Distributed Spatial Kriging for Large Spatial Dataset

Rajarshi Guhaniyogi, Ph.D

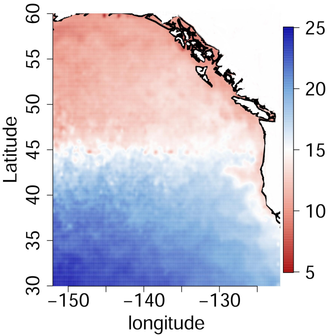
Department of Statistics & Applied Math, University of California Santa Cruz

Joint work with

Sanvesh Srivastava, Cheng Li, Terrance Savistsky

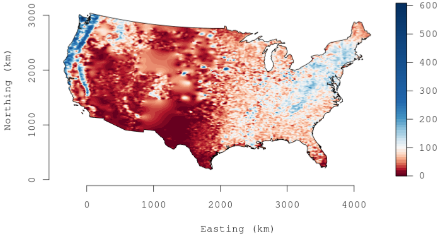
December 5, 2017

Example 1: Pacific Ocean Sea Surface Temperature Data



- Sea surface temperature is a major indicator of climate change.
- Data collected by voluntary observing ships, buoys, military and scientific cruises for last 30 years.
- The figure shows data observed in 1,008,371 locations in 2016 in the Eastern Pacific.

Example 2: Forest Biomass Data



- Prediction of forest biomass is important to understand current carbon stock and flux, bio-feedstock for emerging bio-economies, and impact of deforestation.
- Forest Inventory and Analysis (FIA) under USDA collects data on Biomass regurlarly.
- The figure shows data observed in 114,371 locations in 2012.

Spatial Regression Model: Geostatistical Model

Geostatistical Model

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\beta + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \;\; \epsilon(\boldsymbol{s}) \sim N(0, \tau^2)$$

- $w(\boldsymbol{s})$ is an unknown function that captures local level spatial variation of the response.
- Produce spatial map for $\{y(\boldsymbol{s}) : \boldsymbol{s} \in \mathscr{D}\}$ and $\{w(\boldsymbol{s}) : \boldsymbol{s} \in \mathscr{D}\}$ based on the observed data, i.e. provide $y(s_0)|y(s_1), ..., y(s_n)$ for any unobserved location $\boldsymbol{s}_0$.
- $\mathscr{D}$ is the spatial domain i.e. $\mathscr{D} \subset \mathscr{R}^2$.
- Potentially very rich to understand the spatial impact on the response.

Spatial Gaussian Process

- $\{w(\boldsymbol{s}) : \boldsymbol{s} \in \mathscr{D}\} \sim GP(0, C_{\boldsymbol{\theta}}(\cdot, \cdot))$ implies

$$\boldsymbol{w} = (w(\boldsymbol{s}_1), ..., w(\boldsymbol{s}_n))' \sim N(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\theta}})$$

  for any finite set of locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$.

- $\boldsymbol{C}_{\boldsymbol{\theta}} = (C_{\boldsymbol{\theta}}(\boldsymbol{s}_i, \boldsymbol{s}_j))$ is the $n \times n$ spatial covariance matrix.

- Stationary: $C_{\boldsymbol{\theta}}(\boldsymbol{s}_i, \boldsymbol{s}_j) = C_{\boldsymbol{\theta}}(\boldsymbol{s}_i - \boldsymbol{s}_j)$; Isotropic: $C_{\boldsymbol{\theta}}(\boldsymbol{s}_i, \boldsymbol{s}_j) = C_{\boldsymbol{\theta}}(||\boldsymbol{s}_i - \boldsymbol{s}_j||)$.

