



UNIVERSIDADE
FEDERAL DO CEARÁ



Aprendizagem de Máquina Probabilística

César Lincoln Cavalcante Mattos

2025

Agenda

① Processos Gaussianos

GPs para Regressão

De espaços de atributos para GPs

② GPs para grandes conjuntos de dados

GP esparso variacional

GP com inferência variacional estocástica

③ Otimização Bayesiana

④ Outros tópicos

Classificação

Aprendizagem não-supervisionada e redução de dimensionalidade

Modelagem hierárquica

Modelagem dinâmica

Robótica e controle

Aprendizagem robusta

⑤ Conclusão

⑥ Referências

Ilustração inicial - Teorema de Bayes

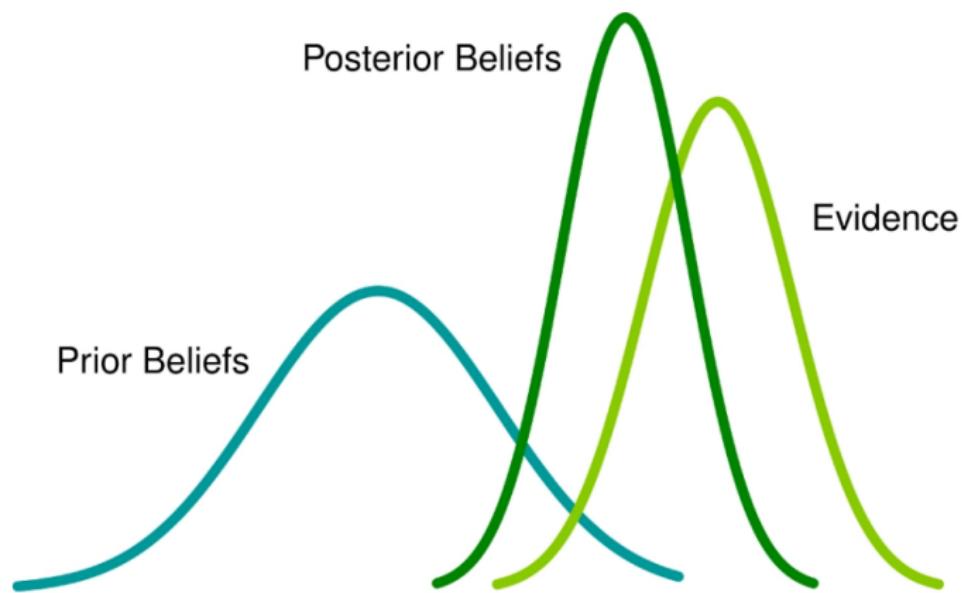
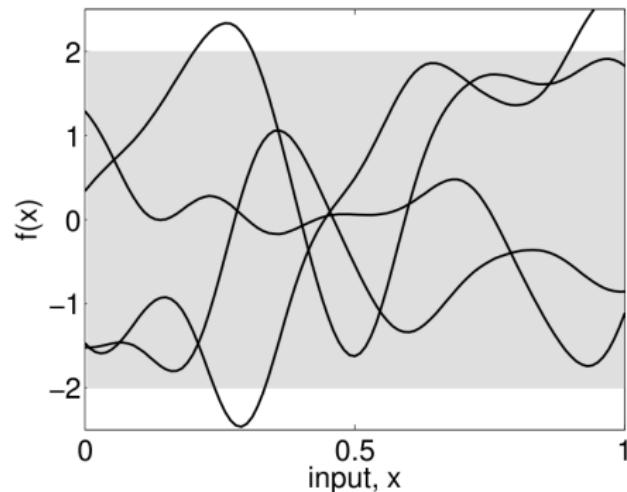
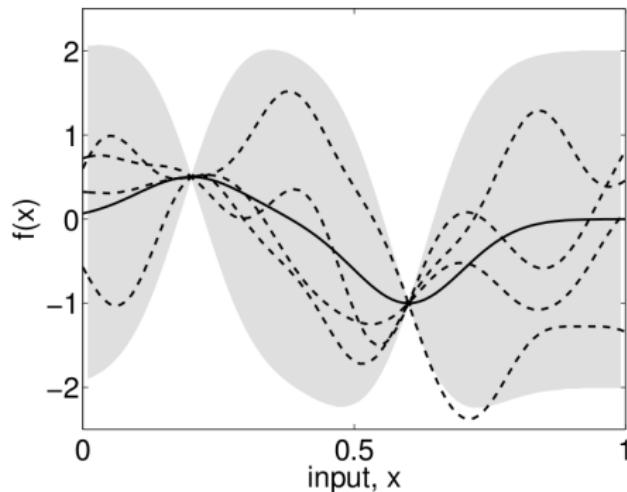


Ilustração inicial - Teorema de Bayes



Amostras da priori (antes da observação dos dados).

- Temos inicialmente infinitas possibilidades de funções contínuas.
- Escolhemos somente aquelas que explicam os dados observados.



Amostras da posteriori (depois da observação dos dados).

Motivação

Por que processos Gaussianos?

- Modelos **paramétricos** agem como um **gargalo** entre os dados e as previsões.
- Equilíbrio entre **ajuste aos dados** e **regularização**.
- (Poucos) **hiperparâmetros** podem ser ajustados diretamente.
- A maior parte do arcabouço teórico deriva de propriedades úteis da **distribuição Gaussiana multivariada**.
- **Modelagem Bayesiana não paramétrica**.

Motivação

Modelagem Bayesiana não paramétrica

- A **incerteza** está relacionada ao grau do **conhecimento** que temos.
- Ausência de um número **finito** de parâmetros.
- A **complexidade** do modelo cresce com o número de **observações**.
- **Processos Gaussianos (Gaussian process, GP)**: abordagem probabilística para a modelagem via máquinas de kernel.

Distribuição Gaussiana Multivariada

Definição

Considere um vetor aleatório $\mathbf{f} \in \mathbb{R}^N$ que segue uma distribuição Gaussiana multivariada:

$$p(\mathbf{f} | \boldsymbol{\mu}, \mathbf{K}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{f} - \boldsymbol{\mu})^\top \mathbf{K}^{-1} (\mathbf{f} - \boldsymbol{\mu})\right).$$

A distribuição é completamente definida pelo seu **vetor de média** $\boldsymbol{\mu} \in \mathbb{R}^N$ e sua **matriz de covariância** $\mathbf{K} \in \mathbb{R}^{N \times N}$.

Duas propriedades importantes

Considere a seguinte coleção de variáveis aleatórias:

$$\mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}), \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}.$$

Marginalização

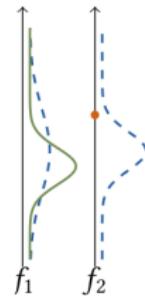
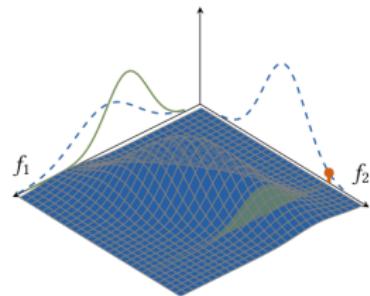
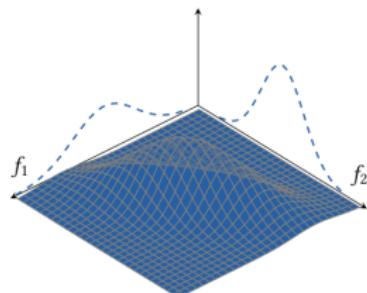
A observação de uma coleção maior de variáveis não afeta a distribuição de subconjuntos menores, o que implica que $f_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{K}_{11})$ e $f_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{K}_{22})$.

Condicionamento

Condicionamento por Gaussianas resulta em uma nova distribuição Gaussiana dada por

$$p(\mathbf{f}_1 | \mathbf{f}_2 = \mathbf{y}) = \mathcal{N}(\mathbf{f}_1 | \boldsymbol{\mu}_1 + \mathbf{K}_{12}\mathbf{K}_{22}^{-1}(\mathbf{y} - \boldsymbol{\mu}_2), \mathbf{K}_{11} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\mathbf{K}_{21}).$$

Propriedades da Gaussiana



$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right)$$

$$f_1 | f_2 \sim \mathcal{N} \left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2} (f_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}\sigma_{21}}{\sigma_2^2} \right)$$

- A observação de uma variável nos fornece informação sobre a outra.

Processos Gaussianos

Definição

- Distribuição sobre **funções**.

Processos Gaussianos

Definição

- Distribuição sobre **funções**.
- Permite a análise de funções (**objetos infinitos**) através de distribuições finitas (**Gaussianas multivariadas**).

Processos Gaussianos

Definição

- Distribuição sobre **funções**.
- Permite a análise de funções (**objetos infinitos**) através de distribuições finitas (**Gaussianas multivariadas**).
- Seja $\mathbf{X} \in \mathbb{R}^{N \times D}$ a reunião de padrões $\mathbf{x}_i \in \mathbb{R}^D$, definimos uma **GP como priori** para o vetor f :

$$\mathbf{f} = f(\mathbf{X}) \sim \mathcal{N}(\mathbf{f} | \mathbf{0}, \mathbf{K}).$$

- Escolhemos um vetor de zeros como **média a priori**, uma escolha comum.
- O vetor de avaliações f é uma única amostra de uma GP.

Amostras de uma GP

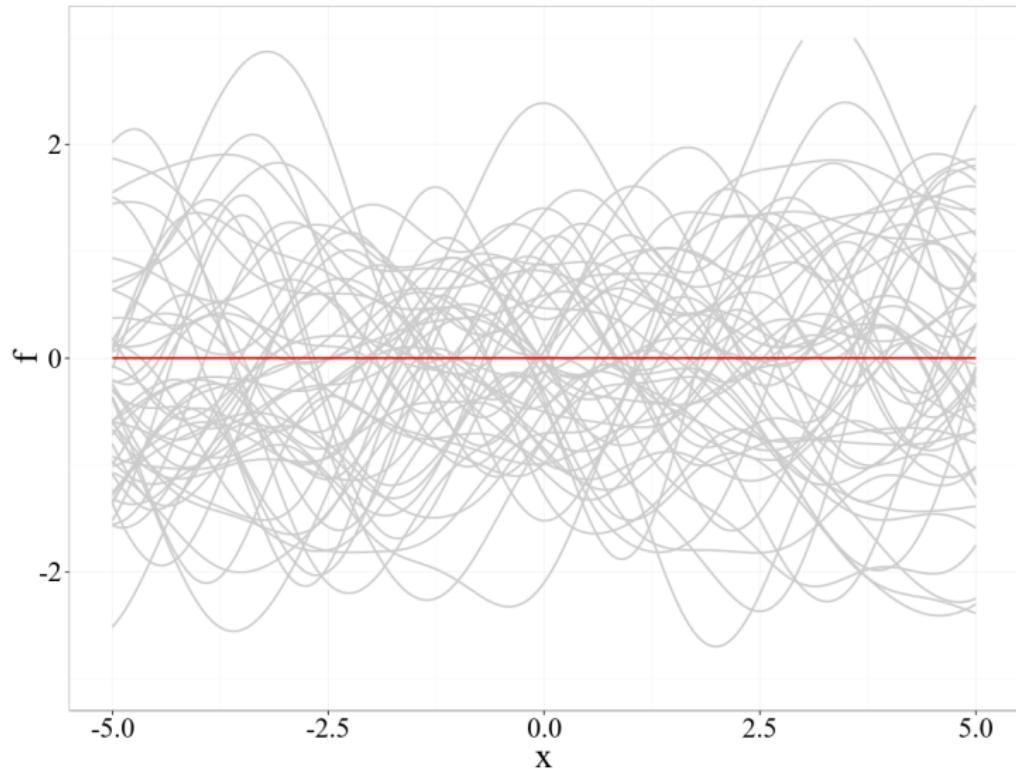


Figura 1: Amostras de uma priori de GP.

Regressão não-linear

- Seja o conjunto de dados

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) |_{i=1}^N\} = (\mathbf{X}, \mathbf{y}),$$

em que $\mathbf{x}_i \in \mathbb{R}^D$ são as entradas, $\mathbf{X} \in \mathbb{R}^{N \times D}$ e $\mathbf{y} \in \mathbb{R}^N$ são as saídas **observadas**.

- Uma tarefa geral de regressão não-linear pode ser expressada por

$$y_i = f(\mathbf{x}_i) + \epsilon_i,$$

em que $\epsilon_i \in \mathbb{R}$ é o ruído de observação.

- Os valores $f_i = f(\mathbf{x}_i)$ não são observados diretamente (eles são **latentes**) e $f(\cdot)$ é desconhecido.

Modelagem padrão com GPs

- Escolha uma **Gaussiana multivariada como priori** para a função desconhecida:

$$\mathbf{f} = f(\mathbf{X}) \sim \mathcal{N}(\mathbf{f} | \mathbf{0}, \mathbf{K}),$$

em que $\mathbf{K} \in \mathbb{R}^{N \times N}$, $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, é a matriz de covariância, obtida pela **função de kernel ou de covariância** $k(\cdot, \cdot)$.

- Note que o comportamento da saída é inteiramente representado pela matriz \mathbf{K} , obtida a partir de relações entre as entradas.

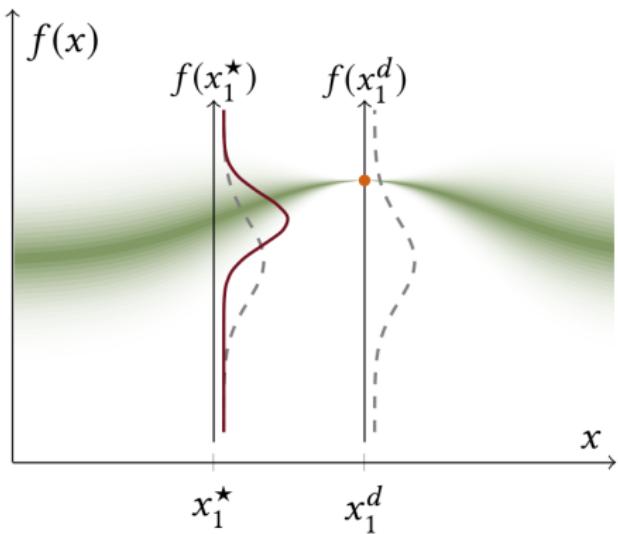
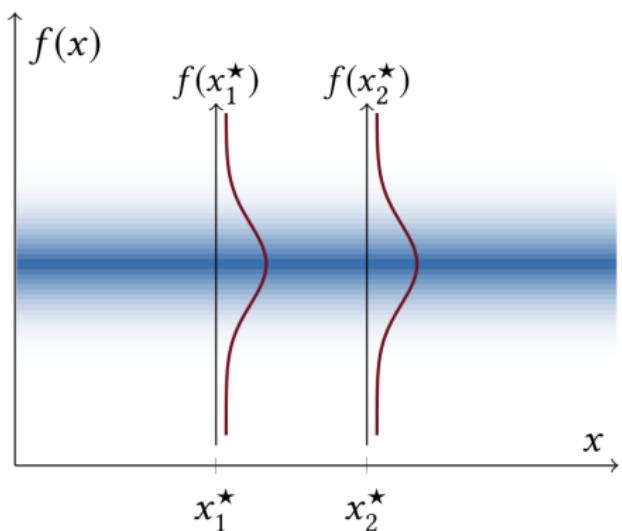
Modelagem padrão com GPs

- Uma escolha comum é a **exponencial quadrática** (*squared exponential* ou *Radial Basis Function* - RBF):

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{d=1}^D w_d^2 (x_{id} - x_{jd})^2\right).$$

- O vetor de **hiperparâmetros** $\theta = [\sigma_f^2, w_1^2, \dots, w_D^2]^\top$ caracteriza a covariância do modelo.

Modelagem padrão com GPs



Modelagem padrão com GPs

- Se o ruído de observação é Gaussiano, i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma_y^2)$, todas as quantidades de interesse são **analíticas**.

Verossimilhança

A verossimilhança dada as observações é Gaussiana:

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 \mathbf{I}),$$

em que $\mathbf{I} \in \mathbb{R}^{N \times N}$ é uma matriz identidade.

Modelagem padrão com GPs

Verossimilhança marginal

A verossimilhança marginal de \mathbf{y} é calculada integrando-se \mathbf{f} :

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f} = \int \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})d\mathbf{f}, \\ &= \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma_y^2 \mathbf{I}). \end{aligned}$$

Modelagem padrão com GPs

Verossimilhança marginal

A verossimilhança marginal de \mathbf{y} é calculada integrando-se \mathbf{f} :

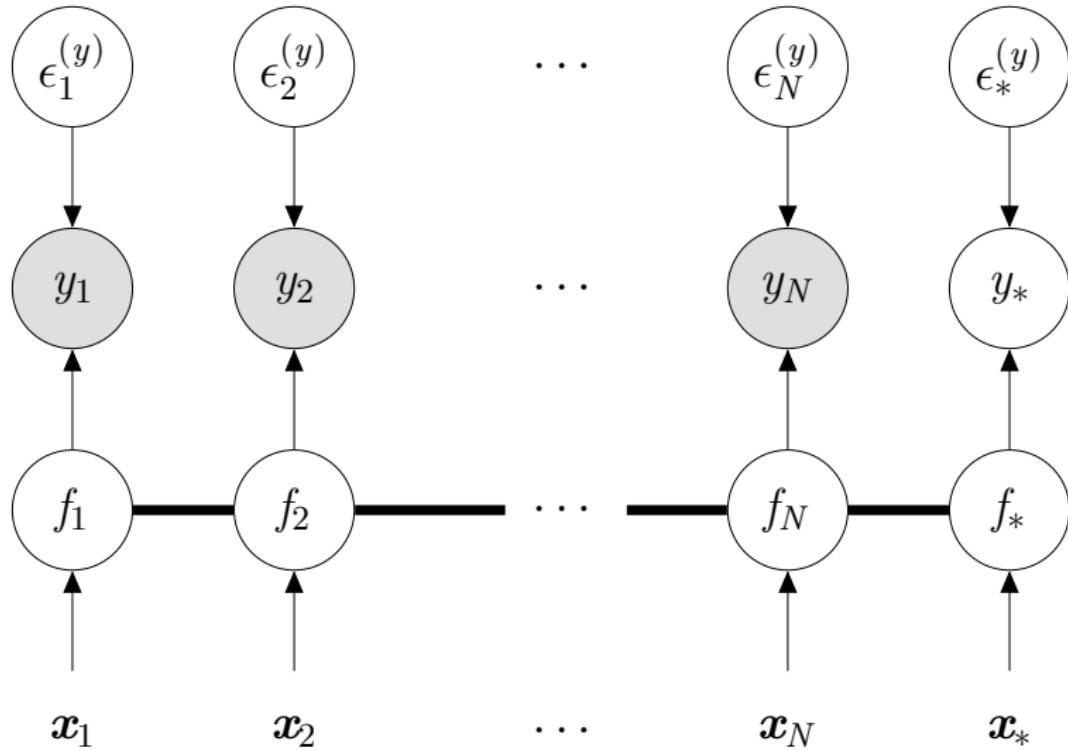
$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f} = \int \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})d\mathbf{f}, \\ &= \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma_y^2 \mathbf{I}). \end{aligned}$$

Posteriori

A posteriori é analítica e também Gaussiana:

$$\begin{aligned} \underbrace{p(\mathbf{f}|\mathbf{y}, \mathbf{X})}_{\text{posteriori}} &= \frac{\overbrace{p(\mathbf{y}|\mathbf{f})}^{\text{verossimilhança}} \overbrace{p(\mathbf{f}|\mathbf{X})}^{\text{priori}}}{\underbrace{p(\mathbf{y}|\mathbf{X})}_{\text{verossimilhança marginal}}} = \frac{\mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 \mathbf{I})\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})}{\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma_y^2 \mathbf{I})}, \\ &= \mathcal{N}(\mathbf{f}|\mathbf{K}(\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1}\mathbf{y}, \mathbf{K} - \mathbf{K}(\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1}\mathbf{K}). \end{aligned}$$

Modelagem padrão com GPs - Modelo gráfico



Fazendo previsões com GPs

Inferência para f_* dado \mathbf{x}_* é obtida via condicionamento:

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma_y^2 \mathbf{I} & \mathbf{k}_{f*} \\ \mathbf{k}_{*f} & k_{**} \end{bmatrix}\right),$$

$$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(f_* | \mu_*, \sigma_*^2), \quad p(y_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(y_* | \mu_*, \sigma_*^2 + \sigma_y^2),$$

$$\mu_* = \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y},$$

$$\sigma_*^2 = k_{**} - \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{k}_{f*},$$

em que $\mathbf{k}_{f*} = [k(\mathbf{x}_*, \mathbf{x}_1), \dots, k(\mathbf{x}_*, \mathbf{x}_N)]^\top$, $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$.

