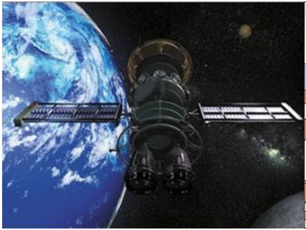
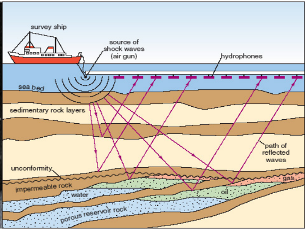
Estimation and prediction in spatial models with block composite likelihoods

Jo Eidsvik,
Department of Mathematical Sciences, NTNU, Norway

With Ben Shaby (Colorado State Univ), Brian Reich (North Carolina State Univ), Matt Wheeler and Jarad Niemi (Iowa State Univ)

Large spatial (spatio-temporal) datasets





Spatial Gaussian model

Model: $Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta} + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$.

- $\boldsymbol{s}$ is a spatial location on a domain of interest.
- $Y(\boldsymbol{s})$ is the response variable.
- $\boldsymbol{X}(s)$ are covariates.
- $\boldsymbol{\beta}$ are regression parameters.
- $w(\boldsymbol{s})$ is a spatially smooth process, covariance parameters; ϕ, σ^2.
- $\epsilon(\boldsymbol{s})$ is independent noise $\epsilon(\boldsymbol{s}) \sim N(0, \tau^2)$
- $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2)$.

Gaussian likelihood

Model: $Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\beta + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$.

Sampling locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$.
Data: $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))'$, $\boldsymbol{X} = (\boldsymbol{X}(\boldsymbol{s}_1), \ldots, \boldsymbol{X}(\boldsymbol{s}_n))'$.

Log likelihood:

$$l(\boldsymbol{Y}; \beta, \boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta)$$

Challenges for **parameter estimation** and **prediction**:

1. Build and store $n \times n$ matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma} = \boldsymbol{C} + \tau^2 \boldsymbol{I}_n$
2. Compute $\log|\boldsymbol{\Sigma}|$ and $\boldsymbol{\Sigma}^{-1}$ or $(\boldsymbol{Y} - \boldsymbol{X}\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta)$

Some approaches: tapering, fixed rank Kriging, Markov random fields, predictive process, approximate likelihood methods.

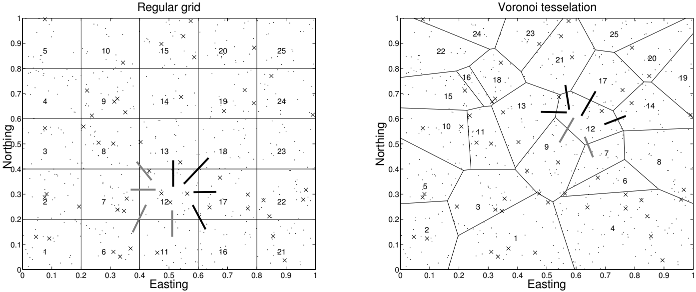
Use model for composites

- Lindsay (1988), Curriere and Lele (1999), Varin (2008).
- Scales with dimension. Useful for parallel computing.
- Use joints for subsets of data, not full joint. M blocks. $n = \sum_{k=1}^{M} n_k$.
- Block-pair variables: $\boldsymbol{Y}_{kl} = (\boldsymbol{Y}'_k, \boldsymbol{Y}'_l)'$, $\boldsymbol{X}_{kl} = (\boldsymbol{X}'_k, \boldsymbol{X}'_l)'$, size $(n_k + n_l) \times (n_k + n_l)$ covariance $\boldsymbol{\Sigma}_{kl} = \boldsymbol{\Sigma}_{kl}(\boldsymbol{\theta})$.
- Log composite likelihood:

$$l_{CL}(\boldsymbol{Y}; \beta, \boldsymbol{\theta}) = \sum_{k=1}^{M-1} \sum_{l>k} \left\{ -\frac{1}{2} \log |\boldsymbol{\Sigma}_{kl}| - \frac{1}{2} (\boldsymbol{Y}_{kl} - \boldsymbol{X}_{kl}\beta)' \boldsymbol{\Sigma}_{kl}^{-1} (\boldsymbol{Y}_{kl} - \boldsymbol{X}_{kl}\beta) \right\}$$

