

# Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets

Abhirup Datta, Sudipto Banerjee<sup>1</sup>, Andrew O. Finley, Alan E. Gelfand

<sup>1</sup>Department of Biostatistics, UCLA Fielding School of Public Health

February 21, 2025

Presented by Christine Shen

# Outline of the presentation

## 1. Introduction

- How to use Gaussian Process in Bayesian spatial models
- Computational issues and existing methods

## 2. Nearest Neighbor Gaussian Process (NNGP)

## 3. Application of NNGP to Bayesian spatial models

## 4. Simulation Results

## 5. Conclusion

# What is a Gaussian Process

A Gaussian Process (GP) is

- a stochastic process, i.e., a collection of random variables indexed by  $\mathbf{s} \in \mathcal{D}$ , where any finite number of them have a joint Gaussian distribution
- an extension of the multivariate Gaussian to infinite dimensions.
  - E.g., we are familiar with random variables indexed by integers:  $X_1, \dots, X_n \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Now consider random variables  $\{X(\mathbf{s})\}$ , indexed by time, or location (spatial), or time and location (spatial-temporal)...

# How to use GP in Bayesian spatial models

Consider data on  $n$  patients who have had upper extremity fractures. For each patient, we know their gender, age, address, and time to readmission since initial fracture (assume no censoring). We are interested in studying the geospatial pattern of patients' readmission risks, controlling for gender and age. We posit the following Bayesian model

$$\log y_i = \mathbf{x}_i^T \boldsymbol{\beta} + w(\mathbf{s}_i) + \epsilon_i, \quad \epsilon_i \sim \text{i.i.d. } N(0, \sigma^2), \quad \text{where}$$

- $y_i$  is the time to readmission for patient  $i$
- $\mathbf{x}_i \in \mathbb{R}^3$  are the covariates for patient  $i$  including an intercept, gender, and age
- $\mathbf{s}_i \in \mathbb{R}^2$  is the location for patient  $i$
- $w(\mathbf{s}_i)$  is the spatial intercept.  $w(\mathbf{s}) \sim GP(0, C(\cdot, \cdot | \boldsymbol{\theta}))$
- With priors for  $\boldsymbol{\beta}$ ,  $\sigma^2$ , and  $\boldsymbol{\theta}$

# Computational issues with GP

GP is not scalable for a full Bayesian analysis. For a dataset of size  $n$ , the GP prior is effectively

$$\begin{pmatrix} w(\mathbf{s}_1) \\ \vdots \\ w(\mathbf{s}_n) \end{pmatrix} \sim N(0, \Sigma), \quad \Sigma \in \mathbb{R}^{n \times n},$$

i.e., we need to work with an  $n \times n$  covariance matrix for each MCMC iteration, which requires  $\mathcal{O}(n^3)$  floating point operations (flops), and  $\mathcal{O}(n^2)$  memory.

# Scalable GP methods

Existing scalable GP methods broadly fall under two categories.

## Low-rank methods

- Approximate GP on a low-dimensional subspace via  $r$  basis functions,  $r \ll n$
- E.g., process convolution ([Higdon, 2002]),  $\mathcal{O}(r^2)$  flops
- E.g., Hilbert space approximation (HSGP, [Riutort-Mayol et al., 2023]),  $\mathcal{O}(rn + r)$  flops
- Drawbacks:
  - 1 As  $n$  grows,  $r$  inevitably increases
  - 2 Perform poorly when neighboring observations are strongly correlated, and the spatial signal dominates the noise ([Stein, 2014])
  - 3 Models are degenerate

## Sparsity methods

Existing scalable GP methods broadly fall under two categories.

## Low-rank methods

## Sparsity methods

- Sparsity in  $C(\theta)$ , e.g., covariance tapering ([Furrer et al., 2006])
- Sparsity in  $C(\theta)^{-1}$ , e.g., Vecchia approximation ([Vecchia, 1988])
- Drawbacks
  - 1 There might not be a corresponding stochastic process
  - 2 Do not naturally extend to new random variables at arbitrary locations

# Nearest Neighbor Gaussian Process

[Datta et al., 2016] introduced Nearest Neighbor Gaussian Process (NNGP). It is

- a scalable method which relies on sparsity
- a valid spatial process

We will look at how NNGP

- introduces sparsity in finite-dimensional probability models
- extends finite-dimensional models to valid spatial processes



# Finite-dimensional probability models with sparsity

Consider a  $q$ -variate gaussian process over  $\mathbb{R}^d$ :

$$w(\mathbf{s}) \sim GP(0, C(\cdot, \cdot \mid \boldsymbol{\theta})), \quad \text{where } w(\mathbf{s}) \in \mathbb{R}^q, \quad \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d,$$

and  $C$  is an *isotropic* covariance function with parameter  $\boldsymbol{\theta}$ . Let  $\mathcal{S} = \{\mathbf{s}_1, \dots, \mathbf{s}_k\} \subset \mathcal{D}$  be a fixed collection of distinct locations in  $\mathcal{D}$ , which we will call the *reference set*.

