

Scalable Gaussian Process approximations

Botond Szabó (Bocconi University)

Bocconi Association for Mathematical Sciences (BAMS) seminar

Milano, Italy, 04. 03. 2026.



**Università
Bocconi**
MILANO

Co-authors

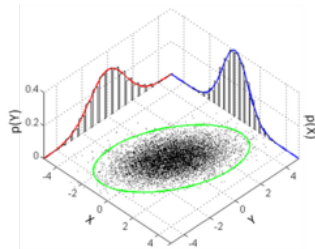
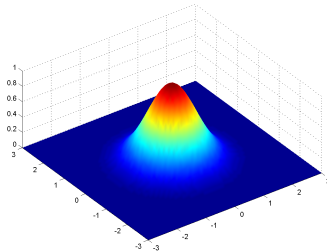
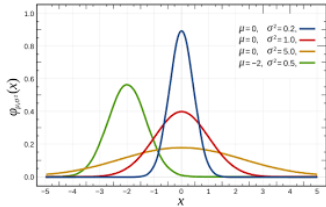
Harry van Zanten (VU Amsterdam), Dennis Nieman (Humboldt U. Berlin), Bernhard Stankewitz (Potsdam), Aad van der Vaart (Delft), Amine Hadji (Leiden), Geerten Koers (Leiden), Thibault Randrianarisoa (Toronto), Yichen Zhu (Hong Kong University), Lasse Vuursteen (Duke), Omar Kassi (Bocconi), Nicola Gorini (Bocconi)

Outline

- Introduction: Gaussian Processes and GP regression
- Frequentist Bayesian paradigm
- Approximation methods for GP
 - Distributed Bayes
 - Variational Bayes
 - Vecchia approximations
- Other models
- Summary

Gaussian Processes and applications

Gaussian distribution in 1D and 2D



Applications:

- **Modelling:** describes human height, weight, test scores, biological measurements, model asset returns, stock market volatility,...
- **Testing:** compute p-values, confidence intervals, t-test, ANOVA,...
- **Machine learning:** basis of classifiers like Naive Bayes, techniques like Gaussian Mixture Models (clustering), and Gaussian processes

Gaussian processes

Gaussian Process: A stochastic process $(W_t : t \in T)$ is called Gaussian if all its **finite-dimensional** marginals are multivariate-**normally** distributed.

They are **determined** by their **mean** function $m : T \rightarrow \mathbb{R}$ and **covariance** function $k : T \times T \rightarrow \mathbb{R}$:

- $m(t) = EW_t, \quad t \in T,$
- $k(t, s) = \text{cov}(W_s, W_t).$

Applications: Machine learning (e.g. classifying hand written digits, learning the inverse dynamics of robotic arms), Epidemiology (e.g. prevalence of malaria, Malaria Atlas Project), Climate Sciences (e.g. modeling ice sheet-climate interactions), Astronomy (e.g. background radiation), Finance (e.g. stock prices), Diffusion models (e.g. image generation),...

Example: Brownian motion

Definition (Brownian motion): A stochastic process $\{W_t\}_{t \geq 0}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *standard Brownian motion* (or *Wiener process*) if it satisfies the following properties:

- 1 **Initial condition:** $W_0 = 0$ almost surely.
- 2 **Independent increments:** For any $0 \leq t_0 < t_1 < \dots < t_n$, the increments

$$W_{t_1} - W_{t_0}, W_{t_2} - W_{t_1}, \dots, W_{t_n} - W_{t_{n-1}}$$

are independent random variables.

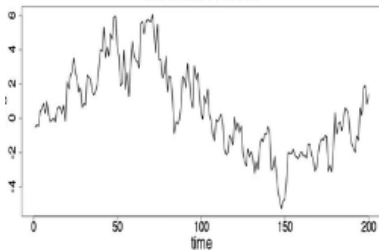
- 3 **Gaussian increments:** For all $0 \leq s < t$, $W_t - W_s \sim \mathcal{N}(0, t - s)$.
- 4 **Continuous paths:** $t \mapsto W_t(\omega)$ is continuous in t (with prob 1).

Properties:

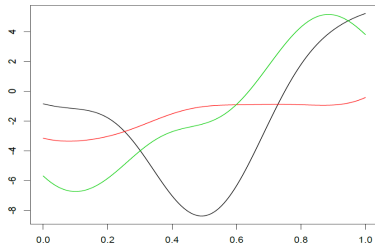
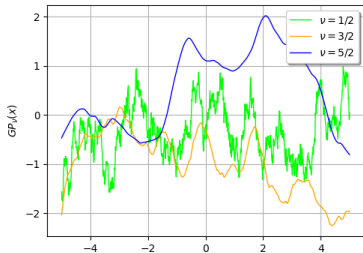
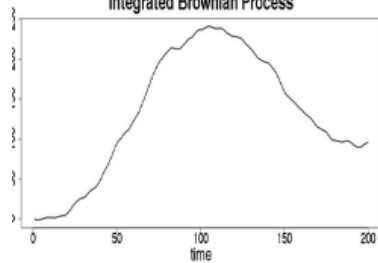
- mean $m(t) = 0$, covariance $k(t, s) = t \wedge s$.
- Sample paths are α -Holder smooth (for any $\alpha < 1/2$).

GP examples (BM, IBM, Matern, SE)

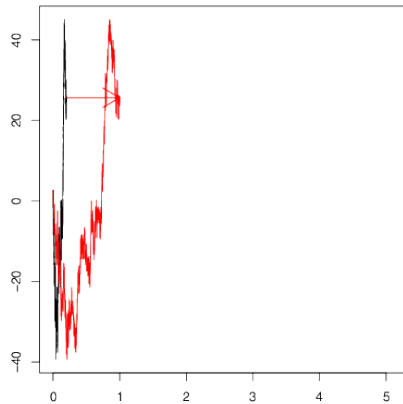
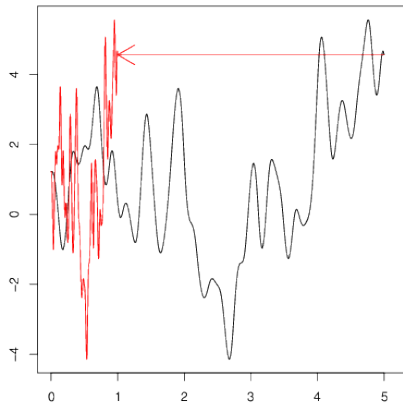
Brownian Process



