

# Additive Gaussian Process Models for Spatial and Sptio-temporal Analysis

Sahoko Ishida

Department of Statistics  
London School of Economics and Political Science

June 2022

- 1 Introduction
- 2 Gaussian Process Models
- 3 Spatio-temporal analysis with GP
- 4 Application
- 5 Discussion

# Spatial and spatio-temporal Data

Geostatistic data Set of locations + observed values

- Environmental: pollutant in the air, lake, soils
- Meteorological: temperature, rainfalls

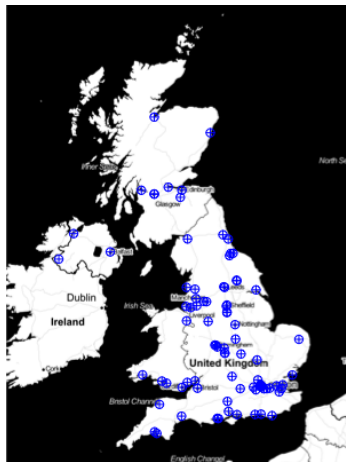
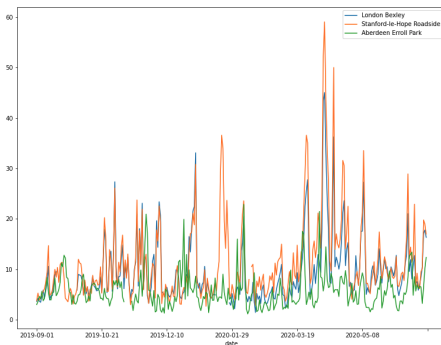


Figure 1: PM<sub>2.5</sub>( $\mu\text{g}/\text{m}^3$ ) - Daily Average



# Spatial and spatio-temporal Data

Measured in area (regions, boroughs, counties)

- Number of crime cases per region
- Number of disease cases

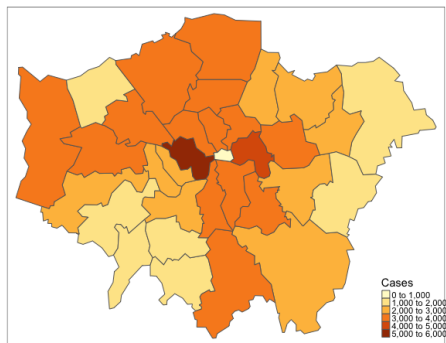


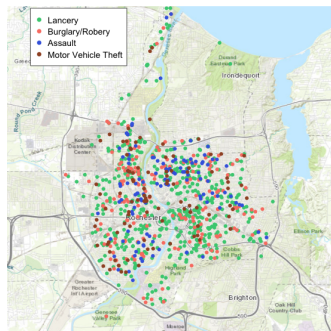
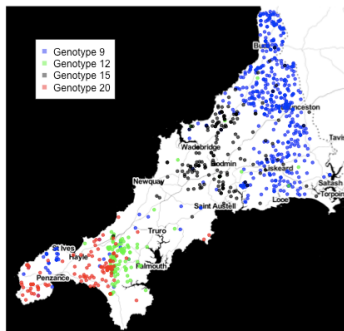
Figure 2: Caption: Crime cases in London 2020-2021 - Monthly average

# Spatial and spatio-temporal Data

Spatial point patterns: location and time of events (univariate / multivariate)

- Bovine Tuberculosis in Cornwall - genotypes responsible for causing the disease
- Crime in Rochester - crime categories

Multivariate spatial point patterns are also called "marked" spatial point patterns



## Common research questions

- Describe spatial patterns of observed (pollutant level, cases, segregation)
- Changes of the pattern over the time
- Prediction at locations where samples were not taken
- Investigate the effect of covariates while controlling for spatial/time effect

## Different models have been analysed by different methods / perspective

- Geo-statistics - Kriging
- Areal - random effect models, conditional auto-regressive model
- Point patterns - Spatial point processes (Inhomogeneous Poisson, Cox etc.)

- ① Provide unified framework for analysing various types of spatial and spatio-temporal data using Gaussian Process
- ② Flexible, interpretable and scalable model
- ③ Proper treatment of spatial and temporal interaction effect

# Outline

- 1 Introduction
- 2 Gaussian Process Models**
- 3 Spatio-temporal analysis with GP
- 4 Application
- 5 Discussion



# Bayesian Linear Regression

- Consider a regression model for  $i = 1, \dots, n$

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

where  $y_i \in \mathbb{R}$ ,  $\mathbf{x}_i \in \mathcal{X}$  and  $(\epsilon_1, \dots, \epsilon_n)^\top \sim \mathbf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ .

- Bayesian Linear Regression

$$f(\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

with a prior

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top \sim \mathbf{N}_p(\mathbf{0}, \mathbf{B})$$

Interests:

- the posterior distribution of the parameters  $\beta$

$$p(\beta|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\beta, \mathbf{X})p(\beta)}{\int p(\mathbf{y}|\beta, \mathbf{X})p(\beta)d\beta}$$

- the predictive distribution for given a new data  $\mathbf{x}_{new}$

$$p(y_{new}|\mathbf{y}, \mathbf{X}, \mathbf{x}_{new}) = \int p(y_{new}|\beta, \mathbf{x}_{new})p(\beta|\mathbf{X}, \mathbf{y})d\beta$$

Idea: Directly put a prior on the function  $f$ . Specifically,

$$f \sim GP(0, k)$$

where  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is some covariance (kernel) function.  
This allows

- Flexible relationship with covariates and response
- Auto-correlation

## Gaussian Process

$f(\cdot)$  is called Gaussian Process(GP), if

$$(f(x_1), \dots, f(x_n)) \sim MVN(\mathbf{f}_0, \mathbf{K})$$

where  $\mathbf{f}_0 = (f_0(x_1), \dots, f_0(x_n))^T$  and  $\mathbf{K}$  is  $n \times n$  matrix with  $(i,j)$ -th element  $k(x_i, x_j) = \text{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j))$

A GP is completely specified by its mean function  $f_0$  and covariance function  $k$ .

# Covariance Functions / Kernel

A covariance function is a positive definite function satisfying for all  $a_1, \dots, a_n \in \mathbb{R}, \mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$

$$\sum_{i,j=1}^n a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

It

- Measures similarity, can be defined in many types of data (strings, graphs, functions)
- Defines a space of function from  $\mathcal{X}$  to  $\mathbb{R}$  (reproducing kernel Hilbert space)

Example : [▶ link](#)

Revisiting the regression model

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

with a GP prior  $f \sim GP(m, k)$ .

Alternative representation of the prior:

$$\begin{aligned}\mathbf{f} = (f(x_1), \dots, f(x_n))^{\top} &\sim \mathbf{N}_n(\mathbf{m}, \mathbf{K}) \\ \mathbf{y} = (y_1, \dots, y_n)^{\top} &\sim \mathbf{N}_n(\mathbf{m}, \mathbf{K} + \sigma^2 \mathbf{I}_n)\end{aligned}$$

# Regression with Gaussian Process Priors

Given a new data  $\mathbf{x}_{new}$ , we have

$$f_{new} | \mathbf{x}_{new} \sim \mathbf{N}(m(\mathbf{x}_{new}), k(\mathbf{x}_{new}, \mathbf{x}_{new})).$$

and

$$\begin{bmatrix} \mathbf{y} \\ f_{new} \end{bmatrix} \sim \mathbf{N}_{n+1} \left( \begin{bmatrix} \mathbf{m} \\ m_{new} \end{bmatrix}, \begin{bmatrix} \mathbf{K} + \sigma^2 \mathbf{I}_n & \mathbf{k}_{new} \\ \mathbf{k}_{new}^\top & k(\mathbf{x}_{new}, \mathbf{x}_{new}) \end{bmatrix} \right).$$

where  $f_{new} = f(\mathbf{x}_{new})$  and  $\mathbf{k}_{new} = (k(\mathbf{x}_1, \mathbf{x}_{new}), \dots, k(\mathbf{x}_n, \mathbf{x}_{new}))^\top$ .

