MAST90125: Bayesian Statistical learning

Lecture 22 & 23: Introduction to Gaussian processes

Feng Liu and Guoqi Qian

What have we learned so far

- We have learned computational techniques for estimating or approximating posterior distributions when we cannot perform inference analytically. We paid particular attention to MCMC techniques such as,
  - Metropolis-Hastings
  - Gibbs sampling
  - Hamiltonian Monte Carlo.

  and applied these techniques to regression type models, including generalised linear models.
- We will not introduce any further computational techniques for performing Bayesian inference from now on. Rather, we will consider a non-regression model: Gaussian processes.

What is a Gaussian process

- A Gaussian process is a collection of random variables, any finite number of which have Gaussian distribution.
- Mathematically, for any set S, a Gaussan process (GP) on S is a set of random variables $(f_x, x \in S)$ such that, for any $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in S$, $(f_{x_1}, \ldots, f_{x_n})$ is (multivariate) Gaussian.

What is a Gaussian process

- A Gaussian process is a collection of random variables, any finite number of which have Gaussian distribution.
- Mathematically, for any set S, a Gaussan process (GP) on S is a set of random variables $(f_x, x \in S)$ such that, for any $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in S$, $(f_{x_1}, \ldots, f_{x_n})$ is (multivariate) Gaussian.
- (**Gaussian process can be determined by mean function and variance-covariance function**) For any set S, any mean function $\mu : S \to \mathbb{R}$ and any covariance function (also called kernel) $k : S \times S \to \mathbb{R}$, there exists a GP $f(x)$ such that $\mathbb{E}[f(x)] = \mu(x)$, and $cov(f(x_i), f(x_j)) = k(x_i, x_j)\ \forall x_i, x_j \in S$. It denotes $f \sim \mathcal{GP}(\mu, k)$.

What is unique about a Gaussian process

- So what restrictions are placed on x?
  - Does x need to be a scalar? No.
  - Does x need to be observed for the prior to be defined? No.
- So what can we say about $\mu(x)$?
  - $\mu(x)$ is a random function.
  - This in turn highlights how general the Gaussian process is. For example, if x is a scalar, then $\mu(x)$ could be any curve.

What is unique about a Gaussian process

- So what restrictions are placed on x?
  - Does x need to be a scalar? No.
  - Does x need to be observed for the prior to be defined? No.
- So what can we say about $\mu(x)$?
  - $\mu(x)$ is a random function.
  - This in turn highlights how general the Gaussian process is. For example, if x is a scalar, then $\mu(x)$ could be any curve.
- In this lecture, we will first consider the Gaussian process prior. In the next lecture, we will show the inference based on the Gaussian process.

What is a Gaussian process prior

Now, assume a GP model

$$\boldsymbol{y} \sim \mathcal{GP}(\mu, \mathrm{cov}).$$

- What do you think is meant if we write

$$p(\mu) = \mathrm{GP}(m, k)?$$

- It looks like a prior. As you may have guessed, GP stands for Gaussian process, but what is a Gaussian process prior?

$$p(\mu(x)) = \mathcal{N}(m(x), k(x, x')),$$

so $m(x)$ must be the mean of a normal distribution, $k(x, x')$ the variance of a normal distribution.

Where is data involved?

- After defining a Gaussian process prior, we have a wide variety of choices for how observed data $\mathbf{y} = (y_1, \ldots, y_n)$ is generated conditional on $\mathbf{x} = (x_1, \ldots, x_n)$. For instance, we could have
  - The Gaussian process model: $\mathbf{y}|\boldsymbol{\mu}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a variance -covariance matrix. Often $\boldsymbol{\Sigma}$ will simplify to,
    - $\mathbf{y}|\boldsymbol{\mu}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \sigma^2\mathbf{I})$
  - The latent Gaussian process model: $\mathbf{y}|\boldsymbol{f} \sim \mathcal{D}(\boldsymbol{f}); \boldsymbol{f}|\boldsymbol{\mu}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma})$, where $\mathcal{D}$ is some distribution.

- Note: The observed data, $\mathbf{y}$ is a vector of length n . This means $\boldsymbol{\mu}(\mathbf{x})$ is an $n \times 1$ vector, which implies $m(\mathbf{x})$ is an $n \times 1$ vector, and $k(\mathbf{x}, \mathbf{x})$ is an $n \times n$ matrix.

Have we previously encountered Gaussian processes?

- Even though we do not think of these models as Gaussian processes, we have already considered Gaussian processes in this course. Where?
  - Linear models can be viewed as Gaussian process models.
  - Generalised linear models can be viewed as latent Gaussian process models.
- We will now show how linear models can be viewed as Gaussian processes.

