Scalable methods for Gaussian process regression

Botond Szabó (Bocconi University)

SFB Colloquium

Potsdam, Germany, 18. 10. 2024.

Co-authors

Harry van Zanten
(VU Amsterdam)

Dennis Nieman
(Humboldt U. Berlin)

Bernhard Stankewitz
(Potsdam)

Also: Aad van der Vaart (Delft), Amine Hadji (Leiden), Thibault Randrianarisoa (Toronto), Yichen Zhu (Bocconi)

Outline

- Introduction: Gaussian Processes and GP regression
- Approximation methods for GP
- Frequentist Bayesian paradigm
- Variational Bayes for GP
- Iterative GP methods
- Summary

Gaussian Processes and applications

Gaussian processes

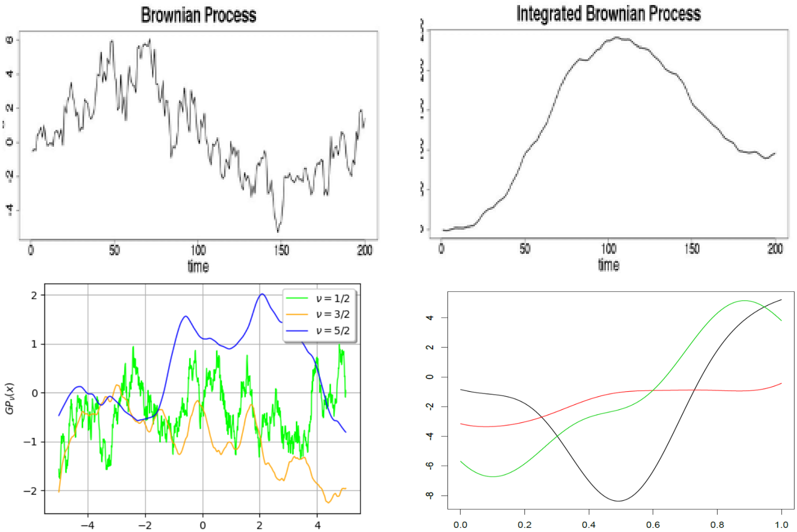
Gaussian Process: A stochastic process $(W_t : t \in T)$ is called Gaussian if all its finite-dimensional marginals are multivariate-normally distributed.

They are determined by their mean function $m : T \to \mathbb{R}$ and covariance function $r : T \times T \to \mathbb{R}$:

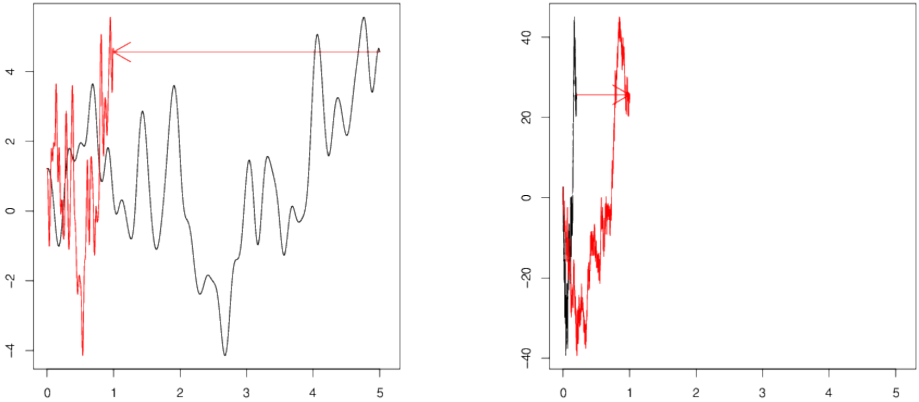
- $m(t) = EW_t, \quad t \in T,$
- $r(t, s) = cov(W_s, W_t).$

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

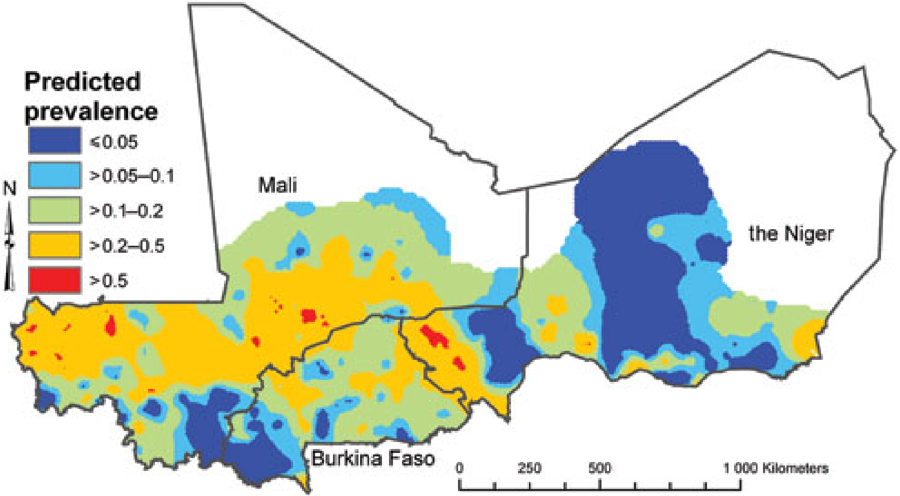
GP examples (BM, IBM, Matern, SE)