# Regression with Gaussian Process Priors

Predictive distribution is also Gaussian

Using conditional distribution of multivariate normal

$$f_{new} | \mathbf{X}, \mathbf{x}_{new}, \mathbf{y} \sim N \left( m_{new} + \mathbf{k}_{new}^{\top} (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} (\mathbf{y} - \mathbf{m}), \right. \\ \left. k(\mathbf{x}_{new}, \mathbf{x}_{new}) - \mathbf{k}_{new}^{\top} (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{k}_{new} \right).$$

▶ posterior



# Constructing a new kernel from existing ones

Given a positive constant, and valid kernels  $k_1$  and  $k_2$  on  $\mathcal{X}$  all of the below  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  are valid kernel

- Adding a positive constant:

$$k(\mathbf{x}, \mathbf{x}') = \alpha + k_1(\mathbf{x}, \mathbf{x}')$$

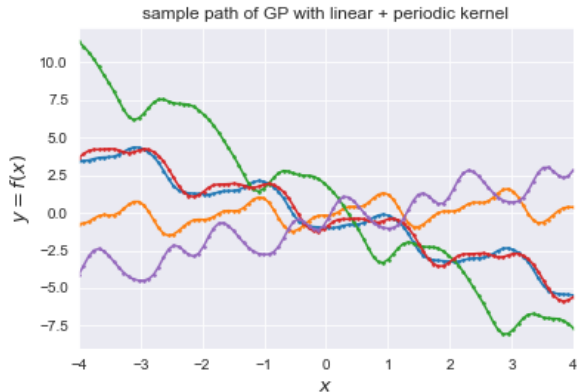
- Sum:

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$$

- Product:

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$$

# Constructing a new kernel from existing ones



# Constructing a new kernel from existing ones

It is not necessary that  $k_1$  and  $k_2$  are defined on the same set.

E.g. we have  $k_1 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  and  $k_2 : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$  then

$$k((\mathbf{x}, \mathbf{s}), (\mathbf{x}', \mathbf{s}')) = 1 + k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{s}, \mathbf{s}') + k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{s}, \mathbf{s}')$$

is a kernel.

# Additive Gaussian Processes

Consider the following regression model for  $i = 1, \dots, n$ :

$$\begin{aligned}y_i &= f(\mathbf{x}_i, \mathbf{s}_i) + \epsilon_i \\f(\mathbf{x}_i, \mathbf{s}_i) &= a + f_x(\mathbf{x}_i) + f_s(\mathbf{s}_i) + f_{xs}(\mathbf{x}_i, \mathbf{s}_i)\end{aligned}$$

with  $a \sim N(0, 1)$  and zero-mean GP priors on each function

$$f_x \sim GP(0, k_1)$$

$$f_s \sim GP(0, k_2)$$

$$f_{xs} \sim GP(0, k_1 k_2)$$

# Additive Gaussian Processes

Overall function  $f$  follows zero-mean GP with kernel defined by

$$k((\mathbf{x}, \mathbf{s}), (\mathbf{x}', \mathbf{s}')) = 1 + k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{s}, \mathbf{s}') + k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{s}, \mathbf{s}')$$

Alternatively, we can write  $\mathbf{f} \sim \mathbf{N}_n(\mathbf{0}, \mathbf{K})$  where

$$\mathbf{K} = \mathbf{1}_n \mathbf{1}_n^\top + \mathbf{K}_x + \mathbf{K}_s + \mathbf{K}_x \odot \mathbf{K}_s$$

where  $\odot$  is an element-wise product operator. 

# GLM with Gaussian Process Priors

Regression with GP priors can be extended to model various types of responses including (ordered or un-ordered) categorical and counts. Given a sample  $(y_i, \mathbf{x}_i)$  for  $i = 1, \dots, n$ , we consider a model

$$g(\mathbb{E}[y_i]) = f(\mathbf{x}_i)$$

where  $g(y)$  is a link function and we put GP prior on  $f$ ,  $f \sim GP(0, k)$ .

## Example - Counts and Poisson likelihood

Suppose that the response variable is counts,  $y_i \in \{0, 1, 2, \dots\}$  with likelihood

$$y_i \sim \text{Poisson}(\lambda_i)$$

Our model is

$$\log(\mathbb{E}(y_i)) = \log(\lambda_i) = f(\mathbf{x}_i)$$

where  $f \sim GP(0, k)$ .

Obtaining predictive distribution  $p(y_{new}|\mathbf{y}, \mathbf{X}, \mathbf{x}_{new})$  is three fold.

