Radial Neighbors for Provably Accurate Scalable Approximations of Gaussian Processes

Yichen Zhu*, Michele Peruzzi*, Cheng Li+ and David B. Dunson*

Duke University, Durham, NC, USA*, National University of Singapore, Singapore+

yichen.zhu@duke.edu

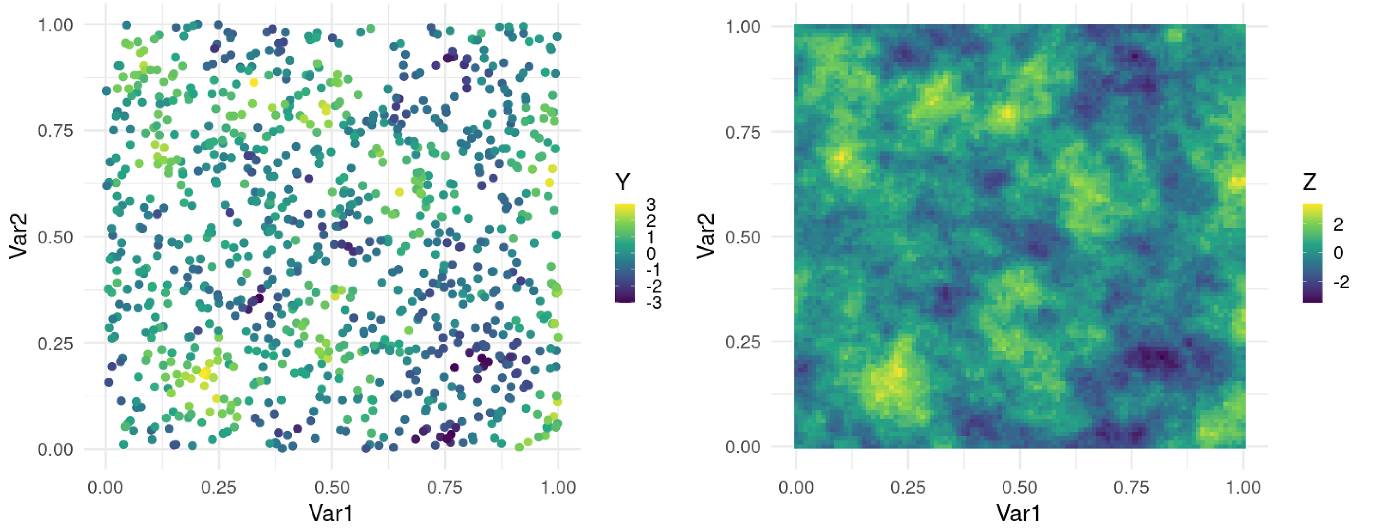
1. Introduction

Gaussian process (GP) regressions are widely used in geostatistics:

$$Y(s) = X(s)^\top \beta + Z(s) + \epsilon(s),$$

where s is the spatial location, $X(s)$ is the vector of covariates at s, β contains regression coefficients, $Z(s)$ is the latent effect following Gaussian process priors and $\epsilon(s)$ is the nugget effect (white noise).

- A common problem is to infer the latent process $Z(s)$ on the whole space (right figure) from $Y(s)$ at training locations (left figure).
- The computational complexity for GP regression is $O(n^3)$.



2.1. Vecchia Approximation

Let the union of training and testing data be $\mathcal{D} = \{w_1, \dots, w_n\}$.

- The joint density of Z on $\mathcal{D}$ can be decomposed into products of unidimensional conditional densities using Bayes rule;
- Vecchia approximations replace each conditional set $\{w_j, j < i\}$ with a much smaller parent set $\mathrm{pa}(w_i)$;
- This results in a new process $\hat{Z}_\mathcal{D}$ scalable to large datasets.

$$p(Z_\mathcal{D}) = p(Z_{w_1}) \prod_{i=2}^{n} p(Z_{w_i} | Z_{w_j, j<i}) \approx p(\hat{Z}_{w_1}) \prod_{i=2}^{n} p(\hat{Z}_{w_i} | \hat{Z}_{\mathrm{pa}(w_i)}).$$

Many existing Vecchia approximation methods are **sensitive** to the specification of certain graph structures; they also have **little theoretical guarantees**.