Fazendo previsões com GPs

Inferência para f_* dado \mathbf{x}_* é obtida via condicionamento:

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma_y^2 \mathbf{I} & \mathbf{k}_{f*} \\ \mathbf{k}_{*f} & k_{**} \end{bmatrix}\right),$$

$$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(f_* | \mu_*, \sigma_*^2), \quad p(y_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(y_* | \mu_*, \sigma_*^2 + \sigma_y^2),$$

$$\mu_* = \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y},$$

$$\sigma_*^2 = k_{**} - \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{k}_{f*},$$

$$\text{em que } \mathbf{k}_{f*} = [k(\mathbf{x}_*, \mathbf{x}_1), \dots, k(\mathbf{x}_*, \mathbf{x}_N)]^\top, \quad k_{**} = k(\mathbf{x}_*, \mathbf{x}_*).$$

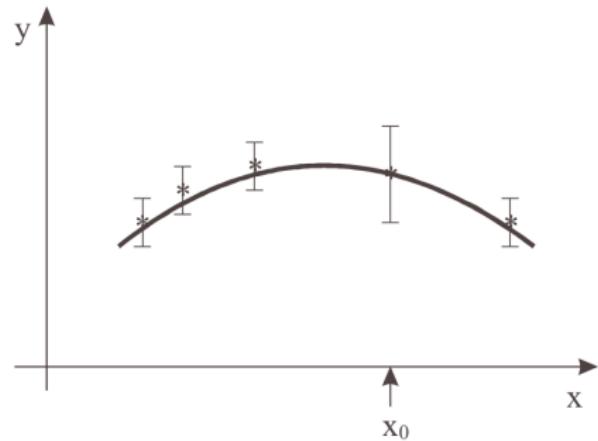
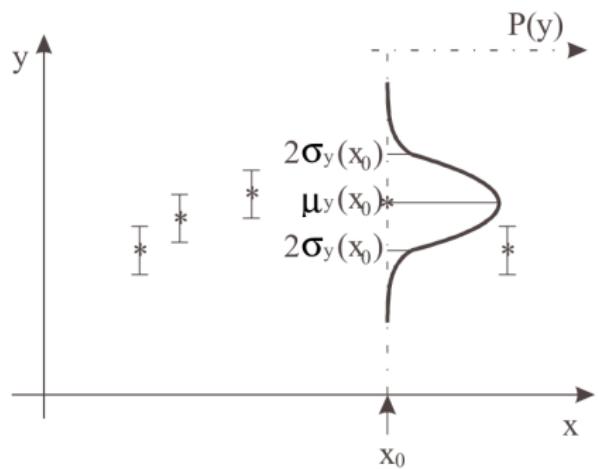
Implementação mais estável via Cholesky

$$\mathbf{K} + \sigma_y^2 \mathbf{I} = \mathbf{L} \mathbf{L}^\top, \quad (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} = \mathbf{L}^{-\top} \mathbf{L}^{-1},$$

$$\mu_* = (\mathbf{L}^{-1} \mathbf{k}_{f*})^\top \mathbf{L}^{-1} \mathbf{y},$$

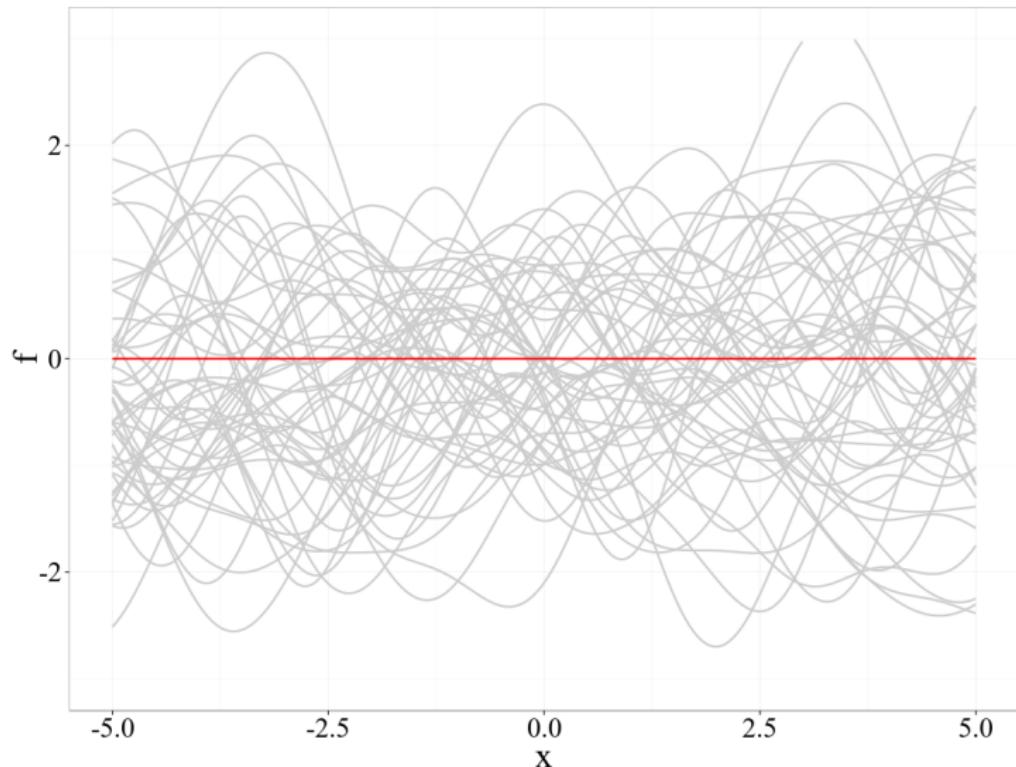
$$\sigma_*^2 = k_{**} - (\mathbf{L}^{-1} \mathbf{k}_{f*})^\top \mathbf{L}^{-1} \mathbf{k}_{f*}.$$

Distribuição preditiva de um GP



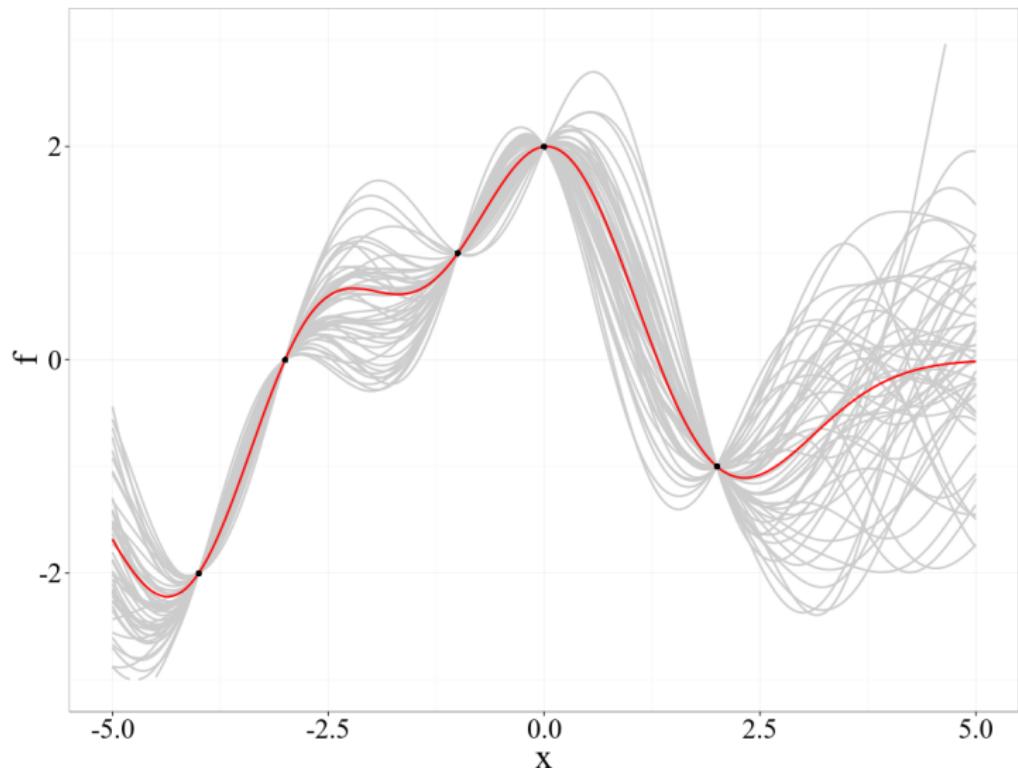
Note que cada predição é uma distribuição Gaussiana bem definida.

Amostras de um GP



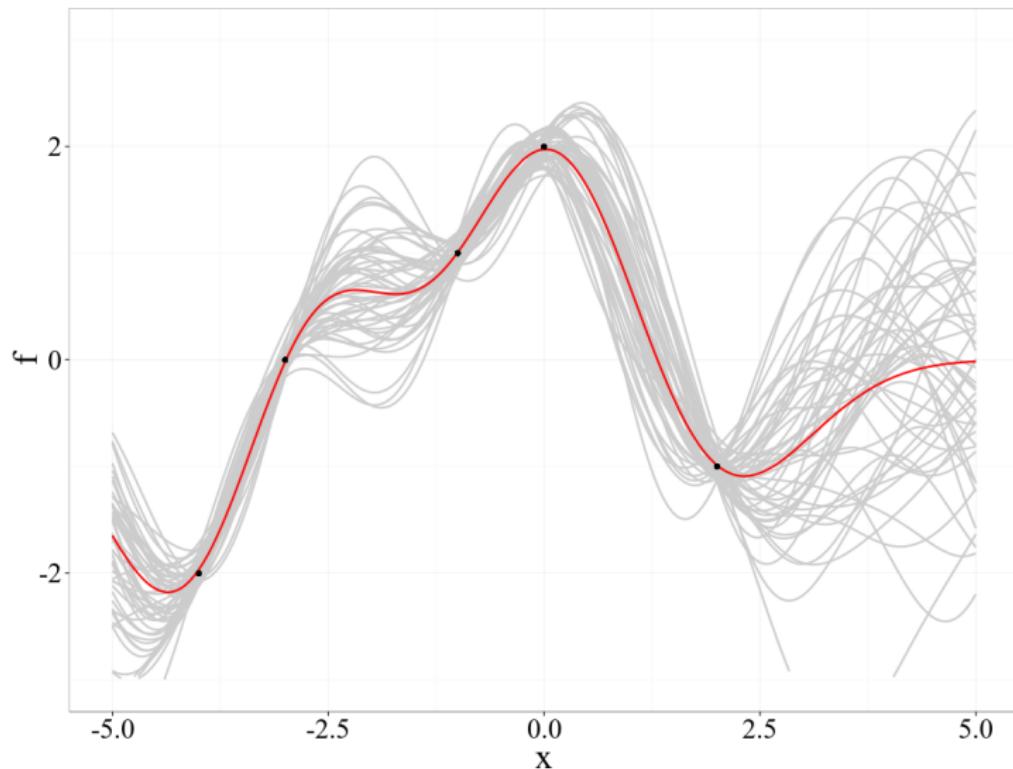
Amostras da priori de um GP com $\sigma_f^2 = 1$, $w = 1$ e $\sigma_y^2 = 0$.

Amostras de um GP



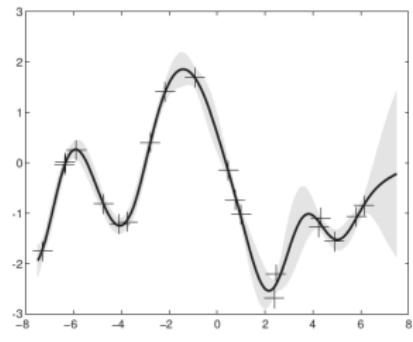
Amostras da posteriori (após y) sem ruído ($\sigma_y^2 = 0$).

Amostras de um GP

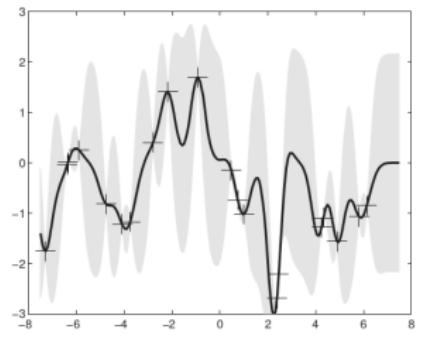


Amostras da posteriori (após y) com ruído ($\sigma_y^2 = 0.01$).

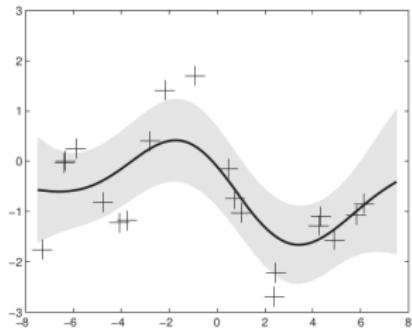
Amostras de um GP



(a)



(b)



(c)

Amostras da posteriori de um GP. Em b) temos lengthscales menores ($\frac{1}{w}$), em c) temos ruído de observação com variância maior.

Seleção de modelos de GP

Otimização dos hiperparâmetros

O vetor de hiperparâmetros $\boldsymbol{\theta} = [\sigma_f^2, w_1^2, \dots, w_D^2, \sigma_y^2]^\top$ pode ser otimizado via **maximização da log-verossimilhança marginal** $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$, a chamada *evidência* do modelo:

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2} \underbrace{\log |\mathbf{K} + \sigma_y^2 \mathbf{I}|}_{\text{capacidade do modelo}} - \frac{1}{2} \underbrace{\mathbf{y}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y}}_{\text{ajuste aos dados}} - \frac{N}{2} \log(2\pi).$$

Seleção de modelos de GP

Otimização dos hiperparâmetros

O vetor de hiperparâmetros $\boldsymbol{\theta} = [\sigma_f^2, w_1^2, \dots, w_D^2, \sigma_y^2]^\top$ pode ser otimizado via **maximização da log-verossimilhança marginal** $\mathcal{L}(\boldsymbol{\theta}) = \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$, a chamada *evidência* do modelo:

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2} \underbrace{\log |\mathbf{K} + \sigma_y^2 \mathbf{I}|}_{\text{capacidade do modelo}} - \frac{1}{2} \underbrace{\mathbf{y}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y}}_{\text{ajuste aos dados}} - \frac{N}{2} \log(2\pi).$$

Implementação mais estável via Cholesky

$$\mathbf{K} + \sigma_y^2 \mathbf{I} = \mathbf{L} \mathbf{L}^\top, \quad \boldsymbol{\alpha} = \mathbf{L}^{-1} \mathbf{y},$$

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_i \log L_{ii} - \frac{1}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{N}{2} \log(2\pi).$$

Seleção de modelos de GP

Otimização dos hiperparâmetros

- A otimização segue **gradientes analíticos**:

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} = & -\frac{1}{2} \operatorname{Tr} \left((\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \frac{\partial (\mathbf{K} + \sigma_y^2 \mathbf{I})}{\partial \theta_i} \right) \\ & + \frac{1}{2} \mathbf{y}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \frac{\partial (\mathbf{K} + \sigma_y^2 \mathbf{I})}{\partial \theta_i} (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y}.\end{aligned}$$

- **Não é preciso fazer validações cruzadas ou grid-search** na seleção de modelos!

Seleção de modelos de GP

Otimização dos hiperparâmetros

- A otimização segue **gradientes analíticos**:

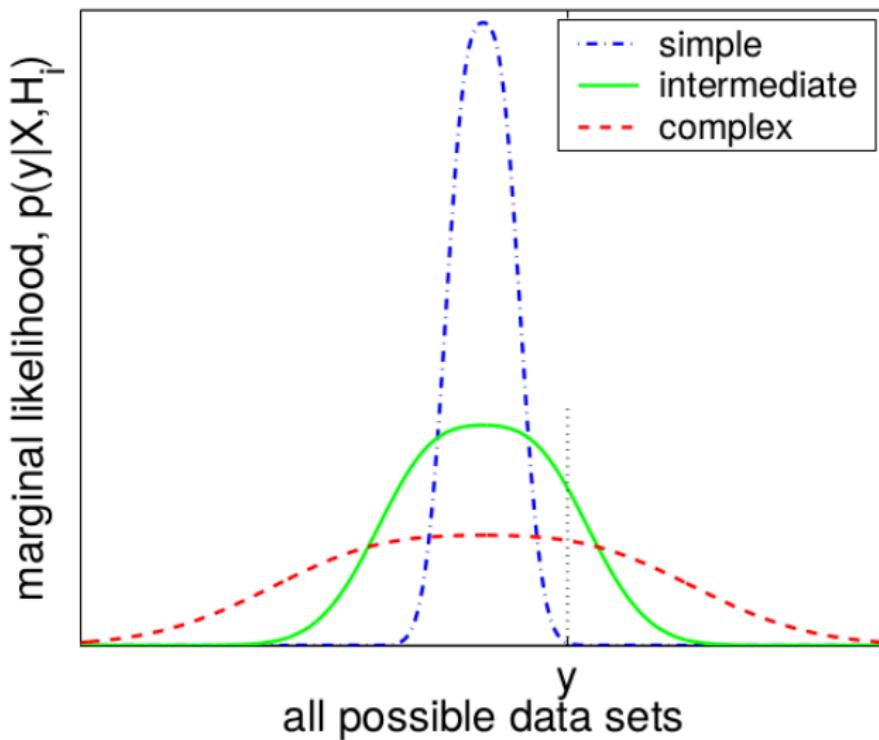
$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} = & -\frac{1}{2} \operatorname{Tr} \left((\boldsymbol{K} + \sigma_y^2 \boldsymbol{I})^{-1} \frac{\partial (\boldsymbol{K} + \sigma_y^2 \boldsymbol{I})}{\partial \theta_i} \right) \\ & + \frac{1}{2} \boldsymbol{y}^\top (\boldsymbol{K} + \sigma_y^2 \boldsymbol{I})^{-1} \frac{\partial (\boldsymbol{K} + \sigma_y^2 \boldsymbol{I})}{\partial \theta_i} (\boldsymbol{K} + \sigma_y^2 \boldsymbol{I})^{-1} \boldsymbol{y}.\end{aligned}$$

- **Não é preciso fazer validações cruzadas ou grid-search** na seleção de modelos!

Automatic Relevance Determination (ARD)

Os hiperparâmetros otimizados w_1^2, \dots, w_D^2 podem indicar a relevância das dimensões de entrada, pois dimensões menos relevantes tendem a ser associadas a valores de w_d^2 menores.

Seleção Bayesiana de modelos



A evidência privilegia modelos de complexidade intermediária.

Inferência com modelos de GP para regressão

Resumo do algoritmo

- Passo de estimação (seleção de modelos)

① Inicialize os hiperparâmetros $\boldsymbol{\theta} = \left[\sigma_f^2, w_1^2, \dots, w_D^2, \sigma_y^2 \right]^\top$;

→ Exemplo: $\sigma_f^2 = \mathbb{V}[\mathbf{y}]$, $w_d^2 = \frac{1}{\mathbb{V}[\mathbf{X}_{:d}]}$, $\sigma_y^2 = 0.01\sigma_f^2$.

② Repita até convergir ou por um número máximo de iterações:

① Calcule a evidência do modelo $\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$;

② Calcule os gradientes analíticos $\frac{\partial \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$;

③ Atualize $\boldsymbol{\theta}$ a partir dos gradientes calculados (e.g. via BFGS);

④ Retorne os hiperparâmetros otimizados $\hat{\boldsymbol{\theta}}$.

