A summary of some mainstream spatial statistical methods for large datasets

Yewen Chen

October 13, 2020

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 GAUSSIAN RANDOM FIELD AND ITS CHALLENGE

Gaussian spatial processes has been popular for decades in spatial data contexts like geostatistics where they are known as kriging, and in computer experiments where they are deployed as surrogate models or emulators. More recently, they have become a popular prediction engine in the machine learning literature. The reasons are many, but the most important are probably that: the Gaussian structure affords a large degree of analytic capability not enjoyed by other general-purpose approaches to nonparametric nonlinear modeling; and because they perform well in out-of-sample tests.

Assume there is a response or dependent variable $Y(s)$ at a generic location $s \in \mathbf{D} \subset \mathcal{R}^2$ along with a $p \times 1$ vector of spatially referenced predictors $\mathbf{X}(s)$. A spatial regression model has the form

$$y(s) = \boldsymbol{\beta}\mathbf{X}(s) + w(s) + \epsilon(s) \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. The residual from the regression is decomposed into two independent parts: a spatial process, $w(s)$, modelling spatial association, and an independent process, $\epsilon(s)$, also known as the nugget effect, modelling measurement error. The spatial process $w(s)$ in (1) is often referred to as spatial random effects, capturing the effect of unmeasured or unobserved covariates with spatial pattern.

The nugget effect $\epsilon(s)$ is often assumed to follow a normal distribution with variance τ^2 for every location s. The most common specification for $w(s)$ is $w(s) \sim GP(0, \mathbf{C}(\cdot,\cdot))$, a zero-mean Gaussian process with a valid covariance function $\mathbf{C}(s, s')$. It is often reasonable to assume a constant process variance and thus we specify $\mathbf{C}(s, s') = \sigma^2\rho(s, s'; \theta)$, where $\rho(s, s'; \theta)$ is a correlation function and θ is a vector of correlation parameters which needs to be estimated from a finite number of observations, $\mathbf{Y} = (y(s_1).\cdots, y(s_n))'$.

According to the above assumptions, $y(s)$ follows a spatial Gaussian process, and thus we have the log-likelihood function for $\Theta = (\boldsymbol{\beta}, \tau^2, \sigma^2, \theta)$:

$$l(\Theta) \propto -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \boldsymbol{\beta}\mathbf{X})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\beta}\mathbf{X}) \tag{2}$$

where $\boldsymbol{\Sigma} = \mathbf{C} + \tau^2\mathbf{I}$. From Bayesian point of view, the posterior distributions of the parameter Θ are derived by

$$p(\Theta|\mathbf{y}) \propto \overbrace{p(\mathbf{y}|\boldsymbol{w}, \boldsymbol{\beta}, \tau^2)}^{\texttt{data model}} \underbrace{p(\boldsymbol{w}|\theta)}_{\texttt{process model}} \pi(\Theta) \tag{3}$$

From (3), the real challenge lies in the process model $p(\boldsymbol{w}|\theta)$ rather than the data model $p(\mathbf{y}|\boldsymbol{w}, \boldsymbol{\beta}, \tau^2)$ in terms of computation, that is because covariance matrix of the process model, typically, is dense while it is just a diagonal matrix for the data model. Therefore, we turn to study $\boldsymbol{w}$ in the process model.

As we all know, the best linear unbiased prediction (BLUP) $\hat{w}(s_0)$ at an unobserved location s_0 can be obtained by the **Kriging equation**:

$$\hat{w}(s_0) = \mathbf{c}'_{s_0}\mathbf{C}^{-1}\boldsymbol{w} \tag{4}$$

where $\mathbf{c}_{s_0} = (\mathbf{C}(s_0, s_1), \ldots, \mathbf{C}(s_0, s_n))'$.

However, with large or massive data (e.g., $n = 10^5$ or more), direct implementation of these statistical process, including parameter etimation by (2) or (3) and interpolation by (4), becomes computationally prohibitive, since evaluation of the likelihood of the process model in (3) (or the likelihood in (2)) or solution of the Kriging equation (4) involve the Cholesky factorization of an $n \times n$ covariance matrix for data of size n, which requires $O(n^3)$ operations and $O(n^2)$ memory in general (see, e.g., **?**, **?**, **?**, **?**).

2 SEVERAL APPROACHES TO OVERCOME THE COMPUTATIONAL PROBLEM

The approximation of the likelihood in either the spatial or spectral domain and the lower-dimensional representation are fundamental solutions to overcome computational obstacles. These solutions focus on model development, the design of efficient and parallel algorithms, and the improvement and efficient use of modern computing platforms (e.g., Using TensorFlow and GPU).