- Examples of spatial covariance function: exponential covariance function, $\boldsymbol{\theta} = \{\sigma^2, \phi\}$

$$\boxed{C_{\sigma^2,\phi}(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sigma^2 \exp(-\phi||\boldsymbol{s}_i - \boldsymbol{s}_j||)}.$$

Full Likelihood from Gaussian Process (GP) model

- $\boldsymbol{y} = y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n)$ are observed data and $\boldsymbol{x}(\boldsymbol{s}_1), ..., \boldsymbol{x}(\boldsymbol{s}_n)$ are the corresponding predictors.
- Let $\boldsymbol{X} = [\boldsymbol{x}(\boldsymbol{s}_1) : \cdots : \boldsymbol{x}(\boldsymbol{s}_n)]'$ be the predictor matrix.
- Model: $\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{C}_{\boldsymbol{\theta}} + \tau^2 \boldsymbol{I})$.
- Estimating parameters $\beta, \boldsymbol{\theta}$ from the likelihood

$$-\frac{1}{2}\log(det(\boldsymbol{C}_{\boldsymbol{\theta}} + \tau^2 \boldsymbol{I})) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{C}_{\boldsymbol{\theta}} + \tau^2 \boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)$$

- Bayesian Inference: Prior on $\{\beta, \boldsymbol{\theta}\}$

Challenges

- Store $\boldsymbol{C}_{\boldsymbol{\theta}} + \tau^2 \boldsymbol{I}$
- Compute $\texttt{Chol}(\boldsymbol{C}_{\boldsymbol{\theta}} + \tau^2 \boldsymbol{I}) = \boldsymbol{L}\boldsymbol{L}'$.

Literature on Spatial Big Data

- Low rank model (Wabha, 1990; Higdon, 2001; Kamman & Wand, 2003; Paciorek, 2007; Lemos and Sanso, 2006; Banerjee et al., 2008; Cressie & Johannesson, 2008; Finley et al., 2009; Gramacy and Lee, 2008; Guhaniyogi et al., 2011 & 2013; Sang et al. 2012; Katzfuss, 2016).
- Multiscale approaches (Nychka, 2002; Johannesson et al., 2007; Tzeng and Huang, 2015; Nychka et al., 2015; Katzfuss, 2016; Guhaniyogi & Sanso, 2017).
- Spectral approximations and composite likelihoods (Funetes, 2007; Eidvisk, 2016).
- Sparsity: (Solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$: (a) $\boldsymbol{A}$ sparse (b) $\boldsymbol{A}^{-1}$ sparse)
  (i) Covariance tapering (Kaufman et al., 2008; Shaby and Ruppert, 2012; Sang et al., 2012).
  (ii) `INLA` (Rue et al., 2009), `lagp` (Gramacy and Apley, 2015), nearest neighbor processes (Stein et al., 2004; Stroud et al., 2014; Datta et al., 2016).

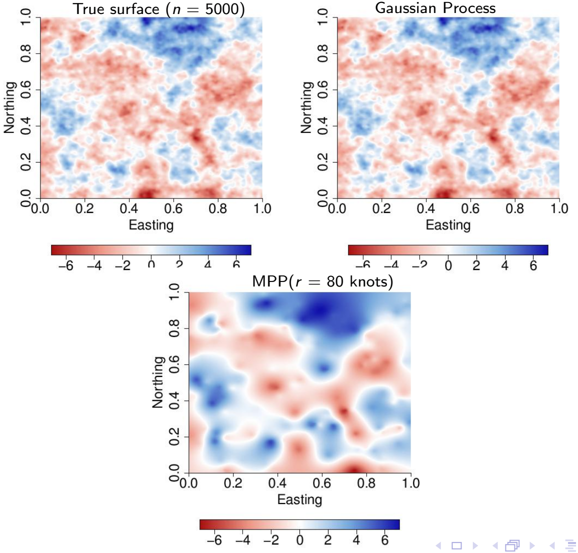
Low Rank Model

- Approximate $\boldsymbol{C}_{\boldsymbol{\theta}} \approx \boldsymbol{B}_{\boldsymbol{\theta}} \boldsymbol{C}_{\boldsymbol{\theta}}^{*-1} \boldsymbol{B}'_{\boldsymbol{\theta}} + \boldsymbol{D}_{\boldsymbol{\theta}}$
- $\boldsymbol{B}_{\boldsymbol{\theta}}$ is the $n \times r$ spatial basis matrix $r << n$.
- $\boldsymbol{C}_{\boldsymbol{\theta}}^{*}$ is an $r \times r$ spatial covariance matrix.
- $\boldsymbol{D}_{\boldsymbol{\theta}}$ is either sparse or diagonal.
- Different choices of basis functions leads to different low rank models.
- The computational complexity $O(r^3 + nr^2) \leq O(n^3)$.

