Smoothed Full-Scale Approximation of Gaussian Process Models for Computation of Large Spatial Data Sets

Bohai Zhang, Huiyan Sang and Jianhua Z. Huang

Nankai University and Texas A&M University

Supplementary Material

The supplementary material contains the proofs of Theorem 1 (Section S1.1), Proposition 1 (Section S1.2) and Theorem 2 (Section S1.3), additional numerical studies on the ordering of blocks for SFSA (Section S2.1), prediction performance of SFSA compared with other competing methods for a large (100,000) 2D spatial example (Section S2.2), performance of both parameter estimation and prediction of SFSA for a medium size 2D spatial data (Section S2.3), the Bayesian analysis results of a precipitation dataset (Section S2.4), and a discussion on selecting the nearest neighboring blocks based on the residual correlations (Section S3).

S1. Proof of Theorems

S1.1. Proof of Theorem 1

We provide the proof of Theorem 1 here. Without loss of generality, let $\boldsymbol{\beta} = \mathbf{0}$ for notation simplicity. We first prove that the approximated density in (2.7) is Gaussian. Let U denote $\mathcal{C}(S, S^*)\mathcal{C}(S^*, S^*)^{-1}$, U_k denote $\mathcal{C}(S_k, S^*)\mathcal{C}(S^*, S^*)^{-1}$, and $U_{N(k)}$ denote $\mathcal{C}(S_{N(k)}, S^*)\mathcal{C}(S^*, S^*)^{-1}$; then $\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\theta})$ is proportional to:

$$\exp\{-\frac{1}{2}\sum_{k=1}^{K}(\mathbf{y}_k - U_k\mathbf{w}^* - \Sigma_{k,N(k)}\Sigma_{N(k)}^{-1}(\mathbf{y}_{N(k)} - U_{N(k)}\mathbf{w}^*))^T$$

$$\times \quad \Sigma_{k|N(k)}^{-1}(\mathbf{y}_k - U_k\mathbf{w}^* - \Sigma_{k,N(k)}\Sigma_{N(k)}^{-1}(\mathbf{y}_{N(k)} - U_{N(k)}\mathbf{w}^*))\} \cdot \prod_{k=1}^{K} |\Sigma_{k|N(k)}|^{-\frac{1}{2}},$$

where $\Sigma_{k|N(k)} = \Sigma_k - \Sigma_{k,N(k)}\Sigma_{N(k)}^{-1}\Sigma_{k,N(k)}^T$, $\Sigma_k = \mathcal{C}_s(S_k, S_k) + \tau^2 I_{n_k}$, $\Sigma_{k,N(k)} = \mathcal{C}_s(S_k, S_{N(k)})$, and $\Sigma_{N(k)} = \mathcal{C}_s(S_{N(k)}, S_{N(k)}) + \tau^2 I_{n_{N(k)}}$. Next, we introduce nota-

tions for obtaining the quadratic term of the Gaussian density. Let

$$B_{k,l} = \begin{cases} I_{n_k}, & \text{if } l = k; \\ \left[-\Sigma_{k,N(k)}\Sigma_{N(k)}^{-1}\right](\cdot, n_{(l-1)}+1 : n_{(l)}), & \text{if } l \in N(k); \\ \mathbf{0}, & \text{otherwise,} \end{cases}$$

where $n_{(l)} = \sum_{1\le i\le l, i\in N(k)} n_i$, and recall that $N(k)$ denotes the neighbor set for the k-th block. Let $B_k^* = (B_{k,1}, \ldots, B_{k,K})$, then we can obtain that

$$\mathbf{y}_k - \Sigma_{k,N(k)}\Sigma_{N(k)}^{-1}\mathbf{y}_{N(k)} = B_k^*\mathbf{y}$$

and

$$U_k - \Sigma_{k,N(k)}\Sigma_{N(k)}^{-1}U_{N(k)} = B_k^*U.$$

Therefore,

$$\begin{aligned} \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\theta}) &\propto \exp\{-\frac{1}{2}\sum_{k=1}^{K}(\mathbf{y}-U\mathbf{w}^*)^T B_k^{*T}\Sigma_{k|N(k)}^{-1}B_k^*(\mathbf{y}-U\mathbf{w}^*)\}\cdot|\Sigma_{con}|^{-\frac{1}{2}} \\ &= \exp\{-\frac{1}{2}(\mathbf{y}-U\mathbf{w}^*)^T B^T\Sigma_{con}^{-1}B(\mathbf{y}-U\mathbf{w}^*)\}\cdot|\Sigma_{con}|^{-\frac{1}{2}}, \end{aligned}$$

where $B = (B_1^{*T}, B_2^{*T}, \ldots, B_K^{*T})^T \in \mathbb{R}^{n\times n}$ and $\Sigma_{con} = \mathrm{diag}\{\Sigma_{1|N(1)}, \ldots, \Sigma_{K|N(K)}\}$. Since $B_{k,l}$ is a nonzero matrix only for $l \le k$, B is an $n\times n$ lower-triangular matrix with ones as its diagonal entries. Hence, $|B| = 1$ and it is clear that

$$\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\theta}) = \mathcal{N}(U\mathbf{w}^*, B^{-1}\Sigma_{con}B^{T^{-1}}).$$

Recall that the marginal likelihood by SFSA is:

$$\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) = \int_{\mathbf{w}^*}\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\theta})\cdot p(\mathbf{w}^*|\boldsymbol{\theta})d\mathbf{w}^*.$$

After integrating out $\mathbf{w}^*$, it can be shown that

$$\begin{aligned}\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \;\propto\; & \exp\{-\frac{1}{2}\mathbf{y}^T B^T(\Sigma_{con}^{-1} - \Sigma_{con}^{-1}BU\Sigma_{\mathbf{w}^*}U^T B^T\Sigma_{con}^{-1})B\mathbf{y}\} \\ & \times|U^T B^T\Sigma_{con}^{-1}BU + C_*^{-1}|^{-\frac{1}{2}} \cdot |\Sigma_{con}|^{-\frac{1}{2}} \cdot |C_*|^{-\frac{1}{2}}, \qquad \text{(S1.1)}\end{aligned}$$

where $C_* \equiv \mathcal{C}(S^*, S^*)$ and $\Sigma_{\mathbf{w}^*}^{-1} = U^T B^T\Sigma_{con}^{-1}BU + C_*^{-1}$. By Sherman-Woodbury-Morrison inversion formula,

$$\Sigma_{con}^{-1} - \Sigma_{con}^{-1}BU\Sigma_{\mathbf{w}^*}U^T B^T\Sigma_{con}^{-1} = (\Sigma_{con} + BUC_*U^T B^T)^{-1}. \qquad \text{(S1.2)}$$

