Estimación funcional de soluciones probabilísticas en modelos de ecuaciones diferenciales ordinarias

Naomi Cedeño[(1)]; Diego Hernan Suntaxi[(1)]; Saba Infante[(1,2)]
Escuela de Ciencias Matemáticas y Computacionales, Universidad Yachay Tech, Ecuador [(1)]
Departamento de Matemáticas, FACyT, Universidad de Carabobo, Venezuela [(2)]
Email: helen.cedeno@yachaytech.edu.ec; diego.suntaxi@yachaytech.edu.ec;sinfante@yachaytech.edu.ec

October 26, 2021

Abstract

Keywords:

1 Introducción

Los modelos matemáticos utilizados para inferir y predecir sobre fenómenos de la vida real por lo general tienen una estructura no lineal, muchos parámetros, observaciones parcialmente observadas y una dinámica compleja. Estos problemas ocurren con frecuencia en muchos campos de las ciencias: biología, ingeniería, física, medicina, agricultura, economía y finanzas entre otras. La estimación de los parámetros en estos modelos se realiza a partir de observaciones medidas con errores y en algunos casos con datos faltantes. Una forma de cuantificar la información experimental es medir las observaciones de las trayectorias generadas por los sistemas dinámicos y luego se utilizan para estimar los parámetros; es decir, se estiman los estados soluciones y parámetros de modelos de ecuaciones diferenciales ordinarias, que describen la dependencia natural entre los estados del sistema y sus tasas de cambio en un dominio espacio-temporal abierto. En la práctica resolver estos problemas resulta complicado, dado que las variables de estados observadas no son exactamente la solución de la ecuación diferencial, porque se miden con errores y en muchos casos los sistemas no son lineales y no se puede encontrar soluciones de forma analítica. Las soluciones de las ecuaciones dado una condición inicial existen y son únicas, si la función f es continua diferenciable, continua Lipschitz con respecto a la variable de estado x, Ramsay, et. al (2007). Huang et. al (2020) desarrollaron un método Bayesiano para estimar los parámetros en una ecuación diferencia ordinaria (ODE) a partir de datos con ruido, utilizando un procedimiento no paramétrico basado en un algoritmo híbrido Monte Carlo que permite ajustar los datos dentro de la estructura de la verosimilitud conjunta. Tronarp et. al (2019) formularon métodos de aproximaciones numéricas probabilísticas a soluciones de ecuaciones diferenciales ordinarias (EDO) como problemas en la regresión del proceso Gaussiano (GP) con funciones de medición no lineales. Ellos desarrollan nuevos algoritmos de filtrado no lineales utilizando procesos Gaussianos. Además, obtienen aproximaciones no Gaussianos al problema de filtrado mediante el enfoque del filtro de partículas. Las soluciones obtenidas son comparadas con otros métodos probabilísticos conocidos. Heinonen et. al. (2018) proponen un paradigma novedoso de modelado no paramétrico de las EDO, el método permiten que funciones diferenciales aprendan a partir de observaciones de los estados desconocidos utilizando un proceso Gaussiano. Schober at. al. (2019) estudiaron las conexiones entre los ODE solvers y los métodos de regresión probabilísticos. Introducen una nueva visión de los probabilistic ODE solvers para hacer inferencia en modelos de ecuaciones diferenciales estocásticas y estiman la solución del problema de valor inicial a partir de observaciones aproximadas de la derivada de la solución, proporcionadas por la dinámica de la ODE. Chkrebtii et. al. (2016) desarrollaron una estructura para hacer inferencia y cuantificar la incertidumbre en modelos definidos por sistemas generales de ecuaciones diferenciales analíticamente intratables. Este enfoque proporciona una alternativa estadística a la integración numérica determinista para la estimación de sistemas dinámicos complejos, y caracteriza de forma probabilísticas en la incertidumbre de la solución introducida cuando los modelos son caóticos están mal condicionados o contienen incertidumbre funcional no modelada. Ellos consideran la estimación de la solución como un problema de inferencia que permite cuantificar la incertidumbre numérica mediante las herramientas de estimación funciones, que pueden propagarse a través de la incertidumbre en los parámetros del modelo y las predicciones a posterior.

En este trabajo se introduce una metodología de estimación Bayesiana basada de datos funcionales, que se obtienen como realizaciones de funciones de variables aleatorias suaves que varían en un continuo, que se recopilan en tiempos discretos con errores de medición; específicamente, se implementa un algoritmo de muestreo y actualización secuencial en el tiempo de procesos Gaussianos, su derivada y los estados soluciones de una ODE. Las soluciones probabilística están basadas en una regresión de un proceso Gaussiano (GP) prior y una expansión del caos polinomial (PCE).
El resto del artículo es como sigue: en Sesión 2 es establece la metodología: se definen los Procesos Gaussianos,

1. Chkrebtii, O; Campbell, D; Girolami, M; Calderhead, B. (2016). Bayesian Uncertainty Quantification for Differential Equations. Bayesian AnalysisTBA, Number TBA, pp. 1–29
2. Schober, M., Särkkä, S., Hennig, P. (2019). A probabilistic model for the numerical solution of initial value problems. Stat. Comput. 29(1), 99–122. https://doi.org/10.1007/s11222-019-09900-1
3. Heinonen, M; Yildiz, C; MannerstrÖm, H; Intosalmi, J; Lähdesmäki, H. (2018). Learning unknown ODE models with Gaussian processes. arXiv:1803.04303v1 [stat.ML] 12.
4. Tronarp, F; Kersting, H; Särkkä, S; Hennig, P. (2019). Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: a new perspective. Statistics and Computing https://doi.org/10.1007/s11222-019-09900-1.
5. Huang, H; Handel, A; Song, X. (2020). A Bayesian approach to estimate parameters of ordinary differential equation. Computational Statistics https://doi.org/10.1007/s00180-020-00962-8.

2 Fórmulación del modelo

Supóngase que se tiene un sistema dinámico de tiempo continuo que modela fenómenos complejos en campos aplicados tales como: el cambio climático, corrientes oceánicas, finanzas, biología, ingeniería, y medicina, que relacionan las derivadas $\dot{x}(t) \in \mathbb{R}^d$ con respecto a variables que varían en espacio-tiempo $(x,t) \in \mathcal{D} \subset \mathbb{R}^d$ con estados $x(t) \in \mathbb{R}^d$ a través de un campo vectorial de funciones suavizadas Lipschitz continua diferenciable en x denotadas por $\mathbf{f} = (f(x_1), \ldots, f(x_d))$, $f : \mathcal{D} \times \Theta \to \mathbb{R}^d$ indexada por parámetros desconocidos $\theta \in \Theta \subset \mathbb{R}^p$. Si se considera un problema de valor inicial; es decir, una ecuación diferencial ordinaria (ODE) que satisface:

$$\frac{dx}{dt} \equiv \dot{x}(t) = f\left(x(t), \theta, t\right), \quad x(t_0) = x_0 \in \mathbb{R}^d, \quad \textit{initial value}, \quad t \in [0,T] \tag{1}$$

Los estados soluciones del sistema $x(t) \equiv (x_1(t), \ldots, x_K(t))$ evolucionan de acuerdo al modelo descrito por la ecuación (1) y generalmente no están disponibles en forma cerrada. Los sistemas que involucran derivadas de x de orden $n > 1$, se pueden reducir a la expresión dada en la ecuación (1), definiendo nuevas variables

$$x_1 = x, \; x_2 = \dot{x}_1, \ldots, \; x_n = \dot{x}_{n-1} \tag{2}$$

Los métodos numéricos son generalmente utilizados para aproximar el problema de valor inicial $x : [0,T] \to \mathbb{R}^d$. Las soluciones del sistema de ecuaciones diferenciales ordinarias (1) dados los valores iniciales x_0 existen y son únicas en una vecindad de $(0, x_0)$. Sin embargo, en la mayoría de los sistemas ODE no se encuentran soluciones analíticas, lo que implica aumenta el costo computacional de la metodología de ajuste de datos, Ramsay et al. (2007), tradicionalmente se sustituye la solución exacta por una solución aproximada $\hat{x}(t)$ que se obtiene usando alguna técnica numérica sobre una partición de una red en un dominio $\mathcal{D}$ y los procesos de inferencia y predicción están basado en esta solución aproximada.
Por ejemplo, consideremos una ecuación diferencial ordinaria tipo:

$$\frac{dx}{dt} = f\left(x(t)\right), \quad x(0) = x_0 \tag{3}$$

dónde x es una función continua que toma valores en $\mathbb{R}^d$, ϕ_t denota el mapa de flujo de la ecuación diferencial ordinaria dada en (3), y $x_t = \phi_t\left(x(0)\right)$, es la solución de la ecuación. Lo métodos numéricos determinísticos clásicos utilizados para encontrar una solución de la ecuación dada en (3) en un intervalo de tiempo $[0,T]$ permite obtener una aproximación de la solución en una malla de puntos $\{t_k = kh\}_{k=1}^K$ con $Kh = T$, dónde $x_k = x(t_t)$ es la solución en la malla basada en la evaluación de la función f y posiblemente sus derivadas de mayor orden en un conjuto de puntos finitos que son generados por la técnica de integración numérica utilizada. Está metodología genera una única solución discreta $\{x_t\}_{t=1}^T$ posiblemente con error de estimación alto, y no cuantifica la incertidumbre sobre el resto de la trayectoria de la solución.
Sea $\mathcal{X}_{a,b}$ denota un espacio de Banach $\mathcal{C}\left([a,b]; \mathbb{R}^d\right)$. La solución de la ecuación (3) sobre un intervalo

$[0,T]$ puede ser interpretada como una medida δ_x de Dirac sobre el espacio $\mathcal{X}_{0,T}$, dónde un elemento x es solución de la ecuación diferencial ordinaria. Se puede definir una medida de probabilidad μ^h sobre el espacio $\mathcal{X}_{0,T}$ de la cual se puede generar muestras aleatorias dentro y fuera de la malla, ahora se puede cuantificar el tamaño del paso de la discretización, y a través de la distribución de muestreo se puede cuantificar la incertidumbre de la solución restante de la ODE, para mayores detalles ver Conrad et. al. (2015) y sus referencias.
La forma integral desde el punto de vista probabilístico: de la ecuación (3)

$$x(t) = x_0 + \int_0^t f\left(x(s)\right) ds \tag{4}$$

las soluciones en la malla satisfacen

$$x_{k+1} = x_k + \int_{t_k}^t f\left(x(s)\right) ds, \quad t \in [t_k, t_{k+1}] \tag{5}$$

Entonces se puede escribir:

$$x_{k+1} = x_k + \int_{t_k}^t g\left(s\right) ds, \quad t \in [t_k, t_{k+1}] \tag{6}$$

dónde $g\left(t\right) = f\left(x(t)\right)$ es una función desconocida en el tiempo t. La función $g\left(t\right)$ se puede aproximar por un método numérico

$$g^h(s) = \frac{d}{d\tau}\left[\psi_\tau\left(X_k\right)\right]_{\tau=s-t_k}, \quad s \in [t_k, t_{k+1}], \quad X_{k+1} = \psi_k\left(X_k\right) \tag{7}$$

dónde $\psi : \mathbb{R}^d \to \mathbb{R}^d$ denota un integrador numérico clásico determinista de un paso temporal h, una clase que incluye todos los métodos de integración numérica de ODE.
La función g se puede aproximar en forma estocástica de la siguiente manera:

$$g^h(s) = \frac{d}{d\tau}\left[\psi_\tau\left(X_k\right)\right]_\tau + \omega_k\left(\tau\right), \quad \omega_k \sim N\left(0, K^h(s,t)\right), \quad \tau = s - t_k, \quad s \in [t_k, t_{k+1}] \tag{8}$$

dónde $\{\omega_k\}$ es una secuencia de funciones de variables aleatorias Gaussianas definidas en el intervalo $[0,h]$. La covarianza $K^h(s,t)$ se elige de tal manera que este cercana a cero con h a una velocidad establecida, y para asegurar que $\omega \in \mathcal{X}_{0,T}$ casi seguro. Las funciones $\{\omega_k\}$ representan la incertidumbre sobre la función g.
En la práctica los estados soluciones del sistema son desconocidos y sólo se conocen parcialmente con cierto un error ϵ_t que no permite obtener la medición exacta de x_t. Supóngase que se tiene una red de puntos discretos $0 < t_0 < t_1 < \ldots < t_n$ donde se obtienen mediciones de observaciones $x_0, x_1, \ldots, x_n$ dónde $x_i \equiv x_{t_i}$ denotados por $y_1, y_2, \ldots, y_n$ dónde los $y_i \equiv y_{t_i}$ y los y_i son tomados de un modelo de probabilidad con ϵ_i

$$y_i = h\left(x_i, \epsilon_i\right), \quad i = 1, \ldots, n \tag{9}$$

dónde $h(.)$ es una función vectorial conocida de valor real, los ϵ_i son variables independientes idénticamente distribuidas tomadas de una distribución de probabilidad y los $\{y_i\}_{i=1}^n$ son condicionalmente independiente dado los $\{x_i\}_{i=1}^n$. Para el $k-$ésimo estado observado, las n observaciones se obtienen a través del siguiente modelo estadístico:

$$y(t) = h\left(x(t)\right) + \epsilon(t), \quad \epsilon(t) \sim N\left(0, \sigma_\epsilon^2\right) \tag{10}$$

entonces se tiene un vector de estados latentes $X \equiv \left(x(t_1), x(t_2), \ldots, x(t_n)\right)^T$ dónde $k-$ésimo estado es dado por $x_k = \left(x_k(t_1), x_k(t_2), \ldots, x_k(t_n)\right)^T$ y un vector de observaciones $Y \equiv \left(y(t_1), \ldots, y(t_n)\right)^T$, dónde la $k-$ésima observación es $y_k = \left(y_k(t), \ldots, y_k(t)\right)^T$, $k \in \{1, 2, \ldots, K\}$. El modelo espacio estado no lineal continuo-discreto en forma compacta sería:

$$\begin{aligned} &\frac{dx}{dt} \equiv \dot{x}(t) = f\left(x(t), t\right), \quad x(t_0) = x_0, \quad \textit{initial value}, \quad t \in [0,T] \\ &y(t) = h\left(x(t)\right) + \epsilon(t), \quad \epsilon(t) \sim N\left(0, \sigma_\epsilon^2\right) \end{aligned} \tag{11}$$

