Embarrassingly Parallel Inference for Gaussian Processes

Michael Zhang, Sinead Williamson

The University of Texas at Austin
Department of Statistics and Data Sciences
College of Natural Sciences

Introduction

- A typical problem in statistics and machine learning is learning a latent function given noisy observations.
- Ex: Regression, classification, optimization.
- Learning a general, non-linear function is not a trivial task.

Introduction

- In the Bayesian context, we can place prior over latent function, $P(f)$
- Prediction and inference requires intractable integrals.
- We have a solution for Bayesian non-parametric function learning.

Gaussian Processes

- Gaussian processes (GP) are distributions over functions.
- A GP-distributed function at any finite set of points is multivariate normal:

$$f|X \sim \mathsf{N}\left(m(X), \Sigma(X, X')\right)$$

- GPs are often used as priors non-linear functions.

Gaussian Processes

- Ex: Regression

$$f \sim \mathsf{GP}(0, \Sigma), \quad f|X \sim \mathsf{N}\left(0, \Sigma(X, X')\right), \quad Y|X, f, \sigma^2 \sim \mathsf{N}(f(X), \sigma^2 I)$$

- Posterior is available in closed form.
- To fit GP model in we need to learn kernel hyperparameters, θ.

Gaussian Process Inference

- We learn hyperparameters by optimizing w.r.t. marginal likelihood:

$$P(Y|X) \propto \left|\Sigma(X,X') + \sigma^2 I\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}Y^T \left[\Sigma(X,X') + \sigma^2 I\right]^{-1} Y\right\}$$

- Inference costs $O(N^3)$.

Sparse Gaussian Process Inference

- From Bauer et al. (2016):

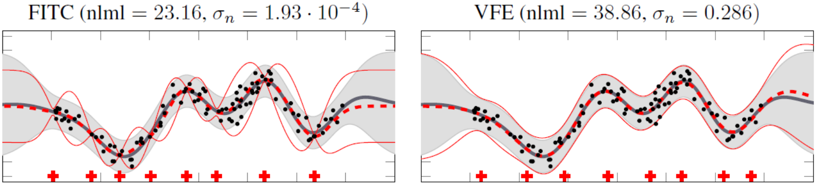
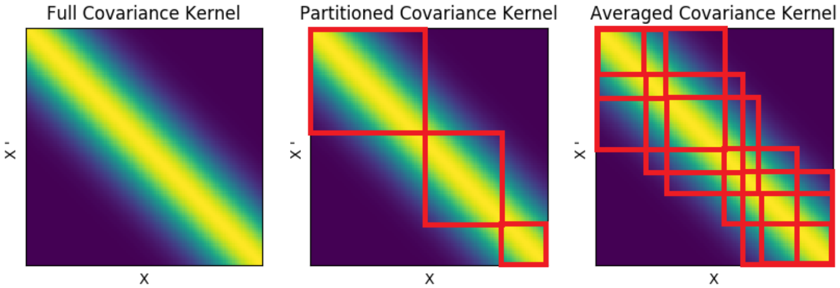


Figure 1: Behaviour of FITC and VFE on a subset of 100 data points of the Snelson dataset for 8 inducing inputs (red crosses indicate inducing inputs; red lines indicate mean and 2σ) compared to the prediction of the full GP in grey. Optimised values for the full GP: nlml $= 34.15$, $\sigma_n = 0.274$

Local Gaussian Process Methods

- Idea behind partitioning and averaging:



Problem Statement

- There are problems in scalable GP inference modeling general types of functions
- Averaging over partitions in local GP methods can solve these problems but is slow.
- We propose a method that averages over partitions quickly.

Importance Sampling

- Suppose we are interested in the integral:

$$\bar{f} = \int f(x)p(x)\,\mathrm{d}x$$

  but we cannot compute the integral easily.
- I.e. integrating partition from posterior.

Importance Sampling

- If we have a proposal distribution $q(x)$ we can approximate $\bar{f}$ with:

$$\bar{f} = \int \frac{f(x)p(x)}{q(x)} q(x)\,\mathrm{d}x \approx \frac{1}{J}\sum_{j=1}^{J} f(x^{(j)})\frac{p(x^{(j)})}{q(x^{(j)})}$$

  by drawing J samples from $q(x)$
- p/q is also known as the importance sampling weight, w

Importance Gaussian Process Sampler

- We assume input is GMM:

$$x_i \sim \text{Normal}(\mu_{z_i}, \Gamma_{z_i}), \quad (\mu_k, \Gamma_k) \sim \text{Normal-Inv. Wishart}(\cdot)$$

$$z_i \sim \text{Categorical}(\pi), \quad \pi \sim \text{Dirichlet}(\alpha)$$

- Each proposal for IS is a partition:

$$Z_j \sim P(Z|X,\pi) := q$$

- We fit K separate GP models to the partitioned data.

