Data Sketching and Stacking: A Confluence of Two Strategies for Predictive Inference in Gaussian Process Regressions with High-Dimensional Features

Samuel Gailliot
Department of Statistics, Texas A&M University
and
Rajarshi Guhaniyogi
Department of Statistics, Texas A&M University
and
Roger D. Peng
Department of Statistics and Data Sciences, University of Texas at Austin.

January 22, 2024

Abstract

This article focuses on drawing computationally-efficient predictive inference from Gaussian process (GP) regressions with a large number of features when the response is conditionally independent of the features given the projection to a noisy low dimensional manifold. Bayesian estimation of the regression relationship using Markov Chain Monte Carlo and subsequent predictive inference is computationally prohibitive and may lead to inferential inaccuracies since accurate variable selection is essentially impossible in such high-dimensional GP regressions. As an alternative, this article proposes a strategy to sketch the high-dimensional feature vector with a carefully constructed sketching matrix, before fitting a GP with the scalar outcome and the sketched feature vector to draw predictive inference. The analysis is performed in parallel with many different sketching matrices and smoothing parameters in different processors, and the predictive inferences are combined using *Bayesian predictive stacking.* Since posterior predictive distribution in each processor is analytically tractable, the algorithm allows bypassing the robustness issues due to convergence and mixing of MCMC chains, leading to fast implementation with very large number of features. Simulation studies show superior performance of the proposed approach with a wide variety of competitors. The approach outperforms competitors in drawing point prediction with predictive uncertainties of outdoor air pollution from satellite images.

Keywords: Bayesian predictive stacking; feature sketching; Gaussian processes; high-dimensional features; manifold regression; posterior consistency.

1 Introduction

We focus on the problem of drawing predictive inference of a random variable from a high-dimensional feature vector using "sketching" of the feature vector when it truly lies on a low-dimensional noisy unknown manifold. In recent years, there has been a growing literature on "data sketching," which involves sketching or compressing the original data before analysis (Halko et al., 2011; Mahoney et al., 2011; Woodruff et al., 2014; Guhaniyogi and Dunson, 2015, 2016). However, our approach differs from the existing data sketching literature in two key aspects. Firstly, while most data sketching approaches aim to reduce the number of data samples, our approach is distinct in that it maintains the same number of samples but instead reduces the dimensionality of the feature vector. Secondly, the majority of research in data sketching focuses on performance evaluation of ordinary and high-dimensional penalized regression methods with sketched data (Zhang et al., 2013; Dobriban and Liu, 2018; Drineas et al., 2011; Ahfock et al., 2017; Huang, 2018), with only a few recent articles considering application of data sketching in Bayesian high-dimensional linear and non-linear regressions (Guhaniyogi and Scheffler, 2021; Guhaniyogi and Dunson, 2015, 2016). In contrast, our approach leverages the benefits of data sketching to deliver scalable predictive inference in non-parametric regressions with a limited sample size and a large number of features, when the features lie on a noisy low-dimensional manifold.

We consider a regression framework with an outcome $y \in \mathcal{Y} \subseteq \mathbb{R}$ and a feature vector $\boldsymbol{x} = (x_1, ..., x_p)^T$ when $\boldsymbol{x}$ resides on a noisy unknown manifold, i.e., $\boldsymbol{x} = \boldsymbol{\phi}(\boldsymbol{o}) + \boldsymbol{\eta}$, where $\boldsymbol{o} = (o_1, ..., o_d)^T$ is d-dimensional co-ordinates for a manifold $\mathcal{O} \subseteq \mathbb{R}^p$, $\boldsymbol{\phi}(\cdot) : \mathbb{R}^d \to \mathbb{R}^p$ is a mapping function such that $\boldsymbol{\phi}(\boldsymbol{o}) \in \mathcal{O}$ and $\boldsymbol{\eta}$ is p-dimensional noise. Often the complex dependence between y and $\boldsymbol{x}$ is encoded via co-ordinates of the low-dimensional manifold, i.e.,

$$y = h(\boldsymbol{o}) + \epsilon, \tag{1}$$

where h is a complex function encoding the true relationship between response and co-ordinates of the manifold and ϵ is the error. Since the manifold $\mathcal{O}$ is unobserved, the co-ordinates $\boldsymbol{o}$ is typically unknown. Hence, the common practice is to estimate complex dependencies between y and $\boldsymbol{x}$

through a non-linear regression model given by,

$$y = f(\boldsymbol{x}) + \epsilon, \tag{2}$$

where f is an unknown regression function and ϵ is the residual. When dealing with high-dimensional features, Gaussian process (GP) priors with an automatic relevance determination (ARD) kernel are commonly used to estimate the underlying function f with sufficient sparsity assumption in the relationship between y and $\boldsymbol{x}$ (Zhao et al., 2018; Jensen et al., 2021). The estimated f is then employed to predict the response variable. However, when the number of features reaches the order of a few thousand, estimation of f with GP-ARD framework is often inaccurate, leading to unsatisfactory predictive inference. This article proposes an alternative approach that exclusively focuses on drawing predictive inference on the response variable y, including both point prediction and uncertainty estimation, using GP regression. We review below a list of existing strategies to draw predictive inference on y in non-linear regressions before introducing our approach.

In the literature, a significant line of work follows a two-stage approach for dealing with high-dimensional features in non-linear manifold regression tasks. In this approach, the first stage involves constructing a lower-dimensional representation of the high-dimensional features using manifold learning techniques. Some popularly employed parsimonious manifold learning algorithms include Isomap (Tenenbaum et al., 2000), Diffusion Maps (Coifman and Lafon, 2006), and Laplacian eigenmap (Belkin and Niyogi, 2003). These algorithms enable the reduction of dimensionality while preserving the essential characteristics of the data. Additionally, there are model-based approaches that estimate the unknown Riemannian manifold structure within the feature vector. These methods utilize techniques such as local PCA (Weingessel and Hornik, 2000; Arias-Castro et al., 2017) or geometric multiresolution analysis (Maggioni et al., 2016), and, more recently, spherical basis functions (Li et al., 2022). Non-linear regression models in the second stage are based on these projected features in lower-dimensions. However, it is important to note that such two-stage approaches rely on learning the manifold structure embedded in the high-dimensional features. While this can be valuable for understanding the underlying data structure, it adds unnecessary computational burden when the primary focus is on prediction rather than inference.

An alternative line of research focuses on estimating the unknown function f using tree-based approaches or deep neural network methods. Tree-based approaches, such as CART (Denison et al., 1998), BART (Chipman et al., 2010), and random forest (Breiman, 2001) are based on finding the best splitting attribute, which can become less efficient as the number of features (p) increases. While there is a growing literature on variable selection within tree-based methods, such as BART (Bleich et al., 2014) and its variants (Liu et al., 2021), estimating the true regression function with a large number of features (p of the order of thousands) can pose challenges. Deep neural networks with variable selection architecture (Dinh and Ho, 2020) are also not ideal as they lack predictive uncertainty and struggle to handle the high-dimensional feature space efficiently.

Bayesian modeling approaches are naturally appealing when the focus is on quantifying predictive uncertainty. To this end, the more traditional Bayesian models simultaneously learn the mapping to the lower-dimensional subspace along with the regression function in the coordinates on this subspace. These approaches range from Gaussian process latent variable models (GP-LVMs) (Lawrence and Hyvärinen, 2005; Titsias and Lawrence, 2010) for probabilistic nonlinear principle component analysis to mixture of factor models (Chen et al., 2010). However, such methods pose daunting computational challenges with even moderately large p and sample size due to learning the number and distribution of latent variables, as well as the mapping functions, while maintaining identifiability restrictions.

To enhance the time efficiency of the aforementioned approaches, pre-processing steps are often employed, and two popular pre-processing methods are feature screening and projection. The feature screening approach identifies features that exhibit the strongest marginal association with the response variable. By selecting the features with the highest marginal association, this approach aims to reduce the dimensionality of the problem. Feature screening methods are generally straightforward to implement, and it offers asymptotic guarantees of selecting a superset of important features (Chen et al., 2018). On the other hand, projection approaches aim to construct lower-dimensional feature vectors by combining the original high-dimensional features. One common method in projection approaches is to construct a few principal components (PCs) from the original p-dimensional feature vector.

A naive implementation of the above pre-processing steps is unappealing to our scenario. For example, in a non-parametric regression with a large number of correlated features and low signal-to-noise ratio, it may be important to choose a conservative threshold for screening, which limits the scope of dimension reduction at this stage. On the other hand, construction of PCs are agnostic to the relationship between the response and the feature vector. Instead, we propose an approach that first employs variable screening (Chen et al., 2018) with a conservative threshold to identify a large subset of features, typically a few thousand, having the highest non-linear marginal association with the response. After variable screening, the screened feature vector is further compressed using a short and fat random sketching matrix (Mahoney et al., 2011; Drineas et al., 2012). This matrix has a small number of rows (m) and entries that are independently and identically drawn from a normal distribution. Predictive inference proceeds by fitting a non-parametric Gaussian process (GP) regression model (Williams and Rasmussen, 2006; Gramacy, 2020) to the scalar outcome and the m-dimensional sketched feature vector after fixing values for the weakly identifiable tuning parameters within the covariance kernel of the Gaussian processes. The posterior predictive distribution corresponding to a choice of such tuning parameters and random sketching matrix comes in a closed form without the need to implement MCMC sampling, so that one can obtain the predictive distribution extremely rapidly even in problems with huge numbers of features. To reduce the sensitivity of predictive inference to the choice of the sketching matrix and the tuning parameters in GP regression, the model is fit with multiple different choices of the sketching matrix and parameters. The predictive inferences obtained from such choices are then aggregated using Bayesian predictive stacking (Yao et al., 2018) to improve accuracy and robustness.

Stacking is a model aggregation procedure to combine predictions from many different models (Wolpert, 1992; Breiman, 1996; LeBlanc and Tibshirani, 1996). In recent years, substantial advancements have been made in Bayesian stacking methodology, with notable contributions made in Le and Clarke (2017); Yao et al. (2018); Pavlyshenko (2020); Yao et al. (2022a,b) and the references therein. However, to the best of our knowledge, the application of stacking in the context of predictive inference for high-dimensional manifold regression is currently lacking. While Bayesian model averaging (Raftery et al., 1997) is most popularly used for aggregating predictive inference from multiple models, it may be less suited to the stacking procedure in our settings. To see this, assume that there are S candidate models $\mathcal{M} = \{\mathcal{M}_1, ..., \mathcal{M}_S\}$. Bayesian model comparison typically encounter three different settings: (i) $\mathcal{M}$-closed where a true data generating model exists and is included in $\mathcal{M}$; (ii) $\mathcal{M}$-complete where a true model exists but is not included in $\mathcal{M}$; and (iii) $\mathcal{M}$-open where we do not assume the existence of a true data generating model. Although Bayesian model averaging has the advantage of asymptotically identifying the true data generating model in the first setting, predictive stacking has advantages in the $\mathcal{M}$-complete and $\mathcal{M}$-open settings. Given that the true model may not be included in the class of fitted Gaussian process regression models with randomly sketched features, predictive stacking offers substantial advantages over model averaging.

