Stochastic dynamic models
Interpolation and integration

thierry.chonavel@imt-atlantique.fr

IMT Atlantique

Outline

Introduction

- Interpolation : find a function $\hat{f}$, such that $\hat{f}(\mathbf{x}_i) = y_i$, $i = 1 : n$, that approximates f
- Approximation : find a function $\hat{f}$, such that $\hat{f}(\mathbf{x}_i) \approx y_i$, $i = 1 : n$, that approximates f
- Often the interpolation or approximation
  - assumes regularity hypotheses upon f
  - uses a family or a basis of functions to compute $\hat{f}$
- Gaussian processes offer a rather general approach for interpolation and approximation with a probabilistic point of view on the problem. This point of view is emphasized here.

A statistical approach to curve fitting

- Problem : learn a curve $y = f(\mathbf{x})$ from data $(\mathbf{x}_i, y_i)_{i=1:n}$, where $y_i = f(\mathbf{x}_i)$ or $y_i = f(\mathbf{x}_i) + n_i$ (n : noise process).
- If many points : kernel regression curve could be considered
- If few points : some prior upon f is required
- In parametric models, we assume that $f = f_\theta$ where $\theta \in \Theta \subset \mathbb{R}^p$ is a vector of parameters.
- In a parametric Bayesian approach, some prior $p(\theta)$ is assumed and $p(\theta \mid (\mathbf{x}_i, y_i)_{i=1:n})$ should be inferred to estimate θ.
- In Bayesian non parametric approaches, $p(f \mid (\mathbf{x}_i, y_i)_{i=1:n})$ is inferred.
- How to choose a prior for the trajectories of f ? Gaussian processes offer a rather simple and effective answer.

Gaussian processes (GPs)

Definition (GP)

A stochastic process $z = (z_{\mathbf{x}})_{\mathbf{x}\in\mathbb{R}^d}$ is a Gaussian process with mean and covariance parameter functions $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ and we note $z \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ if

$$\forall n \in \mathbb{N}^*,\ \forall \mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{d\times n},\ \mathbf{z} = [z_{\mathbf{x}_1}, \ldots, z_{\mathbf{x}_n}]^T \sim \mathcal{N}\left(\mathbf{m}(\mathbf{x}), \mathbf{K}(\mathbf{x}, \mathbf{x})\right) \tag{1}$$

with $\mathbf{m}(\mathbf{x}) = [m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n)]^T$, $[\mathbf{K}(\mathbf{x}, \mathbf{x})]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and k a bilinear function of the positive type : letting $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$,

$$\forall \alpha_{1:n},\ k(\sum_{i=1:n} \alpha_i \mathbf{x}_i, \sum_{i=1:n} \alpha_i \mathbf{x}_i) = \boldsymbol{\alpha}^T \mathbf{K}(\mathbf{x}, \mathbf{x})\boldsymbol{\alpha} \geq 0. \tag{2}$$

- For the curve fitting problem, $y_i = f(\mathbf{x}_i) + n_i$ (possibly $n = 0$) and the prior over f is given by a GP : $f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$.

Outline

Inference for Gaussian process priors : noiseless case

- $y_i = f(\mathbf{x}_i)$ for $i = 1 : n$. Let $\mathbf{x}_D = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ and $\mathbf{y}_D = [y_1, \ldots, y_n]^T$. Let $\mathbf{x}_I \in \mathbb{R}^{d \times m}$ denote a vector of points for which we want to infer $\mathbf{y}_I = \mathbf{f}(\mathbf{x}_I) = [f(\mathbf{x}_{I,1}), \ldots, f(\mathbf{x}_{I,m})]^T \in \mathbb{R}^m$ from $p(\mathbf{y}_I \mid \mathbf{x}_I, \mathbf{x}_D, \mathbf{y}_D)$. As $f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$,

$$\begin{bmatrix} \mathbf{y}_I \\ \mathbf{y}_D \end{bmatrix} = \begin{bmatrix} \mathbf{f}(\mathbf{x}_I) \\ \mathbf{f}(\mathbf{x}_D) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}(\mathbf{x}_I) \\ \mathbf{m}(\mathbf{x}_D) \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathbf{x}_I, \mathbf{x}_I) & \mathbf{K}(\mathbf{x}_I, \mathbf{x}_D) \\ \mathbf{K}(\mathbf{x}_D, \mathbf{x}_I) & \mathbf{K}(\mathbf{x}_D, \mathbf{x}_D) \end{bmatrix} \right) \tag{3}$$