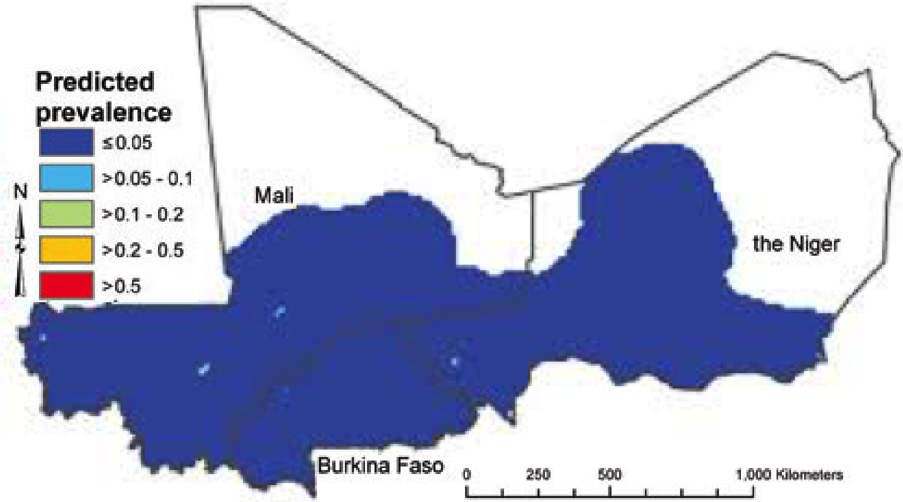
Recaling of GP



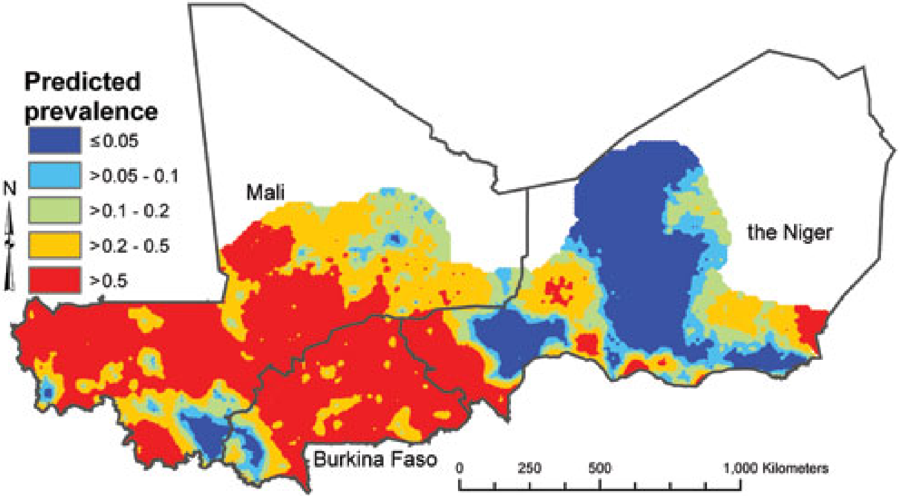
Prediction of infection: posterior mean



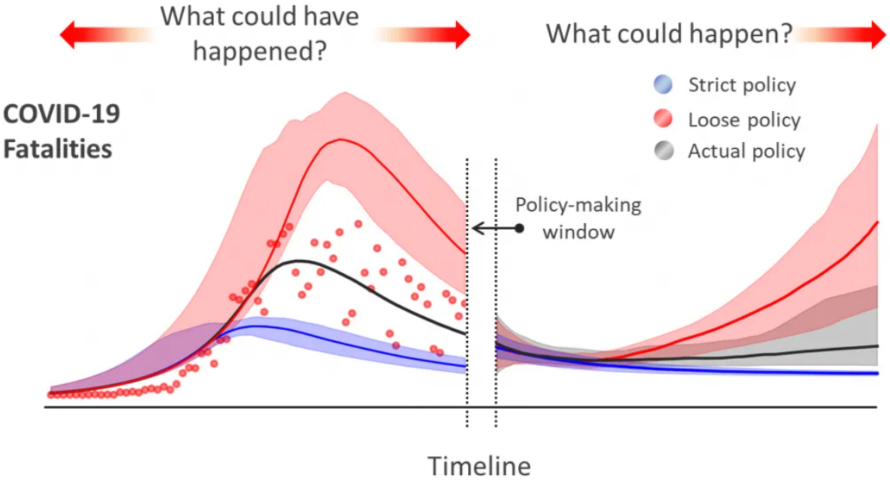
Prediction of infection: lower credible band



Prediction of infection: upper credible band



Covid: impact of policy



Gaussian Process Regression

Gaussian process regression

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

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

where f_0 is the unknown function of interest.

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

Gaussian process regression

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

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

where f_0 is the unknown function of interest.

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

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

$$\begin{aligned} x &\mapsto K_{x\boldsymbol{f}}(\sigma^2 I + K_{\boldsymbol{ff}})^{-1}\boldsymbol{y}, \\ (x,z) &\mapsto k(x,z) - K_{x\boldsymbol{f}}(\sigma^2 I + K_{\boldsymbol{ff}})^{-1}K_{\boldsymbol{f}z}, \end{aligned}$$

Here we denote $\boldsymbol{y} = (y_1, \dots, y_n)^T$, $\boldsymbol{f} = (f(x_1), \dots, f(x_n))^T$,
$K_{x\boldsymbol{f}} = \mathrm{cov}_\Pi(f(x), \boldsymbol{f}) = (k(x,x_1), \dots, k(x,x_n))$, $K_{\boldsymbol{ff}} = \mathrm{cov}_\Pi(\boldsymbol{f}, \boldsymbol{f}) = [k(x_i, x_j)]_{1\le i,j\le n}$.

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.

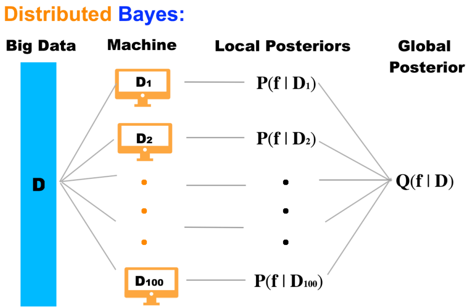
Problem: Standard MCMC methods are also slow, computationally too costly for large data sets.

