A Divide-and-Conquer Bayesian Approach to Large-Scale Kriging

Cheng Li

DSAP, National University of Singapore

Joint work with Rajarshi Guhainiyogi (UC Santa Cruz), Terrance D. Savitsky (US Bureau of Labor Statistics), and Sanvesh Srivastava (U Iowa)
Outline

Background of GP

D&C Bayes

Method

Theory

Experiments

Future Work
Background: Gaussian Process

**Gaussian Process** is a commonly tool to model functions (surfaces).

- Let $s_1, \ldots, s_n$ be $n$ arbitrary different locations in a spatial domain $D$.
- $f(\cdot)$ is the mean function (which we want to model).
- $f_i = f(s_i)$ ($i = 1, \ldots, n$) are the functional values at the location $s_i$.
- A Gaussian process assumes that $f^\top = (f_1, \ldots, f_n)$ jointly follows $\mathcal{N}(f, C)$.
- $(C)_{ij} = C(s_i, s_j)$, $1 \leq i, j \leq n$. $C(\cdot, \cdot)$ is the covariance kernel function.
- Examples of covariance kernel function: $C(s, s') = \tau^2 e^{-\phi \|s-s'\|^2}$, $C(s, s') = \tau^2 e^{-\phi \|s-s'\|}$, $C(s, s') = \frac{\tau^2}{2^{\nu-1} \Gamma(\nu)} (\phi \|s-s'\|)^\nu K_\nu (\phi \|s-s'\|)$. 

![Graph and 3D plot of Gaussian Process](image)
Consider fitting a function $f$ with noise:

- $y(s) = f(s) + \epsilon(s)$.
- Observed locations $S = (s_1, \ldots, s_n)$; $y^\top = (y(s_1), \ldots, y(s_n))$; $f^\top = (f(s_1), \ldots, f(s_n))$.
- We impose a Gaussian process (GP) prior on $f$: $f \sim \mathcal{N}(0, C(S, S))$. $(C(S, S))_{ij} = C(s_i, s_j)$, $1 \leq i, j \leq n$.
- $\epsilon(s)$ is a white noise process independent of $f$; $\text{Var}(\epsilon(s)) = \sigma^2$.
- Let $f^* = f(s^*)$ be the functional value at a testing location $s^*$. Then the joint distribution of $(y, f^*)$ is

$$
\begin{bmatrix}
y \\
f^*
\end{bmatrix} \sim \mathcal{N}
\left(
\begin{bmatrix}
0 \\
0
\end{bmatrix},
\begin{bmatrix}
C(S, S) + \sigma^2 I_n & C(S, s^*) \\
C(s^*, S) & C(s^*, s^*)
\end{bmatrix}
\right)
$$

- The GP posterior of $f^*$ given the observed data $y$ is Gaussian, with

$$
E[f^* \mid y] = C(s^*, S) \left[ C(S, S) + \sigma^2 I_n \right]^{-1} y,
$$

$$
\text{Var}[f^* \mid y] = C(s^*, s^*) - C(s^*, S) \left[ C(S, S) + \sigma^2 I_n \right]^{-1} C(S, s^*).
$$
Inverting the matrix $C(S, S) + \sigma^2 I_n$ requires $O(n^3)$ computational cost and $O(n^2)$ storage cost.

For Bayesian methods, learning the parameters in the model (such as the parameters in the kernel $C(\cdot, \cdot)$ and $\sigma^2$) often requires inverting the matrix $C(S, S) + \sigma^2 I_n$ repeatedly.

For commonly used kernels, such as squared exponential kernel and Matérn kernel, the matrix $C(S, S)$ tends to have singularity problem when $n$ becomes large, say $n \gtrsim 10^4$.

This is one of the main motivations that drive the research on GP.
Existing Methods to Speed up GP

The general idea is to simplify the structure of $C(\cdot, \cdot)$.

- **Low-rank methods**: Fit the spatial surface $f$ using $r$ chosen basis functions or “knots”.
  Fixed-rank kriging (Cressie & Johannesson 08’JRSSB),
  Predictive process (Banerjee et al. 08’JRSSB, Finley et al. 09’ CSDA).

- **Compacted supported covariance kernel**: Covariance tapering (Kaufman, Schervish & Nychka 08’JASA).

- **Sparsity on inverse covariance matrix**: Composite likelihood (Stein, Chi & Welty 04’JRSSB),
  Nearest Neighbor GP (Datta et al. 16’JASA, Datta et al. 16’AOAS)

- **Block structured covariance**: Fitting GP on sub-regions (Nychka et al. 15’JCGS, Katzfuss 17’JASA, Guhaniyogi and Banerjee 18’JCGS).

- **Related methods in machine learning**: Subset of regressors (Quinonero-Candela & Rasmussen 05’JMLR); Inducing point methods (Wilson & Nickisch 15’ICML), etc.
Problems with Existing Methods

- Low-rank methods are popular for large spatial datasets. In this work, we compare mainly with the modified predictive process (MPP) (Finley et al. 09’CSDA).