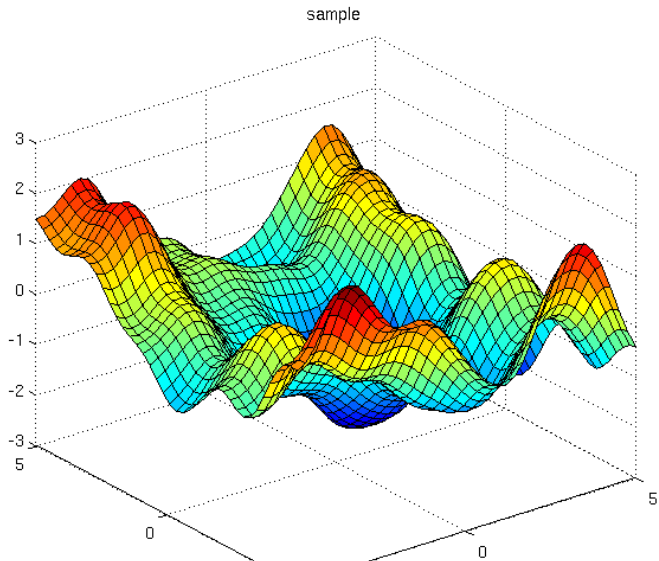
Integrated Brownian Process



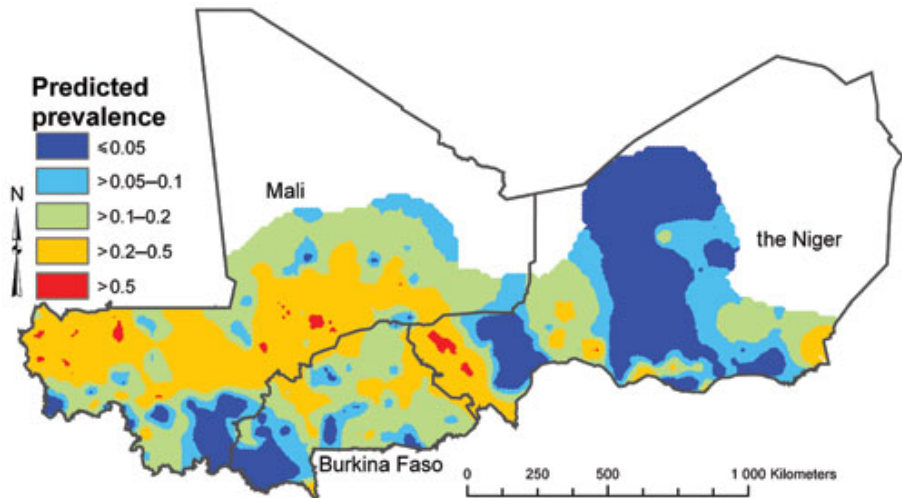
Recaling of GP



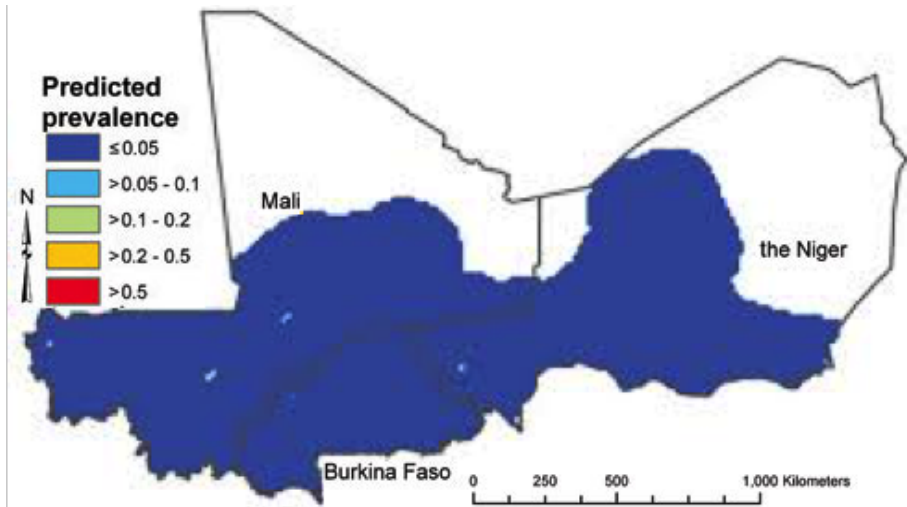
GP in 2D



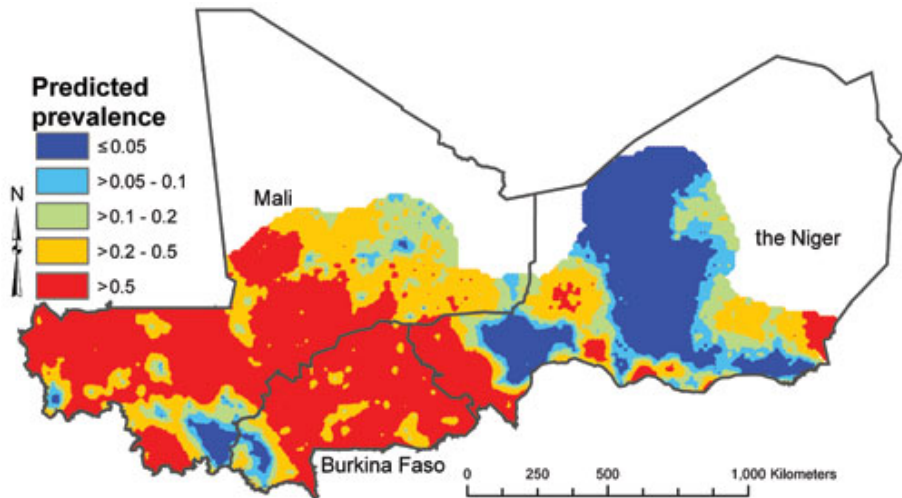
Prediction of infection: posterior mean



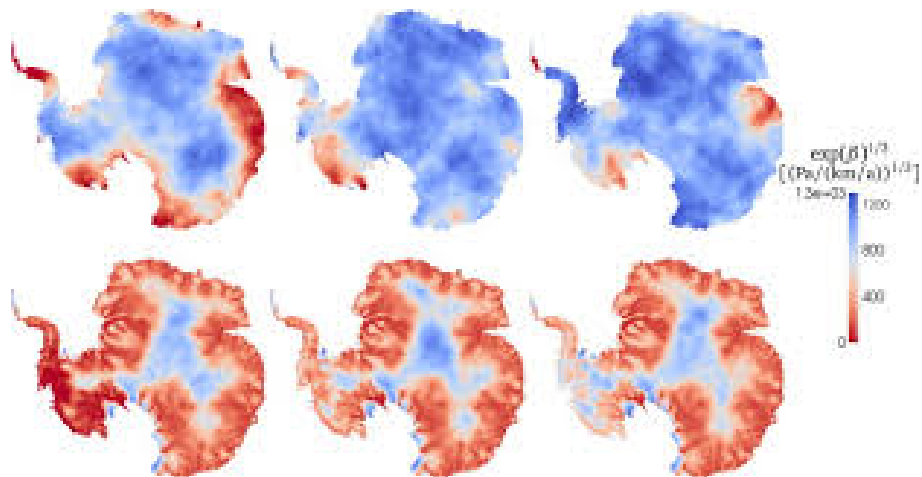
Prediction of infection: lower credible band



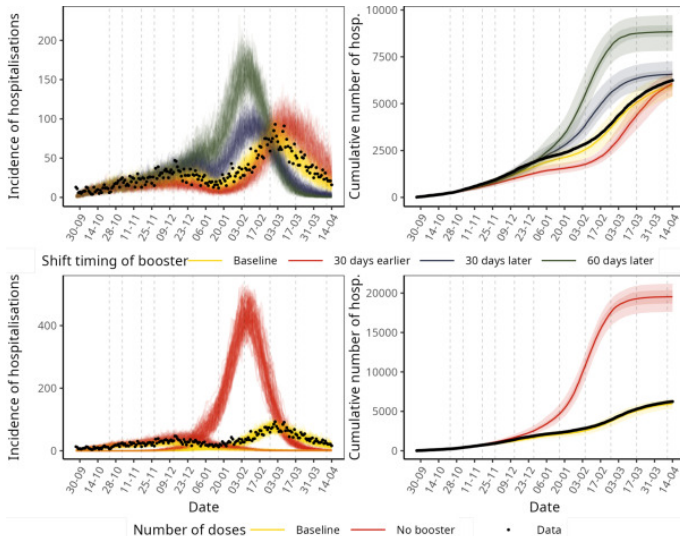
Prediction of infection: upper credible band