Therefore,

$$B^T(\Sigma_{con}^{-1} - \Sigma_{con}^{-1}BU\Sigma_{\mathbf{w}^*}U^T B^T\Sigma_{con}^{-1})B = (B^{-1}\Sigma_{con}B^{T^{-1}} + UC_*U^T)^{-1}.$$

By using the fact that $|B| = 1$ and the Sylvester's theorem, we can show that

$$\begin{aligned} & |B^{-1}\Sigma_{con}B^{T^{-1}} + UC_*U^T| \\ = \; & |B^{-1}| \cdot |\Sigma_{con}| \cdot |I_n + \Sigma_{con}^{-1}BUC_*U^T B^T| \cdot |B^{T^{-1}}| \\ = \; & |\Sigma_{con}| \cdot |I_m + U^T B^T\Sigma_{con}^{-1}BUC_*| \\ = \; & |\Sigma_{con}| \cdot |U^T B^T\Sigma_{con}^{-1}BU + C_*^{-1}| \cdot |C_*|. \qquad \text{(S1.3)}\end{aligned}$$

Thus, the marginal likelihood by SFSA follows the Gaussian distribution given in Theorem 1. It is also straightforward to prove the positive definiteness of the covariance matrix $C_{\mathbf{y}}^{\dagger}$ in Theorem 1. Recall that Σ_{con} is obtained based on the residual covariance function $\mathcal{C}_s(\mathbf{s}, \mathbf{s}') + \tau^2\delta(\mathbf{s}, \mathbf{s}')$, where $\delta(\cdot, \cdot)$ is the Kronecker delta function, it is hence positive definite by the Schur complement rule. In addition, B is a lower triangular matrix of a full rank n, and the predictive process covariance, $\mathcal{C}(S, S^*)\mathcal{C}(S^*, S^*)^{-1}\mathcal{C}(S, S^*)^T$, is positive semi-definite. Therefore, the approximated data covariance matrix, $C_{\mathbf{y}}^{\dagger}$, is positive definite.

S1.2. Proof of Proposition 1

Without loss of generality, we assume $\boldsymbol{\beta} = 0$ for notation simplicity. Let the partition rule $\mathcal{P}$ partition the predictive location set S_p into K disjoint blocks $S_{p,k}$, $k = 1, \ldots, K$, with the corresponding observation vector $\mathbf{y}_p$ partitioned as $\mathbf{y}_p = \cup_{k=1}^K \mathbf{y}_{p,k}$; suppose that each $\mathbf{y}_{p,k}$ has a size $n_{p,k}$ such that $\sum_{k=1}^K n_{p,k} = n_p$.

Then according to the assumptions of SFSA, the joint density of $\mathbf{y}_p$ and $\mathbf{y}$ is approximated as follows:

$$\begin{aligned}\tilde{p}(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\theta}) &= \int \tilde{p}(\mathbf{y}_p|\mathbf{y}, \mathbf{w}^*, \boldsymbol{\theta}) \cdot \tilde{p}(\mathbf{y}|\mathbf{w}^*, \boldsymbol{\theta}) \cdot p(\mathbf{w}^*|\boldsymbol{\theta}) d\mathbf{w}^* \\ &= \int \prod_{k=1}^{K} p(\mathbf{y}_{p,k}|\mathbf{y}_k, \mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\theta}) \cdot \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\theta}) \cdot p(\mathbf{w}^*|\boldsymbol{\theta}) d\mathbf{w}^*.\end{aligned}$$

Let $U_{p,k} = \mathcal{C}(S_{p,k}, S^*)\mathcal{C}(S^*, S^*)^{-1}$ and $B_{p,k} = (B_{p,k,1}, \ldots, B_{p,k,K})$, where $B_{p,k,l}$ has the similar definition to $B_{k,l}$ in (2.8), encoding the residual dependence information of $\mathbf{y}_{p,k}$ given its neighbors, $\mathbf{y}_k$ and $\mathbf{y}_{N(k)}$, for the l-th block, $l = 1, \ldots, K$. The Gaussian density $p(\mathbf{y}_{p,k}|\mathbf{y}_k, \mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\theta})$ has the following quadratic term:

$$(\mathbf{y}_{p,k} - U_{p,k}\mathbf{w}^* + B_{p,k}(\mathbf{y} - U\mathbf{w}^*))^T \Sigma_{p,k|N(k)}^{-1} (\mathbf{y}_{p,k} - U_{p,k}\mathbf{w}^* + B_{p,k}(\mathbf{y} - U\mathbf{w}^*)),$$

where $\Sigma_{p,k|N(k)}$ is the residual conditional variance of $\mathbf{y}_{p,k}$ given its neighbors. Let $B_{p,k}^* = (\mathbf{0}, \ldots, I_{n_{p,k}}, \ldots, \mathbf{0}, B_{p,k}) \in \mathbb{R}^{n_{p,k} \times (n+n_p)}$, $\tilde{\mathbf{y}} = (\mathbf{y}_{p,1}^T, \ldots, \mathbf{y}_{p,K}^T, \mathbf{y}^T)^T \in \mathbb{R}^{(n+n_p)\times 1}$, and $\tilde{U} = (U_{p,1}^T, \ldots, U_{p,K}^T, U^T)^T \in \mathbb{R}^{(n+n_p)\times m}$; then the quadratic term of $p(\mathbf{y}_{p,k}|\mathbf{y}_k, \mathbf{y}_{N(k)}, \mathbf{w}^*, \boldsymbol{\theta})$ can be written as:

$$(\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*)^T B_{p,k}^{*T} \Sigma_{p,k|N(k)}^{-1} B_{p,k}^* (\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*).$$