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

Scaling up Gaussian Processes

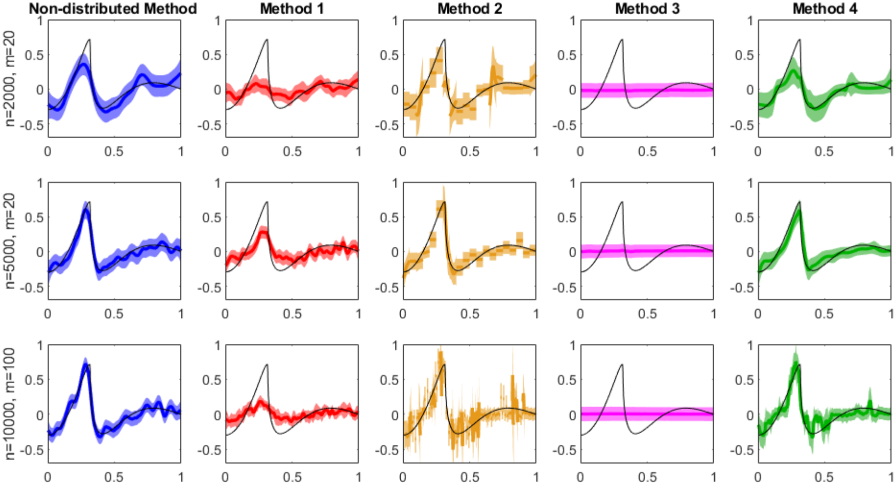
Distributed methods

Distributed Bayes:

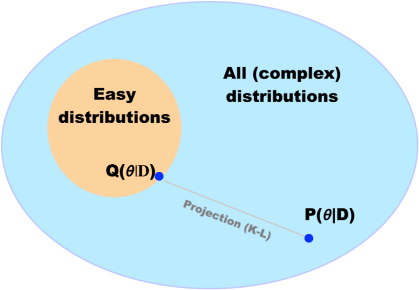


	Product of Experts	Mixture of Experts
Data segregation:	randomly	local blocks
Posterior aggregation:	"averaging"	"sticking together"

Distributed GP



Variational Bayes



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

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

e.g. using gradient descent, coordinate ascent.

Vecchia approximation

Joint distribution with conditionals:

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

Vecchia approximation: sparsify the conditions

$$p(W_{\mathcal{X}_n}) \approx p(W_{X_1}) \prod_{i=2}^{n} p(W_{X_i} | Z_{\mathrm{pa}(X_i)}),$$

where $\mathrm{pa}(X_i)$ 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 $O(m^3 n)$.

Probabilistic numerical methods

- **Computation aware GPs:** methods from probabilistic numerics.
- **Idea:** represent uncertainty resulting from limited computational resources
- **Goal:** learning representer weights $W^* = K_\sigma^{-1}\boldsymbol{y}$.
- **Examples of methods:** Lanczos iteration, conjugate gradient descent.
- **Software:** GPyTorch Gardner et al (2018).

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), $Y \mid f \sim P_f$
Goal:	Recover f_0: Estimator $\hat{f}(Y)$	Update our belief about f: Posterior: $f \mid Y$

Bayes vs. Frequentist

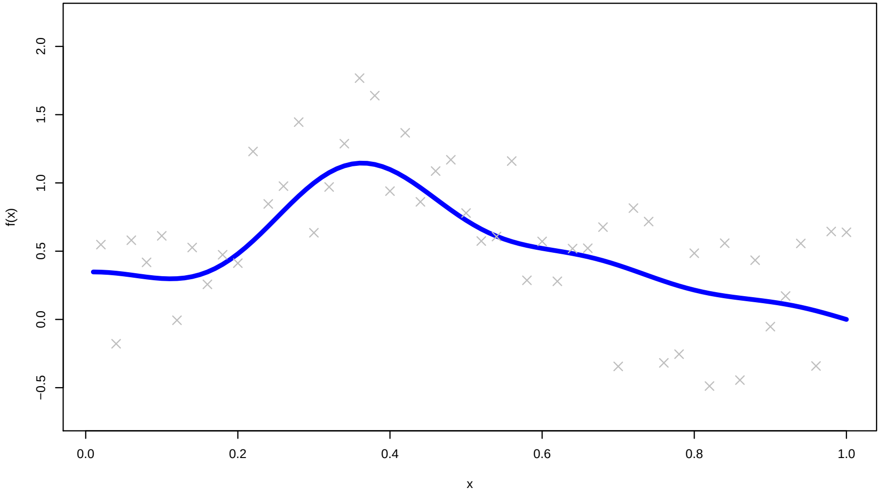
Statistical model: Data Y is generated by $\mathcal{P} = \{P_f : f \in \Theta\}$.

Schools:	**Frequentist**	**Bayes**
Model:	$Y \sim P_{f_0}, f_0 \in \Theta$	$f \sim \Pi$ (prior), $Y \mid f \sim P_f$
Goal:	Recover f_0: Estimator $\hat{f}(Y)$	Update our belief about f: Posterior: $f \mid Y$