Developed models, official websites of scholars:

1) Conditional distributions: Michael L. Stein (The restricted likelihood, 2004);

2) Sparse covariance by tapering method: Reinhard Furrer(2006); Douglas Nychka (bias correction, 2008); Michael L. Stein (Statistical Properties, 2013);

3) Low-Rank methods: Noel Cressie (FRK, 2008), Sudipto Banerjee (Predictive process, 2008); Matthias Katfuss (multiresolution approximation where the basis functions are automatically chosen, 2017);

4) Sparse precision: Douglas Nychka (lattice kriging, 2015) by basis-function; Finn Lindgren (GRMF Approximations, 2011) by SPDE; Abhi Datta (NNGP, 2016a, 2016b, 2016c, 2019) by conditional distributions;

5) Spectral method: Montserrat Fuentes and Joe Guinness (Circulant embedding, 2017 and Periodic Embeddings, 2007, 2019).

6) Discrete process convolutions. Dave Higdon (2002); Bruno Sansó (2009); Francky Fouedjio (2016).

7) Local approximate Gaussian processes: Robert B. Gramacy, [**?**];

8) Vecchia's approximation: **?**, **?** and **?**.

Developed algorithms and platforms, official websites of scholars:

1) INLA: Haavard Rue.

2) Parallel algorithm: Matthias Katfuss.

3) TensorFlow: Andrew Zammit-Mangion.

Table 1: Application situations for different models

Model	**non-stationary**	**space-time**	**multivariate data**
FRK	√	√	√
Predictive process	√	√	√
MRA	√	√	
Tapering		√	√
Spatial partitioning			
LatticeKrig	√	×	×
SPDE	√	√	√
NNGP	√	√	√
Whittle's approximation	×	×	×

The websites of other scholars: 1) Christopher K. Wikle, 2) Bradley P. Carlin, 4) David Bolin, 5) Jonathan R. Stroud, 6) YongTao Guan, 7) Alan E. Gelfand, 8) Peter J. Diggle, 9) Huiyan Sang, 10) Andrw Finley, 11) Matthew J. Heaton, 12) Ying Sun, 13) Furong Sun, 14) Jonathan R. Bradley.

3 LIKELIHOOD APPROXIMATIONS IN THE SPECTRAL DOMAIN

3.1 Whittle's likelihood

The spectral methods are computationally efficient by avoiding the calculation of determinants and can be easily adapted to model nonstationary processes as a mixture of independent stationary processes.

?, [?] and [?] propose the use of small domain expansions and imputing data in a periodic fashion on the expanded lattice and presented a version of Whittle's approximation to the Gaussian negative log-likelihood by introducing a lattice process which can be used to deal with irregularly spaced data. Additional computational savings were obtained by truncating the spectral representation of the lattice process. The calculation requires $O(m \log_2 m + n)$ operations where m is lattice size.

Comment: The flexibility of the spectral methods mentioned above are disputable for prediction problem respecting out of sample data, since the expanded lattice is pre-difined.

4 LIKELIHOOD APPROXIMATIONS IN THE SPATIAL DOMAIN

4.1 Low rank methods

Comment: The reduced rank based methods usually fail to accurately capture the local, small scale dependence structure (see [?], [?] and [?]).

4.1.1 *Fixed rank Kriging*

FRK ([**?**], [**?**]) aims to approximate the spatial process $w(s)$ in (1) by a linear combination of K ($\ll n$) **multi-resolution bisquare** basis functions:

$$w(s) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} h_{rk}(s) w_{rk}^{\star} \tag{5}$$

where $K = \sum_{r=1}^{R} K_r$ and the coefficients $w^{\star}$ is an K-dimensional Gaussian vector with mean zero and exponential covariance defined by some small basic areal units. which ensures that all estimation and prediction equations only contain inverses of matrices of size $K \times K$.

Comment: Since covariance matrix for every resolutions is dense, that implies FRK can be challenging for large K_r.

4.1.2 *Gaussian predictive processes (GPP)*

With regard to the challenge of computational cost on covariance matrices, **?** proposed a class of models based on the idea of a spatial predictive process which is motivated by Kriging equation in (4). The predictive process projects the original process, $w(s)$ in (1), onto a subspace generated by realizations of the original process at a specified set of locations (or knots), e.g., $s_1^{\star}, \cdots, s_K^{\star}, K \ll n$. The approach is in the same spirit as process modeling approaches using basis functions and kernel convolutions, that is, specifications which attempt to facilitate computations through lower dimensional process representations, i.e.

$$w(s) = \mathbf{c}_s' \boldsymbol{\Sigma}_{w^{\star}}^{-1} w_{rk}^{\star} \tag{6}$$

where $\mathbf{c}_s = \left(\mathbf{C}(s, s_1^{\star}), \ldots, \mathbf{C}(s, s_K^{\star})\right)'$, and then $\mathbf{c}_s' \boldsymbol{\Sigma}_{w^{\star}}^{-1}$ plays the same role as h in (5), and thus can be regarded as some basis functions.

