21: Gaussian Process Models

Taylor

University of Virginia

Introduction

We talk about Gaussian process models in this chapter. Gaussian processes describe random functions, and they can show up in statistical modeling in a few places.

If you would like to dig a little deeper, this is considered a good reference: `http://gaussianprocess.org/gpml/`. The textbook's bibliography for this chapter includes many other sources.

Definitions

It's helpful to initially consider $x_i \in \mathbb{R}^p$ where $p = 1$ or $p = 2$.

We say μ follows a **Gaussian process** with mean function m and covariance function k if for any finite set of nonrandom points $x_1, \ldots, x_n$

$$\mu(x_1), \ldots, \mu(x_n) \sim \text{Normal}((m(x_1), \ldots, m(x_n)), K(x_1, \ldots, x_n)).$$

For short, we write $\mu \sim \text{GP}(m, k)$.

Definitions

It's helpful to initially consider $x_i \in \mathbb{R}^p$ where $p = 1$ or $p = 2$.

We say μ follows a **Gaussian process** with mean function m and covariance function k if for any finite set of nonrandom points $x_1, \ldots, x_n$

$$\mu(x_1), \ldots, \mu(x_n) \sim \text{Normal}((m(x_1), \ldots, m(x_n)), K(x_1, \ldots, x_n)).$$

For short, we write $\mu \sim \text{GP}(m, k)$.

This means $E[\mu(x_i)] = m(x_i)$ and $\text{Cov}(\mu(x_i), \mu(x_j)) = K_{i,j} = k(x_i, x_j)$.

A first example: Gaussian process regression

Let's assume we're regressing univariate y_is on vector-valued x_is. Then we are interested in

$$y_i = \mu(x_i) + \epsilon_i.$$

We could also be interested in the "noiseless" situation, as well.

The μ function can be nonlinear and very flexible!

A first example: Gaussian process regression

$$y_i = \mu(x_i) + \epsilon_i.$$

Picking a prior means we need to pick m and k. We can see that

$$E[y_i \mid x_i] = E[\mu(x_i) \mid x_i] = m(x_i).$$

For m

- can assume $m(x) = 0$ (like assuming regression coefficients have a zero-mean prior)
- can use an informative prior

For k...

A popular choice

Any k function gives you a "similarity" or "nearness" measure for any two pairs of inputs. It needs to be chosen very carefully.

We will often use a **squared exponential kernel**

$$k(x, x') = \tau^2 \exp\left[-\sum_{i=1}^{p} \frac{(x_j - x_j')^2}{2l_j^2}\right]$$

Each l_j determines the wiggliness in the jth direction of the predictors.

The τ^2 parameter is an overall variance for each $\mu(x)$.

Simulating from the prior

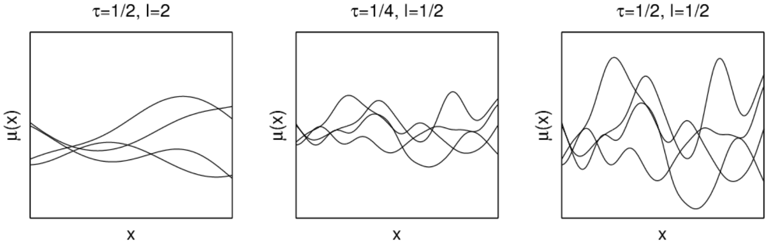


Figure 21.1 *Random draws from the Gaussian process prior with squared exponential covariance function and different values of the amplitude parameter τ and the length scale parameter l.*

More to say about kernel choice:
`https://www.cs.toronto.edu/~duvenaud/cookbook/`

Inference: conditional posterior

We will use a lot of properties of Gaussian random vectors when we conduct inference.

If

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \text{Normal}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

then $x_1 \mid x_2$ is also normally distributed with mean vector

$$\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

and covariance matrix

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Inference: conditional posterior

Let's assume the likelihood is $y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, and for the prior, $m(x) = 0$.

The observed data is $\{x_i, y_i\}$, and the parameters are τ, l, σ^2. To find the conditional posterior $p(\mu(x) \mid x, y, \sigma^2, \tau, l)$, we use

$$\left(\begin{array}{c} y \\ \mu \end{array} \right) \Bigg| x, \sigma^2, \tau, l \sim \text{Normal}\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} K(x,x) + \sigma^2 I & K(x,x) \\ K(x,x) & K(x,x) \end{array} \right) \right)$$

Inference: conditional posterior

Let's assume the likelihood is $y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, and for the prior, $m(x) = 0$.