- Passo de predição

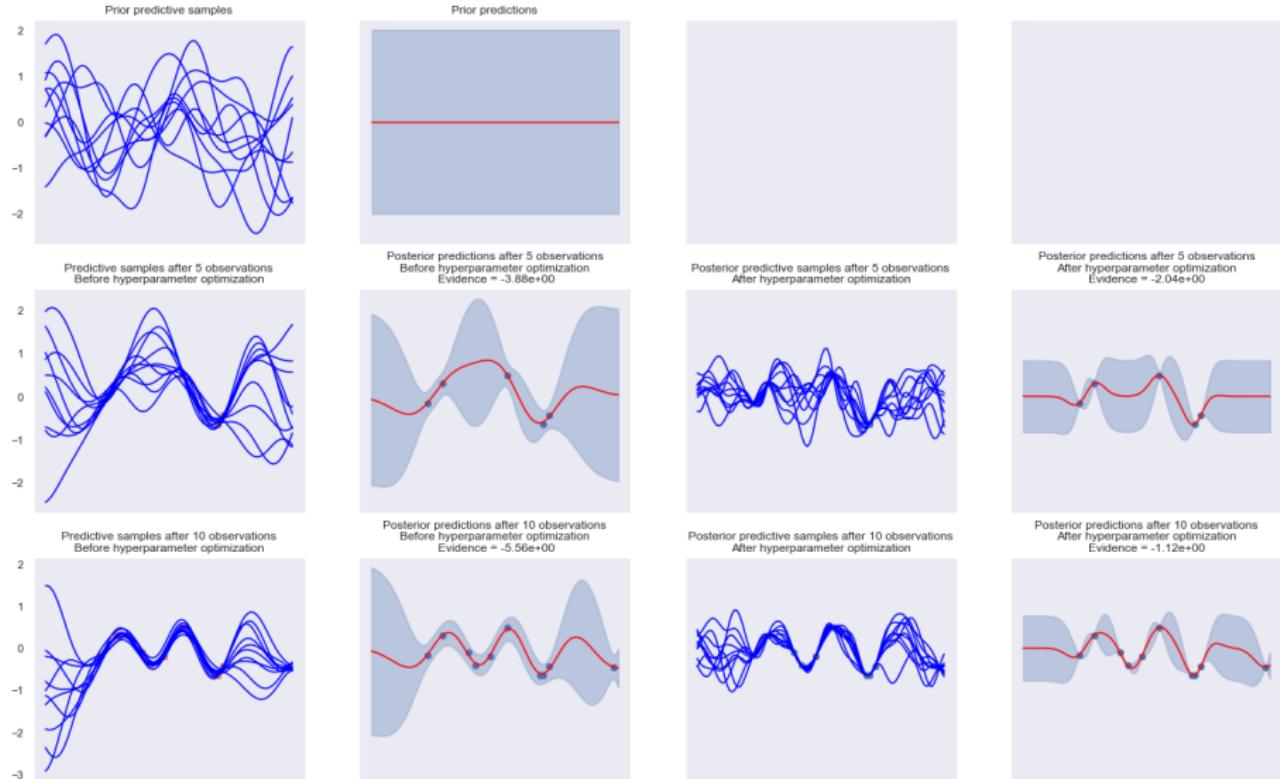
① Dado um novo padrão \mathbf{x}_* , retorne a distribuição preditiva

$$p(y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X}, \hat{\boldsymbol{\theta}}) = \mathcal{N}(y_*|\mu_*, \sigma_*^2 + \sigma_y^2),$$

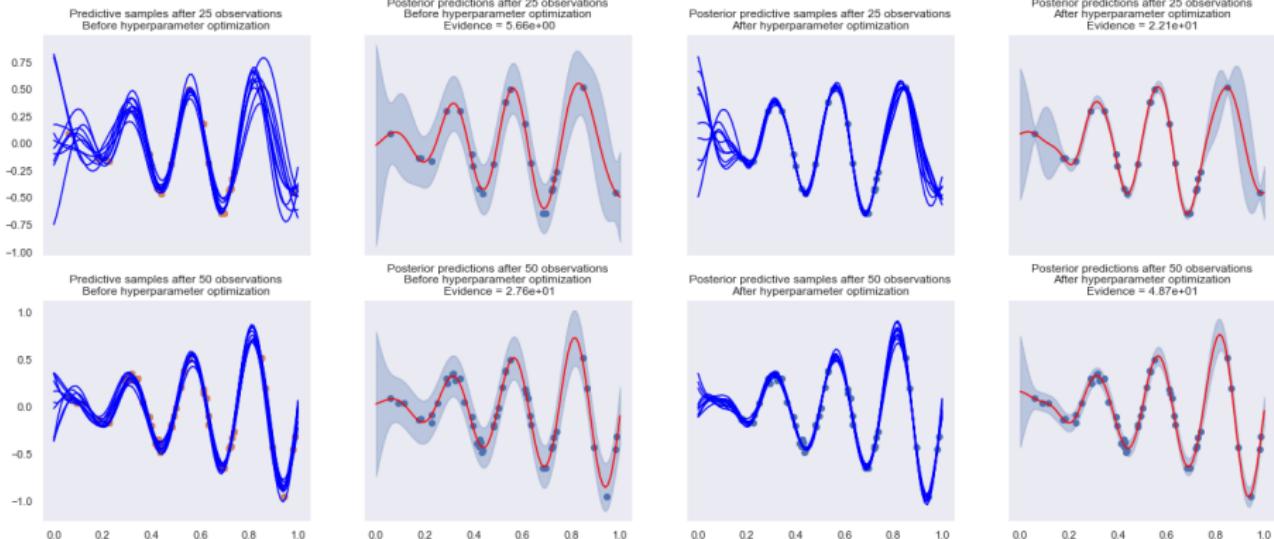
$$\mu_* = \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y},$$

$$\sigma_*^2 = k_{**} - \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{k}_{f*}.$$

GP para regressão com kernel RBF



GP para regressão com kernel RBF



Avaliação de GP para regressão

- É comum avaliarmos um modelo de regressão pelo valor RMSE (*root mean square error*) em dados não usados no treinamento:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N_{\text{teste}}} (y_i - \hat{y}_i)^2}{N}},$$

em que y_i é o valor observado no teste e \hat{y}_i é a predição.

- No caso de modelos de GP, podemos calcular o RMSE usando a média predita $\hat{y}_i = \hat{\mu}_i$.
- O RMSE pode ser usado para avaliar GPs quando comparado a modelos que não fornecem distribuições preditivas.

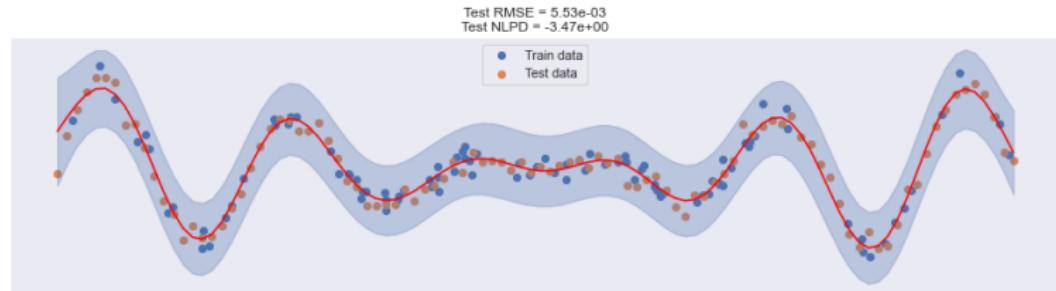
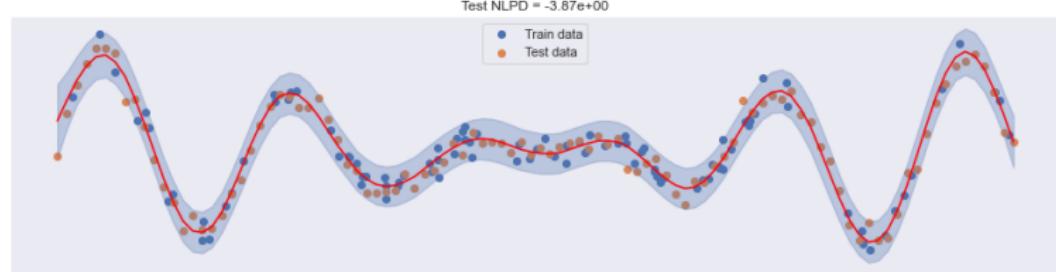
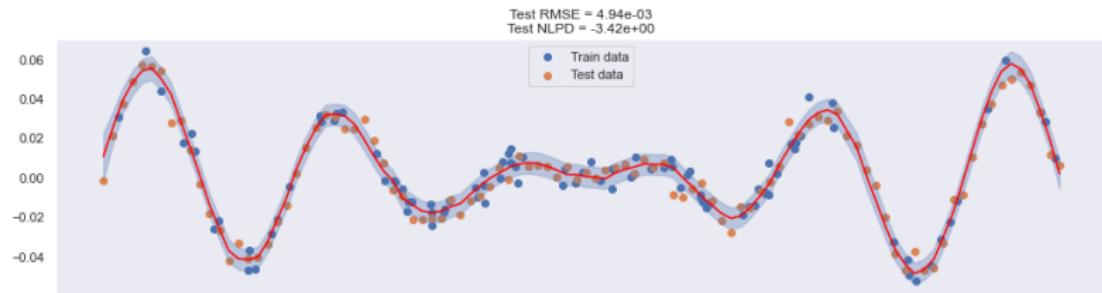
Avaliação de GP para regressão

- Podemos levar em consideração a variância predita $\hat{\sigma}_i^2$, usando a métrica NLPD (quanto menor, melhor):

$$\begin{aligned} \text{NLPD} &= -\frac{1}{N_{\text{teste}}} \sum_{i=1}^{N_{\text{teste}}} \underbrace{\log p(y_i | \mathbf{x}_i, \mathbf{y}, \mathbf{X}, \hat{\boldsymbol{\theta}})}_{\mathcal{N}(y_i | \hat{\mu}_i, \hat{\sigma}_i^2)} \\ &= \frac{1}{2} \log 2\pi + \frac{1}{2N_{\text{teste}}} \sum_{i=1}^{N_{\text{teste}}} \left[\log \hat{\sigma}_i^2 + \frac{(y_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2} \right]. \end{aligned}$$

- Note que a NLPD penaliza modelos muito confiantes ($\hat{\sigma}_i^2$ baixos para erros maiores) e pouco confiantes ($\hat{\sigma}_i^2$ altos para erros menores), realçando os equilibrados.
- A NLPD é preferível na avaliação de modelos que fornecem distribuições preditivas, especialmente quando as observações de teste são ruidosas.

Avaliação de GP para regressão



A função de kernel/covariância

- A escolha da função de kernel/covariância afeta diretamente as **características** do modelo.
- A exponencial quadrática (RBF), por exemplo, considera um grau (elevado) de **suavidade** (*smoothness*).
- Qualquer função que gere uma matriz de covariância **semi-definida positiva** (i.e. $\mathbf{a}^\top \mathbf{K} \mathbf{a} \geq 0, \forall \mathbf{a}$) é válida.

A função de kernel/covariância

- **Novas funções de kernel** podem ser criadas a partir da soma e/ou produto de kernels, possivelmente escalados, **aumentando a expressividade** do modelo:

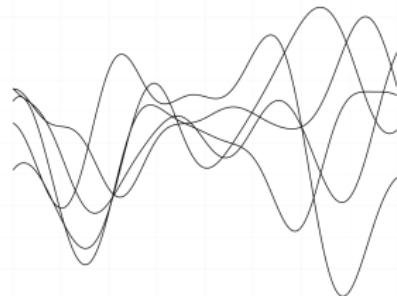
$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_m A_m k_m(\mathbf{x}_i, \mathbf{x}_j) + \prod_n B_n k_n(\mathbf{x}_i, \mathbf{x}_j),$$

em que A_m e B_n são constantes reais positivas.

- Mapeamentos da entrada via uma função arbitrária não-linear $g(\cdot)$ (inclusive para outras dimensões) também gera novos kernels válidos:

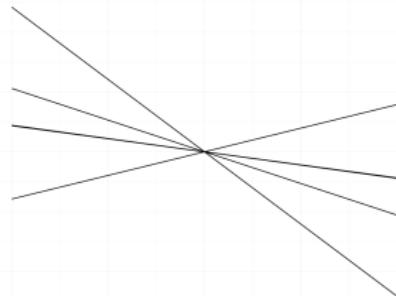
$$k_2(\mathbf{x}_i, \mathbf{x}_j) = k(g(\mathbf{x}_i), g(\mathbf{x}_j)).$$

Amostras de uma GP ($d_{ij} = \|x_i - x_j\|_2$)



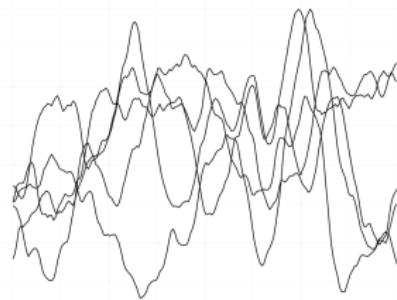
a) Exponencial quadrática

$$k(x_i, x_j) = \sigma_f^2 \exp \left[-\frac{1}{2} w^2 d_{ij}^2 \right]$$



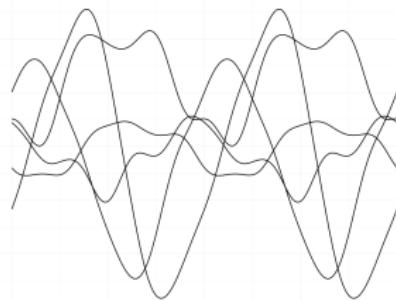
b) Linear

$$k(x_i, x_j) = w^2 x_i x_j$$



c) Matérn 3/2

$$k(x_i, x_j) = \sigma_f^2 (1 + \sqrt{3} w d_{ij}) \exp \left[-\sqrt{3} w d_{ij} \right]$$

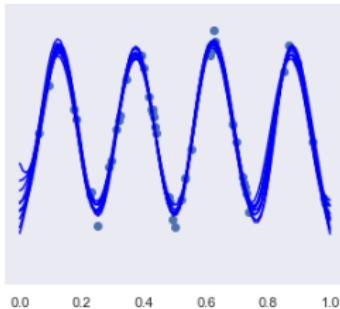


d) Periódica

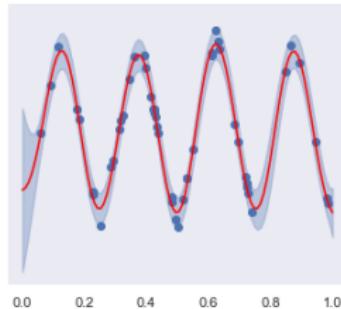
$$k(x_i, x_j) = \sigma_f^2 \exp \left[-2 w^2 \sin^2 \left(\frac{d_{ij}}{2} \right) \right]$$

GP para regressão com kernel Matérn 3/2

Posterior predictive samples after 50 observations
After hyperparameter optimization

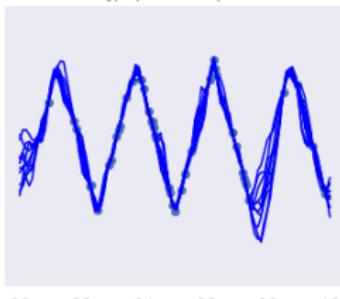


Posterior predictions after 50 observations
After hyperparameter optimization
Evidence = 2.31e+01

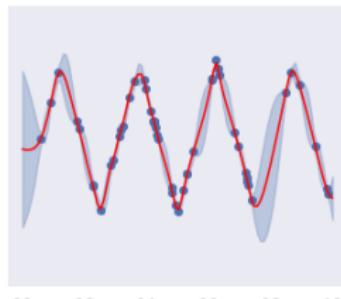


GP com kernel RBF.

Posterior predictive samples after 50 observations
After hyperparameter optimization



Posterior predictions after 50 observations
After hyperparameter optimization
Evidence = 3.12e+01



GP com kernel Matérn 3/2.

De espaços de atributos para GPs

- Seja $\phi : \mathbb{R}^D \rightarrow \mathbb{R}^Q$ uma função de mapeamento e $\mathbf{w}_i \in \mathbb{R}^Q$ um vetor de parâmetros:

$$y_i = \mathbf{w}^\top \phi(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_y^2).$$

De espaços de atributos para GPs

- Seja $\phi : \mathbb{R}^D \rightarrow \mathbb{R}^Q$ uma função de mapeamento e $\mathbf{w}_i \in \mathbb{R}^Q$ um vetor de parâmetros:

$$y_i = \mathbf{w}^\top \phi(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_y^2).$$

- Se considerarmos a priori $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \Sigma_w)$ e a matriz $\Phi = \phi(\mathbf{X})$ em que cada linha é dada por $\phi(\mathbf{x}_i)$, a posteriori será

$$p(\mathbf{w} | \mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} | \mathbf{w}, \mathbf{X}) p(\mathbf{w})}{p(\mathbf{y} | \mathbf{X})} = \mathcal{N}\left(\mathbf{w} \middle| \frac{1}{\sigma_y^2} \mathbf{A}^{-1} \Phi \mathbf{y}, \mathbf{A}^{-1}\right),$$

em que $\mathbf{A} \in \mathbb{R}^{Q \times Q} = \frac{1}{\sigma_y^2} \Phi \Phi^\top + \Sigma_w^{-1}$.

De espaços de atributos para GPs

- Seja $\phi : \mathbb{R}^D \rightarrow \mathbb{R}^Q$ uma função de mapeamento e $\mathbf{w}_i \in \mathbb{R}^Q$ um vetor de parâmetros:

$$y_i = \mathbf{w}^\top \phi(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_y^2).$$

- Se considerarmos a priori $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \Sigma_w)$ e a matriz $\Phi = \phi(\mathbf{X})$ em que cada linha é dada por $\phi(\mathbf{x}_i)$, a posteriori será

$$p(\mathbf{w} | \mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} | \mathbf{w}, \mathbf{X}) p(\mathbf{w})}{p(\mathbf{y} | \mathbf{X})} = \mathcal{N}\left(\mathbf{w} \middle| \frac{1}{\sigma_y^2} \mathbf{A}^{-1} \Phi \mathbf{y}, \mathbf{A}^{-1}\right),$$

em que $\mathbf{A} \in \mathbb{R}^{Q \times Q} = \frac{1}{\sigma_y^2} \Phi \Phi^\top + \Sigma_w^{-1}$.

- Predições são feitas marginalizando os parâmetros:

$$\begin{aligned} p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \int p(f_* | \mathbf{x}_*, \mathbf{w}) p(\mathbf{w} | \mathbf{y}, \mathbf{X}) d\mathbf{w} \\ &= \mathcal{N}\left(f_* \middle| \frac{1}{\sigma_y^2} \phi(\mathbf{x}_*)^\top \mathbf{A}^{-1} \Phi \mathbf{y}, \phi(\mathbf{x}_*)^\top \mathbf{A}^{-1} \phi(\mathbf{x}_*)\right). \end{aligned}$$

De espaços de atributos para GPs

- A distribuição preditiva pode ser reescrita:

$$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N} \left(f_* \middle| \phi(\mathbf{x}_*)^\top \Sigma_w \Phi (\Phi^T \Sigma_w \Phi + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y}, \right.$$
$$\left. \phi(\mathbf{x}_*)^\top \Sigma_w \phi(\mathbf{x}_*) - \phi(\mathbf{x}_*)^\top \Sigma_w \Phi (\Phi^T \Sigma_w \Phi + \sigma_y^2 \mathbf{I})^{-1} \Phi^T \Sigma_w \phi(\mathbf{x}_*) \right).$$

De espaços de atributos para GPs

- A distribuição preditiva pode ser reescrita:

$$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(f_* | \phi(\mathbf{x}_*)^\top \Sigma_w \Phi (\Phi^T \Sigma_w \Phi + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y,}$$
$$\phi(\mathbf{x}_*)^\top \Sigma_w \phi(\mathbf{x}_*) - \phi(\mathbf{x}_*)^\top \Sigma_w \Phi (\Phi^T \Sigma_w \Phi + \sigma_y^2 \mathbf{I})^{-1} \Phi^T \Sigma_w \phi(\mathbf{x}_*)) .$$

- O truque do kernel (*kernel trick*) pode então ser aplicado:

$$\Phi^\top \Sigma_w \Phi = \Phi^T \Sigma_w^{1/2} \Sigma_w^{1/2} \Phi = \Psi^\top \Psi = k(\mathbf{X}, \mathbf{X}) = \mathbf{K},$$

$$\phi(\mathbf{x}_*)^\top \Sigma_w \Phi = k(\mathbf{x}_*, \mathbf{X}) = \mathbf{k}_{*f},$$

$$\Phi^\top \Sigma_w \phi(\mathbf{x}_*) = k(\mathbf{X}, \mathbf{x}_*) = \mathbf{k}_{f*},$$

$$\phi(\mathbf{x}_*)^\top \Sigma_w \phi(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) = k_{**}.$$

De espaços de atributos para GPs

- A distribuição preditiva pode ser reescrita:

$$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(f_* | \phi(\mathbf{x}_*)^\top \Sigma_w \Phi (\Phi^T \Sigma_w \Phi + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y,}$$
$$\phi(\mathbf{x}_*)^\top \Sigma_w \phi(\mathbf{x}_*) - \phi(\mathbf{x}_*)^\top \Sigma_w \Phi (\Phi^T \Sigma_w \Phi + \sigma_y^2 \mathbf{I})^{-1} \Phi^T \Sigma_w \phi(\mathbf{x}_*)) .$$

- O truque do kernel (*kernel trick*) pode então ser aplicado:

$$\Phi^\top \Sigma_w \Phi = \Phi^T \Sigma_w^{1/2} \Sigma_w^{1/2} \Phi = \Psi^\top \Psi = k(\mathbf{X}, \mathbf{X}) = \mathbf{K},$$

$$\phi(\mathbf{x}_*)^\top \Sigma_w \Phi = k(\mathbf{x}_*, \mathbf{X}) = \mathbf{k}_{*f},$$

$$\Phi^\top \Sigma_w \phi(\mathbf{x}_*) = k(\mathbf{X}, \mathbf{x}_*) = \mathbf{k}_{f*},$$

$$\phi(\mathbf{x}_*)^\top \Sigma_w \phi(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) = k_{**}.$$

- Finalmente, a expressão preditiva padrão de GP é obtida:

$$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{k}_{*f}(\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y}, k_{**} - \mathbf{k}_{*f}(\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{k}_{f*}).$$

Agenda

① Processos Gaussianos

GPs para Regressão

De espaços de atributos para GPs

② GPs para grandes conjuntos de dados

GP esparso variacional

GP com inferência variacional estocástica

③ Otimização Bayesiana

④ Outros tópicos

Classificação

Aprendizagem não-supervisionada e redução de dimensionalidade

Modelagem hierárquica

Modelagem dinâmica

Robótica e controle

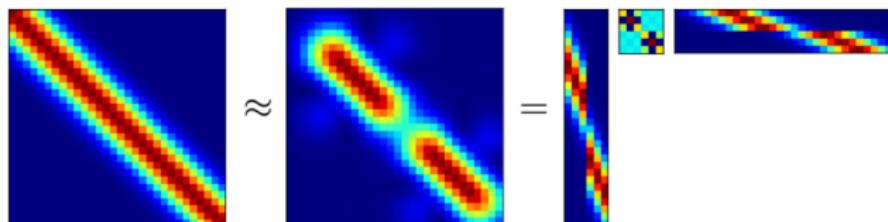
Aprendizagem robusta

⑤ Conclusão

⑥ Referências

GPs para grandes conjuntos de dados

- **Aproximações esparsas** substituem uma matriz de covariância completa K_N por $K_{NM}K_M^{-1}K_{MN}$, em que $M < N$.



