

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

Samuel Gailliot, Rajarshi Guhaniyogi, and Roger D. Peng

19 February 2024

Department of Statistics, Texas A&M University

Introduction

Sketched Gaussian Process Regression

Model Averaging – Stacking of Predictive Distributions

Putting it all Together

Results

Extras

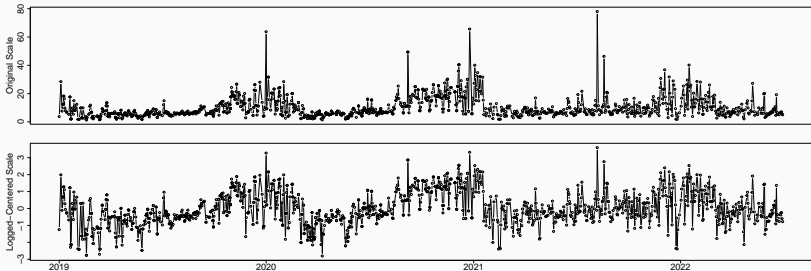
Nonparametric Independence Screening

Introduction

Motivation: Real Data

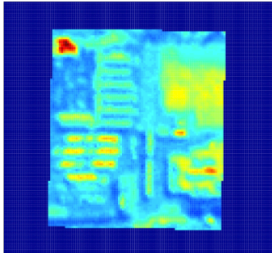
Motivation: Can we use remote sensing data to predict air quality?

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

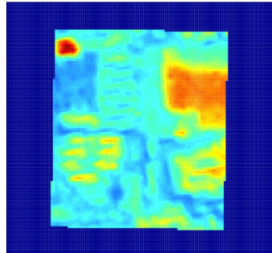


Motivation: Real Data

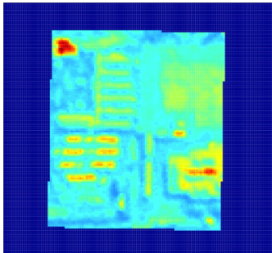
Red



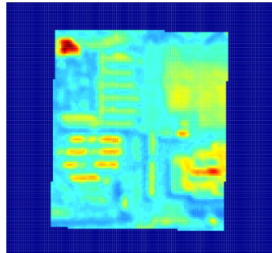
Infrared



Blue



Green



Problem Statement

We want to predict scalar air quality q_t given multi-band satellite image data $\mathcal{X}_t = \{X_t^{(\text{Red})}, X_t^{(\text{Blue})}, X_t^{(\text{Green})}, X_t^{(\text{Infra})}\}$.

$$q_t = f(\mathcal{X}_t) + \epsilon_t$$

depending on the method you choose to estimate f , the input \mathcal{X}_t is either a set of images or four high-dimensional vectors or one *really* high-dimensional vector, etc.

Regardless of representation the information in \mathcal{X}_t stays the same. In this case it is four smoothly varying surfaces.

Manifold Property of Image Data

For GP regression \mathcal{X}_t is a vector of dimension $p = 4 \times p_1 \times p_2 \propto 10^5$.

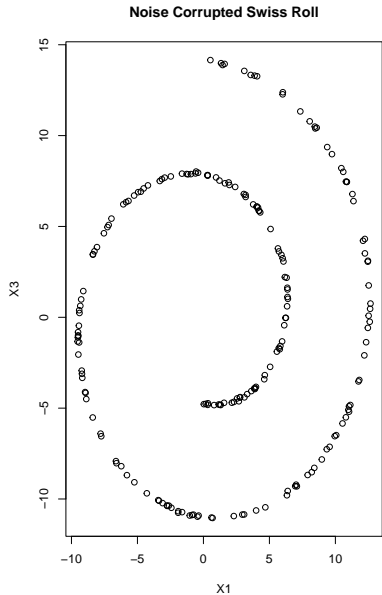
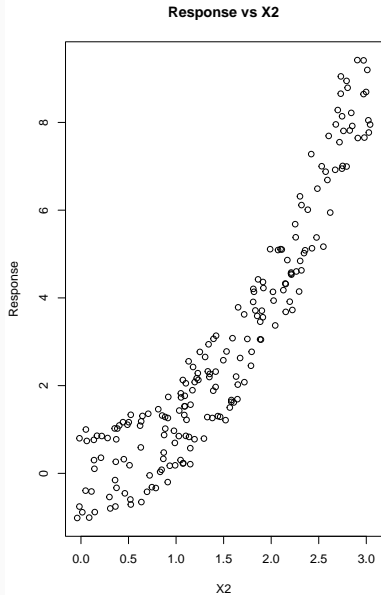
The images themselves have *inherent* dimension $d \approx 1.5$ [Denti, 2021]

When all four images are concatenated into one vector, the inherent dimension remains small, $d < 10$.

