

A Crash Course in Gaussian Processes

May 16, 2019
Thomas Pinder



- 1 Motivation
- 2 Gaussian Processes
- 3 Demonstration
- 4 Problems
- 5 Use Cases of GPs

Motivation

$$\begin{aligned} p(\omega|\mathcal{D}) &= \frac{p(\mathcal{D}|\omega)p(\omega)}{p(\mathcal{D})} \\ &= \frac{p(\mathcal{D}|\omega)p(\omega)}{\int p(\mathcal{D}|\omega)p(\omega)d\omega} \end{aligned}$$



from which we can derive the posterior predictive

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \omega) p(\omega|\mathbf{X}, \mathbf{y})d\omega, \quad (1)$$

where $\mathcal{D} = (\mathbf{X}, \mathbf{y})$.



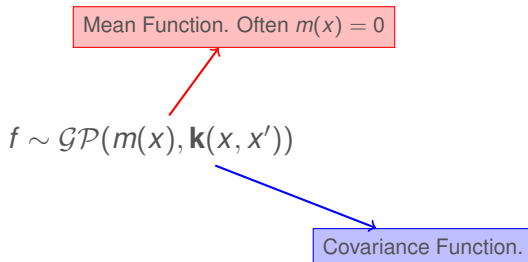
**Trust Me,
I'm An Expert**

Gaussian Processes

What are Gaussian Processes?

Gaussian Processes (GP) can be thought of as prior over functions.

GPs are infinitely parameterised, making them incredibly flexible.



Priors can be expressed through the covariance function and its respective parameters.

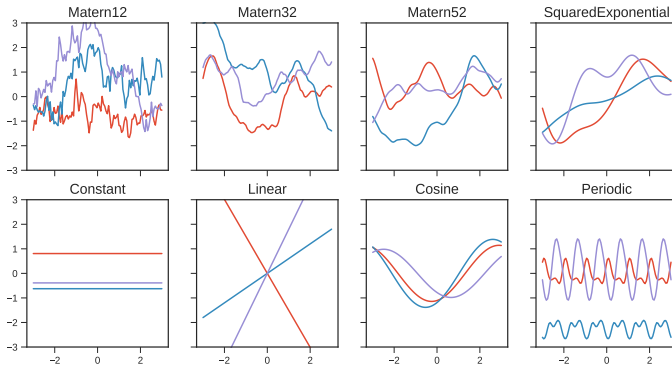


Figure: Series of commonly used covariance functions.

2-Dimensional Visualisation

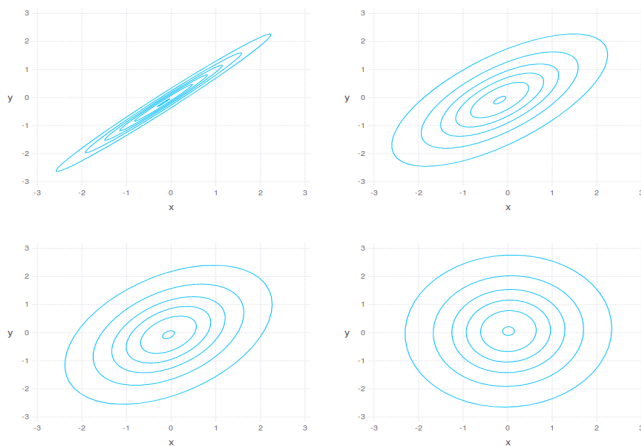


Figure: Range of covariance matrices depicted through ellipses.

We are not limited to just the aforementioned covariance function. We can use any valid kernel function.

Sums of kernels are still valid kernels.

Products of kernels are also still valid kernels.

Example: Matern3/2 + Periodic kernel for time series data with strong periodicities.

Distill.pub

Assume a likelihood $p(y|f)$ and a set of N observations $\mathbf{y} = \{y_1, \dots, y_n\}$ at design locations $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. Our prior is then

$$p(y|x) = \mathcal{N}(y|0, k(x, x)). \quad (2)$$

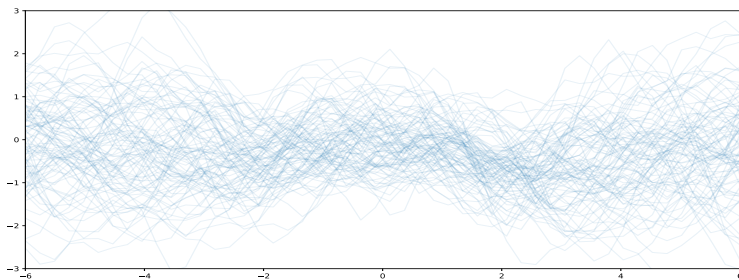


Figure: Prior samples

We observe some new data x_* and wish to predict $y_* = f(x_*)$.

