Vecchia Gaussian Processes: Probabilistic Properties, Minimax Rates and Methodological Developments

Botond Szabó (Bocconi University)

Workshop on Gaussian processes and related topics

Toulouse, France, 09. 07. 2025.

Co-authors

Yichen Zhu
(Bocconi University)

Ismael Castillo
(Sorbonne University)

Thibault Randrianarisoa
(Vector Institute)

Outline

- Gaussian Processes regression
- Scaling up GPs
- Vecchia approximation for GPs
  - Connection with polynomial interpolation
  - Construction of DAG based on norming sets
  - Probabilistic properties
  - Statistical properties (estimation)
- Deep Gaussian Processes (DGP)
- Vecchia approximation for DGPs
- Summary

Gaussian Process Regression

Gaussian process regression

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

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

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), $\ell = 1, ..., n$,

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

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).

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

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

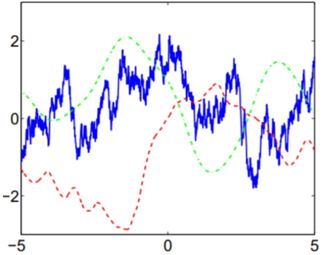
Matérn covariance kernel

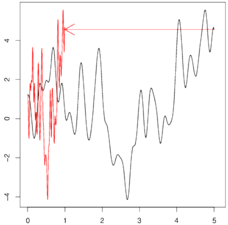
Definition (Matérn) Centered stationary GP $W_t^{\alpha,\tau} = W_{\tau t}^{\alpha}$ with spectral density

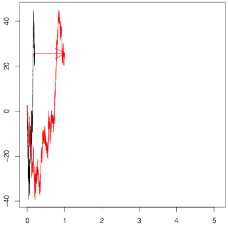
$$\lambda \mapsto \frac{2^d \pi^{d/2} \Gamma(\alpha + d/2)(2\alpha)^\alpha}{\Gamma(\alpha)\tau^{2\alpha}} \left(\frac{2\alpha}{\tau^2} + 4\pi^2 \|\lambda\|^2 \right)^{-(\alpha+d/2)}.$$

Properties:

- Regularity parameter α: sample paths are $\lfloor \alpha \rfloor$-times differentiable ($\alpha \to \infty$ SE).
- Scale parameter τ: shrinking or stretching the paths.







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

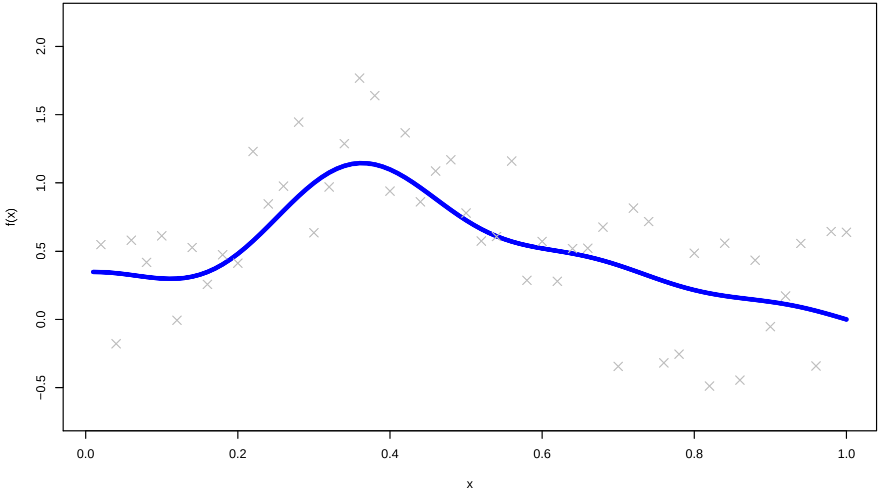
Statistical model: Data Y is generated by $\mathcal{P} = \{P_f : f \in \Theta\}$.

