# Vecchia Gaussian Processes
## Probabilistic Properties, Minimax Rates and Methodology

Botond Szabo and Yichen Zhu

Bocconi University

November 7, 2024

# Gaussian Process Regression

Consider the following Gaussian process regression model:

$$Y(x) = u(x)^T \beta + f(x) + \epsilon(x),$$

- $x$ is the spatial location
- $u(x)$ is the covariates at location $x$ and $\beta$ is the linear regression coefficients
- $f(x)$ is the spatial regression function at $s$ that follows a mean zero Gaussian process:

$$f \sim (Z_x)_{x \in \mathcal{X}}$$

  with covariance function $\mathrm{Cov}(Z_{x_1}, Z_{x_2}) = K(x_1, x_2)$.
- $\epsilon(x) \sim N(0, \sigma^2)$ is the white noise (also called the nugget effect)

# Computational Challenge

The Bayesian prior and hyperprior:

$$f \sim (Z_x|\theta)_{x\in\mathcal{X}}, \quad \beta \sim p(\beta), \quad \theta \sim p(\theta),$$

$$\epsilon(x) \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2), \quad \sigma^2 \sim p(\sigma^2).$$

The joint probability density of $Y_x, Z_x, \beta, \sigma^2$ and $\theta$ at finite dataset $\mathcal{X}_n \triangleq \{X_1, X_2, \cdots, X_n\}$:

$$\sigma^{-2n} \exp\left[-\sum_{i=1}^{n} (Y(X_i) - u(X_i)^T \beta - Z_{X_i})^2\right] p(Z_{\mathcal{X}_n}|\theta) p(\theta) p(\beta) p(\sigma^2).$$

The term $p(Z_{\mathcal{X}_n}|\theta)$ requires evaluation of the $n$-dimensional Gaussian density, which involves computing the precision matrix and the determinant from the covariance matrix at $O(n^3)$ time.

# Vecchia Approximations

Vecchia approximations are designed to approximate the mother Gaussian process $(Z_x|\theta)_{x\in\mathcal{X}}$ with another process $(\hat{Z}_x|\theta)_{x\in\mathcal{X}}$ such that $p(\hat{Z}_{\mathcal{X}_n}|\theta)$ can be computed in $O(n)$ time.

- The joint density of the original Gaussian process $Z$ on $\mathcal{X}_n$:

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

- Vecchia approximations replace each conditional set $\{X_j, j < i\}$ with a much smaller parent set $\mathrm{pa}(X_i)$:

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

$$\hat{Z}_{X_1} \overset{d.}{=} Z_{X_1}, \quad [\hat{Z}_{X_i}|\hat{Z}_{\mathrm{pa}(X_i)} = z] \overset{d.}{=} [Z_{X_i}|Z_{\mathrm{pa}(X_i)} = z], \forall z.$$

# Problems and Challenges

Vecchia approximation was proposed by Vecchia (1988) and received a lot of research attention in the past ten years (Datta et al., 2016; Katzfuss et al., 2020; Katzfuss and Guinness, 2021; Peruzzi et al., 2022).

However, there are major methodological and theoretical problems that remain unsolved for years:

- **Methodology**: How shall we choose the parent set $\mathrm{pa}(X_i), \forall i$ to guarantee good (or optimal) performances?
- **Nonparametric Theory**: What is the rate of convergences (or posterior contraction rate) for Vecchia GPs under ideal DAG structures?
- **Probability**: *How much do we know about Vecchia Gaussian processes as standalone Stochastic processes?*

# Norming Sets

For $\Omega$ a compact subset of $\mathbb{R}^d$, $l \in \mathbb{N}$, denote $\mathscr{P}_l(\Omega)$ as the collection of polynomials on $\Omega$ with orders no greater than $l$.
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

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

## Condition 1

*There exists $c_N > 0$, such that for all $i$ sufficiently large, the parent set $\mathrm{pa}(X_i)$*

- *has cardinality $\binom{\alpha+d}{\alpha}$;*
- *is a norming set for $\mathscr{P}_{\alpha}(C)$ with norming constant $c_N$, where $C \supset \mathrm{pa}(X_i) \cup \{X_i\}$ is a $d$-dimensional cube with side length $r_i$.*