2.2. Radial Neighbors Gaussian Process

Radial neighbors Gaussian process (RadGP) chooses the parent set $\mathrm{pa}(w_i)$ as locations ordered before i and within ρ distance to w_i;

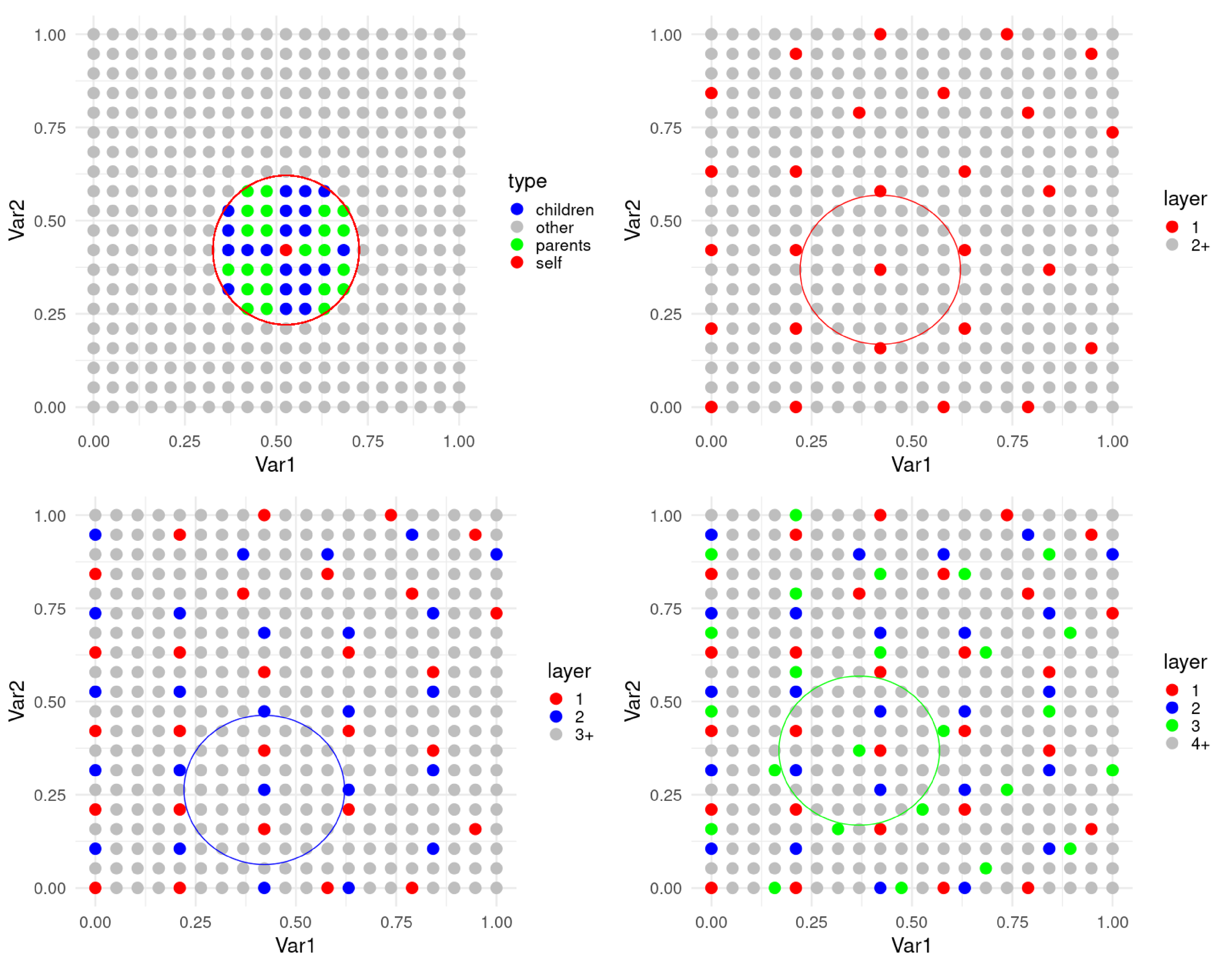
- $\mathrm{pa}(w_i) = \{w_j : \ j < i, \ \|w_i - w_j\|_2 \le \rho\}$;
- The union of parent set and child set covers exactly all locations within ρ radius.