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

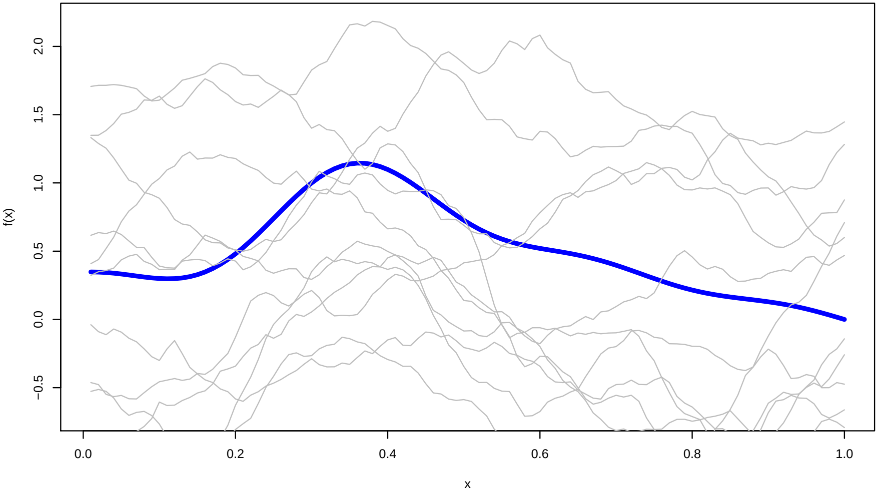
Frequentist Bayes

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

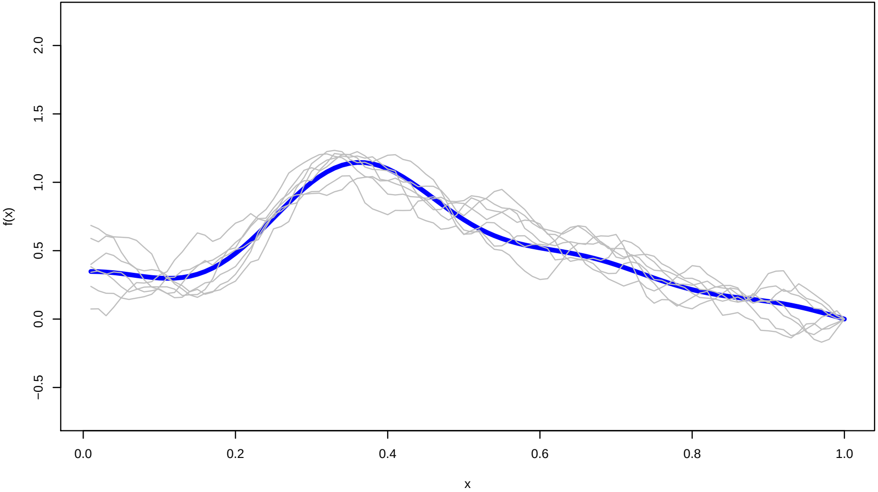
Nonparametric regression



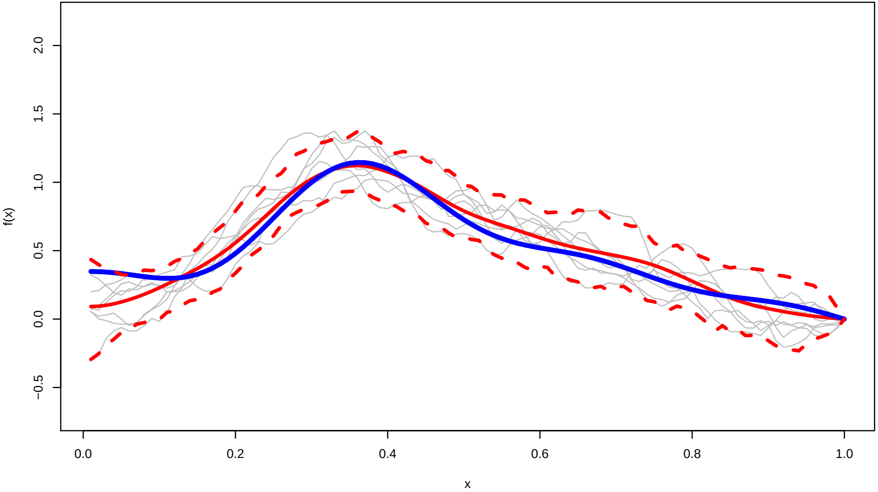
Prior



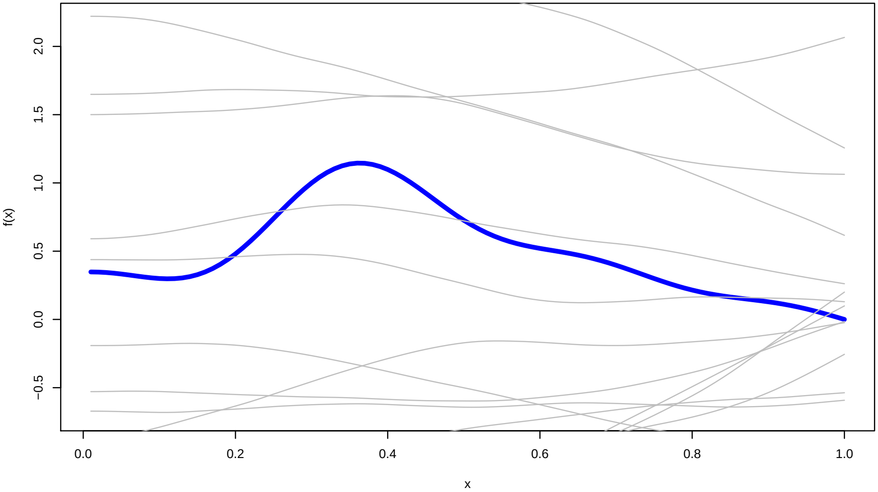
Posterior



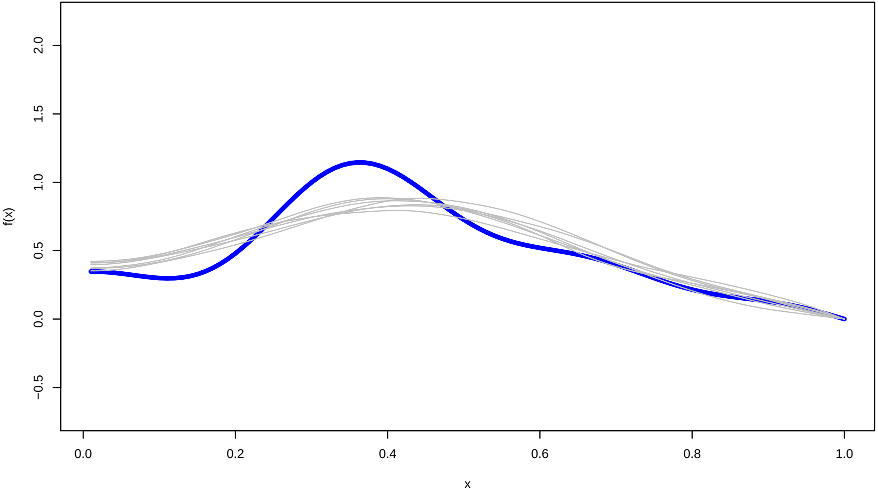
Posterior



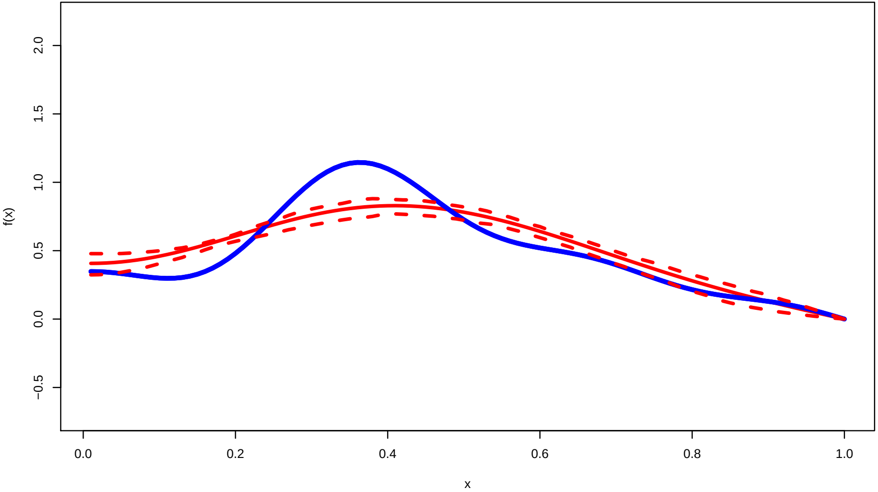
