STAT 576 Bayesian Analysis

Lecture 13: Nonparametric Models

Chencheng Cai

Washington State University

Prior Assumptions

Given a set of paired data $(x_1, y_1), \ldots, (x_n, y_n)$, we often assume that the expected value of y is a function of x:

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

for some unknown function f. The ϵ_i's are assumed to have mean zero.

Prior Assumptions

Given a set of paired data $(x_1, y_1), \ldots, (x_n, y_n)$, we often assume that the expected value of y is a function of x:

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

for some unknown function f. The ϵ_i's are assumed to have mean zero.

- In parametric models (e.g. linear regressions), we assume that μ belongs to a parametric family, e.g. $\mu(x) = \beta_0 + \beta_1 x$. A prior is often placed on the parameters β_0 and β_1.

Prior Assumptions

Given a set of paired data $(x_1, y_1), \ldots, (x_n, y_n)$, we often assume that the expected value of y is a function of x:

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

for some unknown function f. The ϵ_i's are assumed to have mean zero.

- In parametric models (e.g. linear regressions), we assume that μ belongs to a parametric family, e.g. $\mu(x) = \beta_0 + \beta_1 x$. A prior is often placed on the parameters β_0 and β_1.
- In state-space models, the prior is placed on the state variables $(x_1, \ldots, x_n)$ through a transition model.

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}).$$

Prior Assumptions

Given a set of paired data $(x_1, y_1), \ldots, (x_n, y_n)$, we often assume that the expected value of y is a function of x:

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

for some unknown function f. The ϵ_i's are assumed to have mean zero.

- In parametric models (e.g. linear regressions), we assume that μ belongs to a parametric family, e.g. $\mu(x) = \beta_0 + \beta_1 x$. A prior is often placed on the parameters β_0 and β_1.
- In state-space models, the prior is placed on the state variables $(x_1, \ldots, x_n)$ through a transition model.

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}).$$

- In this lecture, we will discuss nonparametric models, where the prior is placed directly on the function μ within a function space.

Gaussian Process

- A **stochastic process** is a collection of random variables indexed by some set, e.g. time or space.
  - Random walk: r.v.s. indexed by time.
  - Brownian motion: r.v.s. indexed by time.
  - Random field: r.v.s. indexed by space.

Gaussian Process

- A **stochastic process** is a collection of random variables indexed by some set, e.g. time or space.
  - Random walk: r.v.s. indexed by time.
  - Brownian motion: r.v.s. indexed by time.
  - Random field: r.v.s. indexed by space.
- A **Gaussian process** is a stochastic process such that any finite collection of r.v.s. has a multivariate normal distribution.

Gaussian Process

- A **stochastic process** is a collection of random variables indexed by some set, e.g. time or space.
  - Random walk: r.v.s. indexed by time.
  - Brownian motion: r.v.s. indexed by time.
  - Random field: r.v.s. indexed by space.
- A **Gaussian process** is a stochastic process such that any finite collection of r.v.s. has a multivariate normal distribution.
- Specifically, if $\{\mu(x) : x \in \mathcal{X}\}$ is a Gaussian process, then for any finite set of indices $x_1, \ldots, x_n \in \mathcal{X}$, the random vector $(\mu(x_1), \ldots, \mu(x_n))$ has a multivariate normal distribution.
- As a special case, for any $x \in \mathcal{X}$, $\mu(x)$ is a normal random variable.

Gaussian Process

- A Gaussian process is completely specified by its mean function $m(x) = E[\mu(x)]$ and covariance function $k(x, x') = \text{Cov}(\mu(x), \mu(x'))$.
- The process is denoted by $\mu(x) \sim \mathcal{GP}(m, k)$.

Gaussian Process

- A Gaussian process is completely specified by its mean function $m(x) = E[\mu(x)]$ and covariance function $k(x, x') = \mathrm{Cov}(\mu(x), \mu(x'))$.
- The process is denoted by $\mu(x) \sim \mathcal{GP}(m, k)$.
- The joint distribution of $\mu(x_1), \ldots, \mu(x_n)$ is given by

$$\begin{pmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}, K(x_1, \ldots, x_n) \right).$$

where $K(x_1, \ldots, x_n)$ is the covariance matrix with (i, j)-th element $k(x_i, x_j)$.