Comment:

1) One advertised advantage of using the GPP approach as opposed to FRK or LatticeKrig is that the GPP basis functions are completely determined by the choice of covariance function $C(\cdot, \cdot)$, and note that the subsequent Multiresolution approximations is in step with GPP in this regard.

2) At the same time, however, when $C(\cdot, \cdot)$ is governed by unknown parameters (which is nearly always the case) the GPP basis functions need to be calculated iteratively rather than once as in FRK or LatticeKrig which will subsequently increase computation time.

4.1.3 *Multi-resolution approximations*

In contrast to FRK or LatticeKrig, the multi-resolution approximation (MRA) basis functions and the prior distribution of the corresponding weights $w_r^{\star}$ are chosen using the predictive-process approach to automatically adapt to any given covariance function $\mathbf{C}_r$, and so the MRA can adjust flexibly to a desired spatial smoothness and dependence structure. The MRA allows the number of basis functions to be approximately the same as the data by the two ways to do. The one is by increasing sparsity of the covariance matrices of the corresponding weights (achieved by tapering method, see [?]), the other is by recursively partitioning the spatial domain (see [?]).

Comment: Two obvious advantages of MRA are the ability to compute in parallel and capture spatial structure from very fine to very large scales.

4.2 Sparse covariance methods

4.2.1 *Tapering*

Including tapering for estimation (e.g., approximation log-likelihood function by a tapered covariance. See ?) and tapering for Kriging (or for interpolation or for prediction, e.g., approximation Kriging equation by replacing the original covariance by a tapered version. See ? and some R packages: spam, 2010; KriSp, 2006; fields, 2017.

Comment: The covariance tapering has shown great computational gains, but it also has its own drawbacks.

1) The covariance tapering may not be effective in accounting for **spatial dependence with long range**.

2) The accuracy of the tapering approximation for **nonstationary problems** remains an open question, and the application of tapering techniques to **multivariate random fields** (e.g., [?]) remains to be explored due to the lack of flexible compactly supported cross-covariance functions, see ?.

4.2.2 *Spatial partitioning*

By dividing region D into m disjoint subregions $(d = 1, 2, \cdots, m)$, and then the modeling approach based on spatial partitioning is to again assume the model in (7) but take on the assumption of independence between observations across subregions, and then

$$y_d = \beta X_d + H_d w + \xi_d + \epsilon_d \tag{7}$$

where H_d is a matrix of spatial basis function for subregions d.

Comment:

1) Notice that, in (7) each subregion shares common β and w parameters which allows smoothing across subregions in spite of the independence assumption. Further, the assumption of independence across subregions effectively creates a block-diagonal structure for Σ and allows the likelihood to be computed in parallel (with one node per subregion), thereby facilitating computation.

2) The key to implementing the spatial partitioning approach is the choice of partition and the literature is replete with various options. A priori methods to define the spatial partitioning include partitioning the region into equal areas ([?]), partitioning based on centroid clustering ([?]), hierarchical clustering based on spatial gradients ([?]). Alternatively, model-based approaches to s-patial partitioning include treed regression([?]) and mixture modeling (?), but these approaches typically require more computation.

4.3 Sparse precision methods

The sparse precision methods focus on basis function and conditional likelihood.

4.3.1 *LatticeKrig*

LatticeKrig (LK, [?]) uses nearly the same setup as is employed by FRK. Here is just a list of the differences:

1) FRK uses families of **bisquare** basis functions allowing different resolutions that are organized on irregular basic areal units (BAUs), and LK uses families of **radial** basis functions that are organized on regular grids of increasing resolution.

2) Covariance matrix, $\mathbf{\Sigma}_r$, of the coefficients $w_r^\star$ in (5) is constrained by a exponential function, the exponential function is defined on the centroids of the BAUs. However, precision matrix, $\mathbf{Q}_r$, of the coefficients $w_r^\star$ in LK is constrained to a spatial autoregressive (SAR, defined on the regular lattice) model and forced to admit some sparse structures by following the ideas of **?**.

3) Furthermore, because $\mathbf{Q}_r$ is sparse, **LK can set** K **to be very large (e.g.,** $K > n$**) without much additional computational cost**. However, the FRK does not achieve this effect.

4.3.2 *Gaussian Markov random field approximations by SPDE*

? rigorously suggest to approximate Gaussian random fields through Markov fields with a huge increase in speed for the simulations and demonstrated empirically that GMRFs could closely approximate some commonly used covariance functions in geostatistics, and whereafter **?** show the construction of the corresponding GMRFs can be used to represent the Matérn field on a triangulated lattice (i.e., the finite element construction).