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y}_* \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix} \right). \quad (3)$$

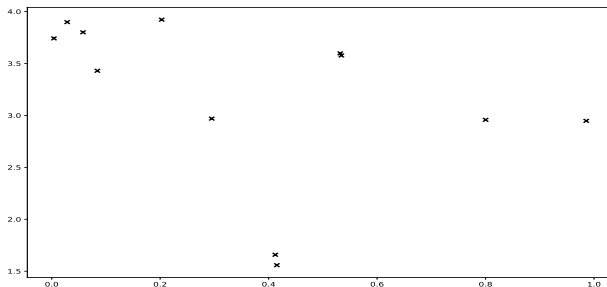


Figure: Observed data

We can then marginalise y_*

$$p(\mathbf{y}_* | \mathbf{x}_*, \mathbf{x}, \mathbf{y}) = \mathcal{N}(\mathbf{y}_* | \mu_*, \Sigma_*) \quad (4)$$

$$\mu_* = \mathbf{K}_*^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \quad (5)$$

$$\Sigma_* = \mathbf{K}_{**} - \mathbf{K}_*^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}_*. \quad (6)$$

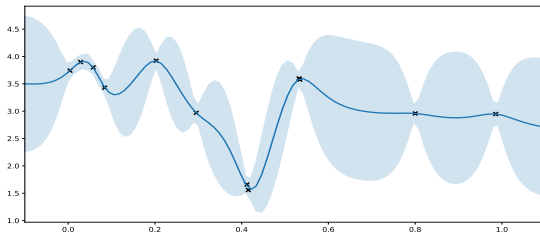


Figure: Posterior Samples

Our GP is parameterised by θ ; the parameters of our mean and covariance function.

With a normal likelihood, we can write down these equations in closed form and maximise the model's likelihood.

We can see this process (Jupyter demo).

Demonstration

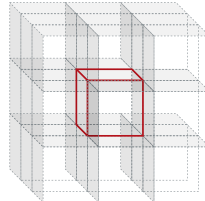
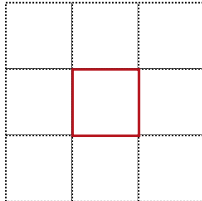
Problems

Returning to Bayes Theorem:

$$\begin{aligned} p(\omega|\mathcal{D}) &= \frac{p(\mathcal{D}|\omega)p(\omega)}{p(\mathcal{D})} \\ &= \frac{p(\mathcal{D}|\omega)p(\omega)}{\int p(\mathcal{D}|\omega)p(\omega)d\omega} \end{aligned}$$

Integration is **really** hard.

Integration is **really** hard.



Modern datasets often contain significantly more than 3 dimensions...

This problem presents itself when the GP's likelihood is non-Gaussian (e.g. Poisson count data, Bernoulli classification).

In these instances we resort to MCMC or, more commonly, variational methods.

Returning to our covariance expression for the GP:

$$\Sigma_* = \mathbf{K}_{**} - \mathbf{K}_*^T (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}_*;$$

Uh ohh...

\mathbf{K} is an $n \times n$ dimensional matrix, meaning that \mathbf{K}^{-1} is of $\mathcal{O}(n^3)$ complexity.

For $n \gtrsim 10000$, this operation is intractable.

A subset of observations \mathbf{Z} such that $|\mathbf{Z}| \ll |\mathbf{y}|$ can be used for regression to reduce the GP's complexity.

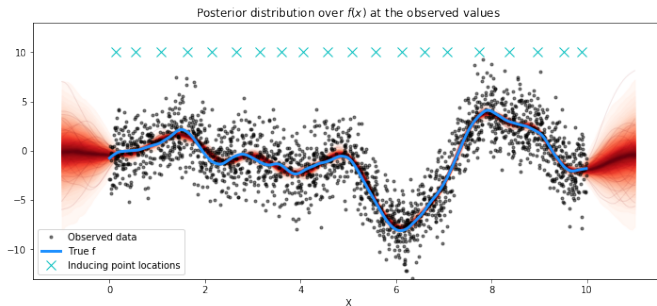


Figure: Example of inducing points

Finding the locations and respective observatory values of \mathbf{Z} is a non-trivial task. Much work has gone into this optimisation task:

- Quiñonero-Candela and Rasmussen, Snelson and Ghahramani
 - FITC and DTC approach.
- Titsias
 - Incorporation of \mathbf{Z} into the ELBO.