El problema se puede formular en términos de un sistema dinámico estocástico de tiempo discreto por el siguiente modelo de espacio estado

$$\begin{aligned} &x_k = f\left(x_{k-1}\right) + \nu_{k-1}, \quad \nu_{k-1} \sim N\left(0, Q\right), \quad (\textit{ system state}) \\ &y_k = h\left(x_k\right) + \epsilon_{k-1}, \quad \epsilon_k \sim N\left(0, R\right), \quad (\textit{system measurement}) \\ &x_0 \sim N\left(m_0^x, k_0^x(s_0, t_0)\right) \end{aligned} \tag{12}$$

dónde $x_k \in \mathbb{R}^{d_x}$, $\nu_{k-1} \in \mathbb{R}^{d_y}$, $y_k \in \mathbb{R}^{d_y}$, $\epsilon_k \in \mathbb{R}^{d_y}$, $f : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$, y $h : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$.
La solución Bayesiana del problema de filtrado puede ser obtenida en términos de la predicción y actualización. La distribución filtrada del estado posterior es

$$p\left(x_k|y_{1:k}\right) = \frac{p\left(y_k|x_k\right) p\left(x_k|y_{1:k-1}\right)}{p\left(y_k|y_{1_k-1}\right)} \tag{13}$$

dónde la verosimilitud $p\left(y_k|x_k\right)$ se obtiene de la ecuación de observación (12), y $y_{1:k} = (y_1, \ldots, y_k)$. La densidad predictiva se obtiene por la ecuación de Chapman–Kolmogorov

$$p\left(x_k|y_{1:k-1}\right) = \int p\left(x_k|x_{k-1}\right) p\left(x_k|y_{1:k-1}\right) dx_{k-1} \tag{14}$$

dónde la densidad de transición $p\left(x_k|x_{k-1}\right)$ se obtiene de la ecuación de estado (12). Esto permite obtener un modelo de las observaciones dado por:

2.1 Gaussian Process Priors

Sea $x = \{x_1, x_2, \ldots, x_n\}$ un conjunto de n vectores de entrada descritos $x \in \mathcal{X}$ asociadas con respuestas observadas $y = \{y_1, y_2, \ldots, y_n\}$, $y_i \in \mathbb{R}$, $i = 1, \ldots, n$. Un proceso Gaussiano (GP) es tal que para algún conjunto finito de evaluaciones de funciones:

$$y = f(x) = \left(f(x_1), f(x_2), \ldots, f(x_n)\right)^T \tag{15}$$

Estamos interesados en modelar la incertidumbre de funciones que tienen comportamiento no lineal. Los Procesos Gausianos (GP) proporcionan una forma natural y poderosa de realizar inferencias en torno a tales funciones (Rasmussen y Williams, (2006)). Los GP definen una distribución sobre funciones f dónde f es una función que asigna algún espacio de entradas $\mathcal{X}$ a $\mathbb{R}$; es decir, $f : \mathcal{X} \to \mathbb{R}$. Formalmente, para algún conjunto finito de elementos tomados de $\mathcal{X}$, f es un GP, definido por:

$$f(x) \sim GP\left(m(x), k\left(x, x^{'}\right)\right)$$

dónde:

$$m(x) = \mathbb{E}\left(f(x)\right) = \mathbb{E}\left(f(x_1), f(x_2), \ldots, f(x_n)\right)^T$$

y

$$k\left(x, x^{'}\right) = \mathbb{C}ov\left(f(x), f(x^{'})\right) = \mathbb{E}\left[\left(f(x) - m(x)\right)\left(f(x^{'}) - m(x^{'})\right)\right]$$

Los funciones admitidas para $m(.)$ y $k(.,.)$ satisfacen la condición de que las marginales siguen la distribución Gaussiana; $\mu(.)$ puede ser alguna función paramétrica y $k(.,.)$ es una función que admite una matriz semi-definida positiva cuando se evalúa en los puntos $x \in \mathcal{X}$. Las funciones de media y covarianza mapean el conjunto de indices a los números reales como sigue:

$$m : \mathcal{X} \to \mathbb{R} \quad , \quad k(.,.) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

Sea x_{new} una matríz con cada fila formada secuencia finita de puntos de entradas x_i^{new} nuevos, con $i = 1, \ldots, n$ ($\{x_1^{new}, x_2^{new}, \ldots, x_n^{new}\}$), la matriz de covarianza se puede estimar cómo sigue:

$$k\left(x_{new}, x_{new}\right) = \begin{pmatrix} k\left(x_1^{new}, x_1^{new}\right) & \ldots & k\left(x_1^{new}, x_n^{new}\right) \\ k\left(x_2^{new}, x_1^{new}\right) & \ldots & k\left(x_2^{new}, (x_n^{new}\right) \\ \vdots & \vdots & \vdots \\ k\left(x_n^{new}, x_1^{new}\right) & \ldots & k\left(x_n^{new}, x_n^{new}\right) \end{pmatrix}$$

La función $k\left(x, x^{'}\right)$ modela la dependencia entre los valores funcionales en diferentes puntos de entrada x y $x^{'}$. Una forma usual de elegir $k\left(x, x^{'}\right)$ es considerar la función de base radial definida por:

$$k\left(x, x^{'}\right) = \sigma_f^2 \exp\left(-\frac{\|x - x^{'}\|}{2\lambda^2}\right) \tag{16}$$

dónde la ecuación (16) proporciona un núcleo de transición que sirve para modelar funciones suaves y estacionarias, λ es una parámetro de escala, y σ_f^2 es la varianza de la señal. Sea $m(x) = 0$, se puede muestrear valores de f en las entrada x_{new} usando un GP como sigue:

$$f_{new}(x_{new}) \sim N\left(0, K\left(x_{new}, x_{new}\right)\right) \quad where \quad f_{new} = f_{new}(x_{new}) = \left(f\left(x_1^{new}\right), \ldots, f\left(x_n^{new}\right)\right)^T$$

Supóngase que se tiene una colección de observaciones $\mathcal{D}_t = \{x_t, y_t\}$ y se quiere hacer predicciones usando los datos nuevos x_{new} tomando $f_{new}(x_{new})$ de la distribución a posterior $p\left(f_{new}(x_{new})|\mathcal{D}_t\right)$. Las observaciones $y_t = f(x_t)$ y la función $f_{new}(x_{new})$ sigue una distribución normal multivariada conjunta:

$$\begin{pmatrix} f \\ f_{new} \end{pmatrix} \sim N\left(\begin{pmatrix} m\left(x\right) \\ m(x_{new}) \end{pmatrix}; \begin{pmatrix} K\left(x,x\right) & K\left(x,x_{new}\right) \\ K\left(x_{new},x\right) & K\left(x_{new},x_{new}\right) \end{pmatrix}\right) \tag{17}$$

con $K\left(x,x\right)$, es un núcleo evaluado en x. La distribución marginal $f_{new}|f,x,y$ sigue siendo Gausiana, esto es:

$$f_{new}|f,x,y \sim N\left(m\left(x_{new},\right)^{post}, K^{post}\left(f(x_{new}),f(x_{new})\right)\right)$$

dónde:

$$m^{post}\left(x_{new}\right) = m\left(x_{new}\right) + K\left(x_{new},x\right)K^{-1}\left(x,x\right)\left[f - m\left(x\right)\right]$$

y

$$K^{post}\left(f(x_{new}),f(x_{new})\right) = K\left(x_{new},x_{new}\right) - K\left(x_{new},x\right)K^{-1}\left(x,x\right)K\left(x,x_{new}\right)$$

Generalizando, si observamos los valores reales de f con un ruido aditivo:

$$y_i = f(x_i) + \epsilon_i \quad , \quad \epsilon_i \sim N\left(0,\sigma_\epsilon^2\right) \quad , \quad y|f \sim N\left(f(x_i),\sigma_\epsilon^2 I\right)$$

dónde I es la matriz identidad. El ruido puede ser incluido en la función de covarianza, como sigue:

$$k\left(f(x_i),f(x_j)\right) = k\left(x_i,x_j\right) + \delta_{ij}\sigma_\epsilon^2$$

dónde δ_{ij} es la función delta de Kronecker. La incertidumbre está ahora presente en las observaciones, y la distribución conjunta sobre los datos desconocidos y los datos conocidos se aumentada en la ecuación de covarianza mediante:

$$\begin{pmatrix} f \\ f_{new} \end{pmatrix} \sim N\left(\begin{pmatrix} m\left(x\right) \\ m(x_{new}) \end{pmatrix}; \begin{pmatrix} K\left(x,x\right)+\sigma_\epsilon^2 I & K\left(x,x_{new}\right) \\ K\left(x_{new},x\right) & K\left(x_{new},x_{new}\right)+\sigma_\epsilon^2 \end{pmatrix}\right) \tag{18}$$

$$f_{new}|f,x,y \sim N\left(m\left(x_{new},\right)^{post}, K^{post}\left(f(x_{new}),f(x_{new})\right)\right)$$

dónde:

$$m^{post}\left(x_{new}\right) = m\left(x_{new}\right) + K\left(x_{new},x\right)\left(K\left(x,x\right)+\sigma_\epsilon^2 I\right)^{-1}\left[f - m\left(x\right)\right]$$

y

$$K^{post}\left(f(x_{new}),f(x_{new})\right) = K\left(x_{new},x_{new}\right) - K\left(x_{new},x\right)\left(K\left(x,x\right)+\sigma_\epsilon^2 I\right)^{-1}K\left(x,x_{new}\right) + \sigma_\epsilon^2$$

2.2 Derivative Gaussian process

La flexibilidad de los procesos GP nos permiten modelar las derivadas de datos funciones ver (Solak et al., (2002)), importante en aplicaciones de ingeniería, en sistemas dinámicos, y en el modelamiento de las soluciones de sistemas de ecuaciones diferenciales. La diferenciación es un operador lineal, así que las derivadas de un GP siguen siendo un proceso GP (Rasmussen and Williams, (2006)); por lo tanto, se puede usar un GP para predecir sobre las derivadas, y también usar observaciones derivadas para hacer predicciones. La media del derivado es igual al derivado de la media del proceso latente:

$$\mathbb{E}\left[\frac{\partial f(x)}{\partial x_d}\right] = \frac{\partial \mathbb{E} f(x)}{\partial x_d} \tag{19}$$

La covarianza sobre los valores de la función también implica covarianzas cruzadas entre derivadas

$$\mathbb{C}ov\left[f(x_i),\frac{\partial f(x_j)}{\partial x_{j,d}}\right] = \frac{\partial k\left(x_i,x_j\right)}{\partial x_{j,d}}, \quad i = 1,2,\ldots,K, \quad j = 1,2,\ldots,d \tag{20}$$

$$\mathbb{C}ov\left[\frac{\partial f(x_j)}{\partial x_{j,d}},f(x_i)\right] = \frac{\partial k\left(x_j,x_i\right)}{\partial x_{i,d}}, \quad \textbf{and} \quad \mathbb{C}ov\left[\frac{\partial f(x_i)}{\partial x_{i,k}},\frac{\partial f(x_j)}{\partial x_{j,d}}\right] = \frac{\partial^2 k\left(x_i,x_j\right)}{\partial x_{j,k}\partial x_{j,d}} \tag{21}$$

dónde $x_{i,k}$ denota el $k-$ésimo elemento de x_i. Este resultado nos permite definir una prior sobre el GP derivado en términos de la prior del GP; es decir,

$$f \sim GP\left(m(.), k_{(f,f)}(.,.)\right) \quad and \quad \nabla f \sim GP\left(m^{'}(.), k^{'}(.,.)\right) \tag{22}$$

La distribución conjunta del proceso GP con incertidumbre y el proceso GP derivado es:

$$\begin{pmatrix} f \\ \nabla f \end{pmatrix} \sim GP\left[\begin{pmatrix} f \\ \nabla f \end{pmatrix}; \begin{pmatrix} m(.) \\ m^{'}(.) \end{pmatrix}, \begin{pmatrix} k\left(x_i, x_j\right) & \frac{\partial k\left(x_i, x_j\right)}{\partial x_{j,d}} \\ \frac{\partial k\left(x_j, x_i\right)}{\partial x_{i,d}} & \frac{\partial^2 k\left(x_i, x_j\right)}{\partial x_{i,k} \partial x_{j,d}} \end{pmatrix}\right]$$

Supongamos que queremos predecir nuevos valores observados $y_* = f_*(x_*) + \epsilon$, $\epsilon \sim N\left(0, \sigma_\epsilon^2\right)$, la media del derivado de la función $f_*(x_*) = (f_*(x_{*,1}), \ldots, f_*(x_{*,d}))^T$, con respecto a la dimensión d es:

$$\mathbb{E}\left[\frac{\partial f_*}{\partial x_{*,d}}\right] = \frac{\partial \mathbb{E}\left(f_*\right)}{\partial x_{*,d}} = \frac{\partial m^{'}\left(x_*\right)}{\partial x_{*,d}} = \frac{\partial k\left(x_*, x\right)}{\partial x_{*,d}} \times \left[k\left(x, x\right) + \sigma_\epsilon^2 I\right]^{-1} y \tag{23}$$

El cálculo de la varianza es dado en Riihimä and Vehtari, (2010):

$$\mathbb{V}ar\left[\frac{\partial f_*}{\partial x_{*,d}}\right] = \frac{\partial^2 k\left(x_*, x_*\right)}{\partial x_{*,d} \partial x_{*,d}} - \frac{\partial k\left(x_*, x\right)}{\partial x_{*,d}} \times \left[k\left(x, x\right) + \sigma_\epsilon^2 I\right]^{-1} \times \frac{\partial k\left(x, x_*\right)}{\partial x_{*,d}} \tag{24}$$

dónde: $x = (x_1, \ldots, x_d)^T$, $x_* = (x_{*,1}, \ldots, x_{*,d})^T$, y $k\left(x_*, x\right) = \left(k\left(x_{*,d}, x_1\right), \ldots, k\left(x_{*,d}, x_d\right)^T\right)$. Ahora se puede escribir la distribución a posterior del proceso derivativo de las observaciones:

$$\frac{\partial f_*}{\partial x_{*,d}} \sim GP\left(\mathbb{E}\left[\frac{\partial f_*}{\partial x_{*,d}}\right], \mathbb{V}ar\left[\frac{\partial f_*}{\partial x_{*,d}}\right]\right) \tag{25}$$

Para el caso particular, si se considera la covarianza exponencial:

$$\mathbb{C}ov\left(f(x_i), f(x_j)\right) = k\left(x_i, x_j\right) = \sigma_f^2 \exp\left(-\frac{1}{2}\frac{\|x^{(i)} - x^{(j)}\|^2}{\lambda^2}\right) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{1}{\lambda_d^2}\left(x_d^{(i)} - x_d^{(j)}\right)^2\right) \tag{26}$$

dónde σ_f^2, y $\lambda = (\lambda_1, \ldots, \lambda_d)$ son los hiperparámetros del modelo GP. Los derivados de las observaciones son:

$$\frac{\partial}{\partial x_d^{(i)}} k\left(x_i, x_j\right) = \sigma_f^2 \times \left(-\frac{1}{\lambda_d^2}\left(x_d^{(i)} - x_d^{(j)}\right)\right) \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{1}{\lambda_d^2}\left(x_d^{(i)} - x_d^{(j)}\right)^2\right) \tag{27}$$

y

$$\begin{aligned} \frac{\partial^2}{\partial x_g^{(i)} \partial x_h^{(j)}} k\left(x_i, x_j\right) = \sigma_f^2 \times \frac{1}{\lambda_g^2} \times \left(\delta_{gh} - \frac{1}{\lambda_h^2}\left(x_h^{(i)} - x_h^{(j)}\right)\left(x_g^{(i)} - x_g^{(j)}\right)\right) \\ \times \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{1}{\lambda_d^2}\left(x_d^{(i)} - x_d^{(j)}\right)^2\right) \end{aligned} \tag{28}$$

dónde $\delta_{gh} = 1$ si $g = h$ y $\delta_{gh} = 0$ si $g \neq h$, ver Riihimäki and Vehtari, (2010).