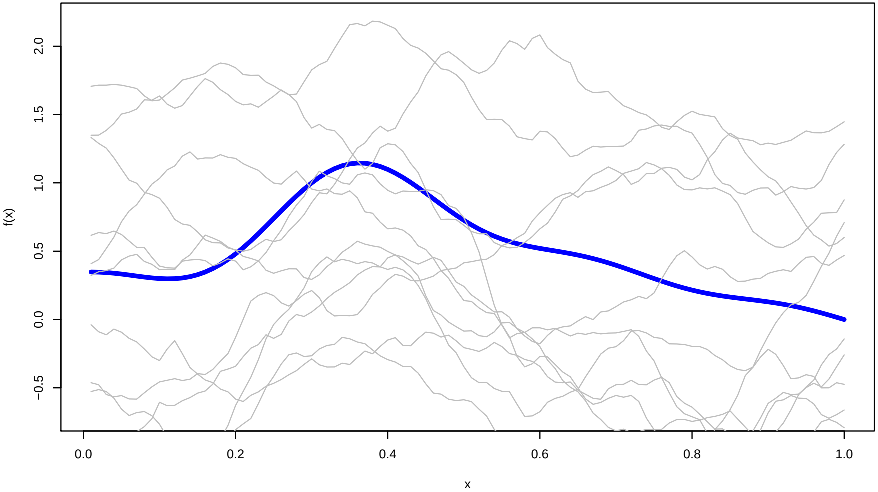
Frequentist Bayes

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

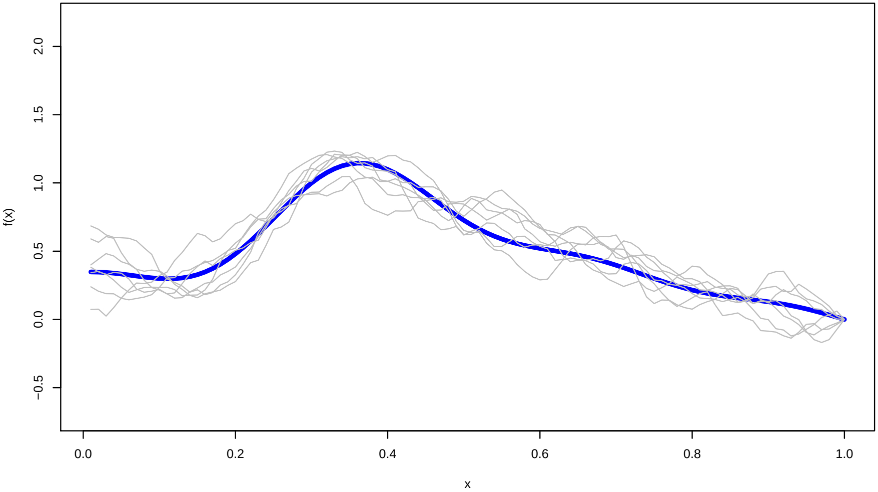
Nonparametric regression



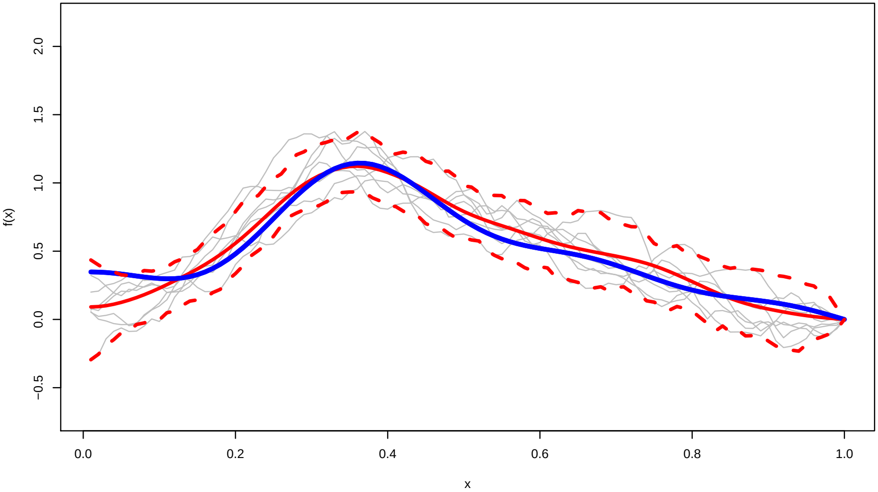
Prior



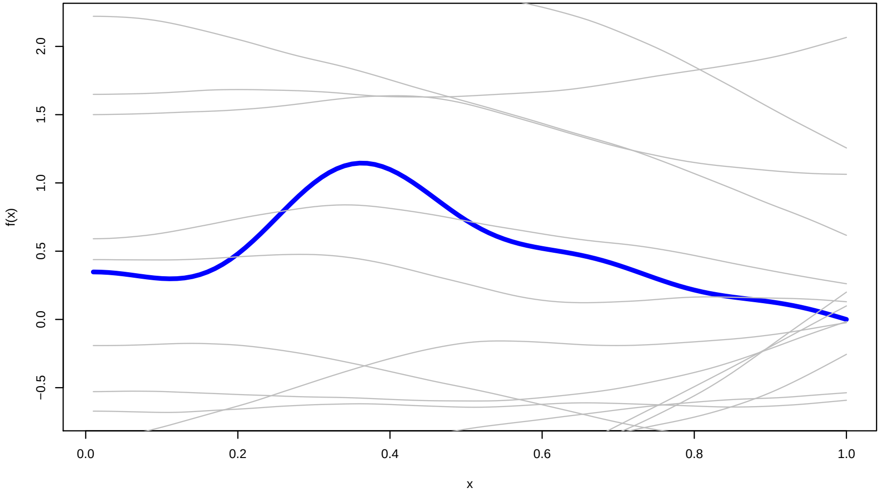
Posterior



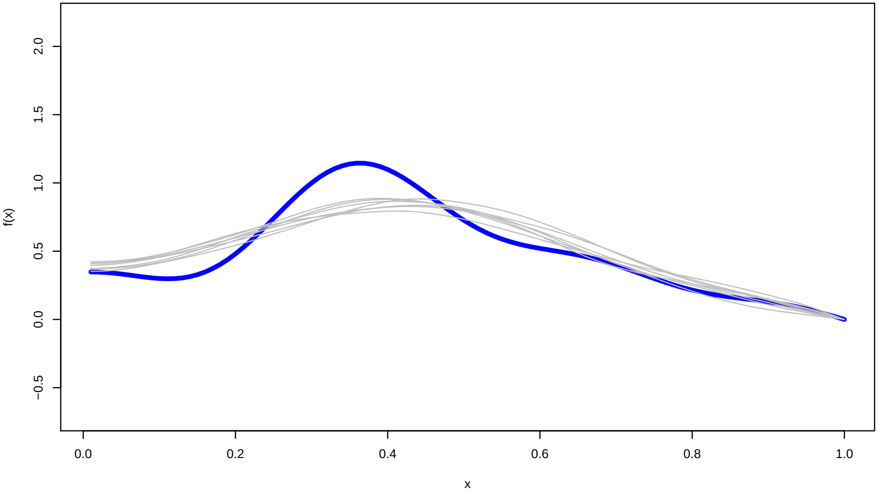
Posterior



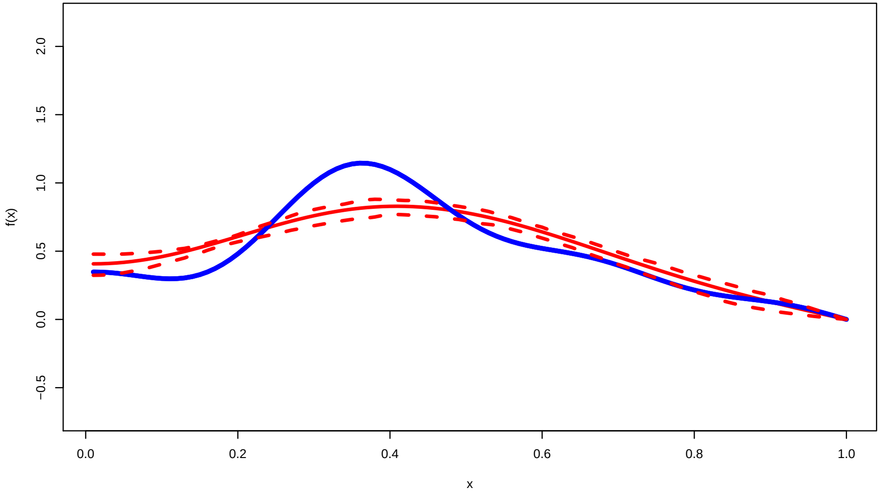
Prior: over-smoothing



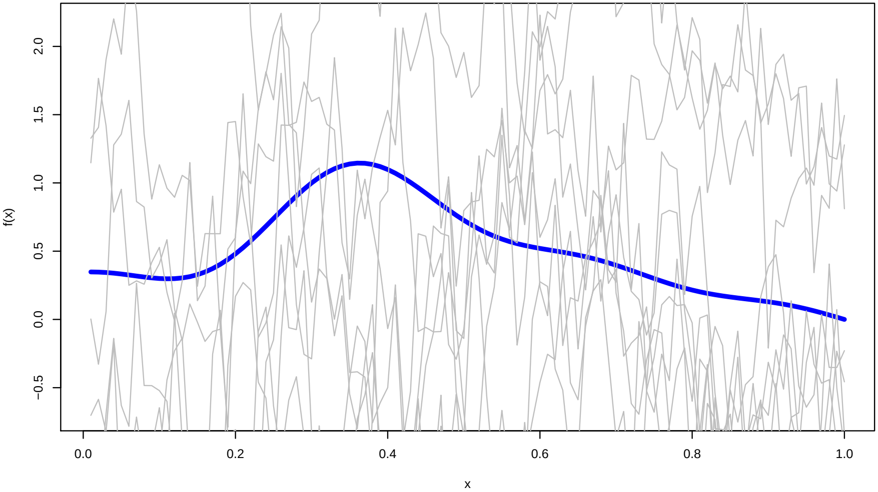
Posterior: over-smoothing



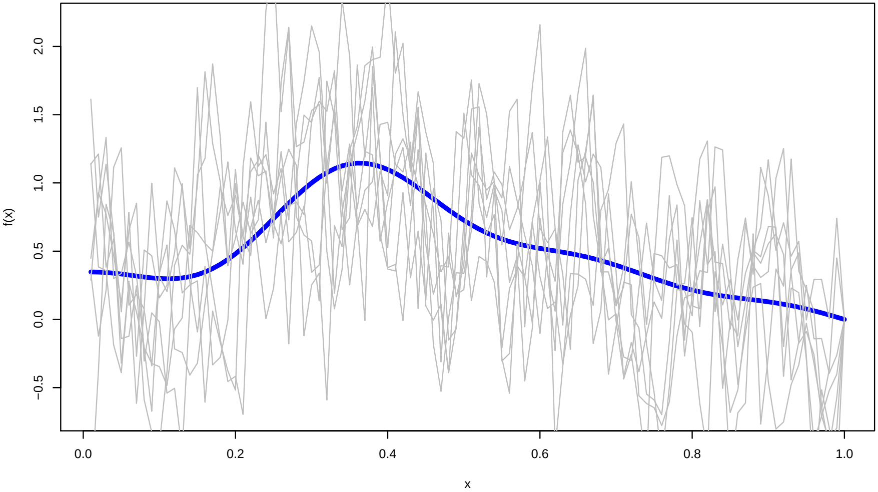
Posterior: over-smoothing



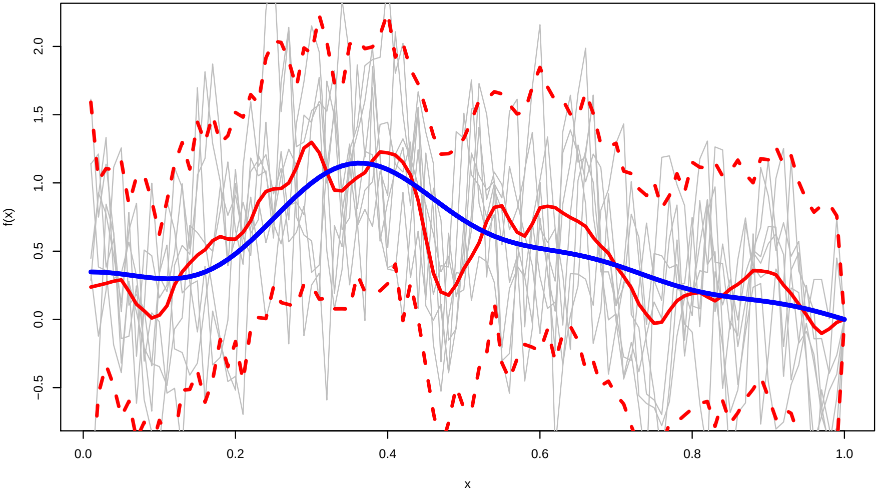
Prior: under-smoothing



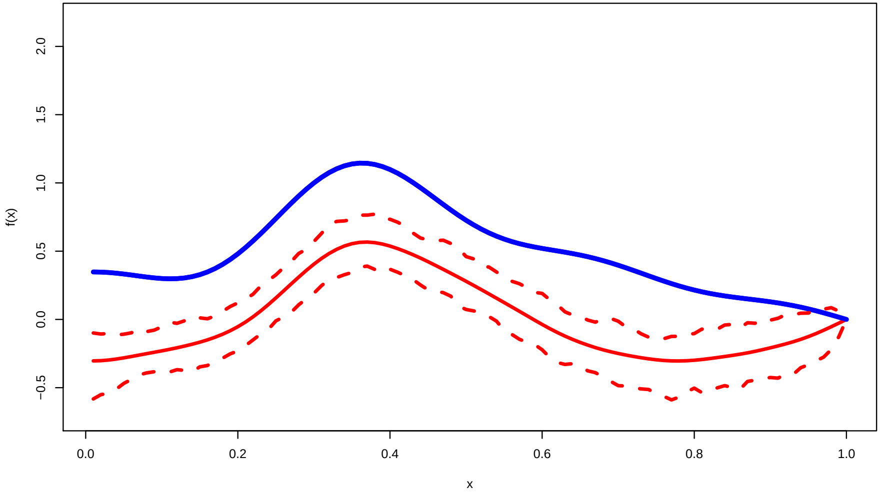
Posterior: under-smoothing



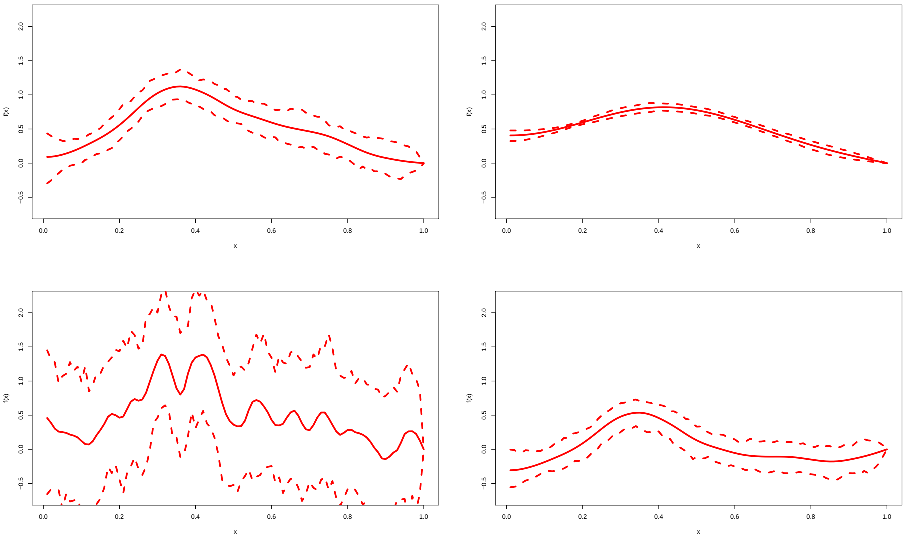
Posterior: under-smoothing



Posterior: misspecified



