MAST90125: Bayesian Statistical Learning

Lecture 23 & 24: Bayesian inference for Gaussian processes

Feng Liu and Guoqi Qian

A re-cap from the last lecture

- In the last lecture, we introduced the Gaussian process prior, and attempted to summarise some of its features.
- However, we did not perform Bayesian inference for any Gaussian process model. This will be the focus of today's lecture. There are two cases to consider,
    - Where observations $\mathbf{y}$ are noisy, i.e. $\mathbf{y} = \boldsymbol{\mu}(\mathbf{x}) + \epsilon$.
    - Where observations $\mathbf{y}$ are noiseless, i.e. $\mathbf{y} = \boldsymbol{\mu}(\mathbf{x})$.

Noiseless observations

- When dealing with noiseless observations $\mathbf{y} = \boldsymbol{\mu}(\mathbf{x})$, what quantities do we want to make inference on?
- Since $\boldsymbol{\mu}(\mathbf{x})$ is a random function of $\mathbf{x}$, our primary interest will be on $\boldsymbol{\mu}(\mathbf{x})$ at those points $\tilde{\mathbf{x}}$ that have not been observed.
- How would we make inference on this? Remember the Gaussian process prior is defined for all possible values of $\mathbf{x}$, so we can write,

$$p\begin{pmatrix}\boldsymbol{\mu}(\mathbf{x})\\ \boldsymbol{\mu}(\tilde{\mathbf{x}})\end{pmatrix} = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{m}(\mathbf{x})\\ \boldsymbol{m}(\tilde{\mathbf{x}})\end{pmatrix}, \begin{pmatrix}\boldsymbol{k}(\mathbf{x},\mathbf{x}) & \boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}})\\ \boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x}) & \boldsymbol{k}(\tilde{\mathbf{x}},\tilde{\mathbf{x}})\end{pmatrix}\right).$$

- So what are we interested in?
  - The distribution of $\boldsymbol{\mu}(\tilde{\mathbf{x}})$ conditional on $\boldsymbol{\mu}(\mathbf{x})$, $p(\boldsymbol{\mu}(\tilde{\mathbf{x}})|\boldsymbol{\mu}(\mathbf{x}))$.

Predicting $\boldsymbol{\mu}(\tilde{\mathbf{x}})$ in the noiseless case

- For $\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\mu}(\tilde{\mathbf{x}})$, the density function is,

$$p\begin{pmatrix}\boldsymbol{\mu}(\mathbf{x})\\ \boldsymbol{\mu}(\tilde{\mathbf{x}})\end{pmatrix} = \frac{e^{-\frac{\left(\boldsymbol{\mu}(\mathbf{x})' - \boldsymbol{m}(\mathbf{x})' \quad \boldsymbol{\mu}(\tilde{\mathbf{x}})' - \boldsymbol{m}(\tilde{\mathbf{x}})'\right)\begin{pmatrix}\boldsymbol{k}(\mathbf{x},\mathbf{x}) & \boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}})\\ \boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x}) & \boldsymbol{k}(\tilde{\mathbf{x}},\tilde{\mathbf{x}})\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{\mu}(\mathbf{x}) - \boldsymbol{m}(\mathbf{x})\\ \boldsymbol{\mu}(\tilde{\mathbf{x}}) - \boldsymbol{m}(\tilde{\mathbf{x}})\end{pmatrix}}{2}}}{(2\pi)^{\frac{n+\tilde{n}}{2}} \det\begin{pmatrix}\boldsymbol{k}(\mathbf{x},\mathbf{x}) & \boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}})\\ \boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x}) & \boldsymbol{k}(\tilde{\mathbf{x}},\tilde{\mathbf{x}})\end{pmatrix}^{\frac{1}{2}}}.$$

- Based on what we have learned from the course so far, what we need to do is extract the component of the kernel that is a function of $\boldsymbol{\mu}(\tilde{\mathbf{x}})$. However you will note that will require us to determine the blocks of the inverse matrix of $\boldsymbol{k}$.
- To do this, the block matrix inverse formula will help

