Large-Scale Low-Rank Gaussian Process Prediction with Support Points

Yan Song[1,2], Wenlin Dai[1*]

[1]Institute of Statistics and Big Data, Renmin University of China, China

and

Marc G. Genton[2]

[2]Statistics Program, King Abdullah University of Science and Technology, Saudi Arabia

July 29, 2024

Abstract

Low-rank approximation is a popular strategy to tackle the "big n problem" associated with large-scale Gaussian process regressions. Basis functions for developing low-rank structures are crucial and should be carefully specified. Predictive processes simplify the problem by inducing basis functions with a covariance function and a set of knots. The existing literature suggests certain practical implementations of knot selection and covariance estimation; however, theoretical foundations explaining the influence of these two factors on predictive processes are lacking. In this paper, the asymptotic prediction performance of the predictive process and Gaussian process predictions are derived and the impacts of the selected knots and estimated covariance are studied. The use of support points as knots, which best represent data locations, is advocated. Extensive simulation studies demonstrate the superiority of support points and verify our theoretical results. Real data of precipitation and ozone are used as examples, and the efficiency of our method over other widely used low-rank approximation methods is verified.

Keywords: Convergence rate; Kernel ridge regression; Kriging; Nyström approximation; Predictive process; Spatial statistics.

1 Introduction

Gaussian processes (GPs; Rasmussen and Williams, 2006) are extensively used for solving regression problems in many fields such as spatial statistics (Stein, 1999; Gelfand et al.,

---

*Corresponding author; email: wenlin.dai@ruc.edu.cn

2010; Cressie, 2015), computer experiments (Sacks et al., 1989; Santner et al., 2003), and machine learning (Rasmussen and Williams, 2006; Liu et al., 2020). The underlying regression function is assumed to be a realization of a GP, whose mean trend and covariance function must be evaluated. Using the conditional multivariate normal distribution, a GP provides each location with a closed-form prediction as well as a predictive interval.

Assume that a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ satisfies the GP model, i.e.,

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and follows a distribution $F_\mathbf{x}$. The $f(\mathbf{x}_i)$'s are the realization of a Gaussian process $Z(\cdot)$ on locations $\{\mathbf{x}_i\}_{i=1}^n$, and the nugget effects ϵ_i's are independent and follow a $\mathcal{N}(0, \tau^2)$ distribution. Under the GP regression, f is considered as a random function and is not distinguished from Z. Assuming $Z(\cdot) \sim \mathrm{GP}(0, c)$, where $c(\cdot, \cdot)$ is a positive definite covariance function. The zero mean assumption simplifies the presentation, as estimating and subtracting the mean is typically computationally feasible, and theoretical results for a general case can be obtained by detrending the process (Burt et al., 2019; Katzfuss et al., 2020). Then, the response vector $\mathbf{y} = (y_1, \ldots, y_n)^\top$ follows a multivariate normal distribution $\mathcal{N}_n(\mathbf{0}, \mathbf{C} + \tau^2\mathbf{I}_n)$, where $\mathbf{C} = \{c(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$. The covariance function c is usually unknown in practice and is represented by a specific parametric form $c(\cdot, \cdot; \boldsymbol{\theta})$, where the parameter vector $\boldsymbol{\theta}$ can be obtained by maximizing the likelihood function with respect to $\boldsymbol{\theta}$: $(2\pi)^{-n/2}|\mathbf{C}(\boldsymbol{\theta}) + \tau^2\mathbf{I}_n|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{y}^\top\{\mathbf{C}(\boldsymbol{\theta}) + \tau^2\mathbf{I}_n\}^{-1}\mathbf{y}\right]$. We avoid theoretical discussions on covariance estimation due to its complexity and methodological dependency, opting instead to demonstrate how an estimated covariance impacts predictions. With the conditional probability density of the multivariate normal distribution, the prediction based on c at any $\mathbf{x} \in \mathcal{X}$ is

$$\hat{f}_c(\mathbf{x}) = \mathbf{c}(\mathbf{x})^\top(\mathbf{C} + \tau^2\mathbf{I}_n)^{-1}\mathbf{y}, \tag{2}$$

where $\mathbf{c}(\mathbf{x}) = \{c(\mathbf{x}, \mathbf{x}_1), \ldots, c(\mathbf{x}, \mathbf{x}_n)\}^\top$. Fitting data of size n with a GP requires $O(n^3)$

operations and $O(n^2)$ memory, and these requirements may easily exhaust computational resources even with a moderately large n.

Many methods are available to tackle the challenge of handling massive datasets with GPs. We refer the readers to Heaton et al. (2019) and Liu et al. (2020) for their comprehensive reviews. The methods of likelihood approximation include the Vecchia approximation (Vecchia, 1988; Stein et al., 2004; Katzfuss et al., 2020; Katzfuss and Guinness, 2021), composite likelihood (Varin et al., 2011; Eidsvik et al., 2014), and hierarchical low-rank approximation (Huang and Sun, 2018). Covariance tapering multiplies the covariance function with a compactly supported one such that the tapered covariance matrix is sparse and can be solved easily (Furrer et al., 2006; Kaufman et al., 2008; Stein, 2013). Rue and Tjelmeland (2002), Rue and Held (2005), and Xu et al. (2015) discussed the methods of Markov random field approximation, which are primarily used for lattice data. Methods based on distributed computing (Paciorek et al., 2015; Deisenroth and Ng, 2015; Katzfuss and Hammerling, 2017; Abdulah et al., 2018a,b, 2019), resampling (Liang et al., 2013; Barbian and Assunção, 2017), and exact approaches using the *ExaGeoStat* software (Abdulah et al., 2018a) were recently proposed.

The low-rank approximation is another important approach for GPs with massive data. It approximates the original GP of mean zero with a low-rank one, which is represented by a linear combination of specified basis functions with random weights, e.g., the fixed-rank kriging (FRK, Cressie and Johannesson, 2008; Zammit-Mangion and Cressie, 2021), LatticeKrig (Nychka et al., 2015), and predictive processes (Banerjee et al., 2008; Finley et al., 2009). Combining low-rank approximations with the concept of covariance tapering can help develop multiple efficient methods (Sang and Huang, 2012; Katzfuss, 2017). In this context, the choice of the basis functions is important to the performance of these

methods. FRK and LatticeKrig suggest multiresolution bases to capture information at different scales. The construction of basis functions involves the delicate selections of multiple tuning parameters, such as the types of basis functions, levels of resolution, and information in the coarsest resolution (Zammit-Mangion and Cressie, 2021; Nychka et al., 2016). Tzeng and Huang (2018) proposed adaptive basis functions to avoid the manual allocation of scales.

Predictive processes simplify the aforementioned problem by developing basis functions with a covariance function and a set of specified knots. In this context, the covariance estimation and knot configuration are of particular interest. Specifically, Banerjee et al. (2008) selected a set of knots $D_k^* = \{\mathbf{x}_i^*\}_{i=1}^k$ well dispersed over $\mathcal{X}$, which may or may not belong to $\{\mathbf{x}_i\}_{i=1}^n$. Then, they projected the original process Z onto a linear space spanned by $\mathbf{z}^* = \{Z(\mathbf{x}_1^*), \ldots, Z(\mathbf{x}_k^*)\}^\top$ and induced the predictive process $\tilde{Z}(\mathbf{x}) = \mathbb{E}\{Z(\mathbf{x})|\mathbf{z}^*\} = \mathbf{c}^*(\mathbf{x})^\top \mathbf{C}^{*-1} \mathbf{z}^*$, where $\mathbf{c}^*(\mathbf{x}) = \{c(\mathbf{x}, \mathbf{x}_1^*), \ldots, c(\mathbf{x}, \mathbf{x}_k^*)\}^\top$ and $\mathbf{C}^* = \{\mathbf{c}^*(\mathbf{x}_1^*), \ldots, \mathbf{c}^*(\mathbf{x}_k^*)\}$. Correspondingly, the process $\tilde{Z} \sim \mathrm{GP}(0, \tilde{c})$, where $\tilde{c}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{c}^*(\mathbf{x}_i)^\top \mathbf{C}^{*-1} \mathbf{c}^*(\mathbf{x}_j)$. Replacing c in (2) with $\tilde{c}$ yields the low-rank prediction

$$\tilde{f}_c(\mathbf{x}) = \mathbf{c}^*(\mathbf{x})^\top \mathbf{C}^{*-1} \mathbf{C}_{nk}^{*\top} (\mathbf{C}_{nk}^* \mathbf{C}^{*-1} \mathbf{C}_{nk}^{*\top} + \tau^2 \mathbf{I}_n)^{-1} \mathbf{y}, \tag{3}$$

where $\mathbf{C}_{nk}^* = \{\mathbf{c}^*(\mathbf{x}_1), \ldots, \mathbf{c}^*(\mathbf{x}_n)\}^\top$. Furthermore, using the Sherman–Morrison–Woodbury formula (Henderson and Searle, 1981), we can rewrite (3) as follows:

$$\tilde{f}_c(\mathbf{x}) = \mathbf{c}^*(\mathbf{x})^\top \mathbf{C}^{*-1} \mathbf{C}_{nk}^{*\top} \{\tau^{-2} \mathbf{I}_n - \tau^{-2} \mathbf{C}_{nk}^* (\tau^2 \mathbf{C}^* + \mathbf{C}_{nk}^* \mathbf{C}_{nk}^{*\top})^{-1} \mathbf{C}_{nk}^{*\top}\} \mathbf{y},$$

which includes the inversion of only a $k \times k$ matrix $\tau^2 \mathbf{C}^* + \mathbf{C}_{nk}^* \mathbf{C}_{nk}^{*\top}$; hence, this step reduces the computational time to $O(nk^2)$. Finally, Banerjee et al. (2008) implemented the entire procedure of parameter estimation and prediction by adopting the Bayesian approach. Finley et al. (2009) proposed a modified predictive process to debias the overestimation of τ^2. Moreover, inspired by strategies in spatial designs, they developed an algorithm to

select those knots that induce a predictive process mimicking the parent process better. Predictive processes are also termed as sparse Gaussian processes in the machine learning literature, where knots D_k^* are termed as inducing points (Titsias, 2009; Hensman et al., 2013; Burt et al., 2019, 2020).