Which one is correct?



Frequentist Bayes

Posterior consistency: for all $\epsilon > 0$

$$\Pi(f : d_n(f, f_0) \leq \varepsilon | Y^{(n)}) \overset{P_{f_0}}{\to} 1.$$

Frequentist Bayes

Posterior consistency: for all $\epsilon > 0$

$$\Pi(f : d_n(f, f_0) \leq \varepsilon | Y^{(n)}) \xrightarrow{P_{f_0}} 1.$$

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

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

Frequentist Bayes

Posterior consistency: for all $\epsilon > 0$

$$\Pi(f : d_n(f, f_0) \leq \varepsilon | Y^{(n)}) \overset{P_{f_0}}{\to} 1.$$

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

$$\Pi(f : d_n(f, f_0) \leq \epsilon_n | Y^{(n)}) \overset{P_{f_0}}{\to} 1$$

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

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

Gaussian process regression: theory

Def (Concentration function):

$$\varphi_{f_0}(\epsilon) = \inf_{h \in \mathbb{H} : \|h - f_0\| \leq \epsilon} \|h\|_{\mathbb{H}}^2 - \log \Pi(f : \|f\| \leq \epsilon).$$

Gaussian process regression: theory

Def (Concentration function):

$$\varphi_{f_0}(\epsilon) = \inf_{h\in\mathbb{H}:\|h-f_0\|\leq\epsilon} \|h\|_{\mathbb{H}}^2 - \log\Pi(f:\ \|f\|\leq\epsilon).$$

Theorem [van der Vaart & van Zanten (2008)] If $\varphi_{f_0}(\epsilon_n)\leq n\epsilon_n^2$ and $h(p_f,p_g)\leq\|f-g\|$ holds then

$$\Pi\big(h(p_f,p_{f_0})\leq M\epsilon_n|\boldsymbol{x},\boldsymbol{y}\big)\overset{P_{f_0}}{\to}1.$$

Application: 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). Also uncertainty quantification and adaptation.