  Then :

$$\mathbf{f}(\mathbf{x}_I) \mid \mathbf{x}_I, \mathbf{x}_D, \mathbf{y}_D \sim \mathcal{N}\left(\mathbf{m}_{post}(\mathbf{x}_I), \mathbf{K}_{post}(\mathbf{x}_I, \mathbf{x}_I)\right) \tag{4}$$

  with

$$\begin{aligned} \mathbf{m}_{post}(\mathbf{x}_I) &= \mathbf{m}(\mathbf{x}_I) + \mathbf{K}(\mathbf{x}_I, \mathbf{x}_D)\mathbf{K}(\mathbf{x}_D, \mathbf{x}_D)^{-1}(f(\mathbf{x}_D) - \mathbf{m}(\mathbf{x}_D)) \\ \mathbf{K}_{post}(\mathbf{x}_I, \mathbf{x}_I) &= \mathbf{K}(\mathbf{x}_I, \mathbf{x}_I) - \mathbf{K}(\mathbf{x}_I, \mathbf{x}_D)\mathbf{K}(\mathbf{x}_D, \mathbf{x}_D)^{-1}\mathbf{K}(\mathbf{x}_D, \mathbf{x}_I) \end{aligned} \tag{5}$$

- Common choices :
  - $m(\mathbf{x}) = 0$.
  - $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp(\frac{-1}{2} \parallel \mathbf{x} - \mathbf{x}' \parallel_{\mathbf{\Sigma}^{-1}})$ with $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$.

Inference for Gaussian process priors : noisy case

- $y_i = f(\mathbf{x}_i) + n_i$ with $\mathbb{E}[n_i^2] = \sigma_n^2$. Then, $\text{cov}(y_i, y_j) = k(x_i, x_j) + \sigma_n^2 \delta_{i,j}$. Then we get the same results as before but with $\mathbf{K}(\mathbf{x}_D, \mathbf{x}_D)$ changed to $\mathbf{K}(\mathbf{x}_D, \mathbf{x}_D) + \sigma_n^2 \mathbf{I}$:

$$f(\mathbf{x}_I) \mid \mathbf{x}_I, \mathbf{x}_D, \mathbf{y}_D \sim \mathcal{N}\left(\mathbf{m}_{post}^n(\mathbf{x}_I), \mathbf{K}_{post}^n(\mathbf{x}_I, \mathbf{x}_I)\right) \tag{6}$$

with

$$\mathbf{m}_{post}^n(\mathbf{x}_I) = \mathbf{m}(\mathbf{x}_I) + \mathbf{K}(\mathbf{x}_I, \mathbf{x}_D)[\mathbf{K}(\mathbf{x}_D, \mathbf{x}_D) + \sigma_n^2 \mathbf{I}]^{-1}(f(\mathbf{x}_D) - \mathbf{m}(\mathbf{x}_D))$$

$$\mathbf{K}_{post}^n(\mathbf{x}_I, \mathbf{x}_I) = \mathbf{K}(\mathbf{x}_I, \mathbf{x}_I) - \mathbf{K}(\mathbf{x}_I, \mathbf{x}_D)[\mathbf{K}(\mathbf{x}_D, \mathbf{x}_D) + \sigma_n^2 \mathbf{I}]^{-1}\mathbf{K}(\mathbf{x}_D, \mathbf{x}_I) \tag{7}$$

Example

- $y = \sin(2x) + \mathcal{N}(0, \sigma_n^2)$
- Noiseless and noisy cases with $\sigma_n = 0$ and $\sigma_n = .1$
- $m(x) = 0$ and $k(x, x') = \sigma_f^2 e^{-\frac{(x-x')^2}{2\sigma_x^2}}$ $(+\sigma_y^2 \delta_{x,x'}$ for data)