2.3 Modelo probabilístico para caracterizar solución de una EDO

En está sesión se introduce un modelo probabilístico que permite caracterizar la estimación del error de la solución de una EDO con valor inicial, dónde los estados soluciones no están definidos en forma explícitos, y no están disponibles en forma cerrada. El modelo considerado es de la forma (11) o (12), dónde se reemplaza la solución exacta $x(t)$ con una representación finita dimensional $x^n(t, \theta)$,

$$\dot{x}(t) = f\left(x(t), t\right), \quad x(t_0) = x_0 \tag{29}$$

$$y(t) = h\left(x^n(t, \theta)\right) + \epsilon(t), \quad \epsilon(t) \sim N\left(0, \sigma_\epsilon^2\right) \tag{30}$$

La ecuación (29) representa un modelo teórico no observado, mientras que la ecuación (30) representa un modelo de observación. Formulamos la aproximación de (29) en puntos discreto $\{t_i\}_{i=1}^n$ cómo un problema de inferencia Bayesiana, dónde se requiere conocer una medida de probabilidad prior, una verosimilitud, y a partir de esta información se define una medida de probabilidad a posterior mediante el Teorema de Bayes. Si $y_{1:t} := (y_1, y_2 \ldots, y_n)$, representa los datos observados, $x_{1:t} := (x_1, x_2 \ldots, x_n)$ representan

los estados desconocidos, $p\left(y_{1:t}|x_{1:t}\right)$ es la verosimilitud, $p_y\left(\theta\right)$ es la distribución prior sobre el espacio solución, la distribución posterior se obtiene:

$$p\left(y_{1:t}|x_{1:t}\right) = \frac{p\left(y_{1:t}|x_{1:t}p_y\left(\theta\right)\right)}{\int p\left(y_{1:t}|x_{1:t}p_y\left(\theta\right)\right)d\theta} \tag{31}$$

Siguiendo la teoría desarrollada en los trabajos de Skilling (1991) y Chkrebttii et al. (2013), se propone modelar la incertidumbre de la solución de una EDO mediante un proceso GP en un espacio de funciones suavizadas en un intervalo $[0, T]$, mediante la implementación de un algoritmo computacional eficiente que permita estimar los estados soluciones $x(t)$ y el derivado de la solución en forma secuencial bajo las condiciones establecidas en el modelo definido en la ecuación (1). La existencia de las soluciones están bien fundamentadas ver Butcher (2008), la condición inicial $x(0) = x_0$ representa la solución exacta en las condiciones de borde, el campo vectorial $f\left(x(t), t, \theta\right)$ proporciona una aproximación a la derivada $\dot{f}(x)$ en un dominio $[0, T]$. Se supone qué existe una constante C es tal que:

$$\|f\left(x\right) - f\left(y\right) \leq C\|x - y\|, \quad \textit{for any} \quad x, y \in \mathbb{R}^d \quad \textit{Lipschitz condition} \tag{32}$$

El procedimiento para estimar la derivada de la curva de las observaciones con ruido, es el siguiente; sea (x_i, y_i), $i = 1, 2, \ldots, n$ las observaciones dónde x_i es la varible de entrada y $y_i = f(x_i) + \epsilon_i$ es la variable respuesta, dónde el error $\epsilon_i \sim N\left(0, \sigma_\epsilon^2\right)$. Interesa estimar $\dot{f}(x)$ sin hacer ningún supuesta. Una forma es suponer que:

$$f(x) \sim GP\left(m(.), k\left(.,.\right)\right) \tag{33}$$

Yaglom (1962) demostró que si $k\left(.,.\right)$ es diferenciable dos veces en el origen, $f(x)$ es diferenciable en media cuadrática; entonces:

$$\dot{f}(x) \sim GP\left(0, \frac{dk\left(x, x^{'}\right)}{dxdx^{'}}\right) \tag{34}$$

Por lo tanto, dada la forma paramétrica de $k\left(.,.\right)$, se pueden usar los datos (x_i, y_i) para estimar los parámetros de $k\left(.,.\right)$ y usar

$$k^{''}(.,.) = \frac{dk\left(x, x^{'}\right)}{dxdx^{'}} \tag{35}$$

para hacer inferencia sobre $\dot{f}(x)$.
Un método alternativo es el siguiente: supóngase que el proceso derivado $\dot{y} = \dot{f}(x)$ es un GP. El proceso integrado con observaciones ruidosas es:

$$f(x) = \int_0^x f(s)ds \tag{36}$$

La función de covarianza de $f(x)$ puede ser obtenida usando una integración doble sobre la función de covarianza de $\dot{f}(x)$, si suponemos que:

$$\dot{y} = \dot{f}(x) \sim GP\left(0, k\left(x, x^{'}\right)\right) \tag{37}$$

entonces,

$$y = f(x) \sim GP\left(0, \int_0^x \int_0^{x^{'}} k\left(s, s^{'}\right) dsds^{'}\right) \tag{38}$$

En contexto de este trabajo, las soluciones de la EDO se aproximan en n puntos $\{t_1, t_2, \ldots, t_n\}$ y se denotan por

$$\hat{x}(t) = \left(\hat{x}(t_1), \hat{x}(t_2), \ldots, \hat{x}(t_n)\right)^T \tag{39}$$

Se evalúa el campo vectorial $y = f\left(\hat{x}(t)\right)$;

$$f(x) = \left(f\left(\hat{x}(t_1)\right), f\left(\hat{x}(t_2)\right) \ldots, f\left(\hat{x}(t_n)\right)\right)^T \tag{40}$$

Ahora bien, la verosimilitud del modelo depende de la ecuación (38), lo que implica que para cada par de localizaciones $(x(t_i), x(t_j))$ se requiere calcular un integral doble y cuando n es grande esto implica un alto costo computacional. Una alternativa es considerar un conjunto de puntos $\{\tau_1, \tau_2, \ldots, \tau_n\}$ y se obtiene:

$$\hat{x}^*(t) = \left(\hat{x}^*(\tau_1), \hat{x}^*(\tau_2), \ldots, \hat{x}^*(\tau_n)\right)^T \tag{41}$$

y se evalúa el campo vectorial $\dot{y} = \dot{f}\left(\hat{x}^*(t)\right)$;

$$\dot{f}(x) = \left(\dot{f}\left(\hat{x}(\tau_1)\right), \dot{f}\left(\hat{x}(\tau_2)\right) \ldots, \dot{f}\left(\hat{x}(\tau_n)\right)\right)^T \tag{42}$$

luego condicionado la distribución conjunta $(y, \dot{y})^T$ sobre $\dot{y}$, permite simplificar el cálculo computacional. Para hacer inferencia se estima el modelo probabilístico que cuantifica el error de estimación cometido al aproximar la solución en una red finita el proceso estocástico de dimensión infinita que representa el modelo continuo. Para llevar a cabo esto se procede como sigue: se fijan los parámetros θ del modelo continuo, los hiperparámetros $\Sigma = \left(\sigma_f^2, \lambda, \sigma_\epsilon^2, x_0\right)$ asociados con los errores de estimación del modelo, y se escribe la verosimilitud del modelo, que representa la solución probabilística del sistema:

$$p\left(x\left(t,\theta\right), y, \dot{y}, \theta, \Sigma\right) \propto p\left(x\left(t,\theta\right)|y, \dot{y}, \theta, \Sigma\right) p\left(y|\dot{y}, \theta, \Sigma\right) p\left(\dot{y}|\Sigma\right) \tag{43}$$

La distribución a posterior se estima por:

$$p\left(x\left(t,\theta\right), \theta, \Sigma|y(t), \dot{y}(t)\right) \propto p\left(x\left(t,\theta\right)|y, \dot{y}, \theta, \Sigma\right) p\left(y|\dot{y}, \theta, \Sigma\right) p\left(\dot{y}|\Sigma\right) p\left(\theta, \Sigma\right) \tag{44}$$

En este estudio se propone estimar la solución de una EDO cómo un problema de inferencia estadística, definido en la estructura de los modelos espacio estado (29) y (30). Se considera una distribución prior para modelar la función de la solución de la EDO considerando un proceso GP y $d-1$ derivados:

$$\left(x(t), \dot{x}(t), x^{(2)}(t), \ldots, x^{(q-1)}(t)\right) : [0,T] \to \mathbb{R}^d \tag{45}$$

que se obtienen de un proceso de Wiener integrado $d-$veces $x = \left(x_t^{(1)}, \ldots, x_t^{(d)}\right)^T$, $t \in [0,T]$; es decir, la dinámicas de x_t representan las soluciones de la EDO.
Consideramos distribuciones a priors dados por una GP como sigue:

$$x(t) \sim GP\left(m(.), k(t,t^{'})\right) \tag{46}$$

dónde $m(.)$ es una función de media, $k(t,t^{'})$ es la función de covarianza. El vector

$$x(t) = \left[x^{(1)}(t), x^{(2)}(t), \ldots, x^{(q-1)}(t)\right]^T \tag{47}$$

dónde $x^{(1)}(t)$ y $x^{(2)}(t)$ modelan $x(t)$ and $\dot{x}(t)$, respectively. El resto de los $q-1$ sub-vectors en $u(t)$ pueden ser usados para modelar los derivados de orden mayor en $x(t)$, estó fue realizado por Schober et al. (2019) y Kersting and Hennig (2016).
La motivación de este trabajo se enfoca sobre la utilización de un método probabilístico que nos permite cuantificar los errores que surgen cuando se aproximan las soluciones en sistemas de ecuaciones diferenciales a partir de esquemas de discretizados. Se introduce una medida prior mediante un campo vectorial aleatorio Gaussiano que refleja las incertidumbre sobre las aproximaciones del modelo. La prior tiene la ventaja que permite incluir información sobre las cantidades de interés (estados soluciones y su derivado), y permite la implementación de algoritmos que facilitan los cálculos computacionales. Siguiendo los trabajos desarrollados por Chkrebtii et al. (2016) y más recientemente, Tronarp et al, (2019), quienes propusieron una distribución prior inicial conjunta de la incertidumbre a través de la solución y su derivado $(\dot{x}(t), x(t))^T$ usando una prior GP, con vector de medias dados por $\dot{m}_0$, m_0, matriz de varianza-covarianza $\dot{k}_0(t,t^{'})$, $k_0(t,t^{'})$, matriz de covarianzas cruzadas $\tilde{k}_0(t,t^{'})$, y $\tilde{k}_0(t^{'},t)$, respectivamente. Se impone las siguientes restricciones sobre la media marginal:

$$m_0(t) = \int_0^t \dot{m}_0(z)dz + x_0$$

y sobre la covarianza del derivado $\dot{k}_0(0,0) = 0$, para asegura hacer cumplir la condición inicial $x(0) = x_0$. La distribución prior conjunta GP inicial para la solución en un vector de tiempos de evaluación t_i y su derivado temporal en un vector de tiempos de evaluación posiblemente diferente t_j, es dado por:

$$\left(\dot{x}(t_i), x(t_j)\right)^T|f_0 \;\; \sim GP\left(\begin{pmatrix} \dot{m}_0(t_i) \\ m_0\left(t_j\right) \end{pmatrix}; \begin{pmatrix} \dot{k}_0(t_i,t_i) & \tilde{k}_0\left(t_i,t_j\right) \\ \tilde{k}_0\left(t_j,t_i\right) & k_0\left(t_j,t_j\right) \end{pmatrix}\right) \tag{48}$$

La matriz $\dot{k}_0\left(t_i,t_j\right)$ tiene entradas dadas por:

$$\dot{k}_0\left(t_i,t_j\right) = \int_0^{t_i}\int_0^{t_j} \dot{k}_0(z,w)dzdw$$

y las matrices de covarianza-cruzada tienen entradas dadas por:

$$\tilde{k}_0\left(t_i,t_j\right)=\int_0^{t_j}\dot{k}_0\left(t_i,z\right)dz,\quad and\quad \tilde{k}_0\left(t_j,t_i\right)=\int_0^{t_j}\dot{k}_0\left(z,t_i\right)dz$$

La solución $x(t)$ y su derivado $\dot{x}(t)$ en el modelo (48) se puede actualizar condicionando la información del modelo sobre la partición $\tau=(\tau_1,\ldots,\tau_n)$. Se evalúa el modelo en

$$f_n=f\left(\tau_n,x(\tau_n),\theta\right)=\dot{x}(\tau_n)$$

de la distribución predictiva a posterior marginal $x\left(\tau_n\right)|\dot{x}(\tau_{n-1})$. Las distribuciones marginales siguen la distribución de un GP, (Solak et al (2003), Chkrebtii et al. (2016), y Overstall et al. (2020)). El vector solución:

$$x(t_j)|\dot{x}(t_i)\sim GP\left(\dot{m}_0,\dot{k}(.,.)\right)\Big)$$

en forma análogo se obtiene la distribución marginal de los derivados:

$$\dot{x}(t_i)|x(t_j)\sim GP\left(m_0,k_0(.,.)\right) \tag{49}$$

La actualización de la ecuación (48) se realiza en forma secuencial, el derivado de la solución exacta en la condición inicial $\dot{x}_0^*\left(\tau_1\right)$ en $\tau_1=0$ se obtiene evaluando la función:

$$f_1=f\left(\tau_1,x_0^*(\tau_1),\theta\right)=\dot{x}^*\left(\tau_1\right) \tag{50}$$

la próxima iteración será:

$$\left(\dot{x}(t_i),x(t_j)\right)^T|f_1\ \sim N\left(\left(\begin{array}{c}\dot{m}_1(t_i)\\ m_1\left(t_j\right)\end{array}\right);\left(\begin{array}{cc}\dot{k}_1(t_i,t_i) & \tilde{k}_1\left(t_i,t_j\right)\\ \tilde{k}_1\left(t_j,t_i\right) & k_1\left(t_j,t_j\right)\end{array}\right)\right) \tag{51}$$

Las medias y covarianzas se actualizan en los vectores de tiempo t_i y t_j:

$$\dot{m}_1(t_i)=\dot{m}_0(t_i)+\dot{k}_0\left(t_i,\tau_1\right)\dot{k}_0\left(\tau_1,\tau_1\right)^{-1}\left[f_1-\dot{m}_0(\tau_1)\right]$$

$$m_1\left(t_j\right)=m_0(t_j)+\dot{k}_0\left(\tau_1,\tau_1\right)^{-1}\tilde{k}_0\left(t_j,\tau_1\right)\left[f_1-\dot{m}_0(\tau_1)\right],$$

$$\dot{k}_1(t_i,t_i)=\dot{k}_0(t_i,t_i)-\dot{k}_0\left(t_i,\tau_1\right)\dot{k}_0\left(\tau_1,\tau_1\right)^{-1}\dot{k}_0\left(\tau_1,t_i\right)$$

y

$$\dot{k}_1(t_j,t_j)=\dot{k}_0(t_j,t_j)-\tilde{k}_0\left(t_j,\tau_1\right)\dot{k}_0\left(\tau_1,\tau_1\right)^{-1}\left(\tilde{k}_0\left(t_j,\tau_1\right)\right)^T$$