Prior: over-smoothing



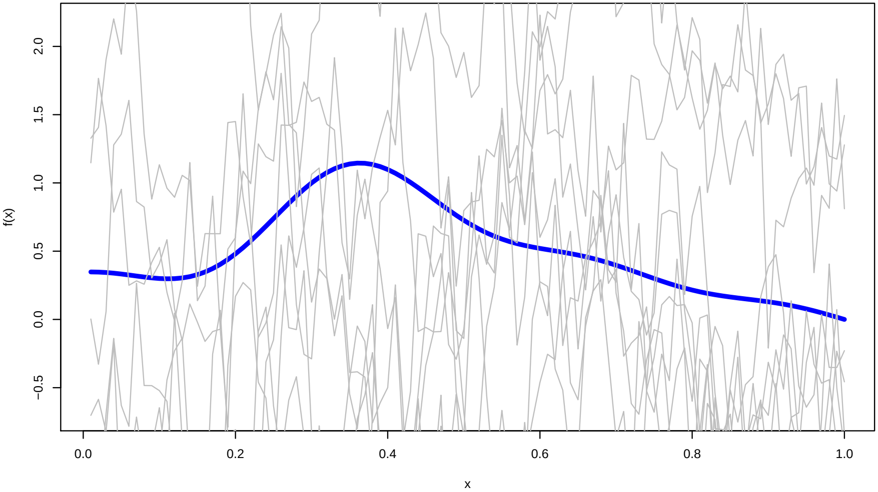
Posterior: over-smoothing



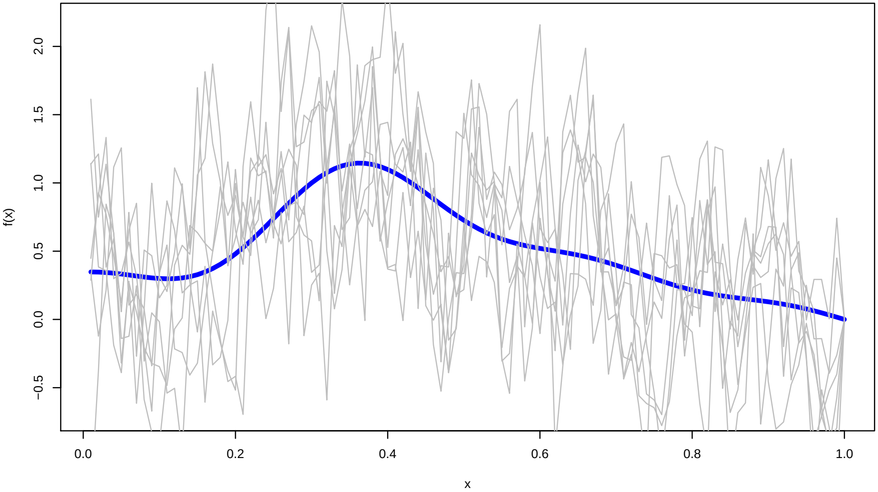
Posterior: over-smoothing



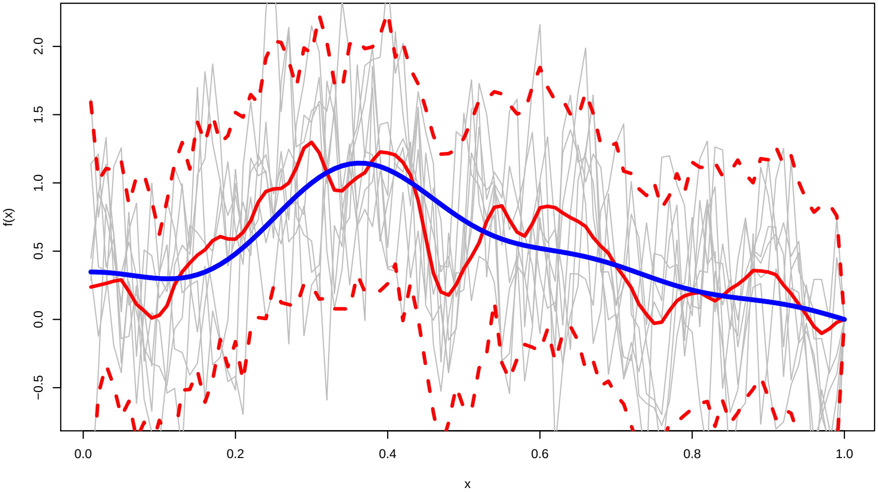
Prior: under-smoothing



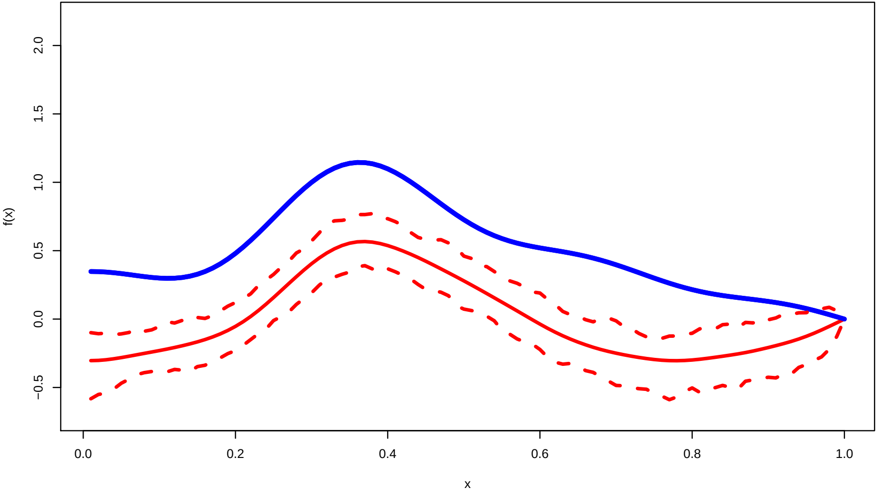
Posterior: under-smoothing



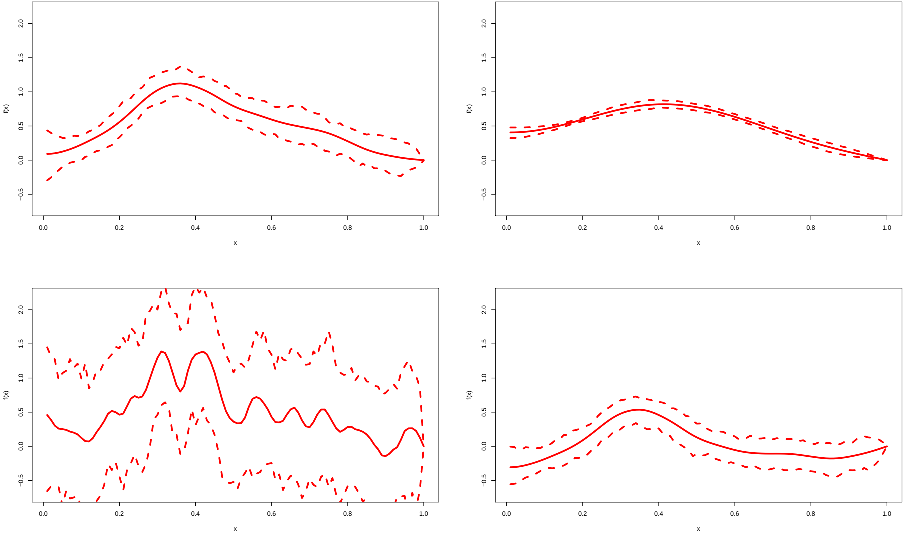
Posterior: under-smoothing



Posterior: misspecified



Which one is correct?