Visualização de aproximações esparsas.

- **Abordagens estocásticas** removem a dependência de N no custo computacional usando otimização em mini-batches (B amostras), permitindo lidar com **dados massivos**.
- **Escalabilidade**: GP completo \rightarrow GP esparso \rightarrow GP estocástico
 - Demanda de processamento:
 $\mathcal{O}(N^3) \rightarrow \mathcal{O}(M^2N) \rightarrow \mathcal{O}(M^3 + M^2B)$.
 - Demanda de armazenamento:
 $\mathcal{O}(N^2) \rightarrow \mathcal{O}(MN) \rightarrow \mathcal{O}(M^2 + MB)$.

GP esparso variacional

- Titsias (2009) propôs uma popular abordagem esparsa variacional para mitigar o alto custo computacional de GPs.
- Começamos aumentando o modelo de GP com M **pontos de indução** $z \in \mathbb{R}^M$ da mesma priori de GP do vetor f .
- Associados a z , temos M **entradas de indução**, ou **pseudo-entradas**, $\zeta_j|_{j=1}^M \in \mathbb{R}^D$, no domínio das entradas $x_i|_{i=1}^N$.

GP esparso variacional

- As distribuições envolvidas no modelo são dadas por:

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 \mathbf{I}),$$

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_f),$$

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{K}_z),$$

$$p(\mathbf{f}, \mathbf{z}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{z} \end{bmatrix} \mid \mathbf{0}, \begin{bmatrix} \mathbf{K}_f & \mathbf{K}_{fz} \\ \mathbf{K}_{ fz}^\top & \mathbf{K}_z \end{bmatrix}\right),$$

em que $\mathbf{K}_z \in \mathbb{R}^{M \times M}$ e $\mathbf{K}_{fz} \in \mathbb{R}^{N \times M}$ são dados por
 $[\mathbf{K}_z]_{jj'} = k(\boldsymbol{\zeta}_j, \boldsymbol{\zeta}_{j'})$ e $[\mathbf{K}_{fz}]_{ij} = k(\mathbf{x}_i, \boldsymbol{\zeta}_j)$.

GP esparso variacional

- As distribuições envolvidas no modelo são dadas por:

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 \mathbf{I}),$$

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_f),$$

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{K}_z),$$

$$p(\mathbf{f}, \mathbf{z}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{z} \end{bmatrix} \mid \mathbf{0}, \begin{bmatrix} \mathbf{K}_f & \mathbf{K}_{fz} \\ \mathbf{K}_{ fz}^\top & \mathbf{K}_z \end{bmatrix}\right),$$

em que $\mathbf{K}_z \in \mathbb{R}^{M \times M}$ e $\mathbf{K}_{fz} \in \mathbb{R}^{N \times M}$ são dados por
 $[\mathbf{K}_z]_{jj'} = k(\boldsymbol{\zeta}_j, \boldsymbol{\zeta}_{j'})$ e $[\mathbf{K}_{fz}]_{ij} = k(\mathbf{x}_i, \boldsymbol{\zeta}_j)$.

- A verossimilhança marginal poder ser escrita por:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{z}, \mathbf{X})p(\mathbf{z})d\mathbf{f}d\mathbf{z}.$$

- Note que o modelo em si não foi alterado, pois \mathbf{z} pode ser marginalizado para recuperarmos o modelo original.

GP esparso variacional

- Como \mathbf{f} e \mathbf{z} possuem distribuições Gaussianas, a distribuição $p(\mathbf{f}|\mathbf{z}, \mathbf{X})$ é analítica, pela propriedade do condicionamento:

$$p(\mathbf{f}|\mathbf{z}, \mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{a}_f, \Sigma_f),$$

$$\mathbf{a}_f = \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{z},$$

$$\Sigma_f = \mathbf{K}_f - \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top.$$

- Note que acima só precisamos inverter a matriz \mathbf{K}_z .
- Como escolhemos $M < N$, as operações mais custosas tornam-se mais escaláveis.

GP esparso variacional

- Introduzimos uma distribuição variacional Q e multiplicamos $\frac{Q}{Q}$ no lado direito da verossimilhança marginal:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{z}, \mathbf{X})p(\mathbf{z})d\mathbf{f}d\mathbf{z},$$
$$p(\mathbf{y}|\mathbf{X}) = \int Q \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{z}, \mathbf{X})p(\mathbf{z})}{Q} d\mathbf{f}d\mathbf{z}.$$

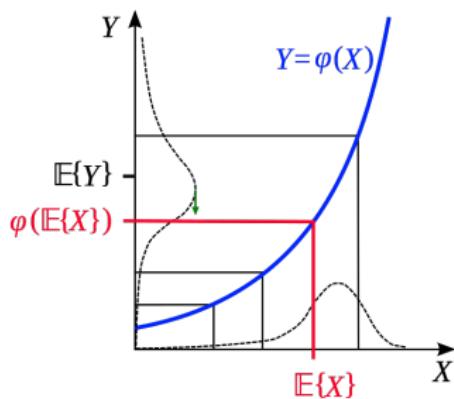
GP esparso variacional

- Introduzimos uma distribuição variacional Q e multiplicamos $\frac{Q}{Q}$ no lado direito da verossimilhança marginal:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{z}, \mathbf{X})p(\mathbf{z})d\mathbf{f}d\mathbf{z},$$
$$p(\mathbf{y}|\mathbf{X}) = \int Q \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{z}, \mathbf{X})p(\mathbf{z})}{Q} d\mathbf{f}d\mathbf{z}.$$

- Seja a desigualdade de Jensen para um função convexa $\phi(\cdot)$:

$$\phi(\mathbb{E}\{x\}) \leq \mathbb{E}\{\phi(x)\}.$$



GP esparso variacional

- Aplicamos $\log(\cdot)$ (uma função côncava!) em ambos os lados, e usamos a desigualdade de Jensen:

$$\log p(\mathbf{y}|\mathbf{X}) \geq \int Q \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{z}, \mathbf{X})p(\mathbf{z})}{Q} d\mathbf{f} d\mathbf{z}.$$

GP esparso variacional

- Aplicamos $\log(\cdot)$ (uma função côncava!) em ambos os lados, e usamos a desigualdade de Jensen:

$$\log p(\mathbf{y}|\mathbf{X}) \geq \int Q \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{z}, \mathbf{X})p(\mathbf{z})}{Q} d\mathbf{f} d\mathbf{z}.$$

- A desigualdade é válida para qualquer Q , mas escolhemos a forma conveniente abaixo:

$$Q = q(\mathbf{z})p(\mathbf{f}|\mathbf{z}, \mathbf{X}).$$

- Substituindo na desigualdade, obtemos:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &\geq \int q(\mathbf{z})p(\mathbf{f}|\mathbf{z}, \mathbf{X}) \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{f} d\mathbf{z}, \\ &\geq \int q(\mathbf{z}) \left[\underbrace{\int p(\mathbf{f}|\mathbf{z}, \mathbf{X}) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{f}}_{\mathcal{L}_1} + \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \right] d\mathbf{z}.\end{aligned}$$

GP esparso variacional

- Inicialmente resolvemos a integral em \mathbf{f} :

$$\begin{aligned}\mathcal{L}_1 &= \int p(\mathbf{f}|\mathbf{z}, \mathbf{X}) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{f} \\&= \int \mathcal{N}(\mathbf{f}|\mathbf{a}_f, \Sigma_f) \log \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 \mathbf{I}) d\mathbf{f} \\&= -\frac{N}{2} \log 2\pi\sigma_y^2 - \frac{1}{2\sigma_y^2} \left(\mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{a}_f + (\mathbf{a}_f)^\top \mathbf{a}_f + \text{Tr}(\Sigma_f) \right) \\&= \log \mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f).\end{aligned}$$

GP esparso variacional

- Substituímos o resultado de \mathcal{L}_1 no limiar original:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &\geq \int q(\mathbf{z}) \left[\underbrace{\log \mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f) + \log \frac{p(\mathbf{z})}{q(\mathbf{z})}}_{\mathcal{L}_1} \right] d\mathbf{z} \\ &\geq \underbrace{\int q(\mathbf{z}) \log \frac{\mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z}}_{\mathcal{L}_2} - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f).\end{aligned}$$

GP esparso variacional

- Substituímos o resultado de \mathcal{L}_1 no limiar original:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &\geq \int q(\mathbf{z}) \left[\underbrace{\log \mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f)}_{\mathcal{L}_1} + \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \right] d\mathbf{z} \\ &\geq \underbrace{\int q(\mathbf{z}) \log \frac{\mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z}}_{\mathcal{L}_2} - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f).\end{aligned}$$

- Comparamos a última integral com a desigualdade de Jensen:

$$\log(\mathbb{E}\{x\}) \geq \mathbb{E}\{\log(x)\}.$$

GP esparso variacional

- Substituímos o resultado de \mathcal{L}_1 no limiar original:

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &\geq \int q(\mathbf{z}) \left[\underbrace{\log \mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f)}_{\mathcal{L}_1} + \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \right] d\mathbf{z} \\ &\geq \underbrace{\int q(\mathbf{z}) \log \frac{\mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z}}_{\mathcal{L}_2} - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f).\end{aligned}$$

- Comparamos a última integral com a desigualdade de Jensen:

$$\log(\mathbb{E}\{x\}) \geq \mathbb{E}\{\log(x)\}.$$

- O limiar é maximizado movendo o $\log(\cdot)$ para fora da integral:

$$\log \int \mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) p(\mathbf{z}) d\mathbf{z} \geq \mathcal{L}_2.$$

GP esparso variacional

- Substituindo no limiar, temos:

$$\log p(\mathbf{y}|\mathbf{X}) \geq \log \int \mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{K}_z) d\mathbf{z} - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f).$$

- Lembramos os valores anteriormente definidos:

$$\mathbf{a}_f = \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{z},$$

$$\Sigma_f = \mathbf{K}_f - \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top.$$

GP esparso variacional

- Substituindo no limiar, temos:

$$\log p(\mathbf{y}|\mathbf{X}) \geq \log \int \mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{K}_z) d\mathbf{z} - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f).$$

- Lembramos os valores anteriormente definidos:

$$\begin{aligned}\mathbf{a}_f &= \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{z}, \\ \Sigma_f &= \mathbf{K}_f - \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top.\end{aligned}$$

- Usamos as propriedades da Gaussiana para marginalizar \mathbf{z} :

$$\begin{aligned}p(\mathbf{a}|\mathbf{b}) &= \mathcal{N}(\mathbf{a}|\mathbf{Ab} + \mathbf{m}, \Sigma), \quad p(\mathbf{b}) = \mathcal{N}(\mathbf{b}|\boldsymbol{\mu}_b, \Sigma_b), \\ p(\mathbf{a}) &= \int p(\mathbf{a}|\mathbf{b})p(\mathbf{b})d\mathbf{b} = \mathcal{N}(\mathbf{a}|\mathbf{A}\boldsymbol{\mu}_b + \mathbf{m}, \Sigma + \mathbf{A}\Sigma_b\mathbf{A}^\top).\end{aligned}$$

- O ELBO (*evidence lower bound*) final então torna-se:

$$\log p(\mathbf{y}|\mathbf{X}) \geq \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma_y^2 \mathbf{I} + \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top) - \frac{1}{2\sigma_y^2} \text{Tr}(\mathbf{K}_f - \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top).$$

GP esparso variacional

- Os hiperparâmetros do kernel, a variância da verossimilhança e as pseudo-entradas $z_j|_{j=1}^M$ podem ser otimizadas a partir da maximização do ELBO.
- Observações importantes:
 - Para $M = N$ e $z_j = \mathbf{x}_j, \forall j$, recuperamos a evidência exata.
 - As pseudo-entradas $z_j|_{j=1}^M$ não são parâmetros do modelo, mas parâmetros variacionais, ou seja, do procedimento de inferência.
 - Valores maiores de M não resultam em overfitting, somente melhoram a aproximação feita (mas podem dificultar a otimização).
 - Custo computacional $\mathcal{O}(NM^2)$ e de armazenamento $\mathcal{O}(NM)$.

GP esparso variacional

- A distribuição ótima $q^*(\mathbf{z})$ é obtida voltando ao limiar original:

$$\log p(\mathbf{y}|\mathbf{X}) \geq \int q(\mathbf{z}) \log \frac{\mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f).$$

- Note que o termo do lado esquerdo parece (o negativo de) uma divergência KL:

$$-\text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X})) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} d\mathbf{Z}.$$

GP esparso variacional

- A distribuição ótima $q^*(\mathbf{z})$ é obtida voltando ao limiar original:

$$\log p(\mathbf{y}|\mathbf{X}) \geq \int q(\mathbf{z}) \log \frac{\mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} - \frac{1}{2\sigma_y^2} \text{Tr}(\Sigma_f).$$

- Note que o termo do lado esquerdo parece (o negativo de) uma divergência KL:

$$-\text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X})) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} d\mathbf{Z}.$$

- O KL é sempre não negativo e seu valor mínimo (zero) ocorre quando ambas as distribuições do KL são iguais.
- Assim, a integral será máxima para:

$$q^*(\mathbf{z}) \propto \mathcal{N}(\mathbf{y}|\mathbf{a}_f, \sigma_y^2 \mathbf{I}) p(\mathbf{z}).$$

GP esparso variacional

- Substituímos os termos conhecidos:

$$q^*(\mathbf{z}) \propto \mathcal{N}(\mathbf{y} | \mathbf{a}_f, \sigma_y^2 \mathbf{I}) p(\mathbf{z}),$$

$$\propto \mathcal{N}(\mathbf{y} | \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{z}, \sigma_y^2 \mathbf{I}) \mathcal{N}(\mathbf{z} | \mathbf{0}, \mathbf{K}_z),$$

$$\log q^*(\mathbf{z}) \propto \log \mathcal{N}(\mathbf{y} | \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{z}, \sigma_y^2 \mathbf{I}) + \log \mathcal{N}(\mathbf{z} | \mathbf{0}, \mathbf{K}_z)$$

$$\propto -\frac{1}{\sigma_y^2} (-\mathbf{y}^\top \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{z} + \frac{1}{2} \mathbf{z}^\top \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{z}) - \frac{1}{2} \mathbf{z}^\top \mathbf{K}_z^{-1} \mathbf{z}$$

$$\propto \frac{1}{\sigma_y^2} \mathbf{y}^\top \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{z} - \frac{1}{2} \mathbf{z}^\top (\mathbf{K}_z^{-1} + \frac{1}{\sigma_y^2} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top \mathbf{K}_{fz} \mathbf{K}_z^{-1}) \mathbf{z},$$

$$q^*(\mathbf{z}) = \mathcal{N}(\mathbf{m}, \mathbf{S}),$$

$$\mathbf{S} = (\mathbf{K}_z^{-1} + \frac{1}{\sigma_y^2} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top \mathbf{K}_{fz} \mathbf{K}_z^{-1})^{-1}$$

$$= \mathbf{K}_z (\mathbf{K}_z + \frac{1}{\sigma_y^2} \mathbf{K}_{fz}^\top \mathbf{K}_{fz})^{-1} \mathbf{K}_z,$$

$$\mathbf{m} = \frac{1}{\sigma_y^2} \mathbf{S} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top \mathbf{y}$$

$$= \frac{1}{\sigma_y^2} \mathbf{K}_z (\mathbf{K}_z + \frac{1}{\sigma_y^2} \mathbf{K}_{fz}^\top \mathbf{K}_{fz})^{-1} \mathbf{K}_{fz}^\top \mathbf{y}.$$

GP esparso variacional

- Predições podem ser feitas marginalizando \mathbf{z} usando a distribuição ótima $q^*(\mathbf{z})$:

$$\begin{aligned} p(f_* | \mathbf{x}_*) &\approx \int p(f_* | \mathbf{z}, \mathbf{x}_*) q^*(\mathbf{z}) d\mathbf{z} \\ &= \int \mathcal{N}(f_* | \mathbf{k}_{*z} \mathbf{K}_z^{-1} \mathbf{z}, K_* - \mathbf{k}_{*z} \mathbf{K}_z^{-1} \mathbf{k}_{*z}^\top) \mathcal{N}(\mathbf{z} | \mathbf{m}, \mathbf{S}) d\mathbf{z} \\ &= \mathcal{N}(f_* | \mu_*, \sigma_*^2), \\ \mu_* &= \mathbf{k}_{*z} \mathbf{K}_z^{-1} \mathbf{m}, \\ \sigma_*^2 &= K_* - \mathbf{k}_{*z} \mathbf{K}_z^{-1} \mathbf{k}_{*z}^\top + \mathbf{k}_{*z} \mathbf{K}_z^{-1} \mathbf{S} \mathbf{K}_z^{-1} \mathbf{k}_{*z}^\top, \end{aligned}$$

em que:

$$\mathbf{m} = \frac{1}{\sigma_y^2} \mathbf{K}_z (\mathbf{K}_z + \frac{1}{\sigma_y^2} \mathbf{K}_{fz}^\top \mathbf{K}_{fz})^{-1} \mathbf{K}_{fz}^\top \mathbf{y},$$

$$\mathbf{S} = \mathbf{K}_z (\mathbf{K}_z + \frac{1}{\sigma_y^2} \mathbf{K}_{fz}^\top \mathbf{K}_{fz})^{-1} \mathbf{K}_z.$$

GP esparso variacional

Resumo do algoritmo

- Passo de estimação (seleção de modelos)

① Inicialize $\boldsymbol{\theta} = \left[\sigma_f^2, w_1^2, \dots, w_D^2, \sigma_y^2 \right]^\top$ e $\mathbf{z}_j|_{j=1}^M \in \mathbb{R}^D$;

→ Exemplo: $\mathbf{z}_j|_{j=1}^M \leftarrow \text{Kmeans}(\mathbf{X})$,

$$\sigma_f^2 = \mathbb{V}[\mathbf{y}], \quad w_d^2 = \frac{1}{\mathbb{V}[\mathbf{X}_{:d}]}, \quad \sigma_y^2 = 0.01\sigma_f^2.$$

② Repita até convergir ou por um número máximo de iterações:

① Calcule o ELBO:

$$\mathcal{L} = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma_y^2 \mathbf{I} + \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top) - \frac{1}{2\sigma_y^2} \text{Tr}(\mathbf{K}_f - \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top);$$

② Calcule os gradientes analíticos $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$ e $\frac{\partial \mathcal{L}}{\partial \mathbf{z}_j}|_{j=1}^M$;

③ Atualize $\boldsymbol{\theta}$ e $\mathbf{z}_j|_{j=1}^M$ a partir dos gradientes (e.g. via BFGS);

④ Calcule a distribuição variacional ótima:

$$q^*(\mathbf{z}) = \mathcal{N}(\mathbf{m}, \mathbf{S}),$$

$$\mathbf{m} = \frac{1}{\hat{\sigma}_y^2} \mathbf{K}_z (\mathbf{K}_z + \frac{1}{\hat{\sigma}_y^2} \mathbf{K}_{fz}^\top \mathbf{K}_{fz})^{-1} \mathbf{K}_{fz}^\top \mathbf{y}, \quad \mathbf{S} = \mathbf{K}_z (\mathbf{K}_z + \frac{1}{\hat{\sigma}_y^2} \mathbf{K}_{fz}^\top \mathbf{K}_{fz})^{-1} \mathbf{K}_z.$$

⑤ Retorne os valores otimizados $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{z}}_j|_{j=1}^M$ e a distribuição $q^*(\mathbf{z})$.