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\end{pmatrix}.$$

Predicting $\boldsymbol{\mu}(\tilde{\mathbf{x}})$ in the noiseless case

- Using the block matrix inverse formula, the sub-matrices $k^*_{(\mathbf{x},\mathbf{x})}$, $k^*_{(\mathbf{x},\tilde{\mathbf{x}})}$ and $k^*_{(\tilde{\mathbf{x}},\tilde{\mathbf{x}})}$ of the inverse of $\boldsymbol{k} = \begin{pmatrix} \boldsymbol{k}(\mathbf{x},\mathbf{x}) & \boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}}) \\ \boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x}) & \boldsymbol{k}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}) \end{pmatrix}$ are:

$$
\begin{aligned}
\boldsymbol{k}^*_{(\mathbf{x},\mathbf{x})} &= \boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1} + \boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}})(\boldsymbol{k}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}) - \boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x})\boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}}))^{-1}\boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x})\boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1} \\
\boldsymbol{k}^*_{(\mathbf{x},\tilde{\mathbf{x}})} &= -\boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}})(\boldsymbol{k}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}) - \boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x})\boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}}))^{-1} \\
\boldsymbol{k}^*_{(\tilde{\mathbf{x}},\tilde{\mathbf{x}})} &= (\boldsymbol{k}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}) - \boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x})\boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}}))^{-1}
\end{aligned} \tag{1}
$$

- Substituting the results in (1) into $p\begin{pmatrix}\boldsymbol{\mu}(\mathbf{x}) \\ \boldsymbol{\mu}(\tilde{\mathbf{x}})\end{pmatrix}$ and extracting the component of the joint kernel that is a function of $\boldsymbol{\mu}(\tilde{\mathbf{x}})$, we obtain:

$$
e^{-\frac{(\boldsymbol{\mu}(\tilde{\mathbf{x}})-\boldsymbol{m}(\tilde{\mathbf{x}}))'(\boldsymbol{k}(\tilde{\mathbf{x}},\tilde{\mathbf{x}})-\boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x})\boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}}))^{-1}(\boldsymbol{\mu}(\tilde{\mathbf{x}})-\boldsymbol{m}(\tilde{\mathbf{x}}))-2(\boldsymbol{\mu}(\mathbf{x})-\boldsymbol{m}(\mathbf{x}))'\boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}})(\boldsymbol{k}(\tilde{\mathbf{x}},\tilde{\mathbf{x}})-\boldsymbol{k}(\tilde{\mathbf{x}},\mathbf{x})\boldsymbol{k}(\mathbf{x},\mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x},\tilde{\mathbf{x}}))^{-1}(\boldsymbol{\mu}(\tilde{\mathbf{x}})-\boldsymbol{m}(\tilde{\mathbf{x}}))}{2}} \tag{2}
$$

Predicting $\boldsymbol{\mu}(\tilde{\mathbf{x}})$ in the noiseless case

- From the kernel in (2), we can deduce that,

$$\boldsymbol{\mu}(\tilde{\mathbf{x}}) - \boldsymbol{m}(\tilde{\mathbf{x}}) | \boldsymbol{\mu}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{k}(\tilde{\mathbf{x}}, \mathbf{x})\boldsymbol{k}(\mathbf{x}, \mathbf{x})^{-1}(\boldsymbol{\mu}(\mathbf{x}) - \boldsymbol{m}(\mathbf{x})), \boldsymbol{k}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - \boldsymbol{k}(\tilde{\mathbf{x}}, \mathbf{x})\boldsymbol{k}(\mathbf{x}, \mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x}, \tilde{\mathbf{x}})).$$

- Which means that the posterior distribution of $\boldsymbol{\mu}(\tilde{\mathbf{x}})$ is,

$$\boldsymbol{\mu}(\tilde{\mathbf{x}}) \sim \mathcal{N}(\boldsymbol{m}(\tilde{\mathbf{x}}) + \boldsymbol{k}(\tilde{\mathbf{x}}, \mathbf{x})\boldsymbol{k}(\mathbf{x}, \mathbf{x})^{-1}(\boldsymbol{\mu}(\mathbf{x}) - \boldsymbol{m}(\mathbf{x})), \boldsymbol{k}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) - \boldsymbol{k}(\tilde{\mathbf{x}}, \mathbf{x})\boldsymbol{k}(\mathbf{x}, \mathbf{x})^{-1}\boldsymbol{k}(\mathbf{x}, \tilde{\mathbf{x}})).$$

Noisy observations

- When dealing with noisy observations $\mathbf{y} = \boldsymbol{\mu}(\mathbf{x}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, what quantities do we want to make inference on?