La segunda realización f_2 se obtiene simulando $x(\tau_2)$ en el tiempo τ_2 de la distribución predictiva posterior:

$$x\left(\tau_2\right)|f_1\sim GP\left(m_1\left(\tau_2\right),\dot{k}_1(\tau_1,\tau_2)\right)$$

y se evalúa:

$$f_2=f\left(\tau_2,x_1(\tau_2),\theta\right)=\dot{x}(\tau_2)$$

El valor simulado $x\left(\tau_2\right)|f_1$ no garantiza que f_2 sea igual al derivado en el tiempo τ_2, esto implica que probablemente se comete un error $\epsilon_{\tau_2}^2=\left(f_2-\dot{x}(\tau_2)\right)^2$. En general el error total se puede cuantificar, como sigue:

$$\sum_i^d\epsilon_{\tau_i}^2=\sum_i^d\left(\dot{x}(\tau_i)-f\left(\tau_i,x(\tau_i),\theta\right)\right)^2$$

Una forma natural de obtener el nuevo dato es simular de un GP, como sigue:

$$f_2|f_1\sim GP\left(\dot{\mu}\left(\tau_2\right),\Lambda_1(\tau_2)\right)$$

dónde la media $\dot{\mu}\left(\tau_2\right)=\dot{x}\left(\tau_2\right)$ y la varianza $\Lambda_1(\tau_2)=\dot{k}_1(\tau_2,\tau_2)$.
El proceso continua, y en la siguiente iteración se actualiza la ecuación (51):

$$\left(\dot{x}(t),x(t)\right)^T|f_2,f_1\ \sim N\left(\left(\begin{array}{c}\dot{m}_2(t_i)\\ m_2\left(t_j\right)\end{array}\right);\left(\begin{array}{cc}\dot{k}_2(t_i,t_i) & \tilde{k}_2\left(t_i,t_j\right)\\ \tilde{k}_2\left(t_j,t_i\right) & k_2\left(t_j,t_j\right)\end{array}\right)\right) \tag{52}$$

Las medias y covarianzas marginales evaluadas en los vectores de tiempo t_i y t_j se actualizan por:

$$\dot{m}_2(t_i) = \dot{m}_1(t_i) + \dot{k}_1\left(t_i, \tau_2\right)\left(\dot{k}_1\left(\tau_2, \tau_2\right) + \Lambda_1(\tau_2)\right)^{-1}\left[f_2 - \dot{m}_1(\tau_2)\right]$$

$$m_2\left(t_j\right) = m_1(t_j) + \tilde{k}_1\left(t_j, \tau_2\right)\left(\dot{k}_1\left(\tau_2, \tau_2\right) + \Lambda_1(\tau_2)\right)^{-1}\left[f_2 - \dot{m}_1(\tau_2)\right],$$

$$\dot{k}_2(t_i, t_i) = \dot{k}_1(t_i, t_i) - \dot{k}_1\left(t_i, \tau_2\right)\left(\dot{k}_1\left(\tau_2, \tau_2\right) + \Lambda_1(\tau_2)\right)^{-1}\dot{k}_1\left(\tau_2, t_i\right)$$

y

$$\dot{k}_2(t_j, t_j) = \dot{k}_1(t_j, t_j) - \tilde{k}_1\left(t_j, \tau_2\right)\left(\dot{k}_1\left(\tau_2, \tau_2\right) + \Lambda_1(\tau_2)\right)^{-1}\left(\tilde{k}_1\left(t_j, \tau_2\right)\right)^T$$

Para la $n-$ésima iteración se genera el dato $x(\tau_n)$ en el tiempo τ_n de la distribución marginal a posterior predicitva,

$$x\left(\tau_n\right)|f_{n-1}, \ldots, f_1 \sim GP\left(\dot{m}\left(\tau_n\right), \Lambda_{n-1}\left(\tau_n\right)\right)$$

donde $\Lambda_{n-1}(\tau_n) = \dot{k}_{n-1}(\tau_n, \tau_n)$.
La actualización del modelo predictivo definido en la ecuación (51), se lleva acabo evaluando en los tiempos t_i y t_j como sigue:

$$\left(\dot{x}(t), x(t)\right)^T|f_n, \ldots, f_1 \sim N\left(\begin{pmatrix} \dot{m}_n(t_i) \\ m_n\left(t_j\right) \end{pmatrix}; \begin{pmatrix} \dot{k}_n(t_i, t_i) & \tilde{k}_n\left(t_i, t_j\right) \\ \tilde{k}_n\left(t_j, t_i\right) & k_n\left(t_j, t_j\right) \end{pmatrix}\right) \tag{53}$$

Las medias y covarianzas marginales evaluadas en los vectores de tiempo t_i y t_j se actualizan por:

$$\dot{m}_n(t_i) = \dot{m}_{n-1}(t_i) + \dot{k}_{n-1}\left(t_i, \tau_n\right)\left(\dot{k}_{n-1}\left(\tau_n, \tau_n\right) + \Lambda_{n-1}(\tau_n)\right)^{-1}\left[f_n - \dot{m}_{n-1}(\tau_n)\right]$$

$$m_n\left(t_j\right) = m_{n-1}(t_j) + \left(\dot{k}_{n-1}\left(\tau_n, \tau_n\right) + \Lambda_{n-1}(\tau_n)\right)^{-1}\tilde{k}_n\left(t_j, t_i\right)\left[f_n - \dot{m}_{n-1}(\tau_n)\right]$$

$$\dot{k}_n(t_i, t_j) = \dot{k}_{n-1}(t_i, t_j) - \dot{k}_{n-1}\left(t_i, \tau_n\right)\left(\dot{k}_{n-1}\left(\tau_n, \tau_n\right) + \Lambda_{n-1}(\tau_n)\right)^{-1}\left(\tilde{k}_{n-1}\left(\tau_n, t_j\right)\right)^T$$

y

$$\dot{k}_n(t_i, t_j) = \dot{k}_{n-1}(t_i, t_j) - \left(\dot{k}_{n-1}\left(\tau_n, \tau_n\right) + \Lambda_{n-1}(\tau_n)\right)^{-1}\tilde{k}_{n-1}\left(t_i, \tau_n\right)\left(\tilde{k}_{n-1}\left(t_j, \tau_n\right)\right)^T$$

Chkrebtii et al. (2016), y Overstall et al. (2020) propusieron el siguiente algoritmo secuencial para actualizar y muestrear en los tiempos $\mathbf{t} = (t_1, \ldots, t_n)^T$, un proceso Gaussiano bivariado entre la solución $x(t)$ y su derivado $\dot{x}(t)$, que representan los estados de evolución $x(t) = (x(t_1), \ldots, x(t_n))^T$ que gobiernan un sistema de ecuaciones diferenciales $\dot{x}(t) = f\left(t, x(t), \theta\right)^T$, y una red de puntos $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^T$, utilizados para evaluar los derivados $\dot{x}(t) = (\dot{x}(\tau_1), \ldots, \dot{x}(\tau_n))^T$, respectivamente.

Algorithm: 1

1. Sean: $\mathbf{x}(0) = \mathbf{x}_0$ un valor inicial, $\tau_1 = t_0$, $\boldsymbol{\Lambda}_1 = 0$, y $f_1 = f\left(x_0, t_0, \theta\right)$.
2. Para $r = 1, \ldots, n-1$, hacer
   (a) Hacer $\boldsymbol{\tau}_r = (\tau_1, \ldots, \tau_r)^T$, $\mathbf{t}_r = (t_1, \ldots, t_r)^T$
   (b) Calcular:

$$\mathbf{B}_r = \left(\dot{k}_0\left(\boldsymbol{\tau}_r, \boldsymbol{\tau}_r\right) + \boldsymbol{\Lambda}_r\right)^{-1} \tag{54}$$

$$\mathbf{a}_r = \mathbf{B}_r \times \tilde{k}_0\left(\boldsymbol{\tau}_r, \tau_{r+1}\right) \tag{55}$$

$$k_r = k_0\left(\tau_r, \tau_r\right) - \tilde{k}_0\left(\tau_{r+1}, \boldsymbol{\tau}_r\right) \times \mathbf{B}_r \times \tilde{k}_0\left(\boldsymbol{\tau}_r, \tau_{r+1}\right) \tag{56}$$

$$\dot{k}_{r+1} = \dot{k}_0\left(\tau_{r+1}, \tau_{r+1}\right) - \dot{k}_0\left(\tau_{r+1}, \boldsymbol{\tau}_r\right) \times \mathbf{B}_r \times \dot{k}_0\left(\boldsymbol{\tau}_r, \tau_{r+1}\right) \tag{57}$$

$$\boldsymbol{\Lambda}_{r+1} = diag\left(\boldsymbol{\Lambda}_r, \dot{k}_{r+1}\right) \tag{58}$$

(c) Calcular:

$$\mathbf{m}_r = \mathbf{x}_0 + \mathbf{F}_r^T \times \mathbf{a}_r$$

dónde $\mathbf{F}_r$ es una matriz de orden $r \times s$ formada por los elementos de f; esto es, la $i-$ésima fila es dada por f_i, para $i = 1, \ldots, n-1$.

(d) Muestrear:

$$\mathbf{x}\left(\tau_{r+1}\right) \sim GP\left(m_r, k_r I_S\right) \tag{59}$$

y se evalúa:

$$f_{\tau+1} = f\left(\mathbf{x}(\tau_{r+1}), \tau_{r+1}, \theta\right) \tag{60}$$

3. Calcular:

$$\mathbf{B}_n = \left(\dot{k}_0\left(\boldsymbol{\tau}_n, \boldsymbol{\tau}_n\right) + \boldsymbol{\Lambda}_n\right)^{-1} \tag{61}$$

$$\mathbf{A}_n(\mathbf{t}) = \mathbf{B}_n \times \tilde{k}_0\left(\boldsymbol{\tau}, \mathbf{t}\right) \tag{62}$$

$$\mathbf{M}_n(\mathbf{t}) = 1_m \times \mathbf{x}_0^T + \mathbf{A}_n(\mathbf{t}) \times \mathbf{F}_n \tag{63}$$

dónde 1_m es el $m-$vector con todas las entradas iguales a uno y F_n es una matriz $n \times s$ con la $k-$ésima fila dada por f_k, $k = 1, \ldots, n$, y

$$k_n\left(\mathbf{t}, t\right) = k_0(\mathbf{t}) - \tilde{k}_0\left(t, \boldsymbol{\tau}\right) \mathbf{B}_N \tilde{k}_0\left(\boldsymbol{\tau}, \mathbf{t}\right)$$

4. Para $l = 1, \ldots, s$, muestrear

$$x_l(t_1), \ldots, x_l(t_n) | \dot{x}_l(\tau_1), \ldots, \dot{x}_l(\tau_n) \sim N\left(\mathbf{M}_N(t) \times \mathbf{e}_l, k_n\left(\mathbf{t}, \mathbf{t}\right)\right)$$

dónde $\mathbf{e}_l$ es el $l-$ésimo vector unitario.

dada una solución inicial $\mathbf{x}_0$ y los parámetros θ, el modelo prior GP puede actualizar las soluciones usando las evaluaciones de la derivada sobre la red de puntos $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^T$, mediante la implementación del algoritmo 1, condicionando secuencialmente sobre $f\left(x, \tau_{r+1}; \theta\right)$ calculando para el estado solución x_l muestreando de la distribución a posterior en el punto τ_r. La distribución marginal GP para $x_l(t)$ es dada por:

$$m_{nl}(\mathbf{t}) = m_{0l} + \tilde{k}_0\left(\mathbf{t}, \boldsymbol{\tau}\right) \times \mathbf{B}_n \times \mathbf{F}_n \times \mathbf{e}_l$$

y

$$k_n\left(\mathbf{t}, \mathbf{t}^{'}\right) = k_0\left(\mathbf{t}, \mathbf{t}^{'}\right) - \tilde{k}_0(\mathbf{t}, \boldsymbol{\tau}) \times \mathbf{B}_n \times \tilde{k}_0(\boldsymbol{\tau}, \mathbf{t})$$

para $l = 1, \ldots, s$, dónde $\mathbf{F}_n$ es la matriz de las evaluaciones de las derivadas y $\mathbf{B}_n$ es la matriz de covarianza de las derivadas actualizadas.

2.4 Método de estimación funcional en modelos de ODE

Regresando al escenario que nos interesa, encontrar soluciones aproximadas del sistema dinámico $\dot{X}(t) = f\left(X(t), t\right)$, $t \in [0, T]$, $X(0) = x_0$. Las soluciones de este sistema dinámico se pueden considerar como realizaciones de un proceso estocástico que varía sobre un continúo, de los cuales sólo se tienen observaciones medidas con errores sobre una red de puntos discretos. Para tratar de estimar estás soluciones proponemos un modelo Bayesiano jerárquico. Supóngase que los datos funcionales contienen n trayectorias independientes, denotadas por:

$$\{Y_i(t_{ij}); \quad i = 1, \ldots, n; \quad j = 1, \ldots, p_i; \quad t_{ij} \in [0, T]\} = \{Y_1(t_{ij}), Y_2(t_{ij}), \quad , Y_n(t_{ij}); \quad t_{ij} \in [0, T]\} \tag{64}$$

La $i-$ésima trayectoria tiene p_i observaciones tomadas en una red de puntos: $t_i = (t_{1i}, \ldots, t_{ip_i})$. Supóngase que $i-$ésima trayectorias $Y_i(.)$ dependen de un proceso subyacente de datos funcionales latente verdadero pero desconocido $X_i(.)$, a través de un modelo funcional Bayesiano jerárquico dado por:

$$\begin{aligned}
&\dot{X}(t_i) = f\left(X(t_i), t_i\right), \quad t_i \in [0,T], \quad X(0) = x_0 \\
&Y_i(t_i) = X_i(t_i) + \epsilon_i(t_i), \quad \epsilon_i(t_i) \sim N\left(0, \sigma_\epsilon^2\right), \quad Y_i(t_i) \sim N\left(X_i(t_i), \sigma_\epsilon^2 \mathbf{I}\right) \\
&X_i(t_i)|\mu_X, \Sigma_X \sim GP\left(\mu_X(.), \Sigma_X(.,.)\right) \\
&\mu_X|\Sigma_X \sim N\left(\mu_0; \frac{1}{\kappa_0}\Sigma_X\right) \\
&\Sigma_X \sim IW\left(\nu_0, \Lambda_0^{-1}\right) \\
&\sigma_\epsilon^2 \sim IG\left(\alpha_\epsilon, \beta_\epsilon\right)
\end{aligned} \tag{65}$$

dónde $IG(.,.)$ denota la distribución inversa Gamma con parámetro de forma α_ϵ, y β_ϵ denota un parámetro de escala, IW denota la distribución inversa Wishart, ν_0 denota un parámetro de forma, $\kappa_0 > 0$, y Λ_0^{-1} es un parámetro de escala. La distribución IW se define sobre una red de puntos finitos $t = \{t_1, t_2, \quad, t_p\}$, con p puntos, dónde:

$$\Sigma_X(t,t) \sim IW\left(\nu_0, \Lambda_0^{-1}(t,t)\right)$$

Para garantizar una estimación suavizada de la covarianza Yang et al., (2016), proponen:

$$\Lambda_0^{-1}(.,.) = \sigma_s^2 \Sigma(.,.)$$

dónde $\sigma_s^2 \sim IG\left(\alpha_s, \beta_s\right)$ es un parámetro de escalamiento positivo y $\Sigma(.,.)$ es un núcleo de covarianza que puede ser estacionario o no estacionario. La estructura de $\Lambda(.,.)$ puede ser paramétrica o no paramétrica. Una estructura de correlación estacionaria es la paramétrización Matérn, definida por:

$$\Lambda^{-1}(t_i, t_j) = \sigma_s^2 Matern_{cor}\left(|t_i - t_j|; \rho, \nu\right)$$

dónde:

$$Matern_{cor}\left(|t_i - t_j|; \rho, \nu\right) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\sqrt{2\nu}\frac{|t_i - t_j|}{\rho}\right)^\nu K_\nu\left(\sqrt{2\nu}\frac{|t_i - t_j|}{\rho}\right), \quad |t_i - t_j| \geq 0, \quad \rho > 0, \nu > 0$$