Modified Predictive Process

- $\mathscr{S}^* = \{\boldsymbol{s}_1^*, ..., \boldsymbol{s}_r^*\}, \mathscr{S} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$.
- $\boldsymbol{B}_{\boldsymbol{\theta}} = \text{Cov}(\boldsymbol{w}(\mathscr{S}), \boldsymbol{w}(\mathscr{S}^*))$, $\boldsymbol{C}_{\boldsymbol{\theta}}^{*} = \text{Var}(\boldsymbol{w}(\mathscr{S}^*))$.
- $\boldsymbol{D}_{\boldsymbol{\theta}} = \text{diag}\{\boldsymbol{C}_{\boldsymbol{\theta}} - \boldsymbol{B}_{\boldsymbol{\theta}} \boldsymbol{C}_{\boldsymbol{\theta}}^{*-1} \boldsymbol{B}'_{\boldsymbol{\theta}}\}$

Modified Predictive Process (MPP): Computation Cost vis a vis Accuracy

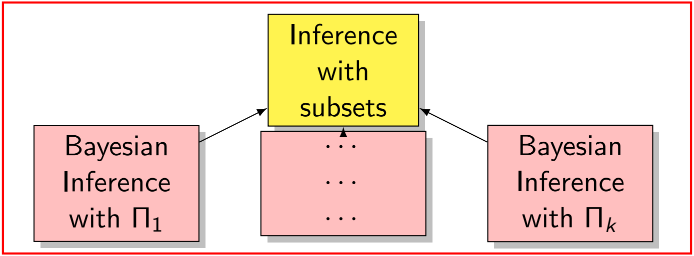


Requirement from Next Generation Spatial Models

- Scalability
- Avoid storage of entire data
- Divide and conquer
- Map Reduce (HADOOP)
- Theoretical support

Posterior on Data Subsets (Subset Posteriors)

- Split the data $\mathscr{S} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}, \mathscr{Y} = \{y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n)\}, \mathscr{X} = \{\boldsymbol{x}(\boldsymbol{s}_1), ..., \boldsymbol{x}(\boldsymbol{s}_n)\}$ into k non-overlapping and exhaustive subsets $\mathscr{S}_j, \mathscr{Y}_j, \mathscr{X}_j$, $j = 1, .., k$.
- Each subset has $m = n/k$ data points drawn randomly from the entire domain.



"Weak learner" of full posterior

$$\Pi_j(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathscr{Y}_j) \propto [p(\mathscr{Y}_j | \mathscr{X}_j, \mathscr{S}_j, \boldsymbol{\beta}, \boldsymbol{\theta})]^k p(\boldsymbol{\beta}, \boldsymbol{\theta})$$

- $p(\mathscr{Y}_j | \mathscr{X}_j, \mathscr{S}_j, \boldsymbol{\beta}, \boldsymbol{\theta})$ is the likelihood of the model under consideration.
- $p(\boldsymbol{\beta}, \boldsymbol{\theta})$ is the prior distribution.
- Π_j's are referred to as the subset posteriors.
- These are stochastic approximations to the full posterior $\Pi(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathscr{Y})$.

How to combine Π_j's optimally?

Combine Subset Posteriors: DISK Pseudo Posterior

- Compute Wasserstein mean $\bar{\Pi}$ of $\Pi_1, ..., \Pi_k$.
- Wasserstein mean $\bar{\Pi}$ is calculated using the simple algorithm. Denote $\boldsymbol{\Omega} = \{\boldsymbol{\beta}, \boldsymbol{\theta}\}$.
- `for` j in 1:k
  (1) Draw s MCMC samples $\boldsymbol{\Omega}_{j1}, ..., \boldsymbol{\Omega}_{js}$ of $\boldsymbol{\Omega}$ from Π_j.
  (2) Calculate α-th `quantile` from the MCMC sample in j-th subset. Denote it by $\boldsymbol{\Omega}_j^{(\alpha)}$.
  (3) $\boldsymbol{\Omega}^{(\alpha)} = \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{\Omega}_j^{(\alpha)}$ is the α-th quantile of $\bar{\Pi}$.
- One $\boldsymbol{\Omega}^{(\alpha)}$ are estimated for a range of α, samples are drawn from $\bar{\Pi}$ using the inverse CDF method.
- $\bar{\Pi}$ is called DISK pseudo posterior and it is used as a substitute to the full posterior distribution.