Hence,

$$\begin{aligned}\tilde{p}(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\theta}) \propto & \int \exp\{-\frac{1}{2}\sum_{k=1}^{K} (\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*)^T B_{p,k}^{*T} \Sigma_{p,k|N(k)}^{-1} B_{p,k}^* (\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*)\} \\ & \times \exp\{-\frac{1}{2}(\mathbf{y} - U\mathbf{w}^*)^T B^T \Sigma_{con}^{-1} B(\mathbf{y} - U\mathbf{w}^*)\} \\ & \times \exp\{-\frac{1}{2}\mathbf{w}^* C_*^{-1} \mathbf{w}^*\} \cdot \prod_{k=1}^{K} |\Sigma_{p,k|N(k)}|^{-\frac{1}{2}} \cdot |\Sigma_{con}|^{-\frac{1}{2}} \cdot |C_*|^{-\frac{1}{2}} d\mathbf{w}^*,\end{aligned}$$

where $C_* \equiv \mathcal{C}(S^*, S^*)$, and Σ_{con} and B are given in Theorem 1. Let $B_p^* = (B_{p,1}^{*T}, \ldots, B_{p,K}^{*T})^T$ and $\Sigma_{p,con} = \mathrm{diag}\{\Sigma_{p,1|N(1)}, \ldots, \Sigma_{p,K|N(K)}\}$; then it can be shown that

$$\sum_{k=1}^{K} (\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*)^T B_{p,k}^{*T} \Sigma_{p,k|N(k)}^{-1} B_{p,k}^* (\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*) = (\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*)^T B_p^{*T} \Sigma_{p,con}^{-1} B_p^* (\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*).$$

Let $B_p = (B_{p,1}^T, \ldots, B_{p,K}^T)^T$, $\tilde{B} = \begin{pmatrix} I_{n_p} & B_p \\ \mathbf{0} & B \end{pmatrix}$ and $\tilde{\Sigma}_{con} = \begin{pmatrix} \Sigma_{p,con} & \mathbf{0} \\ \mathbf{0} & \Sigma_{con} \end{pmatrix}$; since $B_p^* = (I_{n_p},\ B_p)$, it can be shown that

$$\begin{aligned} \tilde{p}(\mathbf{y}_p, \mathbf{y}|\boldsymbol{\theta}) \ \propto \ & \int \exp\{-\frac{1}{2}(\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*)^T \tilde{B}^T \tilde{\Sigma}_{con}^{-1} \tilde{B}(\tilde{\mathbf{y}} - \tilde{U}\mathbf{w}^*) - \frac{1}{2}\mathbf{w}^* C_*^{-1} \mathbf{w}^*\} \\ & \times |\tilde{\Sigma}_{con}|^{-\frac{1}{2}} \cdot |C_*|^{-\frac{1}{2}} d\mathbf{w}^*. \end{aligned}$$

After integrating out $\mathbf{w}^*$, one can obtain that

$$\mathbf{y}_p, \mathbf{y}|\boldsymbol{\theta} \sim \mathcal{N}(\tilde{\mathbf{x}}\boldsymbol{\beta}, \tilde{B}^{-1}\tilde{\Sigma}_{con}\tilde{B}^{T^{-1}} + \tilde{U}C_*\tilde{U}^T),$$

where $\tilde{\mathbf{x}} = (\mathbf{x}_p^T, \mathbf{x}^T)^T$. Since

$$\tilde{B}^{-1} = \begin{pmatrix} I_{n_p} & B_p \\ \mathbf{0} & B \end{pmatrix}^{-1} = \begin{pmatrix} I_{n_p} & -B_p B^{-1} \\ \mathbf{0} & B^{-1} \end{pmatrix},$$

by the properties of the Gaussian distribution, $\mathbf{y}_p|\mathbf{y}, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_p, \Sigma_p)$, where μ_p and Σ_p are given in Proposition 1.

S1.3. Proof of Theorem 2

Since assumptions (5.1) and (5.2) are the block-version assumptions for the nearest neighbor Gaussian process, by using the results in Datta et al. (2016), we can verify that the finite dimensional density defined in (5.3) satisfies the conditions of Kolmogorov consistency theorem and hence leads to a valid spatial process, denoted by $\tilde{w}_s^\dagger(\mathbf{s})$. According to the law of total covariance,

$$\begin{aligned} \tilde{\mathrm{Cov}}(\tilde{w}_s^\dagger(\mathbf{s}), \tilde{w}_s^\dagger(\mathbf{s}')) \ &= \ \tilde{\mathrm{E}}(\tilde{\mathrm{Cov}}(\tilde{w}_s^\dagger(\mathbf{s}), \tilde{w}_s^\dagger(\mathbf{s}')|\tilde{w}_s(S))) \\ &+ \ \tilde{\mathrm{Cov}}(\tilde{\mathrm{E}}(\tilde{w}_s^\dagger(\mathbf{s})|\tilde{w}_s(S)), \tilde{\mathrm{E}}(\tilde{w}_s^\dagger(\mathbf{s}')|\tilde{w}_s(S))), \end{aligned}$$

where $\tilde{\mathrm{Cov}}(\cdot,\cdot)$ and $\tilde{\mathrm{E}}(\cdot)$ are the covariance and expectation operators for the spatial process $\tilde{w}_s^\dagger(\mathbf{s})$. The covariance function of $\tilde{w}_s^\dagger(\mathbf{s})$ can be readily obtained using the law of total covariance. We take the scenario that $\mathbf{s}, \mathbf{s}' \notin S$ but belong to the same block k for illustration. In this case, $\tilde{\mathrm{E}}(\tilde{w}_s^\dagger(\mathbf{s})|\tilde{w}_s(S)) = -B_{\mathbf{s}}\tilde{w}_s(S)$

and $\tilde{\mathrm{E}}(\tilde{w}_s^\dagger(\mathbf{s}')|\tilde{w}_s(S))) = -B_{\mathbf{s}'}\tilde{w}_s(S)$, thus

$$\tilde{\mathrm{Cov}}(\tilde{\mathrm{E}}(\tilde{w}_s^\dagger(\mathbf{s})|\tilde{w}_s(S)), \tilde{\mathrm{E}}(\tilde{w}_s^\dagger(\mathbf{s}')|\tilde{w}_s(S))) = B_{\mathbf{s}}\tilde{\mathrm{Cov}}(\tilde{w}_s(S), \tilde{w}_s(S))B_{\mathbf{s}'}^T = B_{\mathbf{s}}\Sigma_{\mathbf{y}}^\dagger B_{\mathbf{s}'}^T,$$

where $\Sigma_{\mathbf{y}}^\dagger = B^{-1}\Sigma_{con}B^{T^{-1}}$. Following the assumption given in (5.2), we can see that $\tilde{\mathrm{Cov}}(\tilde{w}_s^\dagger(S_{p,k}), \tilde{w}_s^\dagger(S_{p,k})|\tilde{w}_s(S))$, is the residual conditional variance of $\tilde{w}_s(S_{p,k})$ given its neighboring observations $\tilde{w}_s(S_k)$ and $\tilde{w}_s(S_{N(k)})$, denoted by $\Sigma_{p,k|N(k)}$. Thus,

$$\tilde{\mathrm{E}}(\tilde{\mathrm{Cov}}(\tilde{w}_s^\dagger(\mathbf{s}), \tilde{w}_s^\dagger(\mathbf{s}')|\tilde{w}_s(S))) = \Sigma_{p,k|N(k)}(\mathbf{s}, \mathbf{s}').$$