Gaussian Process

- A Gaussian process is completely specified by its mean function $m(x) = E[\mu(x)]$ and covariance function $k(x, x') = \text{Cov}(\mu(x), \mu(x'))$.
- The process is denoted by $\mu(x) \sim \mathcal{GP}(m, k)$.
- The joint distribution of $\mu(x_1), \ldots, \mu(x_n)$ is given by

$$\begin{pmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix}, K(x_1, \ldots, x_n) \right).$$

  where $K(x_1, \ldots, x_n)$ is the covariance matrix with (i, j)-th element $k(x_i, x_j)$.
- Consistent defintion: the distribution of $\mu(x_1), \ldots, \mu(x_m)$ derived from the joint distribution of $\mu(x_1), \ldots, \mu(x_n)$ is the same for any choice of $x_{m+1}, \ldots, x_n$.
- $K(x_1, \ldots, x_n)$ is positive definite for any choice of $x_1, \ldots, x_n$.

Gaussian Process — Covariance

A common choice for $k(x, x')$ is

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

Gaussian Process — Covariance

A common choice for $k(x, x')$ is

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

Show that the covariance matrix $K(x_1, \ldots, x_n)$ is positive definite.

Gaussian Process — Covariance

A common choice for $k(x, x')$ is

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

Show that the covariance matrix $K(x_1, \ldots, x_n)$ is positive definite.

WLOG, we assume $\tau^2 = l^2 = 1$. To show K is positive definite, we need to show that for any vector $u = (u_1, \ldots, u_n)$, $u^T K u \geq 0$.

- Notice that

$$k(x_i, x_j) = \exp\left(-\frac{1}{2}|x_i - x_j|^2\right) = \mathbb{E}\left[e^{i|x_i - x_j|Z}\right]$$

  for $Z \sim \mathcal{N}(0, 1)$.

- Therefore,

$$u^T K u = \sum_{i,j} u_i u_j k(x_i, x_j) = \sum_{i,j} u_i u_j \mathbb{E}\left[e^{i|x_i - x_j|Z}\right] = \mathbb{E}\left[\left(\sum_i u_i e^{ix_i Z}\right)^2\right] \geq 0.$$

Gaussian Process — Basis Functions

The Gaussian Process can also be constructed by basis functions:

$$\mu(x) = \sum_{h=1}^{H} \beta_h b_h(x), \quad \beta = (\beta_1, \ldots, \beta_H) \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta)$$

Gaussian Process — Basis Functions

The Gaussian Process can also be constructed by basis functions:

$$\mu(x) = \sum_{h=1}^{H} \beta_h b_h(x), \quad \beta = (\beta_1, \ldots, \beta_H) \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta)$$

Then μ is a Gaussian process with

$$m(x) = \boldsymbol{b}(x)^T \boldsymbol{\beta}_0, \quad k(x, x') = \boldsymbol{b}(x)^T \boldsymbol{\Sigma}_\beta \boldsymbol{b}(x').$$

Gaussian Process — Inference

Suppose, we have observed the data $(x_1, y_1), \ldots, (x_n, y_n)$, and we assume that

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Given a new point $\tilde{x}$, we want to estimate the expected value of y at $\tilde{x}$, i.e. $\mu(\tilde{x})$.

Gaussian Process — Inference

Suppose, we have observed the data $(x_1, y_1), \ldots, (x_n, y_n)$, and we assume that

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Given a new point $\tilde{x}$, we want to estimate the expected value of y at $\tilde{x}$, i.e. $\mu(\tilde{x})$.

- We assume that μ is a Gaussian process with mean function $m(x) = 0$ and covariance function $k(x, x') = \tau^2 \exp\left(-|x - x'|^2/(2l^2)\right)$.

Gaussian Process — Inference

Suppose, we have observed the data $(x_1, y_1), \ldots, (x_n, y_n)$, and we assume that

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Given a new point $\tilde{x}$, we want to estimate the expected value of y at $\tilde{x}$, i.e. $\mu(\tilde{x})$.

- We assume that μ is a Gaussian process with mean function $m(x) = 0$ and covariance function $k(x, x') = \tau^2 \exp\left(-|x - x'|^2/(2l^2)\right)$.
- The joint distribution of $y = (y_1, \ldots, y_n)$ and $\mu(\tilde{x})$ is given by

$$\begin{pmatrix} y \\ \mu(\tilde{x}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x) + \sigma^2 I & K(x,\tilde{x}) \\ K(\tilde{x},x) & K(\tilde{x},\tilde{x}) \end{pmatrix}\right).$$

Gaussian Process — Inference

Suppose, we have observed the data $(x_1, y_1), \ldots, (x_n, y_n)$, and we assume that

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Given a new point $\tilde{x}$, we want to estimate the expected value of y at $\tilde{x}$, i.e. $\mu(\tilde{x})$.