Surface Interpolation and Prediction at Unobserved Locations

- Let $\boldsymbol{s}_0$ be a location where response has not been observed.
- `for` j in 1:k
  (1) Draw s MCMC samples $y_{j1}(\boldsymbol{s}_0), ..., y_{js}(\boldsymbol{s}_0)$ from $y(\boldsymbol{s}_0)|\mathscr{Y}_j$.
  (2) Calculate α-th `quantile` from the MCMC sample in j-th subset. Denote it by $y_j^{(\alpha)}(\boldsymbol{s}_0)$.
  (3) $y^{(\alpha)}(\boldsymbol{s}_0) = \frac{1}{k}\sum_{j=1}^{k} y_j^{(\alpha)}(\boldsymbol{s}_0)$ is the α-th quantile of DISK pseudo posterior for prediction.
- Surface interpolation is carried out similarly by calculating DISK pseudo posterior for $w(\boldsymbol{s}_0)|\mathscr{Y}$.

DISK: Novelty Over Other Divide & Conquer Techniques

- Aggregation of point estimates through median (Wang and Dunson, 2013; Wang et al., 2015; Minsker, 2014)
- Aggregation of subset posteriors through median (Minsker et al., 2017; Guhaniyogi and Banerjee, 2017)
- Consensus Monte Carlo (CMC) (Scott et al., 2016), Semiparametric Density Product (SDP) (Neiswenger et al., 2014).
- Wasserstein Mean posterior (Srivastava et al., 2015; Li et al., 2017; Savitsky and Srivastava, 2017).
- Both theory and practice are only applicable for i.i.d data.

Simulation Study

- $\boldsymbol{s} = (s_1, s_2)$ are drawn randomly on $[-2, 2]^2$.

Simulation Study

- $\boldsymbol{s} = (s_1, s_2)$ are drawn randomly on $[-2, 2]^2$.
- predictor $x(\boldsymbol{s})$ are sampled iid from N(0,1).

Simulation Study

- $\boldsymbol{s} = (s_1, s_2)$ are drawn randomly on $[-2, 2]^2$.
- predictor $x(\boldsymbol{s})$ are sampled iid from N(0,1).
- Response $y(\boldsymbol{s})$ is simulated from

$$f_0(s) = e^{-(s-1)^2} + e^{-0.8(s+1)^2} - 0.05\sin\{8(s+0.1)\},$$
$$y(s_1, s_2) = \beta_0 + x(s_1, s_2)\beta_1 - f_0(s_1)f_0(s_2) + \epsilon,\ \epsilon \sim N(0, 0.01),$$

- Yields highly nonstationary spatial surface that is difficult to estimate (Gramacy & Apley, 2015).

Simulations

- Simulation 1: $n = 10^4$ locations for model fitting, $l = 2025$ for prediction.
- Simulation 2: $n = 10^6$ locations for model fitting, $l = 2025$ for prediction.