Variational Bayes for Gaussian Processes

Variational GP

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

- Take $u_1, ..., u_m$ 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 $f|(u_1, ..., u_m) \sim GP$

$$x \mapsto K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} \boldsymbol{u},$$
$$(x, z) \mapsto k(x, z) - K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} K_{\boldsymbol{u}z}.$$

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

Variational GP (cont)

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

$$x \mapsto K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} \mu,$$
$$(x, z) \mapsto k(x, z) - K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} (K_{\boldsymbol{uu}} - \Sigma) K_{\boldsymbol{uu}}^{-1} K_{\boldsymbol{u}z},$$

Remarks:

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

Examples: inducing variable methods

Inducing point methods:

- $f(z_1), \ldots, f(z_m)$ with $z_i \in \{x_1, \ldots, x_n\}$. Computational complexity $O(m^2 n)$ after selecting the points z_i.

Examples: inducing variable methods

Inducing point methods:

- $f(z_1), ..., f(z_m)$ with $z_i \in \{x_1, ..., x_n\}$. Computational complexity $O(m^2 n)$ after selecting the points z_i.

Population spectral features method:

- $u_j = \int f \psi_j dG_x$, for ψ_j the eigenfunction of the covariance function k. Computational complexity: $O(m^2 n)$.

Examples: inducing variable methods

Inducing point methods:

- $f(z_1), ..., f(z_m)$ with $z_i \in \{x_1, ..., x_n\}$. Computational complexity $O(m^2 n)$ after selecting the points z_i.

Population spectral features method:

- $u_j = \int f \psi_j dG_x$, for ψ_j the eigenfunction of the covariance function k. Computational complexity: $O(m^2 n)$.

Sample spectral features method:

- $u_j = [f(x_1), ..., f(x_n)]\hat{u}_j$, where $\hat{u}_j$ is the jth eigenvector of $K_{\mathbf{ff}}$. Computational complexity: $O(mn^2)$.

Theory for VB GP regression

General theory for Variational Bayes

Theorem (Ray and Sz (2022))

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

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

for an event A. Then for any distribution Q,

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

Remarks:

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

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

VB posterior contraction

Theorem: For $f_0 : \mathcal{X} \mapsto \mathbb{R}$ assume that

$$(CondGP) \quad \varphi_{f_0}(\epsilon_n) \leq n\epsilon_n^2$$

$$(CondVB) \quad E_X tr(R_{\boldsymbol{ff}}) \leq Cn\epsilon_n^2, \; E_X \|R_{\boldsymbol{ff}}\| \leq C.$$

Then

$$Q^*(h(p_f, p_{f_0}) \leq M_n\epsilon_n) \stackrel{P_{\theta_0}}{\to} 1,$$

where $R_{\boldsymbol{ff}}$ is the covariance matrix of $\boldsymbol{f}|\boldsymbol{u}$.

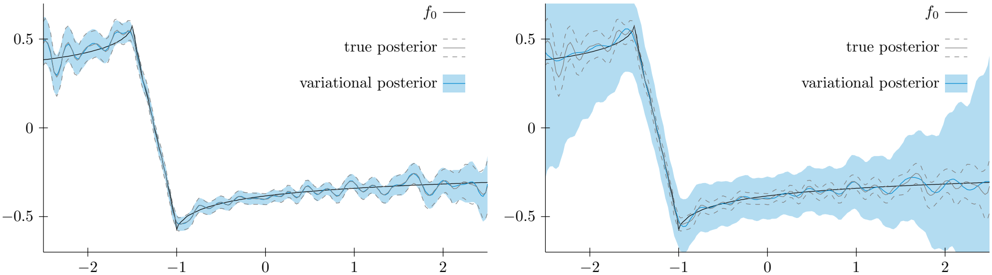
Examples: minimax contraction rates

- For $f_0 \in C^\beta([0,1]^d)$, β-Matérn covariance kernel, and $m \geq n^{\frac{d}{d+2\beta}}$ the contraction rate is $n^{-\beta/(d+2\beta)}$ for the population spectral features method.

- For $f_0 \in C^\beta([0,1])$, squared exponential kernel (with rescaling parameter $b = n^{-1/(1+2\beta)}$), and $m \geq n^{\frac{1}{1+2\beta}}$ the contraction rate is $n^{-\beta/(1+2\beta)}(\log n)^{5/4}$ both for the sample and population spectral features methods.

- For $f_0 \in S^\beta([0,1]^d)$, β-regular sequence prior $\Pi = \sum_{j=1}^\infty j^{-1/2-\beta} Z_j \psi_j$, $Z_j \overset{iid}{\sim} N(0,1)$ and $m \geq n^{\frac{d}{d+2\beta}}$, the posterior mean concentrates with rate $n^{-\beta/(d+2\beta)}$ for the DPP-inducing points method.

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

Iterative, probabilistic numeric methods

Learning the representer weight

Goal: compute representer weights $W^* = K_\sigma^{-1}\boldsymbol{y}$, with $K_\sigma = (K_{\boldsymbol{ff}} + \sigma^2 I)$

Initialization: $W^* \sim N(0, K_\sigma^{-1})$, where $w_0 = 0$ initially best guess and $\Gamma_0 = K_\sigma^{-1}$ excess uncertainty.

Update: Consider policies $s_j \in \mathbb{R}^n$, $j = 1, ..., n$. Iteratively update W^* based on the information

$$\alpha_j := s_j^\top(\boldsymbol{y} - K_\sigma w_{j-1}) = s_j^\top K_\sigma(W^* - w_{j-1})$$

of the residuals projected into direction of s_j. We obtain $W^*|\alpha_j \sim N(w_j, \Gamma_j)$ with

$$w_j = C_j\boldsymbol{y} \quad \text{and} \quad \Gamma_j = K_\sigma^{-1} - C_j,$$

where C_j is the approx of K_σ^{-1} at the jth iteration.

