puntos, estos tienen una distribución conjuntamente gaussiana.

Al aplicar esta definición a nuestro caso anterior, $\mathbb{P}(f)$ será un $\mathcal{GP}$ y para cualquier conjunto finito $\{x_i\}_{i=1}^n \subset \mathcal{X}$, la distribución de $\mathbb{P}(f(\mathbf{x}))$ es Gaussiana multivariada $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))^\top$). En este caso las variables aleatorias representan el valor de la función $f(x_i)$ en la posición $x_i$.

Un $\mathcal{GP}$ queda completamente caracterizado por su función de media $m(\cdot)$ y función de covarianza $K(\cdot, \cdot)$, de esta forma para cualquier conjunto finito podemos encontrar la distribución. Definimos estas funciones como

$$m(x) = \mathbb{E}\{f(x)\}$$
$$K(x, x') = \mathbb{E}\left\{(f(x) - m(x))\left(f(x') - m(x')\right)\right\}$$

# Proceso Gaussiano

**Definition (proceso gaussiano)**

Un proceso gaussiano ($\mathcal{GP}$) es una colección de variables aleatorias, tal que para cualquier subconjunto finito de puntos, estos tienen una distribución conjuntamente gaussiana.

Al aplicar esta definición a nuestro caso anterior, $\mathbb{P}(f)$ será un $\mathcal{GP}$ y para cualquier conjunto finito $\{x_i\}_{i=1}^n \subset \mathcal{X}$, la distribución de $\mathbb{P}(f(\mathbf{x}))$ es Gaussiana multivariada $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))^\top$). En este caso las variables aleatorias representan el valor de la función $f(x_i)$ en la posición $x_i$.

Un $\mathcal{GP}$ queda completamente caracterizado por su función de media $m(\cdot)$ y función de covarianza $K(\cdot, \cdot)$, de esta forma para cualquier conjunto finito podemos encontrar la distribución. Definimos estas funciones como

$$
\begin{aligned}
m(x) &= \mathbb{E}\{f(x)\} \\
K(x, x') &= \mathbb{E}\left\{(f(x) - m(x))\left(f(x') - m(x')\right)\right\}
\end{aligned}
$$

# Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada $\mathcal{X}$ como genérico, un caso común es definir los $\mathcal{GP}$ sobre el tiempo ($\mathbb{R}^+$), es decir que los $x_i$ son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo $\mathbb{R}^d$.

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un $\mathcal{GP}$ define una distribución multivariada para digamos dos variables $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$ entonces también debe definir $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ donde $\mu_1$ es la componente respectiva del vector $\mu$ y $\Sigma_{11}$ la submatriz correspondiente de $\Sigma$. En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

# Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada $\mathcal{X}$ como genérico, un caso común es definir los $\mathcal{GP}$ sobre el tiempo ($\mathbb{R}^+$), es decir que los $x_i$ son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo $\mathbb{R}^d$.

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un $\mathcal{GP}$ define una distribución multivariada para digamos dos variables $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$ entonces también debe definir $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ donde $\mu_1$ es la componente respectiva del vector $\mu$ y $\Sigma_{11}$ la submatriz correspondiente de $\Sigma$. En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

# Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada $\mathcal{X}$ como genérico, un caso común es definir los $\mathcal{GP}$ sobre el tiempo ($\mathbb{R}^+$), es decir que los $x_i$ son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo $\mathbb{R}^d$.

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un $\mathcal{GP}$ define una distribución multivariada para digamos dos variables $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$ entonces también debe definir $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ donde $\mu_1$ es la componente respectiva del vector $\mu$ y $\Sigma_{11}$ la submatriz correspondiente de $\Sigma$. En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

# Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada $\mathcal{X}$ como genérico, un caso común es definir los $\mathcal{GP}$ sobre el tiempo ($\mathbb{R}^+$), es decir que los $x_i$ son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo $\mathbb{R}^d$.

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un $\mathcal{GP}$ define una distribución multivariada para digamos dos variables $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$ entonces también debe definir $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ donde $\mu_1$ es la componente respectiva del vector $\mu$ y $\Sigma_{11}$ la submatriz correspondiente de $\Sigma$. En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

# Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada $\mathcal{X}$ como genérico, un caso común es definir los $\mathcal{GP}$ sobre el tiempo ($\mathbb{R}^+$), es decir que los $x_i$ son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo $\mathbb{R}^d$.