S2. Ordering of Blocks for SFSA and Additional Numerical Results

S2.1. Ordering of blocks for SFSA

We discuss the effect of block ordering on the likelihood approximation for SFSA. Following Guinness (2016), here we consider four block ordering methods: 1) The sorted coordinate ordering that sorts blocks according to their x- and y- coordinates (denoted by "SC"); 2) the maximum-minimum-distance ordering (denoted by "MMD"); 3) the random ordering (denoted by "RAND"); and the center-out ordering that orders the blocks according to their distances to the center of all observation locations (denoted by "CO"). For the SC ordering, we sorted the blocks first according to their y-axis coordinates and then according to their x-axis coordinates; for the CO ordering, the center of observations is defined as the mean of coordinates of all observation locations. Since the Kullback-Leibler (KL) divergence measures the distance between the approximate likelihood and the full likelihood, we compare the performance of different ordering methods in terms of the KL divergence values that are obtained by plugging in the true parameter values.

Table 1 shows the KL divergence results for two simulation examples. The first example is the 2-dimensional example in Section S2.3 with 4000 non-uniformly distributed observations, and the second example is the 2-dimensional example in Section S2.2 with 10^5 uniformly distributed observations; since calculating the determinant of the full covariance matrix is computationally expensive for large sample size, for the second example of 10^5 observations, we randomly selected 10^4 observations for calculating the KL divergence. For each simulation example,

we considered the Gaussian covariance function with different range values.

It is clear that for the non-uniformly distributed observations (example 1), the CO ordering produces the best results, while for the uniformly distributed observations (example 2), the SC ordering performs the best. The MMD ordering works the best for the nearest neighbor Gaussian process (Guinness, 2016) but does not work very well for SFSA, which may be because the ordering of SFSA is for approximating the residual likelihood instead of the original likelihood.

Table 1: The Kullback-Leibler divergence for the approximated likelihood by SFSA with different methods for the block ordering. For SFSA, the number of neighboring blocks $q = 1$.

Example 1	SC	MMD	RAND	CO
$\phi = 0.2$	377.52	429.68	357.35	299.83
$\phi = 0.5$	319.79	366.59	311.97	254.36
$\phi = 2$	83.99	90.34	82.47	77.96
Example 2	SC	MMD	RAND	CO
$\phi = 0.5$	204.96	257.78	234.53	214.75
$\phi = 2$	1.30	1.56	1.61	2.35

S2.2. Prediction for a large 2D spatial data

We generated $100,000$ locations in a $(0, 10) \times (0, 10)$ square domain. The data were realizations from a Gaussian process with mean zero and a Gaussian covariance function with $\sigma^2 = 1$ and $\tau^2 = 0.1$; we varied the range parameters to account for different dependence scales. Two prediction scenarios were considered: 1) Prediction on $10,000$ randomly selected locations (denoted by "MAR") and 2) prediction on locations in a spatial hole $(4, 6) \times (4, 6)$ in the central region of the study domain (denoted by "MBD"); the first scenario accounts for the small-range prediction performance and the second scenario accounts for the large-range prediction performance.

We compare SFSA with a few state-of-the-art methods, including the local GP with adaptive designs (LaGP) (see Gramacy and Apley (2015)), the nearest-neighbor GP (NNGP) proposed by Datta et al. (2016), and the variants of SFSA (including FSA-Block and CBCL); note that CBCL is a block-version of NNGP. For LaGP, we consider three heuristics for selecting the local design points: The active-learning-cohn heuristic (alc) proposed by Cohn (1996), the

nearest-neighbor heuristic (nn), and the mean-squared-prediction-error heuristic (mspe) developed in Gramacy and Apley (2015). For different heuristics, the total number of design points is 100, and the alc and mspe heuristics start with 50 nearest neighbors; we also considered the LaGP with a larger number of nearest neighbors (500 neighbors), denoted by "nn-big". For NNGP, the observed location set was used as the reference set and the observations were ordered according to the sum of their x- and y- coordinates; then 100 nearest neighbors were selected for both parameter-estimation and prediction steps. For SFSA, the blocks are regular blocks with centers on a 20×20 regular grid in the study domain, and the nearest neighboring block is used to correct the residual covariance matrix; the knots are from a regular-grid location set with $m = 225$. As discussed in Section 2.5, the sorted-coordinate (SC) ordering and the center-out (CO) ordering were used for SFSA under the MAR and MBD prediction scenarios, respectively. FSA-Block and CBCL are special cases of SFSA with $q = 0$ and $m = 0$, respectively.

Table 2: Mean Squared Prediction Errors (MSPEs) of SFSA (and its variants), LaGP, and NNGP. The results were obtained based on 20 simulated data sets.