ρ es un parámetro de escala, ν es el orden, $K_\nu(.)$ es una de Bessel modificada del segundo tipo.
Por ejemplo para los tiempos $i = 5$, se tiene:

$$\begin{aligned}
&i = 1; \quad t_1 = \{t_{11}, t_{12}, \ldots, t_{1p_1}\} \\
&i = 2; \quad t_2 = \{t_{21}, t_{22}, \ldots, t_{2p_2}\} \\
&\vdots \\
&i = 5; \quad t_5 = \{t_{51}, t_{52}, \ldots, t_{5p_5}\}
\end{aligned}$$

Las medidas u observaciones obtenidas en esos puntos son:

$$\begin{aligned}
&Y_1(t_1) = \{Y_1(t_{11}), Y_1(t_{12}), \ldots, Y_1(t_{1p_1})\} \\
&Y_2(t_2) = \{Y_2(t_{21}), Y_2(t_{22}), \ldots, Y_2(t_{2p_2})\} \\
&\vdots \\
&Y_5(t_5) = \{Y_5(t_{51}), Y_5(t_{52}), \ldots, Y_5(t_{5p_5})\}
\end{aligned}$$

Los estados soluciones desconocidos que deben estimarse serían:

$$\begin{aligned}
&X_1(t_1) = \{X_1(t_{11}), X_1(t_{12}), \ldots, X_1(t_{1p_1})\} \\
&X_2(t_2) = \{X_2(t_{21}), X_2(t_{22}), \ldots, X_2(t_{2p_2})\} \\
&\vdots \\
&X_5(t_5) = \{X_5(t_{51}), X_5(t_{52}), \ldots, X_5(t_{5p_5})\}
\end{aligned}$$

y los errores que se comenten cuando se miden las observaciones, serían:

$$\begin{aligned}
&\epsilon_1(t_1) = \{\epsilon_1(t_{11}), \epsilon_1(t_{12}), \ldots, \epsilon_1(t_{1p_1})\} \\
&\epsilon_2(t_2) = \{\epsilon_2(t_{21}), \epsilon_2(t_{22}), \ldots, \epsilon_2(t_{2p_2})\} \\
&\vdots \\
&\epsilon_5(t_5) = \{\epsilon_5(t_{51}), \epsilon_5(t_{52}), \ldots, \epsilon_5(t_{5p_5})\}
\end{aligned}$$

Para aproximar $\{X_i(\tau)\}$ se utiliza un proceso prior GP basado en un sistema de funciones base tales como: polinomios cúbicos, B-splines, caos polinomial, fijando una red de puntos $\{\tau_1, \ldots, \tau_L\}$ y evaluando la densidad de los datos en esos puntos. Siguiendo la notación dada Yang et. al., (2017), sea:

$$B(.) = (b_1(.), b_2(.), \ldots, b_K(.))^T$$

denota K funciones base seleccionadas con coeficientes

$$\omega_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{iK})^T$$

entonces,

$$X_i(\tau) = \sum_{k=1}^{K} \omega_{ik} b_k(\tau) = B(\tau)\omega_i$$

Cuando $K = L$, se obtiene:

$$\omega_i = B^{-1}(\tau)X_i(\tau), \quad X_i(\tau) \sim GP\left(\mu_X(\tau), \Sigma_X(\tau, \tau)\right)$$

es una transformación lineal de los $X_i(\tau)$, dónde:

$$\mathbb{E}\left(\omega_i\right) = \mathbb{E}\left(B^{-1}(\tau)X_i(\tau)\right) = B^{-1}(\tau)\mathbb{E}\left(X_i(\tau)\right) = B^{-1}(\tau)\mu_X(\tau) = \mu_\omega(\tau)$$

y

$$\mathbb{C}ov\left(\omega_i\right) = \mathbb{C}ov\left(B^{-1}(\tau)X_i(\tau)\right) = B^{-1}(\tau)\mathbb{E}\left(X_i(\tau)\right)\left(B^{-1}(\tau)\right)^T = B^{-1}(\tau)\Sigma_X(\tau, \tau)\left(B^{-1}(\tau)\right)^T = \Sigma_\omega(\tau, \tau)$$

Entonces el modelo Bayesiano jerárquico en términos de las funciones bases se puede reescribe como:

$$\begin{aligned}
&\omega_i \sim GP\left(\mu_\omega(\tau), \Sigma_\omega(\tau, \tau)\right) \\
&\mu_\omega(\tau)|\Sigma_\omega(\tau, \tau) \sim N\left(B^{-1}(\tau)\mu_0(\tau), \frac{1}{\kappa_0}\Sigma_\omega(\tau, \tau)\right) \\
&\Sigma_\omega(\tau, \tau) \sim IW\left(\nu, \sigma_s^2 B^{-1}(\tau)\Sigma\left(\tau, \tau\right)\left(B^{-1}(\tau)\right)^T\right) \\
&\sigma_s^2 \sim IG\left(\alpha_s, \beta_s\right)
\end{aligned}$$

Para hacer inferencia Bayesiana se requiere calcular la distribución a posterior:

$$p\left(X, \mu_X, \Sigma_X, \sigma_\epsilon^2, \sigma_s^2|Y\right) \propto f_1\left(Y|X, \sigma_\epsilon^2\right) f_2\left(X|\mu_X, \Sigma_X\right) f_3\left(\mu_X|\Sigma_X\right) f_4\left(\Sigma_X\right) f_5\left(\sigma_\epsilon^2\right) f_6\left(\sigma_s^2\right)$$

dónde $f_1(.)$, $f_2(.)$, $f_3(.)$, $f_4(.)$, y $f_5(.)$ son funciones de densidad de probabilidad, y

$$Y = \{Y_1(t_i), \ldots, Y_n(t_n)\}, \quad X = \{X_1(t_i), \ldots, X_n(t_n)\}$$

En términos de las funciones bases se tiene la siguiente equivalencia:

$$p\left(\omega, \mu_\omega, \Sigma_\omega, \sigma_\epsilon^2, \sigma_s^2|Y\right) \propto g_1\left(Y|\omega, \sigma_\epsilon^2\right) g_2\left(\omega|\mu_\omega, \Sigma_\omega\right) g_3\left(\mu_\omega|\Sigma_\omega\right) g_4\left(\Sigma_\omega\right) f_5\left(\sigma_\epsilon^2\right) f_6\left(\sigma_s^2\right) \tag{66}$$

dónde $g_1(.)$, $g_2(.)$, y $g_3(.)$ son funciones de densidad de probabilidad, y

$$\omega = \{\omega_1, \ldots, \omega_n\}, \quad \mu_\omega(\tau) = B^{-1}(\tau)\mu_X(\tau), \quad \Sigma_\omega(\tau, \tau) = B^{-1}(\tau)\Sigma_X(\tau, \tau)\left(B^{-1}(\tau)\right)^T$$

Para aplicar un método Monte Carlo por cadenas de Markov, se requieren obtener las distribuciones marginales. La distribución marginal a posterior de ω_i es:

$$\begin{aligned}
p\left(\omega_i|Y_i(t_i), \mu_\omega, \Sigma_\omega\right) &\propto g_1\left(Y_i(t_i)|\omega, \sigma_\epsilon^2\right) g_2\left(\omega|\mu_\omega, \Sigma_\omega\right) \\
&= N\left(X_i(t_i), \sigma_\epsilon^2\mathbf{I}\right) \times N\left(\mu_\omega, \Sigma_\omega\right) \\
&\propto \frac{1}{|\sigma_\epsilon^2\mathbf{I}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left[Y_i(t_i) - B(t_i)\omega_i\right]^T\left(\sigma_\epsilon^2\mathbf{I}\right)^{-1}\left[Y_i(t_i) - B(t_i)\omega_i\right]\right\} \\
&\times \frac{1}{|\Sigma_\omega|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\omega_i - \mu_\omega\right)^T\left(\Sigma_\omega\right)^{-1}\left(\omega_i - \mu_\omega\right)\right\} \\
&\sim N\left(\mu_{\omega_i|Y_i(.)}, \Sigma_{\omega_i|Y_i(.)}\right)
\end{aligned} \tag{67}$$

dónde:

$$\mu_{\omega_i|Y_i(.)} = \Sigma_{\omega_i|Y_i(.)}\left(\frac{1}{\sigma_\epsilon^2}B(t_i)Y_i(t_i) + \mu_\omega\Sigma_\omega^{-1}\right)$$

y

$$\Sigma_{\omega_i|Y_i(.)} = \left(\frac{1}{\sigma_\epsilon^2} B(t_i) Y_i(t_i) + \Sigma_\omega^{-1}\right)^{-1}$$

La distribución a posterior para $\mu_\omega, \Sigma_\omega|\omega_1, \ldots, \omega_n$ es:

$$p\left(\mu_\omega, \Sigma_\omega|\omega_1, \ldots, \omega_n\right) \propto \prod_{i=1}^{n} g_2\left(\omega_i|\mu_\omega, \Sigma_\omega\right) g_3\left(\mu_\omega|\Sigma_\omega\right) g_4\left(\Sigma_\omega\right)$$

Pero se sabe que:

$$\mu_\omega(\tau)|\Sigma_\omega(\tau,\tau) \sim N\left(B^{-1}(\tau)\mu_0(\tau), \frac{1}{\kappa_0}\Sigma_\omega(\tau,\tau)\right)$$

y

$$\Sigma_\omega(\tau,\tau) \sim IW\left(\nu, \sigma_s^2 B^{-1}(\tau)\Sigma\left(\tau,\tau\right)\left(B^{-1}(\tau)\right)^T\right)$$

entonces se puede deducir que:

$$\begin{aligned} p\left(\mu_\omega|\omega_1, \ldots, \omega_n, \Sigma_\omega\right) &= \int p\left(\mu_\omega, \Sigma_\omega|\omega_1, \ldots, \omega_n\right) d\Sigma_\omega \\ &\sim N\left(\frac{\kappa_0}{n+\kappa_0}\left(\sum_{i=1}^{n}\omega_i + \frac{1}{\kappa_0}B^{-1}(\tau)\mu_0(\tau)\right), \frac{\kappa_0}{n+\kappa_0}\Sigma_\omega\right) \end{aligned} \tag{68}$$

La distribución a posterior marginal de $\Sigma_\omega|\omega_1, \ldots, \omega_n, \mu_\omega$ es:

$$p\left(\Sigma_\omega|\omega_1, \ldots, \omega_n, \mu_\omega\right) = \int p\left(\mu_\omega, \Sigma_\omega|\omega_1, \ldots, \omega_n\right) d\mu_\omega \sim IW\left(\nu_\omega, \Psi_\omega\right) \tag{69}$$

dónde:

$$\nu_\omega = n + 1 + \nu$$

y

$$\begin{aligned} \Psi_\omega = &\sum_{i=1}^{n}\left(\omega_i - \mu_\omega\right)\left(\omega_i - \mu_\omega\right)^T \\ &+ \frac{1}{\kappa_0}\left(\mu_\omega - B^{-1}(\tau)\mu_0(\tau)\right)\left(\mu_\omega - B^{-1}(\tau)\mu_0(\tau)\right)^T \\ &+ \sigma_s^2 B^{-1}(\tau)\Sigma(\tau,\tau)\left(B^{-1}(\tau)\right)^T \end{aligned}$$

Una vez conocida las marginales a posterior se implementa el algoritmo muestreador de Gibbs, para obtener los parámetros estimados una vez garantizada la convergencia a posterior.

1. Se inicializan los parámetros $\left(\mu_X(\tau,\tau), \Sigma_X(\tau,\tau), \sigma_\epsilon^2\right)$.
2. Simulamos de $p\left(\omega_i|Y_i(t_i), \mu_\omega, \Sigma_\omega\right)$, $p\left(\mu_\omega|\omega_1, \ldots, \omega_n, \Sigma_\omega\right)$, y $p\left(\Sigma_\omega|\omega_1, \ldots, \omega_n, \mu_\omega\right)$, usando las ecuaciones (67), (68) y (69), respectivamente.
3. Se aproxima: $p\left(X_i(t_i), \mu_X(t_i), \Sigma_X(t_i,t_i), \Sigma_X(\tau,t_i), \Sigma_X(t_i,\tau), \Sigma_X(\tau,\tau)|\omega_1, \ldots, \omega_n, \mu_\omega, \Sigma_\omega\right)$, por:

$$\begin{aligned} &X_i(t_i) = B(t_i)\omega_i, \\ &\mu_X(t_i) = B(t_i)\mu_\omega, \\ &\Sigma_X(t_i,t_i) = B(t_i)\Sigma_\omega\left(B(t_i)\right)^T, \\ &\left(\Sigma_X(\tau,t_i)\right)^T = \Sigma_X(t_i,\tau) = B(t_i)\Sigma_\omega\left(B(\tau)\right)^T \\ &\Sigma_X(\tau,\tau) = B(\tau)\Sigma_\omega\left(B(\tau)\right)^T \end{aligned}$$

4. La marginal,

$$p\left(\sigma_\epsilon^2|X,Y\right) = p\left(\sigma_\epsilon^2|Y_1(t_1), X_1(t_1), \ldots, Y_n(t_n), X_n(t_n)\right) \propto \prod_{i=1}^{n} f_1\left(Y_i(t_i)|X_i(t_i), \sigma_\epsilon^2\right) f_5\left(\sigma_\epsilon^2\right)$$

se obtienen simulado de:

$$\sigma_\epsilon^2|X,Y \sim IG\left(\alpha_\epsilon + \frac{1}{2}\sum_{i=1}^{n} p_i, \beta_\epsilon + \frac{1}{2}\sum_{i=1}^{n}\left(Y_i(t_i) - X_i(t_i)\right)^T\left(Y_i(t_i) - X_i(t_i)\right)\right)$$

5. La marginal:

$$p\left(\sigma_s^2|\Sigma_X(\tau,\tau)\right) \propto f\left(\Sigma_\omega|\sigma_s^2\right) f\left(\sigma_s^2\right)$$

se obtiene:

$$\sigma_s^2|\Sigma_X(\tau,\tau) \sim IG\left(\alpha_s + \frac{1}{2}(\nu+K-1)K, \beta_s + \frac{1}{2}trace\left(\Sigma(\tau,\tau)\Sigma_\omega^{-1}\right)\right)$$