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un $\mathcal{GP}$ define una distribución multivariada para digamos dos variables $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$ entonces también debe definir $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ donde $\mu_1$ es la componente respectiva del vector $\mu$ y $\Sigma_{11}$ la submatriz correspondiente de $\Sigma$. En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

# Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada $\mathcal{X}$ como genérico, un caso común es definir los $\mathcal{GP}$ sobre el tiempo ($\mathbb{R}^+$), es decir que los $x_i$ son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo $\mathbb{R}^d$.

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un $\mathcal{GP}$ define una distribución multivariada para digamos dos variables $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$ entonces también debe definir $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ donde $\mu_1$ es la componente respectiva del vector $\mu$ y $\Sigma_{11}$ la submatriz correspondiente de $\Sigma$. En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

# Muestreo de un GP

Un $\mathcal{GP}$ define un *prior* sobre funciones, por lo que, antes de ver ningún dato se podría obtener una muestra de este proceso dada una función de media y covarianza.

Consideremos $m(\cdot) = 0$ y función de covarianza (kernel) exponencial cuadrática (o RBF) definida como

$$K_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

Donde en este caso los parámetros son interpretables (y como veremos más adelante pueden ser aprendidos a través de un conjunto de entrenamiento) donde $\sigma^2$ es la varianza de la función, notar que esta es la diagonal de la matriz covarianza. El parámetro $\ell$ es conocido como el *lenghtscale* que determina que tan lejos tiene influencia un punto sobre otro, donde en general un punto no tendrá influencia más allá de $\ell$ unidades alrededor.

A continuación se presenta un muestreo de un *gp* con las características anteriores

# Muestreo de un GP

Un $\mathcal{GP}$ define un *prior* sobre funciones, por lo que, antes de ver ningún dato se podría obtener una muestra de este proceso dada una función de media y covarianza.

Consideremos $m(\cdot) = 0$ y función de covarianza (kernel) exponencial cuadrática (o RBF) definida como

$$K_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

Donde en este caso los parámetros son interpretables (y como veremos más adelante pueden ser aprendidos a través de un conjunto de entrenamiento) donde $\sigma^2$ es la varianza de la función, notar que esta es la diagonal de la matriz covarianza. El parámetro $\ell$ es conocido como el *lenghtscale* que determina que tan lejos tiene influencia un punto sobre otro, donde en general un punto no tendrá influencia más allá de $\ell$ unidades alrededor.

A continuación se presenta un muestreo de un *gp* con las características anteriores

# Muestreo de un GP

Un $\mathcal{GP}$ define un *prior* sobre funciones, por lo que, antes de ver ningún dato se podría obtener una muestra de este proceso dada una función de media y covarianza.

Consideremos $m(\cdot) = 0$ y función de covarianza (kernel) exponencial cuadrática (o RBF) definida como

$$K_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

Donde en este caso los parámetros son interpretables (y como veremos más adelante pueden ser aprendidos a través de un conjunto de entrenamiento) donde $\sigma^2$ es la varianza de la función, notar que esta es la diagonal de la matriz covarianza. El parámetro $\ell$ es conocido como el *lenghtscale* que determina que tan lejos tiene influencia un punto sobre otro, donde en general un punto no tendrá influencia más allá de $\ell$ unidades alrededor.

A continuación se presenta un muestreo de un *gp* con las características anteriores

# Muestreo de un GP

Un $\mathcal{GP}$ define un *prior* sobre funciones, por lo que, antes de ver ningún dato se podría obtener una muestra de este proceso dada una función de media y covarianza.

Consideremos $m(\cdot) = 0$ y función de covarianza (kernel) exponencial cuadrática (o RBF) definida como

$$K_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

Donde en este caso los parámetros son interpretables (y como veremos más adelante pueden ser aprendidos a través de un conjunto de entrenamiento) donde $\sigma^2$ es la varianza de la función, notar que esta es la diagonal de la matriz covarianza. El parámetro $\ell$ es conocido como el *lenghtscale* que determina que tan lejos tiene influencia un punto sobre otro, donde en general un punto no tendrá influencia más allá de $\ell$ unidades alrededor.