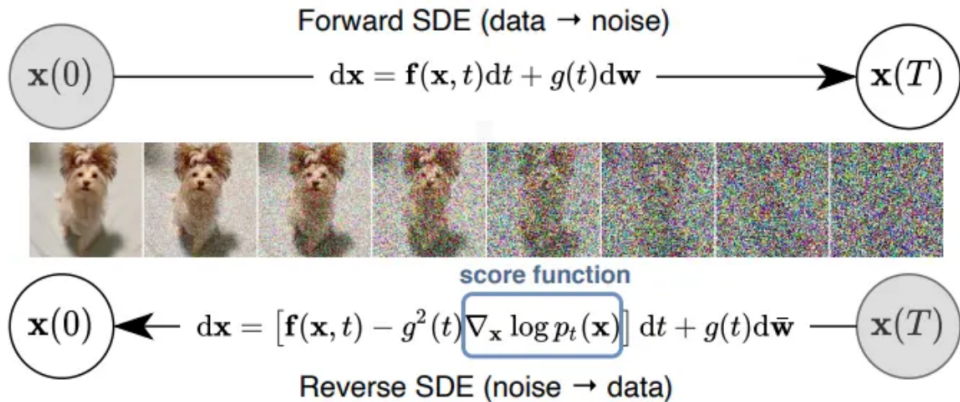
Modeling ice-sheet flow in antarctic



Effect of booster during covid



Generating images with diffusion model



Reverse SDE (noise $\rightarrow$ data): *Source*

Gaussian Process Regression

Gaussian process regression

Model: Assume that we observe the pairs (x_ℓ, y_ℓ) , $\ell = 1, \dots, n$,

$$y_\ell = f_0(x_\ell) + \sigma \varepsilon_\ell, \quad x_\ell \stackrel{iid}{\sim} G_x, \varepsilon_\ell \stackrel{iid}{\sim} N(0, 1),$$

where f_0 is the unknown function of interest.

Bayesian approach: Endow f_0 with $\Pi = GP(0, k)$.

Posterior distribution: Bayes formula

$$\Pi(f \in A | \mathbf{x}, \mathbf{y}) = \frac{\int_A p_f(\mathbf{x}, \mathbf{y}) d\Pi(f)}{\int p_f(\mathbf{x}, \mathbf{y}) d\Pi(f)}.$$

Posterior: GP, analytic form Williams and Rasmussen (2006).

$$\mathbf{x} \mapsto K_{\mathbf{x}\mathbf{f}}(\sigma^2 I + K_{\mathbf{ff}})^{-1} \mathbf{y},$$

$$(x, z) \mapsto k(x, z) - K_{\mathbf{x}\mathbf{f}}(\sigma^2 I + K_{\mathbf{ff}})^{-1} K_{\mathbf{f}z},$$

Here we denote $\mathbf{y} = (y_1, \dots, y_n)^T$, $\mathbf{f} = (f(x_1), \dots, f(x_n))^T$,

$K_{\mathbf{x}\mathbf{f}} = \text{cov}_\Pi(f(x), \mathbf{f}) = (k(x, x_1), \dots, k(x, x_n))$, $K_{\mathbf{ff}} = \text{cov}_\Pi(\mathbf{f}, \mathbf{f}) = [k(x_i, x_j)]_{1 \leq i, j \leq n}$.

Derivation of the posterior distribution

Notations: training $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ and test $\mathbf{x}_* = \{\mathbf{x}_*^{(1)}, \dots, \mathbf{x}_*^{(m)}\}$ inputs.

Distribution of $\mathbf{y}$: $\mathbf{y} = f(\mathbf{x}) + \varepsilon \sim \mathcal{N}_n(0, K + \sigma^2 I)$, where $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$ and $K = k(\mathbf{x}, \mathbf{x})$.

$$\begin{pmatrix} f(\mathbf{x}) \\ f(\mathbf{x}_*) \end{pmatrix} \sim \mathcal{N} \left(0, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix} \right),$$

where

$$K = k(\mathbf{x}, \mathbf{x}), \quad K_* = k(\mathbf{x}, \mathbf{x}_*), \quad K_{**} = k(\mathbf{x}_*, \mathbf{x}_*).$$

Including observation noise, we obtain

$$\begin{pmatrix} \mathbf{y} \\ f(\mathbf{x}_*) \end{pmatrix} \sim \mathcal{N} \left(0, \begin{pmatrix} K + \sigma^2 I & K_* \\ K_*^T & K_{**} \end{pmatrix} \right).$$

Derivation II

Conditional Gaussian distribution: If

$$\begin{pmatrix} a \\ b \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \right),$$

then

$$b \mid a \sim \mathcal{N} \left(\mu_b + \Sigma_{ba} \Sigma_{aa}^{-1} (a - \mu_a), \Sigma_{bb} - \Sigma_{ba} \Sigma_{aa}^{-1} \Sigma_{ab} \right).$$

Posterior mean and variance: Identifying $a = \mathbf{y}$ and $b = f(\mathbf{x}_*)$, we obtain

- **posterior mean:** $\mathbb{E}[f(\mathbf{x}_*) \mid \mathbf{y}] = K_*^T (K + \sigma^2 I)^{-1} \mathbf{y},$
- **posterior covariance:** $\text{Cov}(f(\mathbf{x}_*) \mid \mathbf{y}) = K_{**} - K_*^T (K + \sigma^2 I)^{-1} K_*.$

Computation

Conjugacy: the posterior has an explicit form.

Problem: **Computation** time of the posterior for training $O(n^3)$ and prediction $O(n^2)$.
Memory requirement $O(n^2)$. Becomes **impractical** for large data set.

Computation

Conjugacy: the posterior has an explicit form.

Problem: **Computation** time of the posterior for training $O(n^3)$ and prediction $O(n^2)$. Memory requirement $O(n^2)$. Becomes **impractical** for large data set.

Problem: Standard MCMC methods are also slow, computationally **too costly** for large **data sets**.

Scalable approaches: variational Bayes, probabilistic numerics methods, Vecchia approximation, distributed GP, other sparse/low rank approximation of the covariance/precision matrix (e.g. banding),...

Bayes vs. frequentist statistics

Bayes vs. Frequentist

Statistical model: Data Y is generated by $\mathcal{P} = \{P_f : f \in \Theta\}$.

Schools: **Frequentist**

Bayes

Model: $Y \sim P_{f_0}, f_0 \in \Theta$ $f \sim \Pi$ (prior), $Y|f \sim P_f$

Goal: Recover f_0 : Update our belief about f :

Estimator $\hat{f}(Y)$ **Posterior:** $f|Y$

Bayes vs. Frequentist

Statistical model: Data Y is generated by $\mathcal{P} = \{P_f : f \in \Theta\}$.