In regressions involving high-dimensional features and a large sample size, Guhaniyogi and Scheffler (2021) propose an approach orthogonal to ours which exploits random sketching matrices to reduce the sample size rather than the number of features. A few approaches closely related to ours develop theoretical bound on predictive accuracy when high-dimensional features are sketched with random matrices (Guhaniyogi and Dunson, 2015, 2016; Thanei et al., 2017). These articles tend to include random linear combinations of many unimportant features, diminishing signal in the analysis, which results in less than satisfactory predictive performance with massive-dimensional features. Addressing this issue, Mukhopadhyay and Dunson (2020) proposes novel constructions of projection matrices tailored to deliver more accurate predictive inference. These approaches aim to overcome the challenges associated with sketching high-dimensional features and improve predictive performance. However, they primarily focus on high-dimensional parametric regression, and their applicability to non-parametric regression tasks may require further investigation. Additionally, these approaches address sensitivity to the choice of sketching matrices by aggregating predictive inference over many sketching matrices using Bayesian model averaging technique (Raftery et al., 1997) which is less suitable for prediction than the stacking approach we employ here, as discussed in the last paragraph.

The rest of the article proceeds as follows. Section 1.1 discusses motivating dataset on outdoor air pollution and satellite imagery. Section 2 proposes the model and computational approach for predictive inference in manifold regression with large number of predictors. Section 3 offers empirical evaluation of the proposed approach along with its competitors for simulation studies. Section 4 investigates the proposed approach in drawing predictive inference of outdoor air pollution concentration from satellite images. Finally, Section 5 concludes the article with an eye towards future work.

1.1 Outdoor Air Pollution Application

As a motivation for the development of our methodology, we consider the problem of predicting outdoor air pollution concentrations across the United States. Outdoor air pollution in the U.S. is measured using a network of ground-based monitors managed by local air quality agencies and the U.S. Environmental Protection Agency (Environmental Protection Agency, 1996) (EPA). While the combined network of monitors consists of thousands of locations, the spatial coverage of the network is actually quite sparse, leaving many areas of the country without any ground-level data (Apte et al., 2017). Many dense urban areas only contain one or two monitors, raising a question of whether such measurements are representative of the burden experienced by all members of the population. To address the sparsity of the network, there have been efforts to deploy low-cost sensors across urban areas to fill the gaps. While such approaches have promise, they are still experimental and ad hoc in nature, and the sensors themselves can sometimes introduce new measurement problems (Heffernan et al., 2023).

Remote sensing techniques, which use satellite imagery to predict ground-level concentrations of outdoor air pollution have the potential to address the spatial coverage problem because of their constant monitoring of the entire planet. Traditional approaches have employed aerosol optical depth as a proxy for such pollutants as fine particulate matter (Paciorek et al., 2008), or PM2.5. While the previous generation of Earth observation satellites had excellent spatial coverage, they lacked temporal coverage, typically revisiting an area of the planet only once every one or two weeks. In addition, older satellites tended to have lower resolution, making them difficult to use for predicting air pollution concentrations in dense urban settings. In recent years, there has been a revolution in the deployment of satellite constellations, where hundreds of smaller inexpensive satellites orbit the Earth, providing constant coverage of all areas (Planet Team, 2017). Furthermore, these satellites have much higher resolution, allowing for more detailed examination of areas of interest.

Given the recent emergence of data from satellite constellations, there is still a question of how best to use them for the purpose of predicting ground-level air pollution concentrations. For this application, we focus on predicting fine particulate matter pollution (PM2.5) from multi-band satellite images. For ground-truth information we use the EPA's network of monitors to provide valid PM2.5 measurements. The combination of high-resolution spatial and temporal coverage of the entire U.S. with novel statistical prediction approaches has the potential to dramatically increase the monitoring of outdoor air pollution and its subsequent health effects.

2 Our Approach: Stacked Gaussian Process Regression

Let $\mathcal{D}_n = \{(\boldsymbol{x}_i^T, y_i) : i = 1, ..., n\}$ be a dataset containing n observations each with a p-variate feature $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$ and a scalar-valued response y_i. We assume n is moderately large and p is large. The feature vector $\boldsymbol{x}_i$ lies on an unknown noisy manifold $\mathcal{O} \subseteq \mathbb{R}^p$ with d-dimensional latent co-ordinates $\boldsymbol{o}_i$ (i.e., $\boldsymbol{x}_i = \boldsymbol{\phi}(\boldsymbol{o}_i) + \boldsymbol{\eta}_i$, $\boldsymbol{\phi}(\boldsymbol{o}_i) \in \mathcal{O}$). We assume a nonlinear regression relationship between y_i and $\boldsymbol{x}_i$, and approximate the density of y_i by sketching the high-dimensional feature vector $\boldsymbol{x}_i$ to lower-dimensions using a sketching matrix $\boldsymbol{P}_n$ as follows

$$y_i = f(\boldsymbol{P}_n \boldsymbol{x}_i) + \epsilon_i, \;\; \epsilon_i \overset{i.i.d.}{\sim} N(0, \xi^2), \tag{3}$$

with ξ^2 as the noise variance and $f(\cdot)$ as an unknown continuous function in the Holder class of smoothness s. Discussion on the choice of the sketching matrix $\boldsymbol{P}_n \in \mathbb{R}^{m \times p}$ is provided in Section 2.1.

2.1 Choice of the Sketching Matrix

The sketching matrix $\boldsymbol{P}_n \in \mathbb{R}^{m \times p}$ embeds p-dimensional features $\boldsymbol{x}_i$ into m dimensions while not throwing away excessive amounts of information. The most popular linear embedding is obtained from the singular value decomposition (SVD) of $\boldsymbol{X} = [\boldsymbol{x}_1 : \cdots : \boldsymbol{x}_n]^T$, but they are problematic to estimate when $p >> n$. In contrast, random sketching matrices are often used to embed the high-dimensional features to a random subspace, and appropriate choices of the random matrices allow distances between samples to be approximately preserved (Li and Gu, 2017).

The direct application of sketching matrices on high-dimensional features is unappealing, as it constructs random linear combinations of many unimportant features, diminishing the signal in the analysis. As an alternative approach, we design a sketching matrix that constructs random linear combinations of features with the highest marginal association with the response. To identify these features, we perform nonparametric B-spline regression of y_i onto each component of $\boldsymbol{x}_i$ separately. The order of importance of the features is determined, in descending order, by the residual sum of squares of the marginal nonparametric regressions. Features with a residual sum of squares greater than a user-defined threshold are considered important features related to the response. We adopt a conservative threshold following Fan et al. (2014) to select a large superset of important features, which allows for joint contributions of features in explaining the response.

Let $\mathcal{I}$ correspond to the indices of the features chosen with marginal screening and $\bar{\mathcal{I}}$ be the indices of the features screened out through this procedure, such that $\mathcal{I} \cup \bar{\mathcal{I}} = \{1, ..., p\}$. Let $\boldsymbol{E}_n$ be a permutation matrix such that $\boldsymbol{E}_n \boldsymbol{x} = (\boldsymbol{x}_{\mathcal{I}}^T, \boldsymbol{x}_{\bar{\mathcal{I}}}^T)^T$. We construct a matrix $\boldsymbol{R}_n = [\boldsymbol{R}_{n,1} : \boldsymbol{R}_{n,2}]$ where $\boldsymbol{R}_{n,2} = \mathbf{0}_{m \times (p - |\mathcal{I}|)}$ and $\boldsymbol{R}_{n,1}$ is an $m \times |\mathcal{I}|$ matrix with entries drawn independently from N(0,1), following the literature on random sketching matrices (Baraniuk et al., 2008). The resulting sketching matrix is given by $\boldsymbol{P}_n = \boldsymbol{R}_n \boldsymbol{E}_n$.

2.2 Prior, Posterior and Posterior Predictive Distributions

Following a Bayesian approach, we assign a zero-centered Gaussian process prior on the unknown regression function $f(\cdot)$, denoted by $f(\cdot) \sim GP(0, \sigma^2\delta_\theta)$. Here, δ_θ corresponds to an exponential correlation kernel $\delta_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\theta||\boldsymbol{x}_i - \boldsymbol{x}_j||)$ involving the length-scale parameter θ. The parameter σ^2 is the signal variance parameter and $||\cdot||$ denotes the Euclidean norm. A significant finding by Yang and Dunson (2016) establishes that when features $\boldsymbol{x}_i$ lie on a d-dimensional manifold $\mathcal{O}$, the minimax optimal rate of $n^{-2s/(2s+d)}$ (adapted to the dimension of the manifold) can be achieved in estimating f through an appropriate choice of prior distributions on θ and σ^2. However, in practical scenarios, features may not exactly lie on a manifold due to noise and data corruption, as assumed in our setting. In such instances, the application of random compression, denoted as $\boldsymbol{P}_n\boldsymbol{x}_i$, aids in denoising the features. The de-noised compressed features $\boldsymbol{P}_n\boldsymbol{x}_i$ exhibit a higher concentration around the manifold compared to the original features $\boldsymbol{x}_i$. With this enhanced concentration, the theory presented by Yang and Dunson (2016) suggests that an appropriate GP prior can yield excellent performance. In addition to denoising, the compression of the high-dimensional feature vector has a major advantage in avoiding the estimation of a geodesic distance along the unknown manifold $\mathcal{O}$ between any two feature vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$.

In practical applications, utilizing the recommended prior distributions on θ and σ^2 from Yang and Dunson (2016) requires computationally expensive Markov Chain Monte Carlo (MCMC) sampling, mainly due to the weak identifiability of θ. The posterior computation of θ typically entails meticulous tuning, especially when dealing with high-dimensional features, imposing a significant computational burden. This article introduces an alternative approach that enables exact predictive inference from the model, entirely bypassing the MCMC algorithm in model estimation. The details of the strategy are elaborated below.