A continuación se presenta un muestreo de un *gp* con las características anteriores

# Muestreo de un GP

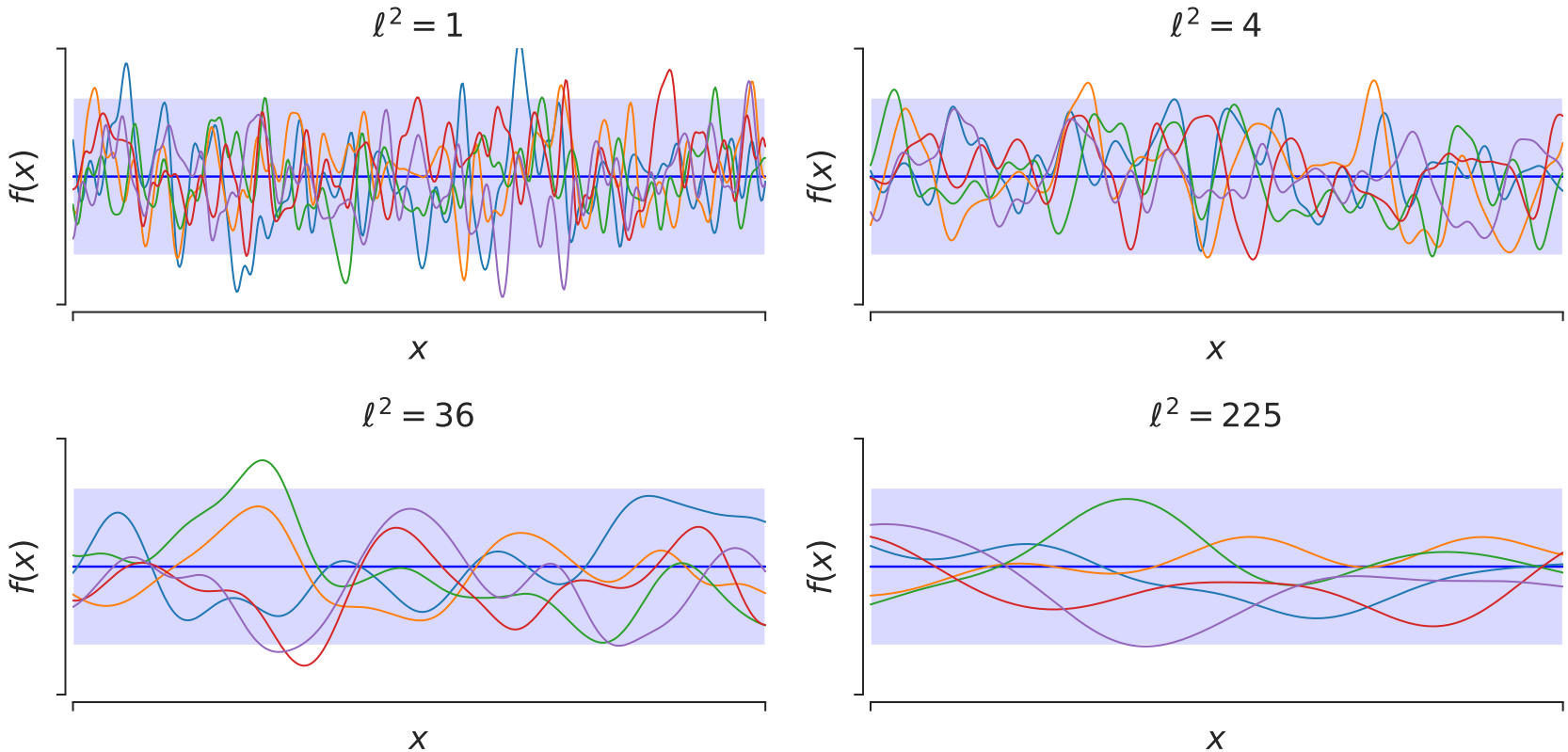


**Fig..** Muestras de un prior $\mathcal{GP}$ con kernel SE, para distintos *lenghtscales* ($\ell$) y función media $m(\cdot) = 0$, la parte sombreada corresponde al intervalo de confianza del 95 %. Se puede ver que a mayor $\ell$ las funciones se van volviendo más suaves.