- 1 the posterior distribution of  $f$

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f}}$$

- 2 the conditional distribution for  $f(x_{new})$

$$p(f_{new}|\mathbf{y}, \mathbf{X}, \mathbf{x}_{new}) = \int p(f_{new}|\mathbf{f}, \mathbf{x}_{new})p(\mathbf{f}|\mathbf{X}, \mathbf{y})d\mathbf{f}$$

- 3 the predictive distribution for the response  $y_{new}$


$$p(y_{new}|\mathbf{X}, \mathbf{y}, \mathbf{x}_{new}) = \int p(y^*|f^*)p(f_{new}|\mathbf{X}, \mathbf{x}_{new}, \mathbf{y})df_{new}.$$

Due to non-Gaussian likelihood, integrals are no longer analytically tractable.

- 1 Approximate the posterior  $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$  with multivariate Gaussian  $\mathbf{N}_p(\boldsymbol{\mu}, \mathbf{V})$ 
  - Numerical Approximation: MCMC
  - Analytical approximation:
    - Laplace Approximation [▶▶ LA](#)
    - Variational Inference [▶▶ VI](#)
- 2 With Gaussian approximation to the posterior, step 2 now has closed form expression
- 3 Simple Monte Carlo simulation works for step 3



Estimating parameters in kernel is also challenging.

- Hierarchical Bayes (Full Bayes): put priors on hyper-parameters
- Naive Bayes: Hyper-parameters are estimated by maximising (approximated) log marginal likelihood
  - Gaussian likelihood 

$$\log p(\mathbf{y}|\mathbf{f}) = -\frac{1}{2}\mathbf{y}^\top (\mathbf{K}^{-1} + \sigma^2\mathbf{I}_n)^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}^{-1} + \sigma^2\mathbf{I}_n| - \frac{n}{2} \log 2\pi$$

- Non-Gaussian likelihood: log-marginal likelihood needs approximation
  - LA - marginal likelihood under Laplace approximation
  - VB - ELBO itself is lower bound for log-marginal likelihood

# Outline

- 1 Introduction
- 2 Gaussian Process Models
- 3 Spatio-temporal analysis with GP**
- 4 Application
- 5 Discussion

When location information is available in the form of geographical coordinate, e.g.  $\mathbf{s}_i = (\textit{longitude}_i, \textit{latitude}_i)$ , we have

$$y_i = f(\mathbf{s}_i) + \epsilon_i$$

where  $f \sim GP(0, k)$  and kernel  $k : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$  is given by

$$k(\mathbf{s}, \mathbf{s}') = \alpha^2 \exp\left(-\frac{1}{2\rho^2} |\mathbf{s} - \mathbf{s}'|^2\right)$$

# Spatial Models - Areal Data

Location information is more commonly available as areal data

Example:

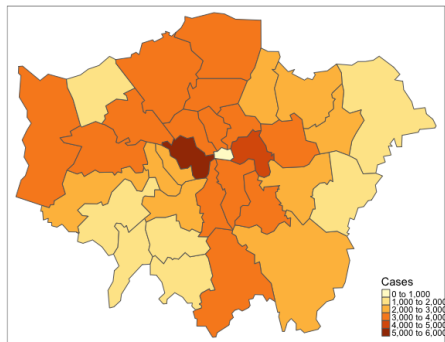


Figure 3: Mean monthly crime cases (2020-2021) per borough

# Spatial Models - Areal Data

- 1 Using centroids (one set of coordinates to represent the location of the area)
- 2 Using weighted adjacency matrix ( $\mathbf{W}$ ) / Graph Laplacian ( $\mathbf{L}$ ) based kernel.
  - Katz kernel

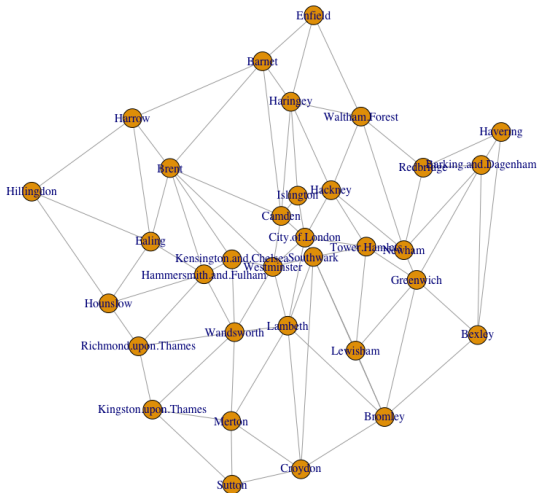
$$\mathbf{K}^{katz} = \sum_{m=0}^{\infty} (\alpha \mathbf{W})^m = [\mathbf{I} - \alpha \mathbf{W}]^{-1}$$

with  $0 \leq \alpha \leq (\rho(\mathbf{W}))^{-1}$  where  $\rho(\mathbf{W})$  is the spectral radius of  $\mathbf{W}$ .  
Regression with katz kernel is similar to conditional auto-regressive (CAR) model.