Importance Gaussian Process Sampler

- The IS weights each proposed model according to:

$$w_j \propto \frac{p(Z|X,Y)}{p(Z|X)} \propto \frac{p(X,Y|Z)p(Z)}{p(X|Z)p(Z)} = p(Y|X,Z)$$

- Because we calculate self-normalized weights, IS has bias of $O(1/J)$.
- Complexity of IGPS: $O(JN^3/K^2)$, for J importance samples.

Importance Gaussian Process Sampler

- Prediction is performed through:

$$P(f_j^*|-) = \sum_{k=1}^{K} P(f_j^*|X_{k,j}, Y_{k,j}, Z_j, X^*, Z_j^*) P(Z_j^*|X_{k,j}, Z_j)$$

- We can also model non-stationary data easily.
- Each proposal is independent–embarrassingly parallel inference

Stochastic Approximation

- We can approximate the full likelihood by sampling a $B << N$ size subset of the data.

$$P(Y|X, Z, f) \approx P(Y^{mb}|X^{mb}, Z^{mb}, f^{mb})^{\frac{N}{B}}$$

- With stochastic approximation, complexity is $O(B^3/K^2)$ per importance sample.

Importance Gaussian Process Sampler

- The algorithm:
  1. Generate J partitions with K clusters.
  2. For each importance sample, fit K independent GP models.
  3. Predict new data with

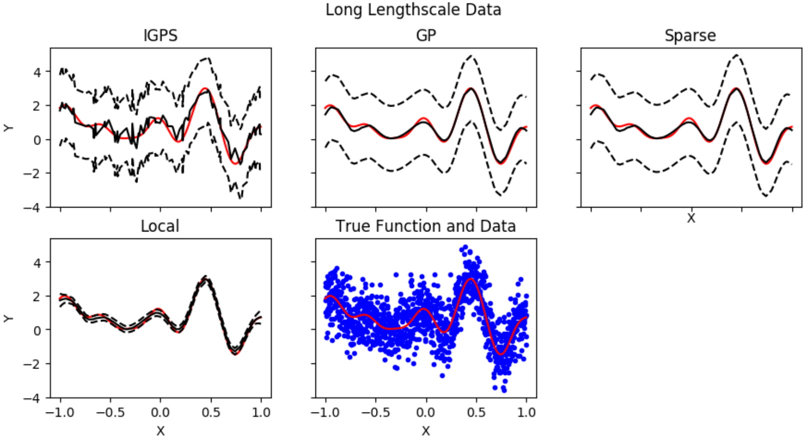
$$P(f_j^*|Z_j,-) = \sum_{k=1}^{K} P(f_j^*|Z_j^*,-)P(Z_j^*|-)$$