# Incorporando información - Evaluación sin ruido

Consideremos las observaciones sin ruido de la forma $\{(x_i, f(x_i))\}_{i=1}^n$ (conocemos el valor real en $X = [x_1, \ldots, x_n]$). Digamos que queremos realizar una predicción en el conjunto $X_*$ de $n_*$ puntos, luego la distribución conjunta es de la forma:

$$\begin{bmatrix} f(X) \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right)$$

El punto clave para realizar predicciones, es el siguiente lema

**Lemma**

*Dado un prior $\mathcal{GP}$ sobre $f(\cdot)$ y una verosimilitud Gaussiana, la posterior sobre $f(\cdot)$ es también un $\mathcal{GP}$. Además, se puede condicionar sobre las observaciones $(X, f(X))$ para obtener*

$$f(X_*)|f(X), X \sim \mathcal{N}(m_{X_*|X}, \Sigma_{X_*|X})$$

*Donde la media y covarianza son:*

$$m_{X_*|X} = m(X_*) + K(X_*, X)K^{-1}(X, X)(f(X) - m(X))$$
$$\Sigma_{X_*|X} = K(X_*, X_*) - K(X_*, X)K^{-1}(X, X)K(X, X_*)$$

# Incorporando información - Evaluación sin ruido

Consideremos las observaciones sin ruido de la forma $\{(x_i, f(x_i))\}_{i=1}^n$ (conocemos el valor real en $X = [x_1, \ldots, x_n]$). Digamos que queremos realizar una predicción en el conjunto $X_*$ de $n_*$ puntos, luego la distribución conjunta es de la forma:

$$\begin{bmatrix} f(X) \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right)$$

El punto clave para realizar predicciones, es el siguiente lema

**Lemma**

*Dado un prior $\mathcal{GP}$ sobre $f(\cdot)$ y una verosimilitud Gaussiana, la posterior sobre $f(\cdot)$ es también un $\mathcal{GP}$. Además, se puede condicionar sobre las observaciones $(X, f(X))$ para obtener*

$$f(X_*)|f(X), X \sim \mathcal{N}(m_{X_*|X}, \Sigma_{X_*|X})$$

*Donde la media y covarianza son:*

$$m_{X_*|X} = m(X_*) + K(X_*, X)K^{-1}(X,X)(f(X) - m(X))$$
$$\Sigma_{X_*|X} = K(X_*, X_*) - K(X_*, X)K^{-1}(X,X)K(X, X_*)$$

# Incorporando información - Evaluación sin ruido

Consideremos las observaciones sin ruido de la forma $\{(x_i, f(x_i))\}_{i=1}^n$ (conocemos el valor real en $X = [x_1, \ldots, x_n]$). Digamos que queremos realizar una predicción en el conjunto $X_*$ de $n_*$ puntos, luego la distribución conjunta es de la forma:

$$\begin{bmatrix} f(X) \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right)$$

El punto clave para realizar predicciones, es el siguiente lema

**Lemma**

*Dado un prior $\mathcal{GP}$ sobre $f(\cdot)$ y una verosimilitud Gaussiana, la posterior sobre $f(\cdot)$ es también un $\mathcal{GP}$. Además, se puede condicionar sobre las observaciones $(X, f(X))$ para obtener*

$$f(X_*)|f(X), X \sim \mathcal{N}(m_{X_*|X}, \Sigma_{X_*|X})$$

*Donde la media y covarianza son:*

$$m_{X_*|X} = m(X_*) + K(X_*, X)K^{-1}(X,X)(f(X) - m(X))$$
$$\Sigma_{X_*|X} = K(X_*, X_*) - K(X_*, X)K^{-1}(X,X)K(X, X_*)$$

# Incorporando información - Evaluación sin ruido

Consideremos las observaciones sin ruido de la forma $\{(x_i, f(x_i))\}_{i=1}^n$ (conocemos el valor real en $X = [x_1, \ldots, x_n]$). Digamos que queremos realizar una predicción en el conjunto $X_*$ de $n_*$ puntos, luego la distribución conjunta es de la forma:

$$\begin{bmatrix} f(X) \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right)$$