# Norming Sets

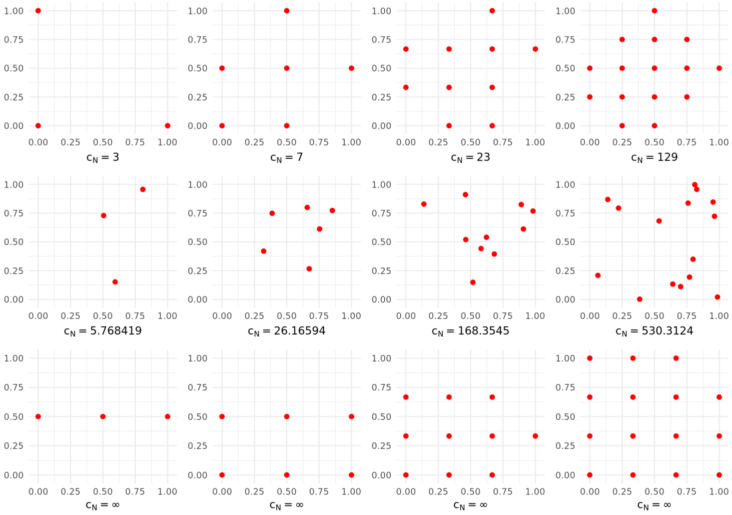


Figure: Norming sets in 2-dimensional space.

# Local Polynomial Behaviors

**Lemma 1**

*For Matérn process with smoothness parameter $\alpha$, under condition 1, for all $X_i$, we have*

$$\mathrm{Var}\{Z_{X_i} - \mathbb{E}[Z_{X_i} | Z_{\mathrm{pa}(X_i)}]\} \asymp {r_i}^{2\alpha}.$$

Let $K$ be the Matérn covariance function and $V$ be the high-dimensional Vandermonde matrix.

**Theorem 1**

*For Matérn process with smoothness parameter $\alpha$, under condition 1, for all $X_i$, we have*

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

# Posterior Contraction

**Theorem 2**

*Suppose Conditions 1 holds and the true regression function f belongs to the unit Hölder ball with smoothness $\beta$. Let $\hat{Z}$ be Vecchia approximation of Matérn process with smoothness $\alpha \geq \beta$. Then there exists a constant $M$, such that conditional on the training data $\mathcal{X}_n$, we have*

$$\Pi(\|f - f_0\|_{\infty,n} > Mn^{-\frac{\beta}{2\alpha+d}} | \mathcal{X}_n) \xrightarrow{P} 0.$$

The minimax rate $n^{-\beta/(2\beta+d)}$ is achieved if and only if the prior smoothness $\alpha$ matches the smoothness of the truth $\beta$.

# Rescaling

Define a Rescaled version of the mother Gaussian process as:

$$Z_x^{\tau,s} = sZ_{\tau x}.$$

Let $\hat{Z}^{\tau,s}$ be the Vecchia approximation of $Z^{\tau,s}$.

**Theorem 3**

*Under the Conditions of Theorem 2, for the process $\hat{Z}^{\tau,s}$, if $\tau^{\alpha}s = n^{\frac{\alpha-\beta}{2\beta+d}}$ and $s \geq 1$, then we have*

$$\Pi(\|f - f_0\|_{\infty,n} > Mn^{-\frac{\beta}{2\beta+d}} | \mathcal{X}_n) \xrightarrow{P} 0.$$

In other words, we can always rescale a smooth process to get minimax rate on a rough function family.