Denote $\boldsymbol{f} = (f(\boldsymbol{P}_n\boldsymbol{x}_1), ..., f(\boldsymbol{P}_n\boldsymbol{x}_n))^T$ as the vector consisting of the function f evaluated at the sketched features $\boldsymbol{P}_n\boldsymbol{x}_1$,...,$\boldsymbol{P}_n\boldsymbol{x}_n$ and $\boldsymbol{C}$ as an $n \times n$ covariance matrix with (i, j)th entry $\delta_\theta(\boldsymbol{P}_n\boldsymbol{x}_i, \boldsymbol{P}_n\boldsymbol{x}_j)$. With $\boldsymbol{y} = (y_1, ..., y_n)^T$ as the response vector, a customary Bayesian hierarchical model is constructed as

$$\boldsymbol{y}|\boldsymbol{f},\xi^2 \sim N(\boldsymbol{f},\xi^2\boldsymbol{I}), \quad (\boldsymbol{f}|\xi^2) \sim N(0,\xi^2\psi^2\boldsymbol{C}), \quad \pi(\xi^2) \propto \frac{1}{\xi^2},$$

where we fix the length-scale parameter θ and the signal-to-noise variance ratio $\psi^2 = \frac{\sigma^2}{\xi^2}$. This ensures closed-form conjugate marginal posterior and posterior predictive distributions. More specifically, the marginal posterior distribution of ξ^2, given the projection matrix $\boldsymbol{P}_n$, θ, ψ^2 and $\mathcal{D}_n$, is inverse gamma with parameters $a = n/2$ and $b = \boldsymbol{y}^T(\psi^2\boldsymbol{C}+\boldsymbol{I})^{-1}\boldsymbol{y}/2$. The marginal posterior distribution of $\boldsymbol{f}$, given $\boldsymbol{P}_n$, ψ^2, θ and $\mathcal{D}_n$, follows a scaled n-variate t distribution with degrees of freedom n, location $\boldsymbol{\mu}_t$ and scale matrix $\boldsymbol{\Sigma}_t$, denoted by $t_n(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, where $\boldsymbol{\mu}_t = (\boldsymbol{I} + \boldsymbol{C}^{-1}/\psi^2)^{-1}\boldsymbol{y}$, $\boldsymbol{\Sigma}_t = (2b/n)(\boldsymbol{I}+\boldsymbol{C}^{-1}/\psi^2)^{-1}$. Consider prediction for the response at n_{new} data points with corresponding covariates $\tilde{\boldsymbol{x}}_1, ..., \tilde{\boldsymbol{x}}_{n_{new}}$. Let $\boldsymbol{C}_{new,new}$ and $\boldsymbol{C}_{new,old}$ denote $n_{new} \times n_{new}$ and $n_{new} \times n$ matrices with (i,j)th elements $\delta_\theta(\boldsymbol{P}_n\tilde{\boldsymbol{x}}_i, \boldsymbol{P}_n\tilde{\boldsymbol{x}}_j)$ and $\delta_\theta(\boldsymbol{P}_n\tilde{\boldsymbol{x}}_i, \boldsymbol{P}_n\boldsymbol{x}_j)$, respectively. The posterior predictive distribution of the response $\tilde{\boldsymbol{y}}_{new} = (\tilde{y}_1, ..., \tilde{y}_{n_{new}})^T$ given $\tilde{\boldsymbol{x}}_1, ..., \tilde{\boldsymbol{x}}_{n_{new}}$, $\boldsymbol{P}_n$, θ, ψ^2 and $\mathcal{D}_n$, marginalizing out $(\boldsymbol{f}, \xi^2)$, follows a scaled n_{new}-variate t-distribution $t_{n_{new}}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$, where

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_t &= \psi^2\boldsymbol{C}_{new,old}(\boldsymbol{I} + \psi^2\boldsymbol{C})^{-1}\boldsymbol{y} \\
\tilde{\boldsymbol{\Sigma}}_t &= (2b/n)\left[\boldsymbol{I} + \psi^2\boldsymbol{C}_{new,new} - \psi^4\boldsymbol{C}_{new,old}(\boldsymbol{I} + \psi^2\boldsymbol{C})^{-1}\boldsymbol{C}_{new,old}^T\right].
\end{aligned} \tag{4}$$

Since the posterior predictive distribution is available in closed form, Bayesian inference can proceed from exact posterior samples.

This tractability is only possible if the length-scale parameter θ and the signal-to-noise variance ratio ψ^2 are fixed. While it is possible to estimate their full posterior distributions through expensive Markov Chain Monte Carlo (MCMC) sampling, these parameters are inconsistently estimable for the general Matern class of correlation functions (Zhang, 2004) often resulting in poorer convergence. Therefore, for the chosen sketching matrix $\boldsymbol{P}_n$, we obtain (θ, ψ^2) such that

$$\max_{\theta,\psi^2} f(\theta, \psi^2 | \boldsymbol{P}_n, \boldsymbol{y}) \propto \max_{\theta,\psi^2} \frac{1}{|\psi^2 \boldsymbol{C} + \boldsymbol{I}|^{\frac{1}{2}}} \frac{2^{\frac{n}{2}} \Gamma(\frac{n}{2})}{[\boldsymbol{y}'(\psi^2 \boldsymbol{C} + \boldsymbol{I})^{-1} \boldsymbol{y}]^{\frac{n}{2}} (\sqrt{2\pi})^n} \tag{5}$$

Our approach will conduct exact predictive inference using the closed form predictive distribution in (4) and stack the predictive inference over different fixed values of $\{\boldsymbol{P}_n, \theta, \psi^2\}$.

2.2.1 Stacking of Predictive Distributions

Let $\mathcal{M}_k$ represent the fitted model (3) with $\boldsymbol{P}_n^{(k)}, \theta^{(k)}, \psi^{2(k)}$ for $k = 1, ..., K$. While the sketching matrix $\boldsymbol{P}_n^{(k)}$ is randomly generated for each k, $\theta^{(k)}$ and $\psi^{2(k)}$ are obtained following equation (5) for the choice of $\boldsymbol{P}_n^{(k)}$. Employing the generalized Bayesian stacking framework proposed by Yao et al. (2018), we implement a stacking procedure over the predictive distribution obtained from each $\mathcal{M}_k$. Let $p(\tilde{\boldsymbol{y}}_{new} | \boldsymbol{y}, \mathcal{M}_k)$ denote the predictive distribution under model $\mathcal{M}_k$, and $p_t(\tilde{\boldsymbol{y}}_{new} | \boldsymbol{y})$ denote the true predictive distribution. Our objective is to determine the distribution in the convex hull $\mathcal{C} = \{\sum_{k=1}^K w_k p(\cdot | \boldsymbol{y}, \mathcal{M}_k) : w_k \in \mathcal{S}_1^K\}$, where $\mathcal{S}_1^K = \{\boldsymbol{w} \in [0,1]^K : \sum_{k=1}^K w_k = 1\}$, that is optimal with respect to some proper scoring function. Using the logarithmic score, which corresponds to the KL divergence, we seek to find the vector $\tilde{\boldsymbol{w}} = (\tilde{w}_1, \ldots, \tilde{w}_K)$ such that

$$\tilde{\boldsymbol{w}} = \max_{\boldsymbol{w} \in \mathcal{S}_1^K} \frac{1}{n} \sum_{i=1}^n \log \left(\sum_{k=1}^K w_k p_{k,-i}(y_i) \right), \tag{6}$$

where $\boldsymbol{y}_{-i} = (y_j : j \neq i,\ j = 1, ..., n)^T$ and $p_{k,-i}(y_i) = p(y_i | \boldsymbol{y}_{-i}, \mathcal{M}_k)$ has a closed from univariate t-distribution with parameters $\tilde{\mu}_{-i}^{(k)}$ and $\tilde{\Sigma}_{-i}^{(k)}$ obtained using the formula for posterior predictive distribution given in equation (4). In practice, calculating the predictive densities $p_{k,-i}(y_i)$ one at a time is computationally expensive as the calculation of $\tilde{\Sigma}_{-i}^{(k)}$ requires inverting an $(n-1) \times (n-1)$ matrix for every $k = 1, ..., K$ and $i = 1, ..., n$. To avoid this, we randomly split the data into $S = 10$ disjoint folds of approximately equal size, $(\boldsymbol{y}_{(1)}, \boldsymbol{X}_{(1)}), \ldots, (\boldsymbol{y}_{(S)}, \boldsymbol{X}_{(S)})$, and compute $\boldsymbol{y}_{(s)} | \boldsymbol{y}_{(1)}, ..., \boldsymbol{y}_{(s-1)}, \boldsymbol{y}_{(s+1)}, ..., \boldsymbol{y}_{(S)}$ for every $s = 1, ..., S$, which follows a multivariate t-distribution with parameters obtained using equation (4). If the ith sample belongs to the sth fold, we will replace $p_{k,-i}(y_i)$ in (6) by $p_{k,(s)}(y_i)$, where $p_{k,(s)}(y_i)$ represents the marginal distribution of y_i from $\boldsymbol{y}_{(s)}|\boldsymbol{y}_{(1)},...,\boldsymbol{y}_{(s-1)},\boldsymbol{y}_{(s+1)},...,\boldsymbol{y}_{(S)}$. This strategy requires inverting an $(n-n/S)\times(n-n/S)$ matrix only S times (assuming that all folds are of equal size), which leads to substantial computational benefits. No analytical solution to this non-convex constrained optimization problem in (6) is available, but first and second derivatives are easily obtained to construct an iterative optimizer. The optimal distribution provides a pseudo posterior predictive distribution given by $\tilde{p}(\tilde{\boldsymbol{y}}_{new}|\boldsymbol{y}) = \sum_{k=1}^{K}\tilde{w}_k t_{n_{new}}(\tilde{\boldsymbol{\mu}}_t^{(k)},\tilde{\boldsymbol{\Sigma}}_t^{(k)})$, where $\tilde{\boldsymbol{\mu}}_t^{(k)}$ and $\tilde{\boldsymbol{\Sigma}}_t^{(k)}$ are obtained from equation (4) by evaluating $\tilde{\boldsymbol{\mu}}_t$ and $\tilde{\boldsymbol{\Sigma}}_t$ at $\boldsymbol{P}_n^{(k)},\theta^{(k)},\psi^{(k)}$. The pseudo posterior predictive distribution is further used to draw point prediction and 95% predictive interval to quantify predictive uncertainty. Figure 1 offers a flowchart outlining the proposed framework.