El punto clave para realizar predicciones, es el siguiente lema

**Lemma**

*Dado un prior $\mathcal{GP}$ sobre $f(\cdot)$ y una verosimilitud Gaussiana, la posterior sobre $f(\cdot)$ es también un $\mathcal{GP}$. Además, se puede condicionar sobre las observaciones $(X, f(X))$ para obtener*

$$f(X_*)|f(X), X \sim \mathcal{N}(m_{X_*|X}, \Sigma_{X_*|X})$$

*Donde la media y covarianza son:*

$$m_{X_*|X} = m(X_*) + K(X_*, X)K^{-1}(X,X)(f(X) - m(X))$$
$$\Sigma_{X_*|X} = K(X_*, X_*) - K(X_*, X)K^{-1}(X,X)K(X, X_*)$$

# Incorporando información - Evaluación sin ruido

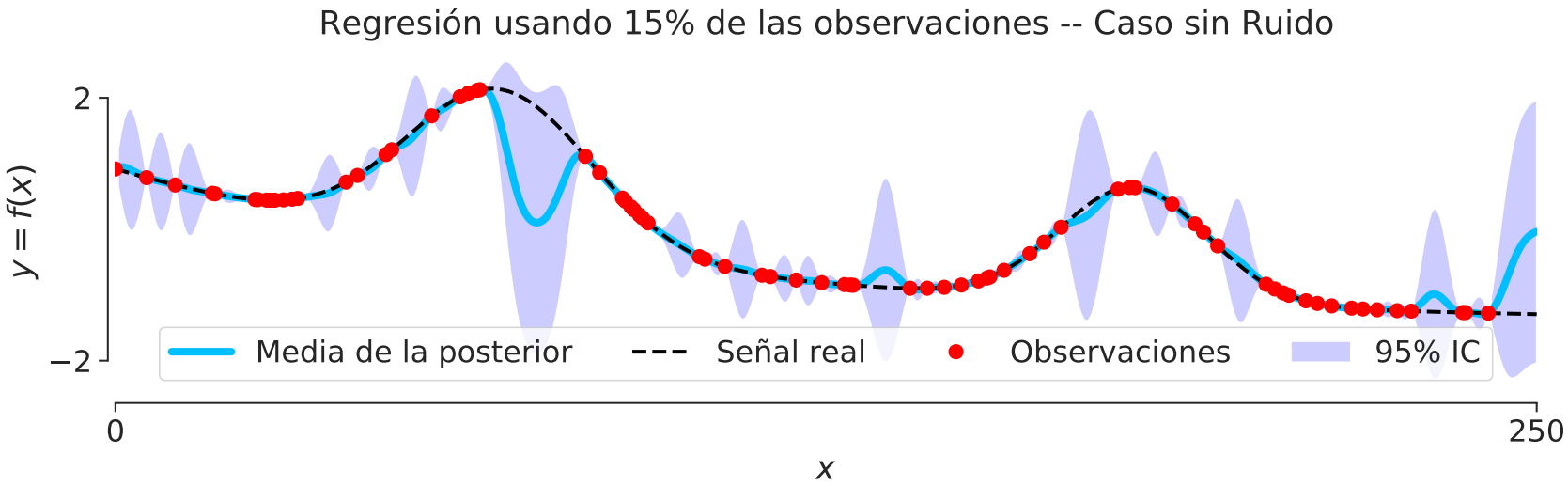


**Fig..** Regresión con $\mathcal{GP}$ para señal sintetica usando el 15 % de los datos muestreados de forma no uniforme, utilizand un $\mathcal{GP}$ de media nula y kernel SE.

# Incorporando información - Evaluación con ruido

En este caso las observaciones son de la forma $y_i = f(x_i) + \eta$ donde $\eta \sim \mathcal{N}(0, \sigma_n^2)$ por lo que ahora nuestro conjunto de observaciones es de la forma $(X, Y)$ donde $Y = f(X) + \eta$.