To obtain such parent sets, we can first compute an "alternating partition" $\mathcal{D} = \cup_{i=1}^{n} \mathcal{D}_i$ such that:

- $\forall i, \forall s_1, s_2 \in \mathcal{D}_i, \|s_1 - s_2\|_2 \ge \rho$;
- Training samples are always allocated to $\mathcal{D}_i$ with smaller indices i than testing samples.

We then compute the parent set for all locations $s \in \mathcal{D}_i$ as

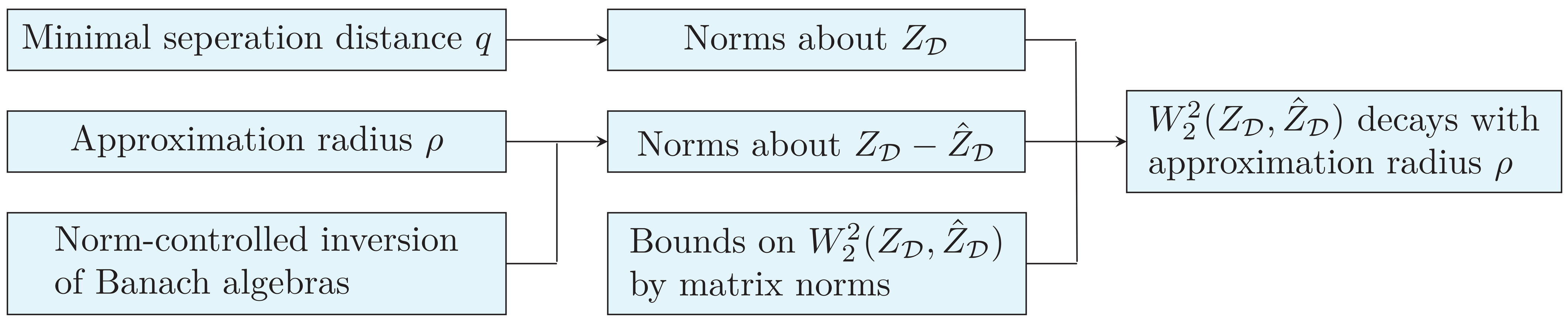
$$\mathrm{pa}(s) = \{s' \in \mathcal{D}_j : \ j < i, \ \|s' - s\|_2 \le \rho\}.$$



Top left: parent set, children set and approximation radius ρ for the location in red; The other three: layer $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$ and ρ radius balls for one location of each layer.

3.1. Overview of Theory

- **Objective**: Bound the **Wasserstein-2 distance** between the radial neighbors Gaussian process $\hat{Z}$ and the original Gaussian process Z (New theoretical guanratees for scalable GP approximation);
- **Quantities Involved**: Decay rate of covariance function; Minimal separation distance of the dataset q; Sample size n;
- **Key Tool**: Theory on Norm-controlled Inversion of Banach Algebra;
- **Results**: W_2 distance decays with the approximation radius ρ, in a similar rate to the covariance function decay rate.



3.2. Main Theory: Rate of Approximation

The Gaussian process $Z(\cdot)$ has the isotropic covariance function $\mathrm{Cov}(Z(s_1), Z(s_2)) = K_0(\|s_1 - s_2\|_2)$.

Case 1 If the covariance function decays faster than any polynomials:

- Define the rate function $v_r(x) = \sum_{k=0}^{+\infty} x^k / (k!)^r$;
- Define the family $\mathscr{Z}_{v_r} = \left\{ Z = (Z_s : s \in \Omega) : K_0(\|s_1 - s_2\|_2) \le \frac{1}{v_r(\|s_1 - s_2\|_2)(1 + \|s_1 - s_2\|_2^{d+1})} \right\}$.