GP regression is able to make good predictions for data with a low *inherent* dimension (d) even when the *apparent* dimension (p) is large. [Yang and Dunson, 2016]

Goal: Present a method which draws 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.

Recurring Example: Swiss Roll



1. Sample manifold coordinates

- $t \sim U(\frac{3\pi}{2}, \frac{9\pi}{2})$
- $h \sim U(0, 3)$

2. Construct high dimensional $\mathbf{x} = (x_1, \dots, x_p)$

- $x_1 = t \cos(t) + \delta_1$
- $x_2 = h + \delta_2$
- $x_3 = t \sin(t) + \delta_3$
- $x_i = \delta_i$ for $i \geq 4$, where $\delta_i \sim N(0, \tau^2)$

3. Responses are simulated have to nonlinear and non-monotonic relationship with the features

- $y_i = \sin(5\pi t) + h^2 + \epsilon_i, \epsilon_i \sim N(0, 0.02^2)$

We consider the following model

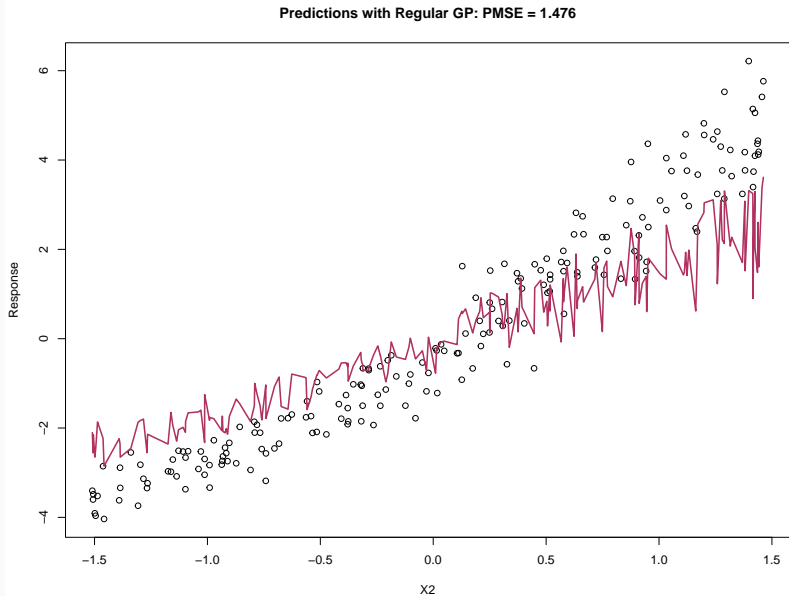
$$y = f(\mathbf{x}) + \epsilon, \quad \epsilon \sim N(0, \tau^2)$$

Where y is a scalar response and $\mathbf{x} = (x_1, \dots, x_p)'$ is a high dimensional feature which resides on a noisy unknown manifold. f is an unknown regression function.

Gaussian process (GP) priors with automatic relevance determination (ARD) kernel are commonly used to estimate f . But this approach struggles when the dimension is of the order of a couple thousand.

We propose a method for ultra-high dimensional GP regression for manifold data.

GP Regression on Swiss Roll: $n = 400$, $p = 10,000$



Sketched Gaussian Process Regression

Johnson-Lindenstrauss Lemma

Johnson-Lindenstrauss Lemma

Given $0 < \epsilon < 1$, a set \mathbf{X} of n points in \mathbb{R}^p , and a number $m > 8 \log(n)/\epsilon^2$, there is a linear map $f: \mathbb{R}^p \rightarrow \mathbb{R}^m$, such that

$$(1 - \epsilon) \|\mathbf{u} - \mathbf{v}\|^2 \leq \|f(\mathbf{u}) - f(\mathbf{v})\|^2 \leq (1 + \epsilon) \|\mathbf{u} - \mathbf{v}\|^2$$

for all $\mathbf{u}, \mathbf{v} \in \mathbf{X}$

Since GP regression is based on distances between points we may work with $f(\mathbf{u})$'s instead of \mathbf{u} 's.

Question: Can we find such an f which is beneficial to our cause?

Answer: Yes! Define $f(\mathbf{u}) = \mathbf{P}\mathbf{u}$. Where $\mathbf{P} = ((P_{ij})) \in \mathbb{R}^{m \times p}$, where $P_{ij} \sim N(0, 1/m)$.