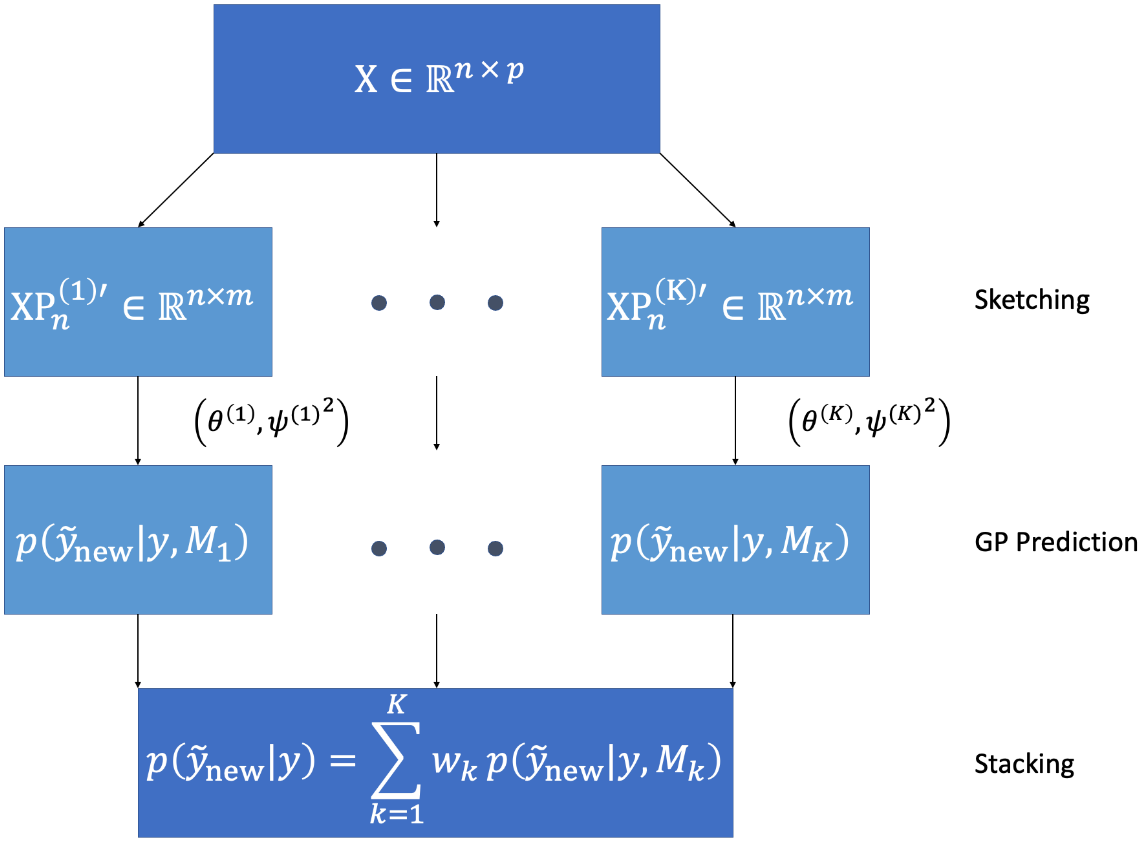

$X \in \mathbb{R}^{n\times p}$
$XP_n^{(1)\prime} \in \mathbb{R}^{n\times m}$
$XP_n^{(K)\prime} \in \mathbb{R}^{n\times m}$
Sketching
$\left(\theta^{(1)},\psi^{(1)^2}\right)$
$\left(\theta^{(K)},\psi^{(K)^2}\right)$
$p(\tilde{y}_{\text{new}}|y,M_1)$
$p(\tilde{y}_{\text{new}}|y,M_K)$
GP Prediction
$p(\tilde{y}_{\text{new}}|y) = \sum_{k=1}^{K} w_k\, p(\tilde{y}_{\text{new}}|y,M_k)$
Stacking

Figure 1: Flowchart representing the Sketched Gaussian process regression framework.

While Bayesian model averaging (BMA) is a common method for combining multiple distributions, its applicability in our context is limited for several reasons. Firstly, the fitted models (3) with randomly sketched features are likely to deviate from the true model, placing us outside the $\mathcal{M}$-closed setting where BMA is optimal. Additionally, stacking is designed to determine weights for optimal prediction, whereas asymptotically, BMA assigns full weight to the "best" single model closest in KL divergence to the true model (Yao et al., 2022a). However, when the true model lies outside the space of fitted models, it may be more advantageous to leverage multiple models in predictive inference. Subsequent empirical experiments demonstrate stacking as a powerful tool for drawing posterior predictive inference in our setting.

3 Simulation Study

We evaluate the performance of the proposed Sketched Gaussian Process (SkGP) regression across various simulation scenarios, exploring different structure of the manifold ($\mathcal{O}$), different feature dimensions (p) and noise levels in the features (τ^2) to analyze their impact. In all simulations, the out-of-sample predictive performance of the proposed SkGP regression is compared with that of uncompressed Gaussian Process (GP), Bayesian Additive Regression Trees (BART) (Chipman et al., 2010), Random Forests (RF) (Breiman, 2001), and deep neural network (NN). We also explore sketched versions of BART and RF, referred to as Sketched BART (SkBART) and Sketched Random Forest (SkRF), respectively, where a single projection matrix is generated to sketch the features, allowing for faster implementation. Each of these methods are applied on $|\mathcal{I}| = 1000$ screened features having highest marginal association with the response. As a default in this analysis, we set $m = 60$. We offer detailed sensitivity analysis with varying choices of the number of screened features $|\mathcal{I}|$ and the sketching dimensions m.

3.1 Simulated Data Generation

During data simulation, we explore specific scenarios where the response distribution follows a nonlinear function of d-dimensional coordinates for a manifold $\mathcal{O} \subseteq \mathbb{R}^p$, embedded in a high-dimensional ambient space. Two distinct choices for $\mathcal{O}$ and their corresponding response distributions are simulated.

$\mathcal{O}$ is a swiss roll and $d = 2$. For the swiss roll, we sample manifold coordinates, $o_1 \sim U(\frac{3\pi}{2}, \frac{9\pi}{2})$, $o_2 \sim U(0, 3)$. A high dimensional feature $\boldsymbol{x} = (x_1, \ldots, x_p)$ is simulated according to $x_1 = o_1 \cos(o_1) + \eta_1,\ x_2 = o_2 + \eta_2,\ x_3 = o_1 \sin(o_1) + \eta_3,\ x_i = \eta_i,\ i \geq 4$. The response y have a non-linear relationship with these features and is simulated following,

$$y = sin(5\pi o_1) + o_2^2 + \epsilon,\ \ \epsilon \sim N(0, 0.02^2), \tag{7}$$

where $\eta_1, \ldots, \eta_p \sim N(0, \tau^2)$. Notably, $\boldsymbol{x}$ and y are conditionally independent given o_1, o_2 which is the low-dimensional signal manifold. In particular, $\boldsymbol{x}$ lives on a (noise corrupted) swiss roll embedded in a p-dimensional ambient space (see Figure 2a), but y is only a function of coordinates along the swiss roll $\mathcal{O}$.

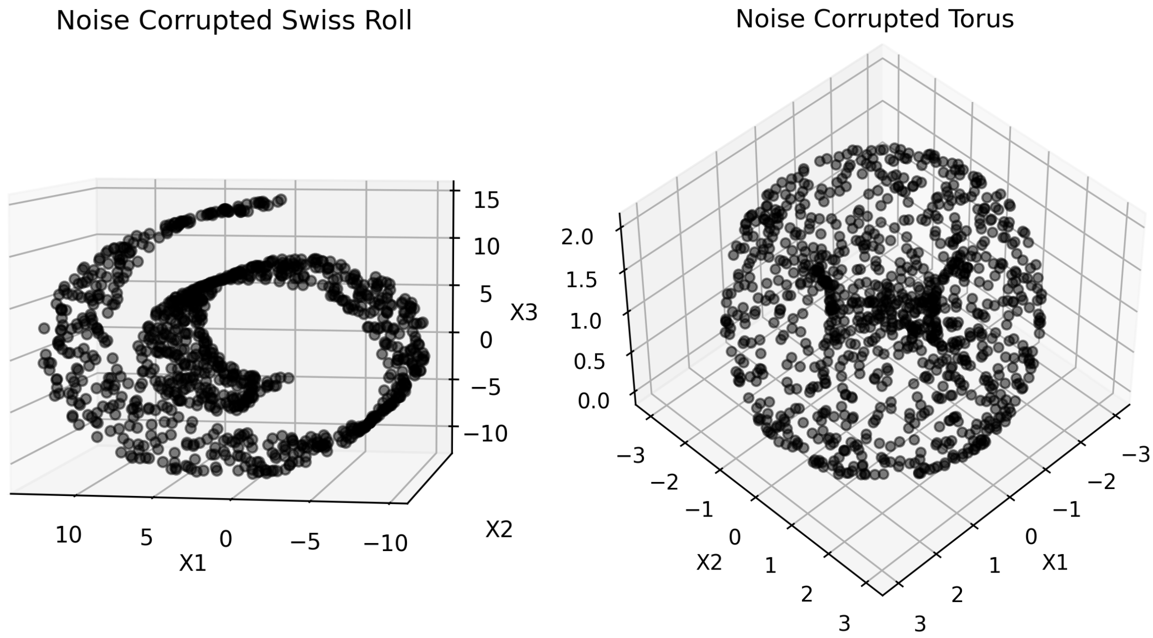
$\mathcal{O}$ is a torus and $d = 3$. For the torus, we consider $x_1 = o_1 + \eta_1, x_2 = o_2 + \eta_2$ and $x_3 = o_3 + \eta_3$ where o_1, o_2, o_3 lie on a three dimensional torus with interior radius 1 and exterior radius 3 (see Figure 2b), such that $(3 - \sqrt{o_1^2 + o_2^2})^2 + o_3^2 = 1$, and set $x_i = \eta_i$ for $i \geq 4$. The feature noise $\eta_1, ..., \eta_n$ are generated i.i.d. from $N(0, \tau^2)$. The response is generated as,

$$y = o_2^2 + \sin(5\pi o_3) + \epsilon,\ \epsilon \sim N(0, 0.1^2).$$

The geodesic distance between two points on both a swiss roll and a torus can substantially differ from their Euclidean distance in the ambient space $\mathbb{R}^p$. The swiss roll, in particular, poses a challenging setup for SkGP, as points on $\mathcal{O}$ that are close in a Euclidean sense can be quite far in a geodesic sense.

To assess the impact of the number of features (p) and noise levels of the features (τ^2) on the performance of the competitors, various simulation scenarios are considered by varying $p = 2000, 10000$ and $\tau^2 = 0.01, 0.03, 0.05, 0.1$. For each of these simulation scenarios, 50 datasets are generated, and metrics such as mean squared prediction error (MSPE), coverage, and lengths of

95% predictive intervals (PI) are calculated across all replicates. All simulations set the sample size $n = 100$ and the number of predicted samples $n_{new} = 100$.



(a)

(b)

Figure 2: Manifolds embedded in noisy ambient dimensions

3.2 Point prediction

Tables 1 and 2 display the MSPE averaged over 50 replications for all the competing methods in the swiss roll and torus examples, respectively. Values in parentheses represent the standard error of MSPE over 50 replicates.