Similar to FRK or LatticeKrig, the spatial process $w(s)$ in model (1) is represented by the basis function, this method use piecewise linear functions on a triangulation of the domain as basis functions $h(\cdot)$ in (5), and then this yields sparse matrices $\mathbf{C}$ and $\mathbf{G}$ such that the appropriate precision matrix for the weights $\boldsymbol{w}^\star$ is given by (see [?])

$$\mathbf{\Omega} = \tau^2(\kappa^4\mathbf{C} + 2\kappa^2\mathbf{G} + \mathbf{G}\mathbf{C}^{-1}\mathbf{G}).$$

One particular advantage of this approach is that assigning the Gaussian distribution $\boldsymbol{w}^\star \sim \mathtt{N}(0, \mathbf{\Omega}^{-1})$ can generate continuously defined functions $\mathcal{X}(s) = \sum_{i=1}^m h_i(s) w_k^\star$ that are approximative solutions to the following SPDE (in a stochastically weak sense):

$$(\kappa^2 - \Delta)^{\alpha/2}\left(\tau(\mathcal{X}(s))\right) = \mathcal{W}, \quad s \in \mathbf{D},$$

where Δ is the Laplacian, κ is the spatial scale parameter, α controls the smoothness of the realisations, τ controls the variance. The right-hand side of the equation, $\mathcal{W}(s)$, is a Gaussian spatial white noise process. Note that the stationary solutions on $\mathcal{R}^d$ is the stationary Gaussian field $\mathcal{X}(s)$ with Matern covariances function given by

$$\operatorname{Cov}\left(\mathcal{X}\left(s_i\right), \mathcal{X}\left(s_j\right)\right) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}}\left(\kappa\left\|s_i - s_j\right\|\right)^{\lambda} K_\lambda\left(\kappa\left\|s_i - s_j\right\|\right) \tag{8}$$

The drawback of ? is that we can only find the explicit form of GMRFs for those Gaussian random fields that have a Matérn covariance structure at certain integer smoothnesses. Subsequently, ? extend this method to fractional order by the rational approximation.

Comment:

1) Unlike FRK or LatticeKrig where we need pre-define basis functions, the basis functions $h(\cdot)$ of GMRFs approximation are **automatically control or chosen by triangulation of the domain**, and thus constrained to be **sparse** generally.

2) The numerical factorization of the precision matrix using INLA can be done at a typical cost of $O(n^{3/2})$ for two-dimensional GMRFs.

3) SPDE can be extended to model Matérn covariances on the sphere, nonstationary locally isotropic Gaussian random fields (e.g., by using spatially varying parameters), Gaussian random fields with oscillating correlation functions, and non-isotropic fields.

4) The SPDE approach also facilitates extensions to multivariate random fields, see ?, ?, ?.

4.3.3 *The nearest neighbor Gaussian process*

Similar to the pre-defined knots for Gaussian predictive processes, the starting point of nearest neighbor Gaussian process (NNGP) approach is to choose a fixed collection (e.g., S) of distinct locations in D, where S need not coincide with or be apart of the observed locations, so its size k need not equal the size of the dataset n (or even larger than the size n). The set S is called as reference set by ?.

A directed acyclic graph is defined on the reference set S, and then the joint distribution of spatial process $w(s)$ from the reference set is represented by the product of conditional densities which is motivated by Vecchia's approximation [**?**] ideas, where a careful choice of suitable conditional sets is required, and this conditional set is constrained to be the m-nearest neighbors by the NNGP approach, thereby facilitating computation in estimation and prediction problem (for more details on those problems, see Appendix of ? and [?] and ?).

Based on these results above, a NNGP can be well-defined from a parent Gaussian process $GP(0, \mathbf{C}(\cdot; \theta))$.

1) Total operations is $O((n+k)m^3)$ where $m(\approx 20)$ is the size of conditioning set or neighbor set, and several processes can run in parallel (e.g, computations of weights in Kriging equation (4)).

2) The other major advantage is that the precision matrix of the NNGP is sparse with at most $km(m+1)/2$ nonzero entries.

3) NNGP can also be extended to large spatio-temporal data (?) and non-stationary process(?), [?].