The idea of low-rank approximations is not limited to the GP model and has been applied to more problems, whereas the basis selection is still being discussed. The Nyström approximation shares some commonality with the low-rank approximation. It approximates a positive definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ by a low-rank one $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S})^- \mathbf{S}^\top \mathbf{A}$, where the "sketching matrix" $\mathbf{S} \in \mathbb{R}^{n \times m}$, with $m < n$, sketches the desirable information from $\mathbf{A}$ and $\mathbf{K}^-$ is the generalized inverse of a matrix $\mathbf{K}$ (Drineas et al., 2008; Mahoney, 2011; Gittens and Mahoney, 2013). This strategy has been used in the GP (Banerjee et al., 2013) and the kernel ridge regression (KRR) models (Rudi et al., 2015; Alaoui and Mahoney, 2015). The matrix $\mathbf{A}$ is prespecified and $\mathbf{S}$ determines the performance of the approximation.

The KRR is closely linked to GPs. In particular, $\hat{f}_c$ is a solution of a KRR obtained by considering f in (1) as a deterministic function:

$$\hat{f}_c = \arg\min_{\eta \in \mathcal{H}_c} \frac{1}{n} \sum_{i=1}^n \{y_i - \eta(\mathbf{x}_i)\}^2 + \frac{\tau^2}{2n} J_c(\eta), \tag{4}$$

where $\mathcal{H}_c = \{\eta = \sum_{r=1}^\infty \alpha_r c(\cdot, \mathbf{x}_r) : \alpha_r \in \mathbb{R}, \mathbf{x}_r \in \mathcal{X}$ such that $\|\eta\|_{\mathcal{H}_c}^2 = \sum_{r,l=1}^\infty \alpha_r \alpha_l c(\mathbf{x}_r, \mathbf{x}_l) < \infty\}$ is the reproducing kernel Hilbert space induced by c with norm $J_c(\eta) = \|\eta\|_{\mathcal{H}_c}^2$. Wahba's representer theorem (Wahba, 1990) ensures that $\hat{f}_c$ lies in a finite subspace of $\mathcal{H}_c$ spanned by $\{c(\cdot, \mathbf{x}_i)\}_{i=1}^n$, i.e., $\mathcal{H}_{n,c}$. The low-rank prediction $\tilde{f}_c$ is the solution of (4) over $\mathcal{H}_{k,c}^* = \text{span}\{c(\cdot, \mathbf{x}_i^*)\}_{i=1}^k$, induced by knots D_k^*. Using a prespecified kernel function c, the performance of the low-rank prediction $\tilde{f}_c$ depends on the choice of knots.

There is a considerable body of literature discussing the selection of knots. Banerjee et al. (2008) and Finley et al. (2009) drew inspiration from spatial designs, such as

space-filling, to select knots. Titsias (2009) proposed a variational formulation that simultaneously estimates the covariance and selects knots by minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the exact posterior distribution of the latent function values. However, the performance of the predictive process with these knots was only numerically tested without theoretical justification. The influence of knots on the asymptotic properties of $\tilde{f}_{\hat{c}}$ should be investigated. Rudi et al. (2015) and Alaoui and Mahoney (2015) chose knots that improve the prediction accuracy of $\tilde{f}_c$. They derived the optimal probabilities that minimize the learning bounds, which are based on a concept called ridge leverage scores (RLS). However, they used a random sampling strategy to obtain knots, which brings randomness and larger uncertainty. Burt et al. (2019) and Burt et al. (2020) investigated how k needs to grow with n so that the KL divergence between the approximate model and the exact posterior can be arbitrarily small. However, they only focused on knots of size k sampled from a fixed-size determinantal point process (k-DPP) and using RLS without considering the influence of a mis-specified covariance.

In this study, we investigate the asymptotic performance of the predictive process and the original process upon linking GP with KRR. Given the true covariance function c, we demonstrate that the predictive process may achieve the same convergence rate as the original GP. The convergence rate is determined by a parameter, γ, which is proposed as a measure of smoothness for the process and for the covariance function. Consequently, we derive a knot-selection criterion, i.e., support points generated by the data locations, which are space-filling, and the best approximation of the locations with respect to the energy distance. Moreover, we demonstrate how the estimated covariance function, $\hat{c}$, affects the convergence rates of the two processes through the smoothness parameter $\hat{\gamma}$ for $\hat{c}$. As a byproduct, the adjusted order of τ^2 is obtained to improve the convergence rates of $\tilde{f}_{\hat{c}}$ and

$\hat{f}_{\hat{c}}$. Extensive numerical studies are conducted to demonstrate the superior performance of the support points and validate our theoretical results. We claim that the convergence rate of $\tilde{f}_{\hat{c}}$ can be identical to that of $\hat{f}_{\hat{c}}$ if D_k^* is close enough to $\{\mathbf{x}_i\}_{i=1}^n$. Thus, the efficiency of the low-rank approximation is validated.

The remainder of this paper is organized as follows. In Section 2, we provide theoretical results and describe the influence of knots and estimated covariance functions on the prediction performance of a predictive process. In Section 3, we present results of our numerical studies to demonstrate our approach and confirm the theoretical results. In Section 4, we describe the application of various low-rank approximations to two real data examples and compare their results. In Section 5, we present the conclusions and discuss future work. The proofs of the theoretical results are provided in the Supplementary Materials.

2 Convergence Rates of Low-Rank Approximations with Support Points

2.1 Known covariance function

Let D_k^* be a set of representative points of $\{\mathbf{x}_i\}_{i=1}^n$ used to develop a solution space for searching the low-rank approximation, hereinafter referred to as *rep-points*. With certain regularity conditions and a known covariance function c, the preferred configuration and size of D_k^* for obtaining $\tilde{f}_c$ with the asymptotically optimal prediction performance are provided in this section.

The selection of D_k^* has been extensively discussed. Banerjee et al. (2008) tested the performances of three strategies, including the use of grids, lattice plus close pairs, and lattice plus infill design, for both parameter estimation and prediction. The latter two

options were inspired from the concepts of spatial design (Diggle and Lophaven, 2006). Compared to the configuration, the size of D_k^* has greater influence on the parameter estimation. Finley et al. (2009) were motivated by spatial design concepts such as the space-filling design (Nychka and Saltzman, 1998) and other optimal designs for various targets (Zhu and Stein, 2005; Xia et al., 2006). They developed an algorithm for sequentially selecting D_k^* such that the predictive process can approximate the parent process better. Given a kernel (covariance) function, Rudi et al. (2015) and Alaoui and Mahoney (2015) proposed randomly selecting subsamples from the full data as per the RLS. Due to the inherent randomness, their D_k^* did not exhibit a space-filling property or a preference for certain points. In general, rep-points that are well dispersed over the region (Heaton et al., 2019) are preferable. Theoretically, we will show herein that the distribution of D_k^* should approximate $F_{\mathbf{x}}$ as closely as possible.

Before presenting the theoretical results, we introduce the necessary notations and conditions. For any $\eta, \xi \in \mathcal{L}_2(\mathcal{X})$, where $\mathcal{L}_2(\mathcal{X})$ is the set of square integrable functions on $\mathcal{X}$, let $\bar{V}(\eta, \xi) = \int_{\mathcal{X}} \eta(\mathbf{x})\xi(\mathbf{x})\mathbf{d}F_{\mathbf{x}}(\mathbf{x})$. We assess the prediction performance of $\tilde{f}_c$ over the whole region $\mathcal{X}$ with $V(\tilde{f}_c - f) \doteq \bar{V}(\tilde{f}_c - f, \tilde{f}_c - f) = \int_{\mathcal{X}}\{\tilde{f}_c(\mathbf{x}) - f(\mathbf{x})\}^2\mathbf{d}F_{\mathbf{x}}(\mathbf{x})$, which is commonly employed in regression literature (Ma et al., 2015; Rudi et al., 2015), acknowledging the impact of $F_{\mathbf{x}}$ on the prediction. Assuming $F_{\mathbf{x}}$ is a finite Borel measure on $\mathcal{X}$ and the covariance c is continuous, Mercer's theorem (Berlinet and Thomas-Agnan, 2004) suggests that $c(\mathbf{x}_i, \mathbf{x}_j)$ has an eigen expansion: $c(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{\infty} \rho_r^{-1}\phi_r(\mathbf{x}_i)\phi_r(\mathbf{x}_j)$ for $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$, where the positive eigenvalues are in the decreasing order of $\rho_1^{-1} \geq \rho_2^{-1} \geq \cdots$, the eigen functions ϕ_r satisfy $\bar{V}(\phi_r, \phi_s) = \delta_{rs}$ and $\int_{\mathcal{X}} c(\mathbf{x}', \mathbf{x})\phi_r(\mathbf{x}')\mathbf{d}F_{\mathbf{x}}(\mathbf{x}') = \rho_r^{-1}\phi_r(\mathbf{x})$ with the Kronecker δ. Then, any $\eta \in \mathcal{H}_c$ can be expanded as $\eta(\mathbf{x}) = \sum_{r=1}^{\infty} \eta_r\phi_r(\mathbf{x})$ with $\eta_r = \bar{V}(\eta, \phi_r)$ and the norm $J_c(\eta) = \sum_{r=1}^{\infty} \rho_r\eta_r^2 < \infty$. Three regularity conditions are

required to analyze $V(\tilde{f}_c - f)$: [R1] $\mathcal{X}$ is a compact set with a Lipschitz boundary and satisfies an interior cone condition (Kanagawa et al., 2018). [R2] The eigenvalue sequence $\{\rho_r^{-1}\}_r$ follows a power law decrease. That is, $\rho_r > \beta r^\gamma$ holds for certain $\gamma > 1$ and $\beta > 0$. [R3] For any r and s, $V(\phi_r\phi_s)$ is bounded and the $(d+1)/2$-th order differential of $\phi_r\phi_s$ belongs to $\mathcal{L}_2(\mathcal{X})$.