- Heat kernel

$$\mathbf{K}^{heat} = \exp(-\alpha \mathbf{L}) = \sum_{m=0}^{\infty} \frac{(-\alpha)^m}{m!} \mathbf{L}^m$$

# Spatial Models - Areal Data



# Spatio-temporal Models

Consider the following regression model for  $i = 1, \dots, n$ :

$$y_i = f(\mathbf{s}_i, t_i) + \epsilon_i$$

where  $s_i = (\text{longitude}_i, \text{latitude}_i)$ ,  $t_i = \text{time}_i$  and with a prior

$$f \sim GP(0, k_{st}).$$

Commonly in the literature,  $k_{st}$  is constructed as

- Additive kernel

$$k_{st}((\mathbf{s}, t), (\mathbf{s}', t')) = k_s(\mathbf{s}, \mathbf{s}') + k_t(t, t')$$

- Multiplicative kernel

$$k_{st}((\mathbf{s}, t), (\mathbf{s}', t')) = k_s(\mathbf{s}, \mathbf{s}')k_t(t, t')$$

# Spatio-temporal Models

- Additive kernel: limited when an interaction effect between space and time is present.
- Multiplicative kernel: fail to incorporate main effects

## ANOVA decomposition kernel

We have  $\mathbf{x} = (x_1, \dots, x_d)^\top \in D$  and kernels  $k_l$  for  $l = 1, \dots, d$  each defined on  $D_1, \dots, D_d$ . For  $D = D_1 \times \dots \times D_d$  the ANOVA kernel  $k_{ANOVA} : D \times D \rightarrow \mathbb{R}$  can be constructed as a product of univariate kernels:

$$k_{ANOVA}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^d (1 + k_l(x_l, x'_l))$$

ANOVA kernel includes a constant term, the  $d$ -th order interaction term and any lower order interaction terms.



# Spatio-temporal Models

ANOVA kernel for Spatio-temporal models

$$\begin{aligned}k_{st}((\mathbf{s}, t), (\mathbf{s}', t')) &= (1 + k_s(\mathbf{s}, \mathbf{s}'))(1 + k_t(t, t')) \\ &= 1 + k_s(\mathbf{s}, \mathbf{s}') + k_t(t, t') + k_s(\mathbf{s}, \mathbf{s}')k_t(t, t')\end{aligned}$$

The function  $f(\mathbf{s}_i, t_i)$  can be decomposed as

$$f(\mathbf{s}_i, t_i) = a + f_s(\mathbf{s}_i) + f_t(t_i) + f_{st}(\mathbf{s}_i, t_i).$$

The priors over a constant  $a$  and each function can be specified in the same manner as the previous example [▶ here](#).

The same idea applies to models with covariates (other than spatial or temporal information). For example, with additional covariates  $\mathbf{x}_i \in \mathcal{X}$ , we can consider a model

$$y_i = f(\mathbf{s}_i, t_i; \mathbf{x}_i) + \epsilon_i$$

with GP prior  $f \sim GP(0, k_{stx})$ .

A few possibilities for the structure of the kernel  $k_{stx}$ :

$$k_x(\mathbf{x}, \mathbf{x}') + (1 + k_s(\mathbf{s}, \mathbf{s}'))(1 + k_t(t, t'))$$

$$(1 + k_x(\mathbf{x}, \mathbf{x}'))(1 + k_s(\mathbf{s}, \mathbf{s}'))(1 + k_t(t, t'))$$

The downside of flexibility of regression with GP is the difficulty of interpretability

- Hyper-parameters:
  - does not always have straight-forward interpretation, works more as tuning parameter / controls for flexibility of GP sample path.
  - some cases where parameters in kernel can be understood intuitively - e.g., period parameters in periodic kernel
- Combining linear regression and a regression with GP

Consider for  $\mathbf{x} \in \mathbb{R}^p$

$$f(\mathbf{x}_i, \mathbf{s}_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + f_s(\mathbf{s}_i)$$

where we specify the priors as  $f_s \sim GP(0, k_s)$  and  $\boldsymbol{\beta} \sim \mathbf{N}_p(\mathbf{0}, \mathbf{B})$ .