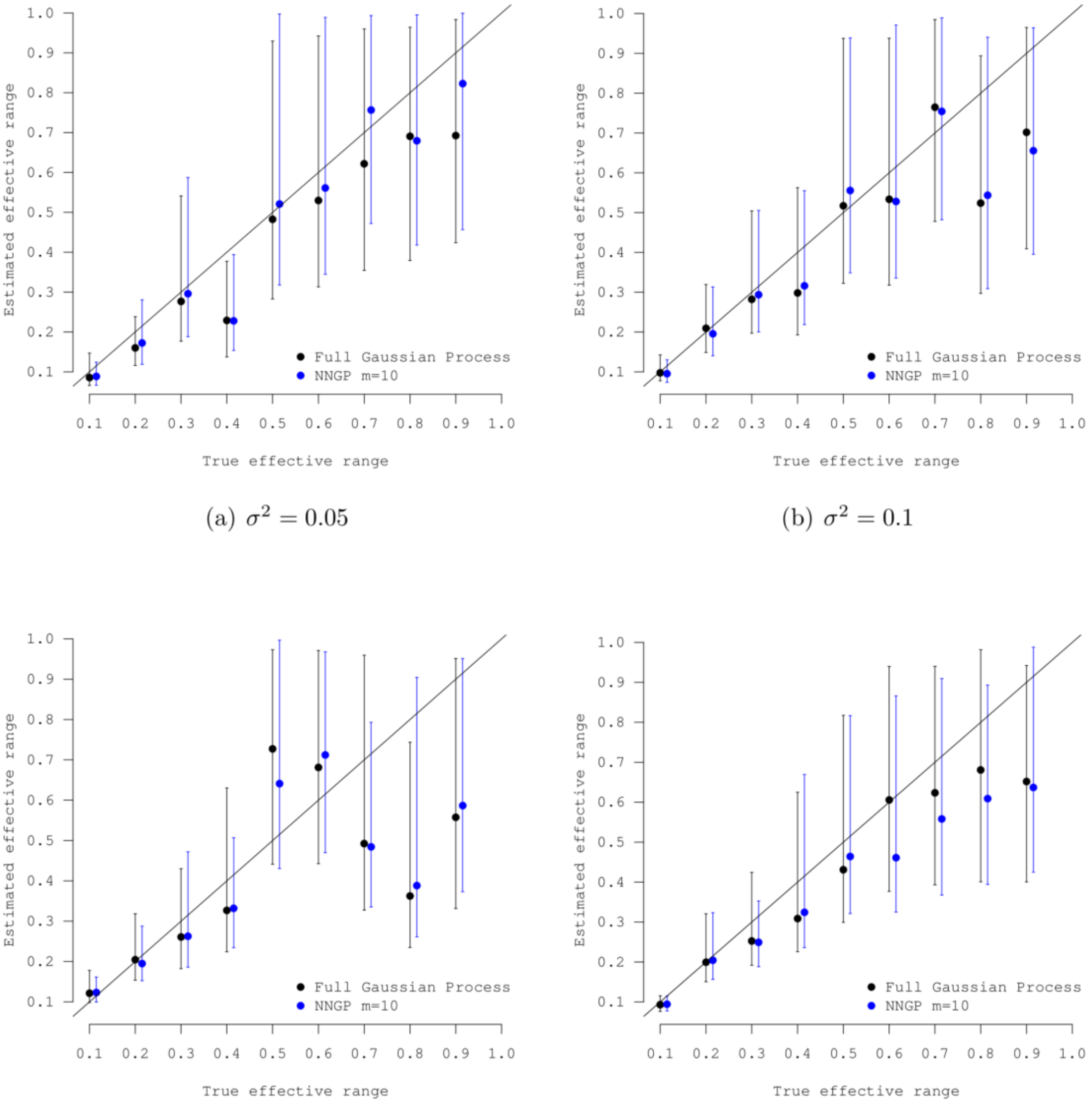
5 TWO EXPLORATORY LIMITATIONS OF THE NNGP APPROACH

5.1 Exploratory analysis

1) Problem 1: Since the joint distribution of spatial process $w(s)$ from the reference set is represented by using the product of conditional densities and those conditional sets are constrained to be the m-nearest neighbors, so m-nearest neighbors may fail to capture all the information about the covariance parameters when there is a true **large scale dependence** in the dataset.

2) Problem 2: From an application point of view, the reference set S cannot be infinite because we need to solve neighbor set for every location, so the performance of the NNGP approximation depends on the size of the spatial dependence range relative to the spacing of the reference set, that maybe lead to the quality of the NNGP approximation gets worse when **the spatial dependence range gets shorter**.

With regard to Problem 1, some simulation results were given by **?** where they generated datasets of size 2500 in a unit square domain and considered estimation cases of different range parameters. Those results can be found in Figure 1. Figure 1 suggests that the NNGP model deliver inference similar to that of a full GP even for slow decaying covariance functions. However, Figure 1 shows also that the NNGP obviously tends to underestimate the estimate of range parameter ϕ with the increase of spatial range ϕ.