Remarks and notations:

- a covariance function is *stationary* if  $C(\mathbf{s}, \mathbf{s}') = C(\mathbf{s} - \mathbf{s}')$ , *isotropic* if  $C(\mathbf{s}, \mathbf{s}') = C(\|\mathbf{s} - \mathbf{s}'\|)$
- $w_{\mathcal{S}} = (w(\mathbf{s}_1)^T, \dots, w(\mathbf{s}_k)^T)^T$  denotes the long stacking vector of the process at locations in  $\mathcal{S}$
- $C_{\mathcal{S}}$  denotes the  $qk \times qk$  covariance matrix with  $C(\mathbf{s}_i, \mathbf{s}_j \mid \boldsymbol{\theta})$  as the  $(i, j)$ th block
- $C_{\mathbf{s}_i, \mathcal{S}}$  denotes the  $q \times qk$  cross covariance matrix between  $w(\mathbf{s}_i)$  and  $w_{\mathcal{S}}$

# Finite-dimensional probability models with sparsity

For the reference set  $\mathcal{S}$ ,

$$w_{\mathcal{S}} \sim N(0, C_{\mathcal{S}}),$$

where  $C_{\mathcal{S}}$  is a dense covariance matrix. We now derive a joint probability model for  $w_{\mathcal{S}}$  with sparse precision matrix.

- 1 The joint density of  $w_{\mathcal{S}}$  can be expressed as product of conditional densities

$$p(w_{\mathcal{S}}) = p(w(\mathbf{s}_1))p(w(\mathbf{s}_2) \mid w(\mathbf{s}_1)) \dots p(w(\mathbf{s}_k) \mid w(\mathbf{s}_1), \dots, w(\mathbf{s}_{k-1}))$$

- 2 We can replace the large conditioning sets with smaller sets of size at most  $m$  as approximation. For each  $\mathbf{s}_i$ , let  $N(\mathbf{s}_i) \subset \mathcal{S} \setminus \{\mathbf{s}_i\}$  be its conditioning set.

$$p(w_{\mathcal{S}}) \approx \tilde{p}(w_{\mathcal{S}}) = \prod_{i=1}^k p(w(\mathbf{s}_i) \mid w_{N(\mathbf{s}_i)})$$

Note that  $\tilde{w}(w_{\mathcal{S}})$  is not guaranteed to induce a valid joint distribution for  $w_{\mathcal{S}}$

# Finite-dimensional probability models with sparsity

- 8 Let  $N_{\mathcal{S}} = \{N(\mathbf{s}_i), i = 1, \dots, k\}$  be the collection of conditioning sets over  $\mathcal{S}$ .  $\mathcal{G} = \{\mathcal{S}, N_{\mathcal{S}}\}$  can be viewed as a directed Graph. For example,

$$m = 2$$

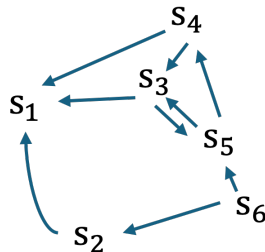
$$N(s_2) = \{s_1\}$$

$$N(s_3) = \{s_1, s_5\}$$

$$N(s_4) = \{s_1, s_3\}$$

$$N(s_5) = \{s_3, s_4\}$$

$$N(s_6) = \{s_2, s_5\}$$



# Finite-dimensional probability models with sparsity

- ③ Let  $N_{\mathcal{S}} = \{N(\mathbf{s}_i), i = 1, \dots, k\}$  be the collection of conditioning sets over  $\mathcal{S}$ .  $\mathcal{G} = \{\mathcal{S}, N_{\mathcal{S}}\}$  can be viewed as a directed Graph. For example,