Condition [R1] indicates that $\mathcal{X}$ is compact and has no $\prec$-shaped region on its boundary (Kanagawa et al., 2018). In [R2], the parameter γ is important to indicate the smoothness of Z and c. The smoother the process, the larger the value of γ. A more intuitive demonstration of γ is reported in Section 3.1. [R3] shows two conditions about $\phi_r\phi_s$ for $\forall r, s$. The first one requires $V(\phi_r\phi_s)$ to be bounded, which virtually calls for a uniformly bounded fourth moment of $\phi_r(\mathbf{x})$ and is a widely-used regularity condition in the literature (Gu, 2013; Ma et al., 2015; Meng et al., 2020). The second one requires the $(d+1)/2$-th order differential of $\phi_r\phi_s$ to be square-integrable. It implies that $\phi_r\phi_s$ cannot be too rough. A similar condition has been used as a regularity condition in Meng et al. (2020). As Gu (2013) commented, the conditions are extremely difficult to verify because ϕ_r has no closed form. However, they appear mild because ϕ_r typically decrease in smoothness but do not grow in magnitude. In Section S3 of the Supplementary Materials, we give an example satisfying [R1]–[R3] to help with the understanding of these conditions. Moreover, this example shows how γ indicates the smoothness of c or Z.

Let the empirical distribution of D_k^* on $\mathcal{X}$ be denoted as F_k^* and let $\mathbf{x}, \mathbf{x}' \overset{i.i.d.}{\sim} F_{\mathbf{x}}$. The energy distance between F_k^* and $F_{\mathbf{x}}$ is $E(F_{\mathbf{x}}, F_k^*) = \frac{2}{k}\sum_{j=1}^k \mathbb{E}(\|\mathbf{x} - \mathbf{x}_j^*\|_2) - \mathbb{E}(\|\mathbf{x} - \mathbf{x}'\|_2) - \frac{1}{k^2}\sum_{i=1}^k\sum_{j=1}^k \mathbb{E}(\|\mathbf{x}_i^* - \mathbf{x}_j^*\|_2)$. This distance was proposed to test the goodness-of-fit of F_k^* to $F_{\mathbf{x}}$ for high-dimensional $\mathbf{x}$ (Rizzo, 2004; Székely and Rizzo, 2013; Mak and Joseph, 2018), and it is used for the first time to assess the performance and influence of rep-points in the

predictive process. Based on the link between GPs and KRRs, we examine the convergence rates of $\hat{f}_c$ and $\tilde{f}_c$. The following theorem indicates that $\tilde{f}_c$ based on rep-points can converge to f at the same rate as $\hat{f}_c$ based on full data; this implication confirms the efficiency of low-rank approximations with rep-points. Moreover, the smoother the process, the faster is the convergence.

Theorem 1. *(Convergence rate of $\tilde{f}_c$) Assume that the regularity conditions [R1]–[R3] hold, τ^2 is a constant, and $\gamma > 2$. As $n \to \infty$, the prediction in (2) has a convergence rate $V(\hat{f}_c - f) = O_p\left\{n^{-(1-1/\gamma)}\right\}$. For $D_k^* \subset \{\mathbf{x}_i\}_{i=1}^n$, if $n^{2/\gamma}E(F_\mathbf{x}, F_k^*) \to 0$ as $k, n \to \infty$, the prediction $\tilde{f}_c$ in (3) achieves an identical convergence rate of $\hat{f}_c$*

$$V(\tilde{f}_c - f) = O_p\left\{n^{-(1-1/\gamma)}\right\}. \tag{5}$$

Although Theorem 1 limits D_k^* to a subset of $\{\mathbf{x}_i\}_{i=1}^n$, it guides the choice of the rep-points. The convergence of $\tilde{f}_c$ to $\hat{f}_c$ versus k requires $E(F_\mathbf{x}, F_k^*) = o_p(n^{-2/\gamma})$. This requirement indicates that the empirical distribution of rep-points F_k^* should approach $F_\mathbf{x}$. Moreover, the faster the decrease in $E(F_\mathbf{x}, F_k^*)$ with an increase in k, the earlier is the convergence of $\tilde{f}_c$ to $\hat{f}_c$ achieved. The abovementioned discussion inspired us to select D_k^* that minimizes $E(F_\mathbf{x}, F_k^*)$; this choice agrees with the definition of support points (SPs; Mak and Joseph, 2018). The exact expression of $F_\mathbf{x}$ may be unknown in practice, and we can use the empirical distribution of $\{\mathbf{x}_i\}_{i=1}^n$, denoted as F_n, to replace $F_\mathbf{x}$. Using the R package *support* (Mak, 2021), we can easily obtain SPs. Note that SPs are generated rather than selected from $\{\mathbf{x}_i\}_{i=1}^n$, indicating that SPs may not belong to $\{\mathbf{x}_i\}_{i=1}^n$.

Figures 1 and S2 show SPs, k-DPPs (obtained by an exact algorithm based on Kulesza et al. (2012)), random subsamples (Rands), and grid points (Grids) for multiple location sets $\{\mathbf{x}_i\}_{i=1}^n$. A comparison of Fig. 1(a) with 1(b) and S2(a) demonstrates that for uniformly distributed $\{\mathbf{x}_i\}_{i=1}^n$, SPs are evenly spaced and cover the whole region well. A comparison

of Figs. 1(c) with 1(d) demonstrates that for nonuniformly distributed $\{\mathbf{x}_i\}_{i=1}^n$, more SPs gather at the data-dense region, and sufficient SPs capture the data-sparse region as well. SPs have a space-filling property and mimic the location set of the full data efficiently, thereby maximizing the representativeness of each point (Mak and Joseph, 2018). This is another reason for terming D_k^* as rep-points.

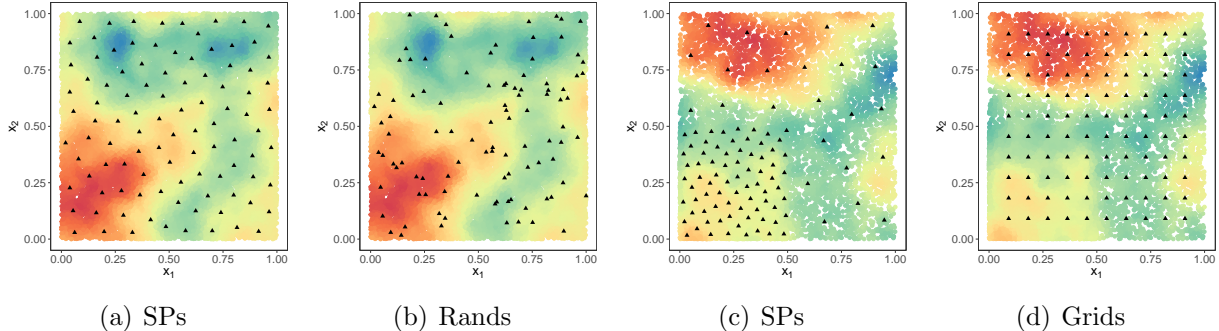| (a) SPs | (b) Rands | (c) SPs | (d) Grids |

Figure 1: Rep-points for spatial data with $n = 20000$ and multiple location sets. In (a) and (b), the location sets $\{\mathbf{x}_i\}_{i=1}^n$ follow a uniform distribution on $[0,1]^2$. In (c) and (d), 75% of $\{\mathbf{x}_i\}_{i=1}^n$ follow a uniform distribution on $[0,0.5]^2$, and 25% of $\{\mathbf{x}_i\}_{i=1}^n$ are uniformly distributed over the remaining region.

Finally, we discuss the order of k required for achieving the convergence of $\tilde{f}_c$. Mak and Joseph (2018) reported the relationship between the energy distance of SPs and k. For a bounded Borel set $\mathcal{X}$ with a nonempty interior, $E(F_{\mathbf{x}}, F_k^*) = O\{k^{-1}(\log k)^{-(1-\omega)/(2d)}\}$, where ω is in $(0, 1)$. We consider $E(F_{\mathbf{x}}, F_k^*) = O(k^{-1})$ as k^{-1} dominates $(\log k)^{-(1-\omega)/(2d)}$. Then, to satisfy the condition of convergence, i.e., $E(F_{\mathbf{x}}, F_k^*) = o_p(n^{-2/\gamma})$, we require that k should have a higher order than $O(n^{2/\gamma})$. We can see that the larger the γ, the smoother is the process, and the smaller is the order of k needed to ensure the convergence of $\tilde{f}_c$.

Note that Burt et al. (2019) and Burt et al. (2020) also derived the order of k needed to make the KL divergence between the approximated model and the exact posterior to be arbitrarily small when inducing points are obtained via an approximated k-DPP and RLS. Their results share some common features with ours, although obtained via different theoretical frameworks. First, both derivations depend on the eigen decomposition of

the covariance c with distribution $F_{\mathbf{x}}$. Therefore, the derived order of k is determined by the smoothness of the process and the characteristics of $F_{\mathbf{x}}$. Second, both results suggest well dispersed rep-points over the domain $\mathcal{X}$. The k-DPP induces strong negative coefficients between the selected points. The support points suggested by us are space-filling. A detailed comparison of their performance as rep-points is presented in Section 3.2, highlighting the efficiency of SPs.

2.2 Estimated covariance function

Usually, the covariance c is unknown and is replaced by an estimated one, i.e., $\hat{c}$. This section examines the influence of the estimated covariance $\hat{c}$ on the convergence rate of the original process and the conditions on rep-points for the predictive process such that its convergence can achieve the same rate. Similar to the studies by Tuo and Wang (2020) and Wang and Jing (2022), we demonstrate the undersmoothed and oversmoothed cases, where $\hat{\gamma}$ is smaller and larger than γ, respectively. Notation $a_n \preceq b_n$ for two positive sequences a_n and b_n indicates that there is a $B > 0$ such that $a_n \leq Bb_n$.