Gauss		SFSA	LaGP				FSA-Block	NNGP	CBCL
range	design		nn	alc	mspe	nn-big			
$\phi = 0.5$	MAR	0.101	0.102	0.101	0.101	0.100	0.101	0.101	0.102
	MBD	0.264	0.363	0.306	0.301	0.282	0.308	0.334	0.351
$\phi = 2$	MAR	0.100	0.101	0.101	0.101	0.101	0.100	0.101	0.101
	MBD	0.104	0.186	0.134	0.134	0.134	0.104	0.175	0.176

The prediction results are summarized in Table 2. For the MAR scenario, all the methods have comparable prediction performances, indicating that they all have very similar performances for small-range predictions. For the MBD scenario, SFSA outperforms the LaGP method with different heuristics on selecting local design points; especially for a larger range value $\phi = 2$, SFSA results in a much smaller MSPE. Hence SFSA has better large-range prediction performances than the local GP approximation. For the LaGP method, the alc and mspe heuristics lead to much smaller prediction errors than the nn heuristic, indicating that choosing some points far away from the prediction location can help improve the prediction accuracy. But using the alc or mspe heuristic is more

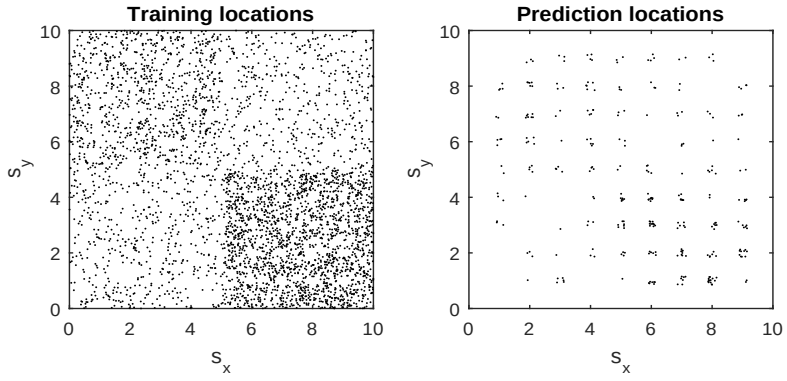
computationally expensive, with similar computational time to the nn heuristic with 500 local design points. The NNGP method has better prediction performances than the LaGP with the nn heuristic, but it is inferior to LaGP with the alc and mspe heuristics.

Under the MBD scenario, FSA-Block does not work very well for a small range $\phi = 0.5$, since its low-rank component cannot give a satisfactory approximation to the original process; it gives similar results to SFSA for a larger range $\phi = 2$ as expected, since the low-rank component approximates the original process quite well in this case. The CBCL method is a block-version of NNGP and yields slightly inferior prediction performances to NNGP. Since SFSA is a composite of FSA-Block and CBCL, it yields more robust prediction performances for different range values, and under different prediction scenarios.

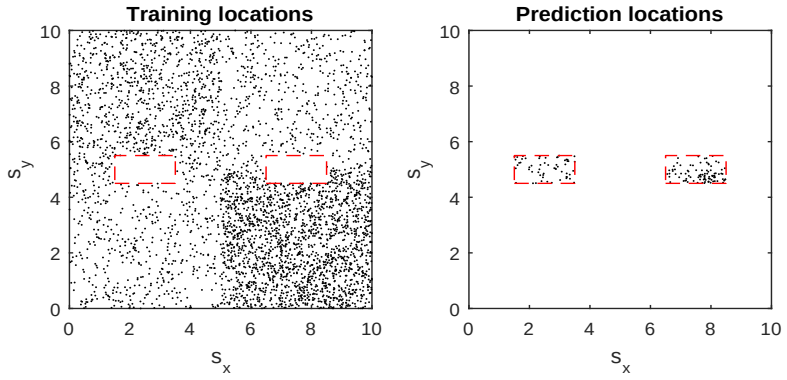
S2.3. Parameter estimation and prediction for a medium size spatial data

In this section, we compare SFSA with its variants FSA-Block and CBCL in terms of both parameter estimation and prediction; the results by the full covariance model (denoted by "FM") are the "gold standard," since it theoretically works the best. We generated 4000 locations in a square domain $\mathcal{S} \equiv [0, 10] \times [0, 10]$. These locations were non-uniformly distributed, with 500 locations in each of the sub-domains $[0, 5] \times [0, 5]$ and $[5, 10] \times [5, 10]$, 1000 locations in the sub-domain $[0, 5] \times [5, 10]$, and 2000 locations in the sub-domain $[5, 10] \times [0, 5]$ (see Figure 1). We considered two prediction settings: 1). Prediction on locations near block boundaries (denoted by "**Boundary**"), where block boundaries were created by a 10×10 regular grid on $\mathcal{S}$ ($s_x = 1, \ldots, 9$ and $s_y = 1, \ldots, 9$ constituted the block boundaries). The locations within 0.15 distances to the crosses of block boundaries were selected for prediction, and the rest of locations were used for parameter estimation; 2). prediction on locations in spatial holes (denoted by "**Hole**"), where locations in two rectangle regions $[1.5, 3.5] \times [4.5, 5.5]$ and $[6.5, 8.5] \times [4.5, 5.5]$ were selected for prediction, and the rest of locations were used for parameter estimation.

For both simulation settings, the regular-grid block boundaries were used to define blocks for all comparison methods, and equally spaced grid knots were



(a) Prediction around block boundaries.



(b) Prediction in holes.

Figure 1: Upper panels show training and prediction locations for the "Boundary" scenario and lower panels show training and prediction locations for the "Hole" scenario. The rectangle boxes in lower panels indicate the hold-out spatial holes.

used for both FSA-block and SFSA; for SFSA and CBCL, the neighbor set $S_{N(k)}$ for the k-th block was specified as the nearest neighboring block ($q = 1$), and the block numbers were ordered from left to right, top to bottom. The simulation data sets were generated from the GP model with $\boldsymbol{\beta} = \mathbf{0}$ and the Matérn covariance function with a nugget effect. We experimented the Matérn covariance function with different smoothness- and range-parameter values for comparing performances of different methods.

Table 3 shows the parameter-estimation results for the simulated data with the Gaussian covariance function under the "Boundary" prediction scenario. We will focus on this scenario, since for parameter estimation, no big differences are

observed between the "Boundary" scenario and the "Hole" scenario. We obtained Relative Efficiency (RE) for parameter estimators by each method, where RE is defined as the ratio of Mean Squared Error (MSE) of an estimator by FM to that by using an approximated inference method. A RE value as close to 1 as possible is preferred, because the full covariance model theoretically leads to estimators of the highest efficiencies (in terms of the smallest MSEs).

Table 3: Parameter-estimation results for the Gaussian covariance function. Relative Efficiencies (REs) of parameter estimates by different methods are reported and the results were obtained based on 200 simulated data sets. For FSA-Block and SFSA, $m = 100$ equally spaced knots were used.