  4. Obtain weights $w_j = \prod_{k=1}^{K} P(Y_{k,j}|X_{k,j},Z_j)$ and normalize
  5. Average using importance weights:

$$P(\bar{f}^*|-) = \sum_{j=1}^{J} w_j P(f_j^*|Z_j,-)$$

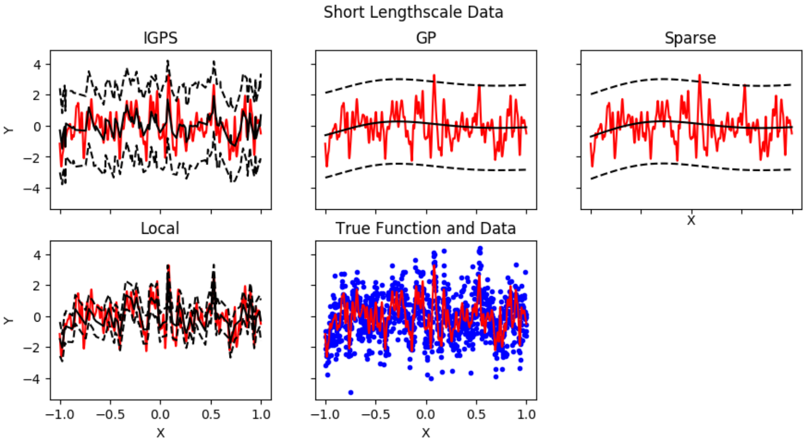
Results

Synthetic data, stationary, long length-scale, $N = 1000$, $K = 10$, $J = 10$



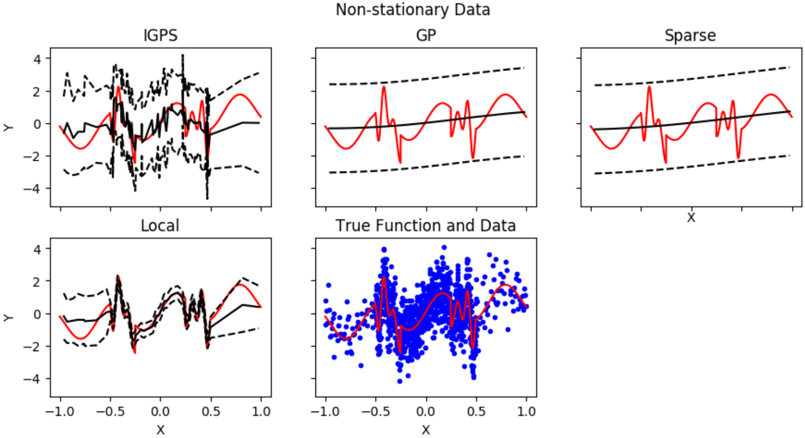
Results

Synthetic data, stationary, short length-scale, $N = 1000$, $K = 10$, $J = 10$



Results

Synthetic data, non-stationary, $N = 1000$, $K = 10$, $J = 10$



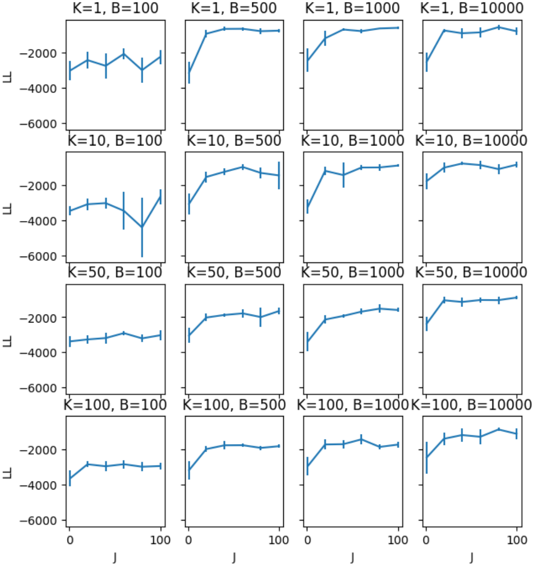
Results

Table: Test set performance on synthetic datasets

Data	**IGPS**	GP	Sparse	Local
Long Lengthscale	-152.41	-143.52	-143.39	-5207.89
Short Lengthscale	-157.16	-172.32	-172.26	-251.50
Non-stationary	-158.21	-181.40	-181.30	-910.54

Results

Sensitivity to settings



Results

Table: Test set log likelihood and MSE for various weighting schemes. Standard errors are in parentheses

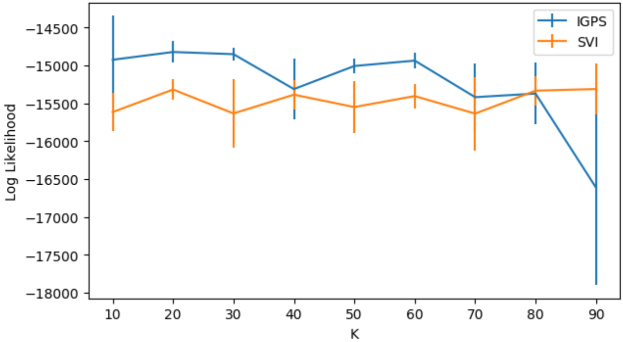
Setting	LL	MSE
IS with SA	-354.96 (28.59)	0.096 (0.003)
IS without SA	-423.93 (41.24)	0.097 (0.002)
Unif. with SA	-726.67 (14.19)	0.19 (0.019)
Unif. without SA	-842.88 (9.82)	0.25 (0.337)

Table: Test set log likelihood and MSE for two different covariate partitioning schemes. Standard errors are in parentheses.

Setting	LL	MSE
GMM	-429.19 (41.57)	0.14 (0.01)
Random Clusters	-789.37 (37.43)	0.53 (0.02)

Results

Large air quality dataset, $N = 209631, J = 100, B = 1000$.

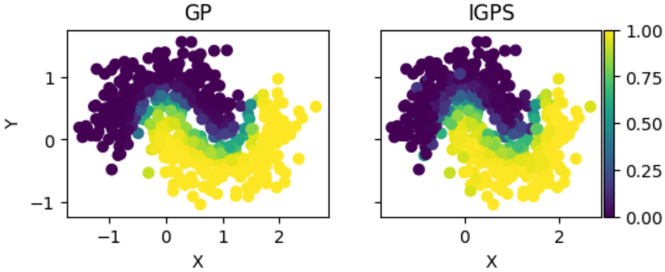


Results

Table: Classification log likelihood results (left) and AUC results (right)

Data	GP	IGPS	Sparse GP
Pima	-128.79	-135.09	-128.61
Parkinsons	-17.00	-22.76	-28.42
WDBC	-15.50	-12.62	-18.01

Data	GP	IGPS	Sparse GP
Pima	0.83	0.81	0.83
Parkinsons	0.86	0.93	0.88
WDBC	0.83	0.91	0.81



Conclusion

- GPs are nice models but not scalable.
- Methods for scalable inference have drawbacks.
- Our proposed method is more general, inference can be parallelized.
- Future interest in applying to complicated GP models.