- MPP reduces the computation complexity of GP from $O(n^3)$ to $O(nr^2)$, where $r$ is the number of “knots” (chosen basis functions).

- However, a very large $n$ also requires a much larger $r$ ($\approx \sqrt{n}$), in order to achieve the same level of accuracy. For example, if $n$ is about 1 million, then the number of knots needs to be a few thousand.

- Nearest Neighbor GP (NNGP, Datta et al. 16’JASA, Datta et al. 16’AOAS) works well for rough surfaces but not as well for smooth surfaces.
Goals

- We need **fast and scalable** computational methods to fit GP to spatial datasets with \( \approx 1 \text{ million observations} \) in reasonable amount of time (e.g. in about 1 day).

- We develop under the Bayesian framework with posterior samplers that can be practically easy to implement.

- We need theoretical guarantees for the validity of uncertainty quantification.
Outline

Background of GP

D&C Bayes

Method

Theory

Experiments

Future Work
Divide-and-Conquer Approach for Big Data

General steps:

- Divide the massive datasets of size $n$ into $k$ non-overlapping subsets;
- Perform statistical inference (estimation, posterior sampling, etc.) on each of the $k$ subsets;
- Combine the $k$ subset results into a global result (averaged estimator, posterior distribution, etc.)

Divide-and-conquer simple and effective:

- Allows parallel computing;
- Significantly reduces computational costs locally. Each machine only handles data of size $m = n/k$. 
Consensus Monte Carlo (CMC, Scott et al. 16')
Semiparametric density product (SDP, Neiswanger et al. 14')
Wasserstein posterior (WASP, Srivastava et al. 15'), PIE (Li et al. 17')
Double-parallel MCMC (DP-MCMC, Xue & Liang 17')

Basic idea: When the data are all independent

\[
\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \\
\propto \left[\prod_{j=1}^{k} j\text{th subset likelihood}\right] \times \text{prior} \\
\propto \prod_{j=1}^{k} \left[\text{\text{jth subset likelihood} \times \text{prior}^{1/k}}\right] \quad (\text{CMC, SDP}) \\
\propto \left[\prod_{j=1}^{k} \left(\text{jth subset likelihood} \right)^{k} \times \text{prior}\right]^{1/k} \quad (\text{WASP, PIE, DP-MCMC})
\]
Divide and Conquer Bayes

Idea: Run MCMC in parallel on disjoint subset data, and combine the subset posterior distributions into a global posterior.
Divide and Conquer Bayes

PIE (Li, Srivastava, and Dunson 17’): Average the subset posterior empirical quantiles. Leading to the “Wasserstein-2 barycenter”, as an approximation to the full data posterior.
PIE Algorithm in 3 Steps

For a 1-dimensional parameter $\xi \in \mathbb{R}$,

- **Step 1:** Run MCMC on $k$ subsets on $k$ machines in parallel. Draw posterior samples from each subset posterior. Get the posterior samples for $\xi$.

- **Step 2:** Find the $q$th empirical quantiles of $\xi$, for each of the $k$ subset posterior distributions. $q \in (0, 1)$ can come from a fine grid on $(0, 1)$.

- **Step 3:** Average the $k$ subset quantiles.

**Output:** Averaged quantiles for $\xi \implies$

- A combined distribution for $\xi$, as an approximation to the true posterior of $\xi$ based on the full data.
- Posterior credible intervals.

For independent data, this method (as well as many other similar ones) is asymptotically valid for approximating the true posterior of $\xi$ given the full data.
Wasserstein Barycenter

- Wasserstein-2 ($W_2$) distance between two measures $\mu$ and $\nu$:

$$W_2(\mu, \nu) = \inf \left\{ \sqrt{\mathbb{E}(X - Y)^2}, \text{law}(X) = \mu, \text{law}(Y) = \nu \right\}$$

- For univariate distribution functions $F_1$ and $F_2$, their $W_2$ distance has an equivalent form:

$$W_2(F_1, F_2) = \sqrt{\int_0^1 \left[ F_1^{-1}(q) - F_2^{-1}(q) \right]^2 dq}$$

where $F_1^{-1}, F_2^{-1}$ are quantile functions of $F_1, F_2$.

- The Wasserstein-2 barycenter of $k$ distributions $\mu_1, \ldots, \mu_k$ is

$$\overline{\mu} = \arg \min_{\mu \in \mathcal{P}_2} \frac{1}{k} \sum_{j=1}^{k} W_2^2 (\mu, \mu_j),$$

where $\mathcal{P}_2$ is the space of all distributions with finite 2nd moment.
Outline