GP esparso variacional

Resumo do algoritmo

- Passo de predição

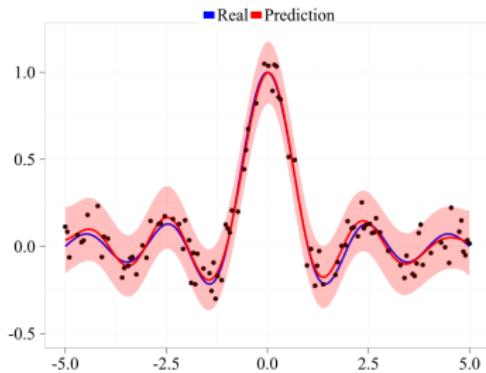
- ➊ Dado um novo padrão \boldsymbol{x}_* , retorne a distribuição preditiva

$$p(y_* | \boldsymbol{x}_*) = \mathcal{N} \left(f_* | \mu_*, \sigma_*^2 + \sigma_y^2 \right),$$

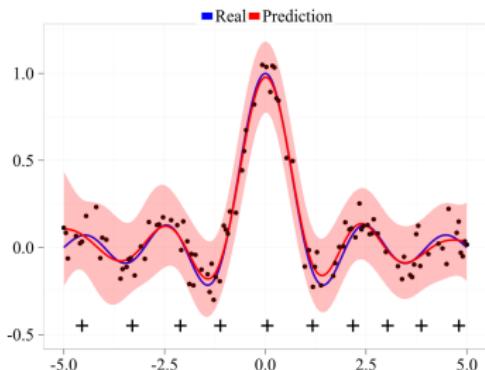
$$\mu_* = \mathbf{k}_{*z} \mathbf{K}_z^{-1} \mathbf{m},$$

$$\sigma_*^2 = K_* - \mathbf{k}_{*z} \mathbf{K}_z^{-1} \mathbf{k}_{*z}^\top + \mathbf{k}_{*z} \mathbf{K}_z^{-1} S \mathbf{K}_z^{-1} \mathbf{k}_{*z}^\top.$$

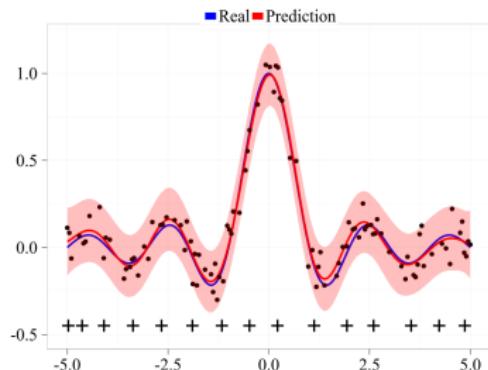
GP esparso variacional



(a) GP padrão.



(b) GP esparso variacional ($M = 10$).



(c) GP esparso variacional ($M = 20$).

Inferência variacional estocástica

- O framework geral de *stochastic variational inference* (SVI) foi proposto por Hoffman *et al.* (2013) para escalar métodos variacionais.
- A ideia principal consiste em usar técnicas de aprendizagem estocástica na otimização do ELBO.
- Em resumo, o SVI considera os seguintes passos:
 - ① Amostre uma certa quantidade de exemplos (um minibatch) do conjunto de treinamento;
 - ② Otimize os parâmetros variacionais locais (caso eles existam);
 - ③ Forme os parâmetros variacionais globais intermediários;
 - ④ Atualize estocasticamente os parâmetros variacionais globais.

Inferência variacional estocástica

- O framework geral de *stochastic variational inference* (SVI) foi proposto por Hoffman *et al.* (2013) para escalar métodos variacionais.
- A ideia principal consiste em usar técnicas de aprendizagem estocástica na otimização do ELBO.
- Em resumo, o SVI considera os seguintes passos:
 - ① Amostre uma certa quantidade de exemplos (um minibatch) do conjunto de treinamento;
 - ② Otimize os parâmetros variacionais locais (caso eles existam);
 - ③ Forme os parâmetros variacionais globais intermediários;
 - ④ Atualize estocasticamente os parâmetros variacionais globais.
- A partir dos pontos acima, notamos que precisamos de:
 - um conjunto de parâmetros globais;
 - um objetivo variacional fatorado para diferentes observações e (se existirem) os parâmetros variacionais locais correspondentes.

SVI para GPs

- Hensman *et al.* (2013) adaptou o framework SVI para permitir o uso de modelos de GP em dados massivos.
- Note que modelos de GP usuais não possuem os requisitos para a aplicação do SVI.
- A versão esparsa variacional de Titsias possui, com os pontos de indução $\mathbf{z} \in \mathbb{R}^M$ agindo como parâmetros variacionais globais.
- No entanto, o ELBO que derivamos antes não pode ser fatorado:

$$\mathcal{L} = \log \mathcal{N}(\mathbf{y} | \mathbf{0}, \sigma_y^2 \mathbf{I} + \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top) - \frac{1}{2\sigma_y^2} \text{Tr}(\mathbf{K}_f - \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top).$$

SVI para GPs

- Hensman *et al.* reescreveram o ELBO do GP esparso sem a remoção ótima da distribuição variacional $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{S})$:

$$\begin{aligned}\mathcal{L} &= \int q(\mathbf{z}) \left[\log \mathcal{N}(\mathbf{y} | \mathbf{a}_f, \sigma_y^2 \mathbf{I}) - \frac{1}{2\sigma_y^2} \text{Tr}(\mathbf{\Sigma}_f) + \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \right] d\mathbf{z} \\ &= \mathbb{E}_{q(\mathbf{z})} [\log \mathcal{N}(\mathbf{y} | \mathbf{a}_f, \sigma_y^2 \mathbf{I})] - \frac{1}{2\sigma_y^2} \text{Tr}(\mathbf{\Sigma}_f) + \mathbb{E}_{q(\mathbf{z})} \left[\log \frac{p(\mathbf{z})}{q(\mathbf{z})} \right] \\ &= \sum_{i=1}^N \left[\mathbb{E}_{q(\mathbf{z})} [\log \mathcal{N}(y_i | [\mathbf{a}_f]_i, \sigma_y^2)] - \frac{1}{2\sigma_y^2} [\mathbf{\Sigma}_f]_{ii} \right] - \text{KL}(q(\mathbf{z}) | p(\mathbf{z})).\end{aligned}$$

- Lembramos os valores anteriormente definidos:

$$\begin{aligned}\mathbf{a}_f &= \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{z}, \\ \mathbf{\Sigma}_f &= \mathbf{K}_f - \mathbf{K}_{fz} \mathbf{K}_z^{-1} \mathbf{K}_{fz}^\top.\end{aligned}$$

SVI para GPs

- Realizamos as substituições:

$$\mathcal{L} = \sum_{i=1}^N \left[\underbrace{\mathbb{E}_{q(\mathbf{z})}[\log \mathcal{N}(y_i | \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{z}, \sigma_y^2)]}_{\mathcal{L}_1} - \frac{1}{2\sigma_y^2} (K_{ii} - \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{k}_i) \right] \\ - \text{KL}(q(\mathbf{z}) | p(\mathbf{z})).$$

em que \mathbf{k}_i é a i -ésima coluna de \mathbf{K}_{fz}^\top .

SVI para GPs

- Realizamos as substituições:

$$\mathcal{L} = \sum_{i=1}^N \left[\underbrace{\mathbb{E}_{q(\mathbf{z})}[\log \mathcal{N}(y_i | \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{z}, \sigma_y^2)]}_{\mathcal{L}_1} - \frac{1}{2\sigma_y^2} (K_{ii} - \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{k}_i) \right. \\ \left. - \text{KL}(q(\mathbf{z}) | p(\mathbf{z})). \right]$$

em que \mathbf{k}_i é a i -ésima coluna de \mathbf{K}_{fz}^\top .

- Continuamos a desenvolver o termo \mathcal{L}_1 :

$$\begin{aligned} \mathcal{L}_1 &= \mathbb{E}_{q(\mathbf{z})}[\log \mathcal{N}(y_i | \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{z}, \sigma_y^2)] \\ &= -\frac{1}{2} \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} \mathbb{E}_{q(\mathbf{z})}[(y_i^2 - 2y_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{z} + \mathbf{z}^\top \mathbf{K}_z^{-1} \mathbf{k}_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{z})] \\ &= -\frac{1}{2} \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} (y_i^2 - 2y_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbb{E}_{q(\mathbf{z})}[\mathbf{z}]) \\ &\quad - \frac{1}{2\sigma_y^2} \text{Tr}(\mathbf{K}_z^{-1} \mathbf{k}_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbb{E}_{q(\mathbf{z})}[\mathbf{z}^\top \mathbf{z}]) \end{aligned}$$

SVI para GPs

- Continuamos organizando os termos:

$$\mathcal{L}_1 = -\frac{1}{2} \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} (y_i^2 - 2y_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{m})$$

$$-\frac{1}{2\sigma_y^2} \text{Tr}(\mathbf{K}_z^{-1} \mathbf{k}_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} (\mathbf{S} + \mathbf{m} \mathbf{m}^\top))$$

$$\mathcal{L}_1 = -\frac{1}{2} \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} (y_i^2 - 2y_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{m} + \text{Tr}(\mathbf{K}_z^{-1} \mathbf{k}_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{m} \mathbf{m}^\top))$$

$$-\frac{1}{2\sigma_y^2} \text{Tr}(\mathbf{K}_z^{-1} \mathbf{k}_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{S})$$

$$\mathcal{L}_1 = \log \mathcal{N}(y_i | \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{m}, \sigma_y^2) - \frac{1}{2\sigma_y^2} \text{Tr}(\mathbf{K}_z^{-1} \mathbf{k}_i \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{S})$$

$$= \log \mathcal{N}(y_i | \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{m}, \sigma_y^2) - \text{Tr}(\boldsymbol{\Lambda}_i \mathbf{S}),$$

em que $\boldsymbol{\Lambda}_i = \frac{1}{\sigma_y^2} \mathbf{K}_z^{-1} \mathbf{k}_i \mathbf{k}_i^\top \mathbf{K}_z^{-1}$.

SVI para GPs

- Voltando à forma completa do ELBO e considerando minibatches com B amostras, temos:

$$\mathcal{L} = \frac{N}{B} \sum_{i=1}^B \left[\log \mathcal{N}(y_i | \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{m}, \sigma_y^2) - \frac{1}{2\sigma_y^2} (\mathbf{K}_{ii} - \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{k}_i) \right. \\ \left. - \frac{1}{2} \text{Tr}(\mathbf{S} \boldsymbol{\Lambda}_i) \right] - \text{KL}(q(\mathbf{z}) || p(\mathbf{z})).$$

- O termo KL, por só envolver Gaussianas, é analítico:

$$\text{KL}(q(\mathbf{z}) || p(\mathbf{z})) = \frac{1}{2} [\text{Tr}(\mathbf{K}_z^{-1} \mathbf{S}) + \mathbf{m}^\top \mathbf{K}_z^{-1} \mathbf{m} - D + \log |\mathbf{K}_z| - \log |\mathbf{S}|]$$

SVI para GPs

- Voltando à forma completa do ELBO e considerando minibatches com B amostras, temos:

$$\mathcal{L} = \frac{N}{B} \sum_{i=1}^B \left[\log \mathcal{N}(y_i | \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{m}, \sigma_y^2) - \frac{1}{2\sigma_y^2} (\mathbf{K}_{ii} - \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{k}_i) \right. \\ \left. - \frac{1}{2} \text{Tr}(\mathbf{S} \boldsymbol{\Lambda}_i) \right] - \text{KL}(q(\mathbf{z}) || p(\mathbf{z})).$$

- O termo KL, por só envolver Gaussianas, é analítico:

$$\text{KL}(q(\mathbf{z}) || p(\mathbf{z})) = \frac{1}{2} [\text{Tr}(\mathbf{K}_z^{-1} \mathbf{S}) + \mathbf{m}^\top \mathbf{K}_z^{-1} \mathbf{m} - D + \log |\mathbf{K}_z| - \log |\mathbf{S}|]$$

- O ELBO é fatorado pelos pares de observações (\mathbf{x}_i, y_i) .
- Note que esse ELBO é **não-colapsado**, pois depende explicitamente de uma parametrização da distribuição $q(\mathbf{z})$.
- O custo computacional é $\mathcal{O}(M^3 + BM^2)$ e o de armazenamento é $\mathcal{O}(M^2 + MB)$, independente de N .

SVI para GPs

- Os parâmetros variacionais globais $\mathbf{m} \in \mathbb{R}^D$ e $\mathbf{S} \in \mathbb{R}^{(D \times D)}$ devem ser otimizados com os demais via gradientes estocásticos.
- Como a matriz \mathbf{S} é simétrica, na prática otimizamos sua Cholesky $\mathbf{L} \in \mathbb{R}^{(D \times D)}$, tal que $\mathbf{S} = \mathbf{LL}^\top$, que por ser triangular possui menos parâmetros livres.
- Note que os elementos na diagonal de \mathbf{L} devem ser positivos.
- A predição é feita como no framework esparso variacional.
- O novo ELBO é sempre menor (ou igual, caso a distribuição ótima $q^*(z)$ seja usada) que o ELBO colapsado.

SVI para GPs

Resumo do algoritmo

- Passo de estimação (seleção de modelos)

① Inicialize $\boldsymbol{\theta} = \left[\sigma_f^2, w_1^2, \dots, w_D^2, \sigma_y^2 \right]^\top$, $\mathbf{z}_j|_{j=1}^M \in \mathbb{R}^D$, $\mathbf{m} \in \mathbb{R}^D$, $\mathbf{L} \in \mathbb{R}^{(D \times D)}$;

→ Exemplo: $\mathbf{z}_j|_{j=1}^M \leftarrow \text{Kmeans}(\mathbf{X})$,

$$\sigma_f^2 = \mathbb{V}[\mathbf{y}], \quad w_d^2 = \frac{1}{\mathbb{V}[\mathbf{X}_{:d}]}, \quad \sigma_y^2 = 0.01\sigma_f^2.$$

② Repita até convergir ou por um número máximo de iterações:

① Calcule o ELBO a partir de um minibatch com B exemplos:

$$\begin{aligned} \mathcal{L} = \frac{N}{B} \sum_{i=1}^B & \left[\log \mathcal{N}(y_i | \mathbf{k}_i^\top \mathbf{K}_z^{-1} \mathbf{m}, \sigma_y^2) - \frac{1}{2\sigma_y^2} (\mathbf{k}_{ii}^\top \mathbf{K}_z^{-1} \mathbf{k}_i) \right. \\ & \left. - \frac{1}{2} \text{Tr}(\mathbf{S} \boldsymbol{\Lambda}_i) \right] - \text{KL}(q(\mathbf{z}) || p(\mathbf{z})). \end{aligned}$$

② Calcule os gradientes analíticos $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$, $\frac{\partial \mathcal{L}}{\partial \mathbf{z}_j}|_{j=1}^M$, $\frac{\partial \mathcal{L}}{\partial \mathbf{m}}$, $\frac{\partial \mathcal{L}}{\partial \mathbf{L}}$;

③ Atualize $\boldsymbol{\theta}$, $\mathbf{z}_j|_{j=1}^M$, \mathbf{m} e \mathbf{L} a partir dos gradientes estocásticos;

④ Retorne os valores otimizados $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{z}}_j|_{j=1}^M$, $\hat{\mathbf{m}}$ e $\hat{\mathbf{S}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^\top$.

SVI para GPs

Resumo do algoritmo

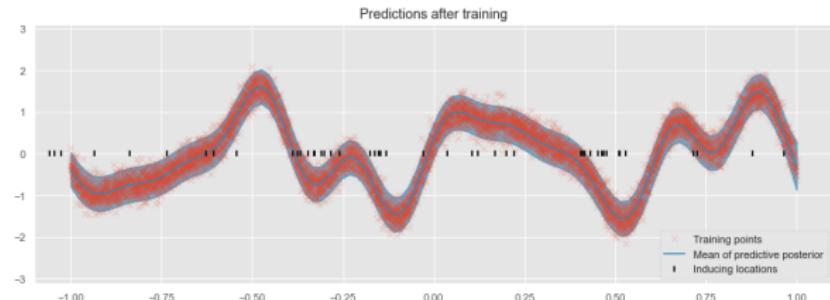
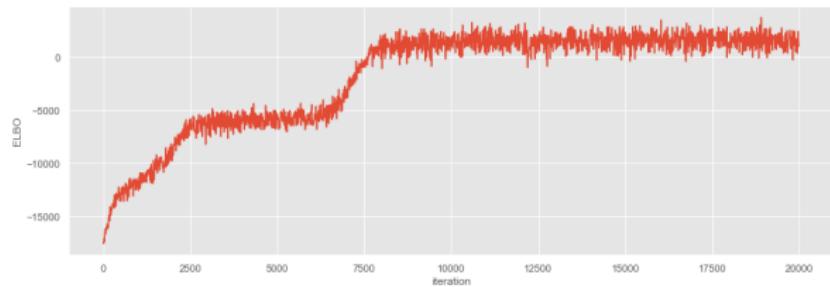
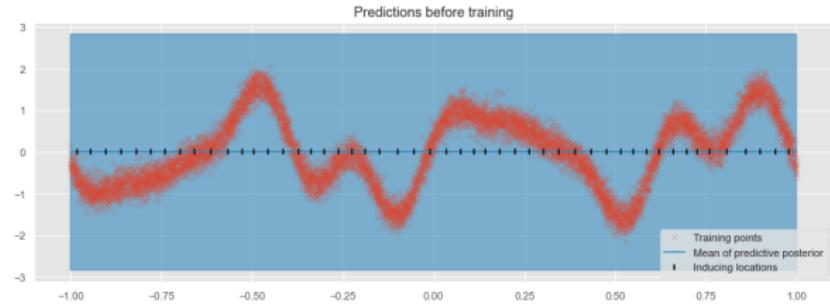
- Passo de predição
 - ➊ Dado um novo padrão \mathbf{x}_* , retorne a distribuição preditiva

$$p(y_* | \mathbf{x}_*) = \mathcal{N}(f_* | \mu_*, \sigma_*^2 + \sigma_y^2),$$

$$\mu_* = \mathbf{k}_{*z} \mathbf{K}_z^{-1} \hat{\mathbf{m}},$$

$$\sigma_*^2 = K_* - \mathbf{k}_{*z} \mathbf{K}_z^{-1} \mathbf{k}_{*z}^\top + \mathbf{k}_{*z} \mathbf{K}_z^{-1} \hat{\mathbf{S}} \mathbf{K}_z^{-1} \mathbf{k}_{*z}^\top.$$

SVI para GPs ($N = 10000, M = 50, B = 100$)



Agenda

① Processos Gaussianos

GPs para Regressão

De espaços de atributos para GPs

② GPs para grandes conjuntos de dados

GP esparso variacional

GP com inferência variacional estocástica

③ Otimização Bayesiana

④ Outros tópicos

Classificação

Aprendizagem não-supervisionada e redução de dimensionalidade

Modelagem hierárquica

Modelagem dinâmica

Robótica e controle

Aprendizagem robusta

⑤ Conclusão

⑥ Referências

Otimização global com GPs

- Considere o problema de otimizar uma função $f(\mathbf{x})$.
- Você desconhece o formato de $f(\cdot)$, mas consegue avaliá-la de maneira ruidosa, observando $y_i = f(\mathbf{x}_i) + \epsilon_i$ para uma entrada \mathbf{x}_i e um ruído ϵ_i .

Otimização global com GPs

- Considere o problema de otimizar uma função $f(\mathbf{x})$.
- Você desconhece o formato de $f(\cdot)$, mas consegue avaliá-la de maneira ruidosa, observando $y_i = f(\mathbf{x}_i) + \epsilon_i$ para uma entrada \mathbf{x}_i e um ruído ϵ_i .
- Exemplos desse cenário:
 - Localizar o melhor ponto para a extração de um minério;
 - Minimizar o uso de recursos em um processo industrial;
 - Configurar um experimento laboratorial;
 - Maximizar a velocidade de acesso em um banco de dados;
 - Selecionar hiperparâmetros para um modelo de aprendizagem de máquina.

Otimização global com GPs

- Considere o problema de otimizar uma função $f(\mathbf{x})$.
- Você desconhece o formato de $f(\cdot)$, mas consegue avaliá-la de maneira ruidosa, observando $y_i = f(\mathbf{x}_i) + \epsilon_i$ para uma entrada \mathbf{x}_i e um ruído ϵ_i .
- Exemplos desse cenário:
 - Localizar o melhor ponto para a extração de um minério;
 - Minimizar o uso de recursos em um processo industrial;
 - Configurar um experimento laboratorial;
 - Maximizar a velocidade de acesso em um banco de dados;
 - Selecionar hiperparâmetros para um modelo de aprendizagem de máquina.
- Sendo a forma de $f(\cdot)$ desconhecida, como vamos minimizá-la ou maximizá-la?

Otimização global com GPs

- Sejam N experimentos que resultam em N pares $(\mathbf{x}_i, y_i)|_{i=1}^N$, em que y_i são observações adicionadas de ruído Gaussiano.
- A função a ser otimizada não possui forma analítica conhecida, então escolhemos uma priori de GP com função de kernel $k(\cdot, \cdot)$:

$$\begin{aligned} p(\mathbf{y}|\mathbf{f}) &= \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_y^2 \mathbf{I}), \\ p(\mathbf{f}|\mathbf{X}) &= \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}), \\ \text{em que } K_{ij} &= k(\mathbf{x}_i, \mathbf{x}_j). \end{aligned}$$

- Como já vimos, esse modelo é analítico e podemos facilmente obter qualquer distribuição de interesse.

Otimização global com GPs

- As predições nesse modelo são distribuições bem definidas:

$$\begin{aligned} p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \mathcal{N}(f_* | \mu_*, \sigma_*^2), \\ \mu_* &= \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y}, \\ \sigma_*^2 &= k_{**} - \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{k}_{f*}. \end{aligned}$$

- A distribuição preditiva informa a incerteza que temos a respeito da função desconhecida no ponto de avaliação \mathbf{x}_* .

