

# Embarrassingly Parallel Inference for Gaussian Processes

Michael Zhang



# Acknowledgments

- Collaboration with Sinead Williamson. Online variant with Bianca Dumitrascu, Barbara Engelhardt as well.
- Just released in Journal of Machine Learning Research.
- Code and paper available at `michaelzhang01.github.io`.

- A typical problem in statistics is learning a latent function given noisy observations.
- Learning a general, non-linear function is not a trivial task.
- In the Bayesian context, we can place Gaussian process prior over latent function,  $P(f)$

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

$$f|X \sim \mathbf{N}(m(X), \Sigma(X, X'))$$

- Normal property makes GPs convenient to use in statistical models.

# Problems with Gaussian Processes

- Problem 1: Inference costs  $O(N^3)$ .
- Problem 2: Popular kernels are stationary, non-stationary kernels and mixtures are slow to use.
- Solution: Fit mixtures of GPs with Monte Carlo methods.

# Sparse Gaussian Process Inference

## ■ From Bauer et al., 2016:

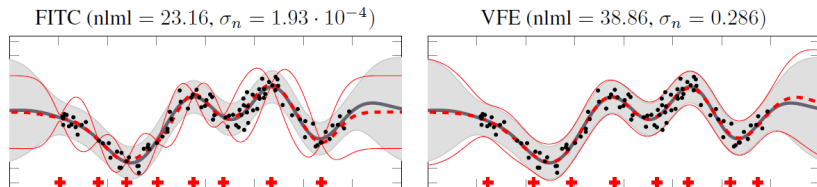
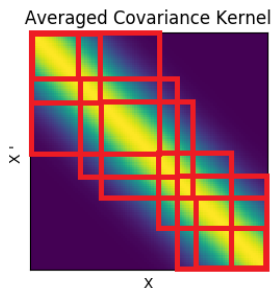
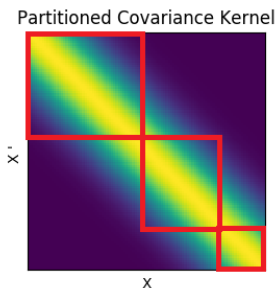
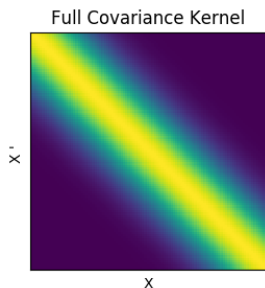


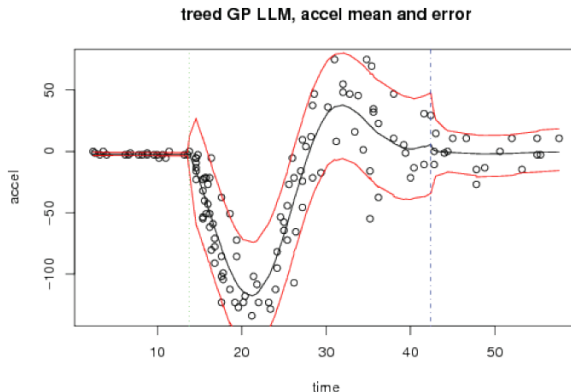
Figure 1: Behaviour of FITC and VFE on a subset of 100 data points of the Snelson dataset for 8 inducing inputs (red crosses indicate inducing inputs; red lines indicate mean and  $2\sigma$ ) compared to the prediction of the full GP in grey. Optimised values for the full GP: nlml = 34.15,  $\sigma_n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

- Idea behind partitioning and averaging:



# Mixtures of Gaussian Processes

- Averaging over partitioned GPs gives us mixture-of-experts.
- Main benefit: Flexible models. Main drawback: Slow, unparallelizable.
- From Gramacy and Lee, 2008:





# 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.

- Suppose we are interested in the integral:

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

but we cannot compute the integral easily.

- I.e. integrating partition from posterior.

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

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

by drawing  $J$  samples from  $q(x)$

- $p/q$  is also known as the importance sampling weight,  $w$

- 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.

- The IS weights each proposed model according to:

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

- Complexity of IGPS:  $O(JN^3/K^2)$ , for  $J$  importance samples.
- Each proposal is independent—embarrassingly parallel inference

- We can approximate the full likelihood by sampling a  $B \ll N$  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 Sampled Mixture of Experts