At iteration m,

$$\Psi_m := \mathbb{P}^{f|W=w} N(w_m, \Gamma_m)(dw),$$

is the Gaussian process with mean and covariance functions

$$x \mapsto k(X,x)^\top C_m \boldsymbol{y},$$
$$(x,x') \mapsto \underbrace{k(x,x') - k(X,x)^\top K_\sigma^{-1} k(X,x')}_{\text{Mathematical uncertainty}} + \underbrace{k(X,x)^\top \Gamma_m k(X,x')}_{\text{Computational uncertainty}}.$$

Remarks:

- Ψ_m is an approximation of the posterior.
- If the policies $(s_j)_{j \le m}$ are lin. indep., then for $m = n$ we have $C_m = K_\sigma^{-1}$.
- Although K_σ^{-1} and Γ_m are computationally prohibitive, the combined uncertainty C_m can be evaluated.

IterGP algorithm Wenger et al. (2022).

Algorithm 1 GP approximation scheme

procedure $\mathrm{ITERGP}(k, X, Y)$
    $C_0 \leftarrow 0 \in \mathbb{R}^{n \times n}$
    for $j = 1, 2, \ldots, m$ **do**
        $s_j \leftarrow \mathrm{POLICY}()$
        $d_j \leftarrow (I - C_{j-1} K_\sigma) s_j$
        $\eta_j \leftarrow s_j^\top K_\sigma d_j$
        $C_j \leftarrow C_{j-1} + \eta_j^{-1} d_j d_j^\top$
    end for
    $\mu_m(\cdot) \leftarrow k(X, \cdot)^\top C_m Y$
    $k_m(\cdot, \cdot) \leftarrow k(\cdot, \cdot) - k(X, \cdot)^\top C_m k(X, \cdot)$
end procedure
return $\mathrm{GP}(\mu_m, k_m)$

Policy examples

(a) $s_j := e_j$, $j \leq m \rightsquigarrow$ partial Cholesky decomposition of K_σ.

(b) $s_j := \widehat{u}_j$, $j \leq m \rightsquigarrow$ SVD of K_σ.

(c) $s_j := \tilde{u}_j$, $j \leq m \rightsquigarrow$ Lanczos approximation.

(b) $s_j := g_j^{CG}$, $j \leq m \rightsquigarrow$ *CG* applied to $K_\sigma v = Y$.

[1] J. Wenger et al. "Posterior and computational uncertainty in Gaussian processes.". In: *Advances in Neural Information Processing Systems* (2022).

Empirical eigenvector actions

Idea: Consider the SVD of the kernel matrix $K = K_{\mathbf{ff}} = \big(k(x_i, x_j)\big)_{i,j}$:

$$K = \sum_{j=1}^{n} \widehat{\mu}_j \widehat{u}_j \widehat{u}_j^\top. \tag{1}$$

Lemma For the eigenvector actions $s_j := \widehat{u}_j$, $j \leq m$, in IterGP, the approximation C_m of K_σ^{-1} is given by

$$C_m = \sum_{j=1}^{m} \frac{1}{\widehat{\mu}_j + \sigma^2} \widehat{u}_j \widehat{u}_j^\top. \tag{2}$$

Remark Equivalent to the empirical spectral features inducing variables VB method.

Lanczos eigenvector actions

Problem: empirical eigenvector actions are expensive to compute.

Solution: use numerical approximations for the eigenvectors. Standard approach for SVD is Lanczos algorithm (sparse.linalg.svds from SciPy)

Lanczos method: Orthogonal projection method based on the Krylov space

$$\mathcal{K}_m := span\{v_0, Kv_0, \dots, K^{m-1}v_0\}, \qquad m = 1, 2, \dots, n, \tag{3}$$

with initial vector $v_0 \in \mathbb{R}^n$ with $\|v_0\| = 1$.

The approximation of the empirical eigenpairs $(\hat{\mu}_j, \hat{u}_j)_{j\leq m}$ are given by the eigenpairs $(\tilde{\mu}_j, \tilde{u}_j)_{j\leq m}$ of VV^TKVV^T, where $V = [v_1, ..., v_m]$ consisting an orthonormal basis of $\mathcal{K}_m$.

Algorithm 1 Lanczos algorithm

procedure ITERLanczos($K, v_0, \tilde{m}$)
    Initialize v_0 with $\|v_0\| = 1$.
    Compute ONB $v_1, \ldots, v_m$ of $\mathcal{K}_m$.
    $V \leftarrow (v_1, \ldots, v_m)$.
    Compute eigenpairs $(\tilde{\mu}_j, \tilde{u}_j)_{j \leq m}$ of $VV^\top KVV^\top$.
end procedure
return $(\tilde{\mu}_j, \tilde{u}_j)_{j \leq m}$.

Lemma For actions $s_j = \tilde{u}_j$, $j \leq m$, the approximation of K_σ^{-1} in the Lanczos version of IterGP is given by

$$C_m = \sum_{j=1}^{m} \frac{1}{\tilde{\mu}_j + \sigma^2} \tilde{u}_j \tilde{u}_j^\top . \tag{4}$$

Theoretical guarantees

Corollary For $f \in C^{\beta}([0,1]^d)$, for optimally tuned GP priors the corresponding posteriors approximated either by the Lanczos iteration and CG method achieve the minimax contraction rate $n^{-\beta/(d+2\beta)}$ if the number of iterations exceed $m_n \geq \log(n) n^{d/(2\beta+d)}$.

Summary

- GP regression doesn't scale well $O(n^3)$.
- Various scalable approximations without theoretical underpinning, e.g. variational, iterative, distributed, Vecchia, low rank
- Theory for these approaches:
  - For well tunned 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 (2024) Contraction rates for conjugate gradient and Lanczos approximate posteriors in Gaussian process regression. Arxiv
- 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 (2024) Adaptation using spatially distributed Gaussian Processes. Arxiv.