Otimização global com GPs

- As previsões nesse modelo são distribuições bem definidas:

$$\begin{aligned} p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \mathcal{N}(f_* | \mu_*, \sigma_*^2), \\ \mu_* &= \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y}, \\ \sigma_*^2 &= k_{**} - \mathbf{k}_{f*}^\top (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{k}_{f*}. \end{aligned}$$

- A distribuição preditiva informa a incerteza que temos a respeito da função desconhecida no ponto de avaliação \mathbf{x}_* .
- Considerando um problema de minimização global, buscamos no domínio \mathcal{X} de entrada:
 - Regiões com menores valores de média predita μ_* (*exploitation*);
 - Regiões com alta incerteza σ_*^2 (*exploration*).
- Note que partimos do princípio que boas soluções estão próximas entre si no espaço de entrada.

Otimização global com GPs

- Como saber que um ponto pertence a uma região de interesse?

Otimização global com GPs

- Como saber que um ponto pertence a uma região de interesse?
- Desejamos usar toda a informação proveniente de avaliações anteriores na escolha do próximo candidato a ser avaliado.
- Usamos uma **função de aquisição**, que **quantifica a relevância** de um ponto candidato.
- A função de aquisição $a : \mathcal{X} \rightarrow \mathbb{R}^+$ é escolhida para ter um formato facilmente calculável.

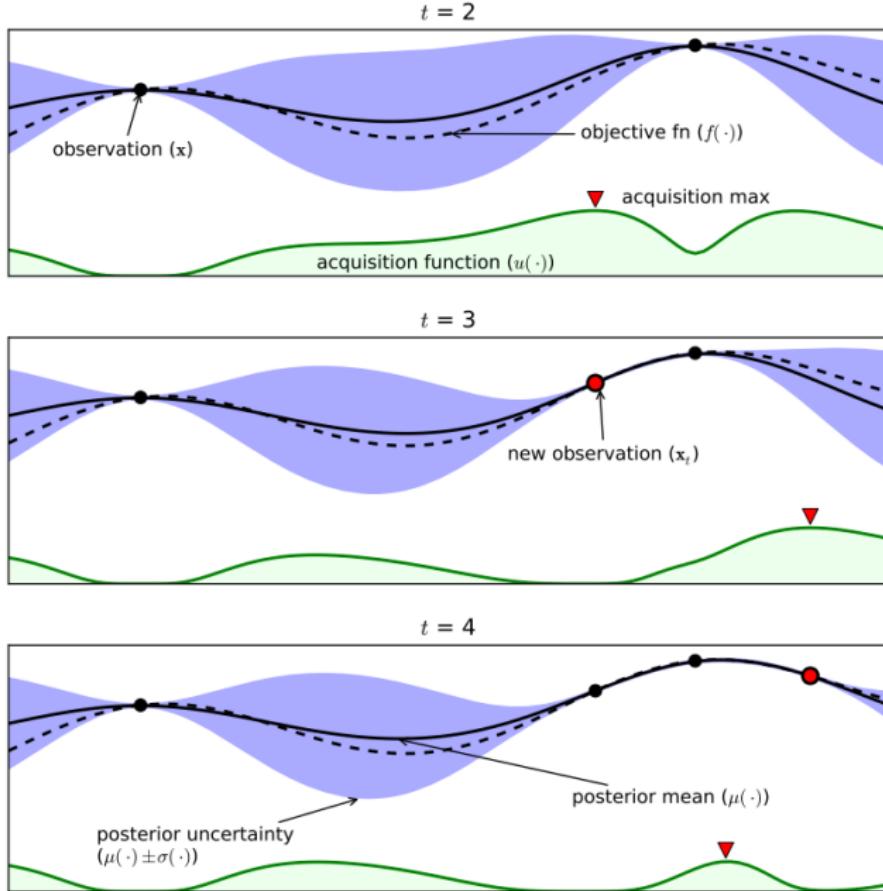
Otimização global com GPs

- Como saber que um ponto pertence a uma região de interesse?
- Desejamos usar toda a informação proveniente de avaliações anteriores na escolha do próximo candidato a ser avaliado.
- Usamos uma **função de aquisição**, que **quantifica a relevância** de um ponto candidato.
- A função de aquisição $a : \mathcal{X} \rightarrow \mathbb{R}^+$ é escolhida para ter um formato facilmente calculável.
- O próximo ponto a ser escolhido para ser avaliado será dado por:

$$\boldsymbol{x}_* = \arg \max a(\boldsymbol{x}).$$

- Exemplos de funções de aquisição usuais:
 - Probability of Improvement;
 - Expected Improvement;
 - GP Lower Confidence Bound.

Ilustração das etapas de otimização Bayesiana



Funções de aquisição

- Probability of Improvement

- Maximiza a probabilidade de obter pontos melhores que a melhor solução atual $\mu^- = \min_i \mu(\mathbf{x}_i)$, em que $\mu(\mathbf{x}_i)$ indica a média predita pelo GP em \mathbf{x}_i .
- Foco em *exploitation*, com equilíbrio determinado pelo hiperparâmetro $\xi \geq 0$ (valor usual $\xi = 0.01$):

$$\text{PI}(\mathbf{x}) = P(f(\mathbf{x}) \leq \mu^- - \xi) = \Phi\left(\frac{\mu^- - \xi - \mu(\mathbf{x})}{\sigma(\mathbf{x})}\right),$$

em que $\sigma(\mathbf{x}_i)$ indica o desvio-padrão predito pelo GP em \mathbf{x}_i e $\Phi(\cdot)$ é função cumulativa de probabilidade da Gaussiana normalizada:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{1}{2}t^2\right) dt.$$

Funções de aquisição

- Expected Improvement

- Maximiza a magnitude da melhoria esperada em relação à melhor solução atual μ^- .
- Considerando o hiperparâmetro $\xi \geq 0$ (valor usual $\xi = 0.01$):

$$EI(\mathbf{x}) = \begin{cases} 0. & \text{se } \sigma(\mathbf{x}) = 0; \\ \tau(\mathbf{x})\Phi\left(\frac{\tau(\mathbf{x})}{\sigma(\mathbf{x})}\right) + \sigma(\mathbf{x})\phi\left(\frac{\tau(\mathbf{x})}{\sigma(\mathbf{x})}\right), & \text{se } \sigma(\mathbf{x}) > 0, \end{cases}$$

em que $\tau(\mathbf{x}) = \mu^- - \xi - \mu(\mathbf{x})$.

Acima, $\Phi(\cdot)$ e $\phi(\cdot)$ são respectivamente as distribuições cumulativa e de densidade da Gaussiana normalizada.

Funções de aquisição

- GP Lower Confidence Bound

- Considera os intervalos de confiança do modelo de GP para escolher o próximo ponto a ser avaliado.
- Considerando o hiperparâmetro $\kappa > 0$:

$$\text{LCB}(\mathbf{x}) = \mu(\mathbf{x}) - \kappa\sigma(\mathbf{x}),$$

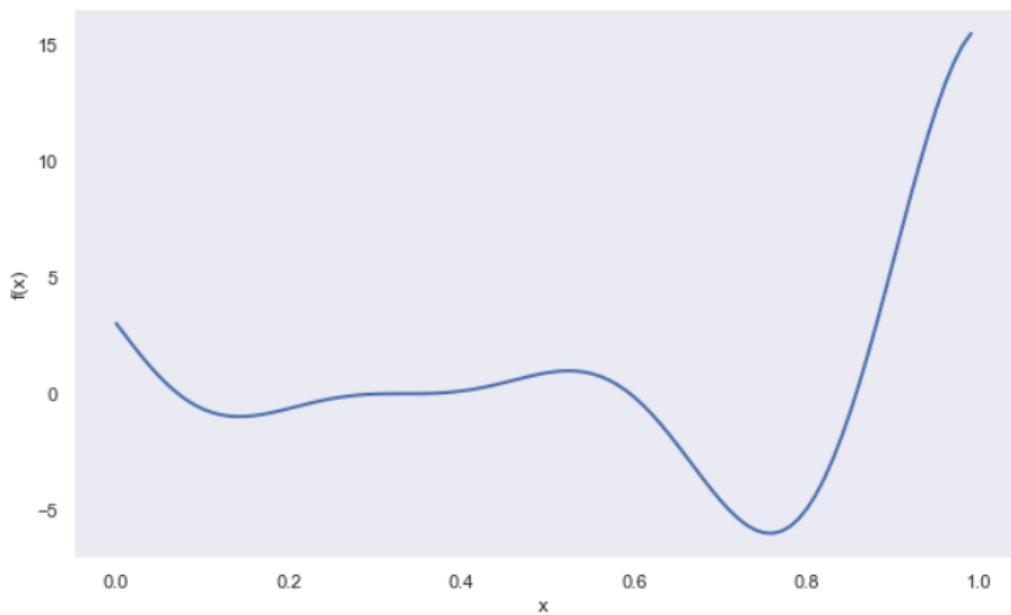
em que $\kappa = \sqrt{\nu\beta_t}$, $\beta_t = 2 \log(t^{D/2+2}\pi^2/3\delta)$, sendo t a iteração do procedimento de otimização e $\nu, \delta > 0$. Valores usuais: $\delta = 0.1$ e $\nu = 0.2$.

Otimização Bayesiana com GPs

Resumo do algoritmo

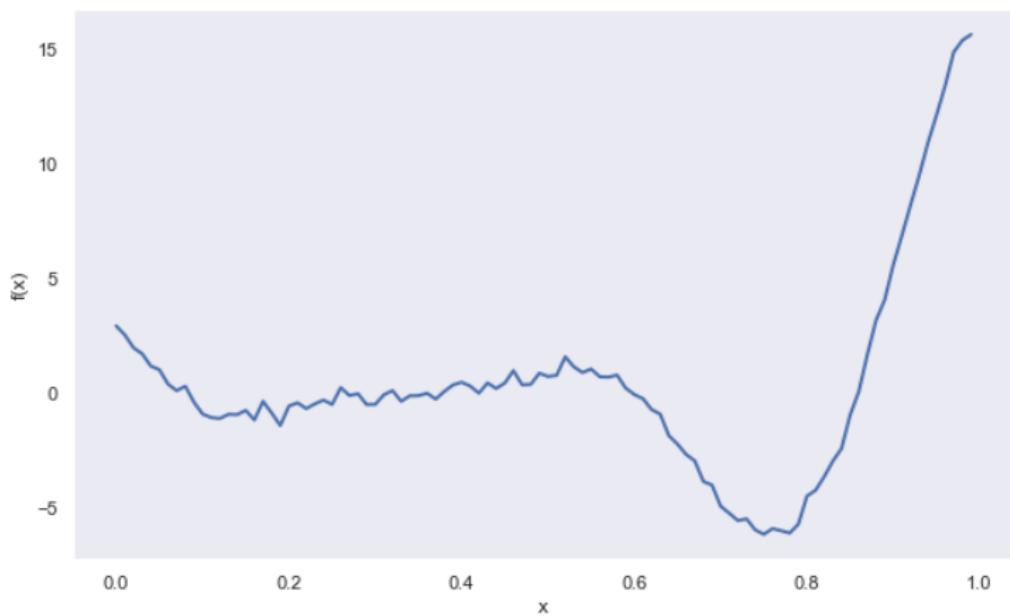
- ① Defina o conjunto de dados inicial $\mathcal{D}_0 = (\mathbf{x}_i, y_i)|_{i=1}^{N_0}$;
- ② Defina uma função de aquisição $a(\mathbf{x})$;
- ③ Inicialize o modelo de GP a partir de \mathcal{D}_0 ;
- ④ Repita até convergir ou alcançar o limite de iterações:
 - ① Encontre o ponto a ser avaliado maximizando a função de aquisição:
$$\mathbf{x}_t = \arg \max_{\mathbf{x}} a(\mathbf{x} | \mathcal{D}_{t-1});$$
 - ② Encontre a avaliação $y_t = f(\mathbf{x}_t) + \epsilon_t$;
 - ③ Aumente o conjunto de dados $\mathcal{D}_t = \{\mathcal{D}_{t-1}, (\mathbf{x}_t, y_t)\}$;
 - ④ Atualize o modelo de GP com o novo conjunto de dados;
- ⑤ Retorne o ponto \mathbf{x}_t com melhor valor correspondente y_t .

Otimização Bayesiana com GPs - 1D



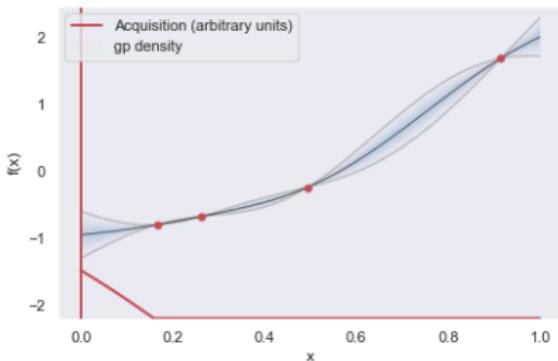
$$f(x) = (6x - 2)^2 \sin(12x - 4)$$

Otimização Bayesiana com GPs - 1D

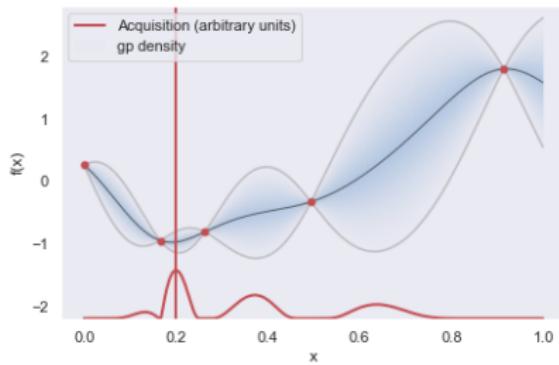


$$f(x) = (6x - 2)^2 \sin(12x - 4) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.25^2)$$

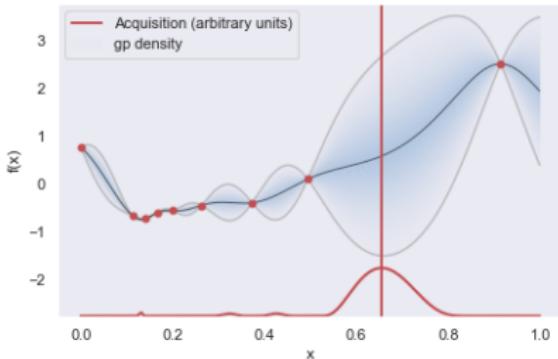
Otimização Bayesiana com GPs - 1D



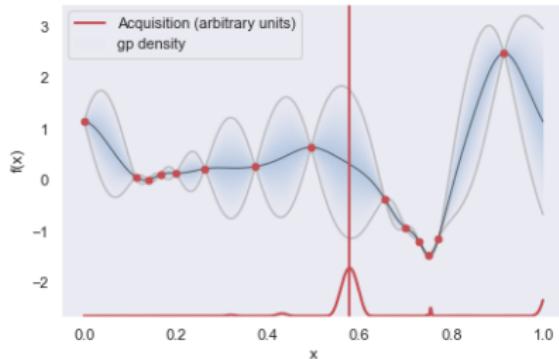
4 pontos iniciais



Depois de 1 iteração

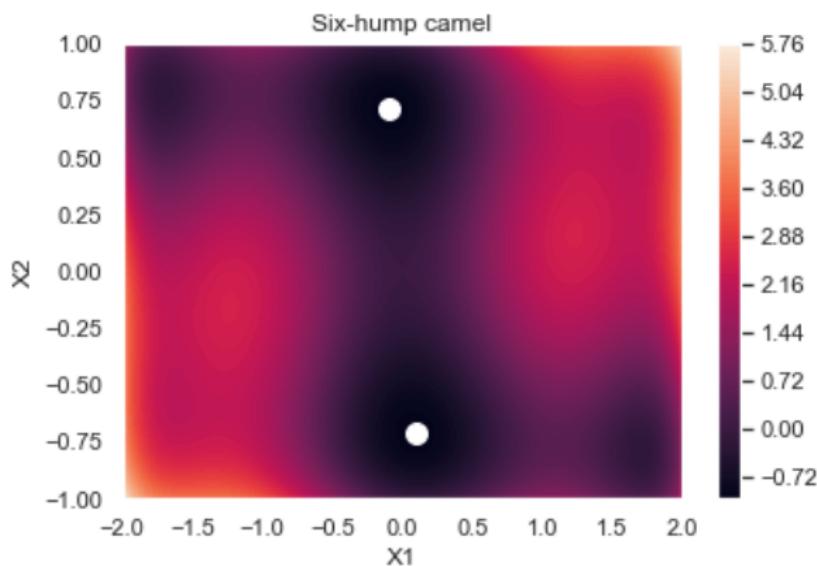


Depois de 5 iterações



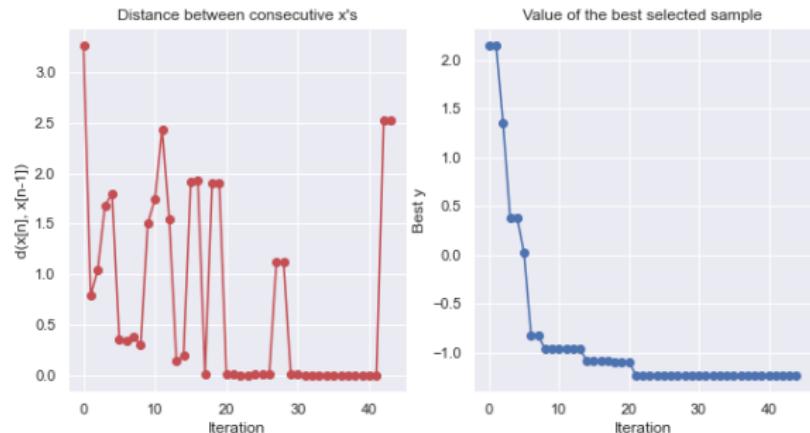
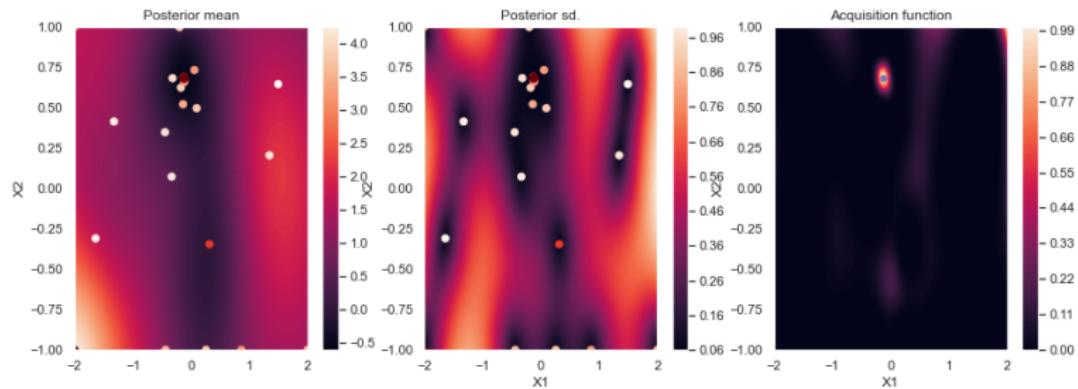
Depois de 10 iterações

Otimização Bayesiana com GPs - 2D

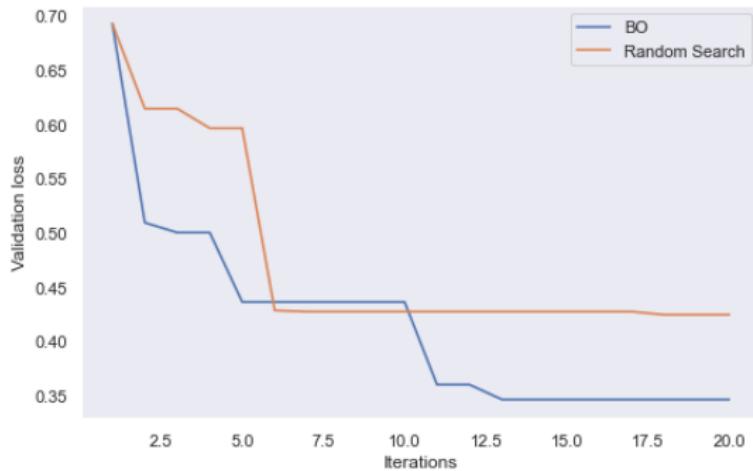


$$f(x_1, x_2) = \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2$$

Otimização Bayesiana com GPs - 2D



Otimização Bayesiana com GPs - MLP

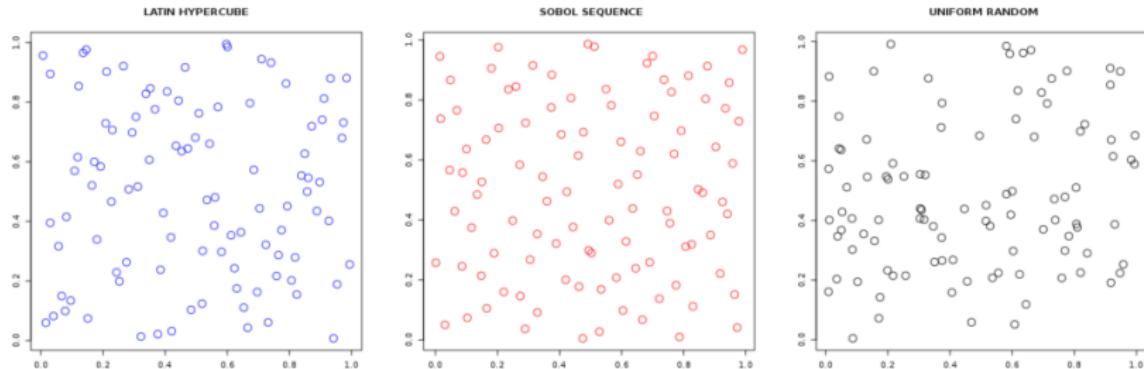


MLP com 2 camadas ocultas para análise de sentimento no IMDB dataset (ReLU, 20 épocas) para 20 avaliações, sendo 5 de inicialização:

	Domínio	BO	Random search
Neurônios ocultos	$2^{[2,9]}$	32, 4	4, 16
Taxa de aprendizagem (SGD)	$10^{[-5, -0.3]}$	5.488e-03	2.356e-02
Fator de momentum	$10^{[-7, -0.001]}$	1.431e-03	1.281e-04
Regularização L2	$10^{[-7, -1]}$	5.223e-05	3.201e-03
Tamanho do minibatch	$2^{[2,9]}$	32	4
Perda no teste, acurácia no teste		0.3466, 0.8543	0.4546, 0.8311

Otimização Bayesiana - Aspectos práticos

- Use um kernel menos suave que o RBF, como Matérn $\frac{5}{2}$ ou $\frac{3}{2}$.
- Use métodos eficientes para otimizar a função de aquisição (busca em árvores, hill-climbing, Monte Carlo, etc).
- Inicialize cuidadosamente o conjunto de dados \mathcal{D}_0 usando amostras via sequência de Sobol ou latin hypercube.
- Se houver conhecimento a priori sobre o problema, escolha pontos iniciais informativos.
- Use uma implementação bem estabelecida: GPyOpt, GPflowOpt, Spearmint, BoTorch, Dragonfly, pybo...