2.5 Karhunen-Loève Expansion

Supóngase que $\{X(t), t \in T\}$ es un proceso estocástico que modela funciones de variables aleatorias, dónde t denota el índice de T que puede representar un dominio espacio o tiempo que a su vez puede ser infinito. Entonces se considera una versión finita del proceso aleatorio $X(t)$, para lo cual se requiere de una discretización del tiempo T en un conjunto finito de índices de puntos y se estudia el proceso en ese dominio; esto es,

$$\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}, \quad t_1, t_2, \ldots, t_n \in T\}$$

La expansión de Karhunen-Loève (KL), (Loève (1977)), es una técnica bastante conocida usada para reducir dimensión y representar procesos estocásticos como los tratados aquí. Sea $\mu_X(t)$ la media del proceso $X(t)$ y sea $k(t,s) = \mathbb{C}ov\left(X(t), X(s)\right)$ la función de covarianza. La expansión KL de $X(t)$ es:

$$X_t(\omega) = \mu_X(t) + \sum_{i=1}^{\infty} \sqrt{\lambda_i}\psi_i(t)X_i(\omega)$$

dónde: $\psi_i(\omega)$ son funciones propias ortogonales y λ_i son los correspondientes valores propios del problema:

$$\int_T k(t,s)\psi_i(s)ds = \lambda_i\psi_i(t), \quad t \in T, \tag{70}$$

dónde $\{\psi_n(t)\}$ son variables aleatorias mutuamente no correlacionadas que satisfacen:

$$\mathbb{E}\left(X_i\right) = 0, \quad \mathbb{E}\left(X_iX_j\right) = \delta_{ij}$$

y se define:

$$X_t(\omega) = \frac{1}{\sqrt{\lambda_i}} \int_T \left(X_t(\omega) - \mu_X(t)\psi_i(t)\right) dt, \quad \forall i$$

En problemas prácticos se utiliza una expansión de la serie finita:

$$X_t(\omega) = \mu_X(t) + \sum_{i=1}^{d} \sqrt{\lambda_i}\psi_i(t)X_i(\omega), \quad d \geq 1$$

Hiu 2010 señala que si $k(t,s) = \exp\left(-\frac{1}{a}|t-s|\right)$, dónde $a > 0$ es la longitud de la correlación y si $T = [-b, b]$, entonces el problema de valores propios definido en (70), puede ser resuelto analíticamente. Los autovalores son:

$$\lambda_i = \begin{cases} \frac{2a}{1+a^2w_i^2} & \text{if i}is\ even \\ \frac{2a}{1+a^2v_i^2} & \text{if i}is\ odd \end{cases} \tag{71}$$

y las correspondiente funciones propias son:

$$\psi_i(t) = \begin{cases} \frac{\sin(w_it)}{\sqrt{b-\frac{\sin(2w_ib)}{2w_i}}} & \text{if i}is\ even \\ \frac{\cos(v_it)}{\sqrt{b-\frac{\sin(2v_ib)}{2v_i}}} & \text{if i}is\ odd \end{cases} \tag{72}$$

dónde ω_i y v_i son las soluciones del sistema de ecuaciones:

$$\begin{cases} aw + \tan(wb) = 0 & \text{if i}is\ even \\ 1 - \tan(vb) = 0 & \text{if i}is\ odd \end{cases} \tag{73}$$

Una forma equivalente de escribir la expasión es la siguiente, si $X_t(\omega)$ se define sobre un espacio de Hilber separable $\mathbb{X}$, y $\{\lambda_j, \psi_j\}_{j\in\mathbb{N}}$ denota los pares de autovalores de $k(.,.)$, entonces se tiene:

$$X_t(\omega) = \mu_X(t) + \sum_{i=1}^{\infty} \sqrt{\lambda_i}\psi_i(t)\zeta_i(\omega) \quad , \quad \zeta_i(\omega) \sim N(0,1)$$

Esta es una representación de la variable aleatoria Gaussiana en términos una base que diagonaliza el operador de covarianza. El Teorema de Mercer supone que para un núcleo $k(t,s)$ continuo, semi-definido positivo sobre un espacio de medida $(\mathcal{X}, \mathcal{F}, \mu)$, existe un conjunto ortogonal $\{\psi_i\}_{i\in N}$ de $L^2(\mathcal{X}, \mathcal{F}, \mu)$, y una secuencia de valores no negativos $\{\lambda_i\}_{i\in N}$ tal que:

$$k(t,s) = \sum_{i=1}^{\infty} \lambda_i \psi_i(t)\psi_i(s)$$

y

$$\int_I k(t,s)\psi_i(s)ds = \lambda_i \psi_i(t) \tag{74}$$

la ecuación (74) es una ecuación integral de Fredholm de segunda clase. La solución analítica para está ecuación sólo existe para casos particulares, se puede calcular para algunos tipos de funciones de covarianza definidas en dominios rectangulares, entonces se necesita recurrir a métodos numéricos o métodos Monte Carlo para obtener una aproximación.
En la práctica, lo que se hace es truncar la expansión en d términos, y aun se retiene un cierto porcentaje de la varianza del proceso; esto es,

$$k(t,s) = \sum_{i=1}^{d} \lambda_i \psi_i(t)\psi_i(s)$$

entonces,

$$X_t(\omega) = \mu_X(t) + \sum_{i=1}^{d} \sqrt{\lambda_i}\psi_i(t)\zeta_i(\omega)$$

2.6 Polynomial Chaos Expansion (PCE) and Gaussian Processes (GP)

Considerase un sistema dinámico cuyo comportamiento está representado por un modelo matemático:

$$Y = f(X), \quad X \in \mathcal{D}_X \subset \mathbb{R}^M$$

dónde: $X = (X_1, \ldots, X_M)^T$ representa los parámetros de entrada del sistema, mientras que Y es la cantidad de interés. Una vez elegido el modelo funcional f, hay que cuantificar las fuentes de incertidumbre, en este caso se utiliza la información disponible base de datos (mediciones), juicio de expertos sobre el problema, esta información es procesada para construir el modelo probabilístico adecuado que ajusta los datos de entrada representados pos un vector aleatorio X modelado por una densidad de probabilidad $f_X(.)$. Cuando los datos se suponen independientes, la distribución conjunta se puede definir por un conjunto de distribuciones marginales, digamos:

$$f_X(x_1, \ldots, x_M) = \prod_{i=1}^{M} f_{X_i}(x_i) \tag{75}$$

En consecuencia, la cantidad de interés $Y = f(X)$ se convierte en una variable aleatoria, cuyas propiedades están implícitamente definidas por la propagación de la incertidumbre descritas por la distribución de probabilidad conjunta $f_X(.)$, a través del modelo matemático.
Supóngase que Y es un proceso estocástico de segundo orden que tiene una varianza finita, entonces Y pertenece a un espacio de Hilbert y se puede representar por:

$$f_{PCE}(X) = \sum_{i=0}^{\infty} \beta_i Z_i \tag{76}$$

dónde Y es una serie infinita, $\{Z_i\}_{i=0}^{\infty}$ es un conjunto numerable de variables aleatorias que forman una base del espacio de Hilbert, $\{\beta_i\}$ representan a los coeficientes de la serie. Los espacios de Hilbert garantizan la existencia de tales bases y su representación; sin embargo, hay muchas formas de representarlas. En el caso de las expansiones del caos polinomial, en el cual los términos de las bases $\{Z_i\}_{i=0}^{\infty}$ son polinomios ortonormales multivariados con vector de entrada X, i.e, $Z_i = \Psi_i(X)$, y la aproximación (76) se puede reescribir truncando la serie en M términos para propósitos prácticos cómo:

$$f_{PCE}(X) = \sum_{i=0}^{M} \beta_i \Psi_i(X) \tag{77}$$

Para la construcción de las bases, se supone que existe un vector aleatorio que tiene componentes independientes denotada por $\{X_i, \quad i = 1, \ldots, n\}$, para cada X_i y dos funciones cualesquiera:

$$\phi_1, \phi_2 : x \in \mathcal{D}_{x_i} \to \mathbb{R}$$

Ahora se define un producto interno funcional:

$$<\phi_1, \phi_2>_i = \int_{\mathcal{D}_{x_i}} \phi_1(x)\phi_2(x) f_{X_i}(x)dx = \mathbb{E}_{f_{X_i}(x)}\left(\phi_1(x)\phi_2(x)\right)$$

Dos funciones de este tipo se dicen ortogonales con respecto a la medida de probabilidad $\mathbb{P}(dx) = f_{X_i}(x)dx$, si:

$$\mathbb{E}\left(\phi_1(x)\phi_2(x)\right) = 0$$

Utilizando la notación previa y el álgebra clásica se puede construir una familia de polinomios ortogonales $\left\{\pi_k^{(i)}, \quad k \in \mathbb{N}\right\}$, que satisface:

$$<\pi_j^{(i)}, \pi_k^{(i)}> = \mathbb{E}\left\{\pi_j^{(i)}(x_i)\pi_k^{(i)}(x_i)\right\} = \int_{\mathcal{D}_{x_i}} \pi_j^{(i)}(x_i)\pi_k^{(i)}(x_i) f_{X_i}(x)dx = a_j^i \delta_{ik} \tag{78}$$

dónde el sub-índice k denota el grado del polinomio $\pi_k^{(i)}$, δ_{ik} denota la delta de Kronecker $\delta_{ik} = 1$, cuando $i = k$, y $\delta_{ik} = 0$ en otro caso, y

$$a_j^i = \|\pi_j^{(i)}\|_i^2 = <\pi_j^{(i)}, \pi_j^{(i)}>_i$$

La familia se puede obtener aplicando el procedimiento de ortogonalización de Gram-Schmidt a la forma canónica de monomios $\left\{1, x, x^2, \ldots\right\}$. Para las distribuciones estándar (Uniforme, Normal, Gamma, Beta), la familia asociada de polinomios ortogonales es bien conocida (Legendre, Hermite, Laguerre, Jacobi).
En general la familia obtenida no suele ser ortonormal. El procedimiento consiste en aplicar una normalización, para obtener la familia ortonormal; esto es:

$$\Psi_j^{(i)} = \frac{\pi_j^{(i)}}{\sqrt{a_j^i}}, \quad i = 1, \ldots, M; \quad j = 1, \ldots, \mathbb{N} \tag{79}$$

En el caso de polinomios multivariados, y se quiere construir una base como la definida en la ecuación (76), se construyen productos tensoriales de polinomios univariados ortonormales, definiendo tuplas o índices múltiples $\alpha \in \mathbb{N}^M$, que representan listas de enteros ordenados:

$$\alpha = (\alpha_1, \ldots, \alpha_M)^T, \quad \alpha_i \in \mathbb{N} \tag{80}$$

Se puede asociar un polinomio multivariado Ψ_α por cualquier índice múltiple α, mediante:

$$\Psi_\alpha(x) = \prod_{i=1}^{M} \Psi_{\alpha_i}^{(i)}(x_i) \tag{81}$$

dónde los polinomios univariados $\left\{\Psi_k^{(i)}, \quad k \in \mathbb{N}\right\}$ son definidos de acuerdo a la $i-$ésima distribución marginal definidas en las ecuaciones (78) y (79). Además, se tiene que los polinomios multivariados con vector de entrada X también son ortonormales, en consecuencia:

$$\mathbb{E}\left\{\Psi_\alpha(X)\Psi_\theta(X)\right\} = \int \Psi_\alpha(X)\Psi_\theta(X) f_X(x)dx = \delta_{\alpha\theta}, \quad \forall \alpha, \theta \in \mathbb{N}^M \tag{82}$$

dónde $\delta_{\alpha\theta}$ es la delta Kronecker. Así que ahora se puede probar que el conjunto de todos los polinomios multivariados con un vector aleatorio de entrada X forma una base en los espacios de Hilbert, en el cual $Y = f(X)$ se puede representar:

$$\hat{f}_{PCE}(X) = \sum_{\theta \in \mathbb{N}^M} \beta_\theta \Psi_\theta(X) \tag{83}$$

Por la propiedad ortogonal de la base del caos polinomial (82) se puede calcular cada coeficiente de expansión como sigue:

$$\mathbb{E}\left\{f(X)\Psi_\alpha(X)\right\} = \mathbb{E}\left\{\beta_\theta \sum_{\theta \in \mathbb{N}^M} \Psi_\alpha(X)\Psi_\theta(X)\right\} = \mathbb{E}\left\{\sum_{\theta \in \mathbb{N}^M} \beta_\theta \delta_{\alpha\theta}\right\} = \beta_\alpha, \quad \alpha = \theta$$

o equivalentemente:

$$\beta_l = \mathbb{E}\left\{f(X)\Psi_l(X)\right\} = \int f(X)\Psi_l(X)f_X(x)dx \approx \sum_{i=1}^{N} \xi_i f_{X_i}(x_i)\Psi_l(X_i), \quad l = 0, \ldots, M$$

dónde la segunda ecuación se obtiene por técnicas de integración numérica, tales como la regla de la cuadratura Gaussiana, $\{X_i, i = 1, \ldots, N\}$ y $\{\xi_i, i = 1, \ldots, N\}$ representan los nodos y pesos, respectivamente; o equivalentemente usando un algoritmo tipo Monte Carlo, como por ejemplo de muestreo de importancia:

$$\begin{aligned}
\beta_l &= \mathbb{E}\left\{f(X)\Psi_l(X)\right\} \\
&= \int f(X)\Psi_l(X)f_X(x)dx \\
&\approx \frac{1}{n}\sum_{i=1}^{N} \Psi_l(x_i)\frac{f_{X_i}(x_i)}{q_{X_i}(x_i)} \\
&= \frac{1}{n}\sum_{i=1}^{N} \xi_i \Psi_l(x_i), \quad \xi_i = \frac{f_{X_i}(x_i)}{q_{X_i}(x_i)}, \quad l = 0, \ldots, M
\end{aligned}$$

dónde $q_X(x)$ se conoce como una función de importancia, que por lo general es fácil de muestrear y tiene un soporte que contiene a $f_X(x)$.
Dado que la serie dada en (83) es infinita, para hacerla útil se debe truncar a M términos, se tiene:

$$\hat{f}_{PCE}(X) = \sum_{i=0}^{M}\left\{\sum_{i=1}^{N} \xi_i f(X_i)\Psi_i(X_i)\right\}\Psi_i(x) = \sum_{i=0}^{M}\beta_i\Psi_i(x)$$

En la práctica los datos experimentales se obtiene con errores; es decir,

$$Y_i = \hat{f}_{PCE}(X_i) + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right), \quad i = 1, \ldots, N$$

entonces $\hat{f}_{PCE}(X_i)$ todavía permanece desconocida.
Ahora se puede conectar la técnica de la expansión del caos polinomial con los procesos Gaussianos, como sigue: para una data de entrada $X = \{X_i, i = 1, \ldots, N\}$ y una respuesta de salida $Y = \{Y_i, i = 1, \ldots, N\}$, dónde:

$$Y = f(X) + \epsilon, \quad \epsilon \sim N\left(0, \sigma_\epsilon^2\mathbf{I}\right), \quad f(X) \sim GP\left(m_{prior}(X), k_{prior}(X, X)\right)$$