Schools: **Frequentist**

Bayes

Model: $Y \sim P_{f_0}, f_0 \in \Theta$ $f \sim \Pi$ (prior), $Y|f \sim P_f$

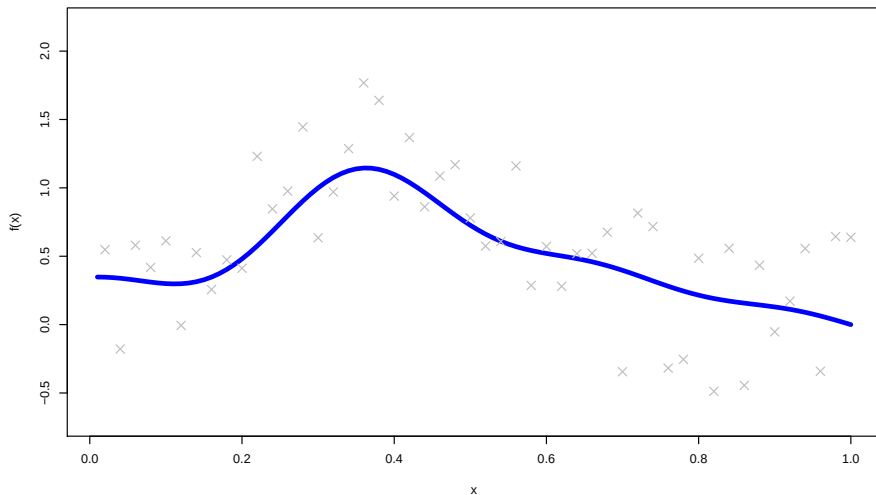
Goal: Recover f_0 : Update our belief about f :

Estimator $\hat{f}(Y)$ **Posterior:** $f|Y$

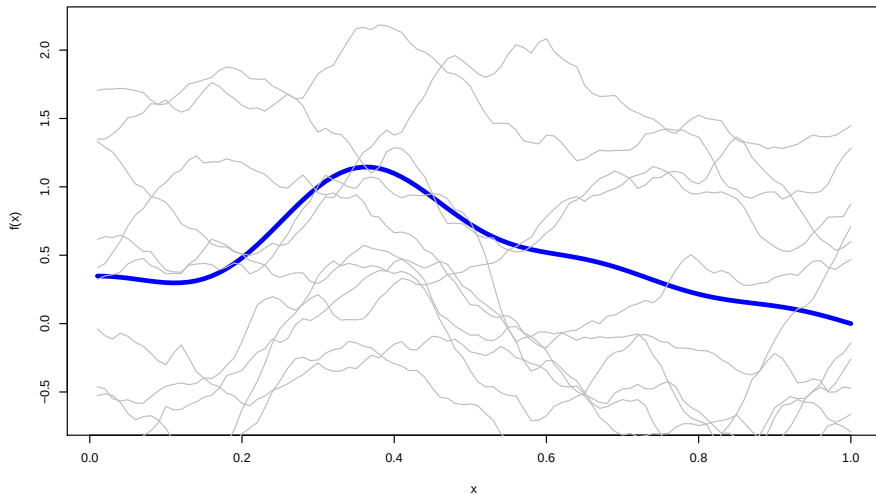
Frequentist Bayes

Investigate **Bayesian** techniques from **frequentist perspective**, i.e. assume that there exists a true f_0 and investigate the behaviour of the posterior $\Pi(\cdot|Y)$.