Gauss	SFSA	FSA-Block	CBCL
$\sigma^2(1)$	0.80	0.67	0.82
$\phi(\mathbf{0.2})$	0.70	0.44	0.71
$\tau^2(0.01)$	1.03	0.90	1.02
$\sigma^2(1)$	0.90	0.73	0.86
$\phi(\mathbf{0.5})$	0.60	0.36	0.56
$\tau^2(0.01)$	0.87	0.78	0.87
$\sigma^2(1)$	0.89	0.60	0.48
$\phi(\mathbf{2})$	0.62	0.32	0.26
$\tau^2(0.01)$	0.94	0.93	0.95

For the Gaussian covariance model (the Matérn covariance with $\nu \to \infty$) with a small range ($\phi = 0.2$ or 0.5), SFSA and CBCL have comparable REs and both of them outperform FSA-Block. This is because the predictive-process component in SFSA/FSA-Block cannot borrow much information from the knots due to the weak correlations between observations; hence in order to increase the parameter-estimation efficiency, it is more effective to borrow information from neighboring locations than to increase the knot number. When the range is relatively large ($\phi = 2$), SFSA leads to the largest relative efficiencies for both the variance and the range parameters among three approaches, since borrowing information from either the neighboring locations or the knot locations is very effective in this case; also both SFSA and FSA-Block outperform CBCL, which may be because CBCL ignores correlations between each block and its non-neighboring blocks, and the loss of dependence information is more severe for a large range value. Since for all three range values, SFSA results in either comparable or higher REs for all covariance parameters, it is more robust to data-dependence structures in terms

of parameter estimation.

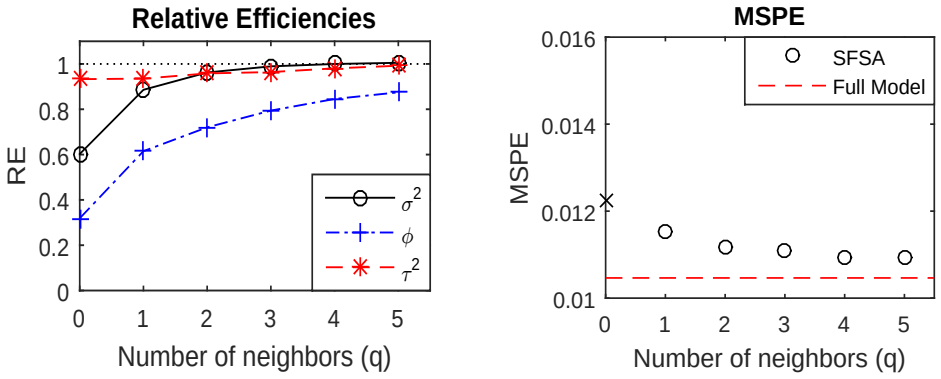
Then we compare prediction performances of these three methods. Table 4 shows Mean Squared Prediction Errors (MSPEs) of each method, under both "Boundary" and "Hole" scenarios. For the "Boundary" scenario, SFSA and CBCL have comparable MSPEs for the Gaussian model with small range values, and both methods have much smaller MSPEs than those by FSA-Block in this case. This is because the smooth component, predictive process part of FSA-Block, does not perform well for small range values with limited number of knots; and conditioning on "similar" neighboring observations for predictions (set $q \geq 1$) is more effective to reduce prediction errors at locations around block boundaries. When ϕ increases to 2, all three methods have comparable MSPEs, and SFSA and FSA-Block have slightly smaller MSPEs than that by CBCL. This indicates that enhancing the predictive-process component in FSA-Block for large range values can help reduce prediction errors around boundaries significantly.

Table 4: Prediction results for the Gaussian covariance function, under both the"Boundary" scenario and the "Hole" scenario. Mean Squared Prediction Errors (MSPE) and their standard errors (in parentheses) were obtained based on 200 simulated data sets. For FSA-Block and SFSA, $m = 100$ equally spaced knots were used.

Gauss		SFSA	FSA-Block	CBCL	FM
$\phi = 0.2$	Boundary	0.063 (0.010)	0.091 (0.015)	0.069 (0.012)	0.023 (0.003)
	Hole	0.321 (0.137)	0.337 (0.144)	0.331 (0.143)	0.267 (0.115)
$\phi = 0.5$	Boundary	0.023 (0.003)	0.032 (0.004)	0.028 (0.004)	0.012 (0.001)
	Hole	0.093 (0.040)	0.116 (0.051)	0.110 (0.051)	0.050 (0.024)
$\phi = 2$	Boundary	0.012 (0.001)	0.012 (0.001)	0.014 (0.001)	0.010 (0.001)
	Hole	0.014 (0.003)	0.016 (0.003)	0.024 (0.006)	0.012 (0.002)

For the "Hole" scenario, SFSA outperforms other two methods for the Gaussian covariance model with different ranges, indicating that combining the strengths of borrowing dependence information from neighboring locations around holes and the knot locations close to holes can further increase the prediction accuracy. Especially for a moderately large range value ($\phi = 0.5$), SFSA leads to much smaller MSPEs than other two methods, since either borrowing information from neighbors (CBCL) or borrowing information from knots (FSA-Block) is not sufficient in this case.

Last, we investigate the effect of the number of neighboring blocks (q) on the



(a) REs v.s. q (b) MSPEs v.s. q

Figure 2: Relative Efficiencies (REs) and MSPEs versus q for SFSA using the Gaussian covariance model with $\sigma^2 = 1$, $\phi = 2$, and $\tau^2 = 0.01$, under the Boundary scenario. The q-nearest neighboring blocks were chosen as the neighbor set for SFSA. The case $q = 0$ corresponds to the FSA-Block result.

performance of parameter estimation and prediction of the SFSA. The left panel in Figure 2 shows the parameter-estimation results for the Gaussian covariance; we used $m = 100$ equally spaced knots and $K = 100$ equally partitioned blocks for the SFSA approach. For this relatively strong data-dependence structure, $q = 3$ (with about 120 neighboring observations for each block) can lead to REs of all estimators larger than 0.8. The right panel in Figure 2 shows how MSPEs decrease with increasing values of q, and $q = 2$ or 3 seems to be a good choice, since further increasing q cannot reduce prediction errors significantly.

We also experimented the Matérn covariance model with $\nu = 1.5$, and similar conclusions hold: When the range is small, SFSA has comparable performance to CBCL, and both methods outperform FSA-Block; when the range is relatively large, SFSA is superior to other two comparison methods in terms of REs.