- Since the random function $\boldsymbol{\mu}(\mathbf{x})$ at the points $\mathbf{x}$ may not be known, presumably we are interested in predicting the random function at the observed points $\mathbf{x}$ as well as at points that have not been observed.

- We know from how the model has been set up that

$$
\begin{aligned}
p(\mathbf{y}|\boldsymbol{\mu}(\mathbf{x})) &= \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma}) \\
p(\boldsymbol{\mu}(\mathbf{x})) &= \mathcal{N}(\boldsymbol{m}(\mathbf{x}), \boldsymbol{k}(\mathbf{x}, \mathbf{x})),
\end{aligned}
$$

which implies that

$$
p(\mathbf{y}) = \mathcal{N}(\boldsymbol{m}(\mathbf{x}), \boldsymbol{k}(\mathbf{x}, \mathbf{x}) + \boldsymbol{\Sigma})
$$

Noisy observations

- Hence we can work with the joint density of $\mathbf{y}$ and $\boldsymbol{\mu}(\tilde{\mathbf{x}})$, just like how we worked with the joint density of $\boldsymbol{\mu}(\mathbf{x})$ and $\boldsymbol{\mu}(\tilde{\mathbf{x}})$ in the noiseless case. The joint distribution is $\mathbf{y}$ and $\boldsymbol{\mu}(\tilde{\mathbf{x}})$ is,

$$p\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\mu}(\tilde{\mathbf{x}}) \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{m}(\mathbf{x}) \\ \boldsymbol{m}(\tilde{\mathbf{x}}) \end{pmatrix}, \begin{pmatrix} \boldsymbol{k}(\mathbf{x}, \mathbf{x}) + \boldsymbol{\Sigma} & \boldsymbol{k}(\mathbf{x}, \tilde{\mathbf{x}}) \\ \boldsymbol{k}(\tilde{\mathbf{x}}, \mathbf{x}) & \boldsymbol{k}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) \end{pmatrix} \right).$$

- Note: The set of points we want to make predictions at $\tilde{\mathbf{x}}$ can include points where we have noisy observations, $\mathbf{y}$.

What else do you want to make inference on?

- In determining the posterior distribution for $\boldsymbol{\mu}(\tilde{\mathbf{x}})$, what did we implicitly assume?
    - That $\boldsymbol{m}(\mathbf{x})$ and $\boldsymbol{k}(\mathbf{x}, \mathbf{x})$ were known.
- If we were dealing with noisy observations, if there is anything we want to make inference on?
    - The variance-covariance matrix $\boldsymbol{\Sigma}$.
- We will now discuss how to perform Bayesian inference for these parameters. For this, we will assume $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ and $\mathbf{y}$ is noisy.
    - In doing this, we will focus on the component of $p(\mathbf{y}|\boldsymbol{\mu}(\mathbf{x}), \sigma^2)p(\boldsymbol{\mu}(\mathbf{x})|\boldsymbol{m}(\mathbf{x}), \boldsymbol{k}(\mathbf{x}, \mathbf{x}))$ that is a function of the additional parameter of interest. We will then discuss whether this has a form that lends itself to conjugacy.