Combining linear regression and a regression with GP - continued  
We can write our prior on  $f$  as  $GP(0, k_{xs})$  where

$$k_{xs}((\mathbf{x}, \mathbf{s}), (\mathbf{x}', \mathbf{s}')) = \mathbf{x}^\top \mathbf{B} \mathbf{x}' + k_s(\mathbf{s}, \mathbf{s}')$$

then we have

$$\bar{\beta} = (B^{-1} + \mathbf{X}(\mathbf{K}_s + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{X}^\top)^{-1} (\mathbf{X}(\mathbf{K}_s + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y})$$

Computational complexity  $O(n^3)$  associated with inversion of matrix  $(\mathbf{K} + \sigma^2 \mathbf{I}_n)$  for Gaussian likelihood, or  $\mathbf{K}$  for non-Gaussian likelihood.

- Low-rank approximation to  $\mathbf{K}$
- Kronecker algebra - applies to special cases in spatio-temporal data

If the data is collected from a set of fixed  $n_s$  locations over a period of  $n_t$  time stamps (repeated measurement / panel data) then  $nm \times nm$  Gram/kernel matrix can be expressed as a kronecker product of a  $n_s \times n_s$  matrix and a  $n_t \times n_t$  matrix:

$$\mathbf{K} = (\tilde{\mathbf{K}}_s) \otimes (\tilde{\mathbf{K}}_t)$$

where  $\otimes$  is a kronecker product operator,  $\tilde{\mathbf{K}}_s = (\mathbf{1}_{n_s} \mathbf{1}_{n_s}^\top + \mathbf{K}_s)$  and  $\tilde{\mathbf{K}}_t = (\mathbf{1}_{n_t} \mathbf{1}_{n_t}^\top + \mathbf{K}_t)$ .

- Inverting  $\mathbf{K}$  (for Non-Gaussian Likelihood) using Cholesky decomposition  $\mathbf{K}^{-1} = \mathbf{L}^{-1\top} \mathbf{L}^{-1}$

$$\mathbf{L} = \mathbf{L}_s \otimes \mathbf{L}_t$$

- Inverting a (symmetric)  $\mathbf{K} + \sigma^2 \mathbf{I}_n$  (for Gaussian Likelihood) using eigen decomposition  $\mathbf{K}^{-1} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top$

$$(\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} = (\mathbf{Q}_s \otimes \mathbf{Q}_t)(\mathbf{\Lambda}_s \otimes \mathbf{\Lambda}_t + \sigma^2 \mathbf{I}_n)^{-1}(\mathbf{Q}_s^\top \otimes \mathbf{Q}_t^\top)$$

The middle matrix is diagonal matrix, hence inversion only requires  $O(n)$  operation. Except for kronecker product computation  $O((n_s \times n_t)^2) = O(n^2)$  matrix multiplication is not required. [▶ example](#)

# Outline

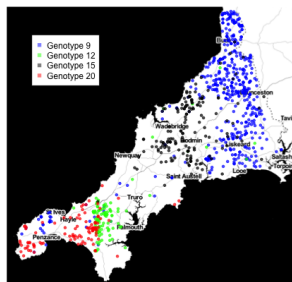
- 1 Introduction
- 2 Gaussian Process Models
- 3 Spatio-temporal analysis with GP
- 4 Application**
- 5 Discussion



# Dataset

Diggle *et al.*(2013): 919 cases of BTB outbreak from 1989 to 2002 in Cornwall, UK.

- Location(2 dimensional spatial coordinates) and year of outbreak
- A mycobacterium bovis genotype which was responsible for the outbreak



Genotypes	Frequency
9	494
12	109
15	166
20	104
total	873

Table 1: Frequency by Genotype

Table 2: 5-fold CV classification error for GP models

Model	kernel	Error
Spatial	SE	0.1408
Spatial	Matérn(1.5)	0.183
Spatial	Matérn(2.5)	0.177
Spatio-temporal	SE+ SE	0.1351
Spatio-temporal	SE*SE	0.4339
Spatio-temporal	(1+SE)(1+SE)	0.1203
Covariates	SE+ SE	0.2130
Covariates	(1+SE)(1+SE)	0.2005

# Outline

- 1 Introduction
- 2 Gaussian Process Models
- 3 Spatio-temporal analysis with GP
- 4 Application
- 5 Discussion**

- Advantages of the proposed method:
  - flexible
  - scalable - using kronecker algebra
  - proper treatment of spatial and temporal interaction effect

Needs further investigation

- More efficient computation (estimation, approximation of kernel matrix)
- Providing tools for easy implementation
- Extension to more types of data - e.g., multivariate response, intensity estimation for univariate spatial/spatio-temporal point patterns
- comparison with related methods (e.g., i-prior)

- squared exponential (S.E.)