S2.4. Analysis of a precipitation dataset

We apply our method to the precipitation dataset in United States in 1962, which contains yearly total precipitation anomalies that are yearly totals standardized by the long-run mean and standard deviation for each of the 7352 weather stations. This precipitation data was collected by the National Climate Data Center and has been analyzed in several studies (e.g., Johns et al. (2003);

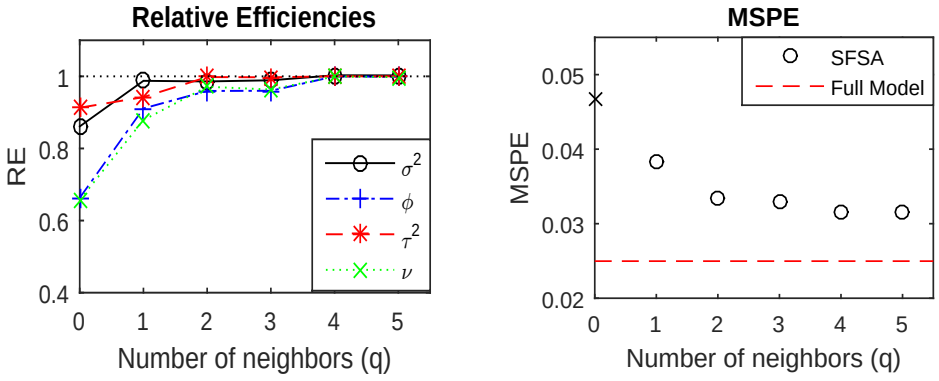
Table 5: Parameter-estimation results for the Matérn covariance function with $\nu = 1.5$. Relative Efficiencies (REs) of parameter estimates by different methods are reported and the results were obtained based on 200 simulated data sets. For FSA-Block and SFSA, $m = 100$ equally spaced knots were used.

Matérn	SFSA	FSA-Block	CBCL
$\sigma^2(1)$	0.99	0.86	0.89
$\phi(\mathbf{0.5})$	0.91	0.66	0.82
$\nu(1.5)$	0.88	0.66	0.85
$\tau^2(0.01)$	0.94	0.91	0.96
$\sigma^2(1)$	0.92	0.81	0.79
$\phi(\mathbf{1})$	0.89	0.71	0.74
$\nu(1.5)$	0.87	0.72	0.73
$\tau^2(0.01)$	0.93	0.86	0.95

Table 6: Prediction results for the Matérn covariance function with $\nu = 1.5$, under both the"Boundary" scenario and the "Hole" scenario. Mean Squared Prediction Errors (MSPE) and their standard errors (in parentheses) were obtained based on 200 simulated data sets. For FSA-Block and SFSA, $m = 100$ equally spaced knots were used.

Matérn		SFSA	FSA-Block	CBCL	FM
$\phi = 0.5$	Boundary	0.038 (0.005)	0.047 (0.006)	0.044 (0.006)	0.025 (0.003)
	Hole	0.117 (0.044)	0.124 (0.050)	0.138 (0.052)	0.097 (0.037)
$\phi = 1$	Boundary	0.018 (0.001)	0.019 (0.002)	0.021 (0.002)	0.014 (0.001)
	Hole	0.032 (0.010)	0.034 (0.011)	0.042 (0.015)	0.028 (0.009)

Kaufman et al. (2008); Sang and Huang (2012)). We used it as a benchmark dataset to compare SFSA with other competing methods. According to Johns et al. (2003), this dataset appears no significant non-stationarity and anisotropy. Therefore, we chose to use the spatial regression model in Section 2.1 with $\boldsymbol{\beta} = 0$ and an isotropic covariance function to fit the data. Since observations are on the sphere, the chordal distance with units of kilometers was used to compute the pairwise distances between weather stations, to ensure positive-definiteness of the covariance function. We partitioned the data into a training dataset of 7000 observations and a prediction dataset of 352 observations, where the prediction dataset contains 143 locations in a randomly specified space hole $(-87, -82) \times (35, 38)$ and 209 locations randomly selected from the remaining locations. The Matérn covariance was used to model the data-dependence structure. Since previous studies (Sang and Huang, 2012) indicated that the smooth-



(a) REs v.s. q (b) MSPEs v.s. q

Figure 3: Relative Efficiencies (REs) and MSPEs versus q for SFSA using the Matérn covariance model with $\sigma^2 = 1$, $\phi = 0.5$, $\nu = 1.5$, and $\tau^2 = 0.01$, under the Boundary scenario. The q-nearest neighboring blocks were chosen as the neighbor set for SFSA. The case that $q = 0$ corresponds to the FSA-Block result.

ness parameter ν is very close to 0.5, we fixed $\nu = 0.5$.

We investigate the performance of SFSA, with comparisons to two of its special cases FSA-Block and CBCL; the full covariance model results will serve as the baseline. For all comparison methods, the K-means clustering algorithm was applied to create data blocks based on the training data. For SFSA and CBCL, since the data locations are non-uniformly spaced, we ordered the data blocks in the dense region of observations first; then the neighbor set of a block was specified as its nearest neighboring data block ($q = 1$). For SFSA and FSA-Block, the K-means clustering algorithm was applied to the training dataset for obtaining 300 cluster centers that were used as the knot set. Both maximum likelihood estimation and Bayesian inference were considered for estimating model parameters. For Bayesian inference, we collected 6000 posterior samples of model parameters after a burn-in period of 1000 iterations; we obtained the Maximum A Posteriori (MAP) estimates of model parameters and the corresponding MSPEs by using the MAP estimates.

Table 7 shows the parameter-estimation and prediction results by each method, with parameter estimates obtained by maximum likelihood estimation (MLE). We can see that the resulting estimates of model parameters by these three methods are all close to the full model results. The SFSA approach produces the

Table 7: Maximum likelihood estimation results using the exponential model (Matérn covariance function with $\nu = 0.5$).