The observed data is $\{x_i, y_i\}$, and the parameters are τ, l, σ^2. To find the conditional posterior $p(\mu(x) \mid x, y, \sigma^2, \tau, l)$, we use

$$\left(\begin{array}{c} y \\ \mu \end{array} \right) \Bigg| x, \sigma^2, \tau, l \sim \text{Normal} \left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} K(x,x) + \sigma^2 I & K(x,x) \\ K(x,x) & K(x,x) \end{array} \right) \right)$$

By properties of multivariate normal random vectors $\mu \mid x, y, \tau, l, \sigma$ is normally distributed with mean and covariance

$$\begin{aligned} E[\mu \mid x, y, \tau, l, \sigma] &= K(x,x)[K(x,x) + \sigma^2 I]^{-1} y \\ Var[\mu \mid x, y, \tau, l, \sigma] &= K(x,x) - K(x,x)[K(x,x) + \sigma^2 I]^{-1} K(x,x) \end{aligned}$$

Inference: conditional posterior

Let's assume the likelihood is $y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, and for the prior, $m(x) = 0$.

The observed data is $\{x_i, y_i\}$, and the parameters are τ, l, σ^2. To find the conditional posterior $p(\mu(x) \mid x, y, \sigma^2, \tau, l)$, we use

$$\begin{pmatrix} y \\ \mu \end{pmatrix} \Bigg| x, \sigma^2, \tau, l \sim \text{Normal}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x) + \sigma^2 I & K(x,x) \\ K(x,x) & K(x,x) \end{pmatrix} \right)$$

By properties of multivariate normal random vectors $\mu \mid x, y, \tau, l, \sigma$ is normally distributed with mean and covariance

$$E[\mu \mid x, y, \tau, l, \sigma] = K(x,x)[K(x,x) + \sigma^2 I]^{-1} y$$
$$Var[\mu \mid x, y, \tau, l, \sigma] = K(x,x) - K(x,x)[K(x,x) + \sigma^2 I]^{-1} K(x,x)$$

What does this simplify to in the case of "noiseless" regression?

Inference: conditional posterior

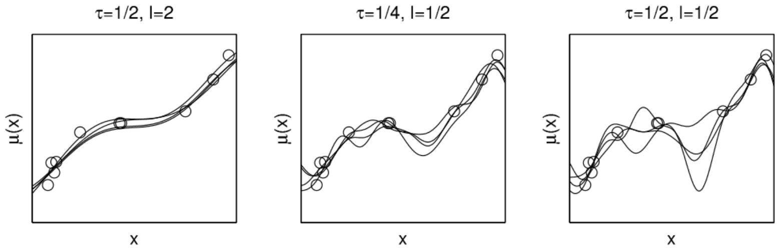


Figure 21.2 *Posterior draws of a Gaussian process $\mu(x)$ fit to ten data points, conditional on three different choices of the parameters τ, l that characterize the process. Compare to Figure 21.1, which shows draws of the curve from the prior distribution of each model. In our usual analysis, we would assign a prior distribution to τ, l and then perform joint posterior inference for these parameters along with the curve $\mu(x)$; see Figure 21.3. We show these three choices of conditional posterior distribution here to give a sense of the role of τ, l in posterior inference.*

Inference: prediction/smoothing at new points

Let's assume the likelihood is $y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, and for the prior, $m(x) = 0$.

Call $\tilde{x}$ unseen data, in addition to $\{x_i, y_i\}$. Then

$$\left(\begin{array}{c} y \\ \tilde{\mu} \end{array} \right) \Big| x, \tilde{x}, \sigma^2, \tau, l \sim \text{Normal}\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} K(x,x) + \sigma^2 I & K(x,\tilde{x}) \\ K(\tilde{x},x) & K(\tilde{x},\tilde{x}) \end{array} \right) \right)$$

By properties of multivariate normal random vectors, $\tilde{\mu} \mid x, y, \tau, l, \sigma$ is normally distributed with

$$E[\tilde{\mu} \mid x, y, \tau, l, \sigma] = K(\tilde{x}, x)[K(x,x) + \sigma^2 I]^{-1} y$$
$$Var[\tilde{\mu} \mid x, y, \tau, l, \sigma] = K(\tilde{x}, \tilde{x}) - K(\tilde{x}, x)[K(x,x) + \sigma^2 I]^{-1} K(x, \tilde{x})$$

Inference: estimating unknown parameters

We need the marginal likelihood for MCMC techniques: $\log p(y \mid \tau, l, \sigma^2)$ equals

$$-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det(K(x,x) + \sigma^2 I) - \frac{1}{2}y^T[K(x,x) + \sigma^2 I]^{-1}y$$

We could use this to do Gibbs sampling or some sort of Metropolis-Hastings technique.

Birthdays and Birthdates example

$$y(t) = \mu(t) + \epsilon_t$$

with

$$\mu(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t).$$

Birthdays and Birthdates example

Slow and fast trends:

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

and

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

There is an identifiability concern. They mention that they put log-t priors on l_1 and l_2 and log-uniform priors on σ_1 and σ_2^2, but they do not give specifics.