Background of GP

D&C Bayes

Method

Theory

Experiments

Future Work
A GP-based Spatial Model

In this work, we consider a generalized GP regression model with spatial predictors:

\[ y(s) = x(s)^T \beta + w(s) + \epsilon(s). \]

\[ \text{Observed locations } S = \{s_1, \ldots, s_n\}; \ y = (y(s_1), \ldots, y(s_n))^T; \]
\[ w = (w(s_1), \ldots, w(s_n))^T. \]

\[ x(s) \text{ is a } p\text{-dimensional predictor vector at location } s. \]
\[ X = (x(s_1), \ldots, x(s_n))^T. \]

\[ \text{We impose a Gaussian process (GP) prior on } w: \ w \sim \mathcal{N}(0, C_\alpha(S, S)). \]
\[ (C_\alpha(S, S))_{ij} = C_\alpha(s_i, s_j), \ 1 \le i, j \le n. \]

\[ \epsilon(s) \text{ is a white noise process independent of } w \text{ and } x; \ \text{Var}(\epsilon(s)) = \sigma^2. \]

\[ \text{Let } s^* \text{ be a “testing location” and } s^* \notin S. \text{ Let} \]
\[ w^* = w(s^*), y^* = y(s^*), x^* = x(s^*). \]

\[ \text{The prior for } \beta \text{ is } \mathcal{N}(\mu_\beta, \Sigma_\beta). \text{ A prior is also imposed on } \alpha, \sigma^2. \]
Posterior Sampling Algorithm

Iterate the following steps:

- Sample $\beta$ given $y, X, \alpha, \sigma^2$ from $N(m_\beta, V_\beta)$, where

$$V_\beta = \left\{ X^T V(\alpha, \sigma^2)^{-1} X + \Sigma_\beta \right\}^{-1},$$

$$m_\beta = V_\beta \left\{ X^T V(\alpha, \sigma^2)^{-1} y + \Sigma_\beta^{-1} \mu_\beta \right\},$$

$$V(\alpha, \sigma^2) = C_\alpha(S, S) + \sigma^2 I_n.$$  

- Sample $(\alpha, \sigma^2)$ given $y, X, \beta$ using the random walk Metropolis-Hastings.

Then $w^*, y^*$ can be drawn as

- $w^* \mid y, X, \beta, \alpha, \sigma^2 \sim N(m^*, V^*)$, where

$$m^* = C_\alpha(s^*, S)V(\alpha, \sigma^2)^{-1}(y - X\beta),$$

$$V^* = C_\alpha(s^*, s^*) - C_\alpha(s^*, S)V(\alpha, \sigma^2)^{-1}C_\alpha(S, s^*).$$

- $y^* \mid y, X, x^*, w^*, \beta, \alpha, \sigma^2 \sim N(x^*\beta + w^*, \sigma^2)$. 
Distributed Kriging (DISK)

What can we do for big $n$ like $n = 10^6$?

We propose a D&C Bayesian method called DISK.

For short, **DISK** = (Existing D&C Bayes) + (Spatially Dependent Data)

We fit the GP-based spatial model on $k$ different subsets of data, and then combine the results.

**Partition scheme**: Uniform sampling - such that every subset is **representative** of the full data.

For example, if PIE (posterior interval estimation, Li et al 17’) is used, then for each parameter in $(\beta, \alpha, \sigma^2, w^*, y^*)$, the marginal posterior is obtained by the following steps:
Partition the locations into $k$ subsets $S = \bigcup_{j=1}^{k} S_j$, $S_j \cap S_{j'} = \emptyset$. Let $n = km$, where $n$ is the total sample size and $m$ is the subset size.

Fit the GP regression model for the $j$th subset ($j = 1, \ldots, k$)

$$y(s_{ji}) = x(s_{ji})^T \beta + w(s_{ji}) + \epsilon(s_{ji}), \quad i = 1, \ldots, m,$$

by using the modified posterior

$$\pi_m(\beta, \alpha, \sigma^2 \mid y_j, X_j) \propto \left\{ \ell_j(\beta, \alpha, \sigma^2) \right\}^k \cdot \pi(\beta, \alpha, \sigma^2),$$

$$\ell_j(\beta, \alpha, \sigma^2) = N \left( X_j \beta, V_j(\alpha, \sigma^2) \right),$$

$$V_j(\alpha, \sigma^2) = C_\alpha(S_j, S_j) + \sigma^2 I_m,$$

where $(y_j, X_j)$ is the $j$th subset data at locations $S_j$.

Take the $q$th empirical quantiles of the marginals of $\pi_m(\beta, \alpha, \sigma^2 \mid y_j, X_j)$ for $j = 1, \ldots, k$, and then average them.

Perform the same combination for the predictive posteriors of $w^*$ and $y^*$. 
Outline

Background of GP

D&C Bayes

Method

Theory

Experiments

Future Work
Theory

- Our theoretical results mainly focus on the posterior of $w^*$, the values function $w$ at a testing location $s^*$.

- We present upper bounds for the Bayes $L_2$ risk, and a Bayesian version of bias-variance tradeoff.

- We first show the results for the simplified model $y(s) = w(s) + \epsilon(s)$ without the $x(s)\top \beta$ part, and then generalize to the full model $y(s) = x(s)\top \beta + w(s) + \epsilon(s)$.

- To ease the technical difficulty, we assume that the finite-dimensional parameters ($\alpha, \sigma^2$) are fixed (because their estimation is notoriously difficult ...)