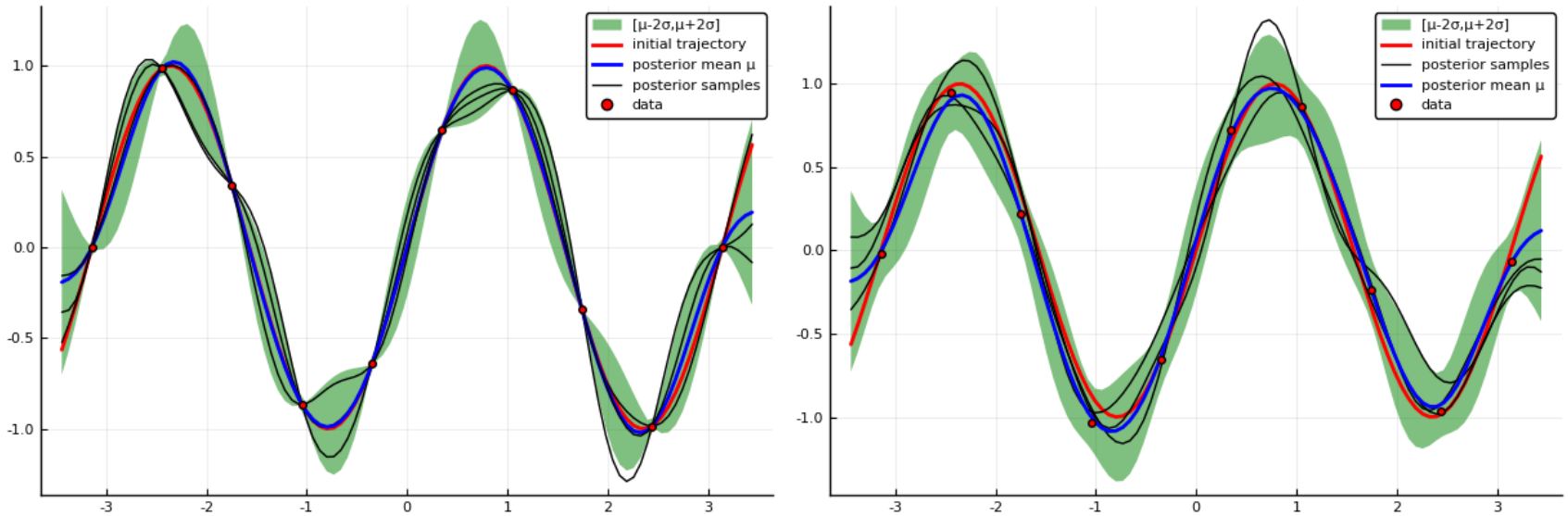

Figure – $\sigma_n = 0$ (left) and $\sigma_n = .1$ (right). $(\sigma_x, \sigma_f, \sigma_y) = (.5, .5, 0)$ (left) and $(\sigma_x, \sigma_f, \sigma_y) = (.5, .5, .1)$ (right)

Outline

Parameter estimation : grid method

- To make a convenient choice for the parameters one can choose to maximize the likelihood $p(\mathbf{y}_D, \mathbf{x}_D \mid \sigma_x, \sigma_f, \sigma_y) \propto p(\mathbf{y}_D \mid \mathbf{x}_D, \sigma_x, \sigma_f, \sigma_y)$: $\log p(\mathbf{y}_D \mid \sigma_x, \sigma_f, \sigma_y) = -\frac{1}{2}\left(\mathbf{y}_D^T \mathbf{K}^{-1} \mathbf{y}_D + \log|\mathbf{K}| + n\log(2\pi)\right)$ with $\mathbf{K}_{ij} = \sigma_f^2 e^{-\frac{(x_i - x_j)^2}{2\sigma_x^2}} + \sigma_y^2 \delta_{i,j}$.
- Grid search : example $y = \sin(2x)$