Both Tables 1 and 2 show that incorporating randomly sketched features into the GP model within the SkGP framework yields strong predictive performance, significantly surpassing the performance of the neural network. For both $p = 2000$ and $p = 10000$, when the manifold is affected by low noise, SkGP significantly outperforms GP, BART, and RF with unsketched features. While SkBART emerges as the second-best performer in scenarios with very low noise in the manifold ($\tau^2 = 0.01$), its performance declines notably with an increase in the noise level in the features. In comparison, SkGP effectively mitigates the impact of noise in the features, but there exists a tipping point (depending on the structure of the underlying manifold $\mathcal{O}$ and sample size n) where noise distorts the manifold excessively, causing SkGP to perform similarly to other competitors. This is observed in the MSPE values corresponding to $\tau^2 = 0.1$. Among the sketched competitors, SkRF exhibits notably inferior performance compared to both SkGP and SkBART in all simulation examples. While theoretically the performance of SkGP should remain similar for both $p = 2000$ and $p = 10000$ when all features lie exactly on a low-dimensional manifold, in practice, we observe a significant decline in the performance of SkGP with an increase in p. This is attributed to the substantial impact of noise corruption on the manifold, influencing predictive performance.

Swiss Roll		Noise			
		0.01	0.03	0.05	0.1
p = 2000	SkGP	0.96 (0.39)	**1.27 (0.53)**	**1.74 (0.431)**	**3.26 (0.61)**
	GP	1.28 (1.03)	1.75 (1.31)	2.49 (1.12)	4.81 (0.85)
	BART	2.31 (0.57)	2.28 (0.49)	2.35 (0.49)	2.57 (0.66)
	SkBART	**0.91 (0.37)**	1.62 (0.51)	2.63 (0.90)	5.17 (1.07)
	RF	6.92 (0.84)	6.87 (0.87)	6.92 (0.91)	6.97 (0.86)
	SkRF	0.99 (0.56)	2.05 (0.85)	3.28 (0.91)	5.84 (0.97)
	NN	3.52 (0.61)	6.66 (0.93)	7.41 (1.13)	8.47 (1.08)
p = 10,000	SkGP	**1.64 (0.48)**	**2.55 (0.48)**	**3.57 (0.57)**	4.54 (0.91)
	GP	5.19 (1.00)	5.65 (1.00)	6.16 (0.99)	7.08 (0.84)
	BART	4.17 (1.12)	4.07 (0.85)	4.33 (1.06)	**4.37 (0.84)**
	SkBART	2.38 (0.99)	5.33 (0.96)	6.34 (0.86)	7.31 (0.87)
	RF	7.32 (0.84)	7.30 (0.89)	7.32 (0.95)	7.35 (0.88)
	SkRF	3.57 (0.86)	5.72 (0.96)	6.66 (0.90)	7.46 (0.88)
	NN	7.32 (1.17)	8.88 (1.58)	10.15 (1.43)	10.80 (1.77)

Table 1: Averaged Mean squared Prediction Error (MSPE) over 50 replications are shown for the competing models in swiss roll example. Standard errors are presented within parenthesis.

3.3 Predictive Uncertainty

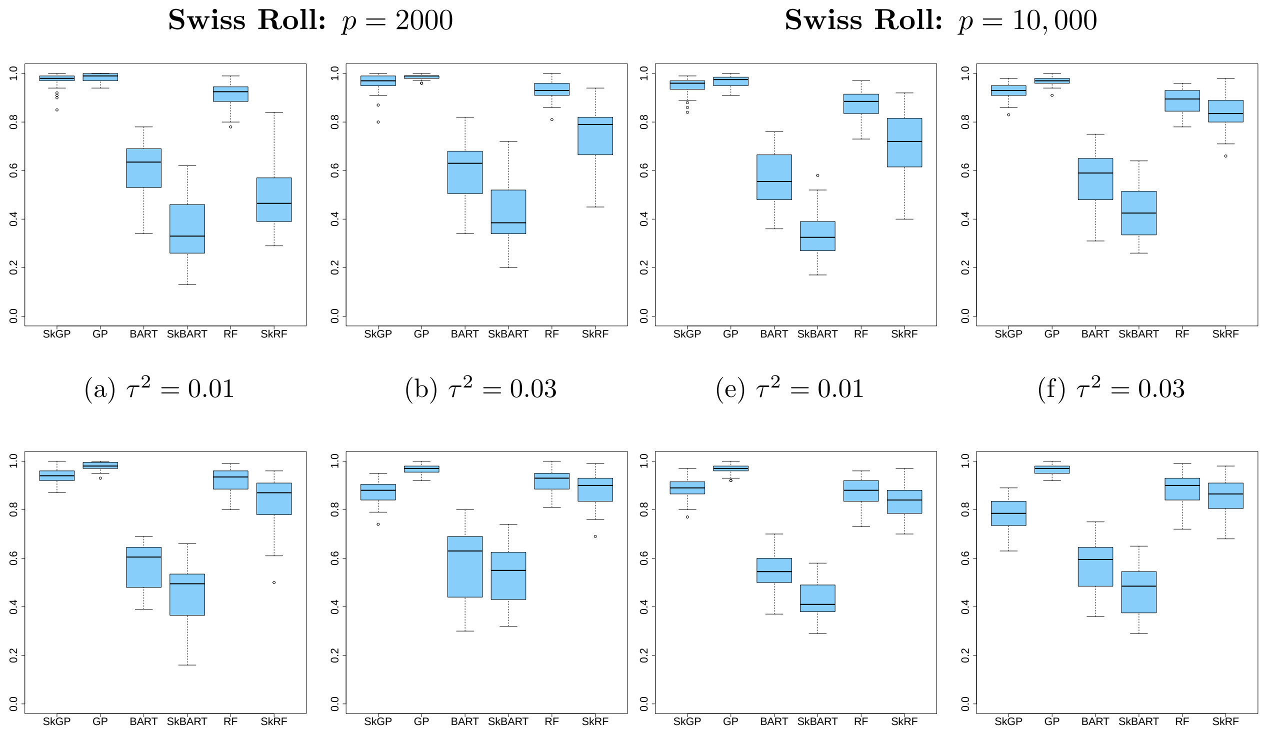
To evaluate quality of predictive uncertainty, we calculate the coverage and length of 95% predictive intervals (PI) for SkGP and other competitors. While frequentist methods, like SkRF and RF, do not inherently provide coverage probabilities with point estimates, we employ a two-stage plug-in approach for them: (i) estimate the regression function in the first stage, and (ii) construct 95% PI based on the normal distribution centered on the predictive mean from the regression model, with variance equal to the estimated variance in the residuals. Coverage probability boxplots over 50 replications for all simulation cases in the swiss roll example and torus example are presented in

Torus		Noise			
		0.01	0.03	0.05	0.1
p = 2000	SkGP	**0.153 (0.036)**	**0.281 (0.072)**	**0.426 (0.108)**	0.884 (0.127)
	GP	0.210 (0.081)	0.308 (0.085)	0.468 (0.104)	**0.817 (0.119)**
	BART	0.932 (0.132)	0.992 (0.137)	0.970 (0.137)	0.916 (0.152)
	SkBART	0.201 (0.084)	0.368 (0.092)	0.645 (0.151)	0.990 (0.139)
	RF	0.957 (0.132)	1.01 (0.124)	0.984 (0.124)	0.937 (0.132)
	SkRF	0.210 (0.091)	0.417 (0.103)	0.641 (0.132)	0.896 (0.127)
	NN	0.547 (0.139)	0.881 (0.122)	0.977 (0.086)	1.082 (0.130)
p = 10,000	SkGP	**0.196 (0.048)**	**0.503 (0.103)**	**0.908 (0.121)**	0.968 (0.136)
	GP	0.237 (0.077)	0.649 (0.110)	0.991 (0.129)	0.956 (0.132)
	BART	0.981 (0.129)	1.04 (0.128)	1.01 (0.129)	0.958 (0.132)
	SkBART	0.228 (0.096)	0.897 (0.152)	1.01 (0.130)	0.995 (0.136)
	RF	0.976 (0.136)	1.03 (0.128)	1.00 (0.125)	**0.953 (0.129)**
	SkRF	0.328 (0.102)	0.79 (0.139)	0.941 (0.131)	0.975 (0.131)
	NN	0.919 (0.114)	1.010 (0.117)	1.053 (0.125)	1.094 (0.118)

Table 2: Averaged Mean squared Prediction Error (MSPE) over 50 replications are shown for the competing models in torus example. Standard errors are presented within parenthesis.

Figures 3 and 4, respectively. Figure 5 displays the median lengths of the 95% PI for all competitors for all simulation cases in both the swiss roll and torus examples.

The results indicate that in all simulation scenarios, the coverage of 95% predictive intervals (PI) for SkGP is close to the nominal level. Although the intervals tend to widen with both increasing noise in the manifold (i.e., higher τ^2) and an increase in the number of features p, the effect is less pronounced with p compared to τ^2. Both BART and SkBART exhibit poor coverage, with significantly narrower PIs. RF and SkRF show undercoverage (around 80% coverage) and wider PIs than SkGP. In the swiss roll example, the coverage of 95% PI for GP is similar to that of SkGP, but the intervals from GP are approximately twice as wide as those from SkGP. In the torus example, GP and SkGP perform similarly for $p = 2000$. However, when $p = 10000$, SkGP demonstrates much narrower predictive intervals than GP, with a similar coverage, as the noise increases in the manifold. Overall, the results suggest that SkGP is more precise in terms of predictive uncertainty and robust compared to its competitors concerning the noise in the manifold.