Compressed GP Regression

Suppose we have the following dataset $\mathcal{D}_n = \{(\mathbf{x}_i^T, y_i) : i = 1, \dots, n\}$. Consisting of n observations of a p -variate feature $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^T$ and a scalar-valued response y_i .

Further suppose that the \mathbf{x}_i live along a noisy d -dimensional manifold and that the relationship between \mathbf{x}_i and y_i can be characterized by

$$y_i = f(\mathbf{x}_i) + \epsilon, \quad \epsilon \sim N(0, \tau^2)$$

We approximate the density of y_i by sketching the high dimensional \mathbf{x}_i down to lower dimension using $\mathbf{P}\mathbf{x}_i$ instead

$$y_i = f(\mathbf{P}\mathbf{x}_i) + \epsilon, \quad \epsilon \sim N(0, \tau^2)$$

Compressed GP Regression: Bayesian Model

We place a GP prior on $f(\cdot)$,

$$f(\cdot) \sim GP(0, \sigma^2 \delta_\theta), \text{ where } \delta_\theta(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\theta \|\mathbf{x}_i - \mathbf{x}_j\|)$$

Denote $\mathbf{f} = (f(\mathbf{P}\mathbf{x}_1), \dots, f(\mathbf{P}\mathbf{x}_n))'$ and $\mathbf{C} \in \mathbb{R}^{n \times n}$ as the covariance matrix, $\mathbf{C}_{ij} = \delta_\theta(\mathbf{P}\mathbf{x}_i, \mathbf{P}\mathbf{x}_j)$

Hierarchical Bayesian Model

$$\mathbf{y}|\mathbf{f}, \tau^2 \sim N(\mathbf{f}, \tau^2 \mathbf{I})$$

$$\mathbf{f}|\tau^2 \sim N(0, \tau^2 \psi^2 \mathbf{C})$$

$$\pi(\tau^2) \propto \frac{1}{\tau^2}$$

for fixed length scale θ and signal to noise ratio $\psi^2 = \sigma^2/\tau^2$

This setup ensures closed-form conjugate posterior distributions.

We obtain the following marginal posterior distributions:

$$\tau^2 | \mathbf{P}, \theta, \psi^2, \mathcal{D}_n \sim IG(n/2, \mathbf{y}^T (\psi^2 \mathbf{C} + \mathbf{I})^{-1} \mathbf{y} / 2)$$

$$\mathbf{f} | \mathbf{P}, \theta, \psi^2, \mathcal{D}_n \sim t_n(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$$

where,

$$\boldsymbol{\mu}_t = (\mathbf{I} + \mathbf{C}^{-1} / \psi^2)^{-1} \mathbf{y}, \quad \boldsymbol{\Sigma}_t = (2b/n)(\mathbf{I} + \mathbf{C}^{-1} / \psi^2)^{-1}.$$

Now consider prediction at n_{new} data points $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{n_{new}}$

The posterior predictive distribution of the response

$\tilde{\mathbf{y}}_{new} = (\tilde{y}_1, \dots, \tilde{y}_{n_{new}})^T$ follows,

$$\tilde{\mathbf{y}}|\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{n_{new}}, \mathbf{P}, \theta, \psi^2, \mathcal{D}_n \sim t_{n_{new}}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$$

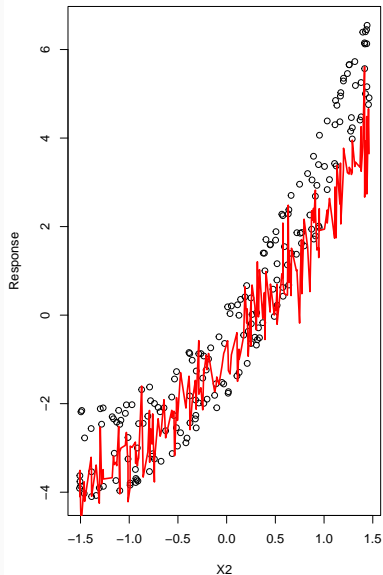
Where

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

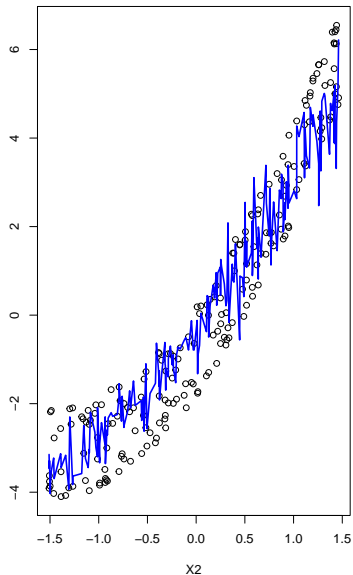
Takeaway: All of the posterior distributions are available in closed form! Bayesian inference can proceed from *exact* posterior samples.