Competitors

Competitors

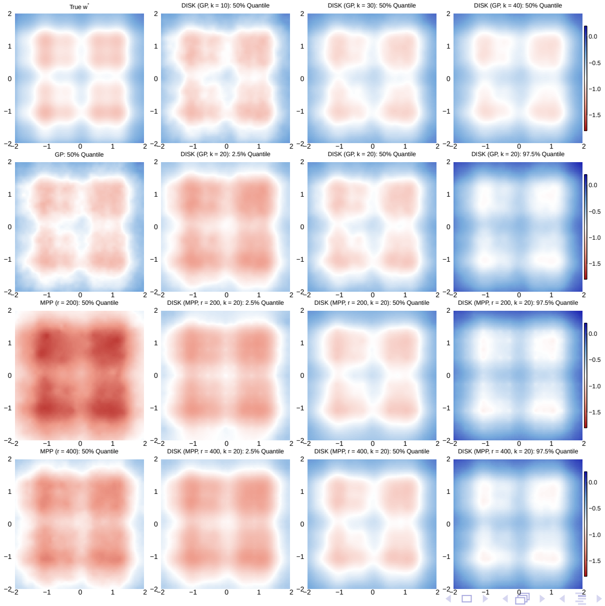
- full GP (GP on the full dataset)
- full MPP (MPP on the full dataset)
- DISK-GP (with different choices of k)
- DISK-MPP (with different choices of k and r)
- SDP (Neiswenger et al., 2014), CMC (Scott et al., 2016)
- laGP (Gramacy & Apley, 2015), NNGP (Datta et al., 2016)
- LatticeKrig (Nychka et al., 2015)

Predicted locations: $\mathscr{S}^0 = \{\boldsymbol{s}_1^0, \dots, \boldsymbol{s}_l^0\}$.

$$\widehat{\text{bias}}^2 = \frac{1}{l}\sum_{i'=1}^{l}\{\hat{w}(\boldsymbol{s}_{i'}^0) - w_0(\boldsymbol{s}_{i'}^0)\}^2, \ \widehat{\text{var}} = \frac{1}{l}\sum_{i'=1}^{l}\widehat{\text{var}}\{w(\boldsymbol{s}_{i'}^0)\},$$

$$L_2\text{-risk} = \widehat{\text{bias}}^2 + \widehat{\text{var}},$$

Simulation 1: Surface Interpolation



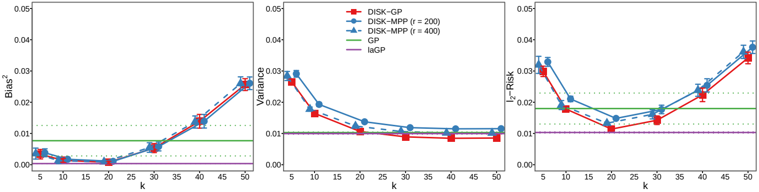
Simulation 1: Inference on Surface

	$Bias^2$	Variance	L_2-risk	95% CI coverage	95% CI Length
laGP	0.0004 (0.0000)	0.1000 (0.0002)	0.0103 (0.0002)	1.00 (0.00)	0.3890 (0.0036)
LatticeKrig	0.0002 (0.0000)	0.0003 (0.0000)	0.0005 (0.0000)	0.98 (0.00)	0.0703 (0.0006)
GP	0.0077 (0.0049)	0.0103 (0.0002)	0.0180 (0.0049)	1.00 (0.00)	0.3943 (0.0006)
MPP (r=400)	0.0623 (0.0369)	0.0105 (0.0002)	0.0727 (0.0370)	0.29 (0.04)	0.3976 (0.0037)
NNGP, NN=25	0.4887 (0.0668)	0.0013 (0.0001)	0.4900 (0.0668)	0.00 (0.00)	0.1398 (0.0032)
CMC-MPP (r=200, k=20)	0.0090 (0.0029)	0.0402 (0.0006)	0.0493 (0.0031)	0.12 (0.10)	0.1026 (0.0009)
CMC-MPP (r=400, k=20)	0.0013 (0.0005)	0.0409 (0.0006)	0.0422 (0.0009)	0.82 (0.09)	0.1005 (0.0009)
DISK-GP (k=20)	0.0008 (0.0005)	0.0106 (0.0004)	0.0221 (0.0005)	1.00 (0.00)	0.4041 (0.0070)
DISK-MPP (r=200, k=20)	0.0009 (0.0004)	0.0131 (0.0002)	0.0270 (0.0004)	1.00 (0.00)	0.4477 (0.0039)
DISK-MPP (r=400, k=20)	0.0007 (0.0004)	0.1180 (0.0002)	0.0243 (0.0005)	1.00 (0.00)	0.4253 (0.0031)