Linear models are Gaussian process models

- In lecture 13, we showed the estimates of linear regression correspond to posterior estimates, if we assume
  - Priors: $p(\beta) \propto 1$ and $p(\tau) \propto \tau^{-1}$
  - Likelihood: $p(\mathbf{y}|\mathbf{X}, \beta) = \mathcal{N}(\mathbf{X}\beta, \mathbf{I}/\tau)$.

- From the likelihood statement, we can deduce that $\boldsymbol{\mu}(\mathbf{X}) = \mathbf{X}\beta$. This just leaves us to determine $p(\boldsymbol{\mu}(\mathbf{X}))$.

- In Assignment 1, you were asked to determine the parameters of the improper normal prior that would be equivalent to a flat prior. If you remember, this was $p(\beta) = \mathcal{N}(\beta_0, \boldsymbol{\Sigma})$, as $\boldsymbol{\Sigma}^{-1} \to \mathbf{0}$ and the choice of β_0 was arbitrary.

- Thus linear regression is a Gaussian process model where

$$p(\boldsymbol{\mu}(\mathbf{X})) = \mathcal{N}(m(\mathbf{X}) = \mathbf{X}\beta_0, k(\mathbf{X}, \mathbf{X}) = \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}') \quad \text{as } \boldsymbol{\Sigma}^{-1} \to \mathbf{0}.$$

Linear models are Gaussian process models

- In lecture 14, we considered the case where
  - Priors: $p(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{K}/\tau_\beta)$, $p(\tau) = \text{Ga}(\alpha_e, \gamma_e)$, $p(\tau_\beta) = \text{Ga}(\alpha_\beta, \gamma_\beta)$
  - Likelihood: $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}/\tau)$.

  Further we noted that special cases of this model corresponded to random effect regression/the linear mixed model.

- As with linear regression, we can deduce that $\boldsymbol{\mu}(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ from the likelihood statement. This just leaves us to determine $p(\boldsymbol{\mu}(\mathbf{X}))$.

- From properties of the normal distribution we know that if $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{K}/\tau_\beta)$, then $\mathbf{X}\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_0, \mathbf{X}\mathbf{K}\mathbf{X}'/\tau_\beta)$

- Thus a regression with a normal prior for $\boldsymbol{\beta}$ is a Gaussian process model where

$$p(\boldsymbol{\mu}(\mathbf{X})) = \mathcal{N}(m(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_0, k(\mathbf{X}, \mathbf{X}) = \mathbf{X}\mathbf{K}\mathbf{X}'/\tau_\beta)$$

Linear models are Gaussian process models

- In lecture 14, we also briefly considered the LASSO, which from a Bayesian perspective assumes the prior $p(\beta_j) = \frac{\gamma}{2} e^{-\gamma|\beta_j|}$.
- We noted that this Laplace or double exponential prior can be written as:
  - $p(\beta_j|\sigma_j^2) = \mathcal{N}(0, \sigma_j^2)$
  - $p(\sigma_j^2) = \text{Exp}(\gamma^2/2)$.
- Hence LASSO is a Gaussian process model with

$$p(\boldsymbol{\mu}(\mathbf{X})) = \mathcal{N}(m(\mathbf{X}) = \mathbf{0}, k(\mathbf{X}, \mathbf{X}) = \mathbf{X}\mathbf{K}\mathbf{X}'),$$

where $\mathbf{K}$ is a diagonal matrix such that $\mathbf{K}_{jj} = \sigma_j^2$.

Are Gaussian processes more flexible?

- While we have just shown that linear models are examples of Gaussian processes, do you think Gaussian processes are restricted to linear models?
- The answer is no. We can come up with a wide variety of possible choices for $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x})$. Some possible ideas for $m(\mathbf{x})$ could be:
  - $m(\mathbf{x}) = \sin(\pi \mathbf{x}' \boldsymbol{\beta})$
  - $m(\mathbf{x}) = \exp(-\alpha x_1 / x_2)$ where $\mathbf{x} = (x_1 \;\; x_2)$ and α is some constant.
  - $m(\mathbf{x}) = \alpha x_1^{-x_2}$ where $\mathbf{x} = (x_1 \;\; x_2)$ and α is some constant.
  - $m(x) = \sum_{i=1}^{\infty} \beta_i b_i(x)$ where $b_i(x)$ is some function of x.
  - $m(\mathbf{x}) = 0$, which is very commonly used in practice.

Possible choices for the covariance function.