- Furthermore, we assume that the testing location $s^*$ is selected independently from the same distribution that generates the training locations $S = \{s_1, \ldots, s_n\}$. 
The Bayes $L_2$ risk of the DISK posterior is defined as

$$\text{Bayes } L_2 \text{ risk} = \mathbb{E}_S \mathbb{E}_{0|S} \mathbb{E}_{s^*} [\mathbb{w}(s^*) - w_0(s^*)]^2,$$

where $\mathbb{E}_{s^*}$ is the expectation w.r.t. the sampled testing location, $\mathbb{E}_{0|S}$ is the DISK posterior expectation given the training locations $S$, and $\mathbb{E}_S$ is the expectation w.r.t. the training location.

This Bayes $L_2$ risk can be decomposed into 3 parts:

$$\text{Bayes } L_2 \text{ risk} = \text{bias}^2 + \text{var}_{\text{mean}} + \text{var}_{\text{DISK}}$$

They are the squared bias, the variance of subset posterior means ($\text{var}_{\text{mean}}$), and the mean of subset variances ($\text{var}_{\text{DISK}}$).

The reason for 2 variance terms is the law of total variance

$$\text{var}(Y) = \text{var}[E(Y|X)] + E[\text{var}(Y|X)].$$
Some Notation

▶ Let \( \mathbb{P}_s \) be the sampling distribution of \( s_1, \ldots, s_n, s^* \).

▶ Let \( \{\phi_i\}_{i=1}^\infty \) be an orthonormal basis of \( L_2(\mathbb{P}_s) \) w.r.t. \( \mathbb{P}_s \).

▶ The kernel has the series expansion \( C_\alpha(s, s') = \sum_{i=1}^\infty \mu_i \phi_i(s) \phi_i(s') \) w.r.t. \( \mathbb{P}_s \) for any \( s, s' \in D \), where \( \mu_1 \geq \mu_2 \geq \ldots \geq 0 \) are the eigenvalues of \( C_\alpha \).

▶ The trace of the kernel \( C_\alpha \) is defined as \( \text{tr}(C_\alpha) = \sum_{i=1}^\infty \mu_i \).

▶ For any \( f \in L_2(\mathbb{P}_s) \), \( f(s) = \sum_{i=1}^\infty \theta_i \phi_i(s) \), where \( \theta_i = \langle f, \phi_i \rangle_{L_2(\mathbb{P}_s)} \).

▶ The reproducing kernel Hilbert space (RKHS) \( \mathcal{H} \) attached to \( C_\alpha \) is the space of all functions \( f \in L_2(\mathbb{P}_s) \) such that the \( \mathcal{H} \)-norm \( \|f\|_\mathcal{H} = \sum_{i=1}^\infty \theta_i^2 / \mu_i < \infty \).
Bayes Bias-Variance Tradeoff

More technical assumptions:

- **Assumption 1**: The true surface \( w_0 \) lies in the RKHS (reproducing kernel Hilbert space) of \( C_\alpha \). (This can be relaxed by using sieve approximation.)

- **Assumption 2**: The covariance kernel \( C_\alpha \) satisfies \( \text{tr}(C_\alpha) < \infty \).

- **Assumption 3**: There are positive constants \( \rho \) and \( r \geq 2 \), such that \( \mathbb{E}_{P_s}[\phi_i^{2r}(s)] \leq \rho^{2r} \) for every \( i = 1, 2, \ldots, \infty \), and \( \text{var} [\epsilon(s)] \leq \sigma_0^2 < \infty \) for all \( s \in D \).

- We derive **upper bounds** for the three terms in the Bayes \( L_2 \) risk, and show that (roughly speaking) for a given total sample size \( n \),

\[
\text{as } k \uparrow, \quad \text{bias}^2 \uparrow, \quad \text{var}_{\text{mean}} \downarrow, \quad \text{var}_{\text{DISK}} \lesssim \text{bias}^2 + \text{var}_{\text{mean}},
\]

where \( k \) is the number of subsets.
Bayes Bias-Variance Tradeoff

\[
\text{bias}^2 \leq \frac{8\sigma_0^2}{n} \| w_0 \|_H^2 + \| w_0 \|_H^2 \inf_{d \in \mathbb{N}} \left[ \frac{8n}{\sigma_0^2} \rho^4 \text{tr}(C_\alpha)\text{tr}(C_\alpha^d) + \mu_1 \left\{ \frac{Ab(m, d, r)}{\sqrt{m}} + \frac{A b(m, d, r) \rho^2 \gamma(\frac{\sigma_0^2}{n})}{\sqrt{m}} \right\} r \right]
\]

\[
\text{var}_{\text{mean}} \leq \frac{2n + 4\| w_0 \|_H^2}{k} \inf_{d \in \mathbb{N}} \left[ \frac{13n}{\sigma_0^2} \rho^4 \text{tr}(C_\alpha)\text{tr}(C_\alpha^d) + \left\{ \frac{Ab(m, d, r) \rho^2 \gamma(\frac{\sigma_0^2}{n})}{\sqrt{m}} \right\} r \right] +
\]

\[
\frac{12 \sigma_0^2}{k \cdot n} \| w_0 \|_H^2 + 12 \frac{\sigma_0^2}{n} \gamma(\frac{\sigma_0^2}{n}),
\]