Swiss Roll: $p = 2000$

Swiss Roll: $p = 10,000$

(a) $\tau^2 = 0.01$

(b) $\tau^2 = 0.03$

(e) $\tau^2 = 0.01$

(f) $\tau^2 = 0.03$

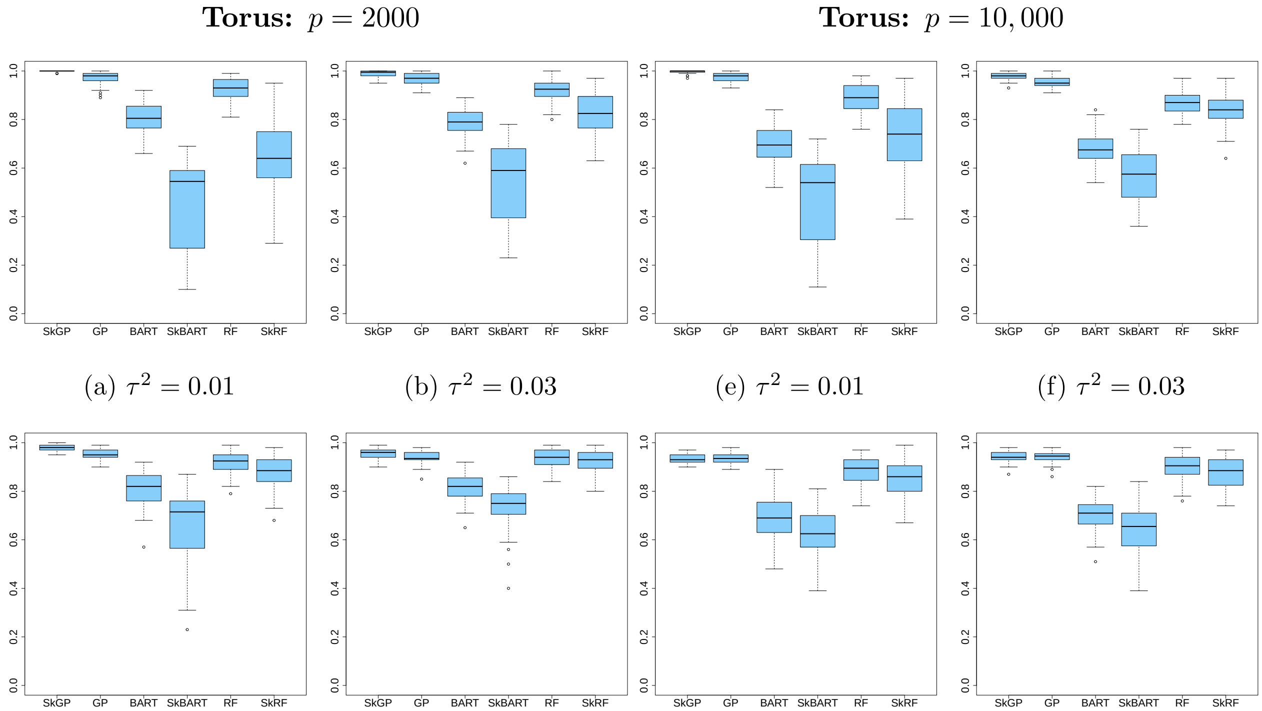
(c) $\tau^2 = 0.05$

(d) $\tau^2 = 0.1$

(g) $\tau^2 = 0.05$

(h) $\tau^2 = 0.1$

Figure 3: Coverage of 95% predictive interval for Swiss Roll Simulations



Torus: $p = 2000$

Torus: $p = 10,000$

(a) $\tau^2 = 0.01$

(b) $\tau^2 = 0.03$

(e) $\tau^2 = 0.01$

(f) $\tau^2 = 0.03$

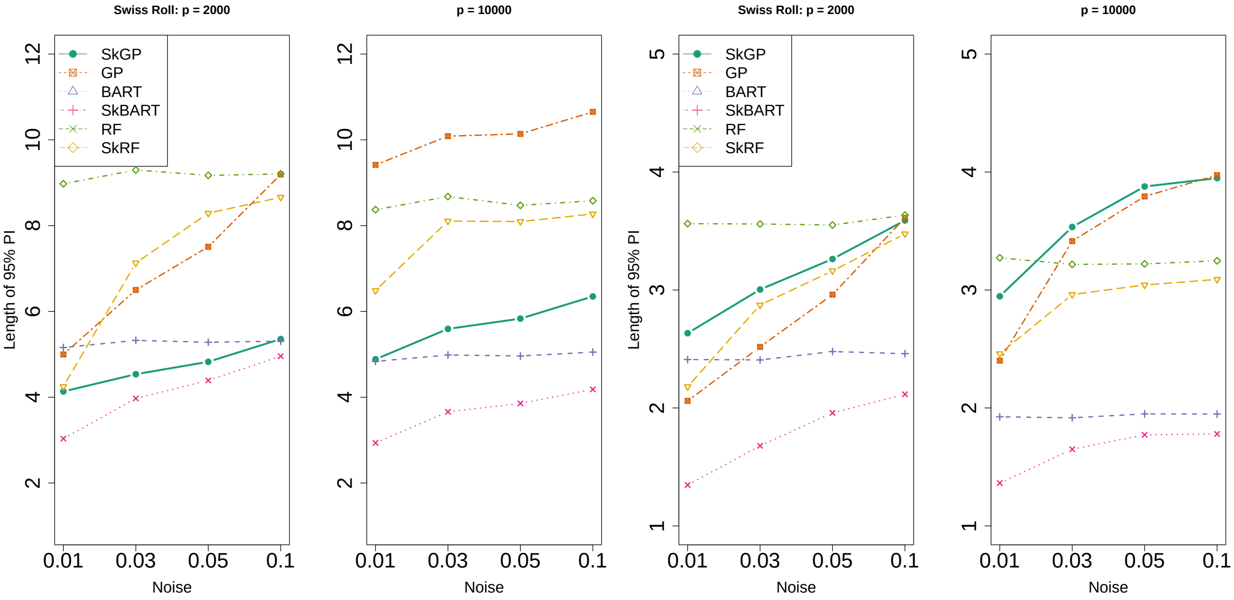
(c) $\tau^2 = 0.05$

(d) $\tau^2 = 0.1$

(g) $\tau^2 = 0.05$

(h) $\tau^2 = 0.1$

Figure 4: Coverage of 95% predictive interval for Torus Simulations



(a) Swiss roll

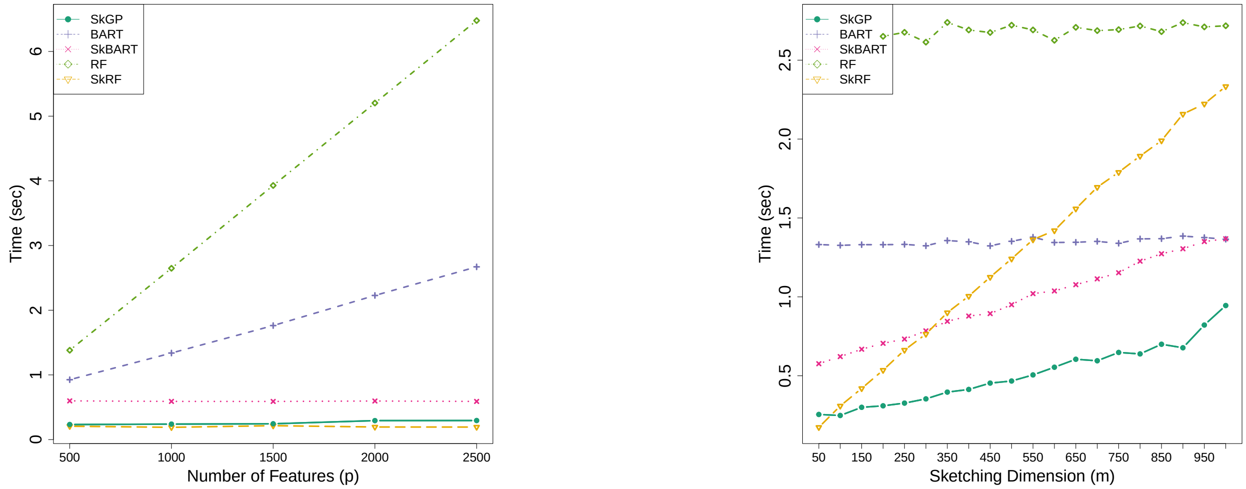
(b) Torus

Figure 5: Length of 95% predictive intervals for all competitors in all simulation settings.

3.4 Computation Time

The main objective in developing SkGP was to improve computational scalability in large p settings. For a specific choice of $\boldsymbol{P}_n, \theta$ and ψ^2, the computation time for SkGP is primarily influenced by two factors: (a) computing the inverse of an $n \times n$ matrix; and (b) multiplying an $m \times |\mathcal{I}|$ matrix with an $|\mathcal{I}| \times n$ matrix. Steps (a) and (b) entail computational complexities of order n^3 and $mn|\mathcal{I}|$, respectively. Since the posterior predictive distribution is available in closed forms, these computations are only needed once for a specific choice of $\boldsymbol{P}_n, \theta$ and ψ^2. We will parallelize the computation across various choices of $\boldsymbol{P}_n, \theta, \psi^2$ on different CPUs. The combination step using stacking requires inverting S matrices each of dimension $(n - n/S) \times (n - n/S)$, incurring a complexity of the order $S(n - n/S)^3$. Since the focus of this article is on moderate n, all computational steps are extremely efficient leading to rapid computation of SkGP.

Figure 6a shows the computation times when the number of features increase and sketching dimension is held fixed ($m = 60$). Computation times for non-sketched tree based methods increase linearly with the number of features, while computation time for sketched methods remain constant, as they only depend on the number of screened features $\mathcal{I}$. Figure 6b shows the computation times as the sketching dimension is increased while the original number of screened features is held constant ($p = 1000$). Considering that the non-sketched tree based methods are not dependent on sketching dimension m, their computation times remain constant. The computation times for the sketched methods increase linearly with sketching dimension. Importantly, SkGP achieves computation time comparable to frequentist approaches, yet being able to allow principled Bayesian predictive inference.



(a) Vary p, fix $m = 60$

(b) Vary m, fix $|\mathcal{I}| = 1000$

Figure 6: The left panel shows the computation time for competitors by fixing the sketching dimension (m), while varying the number of features. The right panel shows the computation time for competitors by fixing the number of screened features $|\mathcal{I}| = 1000$, while varying m.

3.5 Sensitivity to the choice of m and $|\mathcal{I}|$

We present investigation into the choice of the number of features $|\mathcal{I}|$ included through highest marginal association with the response and the dimension of the sketching matrix m applied to this $|\mathcal{I}|$-dimensional feature vector. Figure 7 illustrates the MSPE values for the swiss roll example with varying numbers of included features. Considering the small true dimensions of the swiss roll manifold and the fact that the response is related to $\boldsymbol{x}$ only through the manifold in the swiss roll example, the inclusion of more redundant features in the regression leads to a performance loss, as evidenced by the increasing MSPE values. However, this decline in performance is more pronounced when the swiss roll is affected by noise with higher variance. This aligns with the fact that the accuracy of estimating the regression function depends solely on the intrinsic dimension of the swiss roll and is unaffected by the number of screened features when the features lie on a manifold. When the noise variance is low, resulting in features that approximately lie on the swiss roll, the performance does not change significantly with variations in the number of screened features.