(a) $\sigma^2 = 0.05$

(b) $\sigma^2 = 0.1$

(c) $\sigma^2 = 0.2$

(d) $\sigma^2 = 0.5$

Figure 1: Simulation results of ?.

We considered model (1) with $\beta X = 0.1\mathbf{I}, \sigma^2 = 5, \tau^2 = 1$, and the Matern correlation function for the spatial random effects w, with a constant smoothness parameter $\nu = 0.5$ and varying spatial range parameters.

Three distance criteria:

$$D_{KL}(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \frac{1}{2}\left[\text{trace}\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\right) - p + \log|\hat{\boldsymbol{\Sigma}}| - \log|\boldsymbol{\Sigma}|\right] \tag{9}$$

$$D_B(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \frac{1}{2}\log|\tilde{\boldsymbol{\Sigma}}| - \frac{1}{4}\log|\hat{\boldsymbol{\Sigma}}| - \frac{1}{4}\log|\boldsymbol{\Sigma}|, \quad \tilde{\boldsymbol{\Sigma}} = [\boldsymbol{\Sigma} + \hat{\boldsymbol{\Sigma}}]/2 \tag{10}$$

$$D_F(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \sqrt{\text{trace}\left[(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})'\right]} \tag{11}$$

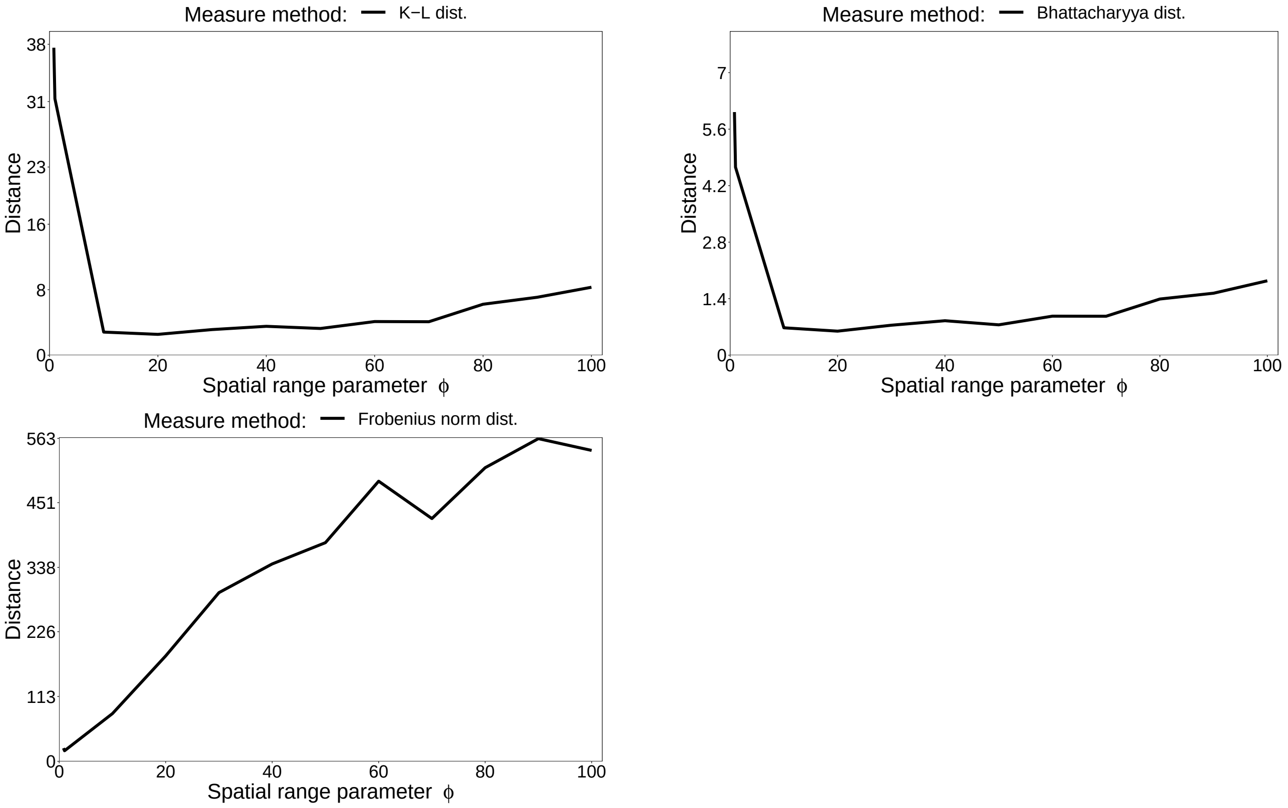


Figure 2: The K-L distance, the Bhattacharyya distance and the Frobenius norm from the estimated covariance matrix to the true covariance matrix for the Matérn family with smoothness parameters $\nu = \frac{1}{2}$ and different spatial range parameters ϕ. And this true covariance matrix at 500 random locations from $[0, 100] \times [0, 100]$ was estimated by the NNGP model with $m = 20$ for 50 simulations.