Agenda

① Processos Gaussianos

GPs para Regressão

De espaços de atributos para GPs

② GPs para grandes conjuntos de dados

GP esparso variacional

GP com inferência variacional estocástica

③ Otimização Bayesiana

④ Outros tópicos

Classificação

Aprendizagem não-supervisionada e redução de dimensionalidade

Modelagem hierárquica

Modelagem dinâmica

Robótica e controle

Aprendizagem robusta

⑤ Conclusão

⑥ Referências

GPs para classificação binária

- A saída observada (a **probabilidade de uma classe**) está relacionada à função latente $f(\cdot)$ via uma verossimilhança não-Gaussiana $p(y_i = 1|f(\mathbf{x}_i))$:

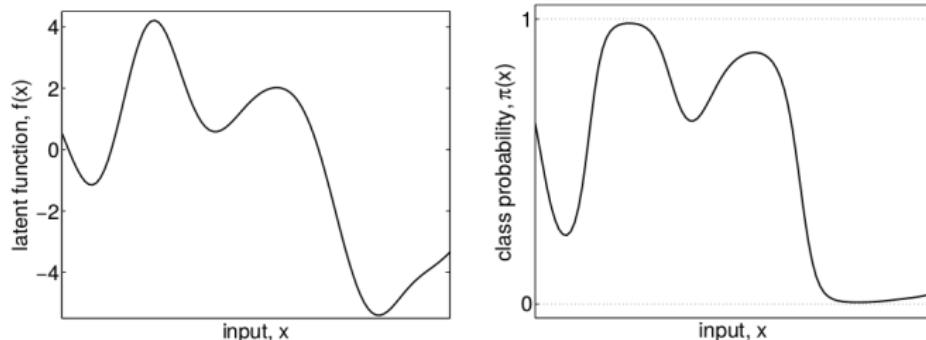
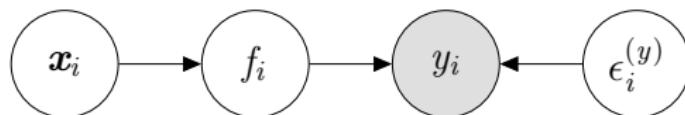


Figura 2: “Achatando” uma função latente para valores de probabilidade.

- São necessários métodos de **inferência aproximada**:
 - Métodos de amostragem (e.g. MCMC);
 - Aproximação de Laplace;
 - Expectation Propagation (EP);
 - Variational Bayes (VB).

GPs para aprendizagem não-supervisionada

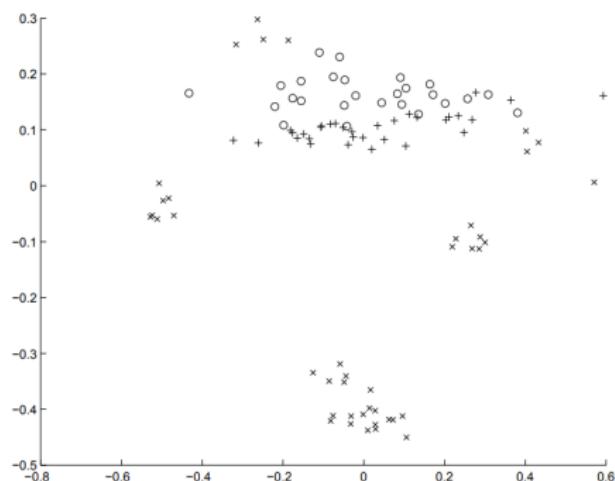
- Dados (ruidosos) observados, mas **com entradas latentes desconhecidas**.
- Relacionado ao problema de **redução de dimensionalidade não-linear**.



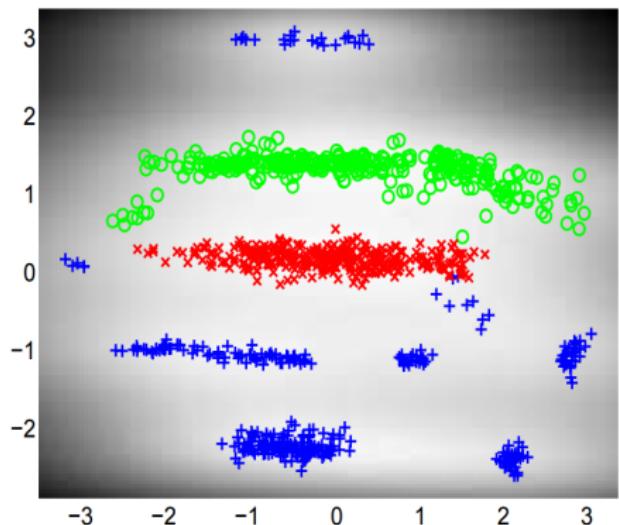
Gaussian Process Latent Variable Model (GPLVM)

- Priori de GP sobre o mapeamento $f(\cdot)$ desconhecido.
- Priori sobre as entradas latentes, e.g.
 $p(\mathbf{X}) = \prod_{i=1}^N \mathcal{N}(\mathbf{x}_i | \mathbf{0}, \sigma_x^2 \mathbf{I})$.
- A propagação da incerteza por uma função não-linear não é analítica.
- Usualmente considera inferência aproximada variacional.

GPs para a visualização de dados de alta dimensão



PCA

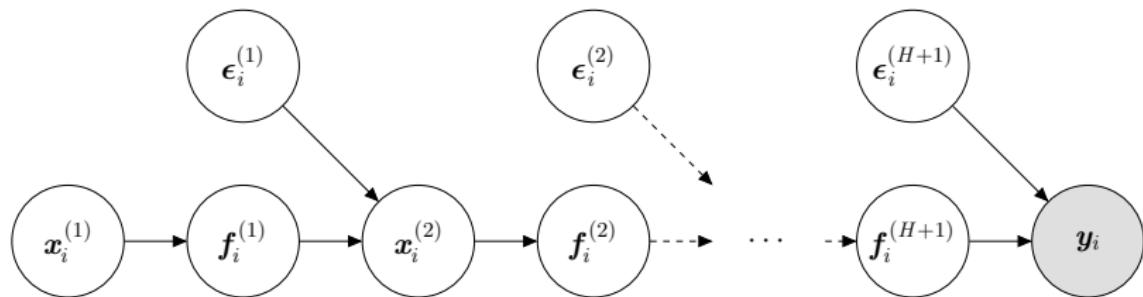


Bayesian GPLVM

Exemplo de redução de dimensionalidade do conjunto de dados Oil, com 12 dimensões.

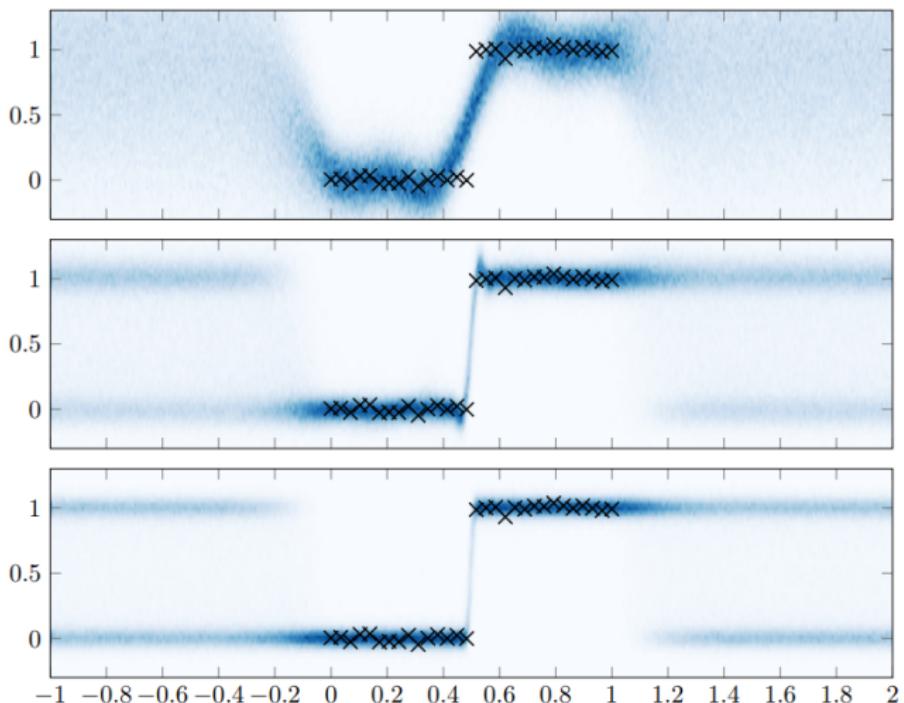
Deep GPs

- Se considerarmos um GPLVM com uma priori de GP em suas entradas, obtemos um modelo de **Deep GP** com duas camadas.
- Múltiplas camadas ocultas permitem uma **estrutura hierárquica poderosa** para aprendizagem profunda não-supervisionada e supervisionada.



Estrutura hierárquica de um modelo de Deep GP.

Deep GP para aprendizagem de funções não-suaves



Aprendizagem de uma função degrau a partir de dados ruidosos com um único GP (topo) e um Deep GPs com 2 (meio) e 3 camadas (abaixo).

Modelos de GP com dinâmicas internas/externas

- **Modelos com dinâmicas externas:** Usa medidas como regressores.
 - **Nonlinear autoregressive with exogenous inputs (NARX) model:**

$$y_i = f(\mathbf{x}_i) + \epsilon_i^{(y)},$$

$$\mathbf{x}_i = [[y_{i-1}, y_{i-2}, \dots, y_{i-L_y}], [u_{i-1}, u_{i-2}, \dots, u_{i-L_u}]]^\top.$$

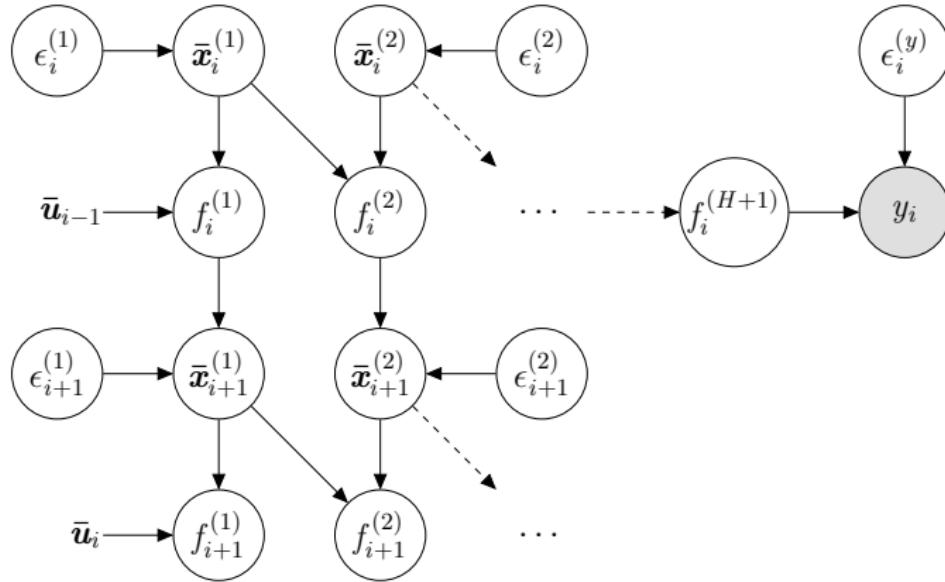
- **Modelos com dinâmicas internas:** Usam estados latentes.
 - **State-space model (SSM):**

$$\mathbf{x}_i = f(\mathbf{x}_{i-1}, u_{i-1}) + \epsilon_i^{(x)},$$

$$y_i = g(\mathbf{x}_i) + \epsilon_i^{(y)},$$

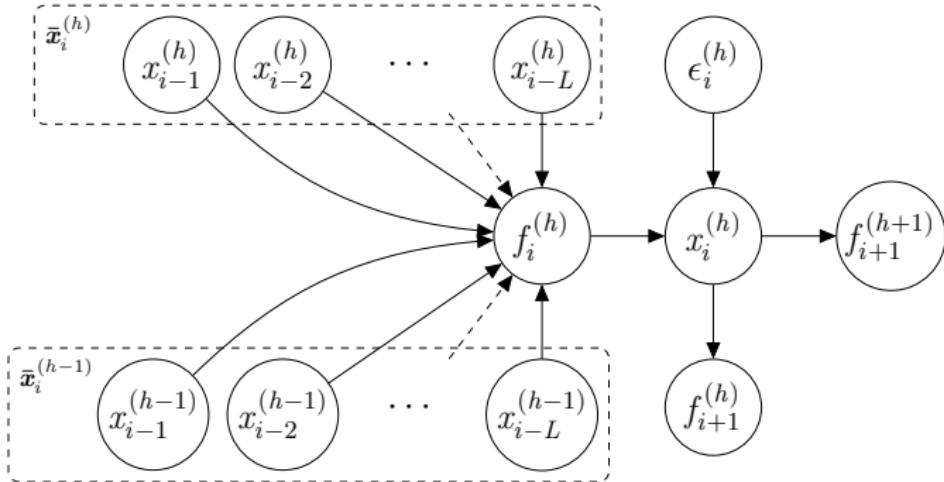
- **Modelos de GP dinâmicos:** Funções desconhecidas recebem priori de GP.
 - **GP-NARX:** Analítico para ruído de observação Gaussiano.
 - **GP-SSM:** Não-analítico devido a presença de entradas latentes.

Recurrent Gaussian Processes (RGPs)



Modelo gráfico de um RGP com H camadas ocultas.

- **Estrutura hierárquica:** Modelagem separada das funções de transição (ocultas) e de observação (emissão).
- **Variáveis latentes dinâmicas:** Sem *feedback* das observações.
- **REVARB (REcurrent VARIational Bayes):**



Detalhamento de uma camada de transição recorrente no RGP.

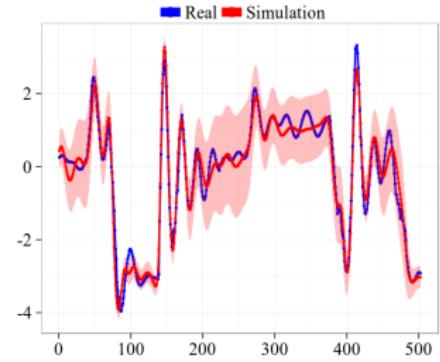
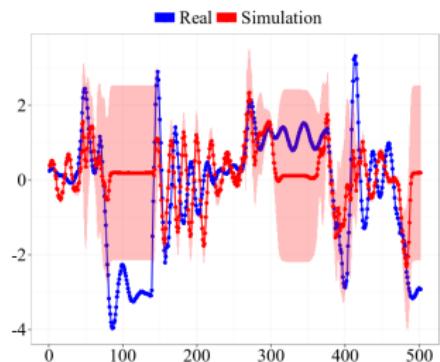
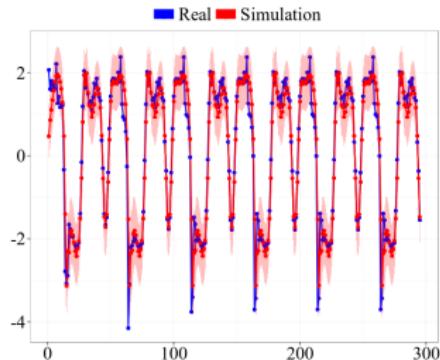
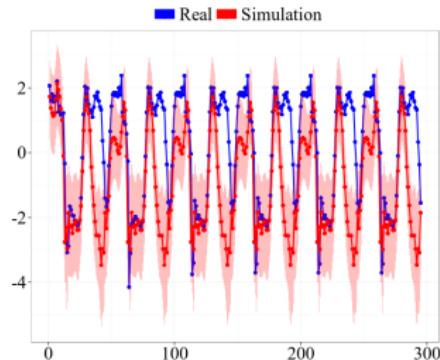
$$p \left(\mathbf{f}^{(h)} \middle| \hat{\mathbf{X}}^{(h)} \right) = \mathcal{N} \left(\mathbf{f}^{(h)} \middle| \mathbf{0}, \mathbf{K}_f^{(h)} \right), \quad 1 \leq h \leq H + 1,$$

$$p \left(x_i^{(h)} \right) = \mathcal{N} \left(x_i^{(h)} \middle| \mu_{0i}^{(h)}, \lambda_{0i}^{(h)} \right), \quad 1 \leq i \leq L,$$

$$p \left(x_i^{(h)} \middle| f_i^{(h)} \right) = \mathcal{N} \left(x_i^{(h)} \middle| f_i^{(h)}, \sigma_h^2 \right), \quad L + 1 \leq i \leq N,$$

$$p \left(y_i \middle| f_i^{(H+1)}, \sigma_{H+1}^2 \right) = \mathcal{N} \left(y_i \middle| f_i^{(H+1)}, \sigma_{H+1}^2 \right), \quad L + 1 \leq i \leq N.$$

RGP para identificação de sistemas (simul. livre)



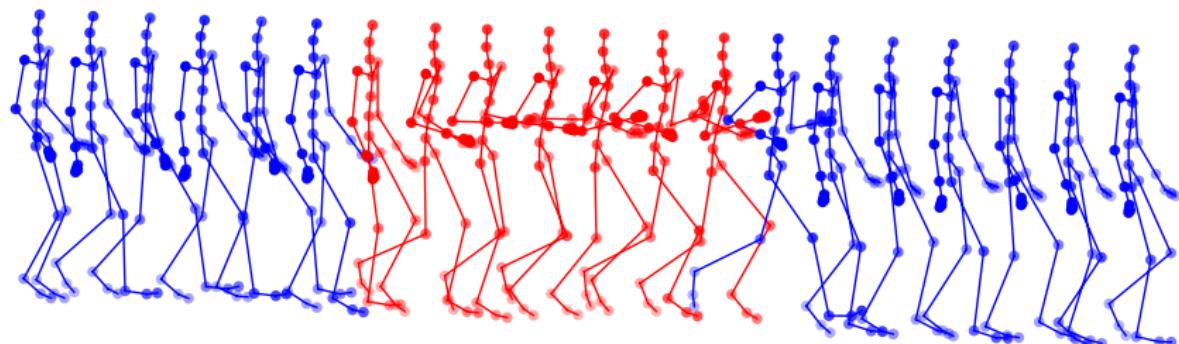
RGP para movimento humano e controle de avatar

- Movimentações de caminhar e correr com 57 dimensões de saída e coordenação com o pé esquerdo como sinal de entrada.

MLP-NARX	GP-NARX	RGP ($H = 2$)
1.2141	0.8987	0.8600

Valores de RMSE para dados de teste em simulação livre.