Gaussian process regression: theory

Theorem For $f_0 \in C^\beta$, $\beta > 1/2$ and Matérn process with regularity $\alpha \geq \beta$ and scale parameter either set to $\tau_n = n^{(\alpha-\beta)/(\alpha(d+2\beta))}$ or endowed with a hyper-prior under mild tail condition, the corresponding posterior achieves the minimax contraction rate, i.e.

$$\sup_{f_0 \in C^\beta(M)} E_{f_0} \Pi_n(f : \|f - f_0\|_n \geq M_n n^{-\beta/(d+2\beta)} | Y) \to 0,$$

for arbitrary $M_n \to \infty$.

Remarks:

- One can endow α with hyper-prior, but τ is computationally better.
- General result for GP priors in van der Vaart & van Zanten (2008).
- Similar results for other GPs, e.g. SE, fractional BM, Riemann-Liouville.
- Similar results for other models, e.g. classification, density estimation.

Problem: GP Computation

Conjugacy: the GP 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.

Problem: GP Computation

Conjugacy: the GP 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.

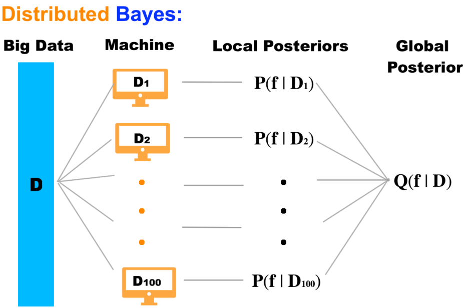
Problem: Standard MCMC methods are also slow, computationally too costly for large data sets.

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

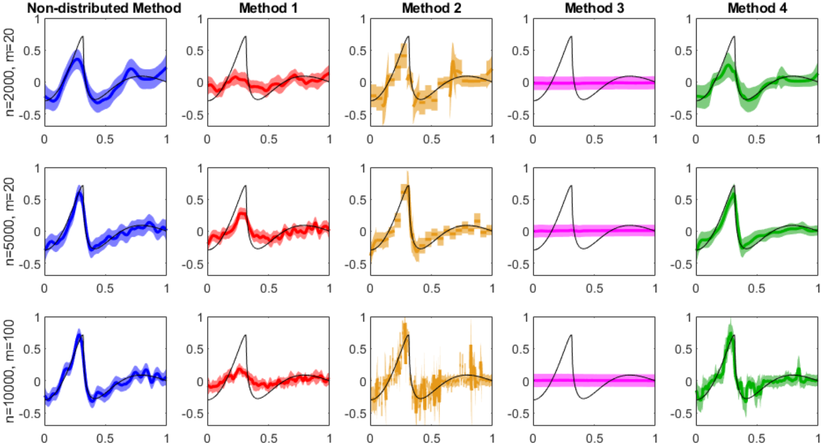
Distributed methods

Distributed Bayes:



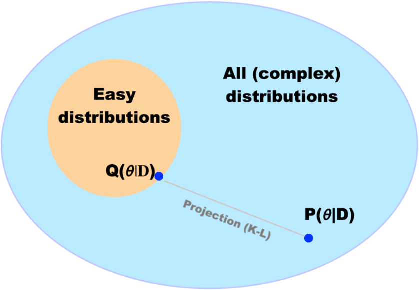
	Product of Experts	Mixture of Experts
Data segregation:	randomly	local blocks
Posterior aggregation:	"averaging"	"sticking together"

Distributed GP



Sz. & van Zanten (2019), Sz., Hadji, vd Vaart (2025)

Variational Bayes



- In VB propose a family of tractable distributions $\mathcal{Q}$ for θ.
- Trade-off: simple vs complex class $\Longleftrightarrow$ 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.

Probabilistic numerics methods

- **Computation aware GPs:** methods from probabilistic numerics, see Wenger et al (2023).
- **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).
- **Theory:** Stankewitz & Sz (2024).

Vecchia approximation of Gaussian Processes

Vecchia Approximations

Consider a mother Gaussian process Z on $\mathcal{X}_n = (X_1, .., X_n)$ with joint density decomposed as

$$p(Z_{\mathcal{X}_n}) = p(Z_{X_1}) \prod_{i=2}^{n} p(Z_{X_i} | Z_{X_j, j<i}).$$

Vecchia Approximations

Consider a mother Gaussian process Z on $\mathcal{X}_n = (X_1, .., X_n)$ with joint density decomposed as

$$p(Z_{\mathcal{X}_n}) = p(Z_{X_1}) \prod_{i=2}^{n} p(Z_{X_i} | Z_{X_j, j<i}).$$

The Vecchia approximations of Gaussian Processes (Vecchia GPs) replace each conditional set $\{X_j, j < i\}$ with a much smaller parent set $\mathrm{pa}(X_i)$

$$p(\hat{Z}_{\mathcal{X}_n}) = p(\hat{Z}_{X_1}) \prod_{i=2}^{n} p(\hat{Z}_{X_i} | \hat{Z}_{\mathrm{pa}(X_i)}),$$

such that

$$[\hat{Z}_{X_i} \mid \hat{Z}_{\mathrm{pa}(X_i)} = z] \overset{d}{=} [Z_{X_i} \mid Z_{\mathrm{pa}(X_i)} = z]$$

Outside of design: for $x \notin \mathcal{X}_n$, $[\hat{Z}_x \mid \hat{Z}_{\mathrm{pa}(x)} = z] \overset{d}{=} [Z_x \mid Z_{\mathrm{pa}(x)} = z]$. $pa(x) \in \mathcal{X}_n$, and $[\hat{Z}_x \mid \hat{Z}_{\mathrm{pa}(x)} = z] \perp\!\!\!\perp [\hat{Z}_y \mid \hat{Z}_{\mathrm{pa}(y)} = z']$.

Methodology: Choose Parent Sets

In view of the joint density of Vecchia Gaussian Processes

$$p(\hat{Z}_{\mathcal{X}_n}) = p(\hat{Z}_{X_1}) \prod_{i=2}^{n} p(\hat{Z}_{X_i} | \hat{Z}_{\mathrm{pa}(X_i)}),$$

and $|\mathrm{pa}(X_i)| \leq m$ the evaluation of this density is $O(nm^3)$.

The principles to choose parent sets are

- The parent sets have bounded cardinality.
- Good approximation property.

Methodology: Choose Parent Sets

There lacks clear guidance on choosing the parent sets:

- **Geometric properties:**
  - It is intuitive to choose close neighbors for parent sets, featured by NNGP Datta et al. (2016).
  - But remote locations are also used, particularly in maximin ordering Katzfuss (2021).
  - It is even proposed to randomly permute dataset before choosing parent sets Guinness (2018).
- **Cardinality m:**
  - Based on theories regarding approximation error, choose $m \asymp (\log n)^b$ for some constant $b > 0$ Schafer et al. (2021), Zhu et al (2024);
  - In practice, m is chosen in adhoc way.