Use Cases of GPs



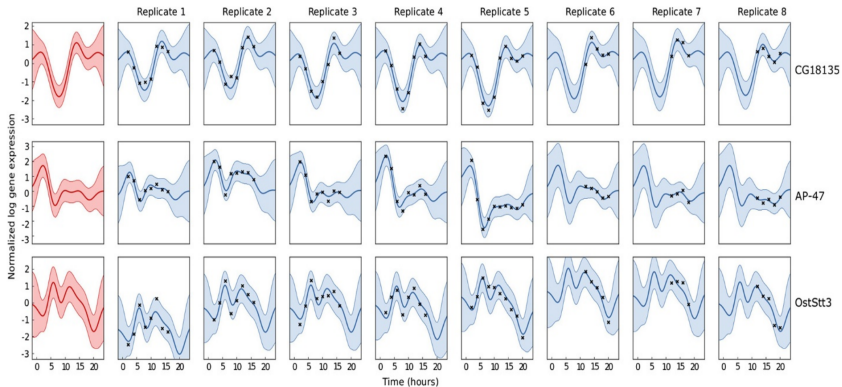


Figure: Depiction of a hierarchical GP applied to gene replication data [2].

A 2-layer hierarchy at time t can be written as

$$\begin{aligned}g_n(t) &\sim \mathcal{GP}(\mathbf{0}, k_g(t, t')); \\f_{nr}(t) &\sim \mathcal{GP}(g_n(t), k_f(t, t'));\end{aligned}$$



Distinct
Kernels

Two points from f_{nr} are jointly Gaussian with 0 mean and covariance $k_g(t, t') + k_f(t, t')$.

Agglomerative clustering and missing data imputation is also possible under this framework.

Deeper hierarchies are a natural extension.

Within spatial statistics we care about conserving the spatial dependency structure of the data.

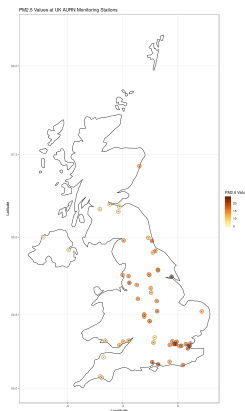


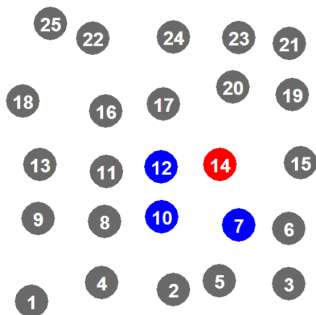
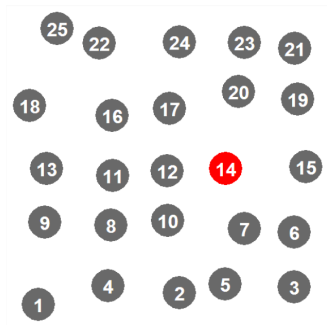
Figure: AURN monitoring stations in the UK, 2014.

Spatial dependencies restricts typical techniques (divide-and-conquer, sparse inputs .etc).

Give a spatial model

$$y(s) = m_{\theta}(s) + w(s) + \epsilon(s) \quad \text{s.t.} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \& \\ w(s) \sim \mathcal{GP}(0, k(s, s')). \quad (7)$$

We can define a neighbourhood set $N(s_i)$ of m points for each spatial location based upon the covariance matrix's top- m correlated locations.



\mathbf{K} is now sparse, meaning that \mathbf{K}^{-1} is tractable through Cholesky decomposition.

Enforcing a neighbourhood structure into a spatial model reduces the complexity to $\mathcal{O}(nm^2)$ where $5 < m < 20$ is usually suffice [1].

Example of the Gibbs sampler computation time on US Biomass data ($n = 10^5$):

Full $\mathcal{GP} = 140hrs$ per iteration

$\mathcal{NNGP} = 6s$ per iteration

Spatiotemporal extensions are natural.




Thank you for listening



Does anyone have any questions?

Contact: t.pinder2@lancaster.ac.uk

ADD TEXT ON COREGIONALISED GPs HERE



-  **Abhirup Datta et al.** “Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets”. In: *Journal of the American Statistical Association* 111.514 (2016), pp. 800–812.
-  **James Hensman, Neil D Lawrence, and Magnus Rattray.** “Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters”. In: *BMC bioinformatics* 14.1 (2013), p. 252.
-  **Joaquin Quiñonero-Candela and Carl Edward Rasmussen.** “A unifying view of sparse approximate Gaussian process regression”. In: *Journal of Machine Learning Research* 6.Dec (2005), pp. 1939–1959.

-  Edward Snelson and Zoubin Ghahramani. “Sparse Gaussian processes using pseudo-inputs”. In: *Advances in neural information processing systems*. 2006, pp. 1257–1264.
-  Michalis Titsias. “Variational learning of inducing variables in sparse Gaussian processes”. In: *Artificial Intelligence and Statistics*. 2009, pp. 567–574.