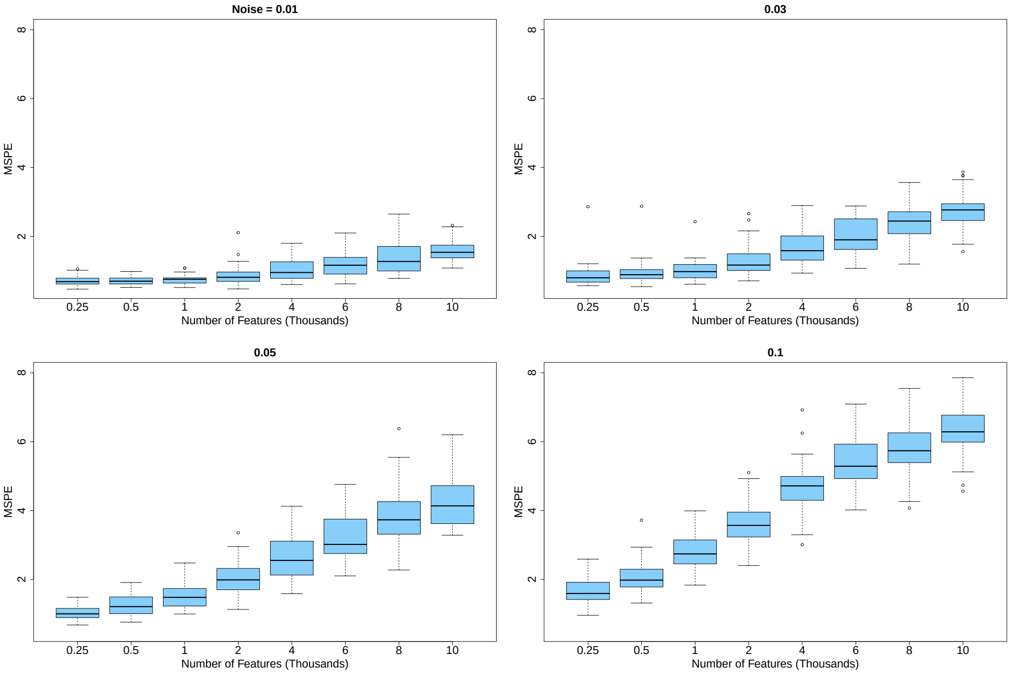


Figure 7: We show the number of included features through the marginal association analysis vis-a-vis predictive accuracy. The plots are presented for the swiss roll example.

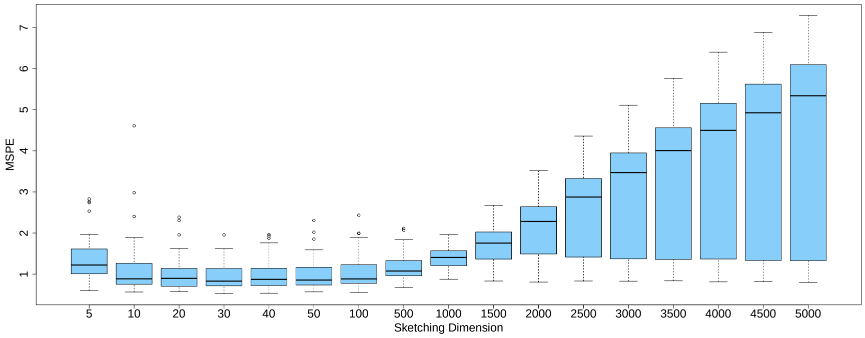


Figure 8: Distribution of MSPE as sketching dimension increases

Figure 8 illustrates the impact of the sketching dimension m on the performance of SkGP.

The figure indicates that as the sketching dimension m increases, MSPE decreases up to a certain point, contingent on the sample size and the structure of the manifold, after which it starts increasing again. This observation is reasonable as, with a low-dimensional manifold $\mathcal{O}$, increasing the sketching dimension m introduces redundant randomly sketched features to the model, leading to a natural decline in performance. In practice, we observed that a sketching dimension of around $m \sim 50$ works well for a diverse range of simulation examples when the intrinsic dimensionality d of the manifold is low.

4 Analysis of Outdoor Air Pollution Data with Satellite Images

We will apply the proposed approach and compare it with relevant competitors in the analysis of air quality using multi-band satellite images over time (see Section 1.1). The air quality dataset consists of measurements taken at the EPA federal reference monitor in Las Vegas, Nevada, spanning almost daily measurements, sometimes with multiple readings in a day, from January 2019 to July 2022. This results in 1667 air quality measurements, as depicted in the first row of Figure 9. For each air quality sample, multi-band satellite images covering the location of the air quality monitor have been acquired, including four wavelength bands: blue, green, red, and near-infrared. The left panel of Figure 10 displays a near-infrared image on a representative day. These data were obtained from Planet using version 1 of their PlanetScope instrument (Planet Team, 2017). Notably, multi-band satellite imagery data are easily obtainable, whereas the installation of monitors measuring air quality is expensive. Hence, a key scientific goal is to predict air quality readings given the high-dimensional multi-band images. To achieve this, at each time point, the $128 \times 128 = 16384$ pixels of the four bands of the multi-band images are vectorized and concatenated into a 65536 dimensional vector.

Although the vectorized images are p dimensional, the estimated intrinsic dimension (ID) for the images is 4.18 with a standard error of 0.1, determined using the two-nearest neighbor (NN) method (Facco et al., 2017). This indicates that the high-dimensional vectorized images lie on a lower-dimensional manifold, motivating the application of our proposed SkGP approach to this data.

Out of 1667 samples, we select every fourth sample point for the test set, resulting in $n = 1334$ training samples and $n_{new} = 333$ test samples. As depicted in Figure 9, the raw air quality monitor data displays characteristics such as non-negativity, heavy-tailed distributions, non-stationary patterns, and periodic behavior. To meet the normality assumption for the error in (3), we apply a log transformation and standardize the response, ensuring a mean of zero and a variance of one. The second row of Figure 9 illustrates the log-transformed and standardized air quality data. Many of the pixels in the satellite images are zero at all times and are included to buffer the image to fit into a square. In our analysis, these zeros are removed. The columns of the image predictor matrix, with zeros removed, have dimension 1334×33068 for the training data, and are pixel-wise standardized to have zero mean and unit variance. We focus our performance comparison on SkGP, BART and SkBART, considering them as the top three competitors based on the simulation studies. While GP is also among the top performers in the simulation studies, it is excluded from the comparison due to its computational demands and memory intensity for this dataset. While the data has a temporal component, our current analysis overlooks its time-varying nature. Incorporating the temporal dynamics and capturing the evolving associations between samples is a direction we plan to explore in future work.

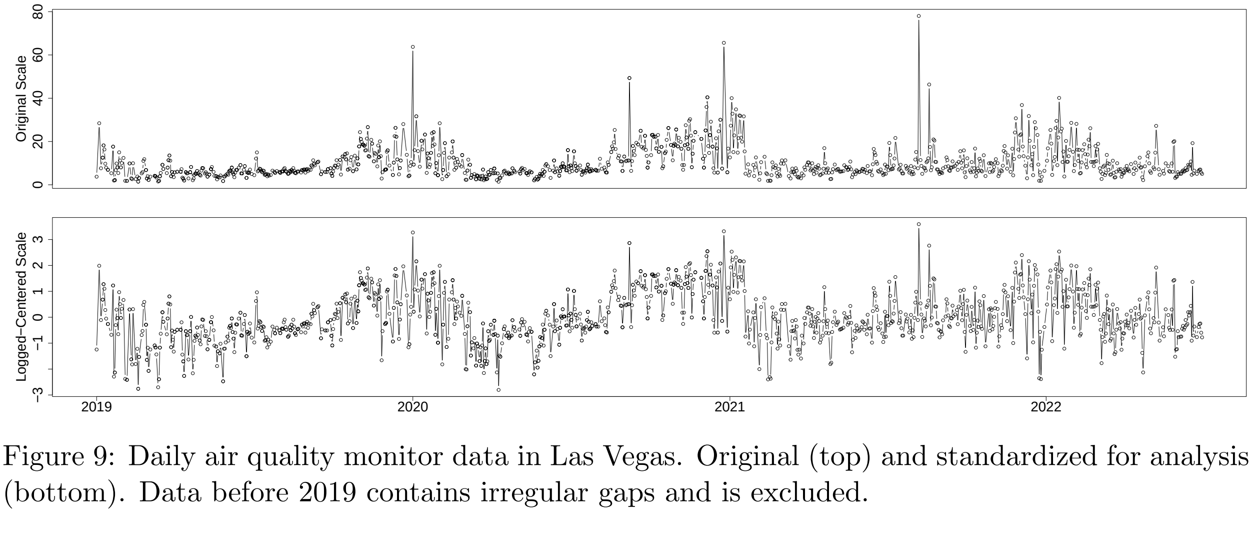


Figure 9: Daily air quality monitor data in Las Vegas. Original (top) and standardized for analysis (bottom). Data before 2019 contains irregular gaps and is excluded.

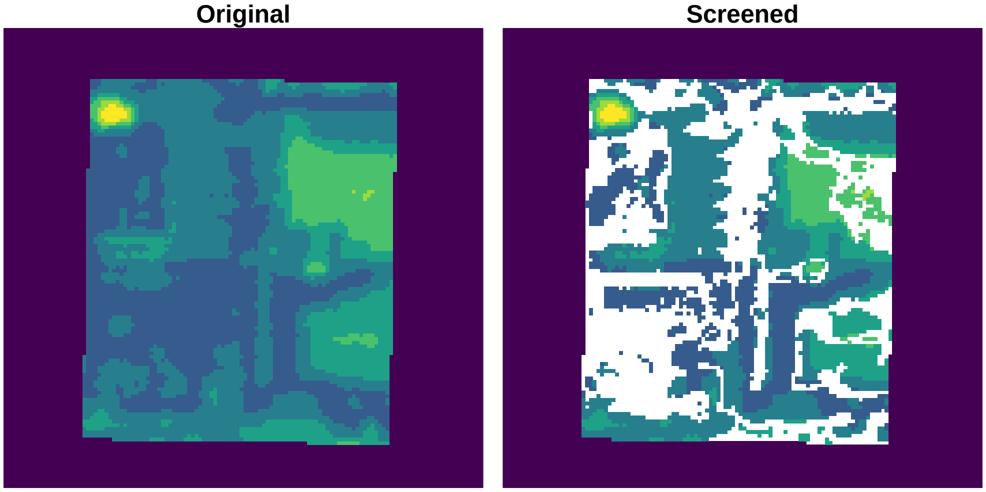


Figure 10: Near infrared image on July 2, 2019. The plot on the left shows the original image. The right plot shows the same image with screened out pixels in white. Interestingly, the independent screening procedure selects contiguous chunks and borders in the image.