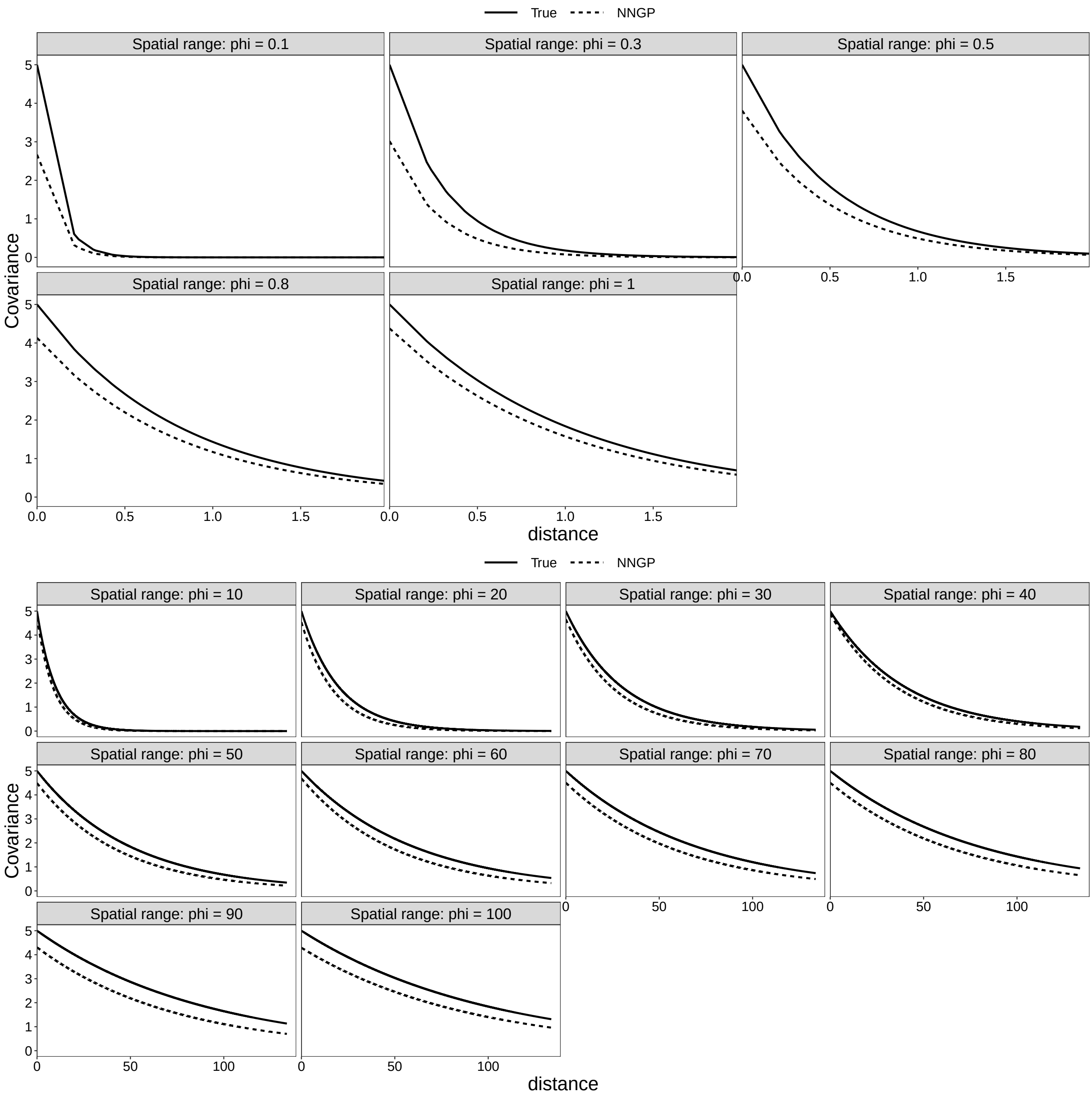


Figure 3: Estimation results of covariance function corresponding to Figure 2.

Figure 2 and 3 suggest that **the large-scale spatial dependencies for a larger domain can not be well-captured** by the NNGP, and at the same time, the top two panel of Figure 2 show that as with predictive process model, the NNGP may also **fail to capture some local information**, that may be related to the selection of reference set.