Method	σ^2	ϕ	τ^2	Log lik	MSPE
FSA-Block, $K = 70$	0.68	180.22	0.11	-5218.69	0.31
CBCL, $K = 70$	0.69	174.37	0.10	-5206.07	0.33
SFSA, $K = 70$	0.67	170.96	0.10	-5179.15	0.30
FSA-Block, $K = 25$	0.68	172.29	0.10	-5190.36	0.30
CBCL, $K = 25$	0.69	169.64	0.10	-5177.63	0.28
SFSA, $K = 25$	0.69	170.86	0.10	-5160.71	0.27
Full Model	0.68	166.84	0.10	-5150.60	0.27

largest log-likelihood value among three comparison methods for a given block number, since it approximates the full covariance model the best and includes other two methods as special cases. In terms of prediction, when the block number $K = 70$, the prediction errors of the SFSA and the FSA-Block methods are much smaller than that of the CBCL approach, and this may be because the additional correction of residual covariance is not sufficient for a relatively small block size. When the block number $K = 25$, the prediction errors of the SFSA and the CBCL methods turn to be much smaller than that of the FSA-Block approach, indicating that the additional correction of residual covariance between neighboring blocks becomes more effective for a larger block size.

Table 8: Bayesian inference results using the exponential model (Matérn covariance function with $\nu = 0.5$). Parameter posterior $50(2.5, 97.5)$ percentiles are reported.

Method	σ^2	ϕ	τ^2	DIC	G	P	D	MSPE
FSA-Block, $K = 70$	0.69(0.58,0.90)	182.11(148.30,243.83)	0.11(0.10,0.12)	10439.50	108.30	105.12	213.42	0.31
CBCL, $K = 70$	0.70(0.60,0.85)	179.34(148.71,224.64)	0.10(0.10,0.12)	10418.12	115.52	107.66	223.18	0.33
SFSA, $K = 70$	0.68(0.58,0.87)	172.05(143.23,231.91)	0.10(0.10,0.12)	10357.64	107.84	105.25	213.09	0.31
FSA-Block, $K = 25$	0.69(0.58,0.88)	176.14(142.39,240.19)	0.11(0.10,0.12)	10392.52	105.39	105.68	211.07	0.30
CBCL, $K = 25$	0.70(0.61,0.93)	174.44(143.23,242.79)	0.10(0.09,0.11)	10361.55	98.57	106.29	204.86	0.28
SFSA, $K = 25$	0.69(0.58,0.97)	174.17(141.37,256.07)	0.10(0.09,0.11)	10329.95	96.94	104.63	201.57	0.28
Full Model	0.70(0.59,0.85)	173.86(141.48,218.46)	0.10(0.09,0.11)	10307.03	95.22	101.04	196.26	0.27

Table 8 gives the Bayesian inference results, which agree with the results by MLE. We observe that the posterior median of each model parameter by SFSA is close to that by the full covariance model. Besides, the Deviance Information Criteria (Gelman et al., 2014) (DIC) value by SFSA is the smallest among the three methods, indicating that it fits this precipitation data the best. We also considered the posterior predictive loss criterion scores (Gelfand and Ghosh,

1998), where "G" denotes the sum of squared biases for the posterior predictive means, "P" denotes the sum of posterior predictive variances, and "D" is the sum of corresponding G and P values; a smaller D value indicates a better fit. The D value of SFSA is the smallest one for different block sizes. The posterior predictive interval results for the hold-out 352 observations are reported in Table 9, and all methods show appropriate 95% posterior predictive interval coverage rates.

Table 9: Predictive interval (PI) coverages and widths for the hold-out 352 observations. The pointwise 95% posterior predictive intervals were obtained as the posterior predictive means plus/minus corresponding 1.96 posterior predictive standard errors.

Method	95% PI cover %	95% PI width
FSA-Block, $K = 70$	0.955	2.100
CBCL, $K = 70$	0.955	2.118
SFSA, $K = 70$	0.957	2.099
FSA-Block, $K = 25$	0.957	2.104
CBCL, $K = 25$	0.966	2.107
SFSA, $K = 25$	0.966	2.093
Full Model	0.957	2.061

SFSA results in a comparable prediction result to that by either FSA-Block or CBCL for a given block size. Therefore, if we do not have any prior knowledge of the optimal block size, using SFSA can provide more robust model-inference and prediction results; otherwise, its special cases, either FSA-Block (SFSA with $q = 0$) or CBCL (SFSA with $m = 0$), can be sufficient in modeling the data.

S3. Selection of the Nearest Neighboring Blocks

In this section, we provide some thoughts on selecting the nearest neighboring blocks based on the residual correlations. In this paper we have focused on the regular-blocks partition, and it is very natural to use the distance between two block centers as the measure of their "closeness".

More optimally, the "closeness" of two blocks can be measured by the correlations of observations in two blocks. Recall that we apply the nearest neighboring blocks approximation to the residual process, and hence the residual correlations between observations in two blocks provide a natural way for measuring the closeness of two blocks. For the moment, assume that the covariance-function parameter $\boldsymbol{\theta}$ is known. For the k-th block, we can calculate the residual covariance matrices $\Sigma_{k,k}$, $\Sigma_{\ell,\ell}$, and $\Sigma_{k,\ell}$ for $1 \leq \ell < k$. Then the Frobenius norm of

the residual correlation matrix $R_{k,\ell} \equiv \text{diag}(\Sigma_{k,k}^{-1/2})\Sigma_{k,\ell}\text{diag}(\Sigma_{\ell,\ell}^{-1/2})$, denoted by $\|R_{k,\ell}\|_F$, can serve as a measure of distances between two blocks. Thus, the q-nearest neighboring blocks for block k are the ones with the first q largest values for $\|R_{k,\ell}\|_F$ by using this criterion.

In practice, the covariance-function parameters $\boldsymbol{\theta}$ are unknown and need to be estimated. One may obtain some rough estimates of parameters from pilot studies to use the above neighbor-selection strategy, or adaptively update neighbor-selection within the MCMC iteration (for Bayesian inference) or the Newton-Raphson type iteration (for frequentist inference).

References

Cohn, D. A. (1996). Neural network exploration using optimal experiment design. *Neural networks 9*(6), 1071–1083.

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*, 800–812.

Gelfand, A. E. and S. K. Ghosh (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika 85*, 1–11.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian Data Analysis*, Volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gramacy, R. B. and D. W. Apley (2015). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics 24*, 561–578.

Guinness, J. (2016). Permutation methods for sharpening Gaussian process approximations. *arXiv preprint arXiv:1609.05372*.

Johns, C. J., D. Nychka, T. G. F. Kittel, and C. Daly (2003). Infilling sparse records of spatial fields. *Journal of the American Statistical Association 98*, 796–806.

Kaufman, C. G., M. J. Schervish, and D. W. Nychka (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association 103*, 1545–1555.

Sang, H. and J. Huang (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*, 111–132.