4.1 Results

The Nonparametric Independence Screening (NIS) method outlined in (Fan et al., 2014) identifies 18640 features out of 33068 features which are marginally related to the air quality. Figure 10 displays the pixels selected by NIS in a representative multi-band image feature. Interestingly, even though the screening procedure is independent for each pixel, contiguous patched of pixels and boundaries around notable imaging patterns are screened out.

All competing models are implemented using $n = 1334$ samples, where each sample comprises air quality measurements as responses and $p = 18640$ features. Predictive inferences are generated for $n_{new} = 333$ holdout samples. Figure 11 displays the point predictions and 95% predictive intervals for SkGP across all time points, effectively capturing the trend in air quality responses. Table 3 highlights the superior performance of SkGP compared to all competitors, as evidenced by its lowest MSPE value. Although all competitors exhibit under-coverage, potentially due to neglecting the time-varying nature of the data, SkGP achieves the highest coverage (close to 80%) with predictive intervals of comparable length to BART or SkBART. Overall, these results underscore SkGP's effectiveness in modeling the non-linear regression relationship between air quality and multi-band satellite images.

Competitor	MSPE	Coverage	Length
SkGP	0.327	0.784	1.165
BART	0.369	0.739	1.300
SkBART	0.536	0.613	1.159

Table 3: Mean squared Prediction Error (MSPE), length and coverage of 95% predictive intervals for the competing methods SkGP, BART and SkBART for air pollution data.

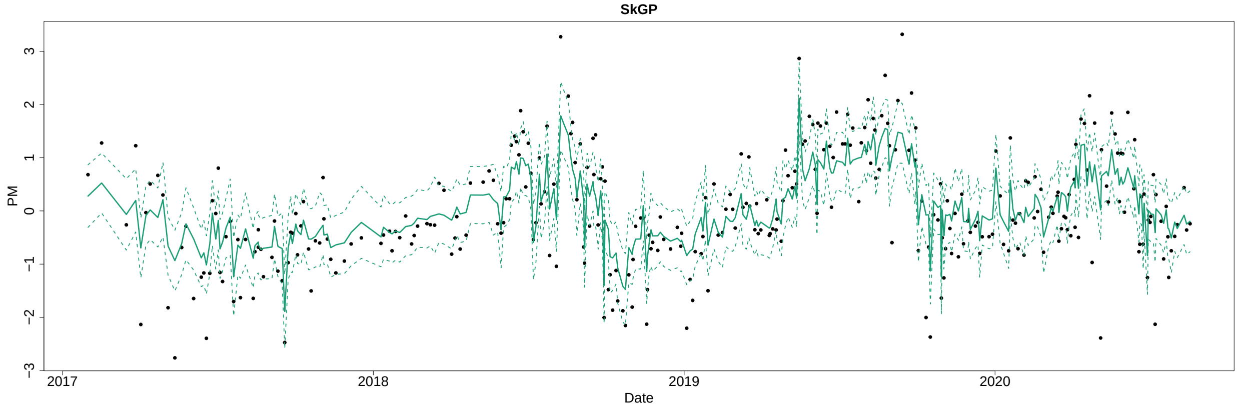


Figure 11: Point prediction and 95% predictive interval at all test samples of air pollution data for SkGP.

5 Conclusion and Future Work

This article is the first to present a novel Bayesian approach for predictive inference of outdoor air quality using high-resolution satellite images, when these images lie on a low-dimensional noise-corrupted manifold. Our methodology exploits two powerful ideas, data sketching and stacking, to eliminate the necessity for computationally demanding manifold structure estimation, providing accurate point predictions and predictive uncertainties. The computation of the posterior predictive distribution does not rely on MCMC sampling, and our framework is amenable to parallel implementation, resulting in substantial reductions in computation and storage costs. Empirical findings underscore the significantly improved point prediction and predictive uncertainty of our approach compared to existing methods. Future research directions will extend our framework to handle large sample sizes using distributed Bayesian inference (Guhaniyogi et al., 2022, 2023), exploring applications to non-Gaussian or multivariate outcomes, and simultaneous estimation of the intrinsic dimensionality of the manifold alongside predictive inference for the outcome.

References

D. Ahfock, W. J. Astle, and S. Richardson. Statistical properties of sketching algorithms. *arXiv preprint arXiv:1706.03665*, 2017.

J. S. Apte, K. P. Messier, S. Gani, M. Brauer, T. W. Kirchstetter, M. M. Lunden, J. D. Marshall, C. J. Portier, R. C. Vermeulen, and S. P. Hamburg. High-resolution air pollution mapping with google street view cars: exploiting big data. *Environmental science & technology*, 51(12): 6999–7008, 2017.

E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local pca. *The Journal of Machine Learning Research*, 18(1):253–309, 2017.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive approximation*, 28:253–263, 2008.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

J. Bleich, A. Kapelner, E. I. George, and S. T. Jensen. Variable selection for bart: an application to gene regulation. 2014.

L. Breiman. Stacked regressions. *Machine learning*, 24:49–64, 1996.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Transactions on Signal Processing*, 58(12):6140–6155, 2010.

Z. Chen, J. Fan, and R. Li. Error variance estimation in ultrahigh-dimensional additive models. *Journal of the American Statistical Association*, 113(521):315–327, 2018.

H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. 2010.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1): 5–30, 2006.

D. G. Denison, B. K. Mallick, and A. F. Smith. A bayesian cart algorithm. *Biometrika*, 85(2): 363–377, 1998.

V. C. Dinh and L. S. Ho. Consistent feature selection for analytic deep neural networks. *Advances in Neural Information Processing Systems*, 33:2420–2431, 2020.

E. Dobriban and S. Liu. A new theory for sketching in linear regression. *arXiv preprint arXiv:1810.06089*, 2018.

P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische mathematik*, 117(2):219–249, 2011.

P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.

Environmental Protection Agency. *Air Quality Criteria for Particulate Matter.* EPA/600/P-95/001aF. Office of Research and Development, Washington DC, 1996.

E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

J. Fan, Y. Ma, and W. Dai. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284, 2014.

R. B. Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences.* CRC press, 2020.

R. Guhaniyogi and D. B. Dunson. Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514, 2015.

R. Guhaniyogi and D. B. Dunson. Compressed gaussian process for manifold regression. *The Journal of Machine Learning Research*, 17(1):2472–2497, 2016.

R. Guhaniyogi and A. Scheffler. Sketching in bayesian high dimensional regression with big data using gaussian scale mixture priors. *arXiv preprint arXiv:2105.04795*, 2021.

R. Guhaniyogi, C. Li, T. D. Savitsky, and S. Srivastava. Distributed bayesian varying coefficient modeling using a gaussian process prior. *The Journal of Machine Learning Research*, 23(1): 3642–3700, 2022.

R. Guhaniyogi, C. Li, T. Savitsky, and S. Srivastava. Distributed bayesian inference in massive spatial data. *Statistical science*, 38(2):262–284, 2023.

N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

C. Heffernan, R. Peng, D. R. Gentner, K. Koehler, and A. Datta. A dynamic spatial filtering approach to mitigate underestimation bias in field calibrated low-cost sensor air pollution data. *The Annals of Applied Statistics*, 17(4):3056–3087, 2023.

Z. Huang. Near optimal frequent directions for sketching dense and sparse matrices. In *International Conference on Machine Learning*, pages 2048–2057. PMLR, 2018.

K. Jensen, T.-C. Kao, J. Stone, and G. Hennequin. Scalable bayesian gpfa with automatic relevance determination and discrete noise models. *Advances in Neural Information Processing Systems*, 34:10613–10626, 2021.

N. Lawrence and A. Hyvärinen. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of machine learning research*, 6(11), 2005.

T. Le and B. Clarke. A bayes interpretation of stacking for m-complete and m-open settings. 2017.

M. LeBlanc and R. Tibshirani. Combining estimates in regression and classification. *Journal of the American Statistical Association*, 91(436):1641–1650, 1996.

D. Li, M. Mukhopadhyay, and D. B. Dunson. Efficient manifold approximation with spherelets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1129–1149, 2022.

G. Li and Y. Gu. Restricted isometry property of gaussian random projection for finite set of subspaces. *IEEE Transactions on Signal Processing*, 66(7):1705–1720, 2017.

Y. Liu, V. Ro$\varepsilon ckov\acute{a}, andY. Wang. Variableselectionwithabcbayesianforests.$ *Journal of the Royal Statistical Soci*

$(3) : 453 - -481, 2021.$

M. Maggioni, S. Minsker, and N. Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. *The Journal of Machine Learning Research*, 17(1):43–93, 2016.

M. W. Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

M. Mukhopadhyay and D. B. Dunson. Targeted random projection for prediction from high-dimensional features. *Journal of the American Statistical Association*, 115(532):1998–2010, 2020.

C. J. Paciorek, Y. Liu, H. Moreno-Macias, and S. Kondragunta. Spatiotemporal associations between goes aerosol optical depth retrievals and ground-level pm2. 5. *Environmental science & technology*, 42(15):5800–5806, 2008.

B. M. Pavlyshenko. Using bayesian regression for stacking time series predictive models. In *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, pages 305–309. IEEE, 2020.

Planet Team. Planet application program interface: In space for life on earth. *San Francisco, CA*, 2017(40):2, 2017. URL https://www.planet.com.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

G.-A. Thanei, C. Heinze, and N. Meinshausen. Random projections for large-scale regression. *Big and Complex Data Analysis: Methodologies and Applications*, pages 51–68, 2017.

M. Titsias and N. D. Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 844–851. JMLR Workshop and Conference Proceedings, 2010.

A. Weingessel and K. Hornik. Local pca algorithms. *IEEE Transactions on neural Networks*, 11 (6):1242–1250, 2000.

C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

D. P. Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Y. Yang and D. B. Dunson. Bayesian manifold regression. 2016.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using stacking to average bayesian predictive distributions (with discussion). 2018.

Y. Yao, G. Pirs, A. Vehtari, and A. Gelman. Bayesian hierarchical stacking: Some models are (somewhere) useful. *Bayesian Analysis*, 17(4):1043–1071, 2022a.

Y. Yao, A. Vehtari, and A. Gelman. Stacking for non-mixing bayesian computations: The curse and blessing of multimodal posteriors. *The Journal of Machine Learning Research*, 23(1):3426–3471, 2022b.

H. Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.

L. Zhang, M. Mahdavi, R. Jin, T. Yang, and S. Zhu. Recovering the optimal solution by dual random projection. In *Conference on Learning Theory*, pages 135–157, 2013.

J. Zhao, L. Chen, W. Pedrycz, and W. Wang. Variational inference-based automatic relevance determination kernel for embedded feature selection of noisy industrial data. *IEEE Transactions on Industrial Electronics*, 66(1):416–428, 2018.