Our Contributions

- **Methodology:**
  - **Problem**: Unclear guidance for choosing parent sets.
  - **Contribution**: Propose Norming Sets as parents, with $m = O(1) \Rightarrow O(n)$ computational complexity.
- **Probabilistic Properties:**
  - **Contribution**: Systematically study the Vecchia GPs as standalone stochastic processes, uncover local polynomial-like behaviors of Vecchia GPs. Derive small deviation bounds.
- **Statistical Theory:**
  - **Problem**: No statistical guarantees for Vecchia GPs.
  - **Contribution**: Prove minimax optimality and adaptation for Vecchia GPs using Norming Sets as parents.

Comparison: a Stationary GP versus a Vecchia GP

Matern GP:

- Stationary GP, marginal distributions have closed form.
- The small ball probability, and subsequently posterior contraction rates, are obtained from studying the RKHS.

Comparison: a Stationary GP versus a Vecchia GP

Matern GP:

- Stationary GP, marginal distributions have closed form.
- The small ball probability, and subsequently posterior contraction rates, are obtained from studying the RKHS.

Vecchia GP:

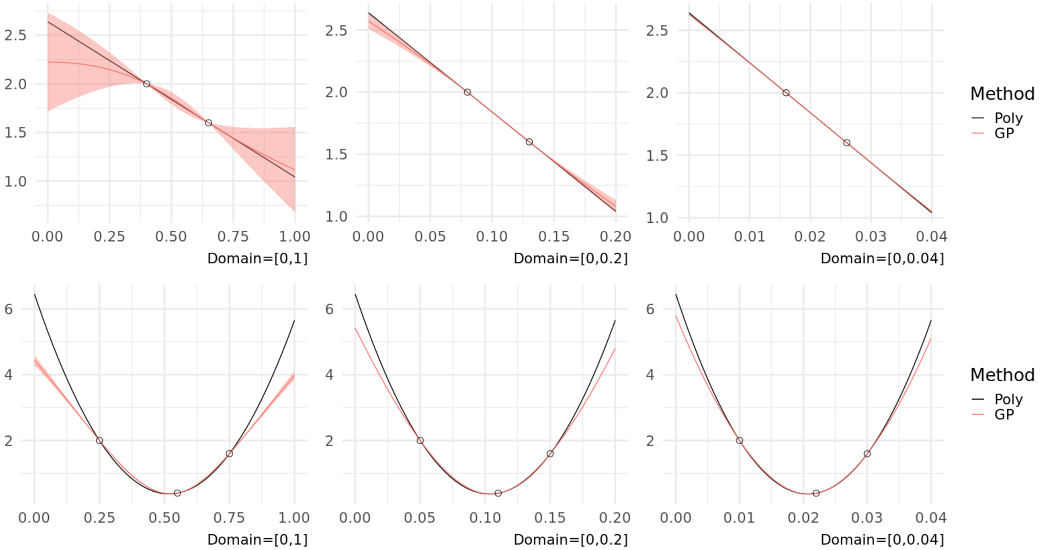
- The joint density is defined through the product of conditionals:

$$p(\hat{Z}_{\mathcal{X}_n}) = p(\hat{Z}_{X_1}) \prod_{i=2}^{n} p(\hat{Z}_{X_i} | \hat{Z}_{\mathrm{pa}(X_i)}),$$

- **No** existing results regarding RKHS, small ball probability, etc.

Key Problem: study the conditional distributions

Connection with polynomial interpolation



Polynomial fit in $d = 1$

Goal: given nodes $A = \{w_1, ..., w_{l+1}\} \subset \mathbb{R}$ and values $z = (z_1, ..., z_{l+1}) \in \mathbb{R}$ fit an l order polynomial $P \in \mathcal{P}_l$, i.e. $P(w_i) = z_i$, for all $i = 1, ..., l+1$.

Solution: There exists a unique solution, called the Lagrange polynomial

$$L(x) = \sum_{j=1}^{l+1} z_j \ell_j(x), \qquad \text{with} \quad \ell_j(x) = \prod_{i \neq j} \frac{x - w_i}{w_j - w_i}.$$

Connection to linear algebra:

$$\begin{bmatrix} 1 & w_1 & w_1^2 & \dots & w_1^l \\ 1 & w_2 & w_2^2 & \dots & w_2^l \\ & & \dots & & \\ 1 & w_{l+1} & w_{l+1}^2 & \dots & w_{l+1}^l \end{bmatrix} \begin{bmatrix} a_0 \\ a_2 \\ \dots \\ a_l \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_{l+1} \end{bmatrix}$$

Polynomial fit in $d \geq 2$

Notations:

- Finite set $A = \{w_1, .., w_m\} \subset \mathbb{R}^d$.
- $\mathscr{P}_l(\mathbb{R}^d)$ the collection of polynomials on $\mathbb{R}^d$ with orders no greater than l.
- $V_A \in \mathbb{R}^{m \times m}$: multidimensional version of the Vandermonde matrix consisting monomials up to order l evaluated at A.
- $v_x \in \mathbb{R}^m$ vector of monomials up to order l evaluated at x.

Lemma (unisolvency): Let $m = \binom{l+d}{l}$ and $z = (z_1, .., z_m)^T \in \mathbb{R}^m$. Then there exists a unique polynomial $P \in \mathcal{P}_l(\mathbb{R}^d)$ satisfying $P(w_i) = z_i$, $i = 1, ..., m$ iff the Vandermonde matrix V_A is invertible. Moreover, this polynomial takes the form

$$P(x) = v_x V_A^{-1} z.$$

See e.g. Wenland (2024).

Norming Sets: definition

Question: When is the Vandermonde matrix V_A invertible?

Definition (Norming set) by Jetter et al (1999):

- For Ω a compact subset of $\mathbb{R}^d$,
- We say a finite set $A = \{w_1, w_2, \cdots, w_m\} \subset \Omega$ is a **norming set** for $\mathscr{P}_l(\Omega)$ with **norming constant** $c_N > 0$ if

$$\sup_{x \in \Omega} |P(x)| \leq c_N \sup_{x' \in A} |P(x')|, \; \forall P \in \mathscr{P}_l(\Omega). \tag{1}$$

Lemma: the Vandermonde matrix V_A is invertable iff A is a norming set.

Conditional expectation

The expectation of conditional distribution is

$$\mathbb{E}[\hat{Z}_{X_i} \mid \hat{Z}_{\mathrm{pa}(X_i)} = z] = z^T K^{-1}_{\mathrm{pa}(X_i),\mathrm{pa}(X_i)} K_{X_i,\mathrm{pa}(X_i)}.$$

Let $r := \mathrm{diam}(\mathrm{pa}(X_i))$ and $l = \underline{\alpha}$.