## ■ The algorithm:

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

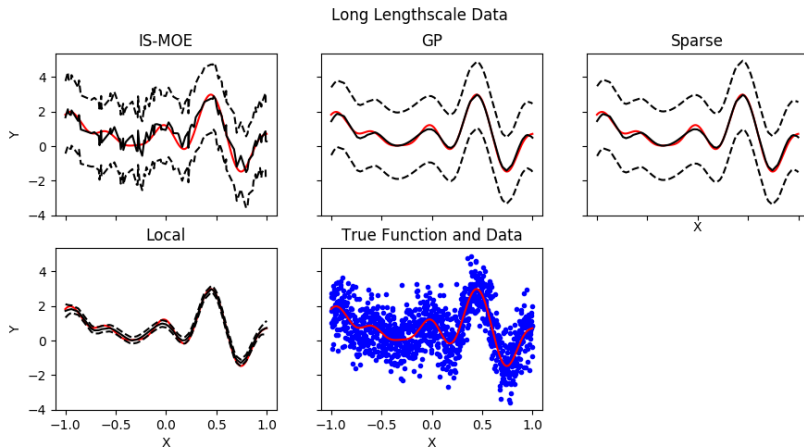
$$P(f_j^*|Z_j, -) = \sum_{k=1}^K P(f_j^*|Z_j^*, -)P(Z_j^*|-)$$

- 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

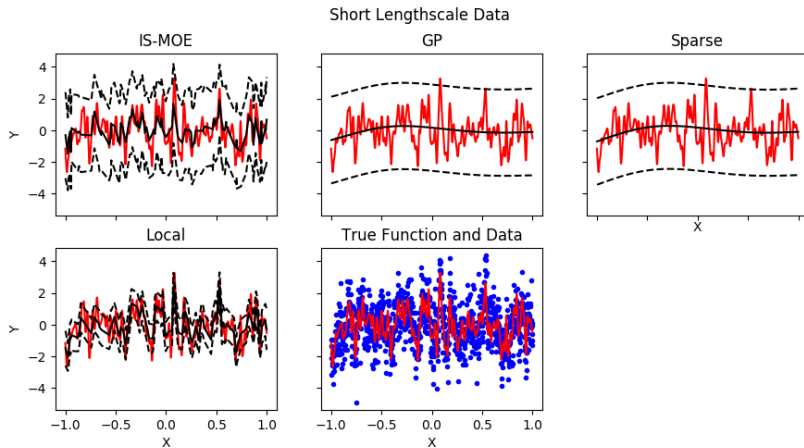
Synthetic data, stationary, long length-scale,  $N = 1000$ ,  $K = 10$ ,  $J = 10$





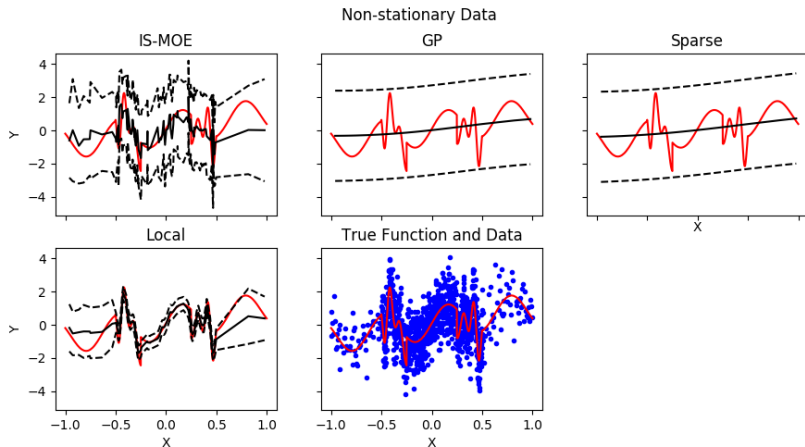
# Results

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



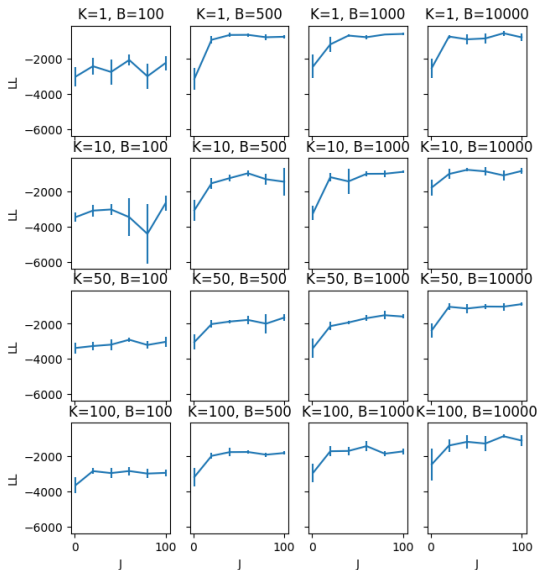
# Results

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



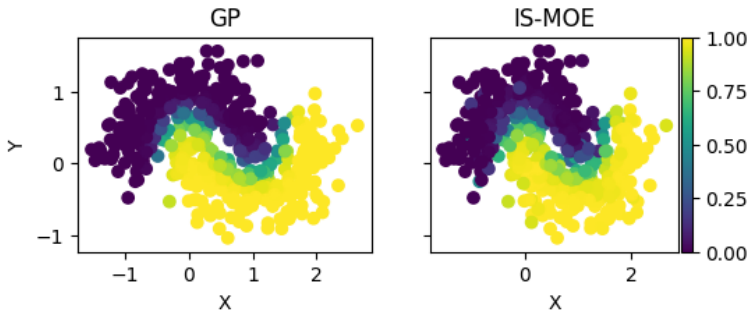
# Results

## Sensitivity to settings



# Classification Results

- Applicable beyond regression:



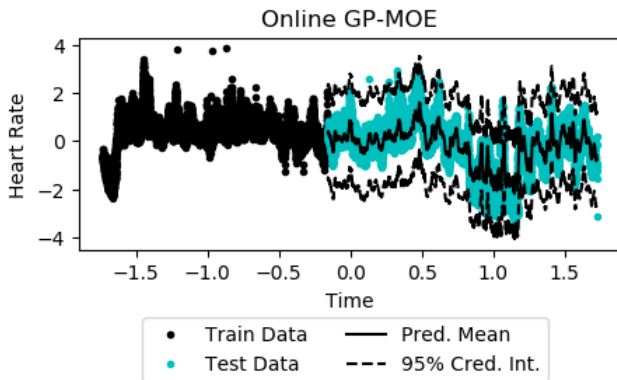
- From “Sequential Gaussian Processes for Online Learning of Nonstationary Functions”.
- Importance Sampling → Sequential Monte Carlo
- Update weight:

$$w_j^{(t)} = w_j^{(t-1)} \cdot \frac{P(Y^{(1:t)}|X^{(1:t)}, Z_j^{(1:t)}, \Theta_j)}{P(Y^{(1:t-1)}|X^{(1:t-1)}, Z_j^{(1:t-1)}, \Theta_j)}$$

- Resample particles when  $ESS < J/2$

# Heart Rate Dataset

- From MIMIC-III dataset:



# Conclusion

- Code and paper available at `michaelzhang01.github.io`.
- GPs are nice models but not scalable.
- Our proposed method is more general, inference can be parallelized.
- Future interest in applying to streaming, multi-output GPs and more complicated models.
- Applications: Patient vital sign tracking.





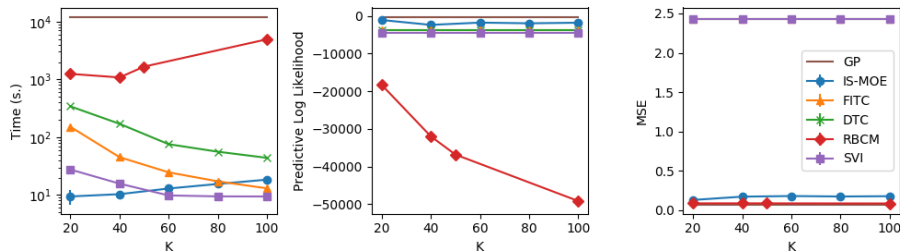
Table: Test set performance on synthetic datasets

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

**Table:** Predictive log likelihood, MSE and total run time for the non-stationary example,  $J = 6000$

	Log Likelihood	MSE	Wall time (s.)
<b>IS-MOE</b>	-154.79	1.28	61.16
BTGP	-145.71	1.04	1116.75
RBCM	-910.55	1.01	13.08

# Timing Results



**Figure:** Wall clock time, predictive log likelihood and MSE results for large synthetic dataset with increasing  $K$ . SVI, FITC and DTC obtain very similar predictive performances.

# Results

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

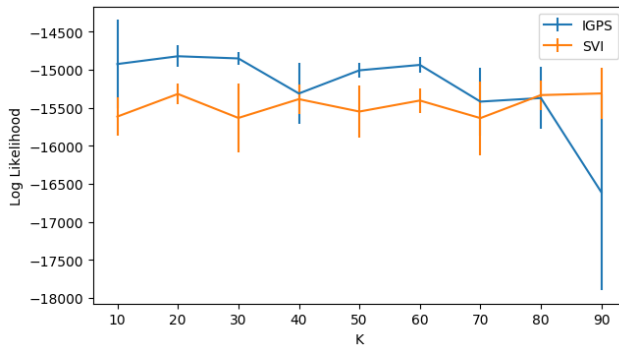
Setting	LL	MSE
IS with SA	-354.96 (28.59)	0.096 (0.003)
IS without SA	-423.93 (41.24)	0.097 (0.002)
Unif. with SA	-726.67 (14.19)	0.19 (0.019)
Unif. without SA	-842.88 (9.82)	0.25 (0.337)

**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,  $N = 209631$ ,  $J = 100$ ,  $B = 1000$ .



# Classification Results

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

Data	GP	IGPS	Sparse GP	Data	GP	IGPS	Sparse GP
Pima	-128.79	-135.09	-128.61	Pima	0.83	0.81	0.83
Parkinsons	-17.00	-22.76	-28.42	Parkinsons	0.86	0.93	0.88
WDBC	-15.50	-12.62	-18.01	WDBC	0.83	0.91	0.81

# Results

Higgs-Boson classification dataset,  
 $N = 1$  million,  $J = 128$ ,  $B = 1000$ ,  $K = 20$ .