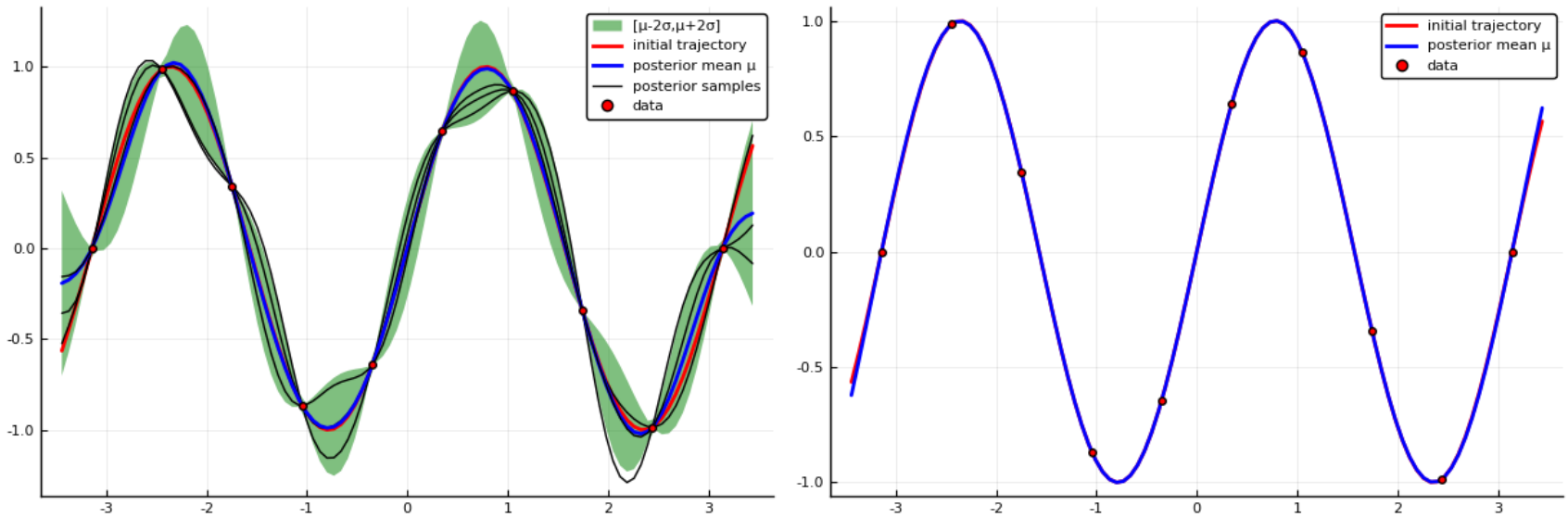

Figure – $(\sigma_x^2, \sigma_f^2) \in [.01, 5] \times [.01, 5]$, 100×100 grid $\rightarrow$ left : $(\sigma_x, \sigma_f) = (.5, .5)$; right : $(\sigma_{x,MV}, \sigma_{f,MV}) = (1.27, 1.82)$

Parameter estimation : descent method

$$L(\theta) = \log p(\mathbf{y}_D \mid \theta) = -\frac{1}{2}\left(\mathbf{y}_D^T \mathbf{K}^{-1} \mathbf{y}_D + \log|\mathbf{K}| + n\log(2\pi)\right)$$

with $\theta = (\sigma_x^2, \sigma_f^2, \sigma_y^2)$. Then

$$\frac{\partial L(\theta)}{\partial \theta_k} = \frac{1}{2} Tr\left((\mathbf{K}^{-1}\mathbf{y}_D\mathbf{y}_D^T\mathbf{K}^{-1} - \mathbf{K}^{-1})\frac{\partial \mathbf{K}}{\partial \theta_k}\right)$$

- Exercise : check the above expression for $\dfrac{\partial L(\theta)}{\partial \theta_k}$.

Parameter estimation : descent method (II)

- For $\mathbf{K}_{ij} = \sigma_f^2 e^{-\frac{(x_i - x_j)^2}{2\sigma_x^2}} + \sigma_y^2 \delta_{i,j}$,

$$\begin{aligned}
\frac{\partial \mathbf{K}}{\partial \sigma_x^2} &= -\frac{1}{2\sigma_x^4}(\mathbf{x}_D \mathbb{1}^T - \mathbb{1}\mathbf{x}_D^T) \odot (\mathbf{x}_D \mathbb{1}^T - \mathbb{1}\mathbf{x}_D^T) \odot (\mathbf{K} - \sigma_y^2 \mathbf{I}) \\
\frac{\partial \mathbf{K}}{\partial \sigma_f^2} &= \frac{1}{\sigma_f^2}(\mathbf{K} - \sigma_y^2 \mathbf{I}) \qquad (8) \\
\frac{\partial \mathbf{K}}{\partial \sigma_y^2} &= \mathbf{I}
\end{aligned}$$

  where $\odot$ is the Hadamard product : $[A \odot B]_{ij} = A_{ij} B_{ij}$

- Usual descent techniques can be used but often there are local minima !

UE Stochastic Dynamic models - Summary

- We have studied
  - Measure theory : definitions and theorems (Radon-Nikodym-Lebesgue theorem)
  - Optimal filtering of stochastic processes : stationary case (Wiener filters), linear state space models and Kalman filter, non linear state space models (particle filters)
  - Stochastic differential equations : complements of probabilities and Brownian motion, Itô integration and Itô formula, analytical and numerical integration of SDEs, parameter estimation for SDEs.
  - Time series analysis : AR, ARMA, ARIMA, ...
  - Interpolation and approximation via Gaussian process priors.
- Topics that we did not cover : prediction theory WSS processes, more advanced Kalman and particle filters, SDEs with jumps, deterministic dynamic models (chaos, ...), models estimation and control.
- What could be done to improve the UE : duration, contents, ... ?