Simulation 1: Parameter Estimation and Predictive Inference

	β	τ^2	MSPE	95% PI Cov.	95% PI Length
laGP	--	--	0.010 (0.000)	0.94 (0.01)	0.39 (0.00)
LatticeKrig	--	--	0.010 (0.000)	0.95 (0.01)	0.39 (0.00)
GP	1.08 (0.50, 1.65)	0.009 (0.009, 0.010)	0.010 (0.000)	0.95 (0.01)	0.39 (0.00)
MPP (r=400)	1.23 (0.61, 1.84)	0.008 (0.008, 0.008)	0.010 (0.000)	0.95 (0.01)	0.40 (0.00)
NNGP, NN=25	0.30 (0.30, 0.30)	0.009 (0.009, 0.009)	0.010 (0.000)	0.95 (0.01)	0.40 (0.01)
CMC-MPP (r=200, k=20)	1.09 (0.89, 1.28)	0.007 (0.006, 0.007)	0.010 (0.000)	0.38 (0.00)	0.10 (0.01)
CMC-MPP (r=400, k=20)	1.02 (0.82, 1.22)	0.007 (0.007, 0.007)	0.010 (0.000)	0.38 (0.00)	0.10 (0.00)
SDP-MPP (r=200, k=20)	1.08 (0.89, 1.27)	0.007 (0.006, 0.007)	--	--	--
SDP-MPP (r=400, k=20)	1.02 (0.83, 1.21)	0.007 (0.007, 0.007)	--	--	--
DISK-GP (k=20)	0.98 (0.82, 1.15)	0.009 (0.008, 0.009)	0.010 (0.000)	0.96 (0.01)	0.42 (0.01)
DISK-MPP (r=200, k=20)	0.98 (0.82, 1.16)	0.008 (0.008, 0.009)	0.010 (0.000)	0.97 (0.01)	0.46 (0.00)
DISK-MPP (r=400, k=20)	0.98 (0.82, 1.16)	0.008 (0.008, 0.009)	0.010 (0.000)	0.97 (0.01)	0.44 (0.00)

Choice of the number of subsets (k) for fixed n



- variance term decreases as a function of k.
- $Bias^2$ term varies within a small window for upto a cetain k, then it keeps on increasing.
- As a result, L_2-risk also increases from that inflexion point.

Assumptions

1. w_0 belongs to the RKHS of $\text{GP}(0, C_\alpha(\cdot,\cdot))$ and $C_\alpha(\mathbf{s},\mathbf{s}') = \sum_{i=1}^{\infty} \mu_i \phi_i(\mathbf{s}) \phi_i(\mathbf{s}')$.
2. $\exists\, \rho > 0$ and $r \geq 2$ s.t. $E\{\phi_i^{2r}(\mathbf{s})\} \leq \rho^{2r}\ \forall\, i$.

Theoretical Result on the $L_2 - risk$ under Assumptions 1-2

1. If C_α is a finite-rank kernel with $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_{d^*} > 0$, $\mu_{d^*+1} = \mu_{d^*+2} = \ldots = 0$ for some constant integer d^*, and for some constant $c > 0$, $k \leq cn^{\frac{r-4}{r-2}} / (\log n)^{\frac{2r}{r-2}}$, then $E\|\overline{w} - w_0\|_2^2 = O\left(n^{-1}\right)$ as $n \to \infty$.
2. If $\mu_i \leq c_{1\mu} \exp\left(-c_{2\mu} i^2\right)$ for some constants $c_{1\mu} > 0, c_{2\mu} > 0$ and all i, and for some constant $c > 0$, $k \leq cn^{\frac{r-4}{r-2}} / (\log n)^{\frac{3r-1}{r-2}}$, then $E\|\overline{w} - w_0\|_2^2 = O\left(\sqrt{\log n}/n\right)$ as $n \to \infty$.
3. If $\mu_i \leq c_\mu i^{-2\nu}$ for some constants $c_\mu > 0, \nu > \frac{r-1}{r-4}$ and all i, and for some constant $c > 0$, $k \leq cn^{\frac{(r-4)\nu-(r-1)}{(r-2)\nu}} / (\log n)^{\frac{2r}{r-2}}$, $E\|\overline{w} - w_0\|_2^2 = O\left(n^{-\frac{2\nu-1}{2\nu}}\right)$ as $n \to \infty$.