\[
\text{var}_{\text{DISK}} \leq \frac{3\sigma_0^2}{n} \gamma(\frac{\sigma_0^2}{n}) + \inf_{d \in \mathbb{N}} \left[ \frac{5n}{\sigma_0^2} \text{tr}(C_\alpha)\text{tr}(C_\alpha^d) + \text{tr}(C_\alpha) \left\{ \frac{Ab(m, d, r) \rho^2 \gamma(\frac{\sigma_0^2}{n})}{\sqrt{m}} \right\} r \right],
\]

where \( \mathbb{N} \) is the set of all positive integers, \( A \) is a positive constant,

\[
b(m, d, r) = \max \left( \sqrt{\max(r, \log d)}, \frac{\max(r, \log d)}{m^{1/2-1/r}} \right),
\]

\[
\gamma(a) = \sum_{i=1}^{\infty} \frac{\mu_i}{\mu_i + a} \text{ for any } a > 0, \quad \text{tr}(C_\alpha^d) = \sum_{i=d+1}^{\infty} \mu_i.
\]
Examples of Covariance Kernels

Implications from the previous results for some kernel classes (with $r > 4$):

- **Finite rank kernels:** if $\mu_{d^*+1} = \mu_{d^*2} = \ldots = 0$ for some integer $d^* \geq 1$, and $k \leq cn^{r-2}/(\log n)^{2r/r-2}$ for some constant $c > 0$, then the Bayes $L_2$ risk is $O(n^{-1})$.

- **Exponentially decaying kernels:** if $\mu_i \leq c_{1\mu} \exp(-c_{2\mu}i^\kappa)$ for some constants $c_{1\mu} > 0$, $c_{2\mu} > 0$, $\kappa > 0$ and all $i \in \mathbb{N}$, and $k \leq cn^{r-2}/(\log n)^{2\kappa(r+\kappa-1)/\kappa(r-2)}$ for some constant $c > 0$, then the Bayes $L_2$ risk is $O((\log n)^{1/\kappa}/n)$.

- **Polynomially decaying kernels:** if $\mu_i \leq c_\mu i^{-2\nu}$ for some constants $c_\mu > 0$, $\nu > \frac{r-1}{r-4}$ and all $i \in \mathbb{N}$, and $k \leq cn^{(r-4)\nu-(r-1)/2}/(\log n)^{2r/r-2}$ for some constant $c > 0$, then the Bayes $L_2$ risk is $O(n^{-2\nu-1/2\nu})$. 
Examples of Covariance Kernels

Some comments:

- The rates for finite rank kernels and exponentially decaying kernels are near minimax optimal up to logarithm factors.

- The rate for polynomially decaying kernels is suboptimal, but the difference is small. The minimax optimal rate is $O\left(n^{-\frac{2\nu}{2\nu+1}}\right)$. The gap between the two rates is $n^{-\frac{1}{2\nu(2\nu+1)}}$.

- The reason for sub-optimality is due to our fixed stochastic approximation (likelihood raised to the fixed power of $k$). One can possibly make it optimal by using an adaptive power in the subset likelihood.
Examples of Covariance Kernels

(Convenient) generalization to the full model \( y(s) = x(s)^T \beta + w(s) + \epsilon(s) \):

- Extra Assumption: \( x(\cdot) \) is a deterministic function. The prior of \( \beta \) is normal \( N(\mu_\beta, \Sigma_\beta) \), and is independent of the prior on \( w(\cdot) \), which is \( \text{GP}(0, C_\alpha(\cdot, \cdot)) \).

- Based on this assumption, we can do a change of kernel: \( \tilde{C}_\alpha(s_1, s_2) = \text{Cov} \{ x(s_1)^T \beta + w(s_1), x(s_2)^T \beta + w(s_2) \} = x(s_1)^T \Sigma_\beta x(s_2) + C_\alpha(s_1, s_2) \).

- Our theory results hold for the mean function \( x(s)^T \beta + w(s) \) in the full model, as long as all the previous conditions on \( C_\alpha \) is satisfied for the new kernel \( \tilde{C}_\alpha \).

- Since we only focus on the predictive loss on the mean function \( x(s)^T \beta + w(s) \), we do not need any identification condition for \( \beta \) and \( w(\cdot) \).
Outline