Usando el Teorema de Bayes, se puede predecir un dato nuevo con su respectivo error (varianza) utilizando la distribución a posterior:

$$f(x)|Y, X, x, \theta \sim N\left(m_{post}(x), k_{post}(x, x)\right)$$

dónde:

$$m_{post}(x) = \mathbb{E}\left(f(x)|Y, X, x, \theta\right) = K_x^T\left(K + \sigma_\epsilon^2\mathbf{I}\right)^{-1}Y \tag{84}$$

y

$$k_{post}(x, x) = K_{xx} - K_x^T\left(K + \sigma_\epsilon^2\mathbf{I}\right)^{-1}K_x$$

dónde:

$$K = k(X, X) \in \mathbb{R}^{N\times N}, \quad K_{ij} = k\left(X_i, X_j\right), \quad K_x = k(X, x) \in \mathbb{R}^{N\times 1}, \quad K_{xx} = k(x, x) \in \mathbb{R}$$

Se puede observar que la ecuación dada en (84) se puede representar como una combinación de N núcleos de funciones:

$$m_{post}(x) = \hat{f}_{GP}(x) = \sum_{i=1}^{N}\beta_i k\left(X_i, x\right), \quad \beta = \left(K + \sigma_\epsilon^2\mathbf{I}\right)^{-1}Y$$

Los PCE y GP pueden estudiarse bajo la misma estructura y combinarse para mejorar la estimación, su vinculo es a través de la reproducción de núcleos de funciones de covarianza en los espacios de Hilbert.
En la literatura reciente se han desarrollado otros métodos de aproximación, utilizando la estructura de modelos de rango reducido combinado con los procesos Gaussianos, ver Svensson et. al. (2016), y Solin and Särkkä (2020) entre otros. La idea consiste en definir la función de covarianza isotrópica $k(x, x') = k(\|\mathbf{r}\|)$,

dónde $\mathbf{r} = \mathbf{x} - \mathbf{x}'$, y $\|.\|$ denota la norma Euclidiana. Cuando $k(\mathbf{r})$ es continua y definida positiva, se define como sigue:

$$k\left(x, x'\right) = \frac{1}{(2\pi)^d} \int \exp\left(i\omega^T \mathbf{r}\right) \mu\left(d\omega\right) \tag{85}$$

dónde μ es una medida positiva. Cuando la medida $\mu(d\omega) = S(\omega)d\omega$ tiene una densidad, esta se llama la densidad espectral $S(\omega)$ y está asociada a la función de covarianza $k(\mathbf{r})$. La definición permite establecer una dualidad de Fourier entre la covarianza y la densidad espectral, que se conoce como el teorema de Wiener–Khintchin, resultando las siguientes identidades:

$$k\left(\mathbf{r}\right) = \frac{1}{(2\pi)^d} \int S(\omega) \exp\left(i\omega^T r\right) d\omega, \quad \mathbf{r} = \mathbf{x} - \mathbf{x}' \tag{86}$$

y

$$S\left(\omega\right) = \int k\left(\mathbf{r}\right) \exp\left(-i\omega^T r\right) d\omega \tag{87}$$

$S\left(\omega\right)$ representa la transformada de Fourier, y $k\left(\mathbf{r}\right)$ es la transformada inversa de Fourier.
Solin y Särkkä (2020), definen un operador de covarianza como un pseudo operador diferencial, asociando a cada función de covarianza $k\left(x, x'\right)$ un operador $\mathcal{K}$, cómo sigue:

$$\mathcal{K}\phi = \int k\left(., x'\right) \phi\left(x'\right) dx' \tag{88}$$

Esta definición permite aproximar el operador covarianza en forma similar a los métodos empleados para aproximar operadores diferenciales y pseudo diferenciales de las ecuaciones diferenciales parciales en espacios de Hilbert.
Por otra parte, en los casos cuando la covarianza es isotrópica el operador es invariante bajo traslación, entonces la representación de Fourier es una función de transferencia con densidad espectral $S(\omega)$ que sigue una distribución Gaussiana, con operador de covarianza asociado:

$$S(\omega) \triangleq S\left(\|\omega\|\right) \tag{89}$$

Ahora se tiene la siguiente expansión polinómica:

$$S\left(\|\omega\|\right) = a_0 + a_1\|\omega\|^2 + a_2\left(\|\omega\|^2\right)^2 + a_3\left(\|\omega\|^2\right)^3 + a_4\left(\|\omega\|^2\right)^4 + \ldots \tag{90}$$

Dado que la función de transferencia corresponde al operador de Laplace $\nabla^2 = -\|\omega\|^2$, y que para una función f regular se satisface:

$$\mathcal{F}[\nabla^2 f](\omega) = -\|\omega\|^2 \mathcal{F}[f](\omega) \tag{91}$$

dónde $\mathcal{F}[.]$ denota la transformada de Fourier de su argumento. Tomando la inversa de la transformada de Fourier se obtiene el operador de la covarianza $\mathcal{K}$:

$$\mathcal{K} = a_0 + a_1\left(-\nabla^2\right) + a_2\left(-\nabla^2\right)^2 + a_3\left(-\nabla^2\right)^3 + a_4\left(\| - \nabla^2\right)^4 + \ldots \tag{92}$$

Luego se define el operador pseudo-diferencial como una serie de operadores diferenciales y se aproxima el operador Laplaciano con un método de Hilbert. Para llevar acabo la aproximación del operador covarianza en el espacio de Hilbert, se considera una función de covarianza $k\left(x, x'\right)$ y el producto interno:

$$\langle f(x), g(x)\rangle = \int_\Omega f(x)g(x)\omega(x)dx \tag{93}$$

dónde $\Omega \subset \mathbb{R}^d$ es un conjunto compacto. El producto interior induce a un espacio de Hilbert de funciones aleatorias. Si se fija una base $\{\phi_j(x)\}$, y un proceso Gaussiano $f(x) = (f(x_1), \ldots, f(x_n))^T$ puede ser expandido en términos de una serie infinita:

$$f(x) = \sum_{j=1}^{\infty} \zeta_j \phi_j(x) \tag{94}$$

donde: ζ_j son independientes y siguen una distribución Gaussiana multivariada, los $\phi_j(x)$ se obtienen a partir de las funciones propias de $k\left(x, x'\right)$ respecto al producto interno definido en la ecuación (93), entonces la serie definida en (94) se transforma en la serie de Karhunen-Loéve.

Para llevar acabo la aproximación del Laplaciano en un espacio de Hilbert, se considera el problema de autovalores para los operadores Laplaciano:

$$-\nabla^2\phi_j(x) = \lambda_j\phi_j(x), \quad x \in \Omega$$
$$\phi_j(x) = 0, \quad x \in \partial\Omega$$

Dado que la versión negativa del operador Laplaciano es definido positivo, el conjunto de funciones propias $\phi_j(.)$ son ortonormales con respecto a la definición dada en (93); es decir:

$$\int_\Omega \phi_i(x)\phi_j(x)dx = \delta_{ij}$$

dónde: $\delta_{ij} = 1$ si $i = j$, $\delta_{ij} = 0$ en otro caso, y todos los auto valores λ_j son números reales positivos. Entonces, el operador Laplaciano negativo se puede expresar en términos de un valor esperado:

$$-\nabla^2 f(x) = \int l\left(x, x'\right) f\left(x'\right) dx' \tag{95}$$

dónde $l\left(x, x'\right)$ es un núcleo de la forma:

$$l\left(x, x'\right) = \sum_{j=1}^{\infty} \lambda_j \phi\left(x\right) \phi\left(x'\right) \tag{96}$$

El desarrollo en potencia del operador definido en (96):

$$l^n\left(x, x'\right) = \sum_{j=1}^{\infty} \lambda_j^n \phi\left(x\right) \phi\left(x'\right), \quad n = 1, 2, \ldots, \tag{97}$$

dónde el operador negativo Laplaciano sería:

$$\left(-\nabla^2\right)^n f(x) = \int l^n\left(x, x'\right) f\left(x'\right) dx' \tag{98}$$

Luego desarrollado en ambos lados la ecuación (98), se tiene:

$$\begin{aligned}\mathcal{K}f(x) &= \left[a_0 + a_1\left(-\nabla^2\right) + a_2\left(-\nabla^2\right)^2 + a_3\left(-\nabla^2\right)^3 + \ldots\right] f(x) \\ &= \int \left[a_0 + a_1 l^1\left(x, x'\right) + a_2 l^2\left(x, x'\right) + a_3 l^3\left(x, x'\right) + \ldots\right] f\left(x'\right) dx'\end{aligned} \tag{99}$$

Ahora haciendo una comparación de (99) con la ecuación definida en (88), y utilizando la formula dada en (97), se puede concluir que la función de covarianza se puede aproximar por:

$$\begin{aligned}k\left(x, x'\right) &\approx a_0 + a_1 l^1\left(x, x'\right) + a_2 l^2\left(x, x'\right) + a_3 l^3\left(x, x'\right) + \ldots \\ &= \sum_{j=1}^{\infty} \left[a_0 + a_1\lambda_j^1 + a_2\lambda_j^2 + a_3\lambda_j^3 + \ldots\right] \phi_j\left(x\right) \phi_j\left(x'\right)\end{aligned} \tag{100}$$

Evaluando la serie de la densidad espectral en $\|\omega\|^2 = \lambda_j$ se obtiene:

$$S\left(\sqrt{\lambda_j}\right) = a_0 + a_1\lambda_j^1 + a_2\lambda_j^2 + a_3\lambda_j^3 + \ldots \tag{101}$$

Sustituyendo (101) dentro de (100) se obtiene la aproximación:

$$k\left(x, x'\right) \approx \sum_{j=1}^{\infty} S\left(\sqrt{\lambda_j}\right) \phi_j\left(x\right) \phi_j\left(x'\right) \tag{102}$$

dónde $S(.)$ es la densidad espectral de la función de covarianza, λ_j es el $j-$ésimo valor propio, y $\phi_j(.)$ es la función del valor propio del operador de Laplace.
Ahora se puede utilizar está metodología en combinación con los GP para hacer predicciones. Recordar que los GP pueden formularse como una predicción de un salida desconocida $f\left(X_{new}\right)$ asociada a una entrada de un dato nuevo conocido $X_{new} \in \mathbb{R}^n$, dado un conjunto de datos experimentales de entrenamiento:

$$\mathbb{D} = \left\{\left(X_i, Y_i\right), \quad i = 1, \ldots, n\right\}$$

La función $f(X_{new})$ evaluada en el nuevo vector de puntos X_{new}, $f\left(X_{new}\right) = \left(f\left(x_{new}^{(1)}\right), \ldots, f\left(x_{new}^{(n)}\right)\right)^T$, se considera como la realización de un GP; esto es, $f(x_{new}) \sim GP\left(0, k\left(x, x'\right)\right)$; además, las observaciones se consideran que son medidas con errores:

$$Y_i = f(X_i) + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma_\epsilon^2\right)$$

Entonces Rasmussen and Williams (2006) demostraron que, la distribución a posterior del dato nuevo es:

$$f\left(X_{new}\right)|\mathbb{D} \sim N\left(\mathbb{E}\left(f(X_{new})\right), \mathbb{V}ar\left(f(X_{new})\right)\right)$$

dónde:

$$\mathbb{E}\left(f(X_{new})\right) = k_{new}^T \left(K + \sigma_\epsilon^2 \mathbf{I}\right)^{-1} Y \tag{103}$$

y

$$\mathbb{V}ar\left(f(X_{new})\right) = k\left(x_{new}, x_{new}\right) - k_{new}^T \left(K + \sigma_\epsilon^2 \mathbf{I}\right)^{-1} k_{new} \tag{104}$$

$K_{ij} = k\left(x_i, x_j\right)$, k_{new} es un vector de dimensión n con la $j-$ésima entrada de $k\left(x_{new}, x_j\right)$, $Y = \left(y_1, \ldots, y_n\right)^T$ es un vector de n observaciones, y $X = \left(x_1, \ldots, x_n\right)^T$.
Para evitar el cálculo de matrices inversas de las ecuaciones (103) y (104), se usa la aproximación obtenida en (102), realizando una proyección del GP sobre un conjunto truncado de m funciones base tal que:

$$f(x) = \sum_{j=1}^{m} \zeta_j \phi_j(x), \quad \zeta_j \sim N\left(0, S\left(\sqrt{\lambda_j}\right)\right) \tag{105}$$

entonces se puede formar una descomposición aproximada de la matriz de covarianza:

$$K \approx \Phi \Lambda \Phi \tag{106}$$

dónde Λ es una matriz diagonal de los valores propios aproximados de tal manera que:

$$\Lambda_{jj} = S\left(\sqrt{\lambda_j}\right) \quad , \quad j = 1, \ldots, m \tag{107}$$

Usando las ecuaciones de inversión de una matriz dados en (103) y (104), se obtiene:

$$\mathbb{E}\left(f(x_{new})\right) \approx \phi_{new} \left(\Phi^T \Phi + \sigma_\epsilon^2 \Lambda^{-1}\right)^{-1} \Phi^T y \tag{108}$$

y

$$\mathbb{V}ar\left(f(x_{new})\right) \approx \sigma_\epsilon^2 \phi_{new}^T \left(\Phi^T \phi + \sigma_\epsilon^2 \Lambda^{-1}\right)^{-1} \phi_{new} \tag{109}$$

dónde: $\phi_{new}(X_{new}) = \left(\phi_1(x_{new}^{(1)}), \ldots, \phi_m(x_{new}^{(n)})\right)^T$ es un vector de dimensión m con la $j-$ésima entrada dada por $\phi_j\left(x_{new}\right)$, ver Solin and Särkka (2020).