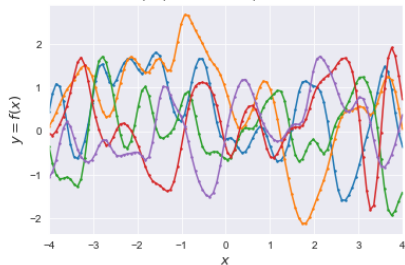
$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\rho^2}\right)$$

- periodic

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{2 \sin^2\left(\frac{\pi\|\mathbf{x}-\mathbf{x}'\|}{p}\right)}{\rho^2}\right)$$

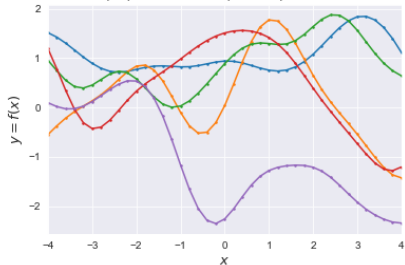
# Kernel Examples

Sample path of GP with periodic kernel

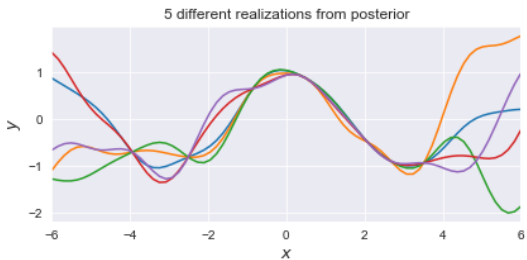
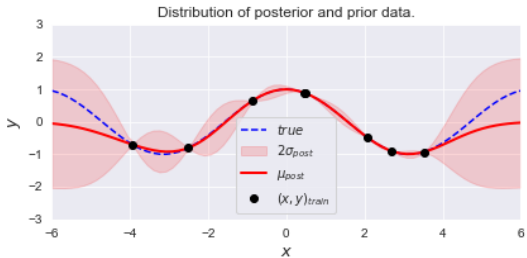


▶ back

Sample path of GP with squared exponential kernel



# Posterior



# Laplace Approximation

Goal: to approximate posterior  $p(\mathbf{f}|\mathbf{y}, \mathbf{X})$  with  $q(\mathbf{f}) \stackrel{D}{=} \mathbf{N}_n(\boldsymbol{\mu}, \mathbf{V})$

The mean vector  $\boldsymbol{\mu}$  is the mode of log-posterior given by

$$\Psi(\mathbf{f}) = -\frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\log |\mathbf{K}| - \frac{n}{2}\log 2\pi + \log p(\mathbf{y}|\mathbf{f}).$$

And the covariance matrix  $\mathbf{V}$  is the inverse of negative Hessian evaluated at the mode. More specifically, we have

$$\begin{aligned}\nabla\Psi(\mathbf{f}) &= \nabla\log p(\mathbf{y}|\mathbf{f}) - \mathbf{K}^{-1}\mathbf{f} \\ \nabla\nabla\Psi(\mathbf{f}) &= \nabla\nabla\log p(\mathbf{y}|\mathbf{f}) - \mathbf{K}^{-1}\end{aligned}$$

The posterior mode can be found by for example, Newton-Raphson Method. [▶ back](#)



Variational Inference aims to find maximizer of Evidence Lower Bound (ELBO)

$$\mathcal{L}(\phi) := -\mathbb{E}_q \left[ \log \frac{q(\mathbf{f}|\phi)}{p(\mathbf{f})} \right] + \mathbb{E}_q [\log p(\mathbf{y}|\mathbf{f})].$$

where  $\phi = (\boldsymbol{\mu}, \mathbf{V})$  represents variational parameters.