Background of GP

D&C Bayes

Method

Theory

Experiments

Future Work
Simulation Model 1: Smooth Surface

- \( \mathcal{D} = [-2, 2] \times [-2, 2] \).
- \( f(s) = e^{-(s-1)^2} + e^{-0.8(s+1)^2} - 0.05 \sin\{8(s + 0.1)\} \), for \( s \in [-2, 2] \).
- The true surface is \( w_0(s) = -f(s_1)f(s_2), \ s = (s_1, s_2) \in \mathcal{D} \).
- The response is \( y(s) = \beta_0 + w_0(s) + \epsilon, \ \epsilon \sim N(0, \sigma^2) \).
- \( \beta_0 = 1, \ \sigma^2 = 0.01 \).
- We use the Matérn kernel with \( \nu = 1/2 \): \( C(s, s') = \tau^2 \exp(-\phi\|s - s'\|) \) for \( s, s' \in \mathcal{D} \).
- \( \sigma^2 \sim \text{InvGamma}(2, 0.1), \ \tau^2 \sim \text{InvGamma}(2, 2), \ \phi \sim \text{Uniform}(0.1, 1) \).
- In 2 simulation scenarios, we use \( n = 10^4 \) and \( n = 10^6 \), respectively. The training locations are sampled uniformly over \( \mathcal{D} \).
- We use \( l = 2025 \) testing locations, also sampled uniformly over \( \mathcal{D} \).
- The Bayes \( L_2 \) risks (and other criteria) are estimated by averaging over the \( l \) testing locations.
- All methods are replicated for 10 Monte Carlo simulations.
Methods Compared

- LatticeKrig (Nychka et al. 15’) using the LatticeKrig package in R, with 3 number of resolutions.

- Locally approximated Gaussian process (laGP, Gramacy and Apley 15’) using the laGP package in R with method set to “nn”.

- Full-rank Gaussian process (GP) using the spBayes package in R with the full data.

- Modified predictive process (MPP, Finley et al. 15’) using the spBayes package in R with the full data.

- Nearest neighbor Gaussian process (NNGP, Datta et al. 15’) using the spNNGP package in R with the number of nearest neighbors (NN) as 5, 15, and 25.

- DISK + CMC (Scott et al. 16’) + (Full GP or MPP).

- DISK + PIE (Li et al. 17’) + (Full GP or MPP).

The first two methods are frequentist; the others are all Bayesian.
Simulation Results When $n = 10^4$

Prediction MSE averaged over $l = 2025$ locations

<table>
<thead>
<tr>
<th>Method</th>
<th>LaGP</th>
<th>LatticeKrig</th>
<th>Full GP</th>
<th>MPP ($r = 200$)</th>
<th>MPP ($r = 400$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNGP</td>
<td>NN= 5</td>
<td>NN= 15</td>
<td>NN= 25</td>
<td>0.3842</td>
<td>0.0727</td>
</tr>
<tr>
<td></td>
<td>0.4233</td>
<td>0.4835</td>
<td>0.4900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISK-CMC(MPP) $r = 200$</td>
<td>$k = 10$</td>
<td>$k = 20$</td>
<td>$k = 30$</td>
<td>0.0436</td>
<td>0.0493</td>
</tr>
<tr>
<td></td>
<td>0.0436</td>
<td>0.0493</td>
<td>0.0422</td>
<td>0.0385</td>
<td>0.0364</td>
</tr>
<tr>
<td>DISK-CMC(MPP) $r = 400$</td>
<td>$k = 10$</td>
<td>$k = 20$</td>
<td>$k = 30$</td>
<td>0.0455</td>
<td>0.0422</td>
</tr>
<tr>
<td></td>
<td>0.0455</td>
<td>0.0422</td>
<td>0.0401</td>
<td>0.0391</td>
<td>0.0392</td>
</tr>
<tr>
<td>DISK-PIE(GP)</td>
<td>$k = 10$</td>
<td>$k = 20$</td>
<td>$k = 30$</td>
<td>0.0332</td>
<td>0.0221</td>
</tr>
<tr>
<td></td>
<td>0.0332</td>
<td>0.0221</td>
<td>0.0232</td>
<td>0.0306</td>
<td>0.0427</td>
</tr>
<tr>
<td>DISK-PIE(MPP) $r = 200$</td>
<td>$k = 10$</td>
<td>$k = 20$</td>
<td>$k = 30$</td>
<td>0.0394</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>0.0394</td>
<td>0.0270</td>
<td>0.0278</td>
<td>0.0352</td>
<td>0.0473</td>
</tr>
<tr>
<td>DISK-PIE(MPP) $r = 400$</td>
<td>$k = 10$</td>
<td>$k = 20$</td>
<td>$k = 30$</td>
<td>0.0369</td>
<td>0.0243</td>
</tr>
<tr>
<td></td>
<td>0.0369</td>
<td>0.0243</td>
<td>0.0250</td>
<td>0.0325</td>
<td>0.0449</td>
</tr>
</tbody>
</table>
Background of GP
D&C Bayes Method Theory Experiments Future Work

True

DISK (GP, k = 10): 50% Quantile
DISK (GP, k = 20): 2.5% Quantile
DISK (GP, k = 20): 50% Quantile
DISK (GP, k = 20): 97.5% Quantile

DISK (MPP, r = 200, k = 20): 2.5% Quantile
DISK (MPP, r = 200, k = 20): 50% Quantile
DISK (MPP, r = 200, k = 20): 97.5% Quantile

DISK (MPP, r = 400, k = 20): 2.5% Quantile
DISK (MPP, r = 400, k = 20): 50% Quantile
DISK (MPP, r = 400, k = 20): 97.5% Quantile
Bias-Variance Tradeoff

When $n = 10^6$, full data GP, LatticeKrig, MPP, NNGP are not applicable. Only laGP is left to compare with DISK.