Predictions on Swiss Roll: $n = 400$, $p = 10,000$

Model 1: PMSE = 1.426



Model 2: PMSE = 0.956



Model Averaging – Stacking of Predictive Distributions

Predictions are dependent on the random sketching matrix \mathbf{P} and on the parameters ψ^2 and θ .

In order to lessen dependence on any single random model $M = \{\mathbf{P}, \theta, \psi^2\}$ we average over many of them, M_1, \dots, M_K

Previous work used Bayesian Model Averaging (BMA) but this approach is unsatisfactory for multiple reasons.

[Guhaniyogi and Dunson, 2016]

We propose to use stacking of predictive distributions instead.

The relationship between the true data generator and the model list $\mathcal{M} = \{M_1, \dots, M_K\}$ falls into one of three categories:

1. \mathcal{M} -Closed: The true data generator is one of the models in \mathcal{M}
2. \mathcal{M} -Complete: The true model exists but is not a member of \mathcal{M} .
We still wish to use the models in \mathcal{M} for some reason.
Tractability, computational ease, etc.
3. \mathcal{M} -Open: We know the true model is not in \mathcal{M} and we cannot specify it.

BMA is appropriate in the \mathcal{M} -Closed case but in the \mathcal{M} -Complete and \mathcal{M} -Open cases BMA will asymptotically select the single model closest to the true model in KL-divergence. [Yao et al., 2018]

We need an averaging method which privileges the predictive performance of the models.

"Scoring rules assess the quality of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materialized" [Gneiting and Raftery, 2007]

Since we know the posterior predictive distribution of $\tilde{\mathbf{y}}$ analytically, we can efficiently evaluate proper scoring functions. We will use this to construct averaging weights.

Stacking

Stacking (of means) is a direct, two-step method for obtaining point estimates from multiple models. Given data (y_i, \mathbf{x}_i) , $i = 1, \dots, n$ and parametric models M_k having form $f_k(\mathbf{x}|\theta_k)$.

1. Fit each model and obtain the LOO predictor for each data point

$$f_k^{(-i)}(x_i) = E[y_i | \theta_k, y_{-i}, M_k]$$

2. Solve for the model weights by minimizing the LOO squared error

$$\hat{w} = \arg \min_w \sum_{i=1}^n \left(y_i - \sum_k w_k \hat{f}_k^{(-i)}(x_i) \right)^2$$

Then the point prediction at a new data point \mathbf{x}_{new} is given by,

$$\hat{y}_{new} = \sum_{k=1}^K \hat{w}_k f_k(\mathbf{x}_{new} | \theta_k)$$

Stacking of Predictive Distributions

Our model returns a full distribution, not just a point estimate.

We can repeat the stacking process above, but instead of minimizing the squared error when finding \hat{w} we can maximize the log score.

1. Obtain the LOO predictive densities for each model k and data point i

$$p_{k,-i}(y_i) = t_{n-1}(y_i | \boldsymbol{\mu}_{k,-i}, \boldsymbol{\Sigma}_{k,-i})$$

2. Solve for the model weights by maximizing the score over all the data

$$\max_{w \in \Delta^K} \frac{1}{n} \sum_{i=1}^n \log \sum_{k=1}^K w_k p_{k,-i}(y_i)$$

Where Δ^K denotes the K -dimensional probability simplex.