Theorem 1 *For the family $\mathscr{Z}_{v_r}$ with $r > 1$, if $0 < q < 1$, then*

$$\sup_{Z \in \mathscr{Z}_{v_r}} W_2^2(Z_\mathcal{D}, \hat{Z}_\mathcal{D}) \lesssim \frac{n}{v_r(\rho/\sqrt{d})} \{\phi_0(c_2/q)\}^{-5} q^{-d} v_{r-1}(c_3 \{\phi_0(c_2/q)\}^{-1}).$$

Else if $q \ge 1$, then $\sup_{Z \in \mathscr{Z}_{v_r}} W_2^2(Z_\mathcal{D}, \hat{Z}_\mathcal{D}) \lesssim n / v_r(\rho/\sqrt{d})$.

Case 2 Else if the covariance function decays no faster than some polynomials:

- Define the rate function $c_r(x) = (1 + |x|)^r$;
- Define the family $\mathscr{Z}_{c_r} = \left\{ Z = (Z_s : s \in \Omega) : K_0(\|s_1 - s_2\|_2) \le \frac{1}{c_r(\|s_1 - s_2\|_2)} \right\}$.

Theorem 2 *For the family $\mathscr{Z}_{c_r}$ with $r \ge d + 1$, if $0 < q < 1$, then,*

$$\sup_{Z \in \mathscr{Z}_{c_r}} W_2^2(Z_\mathcal{D}, \hat{Z}_\mathcal{D}) \lesssim \frac{n}{(1 + \rho/\sqrt{d})^{-(r-d-1)}} q^{(r-8)d} \{\phi_0(c_2/q)\}^{-(r+9/2)} (c_1 c_5 d 2^{d-1} \pi / \sqrt{6})^r.$$

Else if $q \ge 1$, then $\sup_{Z \in \mathscr{Z}_{c_r}} W_2^2(Z_\mathcal{D}, \hat{Z}_\mathcal{D}) \lesssim n(1 + \rho/\sqrt{d})^{-(r-d-1)} \{c_1 c_5 d 2^{d-1} \phi_0(c_2/q) \pi / \sqrt{6}\}^r$.

Summary of Sufficient Conditions on Approximation Radius ρ to Guarantee $W_2^2(Z_\mathcal{D}, \hat{Z}_{\mathcal{D}'}) \to 0$

Covariance function $K_0(\|\Delta s\|_2)$	Lower bounds for ρ
Matérn: $\frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)} (\alpha\|\Delta s\|)^\nu \mathcal{K}_\nu(\alpha\|\Delta s\|_2)$	$\rho \gtrsim \frac{\sqrt{d}}{\alpha} \left[c_{m,1} \left(1 + \frac{c_2^2}{\alpha^2 q^2}\right)^{\nu + \frac{d}{2}} \ln \left\{ c_{m,1} n q^{-d} \left(1 + \frac{c_2^2}{\alpha^2 q^2}\right)^{5(\nu + \frac{d}{2})} \right\} \right]^3$
Gaussian: $\exp(-a\|\Delta s\|_2^2)$	$\rho \gtrsim \frac{\sqrt{d}}{\alpha} \left[e^{\frac{c_2^2}{4aq^2}} \left\{ \ln(nq^{-d}) + \frac{c^2}{4aq^2} \right\} \right]^3$
G-Cauchy: $\sigma^2 \left\{1 + (\|s_1 - s_2\|/\alpha)^\delta\right\}^{-\lambda/\delta}$	$\rho \gtrsim q^{-\left\{\frac{25}{2}\lambda d + \delta(\lambda + \frac{9}{2})\right\} / \{\lambda - (d+1)\}}$

4. Simulations

We focus on inferring dependence among testing latent effects.

- Let the covariance function be exponential $\tau^2 \exp(-\phi\|\Delta s\|_2^2)$ with unknown τ^2 and ϕ;
- Each method outputs posterior samples over some local regions;
- The W_2 distances between summary statistics (mean, standard deviation, median and mean of relu) of these posterior samples and posterior samples of true Gaussian process are computed;
- Smaller values indicate better approximations of the true Gaussian process.