Split and conquer - examples of blocking

$$l_{CL}(\boldsymbol{Y};\beta,\boldsymbol{\theta}) = \sum_{k=1}^{M-1} \sum_{l \in N_k^{\rightarrow}} \left\{ -\frac{1}{2}\log|\boldsymbol{\Sigma}_{kl}| - \frac{1}{2}(\boldsymbol{Y}_{kl} - \boldsymbol{X}_{kl}\beta)'\boldsymbol{\Sigma}_{kl}^{-1}(\boldsymbol{Y}_{kl} - \boldsymbol{X}_{kl}\beta) \right\}$$



Estimation : Maximum composite likelihood

$$(\hat{\theta}, \hat{\boldsymbol{\beta}}) = \text{argmax}_{\boldsymbol{\theta}, \boldsymbol{\beta}} \{ l_{CL}(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{\theta}) \}.$$

- Iterative algorithm $p = 0, 1, \ldots$.
- Estimating equations: $\frac{dl_{CL}(\boldsymbol{Y}; \hat{\boldsymbol{\beta}}_p, \hat{\theta}_p)}{d\boldsymbol{\beta}} = 0$, $\frac{dl_{CL}(\boldsymbol{Y}; \hat{\boldsymbol{\beta}}_p, \hat{\theta}_p)}{d\boldsymbol{\theta}} = 0$.
- $\hat{\boldsymbol{\beta}}_p = \boldsymbol{A}^{-1}\boldsymbol{b}$,
  $\boldsymbol{A} = \boldsymbol{A}(\hat{\theta}_p; X_{kl}\text{pairs})$, $\boldsymbol{b} = \boldsymbol{b}(\hat{\theta}_p; X_{kl}\text{pairs}, Y_{kl}\text{pairs})$.
- Fisher-scoring: $\hat{\theta}_{p+1} = \hat{\theta}_p - E\left(\frac{d^2 l_{CL}(\boldsymbol{Y}; \hat{\boldsymbol{\beta}}_p, \hat{\theta}_p)}{d\boldsymbol{\theta}^2}\right)^{-1} \frac{dl_{CL}(\boldsymbol{Y}; \hat{\boldsymbol{\beta}}_p, \hat{\theta}_p)}{d\boldsymbol{\theta}}$

Analytical derivatives

Computed for each block-pair.

$$
\begin{aligned}
\frac{d \log |\boldsymbol{\Sigma}_{kl}|}{d\theta_r} &= \operatorname{trace}(\boldsymbol{\Sigma}_{kl}^{-1} \frac{d\boldsymbol{\Sigma}_{kl}}{d\theta_r}) \\
\frac{d \boldsymbol{Z}'_{kl} \boldsymbol{\Sigma}_{kl}^{-1} \boldsymbol{Z}_{kl}}{d\theta_r} &= -\boldsymbol{Z}'_{kl} \boldsymbol{\Sigma}_{kl}^{-1} \frac{d\boldsymbol{\Sigma}_{kl}}{d\theta_r} \boldsymbol{\Sigma}_{kl}^{-1} \boldsymbol{Z}_{kl} \\
\boldsymbol{Z}_{kl} &= (\boldsymbol{Y}_{kl} - \boldsymbol{X}_{kl} \boldsymbol{\beta})
\end{aligned}
$$

Asymptotic properties: Godambe sandwich

$\hat{\theta} \to N(\theta, G^{-1})$. (Varin, 2008).