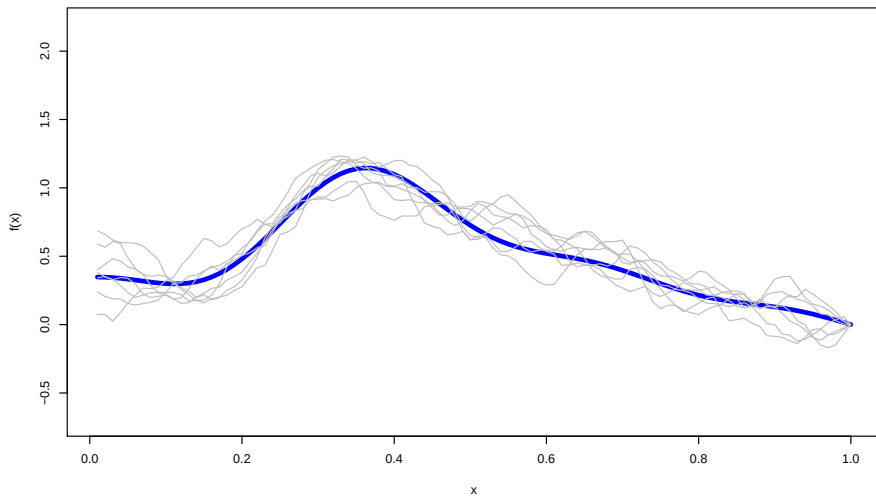
Nonparametric regression



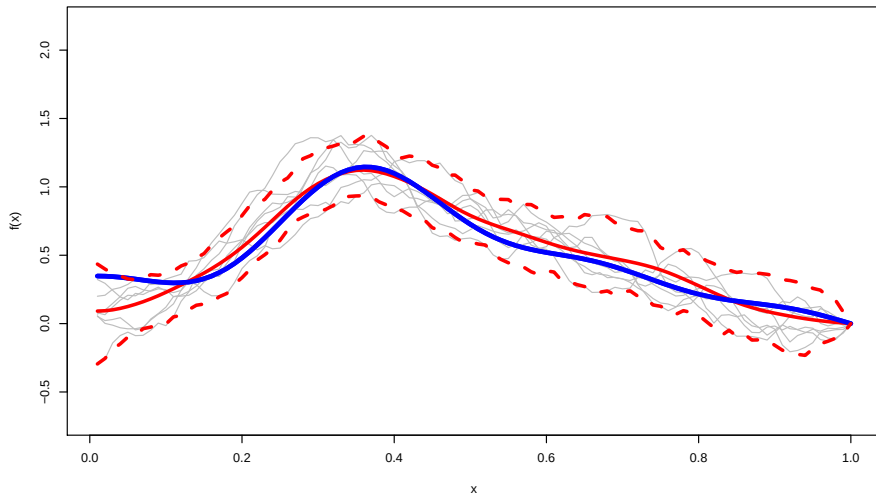
Prior



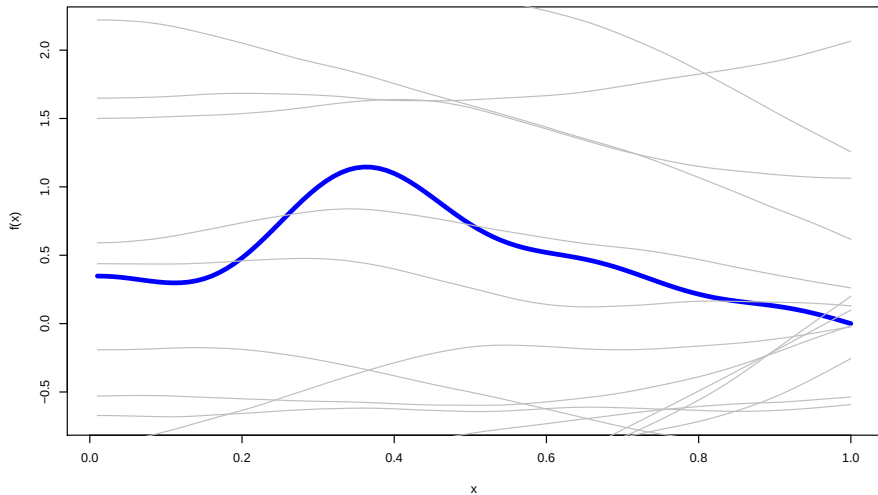
Posterior



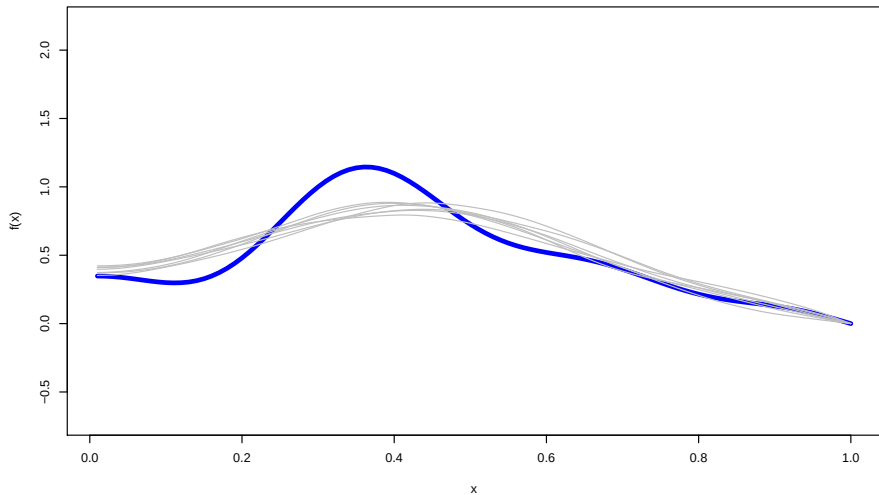
Posterior



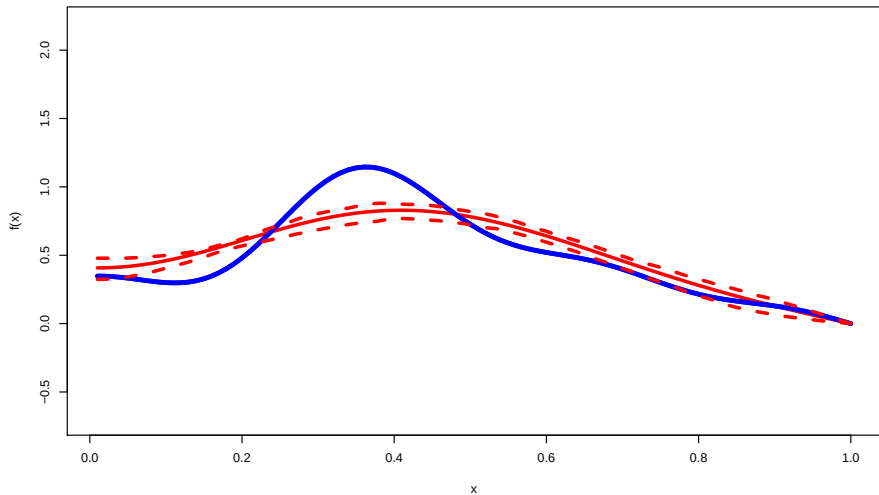
Prior: over-smoothing



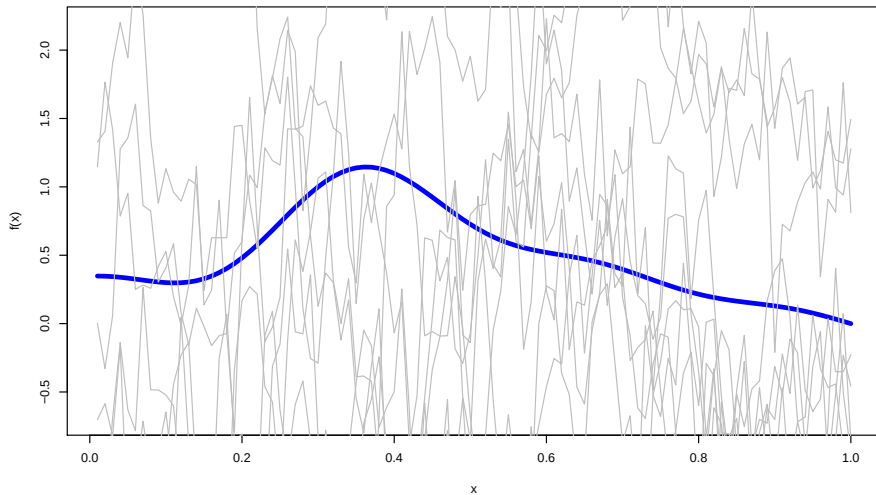
Posterior: over-smoothing



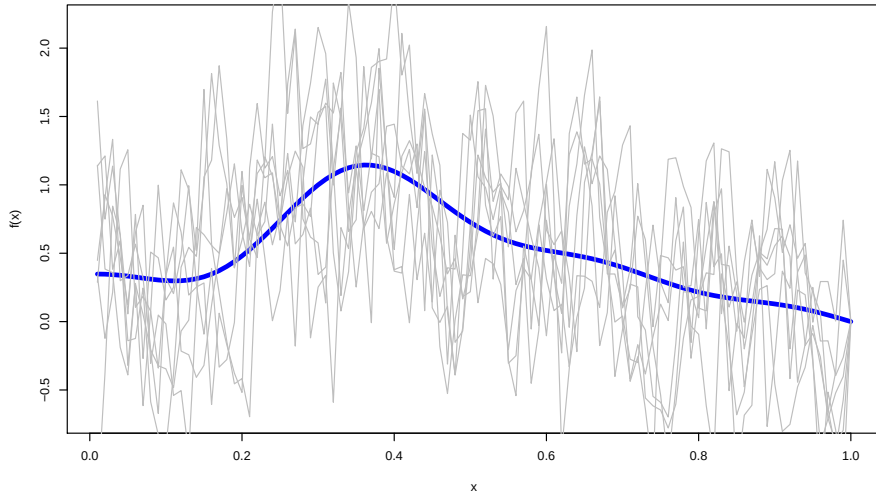
Posterior: over-smoothing



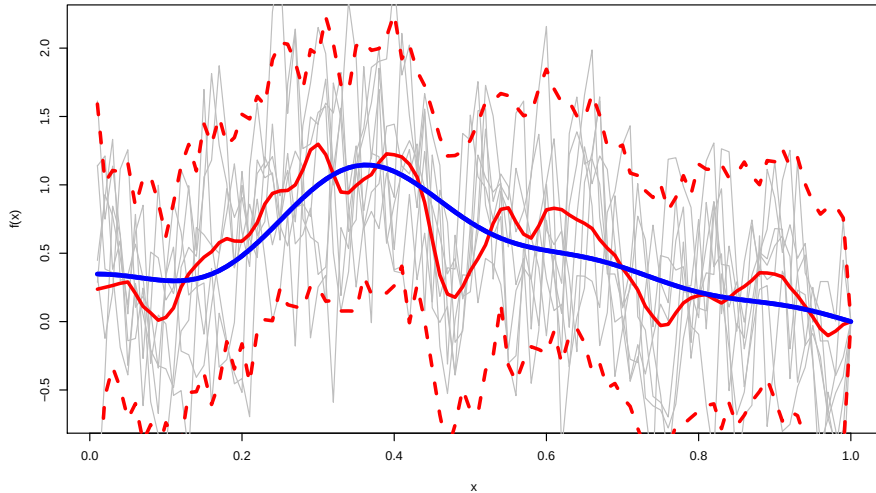
Prior: under-smoothing



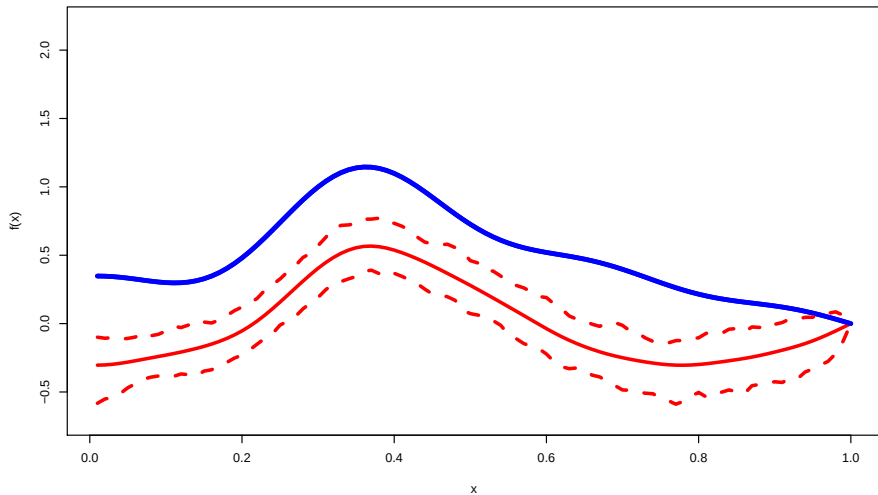
Posterior: under-smoothing



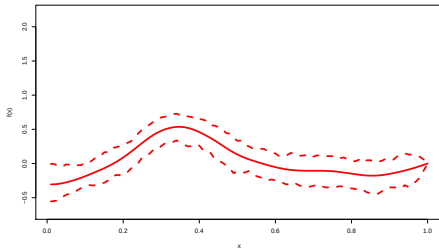
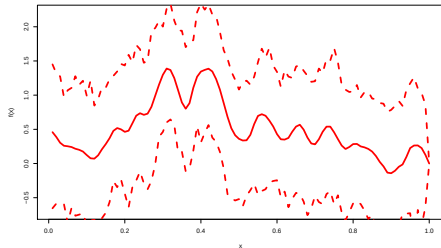
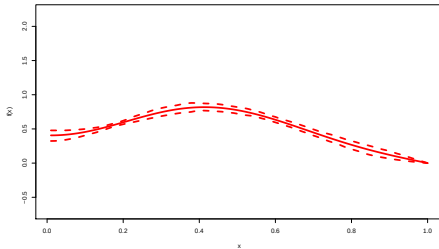
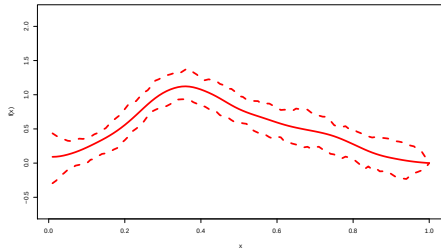
Posterior: under-smoothing



Posterior: misspecified



Which one is correct?



Frequentist Bayes

Posterior consistency: for all $\epsilon > 0$

$$\Pi(f : \|f - f_0\| \leq \epsilon | Y^{(n)}) \xrightarrow{P_{f_0}} 1.$$

Frequentist Bayes

Posterior consistency: for all $\epsilon > 0$

$$\Pi(f : \|f - f_0\| \leq \epsilon | Y^{(n)}) \xrightarrow{P_{f_0}} 1.$$

Posterior contraction rate: The **fastest** $\epsilon_n > 0$:

$$\Pi(f : \|f - f_0\| \leq \epsilon_n | Y^{(n)}) \xrightarrow{P_{f_0}} 1$$

Frequentist Bayes

Posterior consistency: for all $\epsilon > 0$

$$\Pi(f : \|f - f_0\| \leq \epsilon | Y^{(n)}) \xrightarrow{P_{f_0}} 1.$$

Posterior contraction rate: The **fastest** $\epsilon_n > 0$:

$$\Pi(f : \|f - f_0\| \leq \epsilon_n | Y^{(n)}) \xrightarrow{P_{f_0}} 1$$