$$m = 2$$

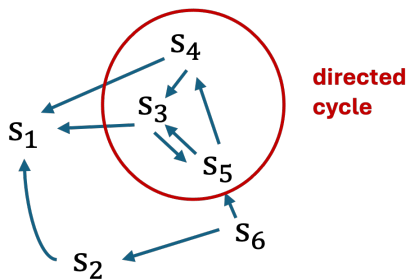
$$N(s_2) = \{s_1\}$$

$$N(s_3) = \{s_1, s_5\}$$

$$N(s_4) = \{s_1, s_3\}$$

$$N(s_5) = \{s_3, s_4\}$$

$$N(s_6) = \{s_2, s_5\}$$



# Finite-dimensional probability models with sparsity

- ③ Let  $N_{\mathcal{S}} = \{N(\mathbf{s}_i), i = 1, \dots, k\}$  be the collection of conditioning sets over  $\mathcal{S}$ .  $\mathcal{G} = \{\mathcal{S}, N_{\mathcal{S}}\}$  can be viewed as a directed Graph. For example,

$$m = 2$$

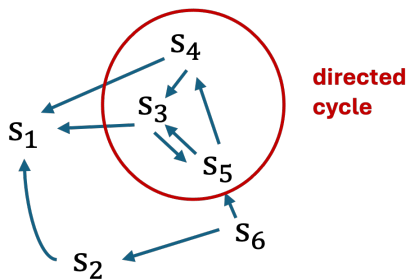
$$N(s_2) = \{s_1\}$$

$$N(s_3) = \{s_1, s_5\}$$

$$N(s_4) = \{s_1, s_3\}$$

$$N(s_5) = \{s_3, s_4\}$$

$$N(s_6) = \{s_2, s_5\}$$



A directed graph without directed cycles is known as a *directed acyclic graph* (DAG).

- If  $\mathcal{G}$  is a DAG,  $\tilde{p}(w_S)$  defines a proper multivariate joint density.

$$\begin{aligned}\tilde{p}(w_S) &= \prod_{i=1}^k p(w(\mathbf{s}_i) \mid w_{N(\mathbf{s}_i)}) \\ &= \prod_{i=1}^k N(w(\mathbf{s}_i) \mid B_{\mathbf{s}_i} w_{N(\mathbf{s}_i)}, F_{\mathbf{s}_i}),\end{aligned}\tag{1}$$

where  $B_{\mathbf{s}_i} = C_{\mathbf{s}_i, N(\mathbf{s}_i)} C_{N(\mathbf{s}_i)}^{-1}$ , and  $F_{\mathbf{s}_i} = C_{\mathbf{s}_i} - C_{\mathbf{s}_i, N(\mathbf{s}_i)} C_{N(\mathbf{s}_i)}^{-1} C_{N(\mathbf{s}_i), \mathbf{s}_i}$ .

It can be shown that  $\tilde{p}(w_S)$  is a multivariate normal, with covariance matrix  $\tilde{C}_S$ , and the precision matrix  $\tilde{C}_S^{-1}$  is sparse. We call  $\tilde{p}(w_S)$  the *nearest neighbor density* of  $w_S$ .

- 5 One way to ensure  $\mathcal{G}$  is a DAG is to always choose neighbor sets  $N(\mathbf{s}_j) \subset \{\mathbf{s}_1, \dots, \mathbf{s}_{j-1}\}$ . In summary, we need to:
- Order locations in the reference set:  $\mathbf{s}_1, \dots, \mathbf{s}_k$
  - For each location  $\mathbf{s}_j$ , select at most  $m$  locations from  $\{\mathbf{s}_1, \dots, \mathbf{s}_{j-1}\}$  to form  $N(\mathbf{s}_j)$

- 5 One way to ensure  $\mathcal{G}$  is a DAG is to always choose neighbor sets  $N(\mathbf{s}_j) \subset \{\mathbf{s}_1, \dots, \mathbf{s}_{j-1}\}$ . In summary, we need to:

- Order locations in the reference set:  $\mathbf{s}_1, \dots, \mathbf{s}_k$
- For each location  $\mathbf{s}_j$ , select at most  $m$  locations from  $\{\mathbf{s}_1, \dots, \mathbf{s}_{j-1}\}$  to form  $N(\mathbf{s}_j)$

[Datta et al., 2016] follows [Vecchia, 1988]'s methods:

- Order the locations based on either one of the two coordinates
- Choose the nearest (at most)  $m$  neighbors based on Euclidean distance (hence NNGP requires isotropic covariance functions)