- We have already showed how linear models can be viewed as Gaussian processes with covariance function,

$$k(\mathbf{X}, \mathbf{X}) = \mathbf{X}\Sigma(\theta)\mathbf{X}',$$

  where $\boldsymbol{\Sigma}$ is an arbitrary positive (semi-)definite matrix, possibly dependent on some additional parameters, $\boldsymbol{\theta}$.
- Other possible choices of covariance function include:
  - White noise, $k(\mathbf{X}_i, \mathbf{X}_{i'}) = \sigma^2 \delta_{\mathbf{X}_i, \mathbf{X}_{i'}}$, where $\delta_{\mathbf{X}_i, \mathbf{X}_{i'}}$ is a Kronecker delta function.
  - Squared exponential, $k(\mathbf{X}_i, \mathbf{X}_{i'}) = \sigma^2 e^{-\sum_{j=1}^{p} (\mathbf{X}_{ij} - \mathbf{X}_{i'j})^2 / l_j^2}$
  - Periodic, $k(t_i, t_{i'}) = \sigma^2 e^{-2 \sin^2(\alpha\pi(t_i - t_{i'}))/l}$

  among others.

Implications of the flexibility of a Gaussian process

- Imagine you want to make predictions of two points, y_i and y_j.
- To make these predictions, you assume $\mathbf{y}$ was generated according to a linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- If the vectors of predictors for observation i, j satisfy $\mathbf{X}_i = \mathbf{X}_j$ for every element, what can you say about the predictions $\hat{y}_i, \hat{y}_j$?
  - The predictions $\hat{y}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} = \mathbf{X}_j\hat{\boldsymbol{\beta}} = \hat{y}_j$ must be identical.
- If the vectors of predictors for observation i, j satisfy $\mathbf{X}_i = \mathbf{X}_j$ for every element, would can you say about the y_i, y_j?
  - According to the assumed model, any difference between y_i and y_j must be due to difference in the residuals $\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j$

Implications of the flexibility of a Gaussian process

- Now imagine you assume data $\mathbf{y}$ was generated according to a Gaussian process, such that for $i = 1, \ldots n$, $y_i = \mu(\mathbf{X}_i) + \epsilon_i$, $\mu(\mathbf{X}_i) \sim \mathcal{N}(m(\mathbf{X}_i), k(\mathbf{X}_i, \mathbf{X}_i))$.
- Based on the model assumed, can we say that if $\mathbf{X}_i = \mathbf{X}_j$ for every element then any difference between y_i and y_j must be due to differences in the residuals ϵ_i, ϵ_j.
  - If y_i, y_j are drawn conditional on the same realisation of a Gaussian process prior, $\boldsymbol{\mu}(\mathbf{X})$, then $\mu(\mathbf{X}_i) = \mu(\mathbf{X}_j)$, if $\mathbf{X}_i = \mathbf{X}_j$ for every element.
- Moreover, if $\mathbf{X}_i = \mathbf{X}_j$ for every element then what can we say about the covariance function $\mathbf{k}(\mathbf{X}, \mathbf{X})$?
  - If $\mathbf{X}_i = \mathbf{X}_j$ are identical, then rows, columns i and j of $\mathbf{k}(\mathbf{X}, \mathbf{X})$ must be identical. This indicates $\mathbf{k}(\mathbf{X}, \mathbf{X})$ is not full-rank, and that elements i and j of $\boldsymbol{\mu}(\mathbf{X})$ are equal.

Implications of the flexibility of a Gaussian process

- On the previous slide, a comment was made about if y_i, y_j are drawn conditional on the same realisation of a Gaussian process prior.

- What does this tell you about $\mu(\mathbf{X})$?
  - $\mu(\mathbf{X})$ is defined for all possible values $\mathbf{X} \in \mathcal{X}$

- What does this tell you about $\mathbf{y}$?
  - $\mathbf{y}$ is conditional on a particular random function evaluated at the points $\mathbf{X}$.

- So if you observe another group of data $\mathbf{y}_2$, and assume the same Gaussian process prior $\mu_2(\mathbf{X}_2) \sim \mathcal{N}(\mathbf{m}(\mathbf{X}_2), \mathbf{k}(\mathbf{X}_2, \mathbf{X}_2))$, what can we say about $\mu_2(\mathbf{X}_2)$?
  - If $\mathbf{y}_2$ is just a continuation of the data $\mathbf{y}$, then $\mu_2(\mathbf{X}_2)$ must be the same function as $\mu(\mathbf{X})$ except evaluated at a different set of points.
  - If $\mathbf{y}_2$ is not a continuation of the data $\mathbf{y}$, then $\mu_2(\mathbf{X}_2)$ would be a different function from $\mu(\mathbf{X})$, even if both are realisations from the same prior.

An example of a Gaussian process

- To conclude this lecture, we generate functions $\boldsymbol{\mu}$ from a Gaussian process. For this example, assume
  - Either $m(x) = \exp(-\alpha x)$.
  - $k(x_i, x_j) = \sigma^2 e^{-\beta \sin^2(\pi(x_i - x_j)/12)}$
- We will consider values for x between $(0, 24)$.
- We will fix σ^2 at one, and vary α, β.