Theorem 2. *(Convergence rate of $\tilde{f}_{\hat{c}}$) Assume that the regularity conditions [R1]–[R3] hold for $\hat{c}$ and $\hat{\gamma}, \gamma > 2$. As $n \to \infty$, $V(\hat{f}_{\hat{c}} - f) = O_p\left\{n^{-\{\min(\hat{\gamma},\gamma)-1\}/\hat{\gamma}}\right\}$. Furthermore, assume that $J_{\hat{c}}(f) \preceq O_p\{n^{-(\gamma-1-\hat{\gamma})/\hat{\gamma}}\}$ when $\hat{\gamma} > \gamma - 1$. Then, for $D_k^* \subset \{\mathbf{x}_i\}_{i=1}^n$, if $n^{2/\hat{\gamma}}E(F_{\mathbf{x}}, F_k^*) \to 0$ as $k, n \to \infty$, the prediction $\tilde{f}_{\hat{c}}$ achieves a convergence rate identical to that achieved by $\hat{f}_{\hat{c}}$*

$$V(\tilde{f}_{\hat{c}} - f) = O_p\left\{n^{-\frac{\min(\hat{\gamma},\gamma)-1}{\hat{\gamma}}}\right\}. \tag{6}$$

Theorem 2 shows the convergence rates of both $\hat{f}_{\hat{c}}$ and $\tilde{f}_{\hat{c}}$. For the undersmoothed case, i.e., $\hat{\gamma} < \gamma$, the convergence rate is $O_p\{n^{-(1-1/\hat{\gamma})}\}$. For the oversmoothed case, i.e., $\hat{\gamma} > \gamma$, the convergence rate becomes $O_p\{n^{-(\gamma-1)/\hat{\gamma}}\}$. The closer $\hat{\gamma}$ is to γ, the faster is the convergence. When $\hat{\gamma} = \gamma$, Theorem 2 becomes Theorem 1. Even in the case of an

estimated covariance, the convergence rate of $\tilde{f}_{\hat{c}}$ can be identical to that of $\hat{f}_{\hat{c}}$ with sufficient rep-points. To achieve convergence, we require $E(F_{\mathbf{x}}, F_k^*) = o_p\{n^{-2/\hat{\gamma}}\}$, indicating that $\tilde{f}_{\hat{c}}$ in the undersmoothed case requires additional rep-points than $\tilde{f}_c$ to achieve convergence; in the oversmoothed case, $\tilde{f}_{\hat{c}}$ achieves convergence faster.

With an estimated covariance, the following theorem suggests the adjustment of τ^2 to improve the convergence rates of $\hat{f}_{\hat{c}}$ and $\tilde{f}_{\hat{c}}$ when k is given and large enough.

Theorem 3. *(Adjusted τ^2) Assume that the regularity conditions [R1]–[R3] hold for $\hat{c}$ and $\hat{\gamma}, \gamma > 2$. If $\tau^2 = O(n^{-\hat{\gamma}/\gamma+1})$, the prediction based on full data achieves the optimal convergence rate, i.e., $V(\hat{f}_{\hat{c}} - f) = O_p\left\{n^{-(1-1/\gamma)}\right\}$ as $n \to \infty$. Furthermore, assume that $J_{\hat{c}}(f) \preceq O_p\{n^{(\gamma-1-\hat{\gamma})/\gamma}\}$ when $\hat{\gamma} > \gamma - 1$. Then, for $D_k^* \subset \{\mathbf{x}_i\}_{i=1}^n$, if $n^{2/\hat{\gamma}}E(F_{\mathbf{x}}, F_k^*) \to 0$ as $k, n \to \infty$, the prediction based on the rep-points achieves the optimal convergence rate*

$$V(\hat{f}_{\hat{c}} - f) = O_p\left\{n^{-(1-1/\gamma)}\right\}.$$

Theorem 3 demonstrates that even with a mis-specified covariance, $\tilde{f}_{\hat{c}}$ and $\hat{f}_{\hat{c}}$ can converge with the same rate as $\tilde{f}_c$ and $\hat{f}_c$, respectively, by properly adjusting the order of τ^2. For the undersmoothed case, the adjusted τ^2 has a higher order than $O(1)$, and vice versa. The larger the $|\hat{\gamma} - \gamma|$, the larger is the required adjustment of τ^2. For $\hat{\gamma} = \gamma$, we do not require to adjust τ^2 and Theorem 3 becomes Theorem 1. Although we can never adjust τ^2 to be of the exact order desired, Theorem 3 provides certain guidance. For the example of the undersmoothed case, we can make τ^2 larger than the initial value to improve the prediction performance of $\tilde{f}_{\hat{c}}$ and $\hat{f}_{\hat{c}}$.

3   Simulations

This section uses simulated data to show the properties of the smoothness parameter γ, demonstrate the advantages of low-rank approximation with SPs under various scenarios,

and confirm our theoretical results. The computations were implemented by R 3.6.3 (R Core Team, 2020) on a machine with Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz and 125 GB RAM.

3.1 Smoothness parameter γ

We use the Matérn covariance function as an example to provide an intuitive demonstration of the smoothness parameter γ. Assume that the process is stationary and c has the following form:

$$c(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\psi} \right)^\nu \mathcal{K}_\nu \left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{\psi} \right), \tag{7}$$

where $\sigma^2 > 0$ is the marginal variance such that $c(\mathbf{x}, \mathbf{x}) = \sigma^2$, $\psi > 0$ is called the range parameter; $\nu > 0$ determines the smoothness of $f(\mathbf{x}_i)$ (or Z) and is termed the smoothness parameter in Matérn; $\mathcal{K}_\nu$ is the modified Bessel function of the second type of order ν.

For obtaining an empirical estimate of γ, we calculate the eigenvalues of an $n_0 \times n_0$ matrix $\mathbf{C}_0 = \{c(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^{n_0}$, i.e., $\lambda_{01}, \ldots, \lambda_{0n_0}$, where $\{\mathbf{x}_i\}_{i=1}^{n_0}$ are uniformly distributed on $[0, 1]^2$. As $\lambda_{0i}^{-1} = O(i^\gamma)$, we fit $\{(\log i, -\log \lambda_{0i})\}_{i=1}^{n_0}$ with a linear regression and obtain the slope as an estimate of γ. Fig. 2 shows how γ changes with the parameters in the
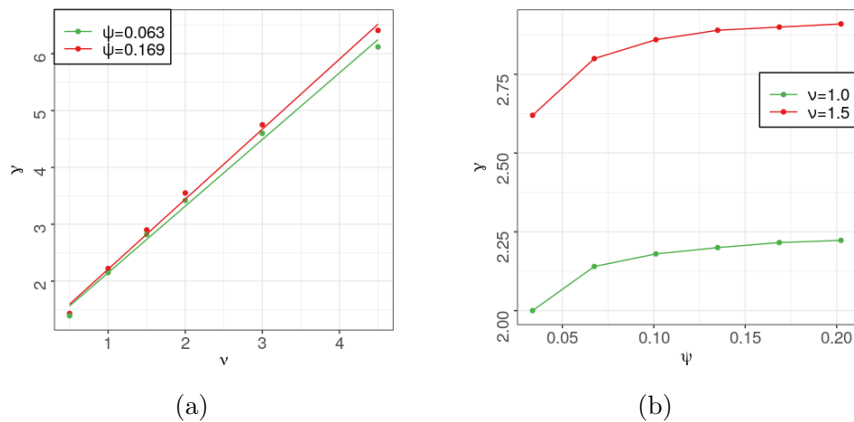

Figure 2: Values of the smoothness parameter γ for the Matérn covariance function (7) under various parameter settings: (a) weak and strong correlations; (b) rough and smooth realizations.

14

Matérn covariance, where σ^2 is fixed at 1.5. In Fig. 2(a), we set ψ to be 0.063 and 0.169 for covering the cases of weak and strong correlations, respectively. Based on other parameters, the smoothness parameter γ linearly grows with the parameter ν; this result is in agreement with the discussion reported by Wang and Jing (2022) that γ is about $2(\nu + d/2)/d$. In Fig. 2(b), γ increases with ψ, though not linearly. Moreover, although ψ varies within a large range, γ varies within 0.5. Compared with ψ, γ is more sensitive to the value of ν.

Two Matérn covariances with consistent parameters $(\sigma_1^2, \psi_1, \nu)$ and $(\sigma_2^2, \psi_2, \nu)$ satisfying $\sigma_1^2/\psi_1^{2\nu} = \sigma_2^2/\psi_2^{2\nu}$ define two equivalent probability measures and thus obtain asymptotically equal predictions (Zhang, 2004). Our simulation and theoretical results taken together demonstrate that the prediction performance is determined by the γ value, and two consistent parameter settings should have an asymptotically equal γ. Table 1 lists the average values of γ among ten replicates under three consistent parameter settings. The average values get closer as n_0 increases. This result supports our claim that γ is a crucial measure of the smoothness of the process. Further validation is provided through an example utilizing another covariance, as demonstrated in Section S3 of the Supplementary Materials.

Table 1: Values of γ under consistent parameter settings obtained from the data of increasing size n_0 of the Matérn covariance function (7).

(σ^2, ψ, ν)	$n_0 = 2000$	4000	6000	8000
$(1.5, 0.169, 1.5)$	2.8899	2.9011	2.9064	2.9026
$(1.0, 0.147, 1.5)$	2.8666	2.8865	2.8954	2.8937
$(2.0, 0.186, 1.5)$	2.9045	2.9102	2.9132	2.9082

3.2 Performance of various rep-points choices

The performance of various rep-points choices under the following four scenarios were tested and compared:

1. **Strong correlation**: $Z \sim \mathrm{GP}(0, c_s)$, where c_s belongs to the Matérn class in (7) with

$\boldsymbol{\theta}_s = (1.5, 0.169, 1.5)^\top$. The effective range of Z is 0.8. Here $F_\mathbf{x}$ is the uniform distribution on $[0, 1]^2$, i.e., $U_{[0,1]^2}$. Four choices of rep-points (SPs, k-DPPs, Grids and Rands) of size $k = (0.10, 0.18, 0.28, 0.40, 0.55, 0.81, 1.12, 1.36) \times n^{2/2.9} \approx (36, 64, 100, 144, 196, 289, 400, 484)$ were selected based on the full data, where 2.9 is the estimated γ for c_s obtained from Table 1. Two mis-specified covariances with $\hat{\boldsymbol{\theta}}_s = (1.5, 0.169, 1.0)^\top$ and $(1.5, 0.169, 3.0)^\top$ were imposed as the undersmoothed and oversmoothed cases, respectively.

2. **Weak correlation**: $Z \sim \mathrm{GP}(0, c_w)$, where c_w belongs to the Matérn class in (7) with $\boldsymbol{\theta}_w = (1.5, 0.063, 1.5)^\top$. The effective range of Z is 0.3, and $F_\mathbf{x} = U_{[0,1]^2}$. SPs, k-DPPs, Grids and Rands of sizes $k = (0.66, 1.21, 1.66, 2.05, 2.63, 3.12, 3.64, 4.02) \times n^{2/2.8} \approx (289, 529, 729, 900, 1156, 1369, 1600, 1764)$ are chosen, where 2.8 is the estimate of γ for c_w obtained from Fig. 2(a). Two mis-specified covariances with $\hat{\boldsymbol{\theta}}_w = (1.5, 0.063, 1.0)^\top$ and $(1.5, 0.063, 3.0)^\top$ were imposed as undersmoothed and oversmoothed cases, respectively.