How would you make inference on $\boldsymbol{m}(\mathbf{x})$

- If we want to make inference on $\boldsymbol{m}(\mathbf{x})$, we can either marginalise $\boldsymbol{\mu}(\mathbf{x})$ out or not.
  - If we marginalise out $\boldsymbol{\mu}(\mathbf{x})$, we are dealing with the likelihood $p(\mathbf{y}|\boldsymbol{m}(\mathbf{x}), \boldsymbol{k}(\mathbf{x}, \mathbf{x})) = \mathcal{N}(\boldsymbol{m}(\mathbf{x}), \boldsymbol{k}(\mathbf{x}, \mathbf{x}) + \sigma^2\mathbf{I})$.
  - If we do not marginalise out $\boldsymbol{\mu}(\mathbf{x})$, we are dealing with the Gaussian process prior $p(\boldsymbol{\mu}(\mathbf{x})|\boldsymbol{m}(\mathbf{x}), \boldsymbol{k}(\mathbf{x}, \mathbf{x})) = \mathcal{N}(\boldsymbol{m}(\mathbf{x}), \boldsymbol{k}(\mathbf{x}, \mathbf{x}))$.

- Can you see any problems?
  - It is likely that $\boldsymbol{k}(\mathbf{x}, \mathbf{x})$ will have parameters that require estimation. Therefore it will be easier to work with conditional posteriors $p(\boldsymbol{m}(\mathbf{x})|\boldsymbol{k}(\mathbf{x}, \mathbf{x}), \cdot)$.
    - By implication, this suggests we want to construct a Gibbs sampler.
  - What about the prior for $p(\boldsymbol{m}(\mathbf{x}))$?
    - The choice of prior for $p(\boldsymbol{m}(\mathbf{x}))$ will depend on whether you assume $\boldsymbol{m}(\mathbf{x})$ is parametric $\boldsymbol{m}(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta})$ or not. If you assume a parametric form $\boldsymbol{m}(\mathbf{x})$, you would want a prior for $\boldsymbol{\theta}$.

How would you make inference on σ^2

- Making inference for σ^2 will be very similar to making inference for the residual variance in regression. To see why, consider the likelihood

$$p(\mathbf{y}|\boldsymbol{\mu}(\mathbf{x}),\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\boldsymbol{\mu}(x_i))^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{(\mathbf{y}-\boldsymbol{\mu}(\mathbf{x}))'(\mathbf{y}-\boldsymbol{\mu}(\mathbf{x}))}{2\sigma^2}}.$$

- If we work with the precision, $\tau = (\sigma^2)^{-1}$, we would get the kernel of a gamma distribution,

$$p(\mathbf{y}|\boldsymbol{\mu}(\mathbf{x}),\tau) \propto \tau^{\frac{n}{2}} e^{-\frac{\tau(\mathbf{y}-\boldsymbol{\mu}(\mathbf{x}))'(\mathbf{y}-\boldsymbol{\mu}(\mathbf{x}))}{2}},$$

which means if we assume a gamma prior for τ, we will obtain a Gamma conditional posterior,

$$p(\tau|\mathbf{y},\boldsymbol{\mu}(\mathbf{x}),\boldsymbol{k}(\mathbf{x},\mathbf{x})) = \mathrm{Ga}(\alpha+n/2,\beta+(\mathbf{y}-\boldsymbol{\mu}(\mathbf{x}))'(\mathbf{y}-\boldsymbol{\mu}(\mathbf{x}))/2).$$

How would you make inference on $\boldsymbol{k}(\mathbf{x},\mathbf{x})$

- Typically it will be assumed that $\boldsymbol{k}(\mathbf{x},\mathbf{x})$ can be written as,

$$\boldsymbol{k}(\mathbf{x},\mathbf{x}) = \sigma_K^2 \boldsymbol{g}(\mathbf{x},\mathbf{x},\boldsymbol{\theta}),$$

  where σ_K^2 is a scale parameter, and $g(\mathbf{x},\mathbf{x},\boldsymbol{\theta})$ controls correlation between different elements.