Uncertainty quantification: Can we get reliable **uncertainty quantification**, i.e. does it hold for $\hat{C} = \{f : \|f - \hat{f}\| \leq \rho_n\}$ with $\Pi(\hat{C} | Y^{(n)}) = 0.95$ that

$$P_{f_0}(f_0 \in \hat{C}) \geq 0.95?$$

Recovery and UQ for GPs

Recovery: Minimax contraction rate $\epsilon_n = n^{-\beta/(1+2\beta)}$ (up to $\log n$ factor) for β -smooth functions for a wide range of (optimally scaled) GP priors (e.g. squared exponential, Matérn, integrated BM,...) and models (regression, classification, density estimation, regression of graphs), see van der Vaart and van Zanten (2008), Sz. et al. (2013).

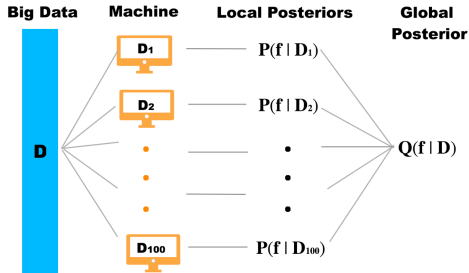
Uncertainty quantification: Reliable uncertainty quantification (under additional regularity assumption for data driven tuning of GPs), Sz. et al. (2015), Rousseau and Sz. (2020)

Scaling up Gaussian Processes

Distributed methods

Distributed Bayes:

Distributed Bayes:



Product of Experts

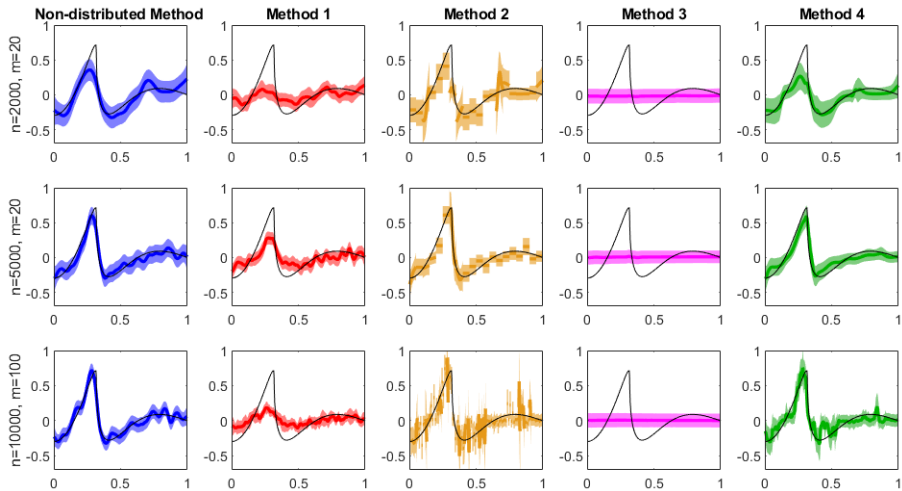
Mixture of Experts

Data segregation:
Posterior aggregation:

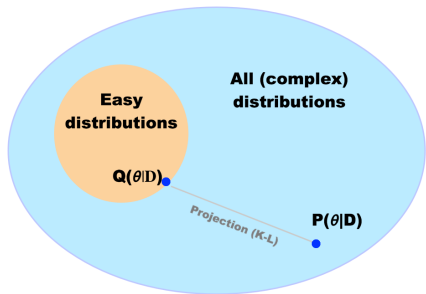
randomly
“averaging”

local blocks
“sticking together”

Distributed GP



Variational Bayes



- In VB propose a family of **tractable** distributions Q for θ .
- **Trade-off**: simple vs complex class $\iff$ speed vs accuracy.
- Solve the following **optimization** problem:

$$\begin{aligned} Q^* &= \arg \min_{Q \in \mathcal{Q}} \text{KL}(Q || \Pi(\cdot | Y)) \\ &= \arg \max_{Q \in \mathcal{Q}} E_Q \log(p(\theta, X)) - E_Q \log(q(\theta)) \end{aligned}$$

e.g. using gradient descent, coordinate ascent.