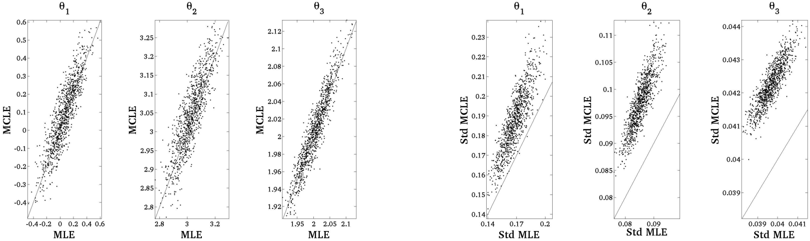
$$
\begin{aligned}
G = G(\hat{\theta}) &= H(\hat{\theta}) J^{-1}(\hat{\theta}) H(\hat{\theta}), \\
H(\hat{\theta}) &= -E\left(\frac{d^2 l_{CL}}{d\theta^2}(\hat{\theta})\right), \quad J(\hat{\theta}) = Var\left(\frac{d l_{CL}}{d\theta}(\hat{\theta})\right)
\end{aligned}
$$

Synthetic data

Unit square domain. $n = 2000$. 2000 replicates of data.
Covariance $\Sigma(h) = \tau^2 I(h = 0) + \sigma^2(1 + \phi h)\exp(-\phi h),\ h = |\boldsymbol{s}_i - \boldsymbol{s}_j|$.
Effective range 1/4th of domain.

Likelihood vs composite with $M = 100$ blocks.

Parameter estimates: log precision, log range, log nugget precision.



Prediction from composite likelihood

$Y_k^a = (\boldsymbol{Y}'_{k0}, \boldsymbol{Y}'_k)'$. $\boldsymbol{Y}_{k0}$ unobserved variables of size n_{k0} in block k.
$X_k^a = (\boldsymbol{X}'_{k0}, \boldsymbol{X}'_k)'$ associated covariates.

$$l_{CL}^k(\boldsymbol{Y}_{k0}) = -\sum_{l \in N_k} \frac{1}{2}[(\boldsymbol{Y}_k^{a'}, \boldsymbol{Y}'_l)' - (\boldsymbol{X}_k^{a'}, \boldsymbol{X}'_l)'\beta]'\boldsymbol{\Sigma}_{0kl}^{-1}[(\boldsymbol{Y}_k^{a'}, \boldsymbol{Y}'_l)' - (\boldsymbol{X}_k^{a'}, \boldsymbol{X}'_l)'\beta]$$

$\boldsymbol{\Sigma}_{0kl}$ is $(n_{k0} + n_k + n_l) \times (n_{k0} + n_k + n_l)$ covariance matrix.

$$\hat{Y}_{k0} = \boldsymbol{X}_{k0}\hat{\beta} + \boldsymbol{A}_0^{-1}\boldsymbol{b}_0$$

$\boldsymbol{A}_0 = \boldsymbol{A}_0(\hat{\theta})$, $\boldsymbol{b}_0 = \boldsymbol{b}_0(\hat{\theta}; X_{kl}\text{pairs}, Y_{kl}\text{pairs})$

Prediction properties

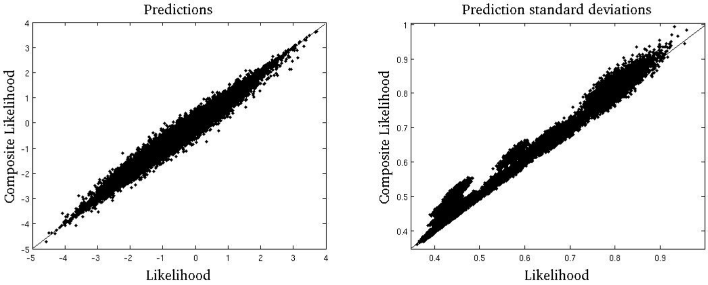
$$\hat{Y}_{k0} - Y_{k0} \sim N(0, G_0^{-1})$$

$$\begin{aligned} G_0 &= H_0 J_0^{-1} H_0, \\ H_0 &= -E\left(\frac{d^2 l_{CL}^k}{dY_{k0}^2}\right), \quad J_0 = \mathrm{Var}\left(\frac{dl_{CL}^k}{dY_{k0}}\right) \end{aligned}$$