3. The stacked estimate of the predictive density is given by

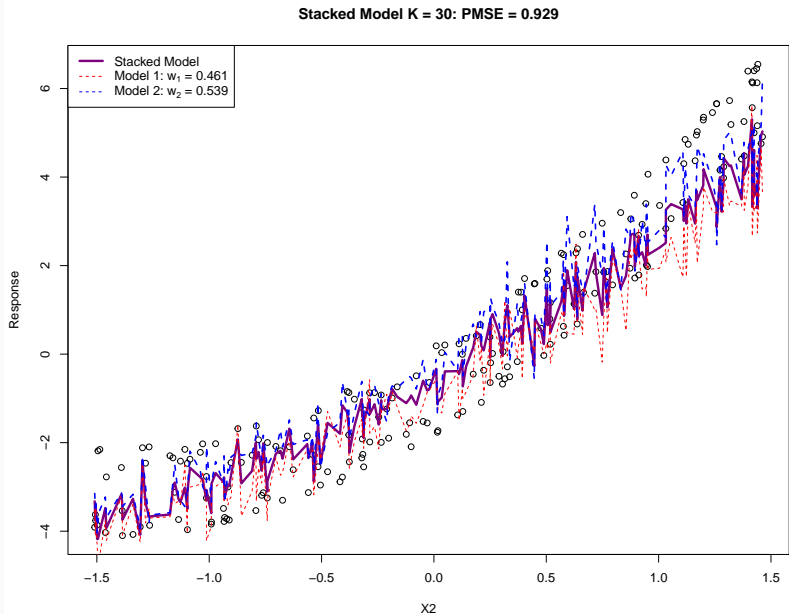
$$\hat{p}(y_{new} | \mathbf{y}) = \sum_{k=1}^K \hat{w}_k p(y_{new} | \mathbf{y}, M_k)$$

Stacking of Predictive Distributions

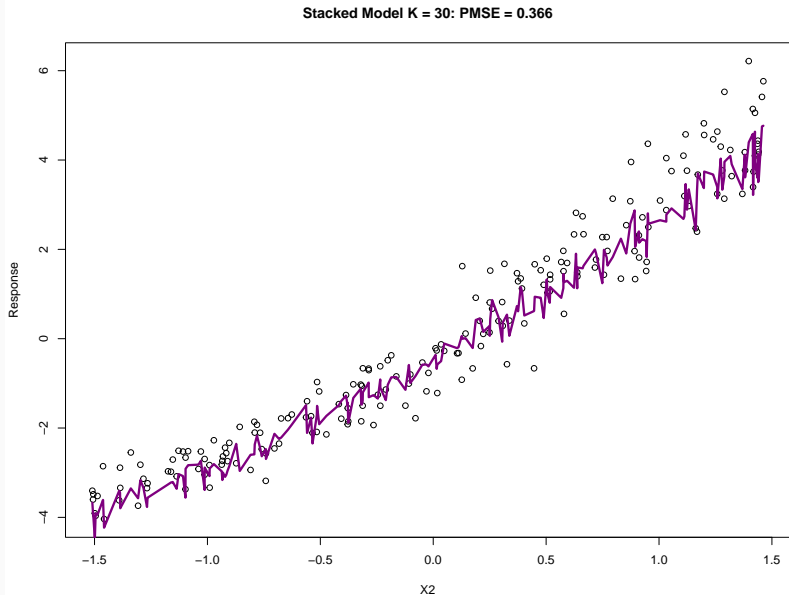
Notes:

1. $\mu_{k,-i}$ and $\Sigma_{k,-i}$ are known in closed form but evaluating them requires inverting and $(n-1) \times (n-1)$ matrix. For LOO stacking we invert Kn $(n-1) \times (n-1)$ matrices. I think I can speed this up by using the Sherman-Woodbury-Morrison formula.
2. Instead of LOO it is possible obtain the predictive densities in K -fold batches. [Zhang et al., 2023] have shown good results in a spatial setting with $K = 10$.
3. Solving for the model weights is done using BFGS the optimizer. It is simple to obtain the gradient of the objective function, but it is a surprisingly difficult optimization problem.