# Adaptation

- Specify the rescaling parameters requires the knowledge of the true smoothness $\beta$.
- It is common in practice to put a hyper prior on $\tau$ and $s$ to obtain a mixture of GPs:

$$\hat{Z}_x^{\mathrm{Pr}} = \int_\tau \hat{Z}^{\tau,s} p(\tau,s) d\tau ds.$$

## Theorem 4

*Under Conditions of Theorem 2, for the mixture of Gaussian process $\hat{Z}^{\mathrm{Pr}}$ satisfying*

$$\log \mathrm{Pr}\left(\tau^\alpha s > n^{\frac{\alpha-\beta}{2\beta+d}}\right) \lesssim -n^{\frac{d}{2\beta+d}},$$

$$\log \mathrm{Pr}\left(\left\{\tau^\alpha s \in \left[n^{\frac{\alpha-\beta}{2\beta+d}}, 2n^{\frac{\alpha-\beta}{2\beta+d}}\right]\right\} \cap \{s \geq 1\}\right) \gtrsim -n^{\frac{d}{2\beta+d}}.$$

*we have*

$$\Pi(\|f - f_0\|_{\infty,n} > Mn^{-\frac{\beta}{2\beta+d}} | \mathcal{X}_n) \xrightarrow{P} 0.$$

# Conlusions and Discussions

We contribute to literature of Vecchia GPs from three aspects:

- **Probabilistic property**: We prove that Matérn processes, as well Vecchia approximations of Matérn processes, conditional on a norming set behave like polynomials.
- **Methodology**: We prove that choose parent sets to be norming sets with exact $\binom{\underline{\alpha}+d}{\underline{\alpha}}$ elements is sufficient to guarantee optimal contraction rates.
- **Nonparametric Theory** Vecchia approximated GPs enjoy the same posterior contraction rates as their mother Gaussian processes, which is minimax optimal if
  - Prior smoothness matches true smoothness
  - Prior is oversmooth but properly rescaled
  - Prior is oversmooth but we put appropriate hyperprior on the rescaling parameters

# Reference

Datta, A., S. Banerjee, A. O. Finley, and A. E. Gelfand (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association 111*(514), 800–812.

Katzfuss, M. and J. Guinness (2021). A general framework for vecchia approximations of gaussian processes. *Statistical Science 36*(1), 124–141.

Katzfuss, M., J. Guinness, W. Gong, and D. Zilber (2020). Vecchia approximations of gaussian-process predictions. *Journal of Agricultural, Biological and Environmental Statistics 25*, 383–414.

Peruzzi, M., S. Banerjee, and A. O. Finley (2022). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association 117*(538), 969–982.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology 50*(2), 297–312.

# Questions?

Funded by the European Research Council (BigBayesUQ, project number: 101041064).

# Nonparametric Estimation Without Approximation

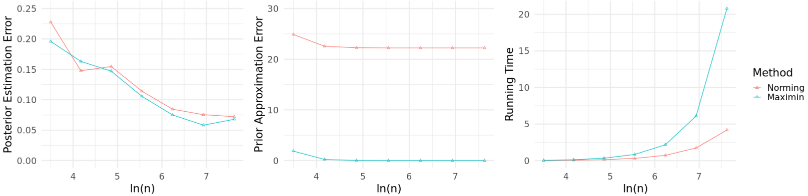


Figure: Qualitative results of applying two Vecchia GP methods under different sample sizes. Left: posterior estimation error measured by $\ell^\infty$ norm between the truth and the posterior mean; Middle: prior approximation error measured by squared Wasserstein distance between marginals of Vecchia GPs and their mother GPs; Right: Run time of MCMC inference measured by seconds.

# Vecchia GPs: Extend to the Whole Domain

Let $\mathcal{X} \subset \mathbb{R}^d$ be the domain for Vecchia Gaussian processes. We extend $\hat{Z}$ from the finite set $\mathcal{X}_n$ to $\mathcal{X}$ as follows:
$\forall X \in \Omega \backslash \mathcal{X}_n$, let the parent set for $X$ be a finite set satisfying $\mathrm{pa}(w) \subset \mathcal{X}_n$. Then for all finite subset $A \subset \Omega \backslash \mathcal{X}_n$,

$$p(\hat{Z}_A | \hat{Z}_{\mathcal{X}_n}) = \prod_{X \in A} p(\hat{Z}_X | \hat{Z}_{\mathrm{pa}(X)}).$$

Vecchia Gaussian processes on the whole domain are conditional independent given the finite set $\mathcal{X}_n$.