Synthetic data

Unit square domain. $n = 2000$. 2000 replicates of data. 200 prediction sites.
Covariance $\Sigma(h) = \tau^2 I(h = 0) + \sigma^2(1 + \phi h)\exp(-\phi h),\ h = |\boldsymbol{s}_i - \boldsymbol{s}_j|$.
Effective range 1/4th of domain.

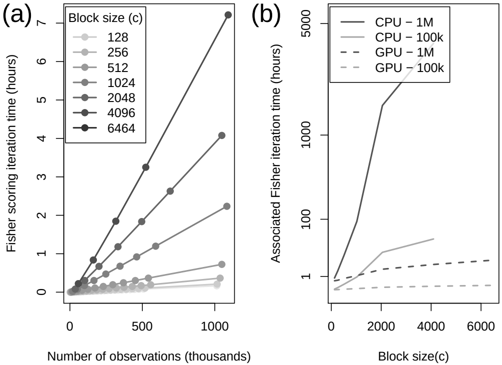
Likelihood vs composite with $M = 100$ blocks.



Results for block size 3^2 and 10^2

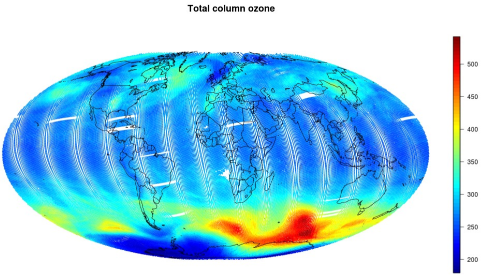
	L	CL, 3×3	CL, 10×10
MSE $\hat{\beta}_1$	0.05	0.05	0.06
MSE $\hat{\beta}_2$	0.17	0.21	0.23
MSE $\hat{\theta}_1$	0.014	0.018	0.018
MSE $\hat{\theta}_2$	0.0036	0.0043	0.0054
MSE $\hat{\theta}_3$	0.0007	0.0008	0.0008
Coverage (0.95) $\hat{\beta}_1$	0.96	0.95	0.96
Coverage (0.95) $\hat{\beta}_2$	0.93	0.92	0.92
Coverage (0.95) $\hat{\theta}_1$	0.93	0.92	0.91
Coverage (0.95) $\hat{\theta}_2$	0.94	0.94	0.91
Coverage (0.95) $\hat{\theta}_3$	0.95	0.95	0.94
MSPE	193	195	204
Mean pred. coverage (0.95)	0.95	0.95	0.95
Computing time (sec)	76	39	12

Parallelization



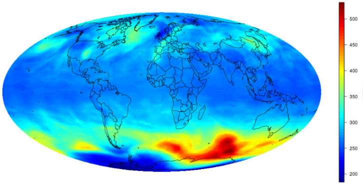
Matrix factorization on GPU. For-loop harder - require splitting the data on resource units.

Satellite dataset: $n \sim 200.000$

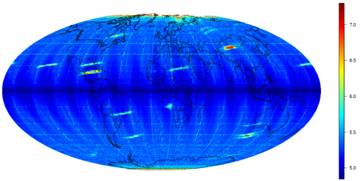


$Y(\boldsymbol{s}) = \beta_0 + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$.
Estimate parameters. Use plug in for prediction.