Variational GP

Variational class: using inducing variables method, see Titsias (2009):

- Take $u_1, \dots, u_m$ linear functionals of f (e.g. $u_i = a_1 f(z_1) + a_2 f(z_2)$ for some $z_1, z_2 \in \mathcal{X}$).
- Then $f|(u_1, \dots, u_m) \sim GP$

$$\begin{aligned}x &\mapsto K_{xu} K_{uu}^{-1} \mathbf{u}, \\(x, z) &\mapsto k(x, z) - K_{xu} K_{uu}^{-1} K_{uz}.\end{aligned}$$

where $K_{xu} = \text{cov}_{\Pi}(f(x), \mathbf{u}) = K_{ux}^T$ and $K_{uu} = [\text{cov}_{\Pi}(u_i, u_j)]_{1 \leq i, j \leq m}$.

Variational GP (cont)

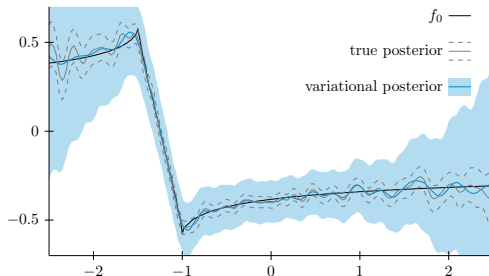
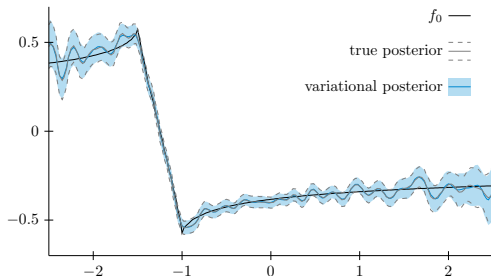
- $Q_{\mu, \Sigma} \in \mathcal{Q} : f|(u_1, \dots, u_m)$ wrt $(u_1, \dots, u_m) \sim N(\mu, \Sigma)$. These are GPs, with

$$\begin{aligned}x &\mapsto K_{xu} K_{uu}^{-1} \mu, \\(x, z) &\mapsto k(x, z) - K_{xu} K_{uu}^{-1} (K_{uu} - \Sigma) K_{uu}^{-1} K_{uz},\end{aligned}$$

Remarks:

- **Exists** optimal μ', Σ' Titsias (2009).
- $Q^* = Q_{\mu', \Sigma'}$ is a particular **rank- m approximation** of $\Pi(\cdot | \mathbf{x}, \mathbf{y})$.
- Upper bound for the expected (wrt the prior) **KL divergence** between $Q_{\mu', \Sigma'}$ and $\Pi(f | \mathbf{x}, \mathbf{y})$, see Burt et al. (2020), optimal contraction rates Nieman et al (2022, 2026) and UQ Nieman et al (2023).

VB: GP with rescaled SE



Number of inducing variables (for $f_0(x) = |x + 1|^\beta - |x + 3/2|^\beta$, $n = 5000$, $x_i \sim N(0, 1)$, $\sigma = 0.2$, $\beta = 0.8$, GP with rescaled SE, Method 2):

$m=80$

$m=40$

General theory for Variational Bayes

Theorem (Ray and Sz (2022))

Suppose there exists $C > 0$ and $M_n \rightarrow \infty$ such that

$$E_{f_0} \Pi(f \notin \Theta_n | Y) 1_A \leq C e^{-M_n}$$

for an event A . Then for any distribution Q ,

$$E_{f_0} Q(f \notin \Theta_n) 1_A \leq \frac{2}{M_n} \left[E_{f_0} KL(Q || \Pi(\cdot | Y)) 1_A + C e^{-M_n/2} \right].$$

Remarks:

- Apply previous with $\Theta_n = \{f : \|f - f_0\|_2 \leq C\epsilon_n\}$
- Generally $KL(Q || \Pi(\cdot | Y)) \rightarrow 0$ is **not required** for optimal **convergence**.

Note: Similar (but more involved) results were derived in Zhang and Gao (2020).

Proof

The duality formula for the KL divergence Boucheron et al (2013)

$$KL(Q\|P) = \sup_f \left[\int f dQ - \log \int e^f dP \right].$$

Hence

$$\int f(\theta) dQ(\theta) \leq KL(Q\|\Pi(\cdot|Y)) + \log \int e^{f(\theta)} d\Pi(\theta|Y).$$

Proof

The duality formula for the KL divergence Boucheron et al (2013)

$$KL(Q\|P) = \sup_f \left[\int f dQ - \log \int e^f dP \right].$$

Hence

$$\int f(\theta) dQ(\theta) \leq KL(Q\|\Pi(\cdot|Y)) + \log \int e^{f(\theta)} d\Pi(\theta|Y).$$

Applying this with $f(\theta) = (M_n/2)1_{\Theta_n^c}(\theta)$, and using $\log(1+x) \leq x$, we get

$$\begin{aligned} (M_n/2)Q(\theta \in \Theta_n^c)1_A &\leq KL(Q\|\Pi(\cdot|Y))1_A + \log(1 + \Pi(\theta \in \Theta_n^c|Y)e^{M_n/2})1_A \\ &\leq KL(Q\|\Pi(\cdot|Y))1_A + e^{M_n/2}\Pi(\theta \in \Theta_n^c|Y)1_A. \end{aligned}$$

Proof

The duality formula for the KL divergence Boucheron et al (2013)

$$KL(Q\|P) = \sup_f \left[\int f dQ - \log \int e^f dP \right].$$

Hence

$$\int f(\theta) dQ(\theta) \leq KL(Q\|\Pi(\cdot|Y)) + \log \int e^{f(\theta)} d\Pi(\theta|Y).$$

Applying this with $f(\theta) = (M_n/2)1_{\Theta_n^c}(\theta)$, and using $\log(1+x) \leq x$, we get

$$\begin{aligned} (M_n/2)Q(\theta \in \Theta_n^c)1_A &\leq KL(Q\|\Pi(\cdot|Y))1_A + \log(1 + \Pi(\theta \in \Theta_n^c|Y)e^{M_n/2})1_A \\ &\leq KL(Q\|\Pi(\cdot|Y))1_A + e^{M_n/2}\Pi(\theta \in \Theta_n^c|Y)1_A. \end{aligned}$$

Conclude by taking expectations wrt P_{θ_0} and using assumption $E_{\theta_0}\Pi(\theta \in \Theta_n^c|Y)1_A \leq Ce^{-M_n}$.

Vecchia approximation

Joint distribution with conditionals:

$$p(W_{X_1}, \dots, W_{X_n}) = p(W_{X_1}) \prod_{i=2}^n p(W_{X_i} | W_{X_1}, W_{X_2}, \dots, W_{X_{i-1}}).$$

Vecchia approximation: sparsify the conditions

$$p(W_{X_n}) \approx p(W_{X_1}) \prod_{i=2}^n p(W_{X_i} | Z_{\text{pa}(X_i)}),$$

where $\text{pa}(X_i)$ denotes the parents of X_i .