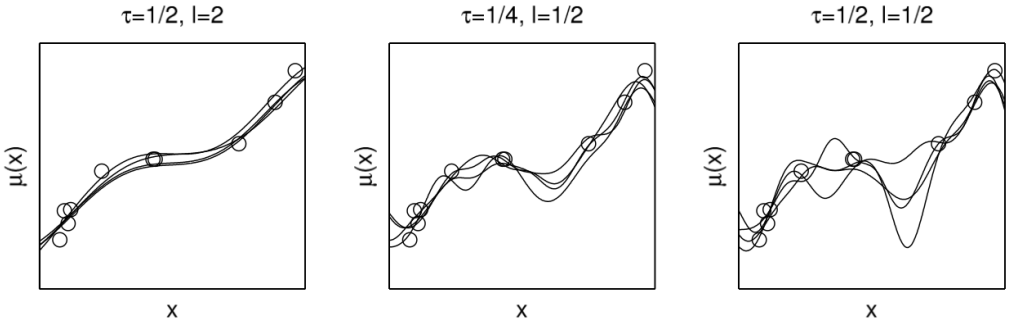
- We assume that μ is a Gaussian process with mean function $m(x) = 0$ and covariance function $k(x, x') = \tau^2 \exp\left(-|x - x'|^2/(2l^2)\right)$.
- The joint distribution of $y = (y_1, \ldots, y_n)$ and $\mu(\tilde{x})$ is given by

$$\begin{pmatrix} y \\ \mu(\tilde{x}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x) + \sigma^2 I & K(x,\tilde{x}) \\ K(\tilde{x},x) & K(\tilde{x},\tilde{x}) \end{pmatrix}\right).$$

- With the properties of conditional distribution of multivariate normal, we can derive the posterior distribution of $\mu(\tilde{x})$ given y.

$$\begin{aligned} &\mu(\tilde{x}) \mid x, y, \tau^2, \sigma^2, l^2 \\ &\sim \mathcal{N}\left(K(\tilde{x},x)(K(x,x) + \sigma^2 I)^{-1} y, K(\tilde{x},\tilde{x}) - K(\tilde{x},x)(K(x,x) + \sigma^2 I)^{-1} K(x,\tilde{x})\right) \end{aligned}$$

Gaussian Process — Example



Gaussian Process — Inference

For a Bayesian procedure, we need to specify the prior distributions for the hyperparameters τ^2, σ^2, and l^2.

Gaussian Process — Inference

For a Bayesian procedure, we need to specify the prior distributions for the hyperparameters τ^2, σ^2, and l^2.
A common choice is

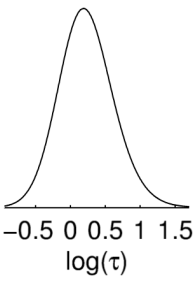
$$p(\log \tau) \propto 1, \quad p(\log \sigma) \propto 1, \quad p(\log l) \propto 1.$$

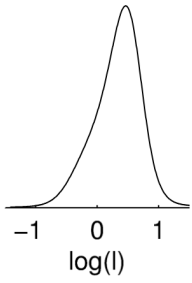
The log-likelihood is

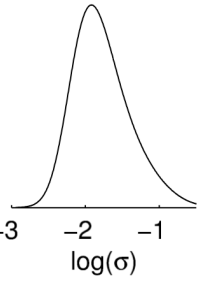
$$\log p(y \mid x, \tau^2, \sigma^2, l^2) = -\frac{1}{2} y^T (K(x,x) + \sigma^2 I)^{-1} y - \frac{1}{2} \log |K(x,x) + \sigma^2 I| - \frac{n}{2} \log(2\pi).$$

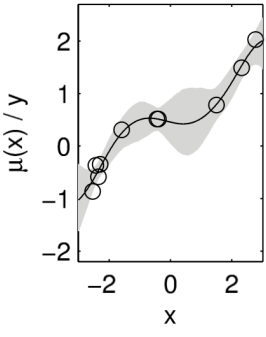
The posterior is now straightforward.

Gaussian Process — Inference









Example — Birth Dates

In this example, we analyze the patterns of birthdays in the United States. The data is the number of births on each day of the year from 1969 to 1988.