# Finite-dimensional probability models with sparsity

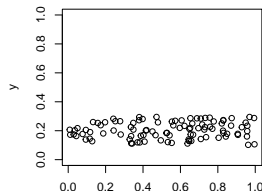
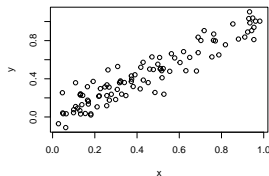
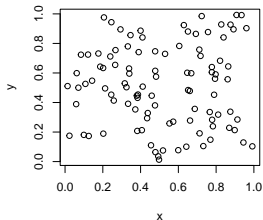
- 5 One way to ensure  $\mathcal{G}$  is a DAG is to always choose neighbor sets  $N(\mathbf{s}_j) \subset \{\mathbf{s}_1, \dots, \mathbf{s}_{j-1}\}$ . In summary, we need to:

- Order locations in the reference set:  $\mathbf{s}_1, \dots, \mathbf{s}_k$
- For each location  $\mathbf{s}_j$ , select at most  $m$  locations from  $\{\mathbf{s}_1, \dots, \mathbf{s}_{j-1}\}$  to form  $N(\mathbf{s}_j)$

[Datta et al., 2016] follows [Vecchia, 1988]'s methods:

- Order the locations based on either one of the two coordinates
- Choose the nearest (at most)  $m$  neighbors based on Euclidean distance (hence NNGP requires isotropic covariance functions)

As an example, consider the following data patterns:



# Extension to a Gaussian process

For any finite set  $\mathcal{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_r\}$  such that  $\mathcal{S} \cap \mathcal{U} = \emptyset$ , the density of  $w_{\mathcal{U}}$  conditional on  $w_{\mathcal{S}}$  can again be decomposed as

$$p(w_{\mathcal{U}} \mid w_{\mathcal{S}}) = p(w(\mathbf{u}_1) \mid w_{\mathcal{S}}) p(w(\mathbf{u}_2) \mid w(\mathbf{u}_1), w_{\mathcal{S}}) \dots p(w(\mathbf{u}_r) \mid w(\mathbf{u}_1), \dots, w(\mathbf{u}_{r-1}), w_{\mathcal{S}})$$

Let  $N(\mathbf{u}_i)$  be the  $m$ -nearest neighbor set of  $\mathbf{u}_i$  in  $\mathcal{S}$ . We can apply the same approximation and obtain the nearest neighbor density of  $w_{\mathcal{U}}$  conditional on  $w_{\mathcal{S}}$

$$\tilde{p}(w_{\mathcal{U}} \mid w_{\mathcal{S}}) = \prod_{i=1}^r p(w(\mathbf{u}_i) \mid w_{N(\mathbf{u}_i)}).$$

We can then use  $\tilde{p}(w_{\mathcal{S}})$  to obtain the nearest neighbor density of  $w_{\mathcal{U}}$ , and extend it to any finite set  $\mathcal{V} \subset \mathcal{D}$ .

# Extension to a Gaussian process

Let  $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$  be any finite set in  $\mathcal{D}$ . Let  $\mathcal{U} = \mathcal{V} \setminus \mathcal{S}$ . Then the nearest neighbor density of  $\mathcal{V}$  can be obtained as

$$\tilde{p}(w_{\mathcal{V}}) = \int \tilde{p}(w_{\mathcal{U}} \mid w_{\mathcal{S}}) \tilde{p}(w_{\mathcal{S}}) \prod_{\mathbf{s}_i \in \mathcal{S} \setminus \mathcal{V}} d(w(\mathbf{s}_i)).$$

It can be shown that these finite dimensional probability densities conform to Kolmogorov's consistency criteria, hence correspond to a valid Gaussian process over  $\mathcal{D}$ , with covariance function  $C^{NN}(\mathbf{v}_1, \mathbf{v}_2 \mid \theta)$

$$= \begin{cases} \tilde{C}(\mathbf{s}_i, \mathbf{s}_j) & \text{if } \mathbf{v}_1 = \mathbf{s}_i, \mathbf{v}_2 = \mathbf{s}_j \\ B_{\mathbf{v}_1} \tilde{C}_{N(\mathbf{v}_1), \mathbf{s}_j} & \text{if } \mathbf{v}_1 \notin \mathcal{S}, \mathbf{v}_2 = \mathbf{s}_j \\ B_{\mathbf{v}_1} \tilde{C}_{N(\mathbf{v}_1), N(\mathbf{v}_2)} B_{\mathbf{v}_2}^T + \delta(\mathbf{v}_1 = \mathbf{v}_2) F_{\mathbf{v}_1} & \text{if } \mathbf{v}_1 \notin \mathcal{S}, \mathbf{v}_2 \notin \mathcal{S}, \end{cases}$$

where  $B_{\mathbf{v}_1}$  and  $F_{\mathbf{v}_1}$  are defined analogously to equation (1).  $C^{NN}$  is continuous with probability 1.