Results from Simulation 2

	laGP	DISK-MPP(r=400,k=500)	DISK-MPP(r=600,k=500)
$Bias^2$	0.00021 (0.00001)	0.00016 (0.00003)	0.00012 (0.00003)
Variance	0.01003 (0.00003)	0.00297 (0.00001)	0.00256 (0.00000)
L_2-risk	0.01024 (0.00003)	0.00314 (0.00003)	0.00268 (0.00003)
95% CI Coverage	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
95% CI Length	0.39047 (0.00061)	0.21316 (0.00024)	0.19774 (0.00019)
Log_{10} Time	0.93 (0.24)	4.98 (0.07)	5.14 (0.11)

	truth	laGP	DISK-MPP (r=400,k=500)	DISK-MPP (r=600,k=500)
β	1	--	1.01 (0.98, 1.04)	1.01 (0.98, 1.04)
τ^2	0.1	--	0.008 (0.008, 0.009)	0.008 (0.008, 0.009)

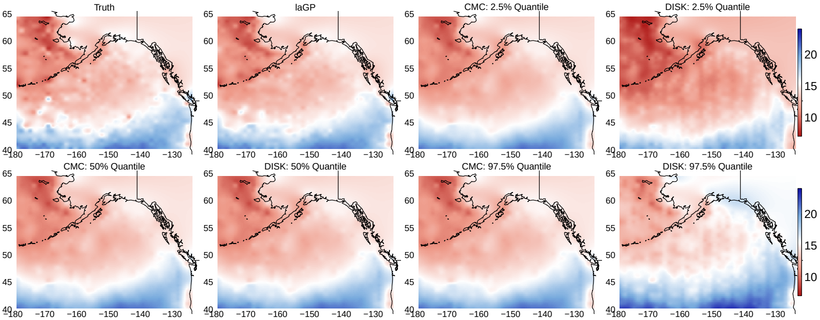
- Notable advantage in terms of L_2 risk compared to laGP.
- Coverage is same with narrower length of the 95% credible interval.
- Provides precise parameter estimates with 95% CIs.
- Enables fitting MPP for 1 million observations in manageable time.

Sea Surface Temperature Data

- Recall the data on SST between $40^0 - 65^0$ N latitude and $120^0 - 180^0$ W. longitude.
- Fit an ordinary linear regression with latitude and longitude as predictors.
- Residual surface showing lots of spatial variability left untapped.
- Use $1,000,000$ for model fitting, rest for prediction.
- Fit: $y(s_1, s_2) = \beta_0 + s_2\beta_1 + w(s_1, s_2) + \epsilon(s_1, s_2)$

	β_0	β_1	τ^2	σ^2	ϕ
CMC-MPP	(31.19, 32.37)	(-0.36, -0.34)	(0.108, 0.112)	(11.78, 12.69)	(0.020, 0.022)
SDP-MPP	(31.19, 32.37)	(-0.36, -0.34)	(0.108, 0.112)	(11.78, 12.69)	(0.020, 0.022)
DISK-MPP	(31.74, 32.95)	(-0.33, -0.31)	(0.182, 0.185)	(11.23, 12.44)	(0.037, 0.041)

Sea Surface Temperature Data



	95% PI Coverage	95% PI Length	MSPE	Log_{10} Time
laGP	0.95 (0.00)	1.35 (0.00)	0.25 (0.00)	0.93 (0.28)
CMC-MPP	0.13 (0.00)	0.14 (0.00)	0.41 (0.00)	4.90 (0.03)
DISK-MPP	0.95 (0.00)	2.67 (0.00)	0.41 (0.00)	5.16 (0.08)

Conclusion

- *Parallelizable* framework for analyzing large spatial data with complex nonstationarity.
- Subset inference can be carried out with *any* spatial model conceptually.
- Enables us to employ powerful spatial models for big data.
- Provides model based estimation, prediction and spatial surface recovery.
- The framework can potentially *scale* any stochastic process model based model.