3. **Consistency**: The settings are the same as those in the first scenario, although parameters of mis-specifed covariances are substituted as $\hat{\boldsymbol{\theta}}_s = (1.0, 0.147, 1.5)^\top$ and $(2.0, 0.186, 1.5)^\top$ to be consistent with the parameters of c_s as discussed in Section 3.1.

4. **Nonuniform $F_\mathbf{x}$**: $Z \sim \mathrm{GP}(0, c_s)$. The covariance c_s is the same as in the first scenario. Here, we have 75% of $\{\mathbf{x}_i\}_{i=1}^n$ uniformly distributed in $[0.0, 0.5]^2$ and the other 25% uniformly distributed in the remaining region of $[0, 1]^2$, as described in Fig. 1(c) and 1(d). Five choices of rep-points were considered, i.e., Grids, k-DPPs, Rands and SPs generated by $\{\mathbf{x}_i\}_{i=1}^n$ (SPs) and by $U_{[0,1]^2}$ (SPUs). The size k was the same as that in the first scenario.

For each scenario, we generate $\{(\mathbf{x}_i, f_i)\}_{i=1}^{n+n_t}$, where $f_i = f(\mathbf{x}_i)$'s are the realizations of the

GP Z and $n = n_t = 5000$. Then, we randomly divide the data into two parts of sizes n and n_t. Next, we perturb the training data with random noises from $\mathcal{N}(0, 0.27)$. Finally, we evaluate $f(\mathbf{x})$ on the remaining n_t locations $\{\mathbf{x}_{ti}\}_{i=1}^{n_t}$ based on the rep-points or full data. The performance of each method is evaluated in terms of the root mean squared prediction error (RMSPE), i.e., $[n_t^{-1} \sum_{i=1}^{n_t} \{\tilde{f}_{\hat{c}}(\mathbf{x}_{ti}) - f(\mathbf{x}_{ti})\}^2]^{1/2}$. We show the average of RMSPEs over 100 replicates of the abovementioned process in Fig. 3.

In Figs. 3(a) and 3(b), the RMSPEs for $\tilde{f}_c$ and $\tilde{f}_{\hat{c}}$ decrease at first and then converge to the levels of $\hat{f}_c$ and $\hat{f}_{\hat{c}}$, respectively, as k increases, regardless of the types of rep-points used. This result agrees with our theoretical results that $\tilde{f}_{\hat{c}}$ and $\hat{f}_{\hat{c}}$ may achieve the same convergence rate if D_k^* (or rep-points) approximate $F_{\mathbf{x}}$ well enough and validates the efficiency of the rep-points. The solid lines representing SPs exhibit the fastest decline, regardless of whether the imposed/estimated covariance is correct or not. This result highlights the superior efficiency of SPs in representing the set $\{\mathbf{x}_i\}_{i=1}^n$. The performance of k-DPPs falls between that of SPs and Grids. Note that the k-DPP algorithm becomes unavailable as ν and k increase. When the covariance is smoother, k-DPPs tend to be more dispersed. However, placing an excessive number of k-DPPs within a fixed domain would disrupt the negative correlations between them. Rands exhibit the worst performance because of the lack of the space-filling property.

To understand the influence of the imposed covariance, we compare curves of the same type but shown in different colors. When k is not extremely small, the predictions given by the true covariance have the lowest RMSPEs, indicating that a wrongly specified covariance can typically weaken the prediction performance. When k is quite small, imposing an oversmoothed covariance function may yield a more accurate prediction. A larger value of $\hat{\gamma}$ (or $\hat{\nu}$) results in the earlier convergence of the corresponding $\tilde{f}_{\hat{c}}$; this agrees with our

(a) Strong correlation

(b) Weak correlation

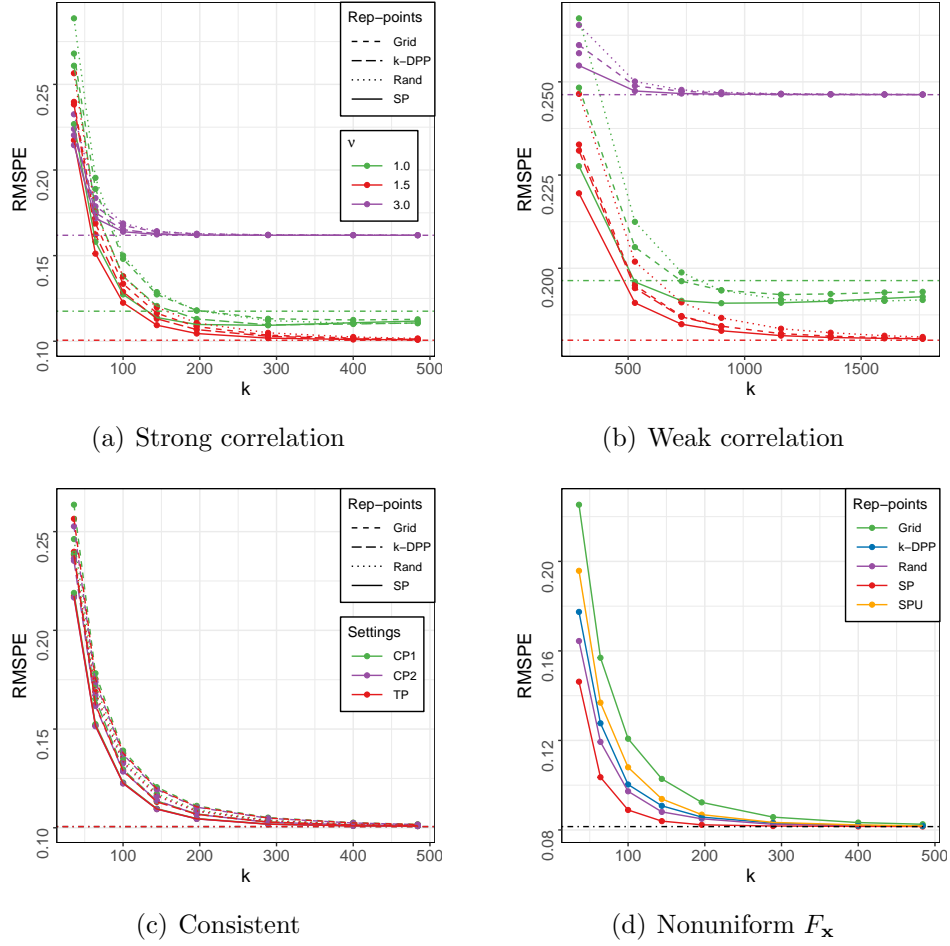(c) Consistent

(d) Nonuniform $F_{\mathbf{x}}$

Figure 3: Comparison of various rep-points under four scenarios (a)–(d). The dashed–dotted lines represent the RMSPEs of predictions based on the full data. The legend in (b) is the same as that in (a). CP1, CP2, and TP in (c) correspond to the consistent parameters $\hat{\boldsymbol{\theta}}_{s1}$, $\hat{\boldsymbol{\theta}}_{s2}$, and true parameter $\boldsymbol{\theta}_s$, respectively, in the third setting.

theoretical results, according to which the required k for the convergence of $\tilde{f}_{\hat{c}}$ is of $O(n^{2/\hat{\gamma}})$.

The predictions shown in Fig. 3(a) have smaller RMSPEs than those shown in Fig. 3(b). The second setting involves a more complex structure with a smaller ψ and hence a smaller γ. Moreover, $\tilde{f}_{\hat{c}}$ and $\tilde{f}_c$ in Fig. 3(b) require a larger k to achieve convergence, further demonstrating the influence of shortening of the range parameter ψ. In particular, define K as the size of rep-points such that $\tilde{f}_c$ based on SPs achieves convergence. Remember that K has a higher order than $O(n^{2/\gamma})$. Then, both ν and ψ influence K. When ψ is fixed, K changes less with ν. For example, K for different lines (in different colors) in Fig. 3(a) are

18

all within 500. However, when we change the value of ψ, K drastically varies. For example, the K for the red solid line in Fig. 3(b) is much larger than that shown in Fig. 3(a).

Figure 3(c) shows the performance of predictions with consistent parameters. Curves represented in different colors but of the same type are approximately coincident. That is, by controlling the choice of rep-points, we can achieve $\tilde{f}_{\hat{c}}$ with consistent $\hat{c}$ to give an asymptotically identical prediction performance. This result agrees with our discussion reported in Section 3.1 that consistent covariance functions have asymptotic identical values of $\hat{\gamma}$, which determine the performance of $\tilde{f}_{\hat{c}}$. Moreover, we confirmed the conclusion of Zhang (2004), i.e., an incorrect but consistent $\hat{c}$ has the same prediction performance as c.

Fig. 3(d) shows the performance of $\tilde{f}_c$ based on various rep-points when $\{\mathbf{x}_i\}_{i=1}^n$ are nonuniformly distributed over a compact region $\mathcal{X}$. $\tilde{f}_c$ based on SPs, k-DPPs, and Rands converge earlier than those based on SPUs and Grids. This is because they follow $F_{\mathbf{x}}$, whereas SPUs and Grids follow $U_{[0,1]^2}$. In other words, the distribution of rep-points should be as close to F_n as possible. SPs meet this requirement by minimizing the energy distance between their empirical distribution and F_n. For an intuitive comparison, we provide $E(F_k^*, F_n)$ for the five rep-points in Table 2. All the energy distances decrease as k grows. SPUs and Grids have larger $E(F_k^*, F_n)$ than SPs and Rands because they follow a distribution different from $F_{\mathbf{x}}$. Even with the same asymptotic distribution, there exist large gaps between the energy distances of various rep-points, particularly for small k. For example

Table 2: $E(F_k^*, F_n)$ for multiple rep-points with $n = 5000$. Furthermore, k is varied in the fourth scenario. The minimum $E(F_k^*, F_n)$ for each k is highlighted in bold.