Predictions on Swiss Roll: $n = 400$, $p = 10,000$

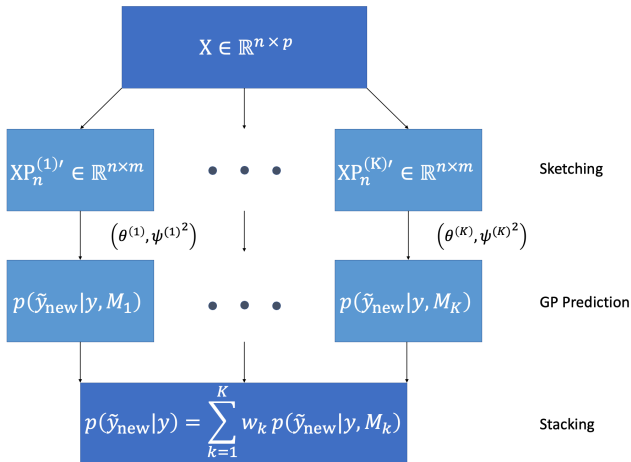


Predictions on Swiss Roll: $n = 400$, $p = 10,000$



Putting it all Together

Full Method



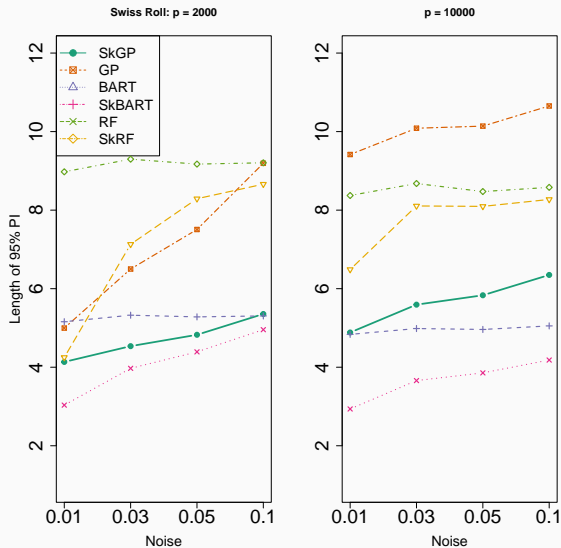
Results

Swiss Roll: Prediction

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.

Swiss Roll: CI Length



Swiss Roll: Run Time

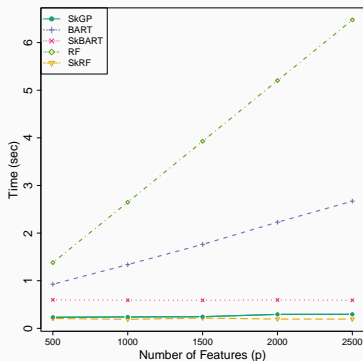


Figure 1: Vary p , fix $m = 60$

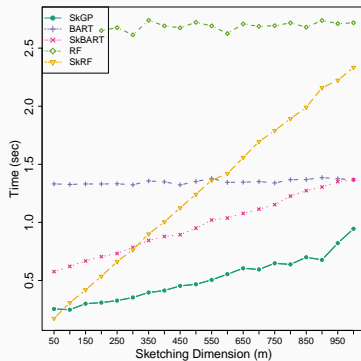


Figure 2: Vary m , fix $|\mathcal{I}| = 1000$

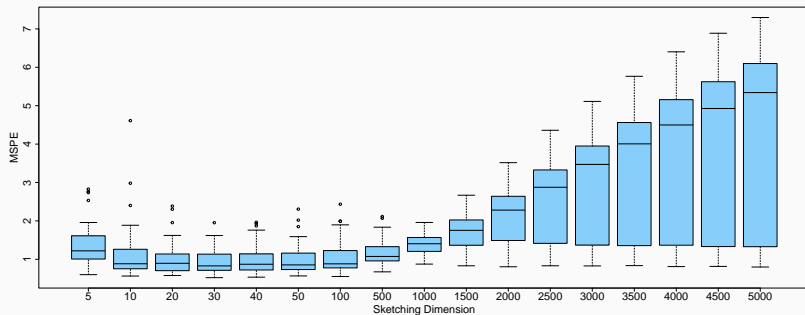


Figure 3: Distribution of MSPE as sketching dimension increases

Real Data: Preprocessing

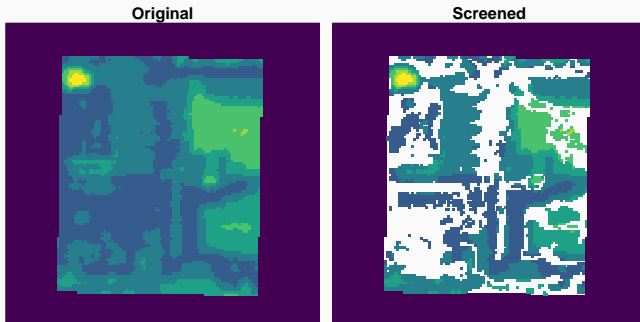


Figure 4: Near infrared image on July 2, 2019. The plot on the left shows the original image. The right plot shows the same image with screened out pixels in white. Interestingly, the independent screening procedure selects contiguous chunks and borders in the image. Screening was performed via the NIS method [Fan et al., 2011].

Real Data: Predictions

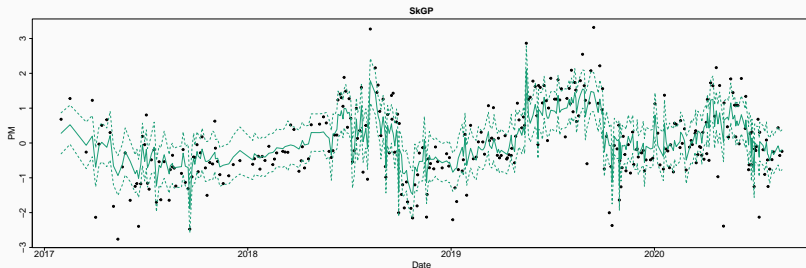


Figure 5: Point prediction and 95% predictive interval at all test samples of air pollution data for SkGP. Every fourth third point is held out for prediction.