Lemma Under the condition that parent set is a norming set,

$$\left\| K^{-1}_{\mathrm{pa}(X_i),\mathrm{pa}(X_i)} K_{\mathrm{pa}(X_i),x_i} - V^{-1}_{\mathrm{pa}(X_i)} v_{X_i} \right\| \lesssim c_N \left(r^{2(\alpha - \underline{\alpha})} + r \right).$$

Flat Limit: Gaussian interpolation approximately polynomial interpolation.

Posterior spread: controlled by the approximation error of Gaussian interpolation with polynomial interpolations.

Norming Sets

Consider the first order polynomial space on $[0,1]^2$ as $\mathscr{P}_1([0,1]^2) = \text{span}\{1, x_1, x_2\}$.
$dim(\mathscr{P}_1([0,1]^2) = 3 \Rightarrow$ norming set has at least three elements.

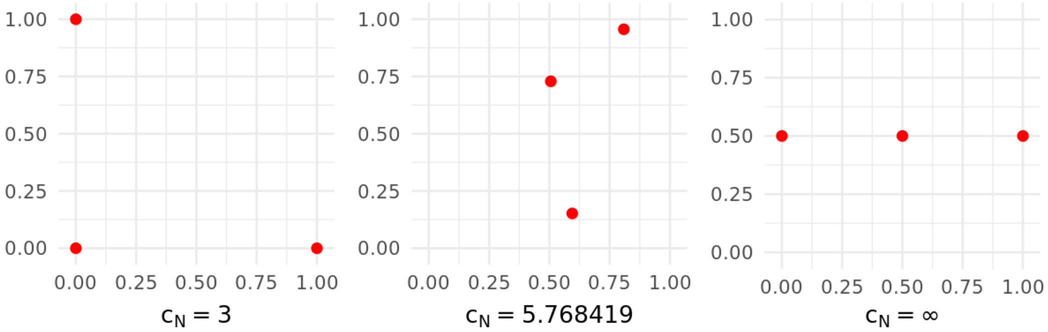


Figure: **Norming constants** w.r.t. $\mathscr{P}_1([0,1]^2)$, for three different sets. "Corner set" Neidinger (2019), random points, non-norming set.

Layered Norming DAGs

Step 1: partition the vertex set $\mathcal{X}_n$ into disjoint layers $\mathcal{N}_0, \mathcal{N}_1, \mathcal{N}_2, \cdots$ (coarse-to-fine).

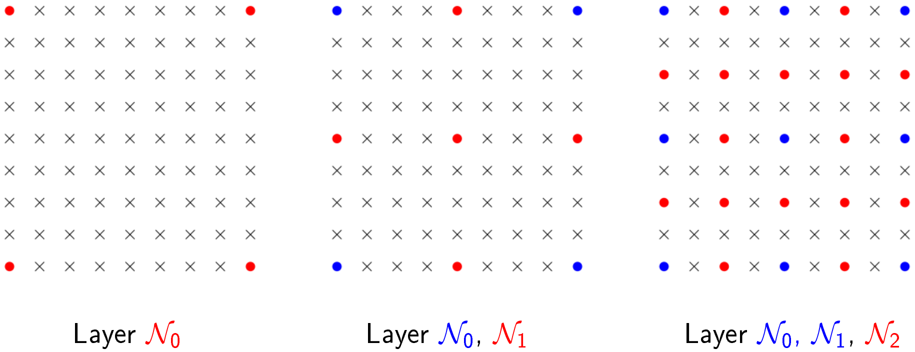


Figure: Illustration of Layers on a 9×9 Grid: Red dots: current layer; Blue dots: all previous layers; Black crosses: all latter layers.

Layered Norming DAGs (cont)

Step 2: for each $X_i \in \mathcal{N}_j, j \geq 1$, $pa(X_i)$ is a Norming Set in $\cup_{\ell=1}^{j-1} \mathcal{N}_\ell$. Order of polynomial space is chosen $l = \underline{\alpha}$.

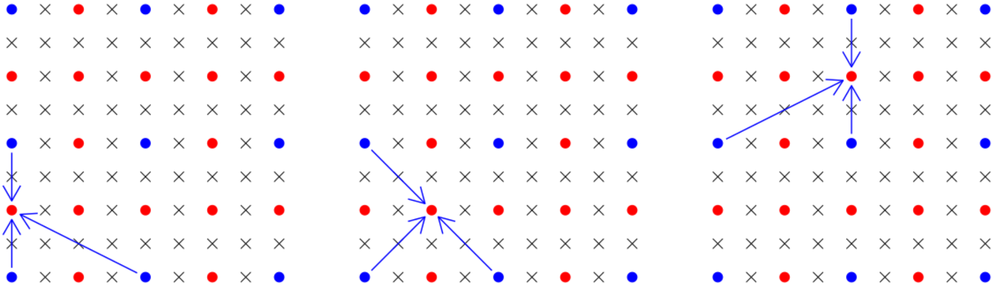


Figure: Illustration of parent sets for $X_i \in \mathcal{N}_2$, with $\underline{\alpha} = 1$. Red dots: current layer $\mathcal{N}_2$; Blue dots: previous layers $\mathcal{N}_0$, $\mathcal{N}_1$; Black crosses: all latter layers. Blue arrows: directed edges from parent sets to children for some $X_i \in \mathcal{N}_2$.

Vecchia GP: small deviation bounds

Concentration function: Let $\mathbb{H}_n^\tau$ denote the RKHS corresponding to the Vecchia GP, then

$$\phi_{f_0,n}^\tau(\epsilon) = \inf_{f \in \mathbb{H}_n^\tau : \|f - f_0\|_\infty \leq \epsilon} \|f\|_{\mathbb{H}_n^\tau}^2 - \log(\|\hat{Z}^\tau\|_\infty < \epsilon).$$

Vecchia GP: small deviation bounds

Concentration function: Let $\mathbb{H}_n^\tau$ denote the RKHS corresponding to the Vecchia GP, then

$$\phi_{f_0,n}^\tau(\epsilon) = \inf_{f \in \mathbb{H}_n^\tau : \|f - f_0\|_\infty \leq \epsilon} \|f\|_{\mathbb{H}_n^\tau}^2 - \log(\|\hat{Z}^\tau\|_\infty < \epsilon).$$

Theorem: For the Layered Norming Set DAG and Matern GP with regularity α and scale parameter τ the Vecchia GP $\hat{Z}^\tau$ satisfies

$$-\log Pr\left(\|\hat{Z}^\tau\|_\infty < \epsilon\right) \lesssim \tau^d \epsilon^{-d/\alpha}.$$