To summarize, given

- a parent Gaussian process with isotropic covariance function  $C$
- a fixed reference set  $\mathcal{S}$ , and
- neighbor sets chosen using Vecchia's method with size  $m$ ,

we have constructed another Gaussian process with covariance function  $C^{NN}$ .

- It is called the *nearest neighbor Gaussian process*
- Its finite-dimensional distributions have sparse precision matrices
- $C^{NN}$  is non-stationary.

# Application to Bayesian spatial models

Recall the Bayesian spatial model at the beginning of the presentation:

$$\log y_i = \mathbf{x}_i^T \boldsymbol{\beta} + w(\mathbf{s}_i) + \epsilon_i, \quad \epsilon_i \sim \text{i.i.d. } N(0, \sigma^2), \quad \text{where}$$

- $y_i$  is the time to readmission for patient  $i$
- $\mathbf{x}_i \in \mathbb{R}^3$  are the covariates for patient  $i$  including an intercept, gender, and age
- $\mathbf{s}_i \in \mathbb{R}^2$  is the location for patient  $i$
- $w(\mathbf{s}_i)$  is the spatial intercept.
  - Instead of a GP prior, we can use an NNGP prior,  $w(\mathbf{s}) \sim GP(0, C^{NN}(\cdot, \cdot | \boldsymbol{\theta}))$
  - In each MCMC iteration, we only need to invert matrices of size up to  $m \times m$
- With priors for  $\boldsymbol{\beta}$ ,  $\sigma^2$ , and  $\boldsymbol{\theta}$

# Computational complexity

Total flop counts per MCMC iteration:

- NNGP:  $\mathcal{O}((n+k)m^3)$  vs GP:  $\mathcal{O}(n^3)$

Storage space:

- NNGP:  $(n+k)(m \times m)$  vs GP:  $n \times n$

# How to choose reference set $\mathcal{S}$

Choosing  $\mathcal{S}$  is similar to choosing basis functions in low-rank methods. However,

- recall in low-rank methods, flop counts scale quadratically in  $r$ , number basis functions
- here flop counts increase linearly in  $k$ , the size of  $\mathcal{S}$ .

So users can choose large  $\mathcal{S}$  if needed.

Two reasonable choices of  $\mathcal{S}$  are:

- 1 All observed data points – this gives computational advantage in Bayesian analysis
- 2 Points on a grid over the domain of interest

Empirical investigations suggest that for large datasets, these two options deliver indistinguishable inference results.

# How to choose reference set $\mathcal{S}$

However, for datasets with large gaps in the observed locations, suggest adding points in the gap to  $\mathcal{S}$  in addition to the observed data.

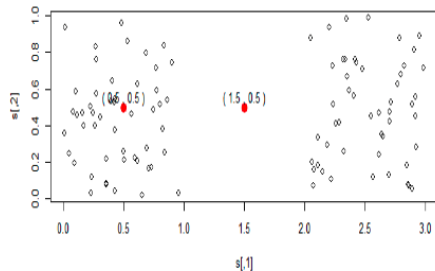


Figure: Figure A4 in Appendix of [Datta et al., 2016]



# How to choose size of neighbor sets $m$

Based on empirical simulation results, [Datta et al., 2016] suggests  $m \in [10, 15]$  is usually sufficient even for large datasets.

# Simulation setup

[Datta et al., 2016] used simulation study to compare the performance of NNGP vs GP.

- Generated observations using 2,500 locations. 2,000 locations were used for training, with the remaining 500 for out-of-model prediction validation.
- Univariate outcomes
- Modeled spatial intercept, with the true surface simulated from Matérn 1/2 kernel (i.e., exponential covariance function)

# Simulation results

**Table 1.** Univariate synthetic data analysis parameter estimates and computing time in minutes for NNGP and full GP models. Parameter posterior summary 50 (2.5, 97.5) percentiles.