- Making inference on σ_K^2 is just like making inference for a variance component in random regression. To see why, extract the component of the Gaussian process prior that is a function of σ_K^2,

$$p(\boldsymbol{\mu}(\mathbf{x})|\boldsymbol{m}(\mathbf{x}),\boldsymbol{g}(\mathbf{x},\mathbf{x},\boldsymbol{\theta}),\sigma_K^2) = \frac{1}{(2\pi\sigma_K^2)^{r/2}\det(\boldsymbol{g}(\mathbf{x},\mathbf{x},\boldsymbol{\theta}))^{1/2}} e^{-\frac{(\boldsymbol{\mu}(\mathbf{x})-\boldsymbol{m}(\mathbf{x}))'\boldsymbol{g}(\mathbf{x},\mathbf{x},\boldsymbol{\theta})^{-}(\boldsymbol{\mu}(\mathbf{x})-\boldsymbol{m}(\mathbf{x}))}{2\sigma_K^2}},$$

  where r is the rank of the matrix $\boldsymbol{g}(\mathbf{x},\mathbf{x},\boldsymbol{\theta})$.

How would you make inference on $\boldsymbol{k}(\mathbf{x}, \mathbf{x})$: σ_K^2

- Just like in the case of σ^2, if we work with the precision $\tau_K = (\sigma_K^2)^{-1}$, we can extract the kernel of a gamma distribution,

$$p(\boldsymbol{\mu}(\mathbf{x})|\boldsymbol{m}(\mathbf{x}), \boldsymbol{g}(\mathbf{x}, \mathbf{x}, \boldsymbol{\theta}), \tau_K) \quad \propto \quad \tau_K^{r/2} e^{-\frac{\tau_K(\boldsymbol{\mu}(\mathbf{x})-\boldsymbol{m}(\mathbf{x}))' \boldsymbol{g}(\mathbf{x},\mathbf{x},\boldsymbol{\theta})^{-}(\boldsymbol{\mu}(\mathbf{x})-\boldsymbol{m}(\mathbf{x}))}{2}}.$$

which means if we assume a gamma prior for τ_K, we will obtain a Gamma conditional posterior,

$$p(\tau_K|\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{m}(\mathbf{x}), \boldsymbol{g}(\mathbf{x}, \mathbf{x}, \boldsymbol{\theta})) = \mathrm{Ga}(\alpha_K + \frac{r}{2}, \beta_K + \frac{(\boldsymbol{\mu}(\mathbf{x}) - \boldsymbol{m}(\mathbf{x}))' \boldsymbol{g}(\mathbf{x}, \mathbf{x}, \boldsymbol{\theta})^{-}(\boldsymbol{\mu}(\mathbf{x}) - \boldsymbol{m}(\mathbf{x}))}{2}).$$

How would you make inference on $\boldsymbol{k}(\mathbf{x},\mathbf{x}) : g(\mathbf{x},\mathbf{x},\boldsymbol{\theta})$

- Unlike with σ^2, σ_K^2 or $\boldsymbol{\mu}(\mathbf{x})$, you cannot guarantee that $g(\mathbf{x},\mathbf{x},\boldsymbol{\theta})$ will be in a form such that you will see any conjugacy properties.

- Moreover, if we consider the component of the joint distribution that is a function of $g(\mathbf{x},\mathbf{x},\boldsymbol{\theta})$,

$$p(\boldsymbol{\mu}(\mathbf{x})|\boldsymbol{m}(\mathbf{x}),\boldsymbol{g}(\mathbf{x},\mathbf{x},\boldsymbol{\theta}),\tau_K) \quad \propto \quad \tau_K^{r/2} e^{-\frac{\tau_K(\boldsymbol{\mu}(\mathbf{x})-\boldsymbol{m}(\mathbf{x}))'\boldsymbol{g}(\mathbf{x},\mathbf{x},\boldsymbol{\theta})^{-}(\boldsymbol{\mu}(\mathbf{x})-\boldsymbol{m}(\mathbf{x}))}{2}},$$

  you will notice that it is the inverse that appears, rather than $g(\mathbf{x},\mathbf{x},\boldsymbol{\theta})$.

- To get around this, we would use a Metropolis step within the overall Gibbs sampler to update the parameters $\boldsymbol{\theta}$.

Shifting to R

- To conclude this lecture, we will simulate a noisy Gaussian process.
- We will then attempt to estimate the parameters using the framework outlined on the previous slides.