Birthdays and Birthdates example

A quasi-periodic weekly effect:

$$f_3(t) \sim GP(0, k_3)$$

$$k_3(t, t') = \sigma_3^2 \exp\left[-\frac{2\sin^2(\pi[t - t']/7)}{l_{3,1}^2}\right] \exp\left[-\frac{(t - t')^2}{2l_{3,2}^2}\right].$$

The kernel k_3 is "high" only when both baby kernels are "high."

Also note $2\sin^2(\pi[t - t']/7) = 1 - \cos\left(\frac{2\pi[t-t']}{7}\right)$ by "product identity."

Birthdays and Birthdates example

A quasi-periodic yearly effect:

$$f_4(t) \sim GP(0, k_4)$$

$$k_4(t, t') = \sigma_4^2 \exp\left[-\frac{2\sin^2(\pi[t-t']/365.25)}{l_{4,1}^2}\right] \exp\left[-\frac{(t-t')^2}{2l_{4,2}^2}\right].$$

Birthdays and Birthdates example

Regular regression parameters

$$f_5(t) = I_{\text{special day}}\beta_a + I_{\text{special day and weekend}}\beta_b$$

where $I_{\text{special day}} = (I_{\text{New Year's Day}}, I_{\text{Valentine's Day}}, \ldots, I_{\text{Christmas}})'$.

$$k_5(t, t') = ?$$

Birthdays and Birthdates example

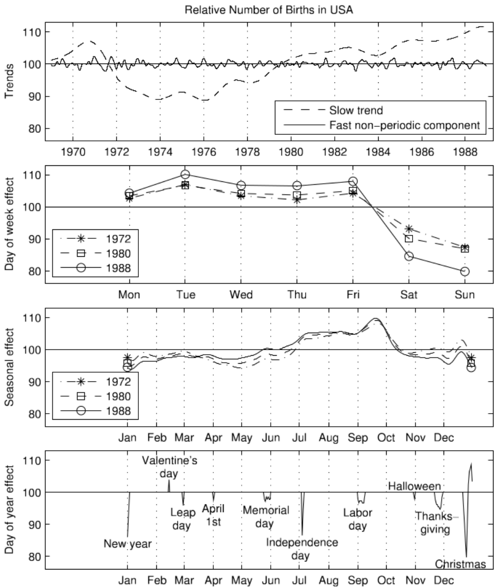
This means

$$\mu(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t) \sim GP(0, k)$$

where

$$k(t, t') = k_1(t, t') + k_2(t, t') + k_3(t, t') + k_4(t, t') + k_5(t, t').$$

Birthdays and Birthdates example



The Catch

All of this seems easy. We're just using normal-normal conjugacy, right?

The Catch

All of this seems easy. We're just using normal-normal conjugacy, right?

Yes, but there are computational difficulties. For large data sets with more than several thousand rows, naively inverting $K(x,x) + \sigma^2 I$ is going to be brutal.

Density estimation example

So far we have discussed Gaussian processes as prior distributions for a function controlling the location and potentially the shape parameter of a parametric observation model.

Density estimation example

So far we have discussed Gaussian processes as prior distributions for a function controlling the location and potentially the shape parameter of a parametric observation model.

To get more flexibility we would like to model also the conditional observation model as nonparametric.

One way to do this is with the **logistic Gaussian proces** (LGP).

Density estimation example

We want a density p for random variables $y_1, \ldots, y_n \overset{\text{iid}}{\sim} p(y \mid f)$.

So far this semester we have assumed p is parametric (conditioning on parameters θ). However, here, we want this density to be based on a nonparametric function f, which is a realization of a Gaussian process.

Density estimation example

We want a density p for random variables $y_1, \ldots, y_n \overset{\text{iid}}{\sim} p(y \mid f)$.

So far this semester we have assumed p is parametric (conditioning on parameters θ). However, here, we want this density to be based on a nonparametric function f, which is a realization of a Gaussian process.

Recall that p needs to be nonnegative, and it needs to integrate to 1:

$$p(y \mid f) = \frac{e^{f(y)}}{\int e^{f(y)} \mathrm{d}y}.$$

Density estimation example

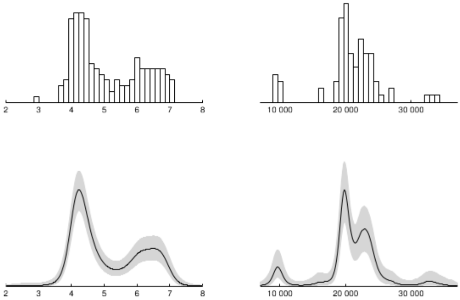


Figure 21.7 *Two simple examples of density estimation using Gaussian processes. Left column shows acidity data and right column shows galaxy data. Top row shows histograms and bottom row shows logistic Gaussian process density estimate means and 90% pointwise posterior intervals.*