		NNGP ( $S \neq \mathcal{T}$ )		NNGP ( $S = \mathcal{T}$ )	
True		$m = 10, k = 2000$	$m = 20, k = 2000$	$m = 10$	$m = 20$
$\beta_0$	1	0.99 (0.71, 1.48)	1.02 (0.73, 1.49)	1.00 (0.62, 1.31)	1.03 (0.65, 1.34)
$\beta_1$	5	5.00 (4.98, 5.03)	5.01 (4.98, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)
$\sigma^2$	1	1.09 (0.89, 1.49)	1.04 (0.85, 1.40)	0.96 (0.78, 1.23)	0.94 (0.77, 1.20)
$\tau^2$	0.1	0.07 (0.04, 0.10)	0.07 (0.04, 0.10)	0.10 (0.08, 0.13)	0.10 (0.08, 0.13)
$\phi$	12	11.81 (8.18, 15.02)	12.21 (8.83, 15.62)	12.93 (9.70, 16.77)	13.36 (9.99, 17.15)
$p_D$	—	1491.08	1478.61	1243.32	1249.57
DIC	—	1856.85	1901.57	2390.65	2377.51
G	—	33.67	35.68	77.84	76.40
P	—	253.03	259.13	340.40	337.88
D	—	286.70	294.82	418.24	414.28
RMSPE	—	1.22	1.22	1.2	1.2
95% CI cover %	—	97.2	97.2	97.6	97.6
95% CI width	—	2.19	2.18	2.13	2.12
Time	—	14.2	47.08	9.98	33.5
		Predictive process	Full		
True		64 knots	Gaussian process		
$\beta_0$	1	1.30 (0.54, 2.03)	1.03 (0.69, 1.34)		
$\beta_1$	5	5.03 (4.99, 5.06)	5.01 (4.99, 5.03)		
$\sigma^2$	1	1.29 (0.96, 2.00)	0.94 (0.76, 1.23)		
$\tau^2$	0.1	0.08 (0.04, 0.13)	0.10 (0.08, 0.12)		
$\phi$	12	<b>5.61 (3.48, 8.09)</b>	13.52 (9.92, 17.50)		
$p_D$	—	1258.27	1260.68		
DIC	—	13677.97	2364.80		
G	—	1075.63	74.80		
P	—	200.39	333.27		
D	—	1276.03	408.08		
RMSPE	—	1.68	1.2		
95% CI cover %	—	95.6	97.6		
95% CI width	—	2.97	2.12		
Time	—	43.36	560.31		

**Figure:** Table 1 of [Datta et al., 2016]

# Simulation results

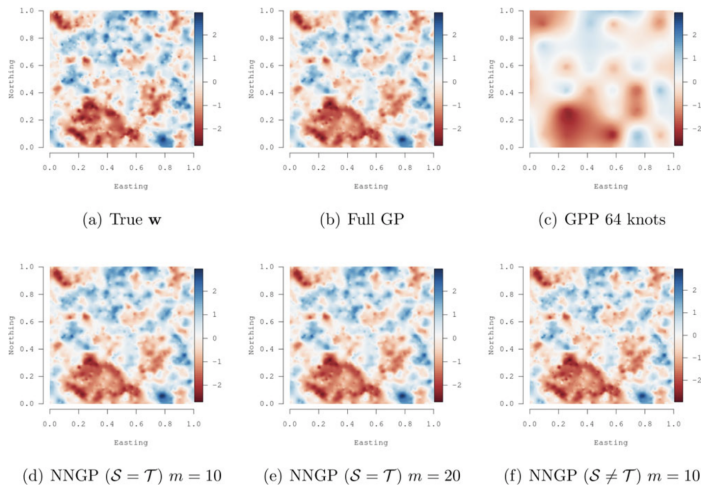


Figure 2. Univariate synthetic data analysis: interpolated surfaces of the true spatial random effects and posterior median estimates for different models.

Figure: Figure 2 of [Datta et al., 2016]

# Conclusion and recommendations

[Datta et al., 2016] introduced the nearest neighbor Gaussian process.

- It is a sparsity-based scalable model that works well for Bayesian hierarchical spatial models.
- It induces a valid Gaussian process.
- It is not only an approximation to GP, but can also be viewed as an independent scalable model.





I would recommend reading and experimenting NNGP

- foundation to a number of modern computational methods for GP
- tutorial codes available online

Potential Limitations

- not recommended for  $d \geq 4$

# References I

-  Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.
-  Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.
-  Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, pages 37–56. Springer.
-  Riutort-Mayol, G., Bürkner, P.-C., Andersen, M. R., Solin, A., and Vehtari, A. (2023). Practical hilbert space approximate bayesian gaussian processes for probabilistic programming. *Statistics and Computing*, 33(1):17.

# References II



Stein, M. L. (2014).

Limitations on low rank approximations for covariance matrices of spatial data.

*Spatial Statistics*, 8:1–19.



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.