Prediction MSE averaged over $l = 2025$ locations

<table>
<thead>
<tr>
<th></th>
<th>DISK(MPP) $(r = 400, k = 500)$</th>
<th>DISK(MPP) $(r = 400, k = 500)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>laGP</td>
<td>0.0102</td>
<td>0.0061</td>
</tr>
<tr>
<td>DISK</td>
<td>0.0061</td>
<td>0.0052</td>
</tr>
</tbody>
</table>
## Simulation Results When $n = 10^4$

### Parametric Inference and Predictive Interval Coverage

<table>
<thead>
<tr>
<th>Method</th>
<th>$\beta$</th>
<th>$\tau^2$</th>
<th>$\sigma^2$</th>
<th>$\phi$</th>
<th>95% PI Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>1.00</td>
<td>-</td>
<td>0.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>laGP</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.94</td>
</tr>
<tr>
<td>LatticeKrig</td>
<td>-</td>
<td>0.01</td>
<td>-</td>
<td>0.08</td>
<td>0.95</td>
</tr>
<tr>
<td>GP</td>
<td>1.08</td>
<td>0.12</td>
<td>0.009</td>
<td>0.115</td>
<td>0.95</td>
</tr>
<tr>
<td>MPP ($r = 200$)</td>
<td>1.56</td>
<td>0.15</td>
<td>0.008</td>
<td>0.119</td>
<td>0.95</td>
</tr>
<tr>
<td>MPP ($r = 200$)</td>
<td>1.23</td>
<td>0.16</td>
<td>0.008</td>
<td>0.120</td>
<td>0.95</td>
</tr>
<tr>
<td>NNGP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NN= 5</td>
<td>0.36</td>
<td>0.29</td>
<td>0.009</td>
<td>0.123</td>
<td>0.94</td>
</tr>
<tr>
<td>NN= 15</td>
<td>0.31</td>
<td>0.17</td>
<td>0.009</td>
<td>0.113</td>
<td>0.95</td>
</tr>
<tr>
<td>NN= 25</td>
<td>0.30</td>
<td>0.16</td>
<td>0.009</td>
<td>0.112</td>
<td>0.95</td>
</tr>
<tr>
<td>DISK+PIE (GP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 20$</td>
<td>0.98</td>
<td>0.22</td>
<td>0.008</td>
<td>0.142</td>
<td>0.96</td>
</tr>
<tr>
<td>DISK+PIE (MPP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r = 200, k = 20$</td>
<td>0.98</td>
<td>0.22</td>
<td>0.008</td>
<td>0.140</td>
<td>0.97</td>
</tr>
<tr>
<td>$r = 400, k = 20$</td>
<td>0.98</td>
<td>0.22</td>
<td>0.008</td>
<td>0.140</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Simulation Results When $n = 10^6$

Parametric Inference and Predictive Interval Coverage

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>$\tau^2$</th>
<th>$\sigma^2$</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Truth</strong></td>
<td>1.00</td>
<td>-</td>
<td>0.01</td>
<td>-</td>
</tr>
<tr>
<td><strong>laGP</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>DISK+PIE (MPP)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r = 400, k = 500$</td>
<td>1.01</td>
<td>0.1609</td>
<td>0.00863</td>
<td>0.134907</td>
</tr>
<tr>
<td>$r = 600, k = 500$</td>
<td>1.01</td>
<td>0.1607</td>
<td>0.00863</td>
<td>0.135020</td>
</tr>
<tr>
<td><strong>MSPE 95% PI Coverage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>laGP</strong></td>
<td>0.0101</td>
<td>0.9450</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DISK+PIE (MPP)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r = 400, k = 500$</td>
<td>0.0099</td>
<td>0.9624</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r = 600, k = 500$</td>
<td>0.0099</td>
<td>0.9590</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $r$ is the number of knots. $k$ is the number of subsets.
- Full GP, LatticeKrig, MPP, and NNGP fail to run for $n = 10^6$. 
Simulation Model 2: Rough Surface

- \( \mathcal{D} = [0, 1] \times [0, 1] \).
- The true surface \( w_0(\cdot) \) is a rough sample path from the Matérn kernel with \( \nu = 1/2 \): \( C(s, s') = \tau^2 \exp(-\phi \|s - s'\|) \) for \( s, s' \in \mathcal{D} \). We fit the model using the same kernel.
- The response is \( y(s) = \beta_0 + w_0(s) + \epsilon \), \( \epsilon \sim N(0, \sigma^2) \).
- \( \beta_0 = 1 \), \( \sigma^2 = 0.1 \), \( \tau^2 = 1 \), \( \phi = 9 \).
- \( \sigma^2 \sim \text{InvGamma}(2, 0.1) \), \( \tau^2 \sim \text{InvGamma}(2, 2) \), \( \phi \sim \text{Uniform}(0.1, 30) \).
- Training sample size \( n = 10^4 \) respectively. The training locations are sampled uniformly over \( \mathcal{D} \).
- Testing sample size \( l = 2025 \), also sampled uniformly over \( \mathcal{D} \).
- The Bayes \( L_2 \) risks (and other criteria) are estimated by averaging over the \( l \) testing locations.
- Only compare DISK(MPP) \( k = 20 \) and laGP.
## Results for Simulation 2