Lo que en nuestro modelo equivale a agregar un término a la función de covarianza

$$cov(Y) = K(X, X) + \sigma_n^2 \mathbb{I}$$

Donde si tenemos el mismo caso anterior, observaciones $(X, Y)$ y queremos evaluar en $X_*$, la conjunta queda

$$\begin{bmatrix} Y \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 \mathbb{I} & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right)$$

**Observación**

*Notemos que el termino de ruido solo es agregado al subbloque correspondiente a las observaciones, no se agrega el termino en los otros subbloques pues buscamos hacer una predicción de la función latente $f(\cdot)$ y no una versión ruidosa de esta.*

# Incorporando información - Evaluación con ruido

En este caso las observaciones son de la forma $y_i = f(x_i) + \eta$ donde $\eta \sim \mathcal{N}(0, \sigma_n^2)$ por lo que ahora nuestro conjunto de observaciones es de la forma $(X, Y)$ donde $Y = f(X) + \eta$.
Lo que en nuestro modelo equivale a agregar un término a la función de covarianza

$$cov(Y) = K(X, X) + \sigma_n^2 \mathbb{I}$$

Donde si tenemos el mismo caso anterior, observaciones $(X, Y)$ y queremos evaluar en $X_*$, la conjunta queda

$$\begin{bmatrix} Y \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 \mathbb{I} & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right)$$

**Observación**

*Notemos que el termino de ruido solo es agregado al subbloque correspondiente a las observaciones, no se agrega el termino en los otros subbloques pues buscamos hacer una predicción de la función latente $f(\cdot)$ y no una versión ruidosa de esta.*

# Incorporando información - Evaluación con ruido

En este caso las observaciones son de la forma $y_i = f(x_i) + \eta$ donde $\eta \sim \mathcal{N}(0, \sigma_n^2)$ por lo que ahora nuestro conjunto de observaciones es de la forma $(X, Y)$ donde $Y = f(X) + \eta$.
Lo que en nuestro modelo equivale a agregar un término a la función de covarianza

$$cov(Y) = K(X, X) + \sigma_n^2 \mathbb{I}$$

Donde si tenemos el mismo caso anterior, observaciones $(X, Y)$ y queremos evaluar en $X_*$, la conjunta queda

$$\begin{bmatrix} Y \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_n^2 \mathbb{I} & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right)$$

**Observación**

*Notemos que el termino de ruido solo es agregado al subbloque correspondiente a las observaciones, no se agrega el termino en los otros subbloques pues buscamos hacer una predicción de la función latente $f(\cdot)$ y no una versión ruidosa de esta.*

# Incorporando información - Evaluación con ruido

En este caso las observaciones son de la forma $y_i = f(x_i) + \eta$ donde $\eta \sim \mathcal{N}(0, \sigma_n^2)$ por lo que ahora nuestro conjunto de observaciones es de la forma $(X, Y)$ donde $Y = f(X) + \eta$.
Lo que en nuestro modelo equivale a agregar un término a la función de covarianza

$$cov(Y) = K(X, X) + \sigma_n^2 \mathbb{I}$$

Donde si tenemos el mismo caso anterior, observaciones $(X, Y)$ y queremos evaluar en $X_*$, la conjunta queda

$$\begin{bmatrix} Y \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 \mathbb{I} & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right)$$

**Observación**

*Notemos que el termino de ruido solo es agregado al subbloque correspondiente a las observaciones, no se agrega el termino en los otros subbloques pues buscamos hacer una predicción de la función latente $f(\cdot)$ y no una versión ruidosa de esta.*

# Incorporando información - Evaluación con ruido

En este caso las observaciones son de la forma $y_i = f(x_i) + \eta$ donde $\eta \sim \mathcal{N}(0, \sigma_n^2)$ por lo que ahora nuestro conjunto de observaciones es de la forma $(X, Y)$ donde $Y = f(X) + \eta$.
Lo que en nuestro modelo equivale a agregar un término a la función de covarianza

$$cov(Y) = K(X, X) + \sigma_n^2 \mathbb{I}$$

Donde si tenemos el mismo caso anterior, observaciones $(X, Y)$ y queremos evaluar en $X_*$, la conjunta queda

$$\begin{bmatrix} Y \\ f(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 \mathbb{I} & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right)$$

**Observación**

*Notemos que el termino de ruido solo es agregado al subbloque correspondiente a las observaciones, no se agrega el termino en los otros subbloques pues buscamos hacer una predicción de la función latente $f(\cdot)$ y no una versión ruidosa de esta.*

# Incorporando información - Evaluación con ruido

Igual que en el caso sin ruido, podemos condicionar esta conjunta a las observaciones y obtenemos el siguiente resultado:

**Lemma**

*Para una evaluación con ruido se tiene que*

$$f(X_*)|Y, X \sim \mathcal{N}(m_{X_*|X}, \Sigma_{X_*|X})$$

*Donde la media y covarianza son:*

$$m_{X_*|X} = m(X_*) + K(X_*, X)[K(X, X) + \sigma_n^2 \mathbb{I}]^{-1}(Y - m(X))$$
$$\Sigma_{X_*|X} = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 \mathbb{I}]^{-1} K(X, X_*)$$

# Incorporando información - Evaluación con ruido

Igual que en el caso sin ruido, podemos condicionar esta conjunta a las observaciones y obtenemos el siguiente resultado:

**Lemma**

*Para una evaluación con ruido se tiene que*

$$f(X_*)|Y, X \sim \mathcal{N}(m_{X_*|X}, \Sigma_{X_*|X})$$

*Donde la media y covarianza son:*

$$m_{X_*|X} = m(X_*) + K(X_*, X)[K(X, X) + \sigma_n^2 \mathbb{I}]^{-1}(Y - m(X))$$
$$\Sigma_{X_*|X} = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 \mathbb{I}]^{-1} K(X, X_*)$$

# Incorporando información - Evaluación con ruido

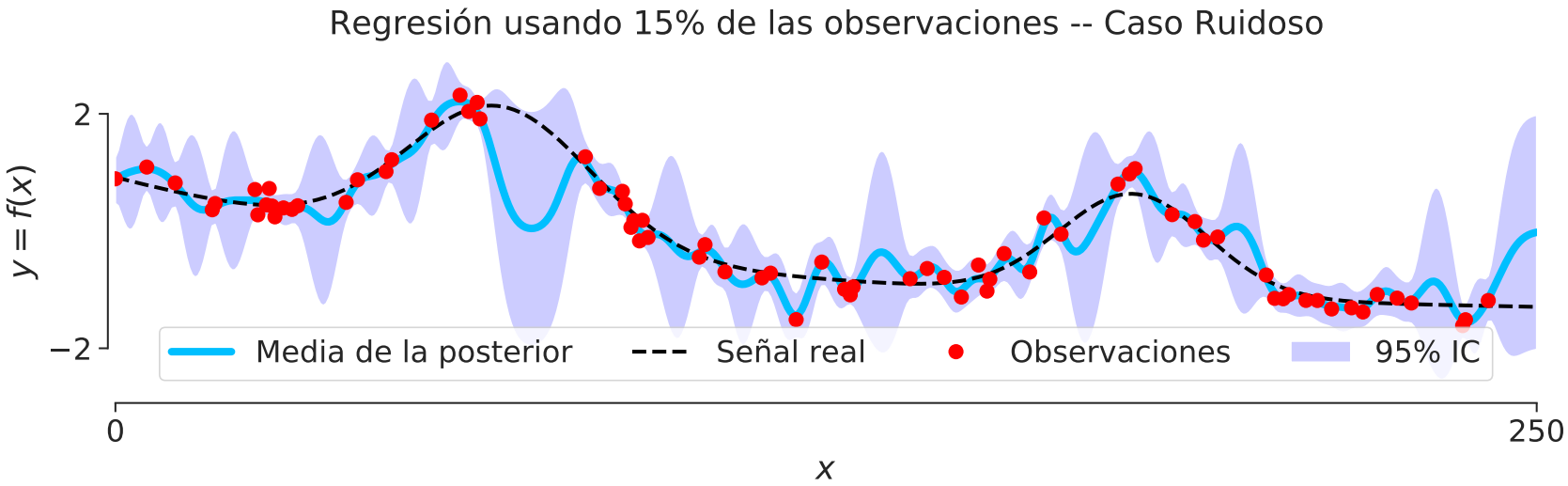


**Fig..** Regresión con $\mathcal{GP}$ para señal sintetica usando el 15 % de los datos muestreados de forma no uniforme y contaminados con ruido Gaussiano, utilizando un $\mathcal{GP}$ de media nula y kernel SE.

# Clase 20 - Procesos Gaussianos I

## Aprendizaje de Máquinas - MA5204

Felipe Tobar

Department of Mathematical Engineering &
Center for Mathematical Modelling
Universidad de Chile

20 de marzo de 2021

UNIVERSIDAD
DE CHILE