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

Conclusion

We have presented a method which allows for scalar on image regression problems to be solved using GP regression.

Our proposed method:

1. draws predictive inference of a random variable from a high-dimensional feature vector using “sketching” of the feature vector when the feature vector lies on a low-dimensional noisy unknown manifold
2. **Is Fast**: Requires no MCMC and is easily parallelized.
3. **Scalable**: Regular GP regression can be replaced with Vecchia approximation or other scalable GP methods.
4. naturally **Quantifies Uncertainty**

Questions?

Thank you!

Extras

Choosing θ and ψ^2

In order to limit sensitivity of the results to any one choice of $\{\mathbf{P}_n, \theta, \psi^2\}$ we propose to average over many models. We generate $k = 1, \dots, K$ sketching matrices $\mathbf{P}_n^{(k)}$ where $\{\mathbf{P}_n^{(k)}\}_{i,j} \sim N(0, 1)$. Then, using the fact that the marginal posterior distribution of $\theta, \psi^2 | \mathbf{P}_n^{(k)}, \mathbf{y}$ is given by

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

We find a pair $(\theta^{(k)}, \psi^{2(k)})$ which maximize this density. Due to intractability of obtaining the gradient of f , optimization is performed via grid search. The grid $\{\theta_1, \dots, \theta_t\} \times \{\psi_1^2, \dots, \psi_s^2\}$ is constructed by sampling $\theta_1, \dots, \theta_t$ uniformly in $[3/d_{max}, 3/d_{min}]$ and $\psi_1^2, \dots, \psi_t^2$ uniformly in $(0, \psi_{max}^2]$, where $d_{max} = \max_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|$, $d_{min} = \min_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|$ and ψ_{max}^2 is a sufficiently large signal to noise ratio, e.g. 10, which we allow to be user specified in our code implementation.

-



Denti, F. (2021).

intrinsic: An r package for model-based estimation of the intrinsic dimension of a dataset.

arXiv preprint arXiv:2102.11425.



Fan, J., Feng, Y., and Song, R. (2011).

Nonparametric independence screening in sparse ultra-high-dimensional additive models.





Journal of the American Statistical Association, 106(494):544–557.



Gneiting, T. and Raftery, A. E. (2007).

Strictly proper scoring rules, prediction, and estimation.

Journal of the American statistical Association, 102(477):359–378.

-  Guhaniyogi, R. and Dunson, D. B. (2016).
Compressed gaussian process for manifold regression.
The Journal of Machine Learning Research, 17(1):2472–2497.
-  Yang, Y. and Dunson, D. B. (2016).
Bayesian manifold regression.
-  Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018).
Using stacking to average bayesian predictive distributions (with discussion).
-  Zhang, L., Tang, W., and Banerjee, S. (2023).
Exact bayesian geostatistics using predictive stacking.
arXiv preprint arXiv:2304.12414.

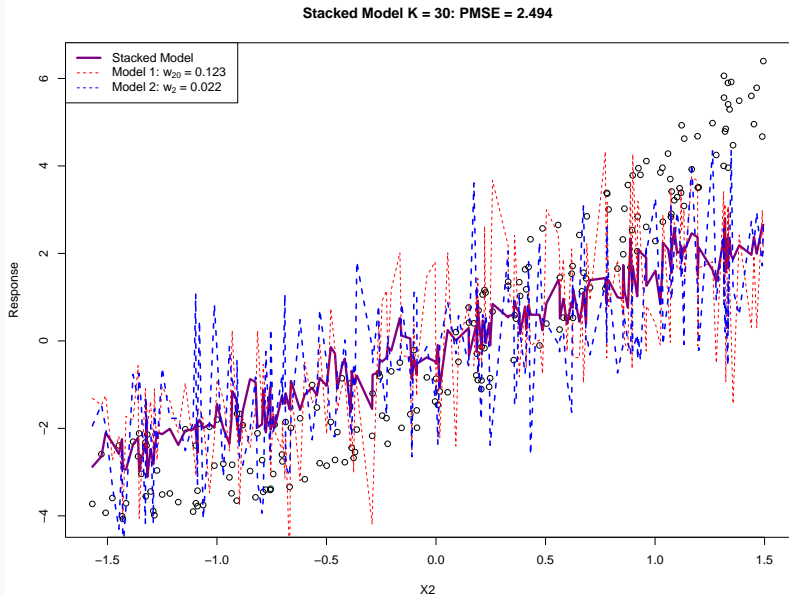
Nonparametric Independence Screening

If the amount of noise in the ultra-high dimensional predictors is too large compression will destroy all of the signal.