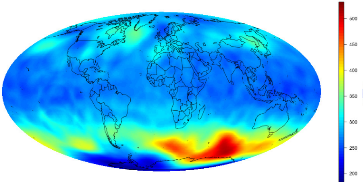
Satellite dataset: Fixed rank kriging and CL



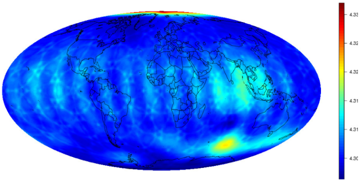
(a) CL predictions



(b) CL standard errors



(c) FRK predictions



(d) FRK standard errors

Satellite dataset: Prediction results for a hold-out set

	FRK, 4	FRK, 3	CL, reg 15	CL, fast 30
MSPE	44.3	88.1	25.7	26.0
Pred cov (0.95)	0.81	0.71	0.96	0.96
Timing (min)	6	2	40	4

Local composite spatial modeling

- Split parameter estimation and spatial prediction
- Just predicting in (pairs of) blocks... overlapping blocks...

Unifying these elements instead:

- Form a *process* on a local graph.
- Ease hierarchical Bayesian modeling.

Predictions on local domains, building a process

$$
\begin{aligned}
p(x_1, x_2, \ldots, x_n) &= p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)\ldots p(x_n|x_{n-1},\ldots,x_1) \\
&= p(x_1)\prod_{i=2}^{n} p(x_i|x_1,\ldots,x_{i-1})
\end{aligned}
$$

Assuming a neighborhood. Some of the conditioning can be ignored.

$$p(x_1, x_2, \ldots, x_n) = p(x_1)\prod_{i=2}^{n} p(x_i|x_j, j \in N_i^{\rightarrow})$$

Where $N_i^{\rightarrow}$ represents the neighborhood of i for the indexes below i. (Similar to a Cholesky factorization.)

Datta et al. (2016) formed a neighborhood graph

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

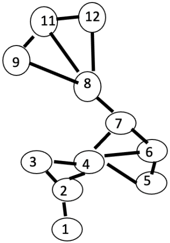
$$\tilde{p}(\mathbf{w}_\mathcal{S}) = \prod_{i=1}^{k} p(\mathbf{w}(\mathbf{s}_i) \mid \mathbf{w}_{N(\mathbf{s}_i)})$$

$$\tilde{p}(\mathbf{w}_\mathcal{S}) = \prod_{i=1}^{k} N(\mathbf{w}(\mathbf{s}_i) \mid \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)}, \mathbf{F}_{\mathbf{s}_i})$$

$$\mathbf{B}_{\mathbf{s}_i} = \mathbf{C}_{\mathbf{s}_i, N(\mathbf{s}_i)}\mathbf{C}^{-1}_{N(\mathbf{s}_i)}$$

$$\mathbf{F}_{\mathbf{s}_i} = \mathbf{C}(\mathbf{s}_i, \mathbf{s}_i) - \mathbf{C}_{\mathbf{s}_i, N(\mathbf{s}_i)}\mathbf{C}^{-1}_{N(\mathbf{s}_i)}\mathbf{C}_{N(\mathbf{s}_i), \mathbf{s}_i}$$

Neighborhood structure is defined by edges between nodes.
And by order of variables.



Nearest neighbor Gaussian process

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

$$\tilde{p}(\mathbf{w}_{\mathcal{U}} \mid \mathbf{w}_{\mathcal{S}}) = \prod_{i=1}^{r} N\left(\mathbf{w}(\mathbf{u}_i) \mid \mathbf{B}_{\mathbf{u}_i}\mathbf{w}_{N(\mathbf{u}_i)}, \mathbf{F}_{\mathbf{u}_i}\right) = N(\mathbf{B}_{\mathcal{U}}\mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}})$$

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

where $\mathcal{U} = \mathcal{V} \setminus \mathcal{S}$.