Rep-points	$k = 36$	64	100	144	196	289	400	484
SPs	**0.001081**	**0.000455**	**0.000237**	**0.000141**	**0.000093**	**0.000056**	**0.000039**	**0.000034**
Rands	0.013513	0.007517	0.004582	0.003171	0.002268	0.001555	0.001244	0.001053
k-DPP	0.017491	0.014829	0.013808	0.012821	0.012957	0.012167	0.011523	0.011610
SPUs	0.030338	0.029755	0.029427	0.029356	0.029288	0.029318	0.029379	0.029305
Grids	0.034862	0.032329	0.031250	0.030650	0.030270	0.029863	0.029729	0.029670

when $k = 484$, the $E(F_n, F_k^*)$ for SPs is only 3.23% and 0.29% of those for Rands and k-DPPs, respectively. SPs are the most efficient in minimizing $E(F_k^*, F_n)$. The performance of k-DPPs is affected by the non-uniform distribution.

3.3 Influence of $\hat{c}$

First, we confirm Theorem 2 by exploring the orders of $V(\tilde{f}_{\hat{c}} - f)$ and $V(\tilde{f}_c - f)$ with respect to n for a sufficiently large k. Let $F_{\mathbf{x}} = U_{[0,1]^2}$, $n = 1000, 1500, \ldots, 7000$, and $n_t = 5000$. For each n, we follow the procedure described in Section 3.2 to generate the training data and to predict on the testing location based on SPs and Rands with $k = \lfloor 1.5 \times n^{2/2.9} \rfloor$. We consider c_s in the first scenario as the true covariance such that $\tilde{f}_{\hat{c}}$ and $\tilde{f}_c$ can converge to $\hat{f}_{\hat{c}}$ and $\hat{f}_c$ with a smaller k. The values of σ^2 and ψ in c_s are fixed; however, ν is set to be 0.5 and 1.0 for undersmoothed covariances and 2.0 and 3.0 for oversmoothed ones.

The MSPEs for $\tilde{f}_{\hat{c}}$ and $\tilde{f}_c$ may well approximate $V(\tilde{f}_{\hat{c}} - f)$ and $V(\tilde{f}_c - f)$, respectively, as per the law of large numbers and because n_t is quite large. Let $\mathrm{MSPE}(n)$ be the MSPE obtained from the training data of size n. Then, by Theorem 2, we have $\log\{\mathrm{MSPE}(n)\} \approx -\frac{\min(\gamma, \hat{\gamma}) - 1}{\hat{\gamma}} \log n + \log C_0$, where C_0 is the multiplicative constant of the order term. Our target is to fit the linear regressions above and analyze the slopes.

In Fig. 4, we show the average $\log\{\mathrm{MSPE}(n)\}$s of 100 replicates (using points) and their fitted lines. For simplicity of discussion, only SPs and Rands are considered. The points for each n and lines in the same color are nearly coincident in most cases, indicating that $k = \lfloor 1.5 \times n^{2/2.9} \rfloor$ is enough for both rep-points in this scenario, and hence $\tilde{f}_{\hat{c}}$ and $\tilde{f}_c$ have converged to $\hat{f}_{\hat{c}}$ and $\hat{f}_c$, respectively. Red lines have the largest negative slopes (NSs) because the prediction under the true covariance has the best performance. Additionally, the value of C_0 for red lines is around 2.4. For undersmoothed cases, the NSs for SP(0.5)
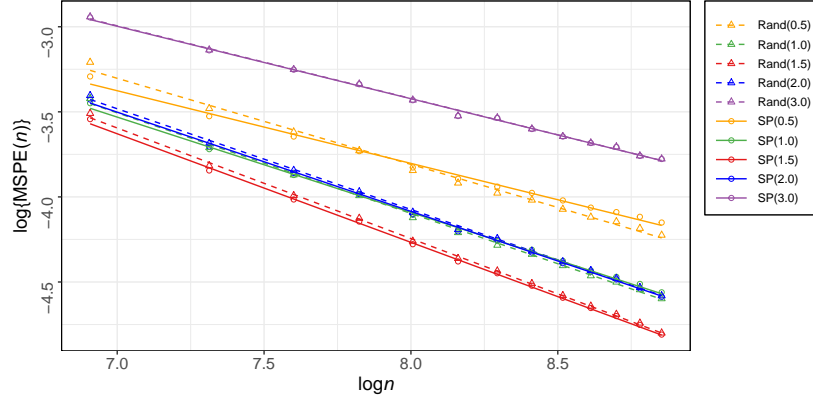
Figure 4: Prediction performance of $\tilde{f}_{\hat{c}}$ and $\tilde{f}_c$ with varying n and a sufficiently large k. Rand(0.5) in the legend indicates the performance of $\tilde{f}_{\hat{c}}$ with Rands and $\nu = 0.5$, and so on for the rest.

and Rand(0.5) are smaller than those for SP(1.0) and Rand(1.0). For oversmoothed cases, the NSs for SP(3.0) and Rand(3.0) are smaller than those for SP(2.0) and Rand(2.0). The worse the estimate of $\hat{c}$, the larger is the $|\hat{\gamma} - \gamma|$; hence, the worse is the behavior of $\tilde{f}_{\hat{c}}$.

As per (5) and (6), we can evaluate the values of $\hat{\gamma}$ and γ through NS. That is, $\hat{\gamma} \approx 1/(1 - \mathrm{NS})$ for undersmoothed cases and $\hat{\gamma} \approx (\gamma - 1)/(1 - \mathrm{NS})$ for oversmoothed cases. The results from 100 replicates are summarized in Fig. 5. For comparison, we provide the empirical estimates of $\hat{\gamma}$ and γ obtained as described in Section 3.1. In most cases, boxes cover the respective black dotted lines well, thereby confirming our theory. For the
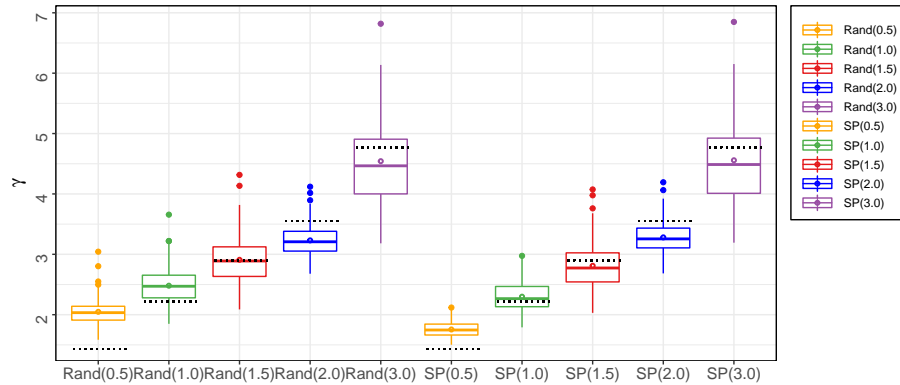

Figure 5: Boxplots for the theoretically obtained $\hat{\gamma}$ and γ. The circles in the middle of boxes are the average values. The black dotted lines are $\hat{\gamma}$ and γ, as reported in Section 3.1.

21

undersmoothed covariance with $\nu = 0.5$, the black dotted lines are considerably lower than their boxes because their smoothness parameters are smaller than 2. Hence our theorems do not apply in these cases. Our theory was further confirmed by setting σ^2 as 0.5 while maintaining the other parameters fixed for both the true and mis-specified covariances. This verification is illustrated in Fig. S1.

Finally, we demonstrate the improvement of $\tilde{f}_{\hat{c}}$ by adjusting τ^2. Still using the first scenario and $\tau^2 = 0.27$, we generate $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $n = 5000$ and $\{(\mathbf{x}_{ti}, f_{ti})\}_{i=1}^{n_t}$ with $n_t = 5000$ in the same way. The predictions $\tilde{f}_c(\mathbf{x}_{ti})$ and $\tilde{f}_{\hat{c}}(\mathbf{x}_{ti})$ are calculated using the adjusted $\tau^2 = 10^{(-3,2,1,0,0.5,1,1.5,2)} \times 0.27$. Fig. 6 summarizes the average log(MSPE)s over 100 replicates versus $a_\tau = \log_{10}(\tau^2/0.27)$. When the correct covariance ($\nu = 1.5$) is used, the red curve achieves the lowest MSPE when $a_\tau = 0$, and this lowest MSPE corresponds to the true value of τ^2. When the covariance is mis-specified ($\nu = 1$ or 3), neither of the two curves reach the minimum MSPE under the true τ^2. However, the performance of the oversmoothed covariance (with $\nu = 3$) improves when a_τ is reduced to -2; the performance of the undersmoothed covariance (with $\nu = 1$) improves when a_τ is increased to 0.5. Furthermore, the lowest MSPEs of $\tilde{f}_{\hat{c}}$ are comparable to those of $\tilde{f}_c$; this result is in agreement with Theorem 3 that $\tilde{f}_{\hat{c}}$ may have the same convergence rate as $\tilde{f}_c$ for an appropriate value of τ^2.
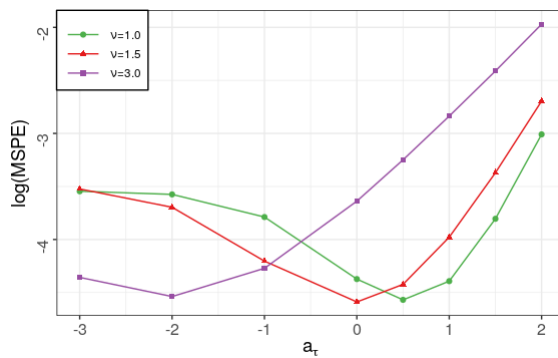

Figure 6: Prediction performance of $\tilde{f}_{\hat{c}}$ and $\tilde{f}_c$ with varying τ^2.

4 Data Examples

This section shows the advantages of low-rank approximations with an estimated covariance and SPs on two real datasets: annual total precipitation anomalies and total column ozone data. We compare our method with other popular methods such as FRK (Zammit-Mangion and Cressie, 2021), LatticeKrig (Nychka et al., 2016), and autoFRK (Tzeng et al., 2021).