Vecchia GP: small deviation bounds

Concentration function: Let $\mathbb{H}_n^\tau$ denote the RKHS corresponding to the Vecchia GP, then

$$\phi_{f_0,n}^\tau(\epsilon) = \inf_{f \in \mathbb{H}_n^\tau : \|f - f_0\|_\infty \le \epsilon} \|f\|_{\mathbb{H}_n^\tau}^2 - \log(\|\hat{Z}^\tau\|_\infty < \epsilon).$$

Theorem: For the Layered Norming Set DAG and Matern GP with regularity α and scale parameter τ the Vecchia GP $\hat{Z}^\tau$ satisfies

$$-\log Pr\left(\|\hat{Z}^\tau\|_\infty < \epsilon\right) \lesssim \tau^d \epsilon^{-d/\alpha}.$$

Lemma: For the Layered Norming Set DAG and Matern GP with regularity α and scale parameter τ,

$$\inf_{f \in \mathbb{H}_n^\tau : \|f - f_0\|_\infty \le \epsilon} \|f\|_{\mathbb{H}_n^\tau}^2 \lesssim \tau^d \epsilon^{-d/\alpha} + \tau^{-2\alpha} \epsilon^{-\frac{2(\alpha-\beta)+d}{\beta}} + \epsilon_n^{-\frac{d}{\beta}}.$$

Vecchia GP: posterior contraction

Theorem: Consider the rescaled Matérn GP as based prior. Then for $f_0 \in C^\beta$, with $\beta \leq \alpha$, and setting either $\tau = n^{\frac{\alpha-\beta}{\alpha(2\beta+d)}}$ or endowing τ with a hyperprior (satisfying mild tail conditions), the posterior corresponding to the (hierarchical) Vecchia GP approximation achieves minimax contraction rate, i.e.

$$\sup_{f_0 \in C^\beta(M)} E_{f_0} \Pi_n^V (f : \|f - f_0\|_n \geq M_n n^{-\frac{\beta}{d+2\beta}} | Y) \to 0,$$

for arbitrary $M_n \to \infty$.

Proof sketch: Solve

$$\phi^\tau_{f_0,n}(\epsilon_n) \leq n\epsilon_n^2.$$

This $\epsilon_n = n^{-\frac{\beta}{d+2\beta}}$ is the posterior contraction rate by general GP theorem vd Vaart, van Zanten (2018).

Vecchia NNGP vs Layered Norming DAG

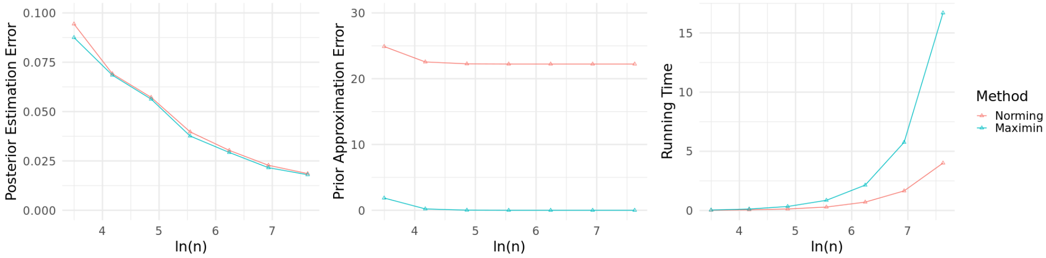


Figure: Comparison for Vecchia GP with **Layered Norming DAGs** and **NNGP with maximin ordering**:

Left: posterior estimation error measured by ℓ^2 norm between the truth and the posterior mean;

Middle: prior approximation error measured by squared Wasserstein distance between marginals of Vecchia GPs and their mother GPs;

Right: Run time of MCMC inference measured by seconds.

Deep Gaussian Processes

Limitations of GPs

Problem: Not appropriate to learn compositional structures, adapt to local regularities and structures.

Def (Generalized additive models):

$$\mathcal{G}(M) = \{f(x_1, ..., x_d) = h(\sum_{j=1}^{d} g_i(x_i)) : \; g_i, h \in \mathrm{Lip}(M) \cap L^\infty(M)\}.$$

Minimax rate Schmidt-Hieber (2020): $\inf_{\hat{f}} \sup_{f \in \mathcal{G}(M)} \|\hat{f} - f\|_2 \asymp^* n^{-1/3}$.

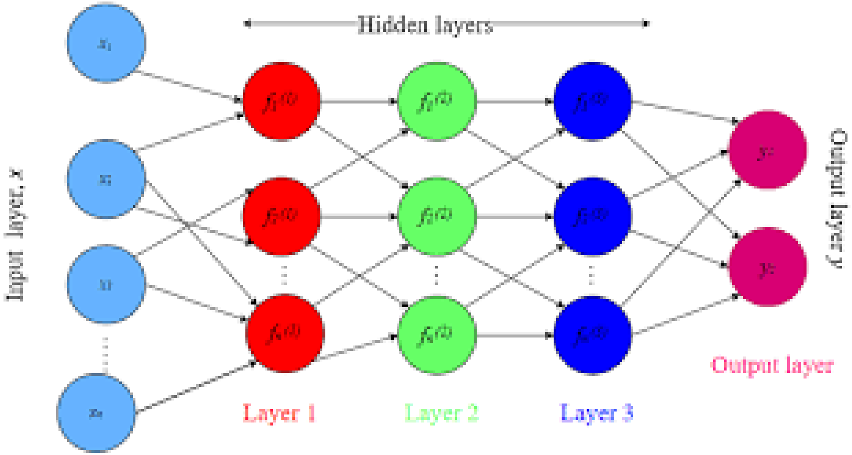
* : up to a poly-log term.

Theorem (Giardano et al. (2022)) For any sequence Π_n of Gaussian process priors, if

$$\sup_{f_0 \in \mathcal{G}(M)} E_{f_0} \Pi_n(f : \; \|f - f_0\|_2 \geq \varepsilon_n | Y) \to 0$$

holds, then $\varepsilon_n \gtrsim n^{-\frac{1}{4} - \frac{1}{4+4d}}$ (suboptimal for $d > 2$).