Node set:

$$\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$$

Outside node set:

$$\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$$

$$\tilde{\mathbf{C}}(\mathbf{v}_1, \mathbf{v}_2; \boldsymbol{\theta}) = \begin{cases} \tilde{\mathbf{C}}_{\mathbf{s}_i,\mathbf{s}_j}, & \text{if } \mathbf{v}_1 = \mathbf{s}_i \text{ and } \mathbf{v}_2 = \mathbf{s}_j \text{ are both in } \mathcal{S}, \\ \mathbf{B}_{\mathbf{v}_1}\tilde{\mathbf{C}}_{N(\mathbf{v}_1),\mathbf{s}_j} & \text{if } \mathbf{v}_1 \notin \mathcal{S} \\ \mathbf{B}_{\mathbf{v}_1}\tilde{\mathbf{C}}_{N(\mathbf{v}_1),N(\mathbf{v}_2)}\mathbf{B}'_{\mathbf{v}_2} + \delta_{(\mathbf{v}_1 =} & \ldots \text{ are not in } \mathcal{S} \end{cases}$$

$$\frac{1}{\prod_{i=1}^{k}\sqrt{\det(\mathbf{F}_{\mathbf{s}_i})}} \exp\left(-\frac{1}{2}\sum_{i=1}^{k}(\mathbf{w}(\mathbf{s}_i) - \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})'\mathbf{F}_{\mathbf{s}_i}^{-1}(\mathbf{w}(\mathbf{s}_i) - \mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})\right)$$

Sparse:

$$\tilde{\mathbf{C}}_{\mathcal{S}}^{-1}$$

Nearest neighbor Gaussian process

$$\frac{1}{\prod_{i=1}^{k}\sqrt{\det(\mathbf{F}_{\mathbf{s}_i})}}\exp\left(-\frac{1}{2}\sum_{i=1}^{k}(\mathbf{w}(\mathbf{s}_i)-\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})'\mathbf{F}_{\mathbf{s}_i}^{-1}(\mathbf{w}(\mathbf{s}_i)-\mathbf{B}_{\mathbf{s}_i}\mathbf{w}_{N(\mathbf{s}_i)})\right)$$

$$(\tilde{\mathbf{C}}_{\mathcal{S}})^{-1} = \mathbf{B}_{\mathcal{S}}'\mathbf{F}_{\mathcal{S}}^{-1}\mathbf{B}_{\mathcal{S}} \qquad \det((\mathbf{B}_{\mathcal{S}}'\mathbf{F}_{\mathcal{S}}^{-1}\mathbf{B}_{\mathcal{S}})^{-1}) = \prod \det(\mathbf{F}_{\mathbf{s}_i})$$

Hierarchical model evaluation:

$$p(\boldsymbol{\theta}) \times \prod_{j=1}^{l} IG\left(\tau_j^2 \,\middle|\, a_{\tau_j}, b_{\tau_j}\right) \times N\left(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_{\beta}, \mathbf{V}_{\beta}\right) \times N\left(\mathbf{w}_{\mathcal{U}} \,|\, \mathbf{B}_{\mathcal{U}}\mathbf{w}_{\mathcal{S}}, \mathbf{F}_{\mathcal{U}}\right)$$
$$\times N\left(\mathbf{w}_{\mathcal{S}} \,\middle|\, \mathbf{0}, \widetilde{\mathbf{C}}_{\mathcal{S}}\right) \times \prod_{i=1}^{n} N\left(\mathbf{y}(\mathbf{t}_i) \,|\, \mathbf{X}(\mathbf{t}_i)'\boldsymbol{\beta} + \mathbf{Z}(\mathbf{t}_i)'\mathbf{w}(\mathbf{t}_i), \mathbf{D}\right),$$

			NNGP ($\mathcal{S} = \mathcal{T}$)	
	True	Full Gaussian Process	Order by y-coordinates	Order by x-coordinates
σ^2	1	0.640 (0.414, 1.297)	0.712 (0.449, 1.530)	0.757 (0.479, 1.501)
τ^2	0.1	0.107 (0.098, 0.117)	0.106 (0.097, 0.114)	0.107 (0.099, 0.117)
ϕ	6	8.257 (4.056, 13.408)	8.294 (3.564, 12.884)	7.130 (3.405, 11.273)