4.1 Annual total precipitation anomalies data

In the first application, we consider the data of annual total precipitation anomalies observed at 7352 weather stations in the United States in 1962. This dataset was discussed by Kaufman et al. (2008) and Sang and Huang (2012). As shown in Fig. S3, the locations are nonuniformly distributed over an irregular region $\mathcal{X}$. Without an obvious mean trend, nonstationarity or anisotropy (Kaufman et al., 2008), we fit the data with a stationary Gaussian process with mean zero. In particular, we randomly divide the data into training data with $n = 7000$ and testing data with $n_t = 352$. Based on the training data, we apply the Vecchia approximation to obtain an approximated covariance estimation. This approximation can be implemented by the R package $GpGp$ (Guinness et al., 2021), where we set the covariance to be Matérn in (7). Furthermore, we use $ExaGeoStat$ (Abdulah et al., 2018a) to calculate the exact covariance estimation as a benchmark. Using these approximated covariance and SPs of $k = 210$, 500, 750 and 1000, we obtain $\tilde{f}_{\hat{c}}(\mathbf{x}_{ti})$s. We calculate the MSPE $= n_t^{-1} \sum_{i=1}^{n_t} \{y_{ti} - \tilde{f}_{\hat{c}}(\mathbf{x}_{ti})\}^2$ to measure the prediction performance of our method, where $\{y_{ti}\}_{i=1}^{n_t}$ are the responses of the testing data.

Moreover, we consider the method of using the estimated covariance and full data (FULL) and three popular low-rank approximation methods mentioned earlier, namely, FRK, LatticeKrig, and autoFRK. The FRK performs inference on basic areal units (BAUs).

It develops multiresolution basis functions and calculates coefficients by using the expectation-maximization (EM) algorithm. We set two and three levels of resolutions and exponential kernel functions. LatticeKrig induces the basis functions with regular lattices of different levels and uses the Markov random field assumption to develop a sparse precision matrix for random weights. The higher the level, the finer are the grid points. This method can be implemented by the R package *LatticeKrig* (Nychka et al., 2016). We test the performance of LatticeKrig with smoothness parameters $\nu = 0.5$ and 1, and spatially autoregressive weights AW=4.1, 5.0, and 6.0, two levels of resolutions, and 15 grids points at the coarsest resolution. For additional details of these parameters, we refer the readers to Nychka et al. (2015). The autoFRK develops adaptive basis functions corresponding to different scales with thin-plate splines. Based on a maximum number B of bases, it selects $B^* \leq B$ bases by the Akaike's information criterion (AIC). The method can be implemented by the R package *autoFRK* (Tzeng et al., 2021). Here we set B as 210, 500, 750 and 1000.

We repeat the above process 100 times and report the average performance of these methods with regard to their MSPEs, computational time (seconds) of covariance estimation and prediction, and the actual number of bases. These results are summarized in Table 3. For FULL and SP(210)–(1000), Table 3 only lists the results of predictions using the estimated covariance calculated by *GpGp*. Based on the exact covariance estimation calculated by the *ExaGeoStat*, FULL and SP(210)–(1000) have MSPEs 0.22096, 0.31711, 0.27879, 0.26108, and 0.25124, respectively. For LatticeKrig, we only listed the best result among all the parameter settings, i.e., the result for $\nu = 1$ and AW= 6.0. FRK has the lowest MSPE when #Bases is small but requires longer time. Moreover, the decrease in its MSPE is limited when #Bases increases. Furthermore, with multiresolution bases, autoFRK behaves well when #Bases is small. However, its computational time rapidly

Table 3: Performance of various methods for the annual total precipitation data. FULL represents the prediction based on full data and estimated covariance obtained by *GpGp*. SP(210)–(1000) represent low-rank approximations based on 210–1000 SPs and estimated covariance obtained by *GpGp*. FRK(2)–(3) represent the FRK method with two and three levels of resolutions, respectively. Furthermore, autoFRK(210)–(1000) represent the autoFRK method with $B = 210$–1000. The approximation with the minimum MSPE is marked in bold.

Methods	FULL	SP(210)	SP(500)	SP(750)	**SP(1000)**	FRK(2)
MSPE	0.22097	0.31694	0.27897	0.26138	**0.25159**	0.28670
Time (s)	159	9	10	11	**13**	15
#Bases	7000	210	500	750	**1000**	210
Methods	FRK(3)	autoFRK(210)	autoFRK(500)	autoFRK(750)	autoFRK(1000)	LatticeKrig
MSPE	0.27809	0.30461	0.26883	0.25975	–	0.29958
Time (s)	383	6	13	21	–	51
#Bases	1936	205	499	748	–	1258

increases and it cannot be applied when B is larger than about 1000. LatticeKrig uses more bases and consumes more time. The performance of our method is less superior when #Bases is small because SPs cannot represent the full data locations well. However, the performance is drastically improved as we increase k. SP(1000) provides a smaller MSPE with less computational time compared to FRK(3) and LatticeKrig, both of which use more bases. Thus, the superiority of our method is verified.

4.2 Total column ozone data

In the second application, we demonstrate the feasibility of application of various methods on a large and nonstationary dataset, which includes $173, 405$ observations of the level-2 total column ozone for October 1, 1988, along with their locations. The data were collected and preprocessed by NASA and were discussed by Cressie and Johannesson (2008) and Meng et al. (2020). The region of interest $\mathcal{X}$ has a regular shape. We fit the data with $Z \sim \mathrm{GP}(\mu, c)$, where μ is a constant and $c(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \Gamma(\nu)^{-1} 2^{1-\nu} (\|\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j\|_2)^\nu \mathcal{K}_\nu(\|\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j\|_2)$; here, μ, σ^2, ν, and the 2×2 symmetric matrix $\mathbf{L}$ are the parameters to be evaluated.

We divide the data into two parts: training data with $n = 150,000$, and testing data with $n_t = 23,405$. We evaluate the covariance function using the R package *GpGp*. The SPs of $k = 189$, 500, 1361, and 1755 are generated for facilitating better comparison with the other methods. We fitted the low-rank approximation based on the estimated covariance and SPs induced by the training data and performed predictions on the testing data. For the total column ozone data, only FRK and LatticeKrig are considered for comparison as autoFRK is not applicable because of the large sample size. For implementing FRK, we set two and three levels of resolutions and Matérn kernel functions with $\nu = 1.5$. For LatticeKrig, we explore the parameter settings described in the previous subsection, and we set $\nu = 0.5$, AW= 4.1, two levels of resolutions, and 15 grid points at the coarsest resolution. Neither FRK nor LatticeKrig allow smoothly setting the number of bases.

The average results over 100 runs are summarized in Table 4. Similar to the first example, the MSPE of FRK slightly decreases but its computational cost dramatically increases with an increase in #Bases. LatticeKrig, while being the fastest, does not have efficient bases. For instance, with the same #Bases, the MSPE of LatticeKrig(2) is more than twice that of SP(1361). With comparable MSPEs, the #Bases of LatticeKrig(3) is more than twice that of SP(1755). The basis functions induced by the SPs are more efficient. The MSPE of SP methods decreases significantly as #Bases grows, with only a

Table 4: Performance of various methods for the total column ozone data. LatticeKrig(2) and (3) mean LatticeKrig with two and three levels of resolutions, respectively. The one with the minimum MSPE is marked in bold.

Methods	FRK(2)	FRK(3)	SP(189)	SP(500)
MSPE	105.94	50.18	181.12	70.00
Time (s)	299	6521	437	462
#Bases	189	1755	189	500
Methods	SP(1361)	**SP(1755)**	LatticeKrig(2)	LatticeKrig(3)
MSPE	38.80	**35.87**	84.85	38.31
Time (s)	531	**563**	132	212
#Bases	1361	**1755**	1361	3840

slight increase in computational time. In our method, the main time-consuming parts are the estimation of covariance parameters and the generation of the SPs. For the first issue, one can choose more efficient algorithms. To address the second issue, one can generate SPs from a random subset of the training data without affecting the space-filling property.

5 Discussion

In this study, we derived the asymptotic performance of a low-rank GP prediction $\tilde{f}_{\hat{c}}$, investigated the influence of rep-points D_k^* and estimated covariance $\hat{c}$ on the convergence rate of $\tilde{f}_{\hat{c}}$. Using the concept of energy distance, we demonstrated that the distribution of rep-points should be as close as possible to that of the full data locations. This insight motivates the utilization of SPs, which have space-fillingness but also can best mimic the full data. We set the order of k such that the convergence rate of $\tilde{f}_{\hat{c}}$ is equal to that of $\hat{f}_{\hat{c}}$ under certain regularity conditions. These conditions are exemplified in Section S3 to facilitate comprehension. With $\hat{\gamma}$ indicating the smoothness of $\hat{c}$, we provided the convergence rates of $\tilde{f}_{\hat{c}}$ and $\hat{f}_{\hat{c}}$ when sufficient rep-points are given, i.e., $V(\tilde{f}_{\hat{c}} - f) = V(\hat{f}_{\hat{c}} - f) = O_p[n^{-\{\min(\hat{\gamma},\gamma)-1\}/\hat{\gamma}}]$. This indicates that the closer the $\hat{\gamma}$ is to γ, the faster is the convergence. Moreover, we demonstrated the performance improvement of $\tilde{f}_{\hat{c}}$ by adjusting the order of the nugget effect. We demonstrated the value of γ and confirmed our theoretical results via extensive numerical studies. Using two examples of real data, we validated the prediction of $\tilde{f}_{\hat{c}}$ based on SPs by comparing our results with the results obtained using other existing low-rank approximation methods.

In some of the existing methods, such as FRK and LatticeKrig, multiresolution bases are desirable to capture information at different scales. In this way, they can better represent the local dependency and evaluate the covariance. In this study, we used the bases induced

by SPs without multiresolution capability because we fixed the covariance and focused only on the prediction accuracy. We preferred to consider the tasks of covariance estimation and prediction separately because they require different configurations of rep-points. Literature in the field of spatial designs suggests future avenues, e.g., samples in clusters behave better for covariance estimation, whereas widely distributed samples are better for prediction (Zhu and Stein, 2005, 2006; Barbian and Assunção, 2017). Therefore, one possible extension of our work is the configuration of rep-points for covariance estimation.

Supplementary Materials

A supplement to the main manuscript, including proofs for Theorems 1–3, an illustrative example for regularity conditions [R1]–[R3], and additional simulated results, is provided in the online supplementary materials. (.pdf file)

The R code, data, and instructions for reproducing the results in the article are available at the GitHub repository: SpatialTemporalStats/LRGPSP_Reproducibility_Materials.

Disclosure Statement

The authors report there are no competing interests to declare.