Compositional structure: picture



Compositional structure: definition

Compositional class:

$$\mathcal{F} = \{f = h_q \circ h_{q-1} \circ \ldots \circ h_0 \, : \, h_i = (h_{ij})_j : [-1,1]^{d_i} \to [-1,1]^{d_{i+1}}, \bar{h}_{ij} \in C_{t_i}^{\beta_i}(M)\},$$

where h_{ij} is allowed to depend on $t_i \leq d_i$ variables $\mathcal{S}_{ij} \subseteq \{1, .., d_i\}$, with $|\mathcal{S}_{ij}| = t_i$ and $\bar{h}_{ij} : [-1,1]^{t_i} \to [-1,1]$,

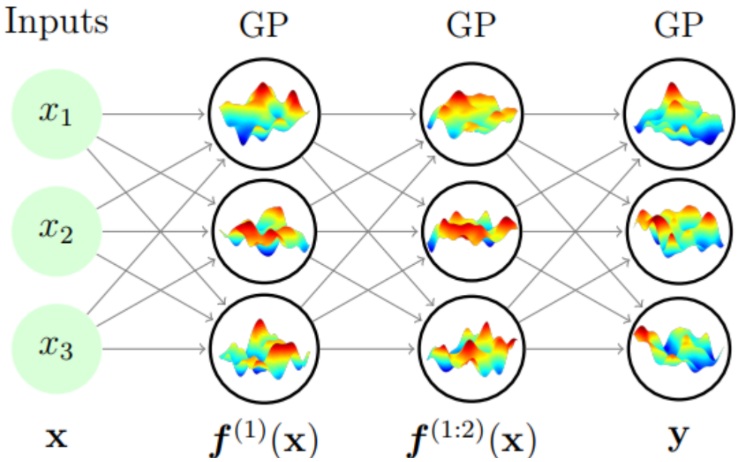
$$x_{\mathcal{S}_{ij}} \mapsto h_{ij}(x_{\mathcal{S}_{ij}}, x_{\mathcal{S}_{ij}^c}).$$

Minimax rate (Schmidt-Hieber (2020)):

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} E_f \|\hat{f} - f\|_2 \asymp^* \max_{i=0,..,q} n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}}, \qquad \text{with } \beta_i^* = \beta_i \prod_{\ell=i+1}^{q} (\beta_\ell \wedge 1).$$

* up to a logarithmic term.

Deep GP: Illustration



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

Deep GP: application

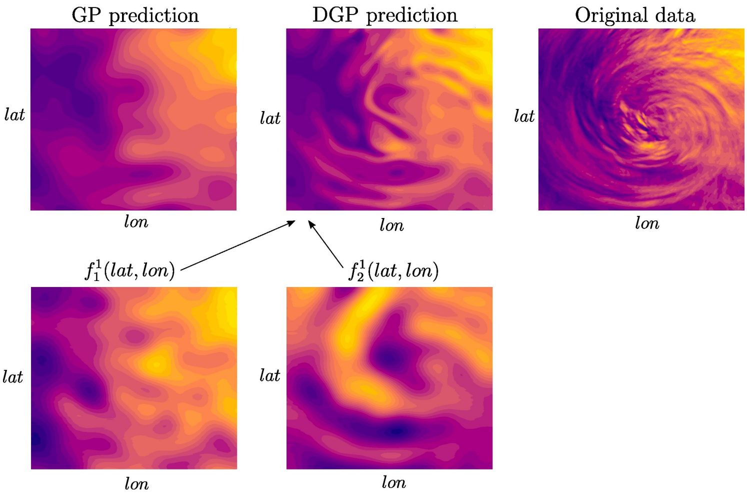


Figure: Modeling a hurricane field with GP vs Deep GP. Svendsen et al (2020): Deep Gaussian processes for biogeophysical parameter retrieval and model inversion

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: : bounded sup norm, close to Holder function

Deep GP: theory

Theorem (Finocchio & Schmidt-Hieber (2022)): Under weak regularity assumptions and suitable GP priors on the edges, the hierarchical construction of the deep GP prior results in nearly minimax posterior contraction in the compositional class, i.e. for some poly-log sequence M_n

$$\sup_{f_0 \in \mathcal{F}(M)} E_{f_0} \Pi_n\Big(f : \|f - f_0\|_2 \geq M_n \max_{i=0,..,q} n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}} \,|\, Y\Big) \to 0.$$

Other approach Castillo & Randrianarisoa (2025): endow the scale parameters of SE GP with another layer of prior. Does automatic edge selection. Similar theoretical results for fractional posteriors (works in high-dimensional models as well).

Vecchia approximation of Deep Gaussian Processes

Deep Vecchia GP: method

Previous method (Sauer et al. (2022): deepgp) Two-layer deep Vecchia $f_2 \circ f_1$

- Layer 1: NNGP approx. of f_1 based on design points $x_1, ..., x_n$.
- Layer 2: NNGP approx. of f_2 based on image of design $f_1(x_1), ..., f_1(x_n)$.

Problem: $f_1(x_1), ..., f_1(x_n)$ can be close to each other resulting in bad fit.

Proposed method: q-layer Vecchia GP $f_q \circ f_{q-1} \circ ... \circ f_1$

- Layer j: Vecchia (Layered Norming DAG) approximation of f_j based on grid points $(i/m, j/m)_{i,j=1,..,m}$
- Gibbs sampler (under construction): complexity $O(q \log n)$ per iteration (due to localized basis structure of Vecchia GP).

Deep Vecchia GP: Theory

Conjecture: Using Layered Norming DAG Vecchia approximation at each layer of the hierarchical deep GP construction of Finocchio & Schmidt-Hieber (2022) with Matérn covariance kernel, the corresponding posterior achieves the near minimax contraction rate for compositional functions, i.e.

$$\sup_{f_0 \in \mathcal{F}(M)} E_{f_0} \Pi_n(f : \|f - f_0\|_2 \geq M_n \max_{i=0,..,q} n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}} | Y) \to 0,$$

for M_n a poly-log factor.

Deep Vecchia GP: extension/ongoing work

- Extend our results to Deep Horseshoe GPs. This allows to get rid of the regularization of the sample paths.

- Consider also the square exponential covariance kernel.

- Prove local adaptation of (Vecchia) Deep GPs.

- Prove pointwise/supremum convergence rates for (Vecchia) Deep GPs

Summary

- Gaussian Processes are popular in applications.
- Good theoretical performance, but computational problems.
- Scalable approximation: Vecchia.
- Vecchia GP based on layered norming DAG: parents set $m = O(1)$ and minimax contraction rate.
- Deep GPs: compositional (deep) structure for GPs (with prior on graph structure).
- Extension of Vecchia approximation to deep GPs: minimax rates, algorithmic aspects in development.

Papers

- B. Szabo, Y. Zhu (2025+) Vecchia gaussian processes: Probabilistic properties, minimax rates and methodological developments. Major revision for AoS.
- H. van Zanten, B. Szabo (2019) An asymptotic analysis of distributed nonparametric methods. JMLR 20 (87), 1-30.
- B. Szabo, A. Hadji, A vd Vaart (2025) Adaptation using spatially distributed Gaussian Processes. JASA (to appear).
- 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
- T. Randrianarisoa, B. Szabo (2023) Variational Gaussian processes for linear inverse problems. NeurIPS 36, 28960-28972.
- B. Stankewitz, B. Szabo (2024) Contraction rates for conjugate gradient and Lanczos approximate posteriors in Gaussian process regression. Arxiv