### Parametric Inference and Predictive Interval Coverage ($k = 20$ for DISK)

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\sigma^2$</th>
<th>$\tau^2$</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Truth</strong></td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td><strong>laGP</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>DISK(MPP)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r = 200$</td>
<td>1.00</td>
<td>0.14</td>
<td>1.00</td>
<td>9.28</td>
</tr>
<tr>
<td></td>
<td>(0.96, 1.03)</td>
<td>(0.12, 0.17)</td>
<td>(0.92, 1.03)</td>
<td>(8.94, 9.68)</td>
</tr>
<tr>
<td>$r = 400$</td>
<td>1.00</td>
<td>0.14</td>
<td>1.00</td>
<td>9.32</td>
</tr>
<tr>
<td></td>
<td>(0.96, 1.03)</td>
<td>(0.11, 0.17)</td>
<td>(0.92, 1.03)</td>
<td>(8.96, 9.73)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MSPE</th>
<th>95% PI Coverage</th>
<th>95% PI Length</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>laGP</strong></td>
<td>0.50 (0.0120)</td>
<td>0.21 (0.0115)</td>
<td>0.38 (0.0003)</td>
</tr>
<tr>
<td><strong>DISK(MPP)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r = 200$</td>
<td>0.90 (0.0210)</td>
<td>0.96 (0.0027)</td>
<td>3.97 (0.0773)</td>
</tr>
<tr>
<td>$r = 400$</td>
<td>0.82 (0.0194)</td>
<td>0.96 (0.0035)</td>
<td>3.86 (0.0717)</td>
</tr>
</tbody>
</table>
Sea Surface Temperature Data

- SST in the west coast of mainland U.S.A., Canada, and Alaska between 40° – 65° north latitudes and 100° – 180° west longitudes.

- The dataset is obtained from NODC World Ocean Database. https://www.nodc.noaa.gov/OC5/WOD/pr_wod.html

- We choose 1,000,800 spatial observations in the selected domain. $10^6$ obs are used as training data, and 800 obs are used as testing data.

- Model: $y(s) = \beta_0 + \beta_1 \times \text{latitude} + w(s) + \epsilon(s), \epsilon(s) \sim N(0, \sigma^2)$.

- For DISK+CMC(MPP) and DISK+PIE(MPP), we show the results with $r = 400$ and $k = 300$. 
The PMSEs of CMC(MPP) and DISK(MPP) are almost the same, and both are comparable to laGP (a freq GP method). DISK(MPP) has wider predictive intervals.

Other Bayesian GP methods do not work, including MPP and NNGP.
### SST Data: Numerical Results

Parametric inference and prediction in SST data. DISK-CMC and DISK-PIE use MPP-based modeling with 400 knots on $k = 300$ subsets.

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\sigma^2$</th>
<th>$\tau^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>laGP</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DISK-CMC</td>
<td>31.78</td>
<td>-0.35</td>
<td>12.22</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td>(31.19, 32.37)</td>
<td>(-0.36, -0.34)</td>
<td>(11.78, 12.69)</td>
<td>(0.108, 0.112)</td>
</tr>
<tr>
<td>DISK-PIE</td>
<td>32.34</td>
<td>-0.32</td>
<td>11.83</td>
<td>0.184</td>
</tr>
<tr>
<td></td>
<td>(31.74, 32.95)</td>
<td>(-0.33, -0.31)</td>
<td>(11.23, 12.45)</td>
<td>(0.182, 0.185)</td>
</tr>
<tr>
<td>MSPE</td>
<td>0.25 (0.00)</td>
<td>0.95 (0.00)</td>
<td>1.35 (0.00)</td>
<td></td>
</tr>
<tr>
<td>95% PI Coverage</td>
<td>0.41 (0.00)</td>
<td>0.13 (0.00)</td>
<td>0.14 (0.00)</td>
<td></td>
</tr>
<tr>
<td>95% PI Length</td>
<td>0.41 (0.00)</td>
<td>0.95 (0.00)</td>
<td>2.67 (0.00)</td>
<td></td>
</tr>
</tbody>
</table>
Outline

Background of GP

D&C Bayes

Method

Theory

Experiments

Future Work
Future Work

- Theory for the finite-dimensional parameters ($\alpha, \sigma^2$), and the joint distribution of both parametric and nonparametric parts.
- Theory for the frequentist coverage of DISK posterior credible intervals.
- The choice of partition distribution; The influence from the number of knots in MPP.
- Extension to spatial-temporal processes.
Future Work

- Theory for the finite-dimensional parameters \((\alpha, \sigma^2)\), and the joint distribution of both parametric and nonparametric parts.

- Theory for the frequentist coverage of DISK posterior credible intervals.

- The choice of partition distribution; The influence from the number of knots in MPP.

- Extension to spatial-temporal processes.

- Manuscript at \texttt{arXiv:1712.09767}.

Thank you!

&

Questions?