References

Abdulah, S., Ltaief, H., Sun, Y., Genton, M. G., and Keyes, D. E. (2018a), "ExaGeo-Stat: a high performance unified software for geostatistics on manycore systems," *IEEE Transactions on Parallel and Distributed Systems*, 29, 2771–2784.

— (2018b), "Parallel approximation of the maximum likelihood estimation for the prediction of large-scale geostatistics simulations," in *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, pp. 98–108.

— (2019), "Geostatistical modeling and prediction using mixed precision tile Cholesky factorization," in *2019 IEEE 26th International Conference on High Performance Computing, Data, and Analytics (HiPC)*, pp. 152–162.

Alaoui, A. and Mahoney, M. W. (2015), "Fast randomized kernel ridge regression with statistical guarantees," in *Advances in Neural Information Processing Systems*, eds. Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., Curran Associates, Inc., vol. 28.

Banerjee, A., Dunson, D. B., and Tokdar, S. T. (2013), "Efficient Gaussian process regression for large datasets," *Biometrika*, 100, 75–89.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 825–848.

Barbian, M. H. and Assunção, R. M. (2017), "Spatial subsemble estimator for large geostatistical data," *Spatial Statistics*, 22, 68–88.

Berlinet, A. and Thomas-Agnan, C. (2004), *RKHS and Stochastic Processes*, Boston, MA: Springer US, pp. 55–108.

Burt, D., Rasmussen, C. E., and Van Der Wilk, M. (2019), "Rates of convergence for sparse variational Gaussian process regression," in *International Conference on Machine Learning*, PMLR, pp. 862–871.

Burt, D. R., Rasmussen, C. E., and Van Der Wilk, M. (2020), "Convergence of sparse

variational inference in Gaussian processes regression," *The Journal of Machine Learning Research*, 21, 5120–5182.

Cressie, N. (2015), *Statistics for Spatial Data*, John Wiley & Sons, Revised edition.

Cressie, N. and Johannesson, G. (2008), "Fixed rank kriging for very large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 209–226.

Deisenroth, M. and Ng, J. W. (2015), "Distributed Gaussian processes," in *Proceedings of the 32nd International Conference on Machine Learning*, eds. Bach, F. and Blei, D., Lille, France: PMLR, vol. 37 of *Proceedings of Machine Learning Research*, pp. 1481–1490.

Diggle, P. and Lophaven, S. (2006), "Bayesian geostatistical design," *Scandinavian Journal of Statistics*, 33, 53–64.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. (2008), "Relative-error CUR matrix decompositions," *SIAM Journal on Matrix Analysis and Applications*, 30, 844–881.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014), "Estimation and prediction in spatial models with block composite likelihoods," *Journal of Computational and Graphical Statistics*, 23, 295–315.

Finley, A., Sang, H., Banerjee, S., and Gelfand, A. (2009), "Improving the performance of predictive process modeling for large datasets," *Computational Statistics & Data Analysis*, 53, 2873–2884.

Furrer, R., Genton, M. G., and Nychka, D. (2006), "Covariance tapering for interpolation of large spatial datasets," *Journal of Computational and Graphical Statistics*, 15, 502–523.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010), *Handbook of spatial statistics (Chapman & Hall CRC Handbooks of Modern Statistical Methods)*, Chapman & Hall CRC Handbooks of Modern Statistical Methods, Taylor and Francis.

Gittens, A. and Mahoney, M. (2013), "Revisiting the Nyström method for improved large-scale machine learning," in *Proceedings of the 30th International Conference on Machine Learning*, eds. Dasgupta, S. and McAllester, D., Atlanta, Georgia, USA: PMLR, vol. 28 of *Proceedings of Machine Learning Research*, pp. 567–575.

Gu, C. (2013), *Smoothing Spline ANOVA Models*, vol. 297, Springer Science & Business Media.

Guinness, J., Katzfuss, M., and Fahmy, Y. (2021), *GpGp: fast Gaussian process computation using Vecchia's approximation*, R package version 0.4.0.

Heaton, M. J., Datta, A., Finley, A., Furrer, R., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., et al. (2019), "A case study competition among methods for analyzing large spatial data," *Journal of Agricultural, Biological and Environmental Statistics*, 24, 398–425.

Henderson, H. V. and Searle, S. R. (1981), "On deriving the inverse of a sum of matrices," *SIAM Review*, 23, 53–60.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013), "Gaussian processes for big data," in *Procaddings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*.

Huang, H. and Sun, Y. (2018), "Hierarchical low rank approximation of likelihoods for large spatial datasets," *Journal of Computational and Graphical Statistics*, 27, 110–118.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018), "Gaussian processes and kernel methods: A review on connections and equivalences," *ArXiv*, abs/1807.02582.

Katzfuss, M. (2017), "A multi-resolution approximation for massive spatial datasets," *Journal of the American Statistical Association*, 112, 201–214.

Katzfuss, M. and Guinness, J. (2021), "A general framework for Vecchia approximations of Gaussian processes," *Statistical Science*, 36, 124–141.

Katzfuss, M., Guinness, J., Gong, W., and Zilber, D. (2020), "Vecchia approximations of Gaussian-process predictions," *Journal of Agricultural, Biological and Environmental Statistics*, 25, 383–414.

Katzfuss, M. and Hammerling, D. (2017), "Parallel inference for massive distributed spatial data using low-rank models," *Statistics and Computing*, 27, 363–375.

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

Kulesza, A., Taskar, B., et al. (2012), "Determinantal point processes for machine learning," *Foundations and Trends® in Machine Learning*, 5, 123–286.

Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P. (2013), "A resampling-based stochastic approximation method for analysis of large geostatistical data," *Journal of the American Statistical Association*, 108, 325–339.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. (2020), "When Gaussian process meets big data: a review of scalable GPs," *IEEE Transactions on Neural Networks and Learning Systems*, 31, 4405–4423.

Ma, P., Huang, J. Z., and Zhang, N. (2015), "Efficient computation of smoothing splines via adaptive basis sampling," *Biometrika*, 102, 631–645.

Mahoney, M. W. (2011), "Randomized algorithms for matrices and data," *Foundations and Trends in Machine Learning*, 3, 123–224.

Mak, S. (2021), *support: Support Points*, R package version 0.1.5.

Mak, S. and Joseph, V. R. (2018), "Support points," *The Annals of Statistics*, 46, 2562–592.

Meng, C., Zhang, X., Zhang, J., Zhong, W., and Ma, P. (2020), "More efficient approximation of smoothing splines via space-filling basis selection," *Biometrika*, 107, 723–735.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015), "A multiresolution Gaussian Process model for the analysis of large spatial datasets," *Journal of Computational and Graphical Statistics*, 24, 579–599.

Nychka, D., Hammerling, D., Sain, S., and Lenssen, N. (2016), "LatticeKrig: multiresolution kriging based on Markov random fields," R package version 8.4.

Nychka, D. and Saltzman, N. (1998), *Design of Air-Quality Monitoring Networks*, New York, NY: Springer US, pp. 51–76.

Paciorek, C. J., Lipshitz, B., Zhuo, W., Prabhat, ., Kaufman, C. G. G., and Thomas, R. C. (2015), "Parallelizing Gaussian process calculations in R," *Journal of Statistical Software*, 63, 1–23.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C. E. and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, The MIT Press.

Rizzo, M. L. (2004), "Testing for equal distributions in high dimension," *InterStat*, 5, 1–6.

Rudi, A., Camoriano, R., and Rosasco, L. (2015), "Less is more: Nyström computational regularization," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, MIT Press, NIPS'15, pp. 1657–1665.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall/CRC, 1st Edition.

Rue, H. and Tjelmeland, H. (2002), "Fitting Gaussian Markov random fields to Gaussian fields," *Scandinavian Journal of Statistics*, 29, 31–49.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and analysis of computer experiments," *Statistical Science*, 4, 409–423.

Sang, H. and Huang, J. Z. (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.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, Spring Science & Business Media.

Stein, M., Chi, Z., and Welty, L. (2004), "Approximating likelihoods for large spatial data sets," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 66, 275–296.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer.

— (2013), "Statistical properties of covariance tapers," *Journal of Computational and Graphical Statistics*, 22, 866–885.

Székely, G. J. and Rizzo, M. L. (2013), "Energy statistics: A class of statistics based on distances," *Journal of Statistical Planning and Inference*, 143, 1249–1272.

Titsias, M. (2009), "Variational learning of inducing variables in sparse Gaussian processes," in *Artificial intelligence and statistics*, PMLR, pp. 567–574.

Tuo, R. and Wang, W. (2020), "Kriging prediction with isotropic Matérn correlations: robustness and experimental designs," *Journal of Machine Learning Research*, 21, 1–38.

Tzeng, S. and Huang, H.-C. (2018), "Resolution adaptive fixed rank Kriging," *Technometrics*, 60, 198–208.

Tzeng, S., Huang, H.-C., Wang, W.-T., Nychka, D., and Gillespie, C. (2021), *autoFRK: automatic fixed rank kriging*, R package version 1.4.3.

Varin, C., Reid, N. M., and Firth, D. (2011), "An overview of composite likelihood methods," *Statistica Sinica*, 21, 5–42.

Vecchia, A. V. (1988), "Estimation and model identification for continuous spatial processes," *Journal of the Royal Statistical Society: Series B (Methodological)*, 50, 297–312.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM.

Wang, W. and Jing, B.-Y. (2022), "Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression," *Journal of Machine Learning Research*, 23, 1–67.

Xia, G., Miranda, M., and Gelfand, A. (2006), "Approximately optimal spatial design approaches for environmental health data," *Environmetrics*, 17, 363–385.

Xu, G., Liang, F., and Genton, M. G. (2015), "A Bayesian spatio-temporal geostatistical model with an auxiliary lattice for large datasets," *Statistica Sinica*, 25, 61–79.

Zammit-Mangion, A. and Cressie, N. (2021), "FRK: An R package for spatial and spatio-temporal prediction with large datasets," *Journal of Statistical Software*, 98, 1–48.

Zhang, H. (2004), "Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics," *Journal of the American Statistical Association*, 99, 250–261.

Zhu, Z. and Stein, M. L. (2005), "Spatial sampling design for parameter estimation of the covariance function," *Journal of Statistical Planning and Inference*, 134, 583–603.

— (2006), "Spatial sampling design for prediction with estimated parameters," *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 24–44.