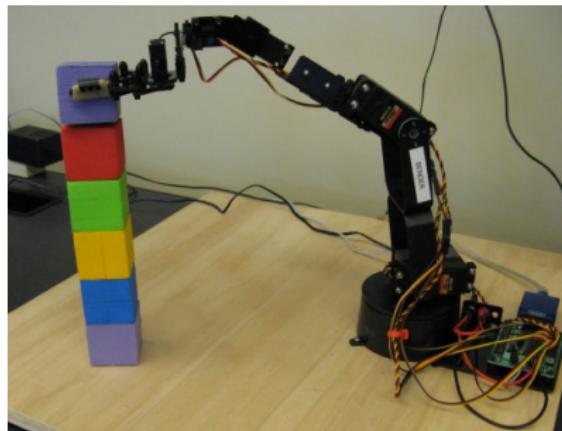
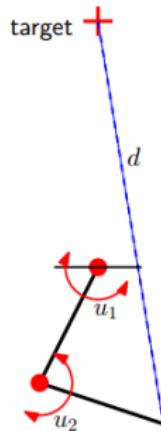
- Velocidade como sinal para controle do movimento do avatar.



Movimento gerado por um modelo RGP.

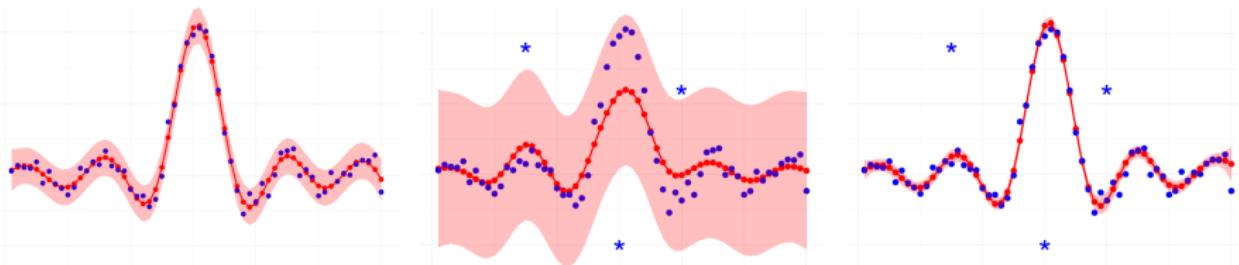
GP para robótica e controle

- Probabilistic Inference for Learning Control (PILCO)
- Aprendizagem autônoma para busca de política de controle baseada no modelo.
- Usa a incerteza fornecida pelo modelo de GP.



Exemplos de sistemas controlados via PILCO. Vídeos disponíveis em
<http://www.youtube.com/user/PilcoLearner>.

Aprendizagem na presença de outliers



(a) GP com ruído Gaussiano.

(b) GP com outliers.

c) GP robusto com outliers.

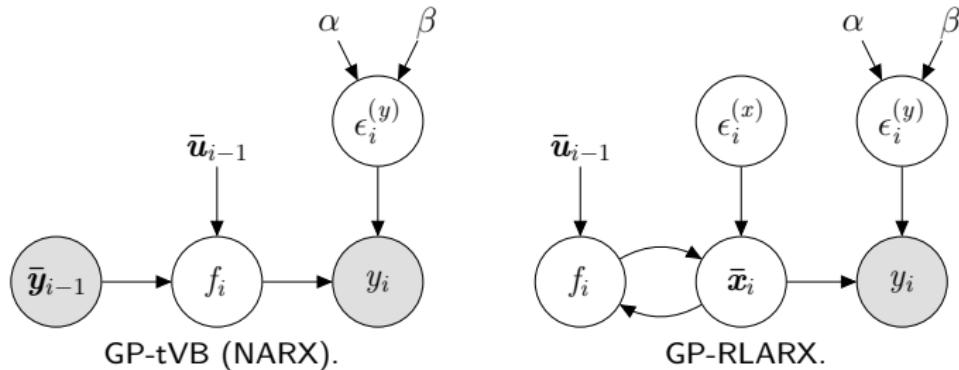
Efeito de outliers em modelos de GP.

- **Abordagem mais comum:** verossimilhança de cauda pesada + inferência aproximada.
 - **GP-tVB:** Verossimilhança Student- t e inferência variacional.
 - **GP-LEP:** Verossimilhança Laplace e Expectation Propagation.

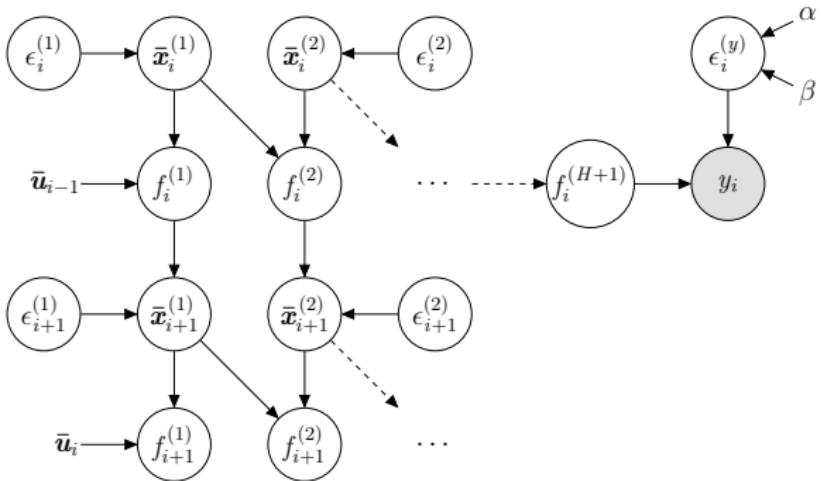
Robust GP Latent Autoregressive (GP-RLARX)

Camada GP de transição com autorregressão latente, função de emissão identidade, verossimilhança Student- t e inferência variacional.

$$\begin{aligned} p(\mathbf{f}|\hat{\mathbf{X}}) &= \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_f), \\ p(x_i) &= \mathcal{N}(x_i|\mu_{0i}, \lambda_{0i}), & 1 \leq i \leq L, \\ p(x_i|f_i) &= \mathcal{N}(x_i|f_i, \sigma_x^2), & L+1 \leq i \leq N, \\ p(y_i|x_i, \tau_i) &= \mathcal{N}(y_i|x_i, \tau_i^{-1}), & L+1 \leq i \leq N, \\ p(\tau_i) &= \Gamma(\tau_i|\alpha, \beta), & L+1 \leq i \leq N, \end{aligned}$$



RGP- t /REVARB- t : RGP robusto com Student- t



$$p \left(\mathbf{f}^{(h)} \middle| \hat{\mathbf{X}}^{(h)} \right) = \mathcal{N} \left(\mathbf{f}^{(h)} \middle| \mathbf{0}, \mathbf{K}_f^{(h)} \right), \quad 1 \leq h \leq H + 1,$$

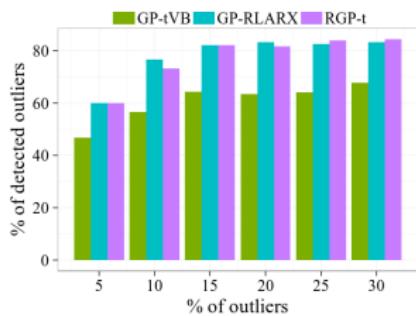
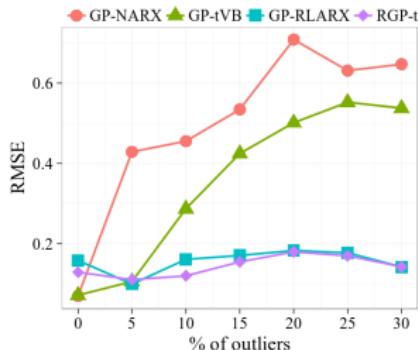
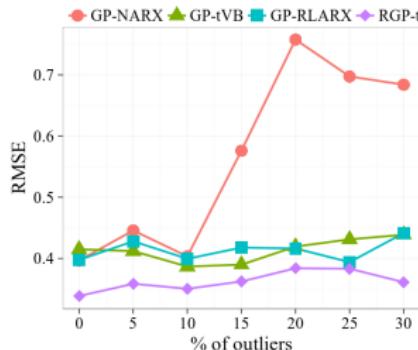
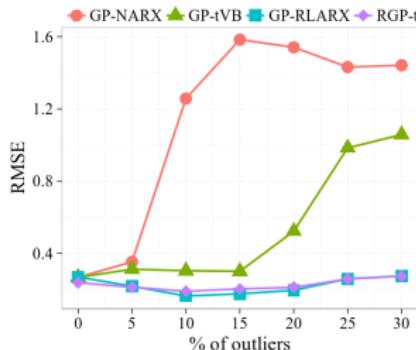
$$p \left(x_i^{(h)} \right) = \mathcal{N} \left(x_i^{(h)} \middle| \mu_{0i}^{(h)}, \lambda_{0i}^{(h)} \right), \quad 1 \leq i \leq L,$$

$$p \left(x_i^{(h)} \middle| f_i^{(h)} \right) = \mathcal{N} \left(x_i^{(h)} \middle| f_i^{(h)}, \sigma_h^2 \right), \quad L + 1 \leq i \leq N,$$

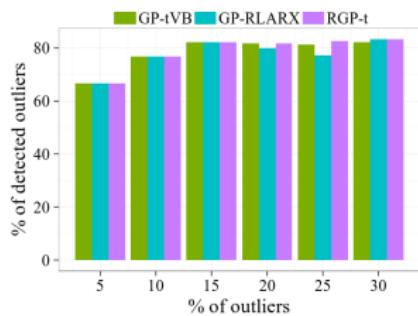
$$p \left(y_i \middle| f_i^{(H+1)}, \tau_i \right) = \mathcal{N} \left(y_i \middle| f_i^{(H+1)}, \tau_i^{-1} \right), \quad L + 1 \leq i \leq N,$$

$$p(\tau_i) = \Gamma(\tau_i | \alpha, \beta), \quad L + 1 \leq i \leq N.$$

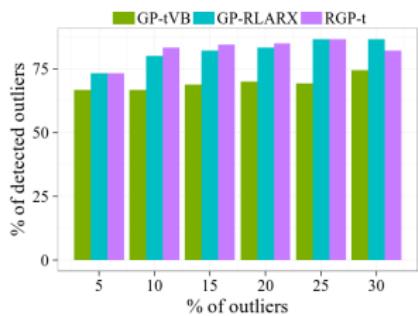
GP para identificação de sistemas com outliers



(a) Artificial 1.



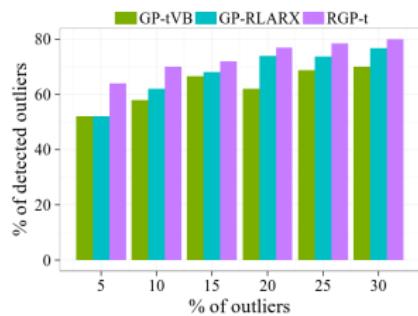
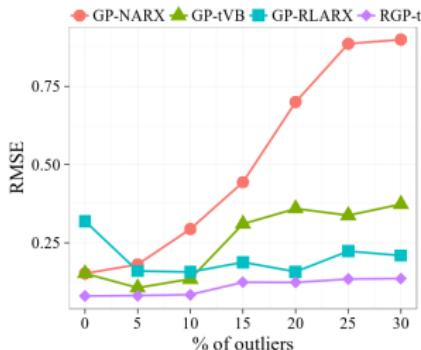
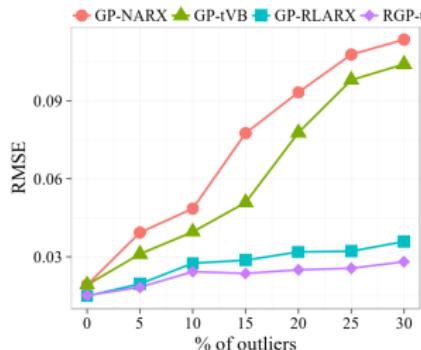
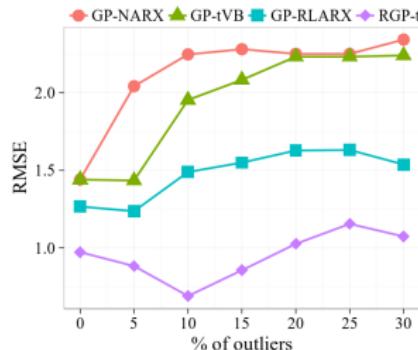
(b) Artificial 2.



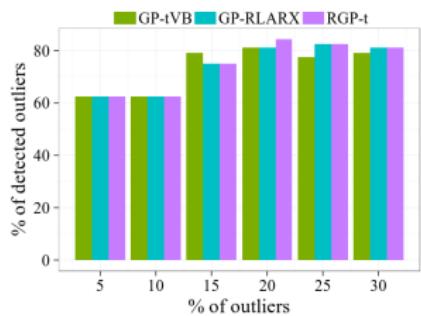
(c) Artificial 3.

Linhas são os valores de RMSE em simulação livre e barras são os percentuais de outliers detectados.

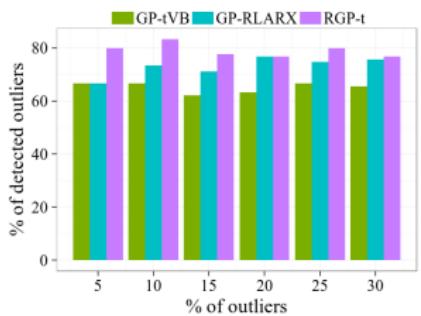
GP para identificação de sistemas com outliers



(d) Artificial 4.



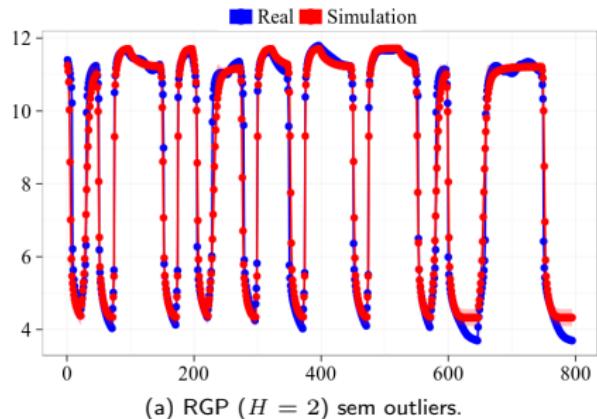
(e) Artificial 5.



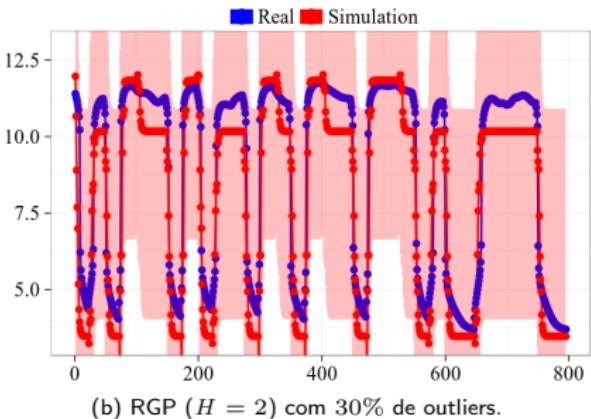
(f) Artificial 6.

Linhas são os valores de RMSE em simulação livre e barras são os percentuais de outliers detectados.

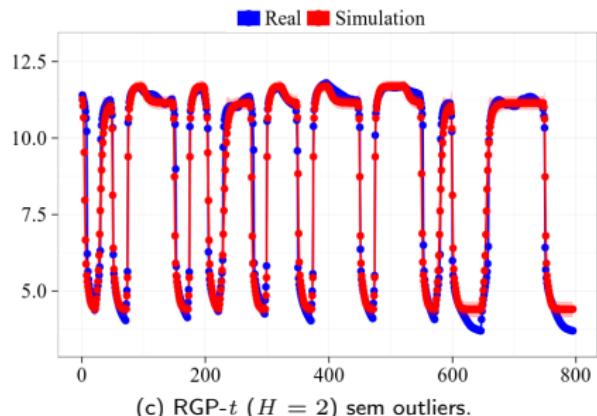
RGP- t para identificação robusta (dados de pH)



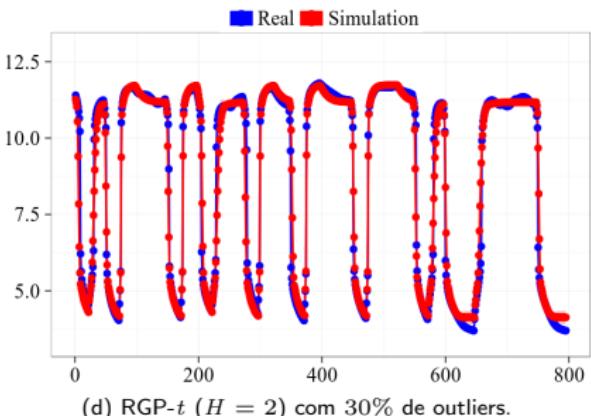
(a) RGP ($H = 2$) sem outliers.



(b) RGP ($H = 2$) com 30% de outliers.

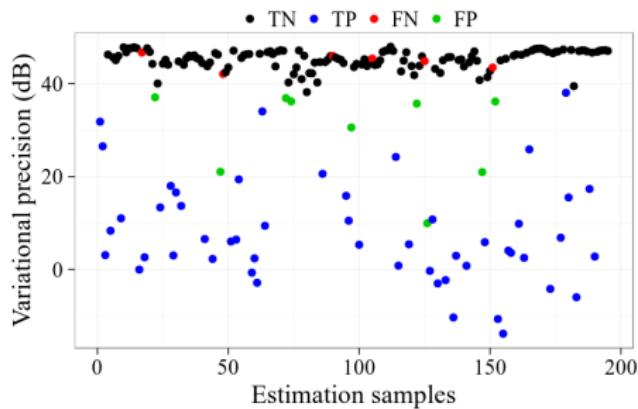


(c) RGP- t ($H = 2$) sem outliers.

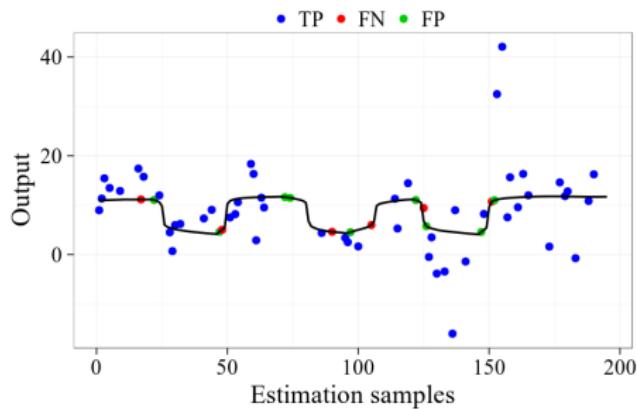


(d) RGP- t ($H = 2$) com 30% de outliers.

RGP- t para identificação robusta (dados de pH)



(a) Precisões variacionais otimizadas.



(b) Outliers detectados.

Detecção de outliers pelo modelo RGP- t com 2 camadas ocultas e inferência variacional (REVARB- t) com 30% de outliers no treinamento.

Agenda

① Processos Gaussianos

GPs para Regressão

De espaços de atributos para GPs

② GPs para grandes conjuntos de dados

GP esparso variacional

GP com inferência variacional estocástica

③ Otimização Bayesiana

④ Outros tópicos

Classificação

Aprendizagem não-supervisionada e redução de dimensionalidade

Modelagem hierárquica

Modelagem dinâmica

Robótica e controle

Aprendizagem robusta

⑤ Conclusão

⑥ Referências

Conclusão

- GP é um poderoso framework de aprendizagem **Bayesiano não-paramétrico**.
- A saída do modelo é um **distribuição de probabilidade bem definida**.
- **Otimização de hiperparâmetros** a partir da verossimilhança marginal.
- **Método versátil**: regressão, classificação, aprendizagem não-supervisionada, modelagem dinâmica, aprendizagem robusta a outliers, otimização global, etc.
- **Grande flexibilidade** a partir de diferentes funções de *kernel* e arquiteturas.

Agenda

① Processos Gaussianos

GPs para Regressão

De espaços de atributos para GPs

② GPs para grandes conjuntos de dados

GP esparso variacional

GP com inferência variacional estocástica

③ Otimização Bayesiana

④ Outros tópicos

Classificação

Aprendizagem não-supervisionada e redução de dimensionalidade

Modelagem hierárquica

Modelagem dinâmica

Robótica e controle

Aprendizagem robusta

⑤ Conclusão

⑥ Referências

Referências bibliográficas

- **Cap. 10** - RASMUSSEN, C.; WILLIAMS, C. **Gaussian Processes for Machine Learning**, 2006.
- **Cap. 15** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Cap. 6** - BISHOP, C. M. **Pattern recognition and machine learning**, 2006.
- MATTOS, C. L. C. **Recurrent Gaussian Processes and Robust Dynamical Modeling**, 2017
(http://www.repositorio.ufc.br/bitstream/riufc/25604/1/2017_tese_clcmattos.pdf).
- MATTOS, C. L. C. and TOBAR, F. **The Art of Gaussian Processes: Classic and Contemporary**, 2021
(https://github.com/GAMES-UChile/The_Art_of_Gaussian_Processes)