We want to remove some of the noisy, useless predictors before we multiply by the compression matrix.

To this end we employ the Nonparametric Independence Screening (NIS) method from [Fan et al., 2011].

Too Much Noise



The purpose of the NIS procedure is to quickly assess the marginal importance of each covariate \mathbf{X}_j , $j = 1, \dots, p$. for the model

$$y_i = f(\mathbf{x}_i) + \epsilon$$

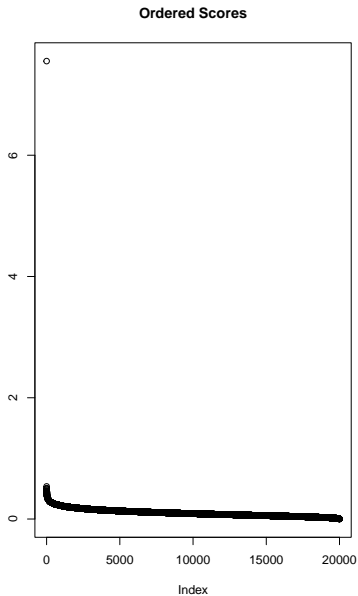
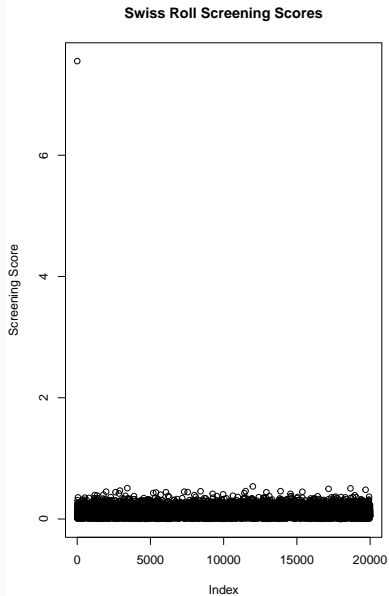
1. Consider the p marginal nonparametric regression problems

$$\min_{g_j} E [(\mathbf{y} - g_j(\mathbf{X}_j))^2]$$

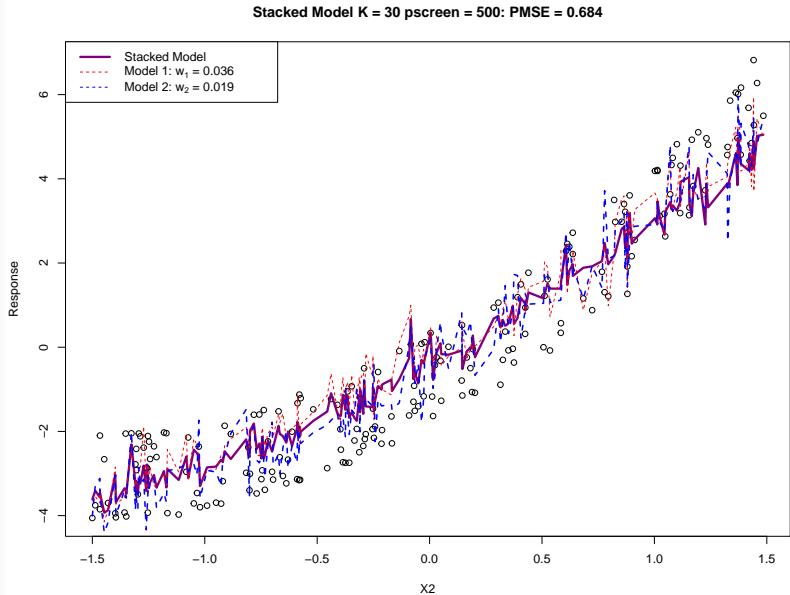
2. To obtain a sample version we use a B-spline basis using a shared number of basis functions across all \mathbf{X}_j .
3. Rank according to the descent order of the residual sum of squares of the componentwise nonparametric regressions.

[Fan et al., 2011] suggest using a permutation test based cutoff for variable selection. In our simulation studies we set a conservatively high p_{screen} .

Swiss Roll Screening



High Noise with Screening



Effect of Screening