Vecchia GP approximation:

- Mother Gaussian Process
- Directed acyclic graph (DAG) providing the parent sets structure

The number of parents are restricted to m . Computational time is $O(m^3 n)$.

Beyond Gaussian Process regression

Other models

Model (density estimation): $X_1, \dots, X_n \sim^{iid} p$ on $[0, 1]^d$.

Other models

Model (density estimation): $X_1, \dots, X_n \sim^{iid} p$ on $[0, 1]^d$.

Prior: $p(x) = e^{W_x} / \int_{[0,1]^d} e^{W_y} dy$

Other models

Model (density estimation): $X_1, \dots, X_n \sim^{iid} p$ on $[0, 1]^d$.

Prior: $p(x) = e^{W_x} / \int_{[0,1]^d} e^{W_y} dy$

Model (binary classification): $(X_1, Y_1), \dots, (X_n, Y_n)$ random vectors, where $X_i \in [0, 1]^d$ and $Y_i \in \{0, 1\}$, satisfying

$$p(Y_1 = 1 | X_1 = x) = p(x).$$

Other models

Model (density estimation): $X_1, \dots, X_n \sim^{iid} p$ on $[0, 1]^d$.

Prior: $p(x) = e^{W_x} / \int_{[0,1]^d} e^{W_y} dy$

Model (binary classification): $(X_1, Y_1), \dots, (X_n, Y_n)$ random vectors, where $X_i \in [0, 1]^d$ and $Y_i \in \{0, 1\}$, satisfying

$$p(Y_1 = 1 | X_1 = x) = p(x).$$

Prior: $p(x) \sim e^{W_x} / (1 + e^{W_x})$

Other models

Model (density estimation): $X_1, \dots, X_n \sim^{iid} p$ on $[0, 1]^d$.

Prior: $p(x) = e^{W_x} / \int_{[0,1]^d} e^{W_y} dy$

Model (binary classification): $(X_1, Y_1), \dots, (X_n, Y_n)$ random vectors, where $X_i \in [0, 1]^d$ and $Y_i \in \{0, 1\}$, satisfying

$$p(Y_1 = 1 | X_1 = x) = p(x).$$

Prior: $p(x) \sim e^{W_x} / (1 + e^{W_x})$

Model (inverse problem): $(X_1, Y_1), \dots, (X_n, Y_n)$ random vectors, where $X_i \in [0, 1]^d$ and $Y_i \in \mathbb{R}$, satisfying

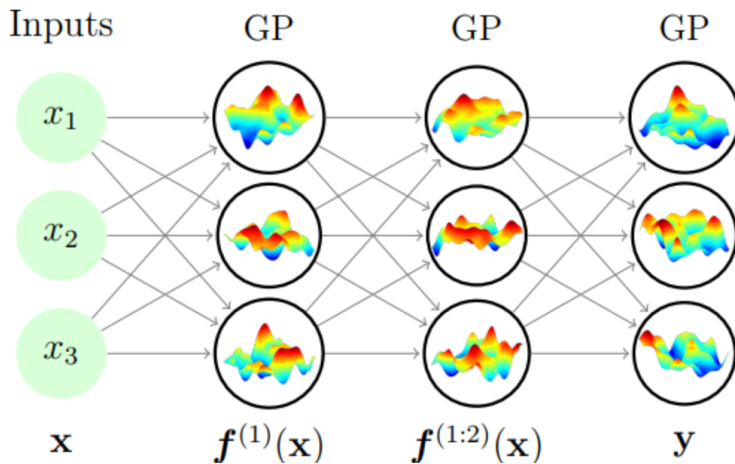
$$Y_i = \mathcal{A}f(X_i) + \varepsilon_i, \quad \varepsilon_i \sim^{iid} N(0, \sigma^2),$$

where $\mathcal{A}$ is a known linear/non-linear operator. Ill-posed inverse problem: the inverse of $\mathcal{A}$ is not continuous.

Prior: $f(x) \sim W_x$.

Posterior: Gaussian if $\mathcal{A}$ is linear, non-Gaussian if $\mathcal{A}$ is non-linear.

Deep GP: Illustration



Titsias & Lawrence (2010), Damianou & Lawrence (2013).

Deep GP: model selection prior

Hierarchical Deep GP construction Finocchio, Schmidt-Hieber (2022):

- ① Put a prior on **composition graph**
 - prior on the number of layers, i.e. **depth**
 - prior on the **width** of the layers
 - prior on compositional **sparsity** (model selection prior)
- ② GP priors on the **edges**
- ③ **Regularization** of sample paths.

Deep GP: application

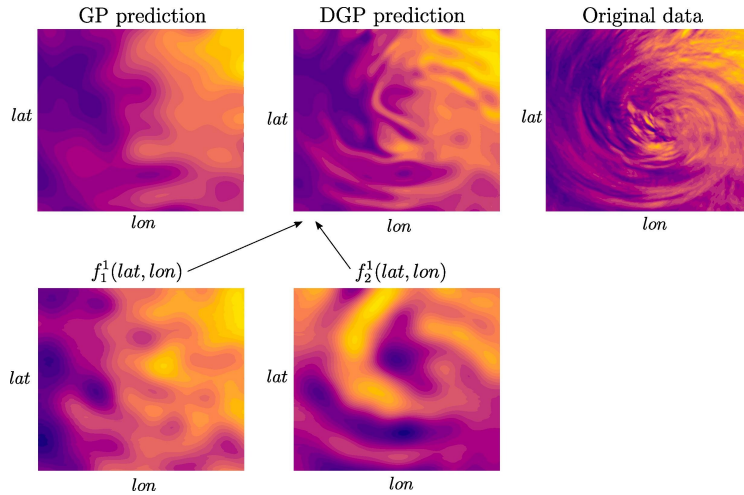


Figure: Modeling a hurricane field with GP vs Deep GP. Svendsen et al (2020): Deep Gaussian processes for biogeophysical parameter retrieval and model inversion

Summary

- GP regression **doesn't scale** well $O(n^3)$.
- Various **scalable** approximations **without theoretical** underpinning, e.g. variational, distributed, Vecchia, low rank
- Theory for these approaches:
 - For well tuned priors and calibrated approximations **optimal rates** can be achieved
 - **Adaptation** is also possible (but one has to be careful).
 - Reliable **uncertainty** quantification: even for VB.

Papers

- K. Ray and B. Szabo (2022) Variational Bayes for high-dimensional linear regression with sparse priors. JASA 117 (539) 1270-1281.
- D. Nieman, B. Szabo, H. van Zanten (2022) Contraction rates for sparse variational approximations in Gaussian process regression. JMLR 23 (205) 1-26.
- D. Nieman, B. Szabo, H. van Zanten (2023) Uncertainty quantification for sparse spectral variational approximations in Gaussian process regression. EJS 17 (2), 2250-2288
- B. Stankewitz, B. Szabo (2026) Contraction rates for conjugate gradient and Lanczos approximate posteriors in Gaussian process regression. Bernoulli (forthcoming)
- H. van Zanten, B. Szabo (2019) An asymptotic analysis of distributed nonparametric methods. JMLR 20 (87), 1-30.
- B. Szabo, A. Hadji, A. van der Vaart (2025) Adaptation using spatially distributed Gaussian Processes. JASA 120 (552) 2784-2795