R code for generating $\mu(x)$

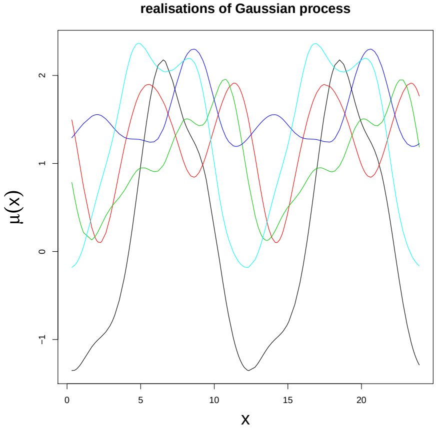
```
#function generating function for Gaussian process prior described on previous slide.
#Inputs are
#x: points where gaussian process was evaluated.
#\alpha: parameter in mean function exp(-\alpha x)
#beta: decay parameter for k
#sigma2: scale parameter for k
#n: number of functions to generate
mu.fun<-function(x,alpha,beta,sigma2,n){
library(mvtnorm)
mx <- exp(-alpha*x) #mean function
np<-length(x)       #number of location to evaluate Gaussian process.
mT<-matrix(x,np,np)
kx<- sigma2*exp(-beta*sin(pi*(mT-t(mT))/12)^2 )

result<-rmvnorm(n,mean=mx,sigma=kx)
return(result)
}

#An example of generating function with $n=5$.
x<-sort(runif(200,0,24)) #generate 200 points for gaussian process to be evaluated at.
test<-mu.fun(x=x,alpha=-0.1,beta=2,sigma2=1,n=5)
#plotting result
plot(x,test[1,],type='l',col=1,ylim=c(min(test),max(test)),ylab=expression(mu(x)),main='realisations of Gaussian process')
for(i in 2:5){lines(x,test[i,],type='l',col=i)}
```

Examples of Gaussian processes

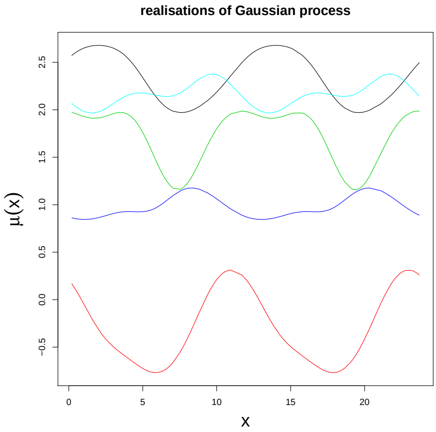
- In this this example, we assume $\alpha = 0$, $\beta = 2$, $\sigma^2 = 1$?



- By setting $\alpha = 0$, we have implied that $m(x) = 1 \; \forall x$.
- Can you see any patterns within each of the five functions?
  - The curves are periodic, with a period of 12. This shows that $k(x, x')$ is not full rank for the range of x values we considered.
- There still appears to be considerable variation in shape between different curves.

Examples of Gaussian processes

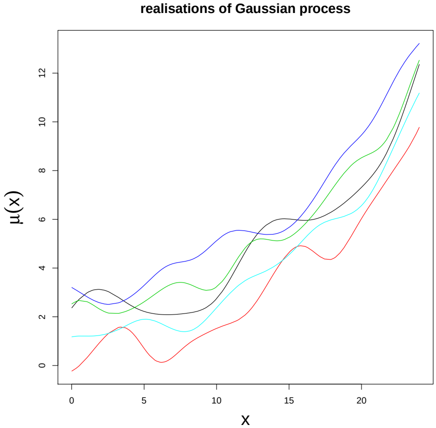
- In this example, we assume $\alpha = 0$, $\beta = 0.3$, $\sigma^2 = 1$?



- Like before, by setting $\alpha = 0$, we have implied that $m(x) = 1\ \forall x$.
- Can you see any patterns within each of the five functions?
  - As expected, the curves are still periodic, with a period of 12.
- By reducing β, we have reduced the rate of decay in correlation. This has reduced variation within each curve $\boldsymbol{\mu}(x)$.

Examples of Gaussian processes

- In this example, we assume $\alpha = -0.1$, $\beta = 1$, $\sigma^2 = 1$?



- By setting $\alpha = -0.1$, we have implied that $m(x)$ will increase with x.
- Can you see any patterns within each of the five functions?
  - By allowing $m(x)$ to be non-constant, the periodicity is more difficult to detect.
- In this particular case, the variation in $m(x)$ likely dominates variation due to $k(x, x')$. The trend in $m(x)$ is clearly seen in each curve generated.