The first term has closed form expression

$$-\mathbb{E}_q \left[ \log \frac{q(\mathbf{f}|\phi)}{p(\mathbf{f})} \right] = \frac{1}{2} \log \left[ |\mathbf{V}\mathbf{K}^{-1}| - \text{tr}|\mathbf{V}\mathbf{K}^{-1}| - \mathbf{m}^\top \mathbf{K}^{-1} \mathbf{m} + n \right]$$

while second term does not in general. Further approximation needed

[▶ back](#)

The name evidence (log marginal likelihood) lower bound comes from the fact that

$$\begin{aligned}\log p(\mathbf{y}) &= \log \int p_y(\mathbf{y}|\mathbf{f})p(\mathbf{f})d\mathbf{f} \\ &= \log \int \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{q(\mathbf{f}|\phi)}q(\mathbf{f}|\phi)d\mathbf{f} \\ &\geq \int \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{q(\mathbf{f}|\phi)}q(\mathbf{f}|\phi)d\mathbf{f} \\ &= - \int \log \frac{q(\mathbf{f}|\phi)}{p(\mathbf{f})}q(\mathbf{f}|\phi)d\mathbf{f} + \int \log p(\mathbf{y}|\mathbf{f})q(\mathbf{f}|\phi)d\mathbf{f} \\ &= -\mathbb{E}_{f \sim q}[\log \frac{q(\mathbf{f}|\phi)}{p(\mathbf{f})}] + \mathbb{E}_{f \sim q}[\log p(\mathbf{y}|\mathbf{f})]\end{aligned}$$

Jensen's inequality applies as logarithm is concave.

# Gaussian Process for Categorical Response

For C class categorical response  $y$  we have C class latent function for each observation,

$$\mathbf{f}_i = (f^{(1)}(\mathbf{x}_i), \dots, f^{(C)}(\mathbf{x}_i))^\top$$

Model

$$\pi_i^{(c)} = p(y_i = c | \mathbf{x}_i) = \frac{\exp f^{(c)}(\mathbf{x}_i)}{\sum_{c'=1}^C \exp f^{(c')}(\mathbf{x}_i)}.$$

- Prior:  $f^{(c)} \sim GP(0, k^{(c)})$ ,

$$k^{(c)}(\mathbf{x}, \mathbf{x}') = \lambda_c^2 \exp\left(-\frac{1}{2\rho_c^2} \|\mathbf{x} - \mathbf{x}'\|^2\right)$$

- C latent processes are uncorrelated

# Gaussian Process for Categorical Response

- Given training points  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$  prior over  $\mathbf{f} = (\mathbf{f}_1^\top, \dots, \mathbf{f}_n^\top)^\top$  has a form  $\mathbf{f}|\mathbf{X} \sim \mathbf{MVN}(\mathbf{0}, \mathbf{K})$  where

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}^{(1)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{K}^{(2)} & \dots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{K}^{(C)} \end{bmatrix}.$$

$\mathbf{K}^{(c)}$  is a  $n \times n$  matrix with  $i, j$ -th element equals  $k^c(\mathbf{x}_i, \mathbf{x}_j)$

- Re-coding response variable:

$$\mathbf{y} = (y_1^{(1)}, \dots, y_n^{(1)}, \dots, y_1^{(C)}, \dots, y_n^{(C)})^\top$$

where  $y_i^{(c)} = I(y_i = c)$

Figure 4: Conditional Probability (LGCP model): genotype 9, 12, 15 and 20

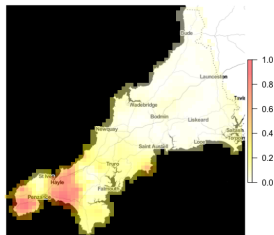
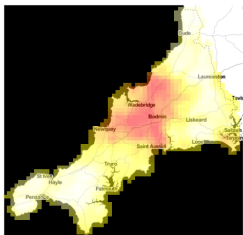
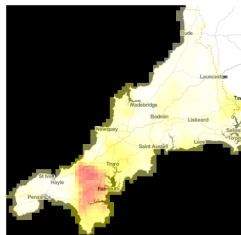
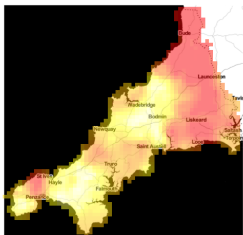


Figure 6: Conditional Probability (GP model): genotype 9, 12, 15 and 20