- This is a time series data, where the index is the number of days from 1969-01-01.
- The series contains periodic patterns, e.g. yearly and weekly patterns.
- The series also contains long term trends.

Example — Birth Dates

We model the time series as an additive model:

$$y(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t) + \epsilon_t,$$

- Long-term trend:

$$f_1(t) \sim \mathcal{GP}(0, k_1), \quad k_1(t, t') = \sigma_1^2 \exp\left(-\frac{|t - t'|^2}{2l_1^2}\right)$$

- Short-term trend:

$$f_2(t) \sim \mathcal{GP}(0, k_2), \quad k_2(t, t') = \sigma_2^2 \exp\left(-\frac{|t - t'|^2}{2l_2^2}\right)$$

Example — Birth Dates

We model the time series as an additive model:

$$y(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t) + \epsilon_t,$$

- Weekly pattern:

$$f_3(t) \sim \mathcal{GP}(0, k_3), \quad k_3(t, t') = \sigma_3^2 \exp\left(-\frac{2\sin^2(\pi(t-t')/7)}{l_{3,1}^2}\right) \exp\left(-\frac{|t-t'|^2}{2l_{3,2}^2}\right)$$

- Yearly pattern:

$$f_4(t) \sim \mathcal{GP}(0, k_4), \quad k_4(t, t') = \sigma_4^2 \exp\left(-\frac{2\sin^2(\pi(t-t')/365.25)}{l_{4,1}^2}\right) \exp\left(-\frac{|t-t'|^2}{2l_{4,2}^2}\right)$$

Example — Birth Dates

We model the time series as an additive model:

$$y(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t) + \epsilon_t,$$

- Special days and its interaction with weekends:

$$f_5(t) = I_{s.d.}(t)\beta_a + I_{s.d.}(t)I_{w.e.}(t)\beta_b$$

  where $I_{s.d.}(t)$ is an indicator function for special days (13 holidays), and $I_{w.e.}(t)$ is an indicator function for weekends.
- $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the residual.

Example — Birth Dates

- Sum of Gaussian processes is still a Gaussian process.
- The model can be fit through a standard GP inference.

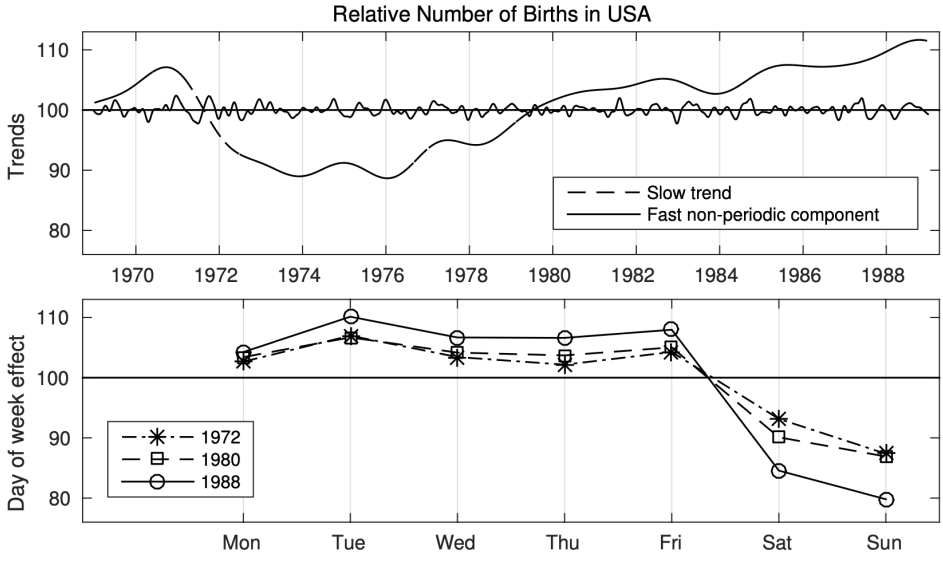
Example — Birth Dates

- Sum of Gaussian processes is still a Gaussian process.
- The model can be fit through a standard GP inference.
- log-t prior for time scales l.
- log-unifom prior for other parameters.

Example — Birth Dates

- Sum of Gaussian processes is still a Gaussian process.
- The model can be fit through a standard GP inference.
- log-t prior for time scales l.
- log-unifom prior for other parameters.
- The model can be further extended by considering weekdays v.s. weekends. See textbook Ch. 21.2.

Example — Birth Dates



Example — Birth Dates