5.2 Improvement project

1) Ideas for Problem 1: Combing a NNGP $w(s)$ used to capture the local, small scale dependence structure with its spatial smooth version $f(s)$ used to capture the large scale dependence structure, for example,

$$y(s) = w(s) + f(s) + \epsilon(s). \tag{12}$$

Denote the reference set as $\mathbf{S} = \{s_j^\star : j = 1, \cdots, K\}$ determined through the grid of the domain and $w(s_j^\star)$ as $w_j^\star$. Assume that $\boldsymbol{w}^\star = \left(w_1^\star, \cdots, w_K^\star\right)' \sim \text{NNGP}\left(0, \tilde{\mathbf{C}}(\cdot;\theta)\right)$ and $h_k, k = 1, \cdots, K$ is a sequence of fixed basis functions, and the $w(s)$ and $f(s)$ are then constructed as follows:

$$w(s) = \mathbf{d}'(s)\boldsymbol{w}^\star \tag{13}$$

and

$$f(s) = \sum_{k=1}^{K} h_k(s) w_k^\star = \mathbf{h}'(s)\boldsymbol{w}^\star \tag{14}$$

where

$$\mathbf{d}'(s) = \begin{cases} \left(\mathbb{1}\left(s=1\right), \cdots, \mathbb{1}\left(s=k\right), \cdots, \mathbb{1}\left(s=K\right)\right), \text{ if } s \in \mathbf{S} \\ \mathcal{G}\left(s; \mathbf{c}'_{s,N(s)} \mathbf{C}^{-1}_{N(s)}\right), \text{ otherwise} \end{cases} \tag{15}$$

and the function of $\mathcal{G}(s; \mathbf{a}_s)$ is to expand the vector $\mathbf{a}_s$ from the low-dimension to high-dimension (e.g., from m to K) with according to the index of neighbor set for the location s, and then to derive a vector $\mathbf{b}_s$ where $\mathbf{b}_s(j) = \mathbf{a}_s(j)$ if j is the neighbor of location s corresponding to $w_j^\star$, 0 otherwise.

In terms of matrix-vector representation, obtaining

$$\boldsymbol{y} = \left(\mathbf{D} + \mathbf{H}\right)\boldsymbol{w}^\star + \epsilon(s) \tag{16}$$

where $\mathbf{D} = \left(\mathbf{d}'(1), \ldots, \mathbf{d}'(n)\right)'$ and $\mathbf{H} = \left(\mathbf{h}'(1), \ldots, \mathbf{h}'(n)\right)'$. And according to the assumptions previously mentioned in the model, we know that

$$\boldsymbol{y} \sim \text{GP}\left(0, \boldsymbol{\Sigma}(\cdot;\theta)\right) \tag{17}$$

where $\boldsymbol{\Sigma}(\cdot;\theta) = \left(\mathbf{D}+\mathbf{H}\right)\tilde{\mathbf{C}}(\cdot;\theta)\left(\mathbf{D}+\mathbf{H}\right)' + \tau^2\mathbf{I}$.

Let $\mathbf{U} = \mathbf{D} + \mathbf{H}$, and by applying the Sherman-Woodbury-Morrison formula for inverse matrices (as $n \gg K$), we have

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\tau^2}\mathbf{I} - \frac{1}{\tau^2}\mathbf{U}\left(\tau^2\tilde{\mathbf{C}}^{-1} + \mathbf{U}'\mathbf{U}\right)^{-1}\mathbf{U}'. \tag{18}$$

On the other hand, the determinant can be computed using

$$\det\{\boldsymbol{\Sigma}\} = \tau^{2(n-K)}\frac{\det\{\tau^2\tilde{\mathbf{C}}^{-1} + \mathbf{U}'\mathbf{U}\}}{\det\{\tilde{\mathbf{C}}^{-1}\}}. \tag{19}$$

Recall that $\tilde{\mathbf{C}}^{-1}$ is a sparse precision matrix with at most $Km(m+1)/2$ nonzero entries, thereby, (18) and (19) can facilitate computations.

2) Ideas for Problem 2: Similar to ?,

$$y(s) = w_l(s) + w_s(s) + \epsilon(s) \tag{20}$$

where w_l is a NNGP and w_s is a residual process with covariance function constrained by tapered function.

3) A more thorough solution for both problems: Similar to LatticeKrig, to build a multi-resolution NNGP approximation, i.e.,

$$y(s) = \sum_{r=1}^{R} \sum_{k=1}^{K_r} h_{ri}(s) w^{\star}_{rk} + \epsilon(s) \tag{21}$$

where $\boldsymbol{w}^{\star}_{r} = \left(w^{\star}_{r,1}, \cdots, w^{\star}_{r,K_r}\right)' \sim \mathrm{NNGP}\left(0, \tilde{C}_r(\cdot;\theta_r)\right)$.

some other websites: Probability, Mathematical Statistics, Stochastic Processes