2.7 Regresión functional no parámetrica

Una tercera alternativa que se propone para estimar los parámetros desconocido de la ODE es usar un modelo de regresión no paramétrico basado en un índice funcional. Este índice mapea la integral ponderada del predictor funcional y el coeficiente en una respuesta uni-dimensional usando una función no lineal f:

$$y_i = f\left(\int_I x_i(t)\beta(t)dt\right) + \epsilon_i \tag{110}$$

dónde f es una función no lineal que debe estimarse en conjunto con $\beta(t)$. El predictor funcional se proyecta en un sólo índice $\int_I x_i(t)\beta(t)dt$. Un enfoque natural no paramétrico es modelar f con un GP; donde ahora mapeamos el producto interno de los predictores funcionales y la función coeficiente a la respuesta a través de un GP, ver (Müller and Stadtmüller, (2005)), Chen et al., (2011), Choi et al., (2011) and Gramacy and Lian, (2010), entre otros).
Para realizar la inferencia, se procede como sigue: se expresan en términos de $X(t)$ y $\beta(t)$ en una expansión de la base:

$$x(t) = \sum_{i=1}^{N_x} \alpha_i \phi_i(t), \quad and \quad \beta(t) = \sum_{i=1}^{N_x} \beta_i \phi_i(t) \tag{111}$$

Para una base ortonormal $\{\phi_i(t)\}_{i=1}^{N_x}$, en un dominio I, dónde α_i y β_i son los coeficientes base del predictor y el parámetro. Así que:

$$\int_I x(t)\beta(t)dt = \sum_{i=1}^{N_x} \alpha_i\beta_i = \alpha^T\beta \tag{112}$$

dónde $\alpha = (\alpha_1, \ldots, \alpha_{N_x})^T$, y $\beta = (\beta_1, \ldots, \beta_{N_x})^T$.
Ahora el modelo se puede reescribir en forma compacta:

$$\begin{aligned} &y = f\left(\alpha^T\beta\right) + \epsilon \quad , \quad \epsilon \sim N\left(0, \sigma_y^2\right) \\ &f(.) \sim GP\left(\mu(.), \Sigma(.)\right) \\ &x(t) = \phi(t)\alpha + \varepsilon \quad , \quad \alpha \sim N\left(\mu_\alpha, \Sigma_\alpha\right) \quad , \quad \varepsilon \sim N\left(0, \sigma_x^2 I\right) \end{aligned} \tag{113}$$

dónde $\phi(t)$ es la matriz base con cada entrada dada por la evaluación de la función base en el punto t_j. $\phi\left(t\right)_{jk} = \phi_k(t_j)$, con $\phi_k(.)$ la $k-$ésima función base, $\alpha \in \mathbb{R}^{N_x}$ son los coeficientes de la base, ε es un ruido Gaussiano de las entradas, y ϵ ruido Gaussiano de las salidas. La media y la covarianza del proceso GP son determinados por la parametrización $\mu\left(\alpha^T\beta\right)$, y $\Sigma\left(\alpha^T\beta, \omega^T\beta\right)$. El núcleo del indíce funcional se puede interpretar como sigue:

$$\begin{aligned} |\alpha^T\beta - \omega^T\beta|^2 &= \left(\beta^T\alpha - \beta^T\omega\right)^T\left(\beta^T\alpha - \beta^T\omega\right) \\ &= \left(\alpha^T\beta - \omega^T\beta\right)\left(\beta^T\alpha - \beta^T\omega\right) \\ &= \left(\alpha^T - \omega^T\right)\beta\beta^T\left(\alpha - \omega\right) \\ &= \left(\alpha - \omega\right)^T A\left(\alpha - \omega\right) \quad , \quad A = \beta\beta^T \end{aligned} \tag{114}$$

Entonces,

$$\Sigma\left(\alpha^T\beta, \omega^T\beta\right) = \Sigma\left(\alpha, \omega\right) = \lambda^2 \exp\left(-\frac{\left(\alpha - \omega\right)^T A\left(\alpha - \omega\right)}{2\theta^2}\right) \tag{115}$$

qué representa un núcleo exponencial al cuadrado.
Para hacer inferencia se considera una matriz de funciones observadas $x = (x_1, \ldots, x_n)^T$, su correspondiente expansión base $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, $\beta = (\beta_1, \ldots, \beta_{N_x})^T$, y un conjunto de respuestas $y = (y_1, \ldots, y_n)^T$, entonces la verosimilitud:

$$p\left(x|\Theta\right) \propto \left(p\left(y|\alpha\beta\right)p\left(x|\alpha\right)p(\alpha)\right)$$

En forma equivalentemente expresada en términos de la log-verosimilitud:

$$\begin{aligned} \ln\left(p\left(x|\Theta\right)\right) &\propto \ln\left(p\left(y|\alpha\beta\right)p\left(x|\alpha\right)p(\alpha)\right) \\ &= -\frac{1}{2}y^T\left(\Sigma\left(\alpha\beta, \alpha\beta\right) + \sigma_y^2\right)^{-1}y - \frac{1}{2}\ln\left|\Sigma\left(\alpha\beta, \alpha\beta\right) + \sigma_y^2\right| \\ &- \frac{1}{2}\ln 2\pi - \sum_{i=1}^{N}\left(\frac{n_i}{2}\log\sigma_x^2 + \frac{1}{2\sigma_x^2}\left(x_i - \phi(t)\alpha_i\right)^T\left(x_i - \phi(t)\alpha_i\right)\right) \\ &- \sum_{i=1}^{N}\left(\frac{1}{2}\ln\left|\Sigma_\alpha\right| + \frac{1}{2}\left(\alpha_i - \mu_\alpha\right)^{-1}\Sigma_\alpha^{-1}\left(\alpha_i - \mu_\alpha\right)\right) \end{aligned}$$

Para considerar un método Bayesiano se require obtener las distribuciones marginales:

$$p\left(y|x\right) = \int_\Omega\int_\Omega\int_\Omega p\left(y|f\right)p\left(f|\alpha\beta\right)p\left(x|\alpha\right)p\left(\alpha\right)dfd\Theta d\alpha$$

Está integral es analiticamente intratable por lo que se requieren utilizar métodos de aproximación tipo Laplace o métodos Monte Carlo.
En los casos cuando se puede representar las entradas funcionales en términos de los coeficientes de la base, no se necesita considerar las distribuciones sobre los parámetros, ver (Chen et al., (2011), McLean et al., (2012), Morris, (2015)), y la log-verosimilitud se simplifica:

$$\ln\left(y|\Theta\right) = -\frac{1}{2}y^T\left(\Sigma\left(\alpha\beta, \alpha\beta\right) + \sigma_y^2\right)^{-1}y - \frac{1}{2}\ln\left|\Sigma\left(\alpha\beta, \alpha\beta\right) + \sigma_y^2\right| - \frac{N}{2}\ln 2\pi$$

Una estimación máxima a posteriori de los parámetros de este modelo $\Theta = (\lambda, \{\beta_i\}_{i=1}^p, \theta)$, dónde θ es un vector de hiperparámetros de $\Sigma(.)$, requiere de las derivadas de la log-verosimilitud:

$$\frac{\partial}{\Theta_i}\{\ln p\left(y|\Theta\right) p\left(\beta\right)\} = -\frac{1}{2} Tr\left(\Sigma^{-1}(.)\frac{\partial \Sigma}{\partial \Theta_i}\right) + \frac{1}{2} y^T \Sigma^{-1}(.)\frac{\partial \Sigma}{\partial \Theta_i}\Sigma^{-1}(.)\Sigma^{-1}(.)y + \frac{\partial}{\Theta_i} p\left(\beta\right)$$

Igualando a cero la derivadas con respecto a cada parámetro, y luego implementando el algoritmo del gradiente descendente se optimizan los valores de los parámetros.
También se puede predecir un dato nuevo usando la media y varianza a posteriori predictiva, dado los coeficientes entrenados $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ y su correspondiente respuestas $y = (y_1, \ldots, y_n)^T$. El predictor nuevo $x_{new}(t)$, con su coeficiente base α_{new} se obtiene como sigue:

$$\mathbb{E}\left(y_{new}\right) = \Sigma\left(\alpha_{new}^T \beta, \alpha\beta\right)\left(\Sigma\left(\alpha\beta, \alpha\beta\right) + \sigma_y^2 I\right)^{-1} y$$

y

$$\mathbb{V}ar\left(y_{new}\right) = \Sigma\left(\alpha_{new}^T \beta, \alpha_{new}^T \beta\right) - \Sigma\left(\alpha_{new}^T \beta, \alpha\beta\right)\left(\Sigma\left(\alpha\beta, \alpha\beta\right) + \sigma_y^2 I\right)^{-1} \Sigma\left(\alpha\beta, \alpha_{new}^T \beta\right)$$

2.8 Medidas para evaluar la performance de los métodos

Para evaluar la performance de los diferentes métodos podemos usar algunas medidas de bondad de ajuste que nos permiten hacer comparaciones: se pueden separa los datos en un conjunto de entrenamiento y^t y datos de evaluación y^e, entoces se puede evaluar la performance de los métodos cuantitativamente, a través de algunas medidas tales como:

1. La ráiz del error cuadrático medio (RMSE):

$$RMSE = \sqrt{\frac{1}{T_e}\sum_{t=1}^{T_e}\left(\hat{y}_t - y_t^e\right)^2}$$

   dónde $\hat{y}_t$ es el dato estimado y y_t^e es el dato verdadero.

2. The mean log likelihood (LL):

$$LL = \frac{1}{T_e}\sum_{t=1}^{T_e} \log N\left(\mathbb{E}\left(\hat{y}_t\right), \mathbb{V}ar\left(\hat{y}_t\right)\right)$$

3. Dado un modelo entrenado y un conjunto de estados latentes de prueba $x_* = (x_{*1}, \ldots, x_{*t})$, podemos construir características estadística de la distribución predicitva $f(x_{*t})$ usando muestras a posteriori de los parametros. Se puede cálcular the standardized mean squared error (SMSE), que se define por:

$$SMSE = \sum_{t=1}^{n_*} \frac{\left(y_{*t} - \mu_{*t}\right)^2}{\mathbb{V}ar\left(y\right)}$$

4. The mean standardized log loss (MSLL), determinada por:

$$MSLL = \frac{1}{n_*}\sum_{t=1}^{n_*} \frac{\left(y_{*t} - \mu_{*t}\right)^2}{\sigma_{*t}^2} + \ln 2\pi\sigma_{*t}^2$$

   dónde

$$\mu_{*t} = \mathbb{E}\left(f(x_{*t})\right) \quad , \quad \sigma_{*t}^2 = \mathbb{V}ar\left(f(x_{*t})\right) + \sigma_\epsilon^2$$

   are the predictive mean and variance for test sample $t = 1, 2, \ldots, n_*$, and y_{*t} is the actual test value. The training data variance is denoted by $\mathbb{V}ar(y)$.

3 Applications

3.1 Kermack-McKendrick SIR model:

The SIR model is deterministic and specified through ordinary differential equations (ODE):

$$\begin{aligned} \dot{S} &= \frac{dS}{dt} = -\beta SI \\ \dot{I} &= \frac{dI}{dt} = \beta SI - \gamma I \\ \dot{R} &= \frac{dR}{dt} = \gamma I \\ S(t) + I(t) + R(t) &= N \end{aligned} \tag{116}$$

The model is categorized as a compartmental model where N host individuals can be in either of the Susceptible (S), Infected/Infectious (I) or Removed/Recovered (R) compartments at any point in time.

3.2 Lorenz Model (Lorenz (1963)):

Is a coupled system of nonlinear differential equations describing fluid dynamics:

$$\begin{aligned} \dot{x} &= \frac{dx}{dt} = s(y - x) \\ \dot{y} &= \frac{dy}{dt} = rx - y - xz \\ \dot{z} &= \frac{dz}{dt} = xy - bz \end{aligned} \tag{117}$$

where: s, r, b are parameters, The standard choice of $(s, r, b) = (10, \frac{8}{3}, 28)$. The state vector $\mathbf{x} = (x, y, z)^T$ represents a position of the particles in phase space. The classical Lorenz attractor is a simple model of convective fluid motion induced by a temperature difference between the upper and lower surfaces. States x, y, and z are proportional to the intensity of the fluid's convective motion, the temperature difference between (hot) rising and (cold) descending currents, and the deviation of the vertical temperature profile from linearity respectively.

3.3 Fitzhugh–Nagumo model:

The Fitzhugh–Nagumo equations describe the behavior of spike potentials in the giant axon of squid neurons (Ramsay et al. (2007)):

$$\begin{aligned} \dot{V}(t) &= \frac{dV(t)}{dt} = c\left(V(t) - \frac{V^3(t)}{3}\right) \\ \dot{R}(t) &= \frac{dR(t)}{dt} = -\left(V(t) - a + bR(t)\right) \end{aligned} \tag{118}$$

where V describes the voltage across an axon membrane and R is a recovery variable summarizing outward currents. The parameter values used to generate the data are $a = 0.2$, $b = 0.1$, $c = 3$ and $V(0) = 1$, $R(0) = 3$.

4 Discusión y Conclusiones

BLABLa...BLA

Bibliography

1. M. Loève. (1977). Probability Theory, 4th ed. Springer-Verlag, New York.
2. Ramsay, J. O; Hooker, G; Campbell, D; and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. J. R. Statist. Soc. B, 69, Part 5, pp. 741–796.
3. Overstall, A; Woods, D; Parker, B. (2020) Bayesian Optimal Design for Ordinary Differential Equation Models With Application in Biological Science, Journal of the American Statistical Association, 115:530, 583-598, DOI: 10.1080/01621459.2019.1617154.

4. Solin, A. (2016). Stochastic differential equation methods for Spatio-Temporal Gaussian process Regression. Doctoral Thesis.
5. Solin A; Särkkä, S. (2020). Hilbert space methods for reduced-rank Gaussian process regression. Statistics and Computing (2020) 30:419–446. https://doi.org/10.1007/s11222-019-09886-w.
6. Rasmussen and C. K. Williams. Gaussian Processes for Machine Learning. the MIT Press, apr 2006. ISBN ISBN-10 0-262-18253-X. URL http://www. gaussianprocess.org/gpml/chapters/.
7. Schober, M., Särkkä, S., Hennig, P. (2019). A probabilistic model for the numerical solution of initial value problems. Stat. Comput. 29(1), 99–122.
8. Kersting, H., Hennig, P. (2016). Active uncertainty calibration in Bayesian ODE solvers. In: 32nd Conference on Uncertainty in Artificial Intelligence, pp. 309–318. Curran Associates, Inc.
9. Oksana Chkrebtii, David A. Campbell, Mark A. Girolami, Ben Calderhead. (2013). Bayesian Uncertainty Quantification for Differential Equations. arXiv:1306.2365v1 [stat.ME] 10.
10. Oksana A. Chkrebtii, David A. Campbell, Ben Calderhead, and Mark A. Girolam. (2016). Bayesian Analysis. TBA, Number TBA, pp. 1–29.
11. Butcher, J. (2008). Numerical Methods for Ordinary Differential Equations. John Wiley and Sons Ltd.
12. Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. Acta Numerica, 19: 451–559. MR2652785. doi: http://dx.doi.org/10.1017/S0962492910000061. 2,4, 6.
13. Skilling, J. (1991). Bayesian Solution of Ordinary Differential Equations, 23–37. Seattle: Kluwer Academic Publishers. 3, 4, 23, 24.
14. Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., and Rasmussen, C. E. (2003. Derivative observations in Gaussian process models of dynamic systems. In Becker, S., Thrun, S., and Obermayer, K. (eds.), Advances in Neural Information Processing Systems 15, 1057–1064. MIT Press. http://papers.nips.cc/paper/2287-derivative-observations-in-gaussianprocess- models-of-dynamic-systems.pdf 6.
15. Riihimäki, J and Vehtari, A. (2010). Gaussian processes with monotonicity information.Journal of Machine Learning Research, 9:645-652. ISSN 15324435.
16. Conrad, P; Girolami, M; Säkkä, S; Stuart, A; and Zygalakis, K. (2015). Probability Measures for Numerical Solutions of Differential Equations arXiv:1506.04592v1 [stat.ME].
17. Svensson, A; Solin, A; Särkkä, S; Schän, T. (2016). Computationally Efficient Bayesian Learning of Gaussian Process State Space Models. Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), 2016, Cadiz, Spain. JMLR: W&CP volume 51.