SDS 383D: Modeling II **Spring 2017**

Section 4: Gaussian processes

4.1 Non-linear functions

4.1.1 Regression view

So far, we've assumed our latent function is a linear function of our data – which is obviously limiting. One way of circumventing this is to project our inputs into some high-dimensional space using a set of basis functions $\phi : \mathbb{R}^d \to \mathbb{R}^N$, and then performing linear regression in that space, so that

$$y_i = \phi(x)^T \beta + \epsilon_i$$

For example, we could project x into the space of powers of x, i.e. $\phi(x) = (1, x, x^2, x^3 \dots)$ to obtain polynomial regression.

Exercise 4.1 *Let* $\mathbf{y}$ *and* $\mathbf{X}$ *be set of observations and corresponding covariates, and* y_* *be the unknown value we wish to predict at covariate* $\mathbf{x}_*$*. Assume that*

$$\begin{aligned} \beta \sim & N(0, \Sigma) \\ \begin{bmatrix} f_* \\ \mathbf{f} \end{bmatrix} = & \begin{bmatrix} \boldsymbol{\phi}_*^T \\ \boldsymbol{\Phi}^T \end{bmatrix}^T \beta \\ \begin{bmatrix} y_* \\ \mathbf{y} \end{bmatrix} \sim & N\left(\begin{bmatrix} f_* \\ \mathbf{f} \end{bmatrix}, \sigma^2 \mathbf{I} \right) \end{aligned}$$

where $\boldsymbol{\phi} := \phi(\mathbf{x})$ *and* $\boldsymbol{\Phi} := \phi(\mathbf{X})$*.*

What is the predictive distribution $p(f_*|\mathbf{y}, \mathbf{x}_*, \mathbf{X})$*? Note: this is very similar to questions we did in Section 1.*

Proof: From Section 1, we know that

$$p(\beta|y, X) \sim \mathcal{N}(\frac{1}{\sigma^2} A^{-1} \Phi y, A^{-1}),$$

where $A = \sigma^{-2} \Phi \Phi^T + \sigma^{-1}$.

Since $f_*|\beta = \phi_* \beta$, $f_*|x_*, X, y \sim \mathcal{N}(\frac{1}{\sigma^2} \phi_*' A^{-1} \Phi y, \phi_*' A^{-1} \phi_*)$.

We can rewrite the equation is the following way:

$$f_*|x_*, X, y \sim \mathcal{N}(\phi_*' \Sigma \Phi (K + \sigma^2 I)^{-1} y, \phi_*' \Sigma \phi_* - \phi_*' \Sigma \Phi (K + \sigma^2 I)^{-1} \Phi^T \Sigma \phi_*), \tag{4.1}$$

where $K = \Phi' \Sigma \Phi$. ■

Note that, in the solution to Exercise 1, we only ever see ϕ or Φ in a form such as $\Phi^T \Sigma \Phi$. We will define $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}^T \Sigma \phi(\mathbf{x}')$. Since Σ is positive definite, we can write:

$$k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})^T \psi(\mathbf{x})$$

where $\psi(\mathbf{x}) = \phi(\mathbf{x})\Sigma^{1/2}$

If (as here) we only ever access ψ via this inner product, we can choose to work instead with $k(\cdot, \cdot)$. This may be very convenient if the dimensionality of $\psi(x)$ is very high (or even infinite... see later). $k(\cdot, \cdot)$ is often refered to as the kernel, and this replacement is referred to as the kernel trick.

Exercise 4.2 *Let's look at a concrete example, using the old faithful dataset on R*

- *`data("faithful",package="datasets")` in R*
- *or available as `faithful.csv` on github if you're not using R.*

Let $\phi(x) = (1, x, x^2, x^3)$. Using appropriate priors on β and σ^2, obtain a posterior distribution over $f := \phi(x)^T\beta$. Plot the function (with a 95% credible interval) by evaluating this on a grid of values.

Proof: For simplicity, we set $\Sigma = I$ and $\sigma^2 = 1$. In this case, $k(\mathbf{x}, \mathbf{y}') = 1 + xy + x^2y^2 + x^3y^3$. The posterior is defined in Equation 4.2. The plot is shown in Figure 4.1.

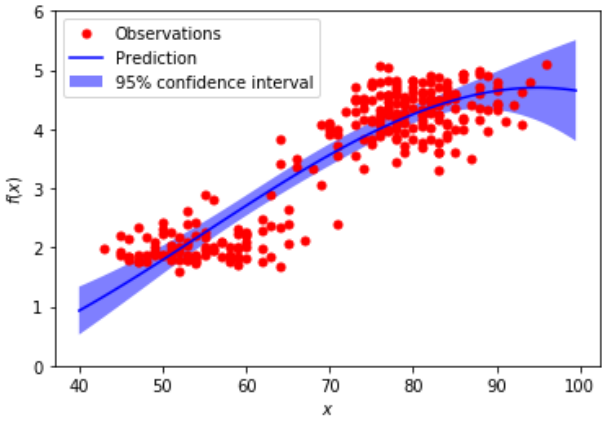


Figure 4.1: Predictions with polynomial kernel

■

4.1.2 Function space view

Look back at the plot from Exercise 2. We specified a prior distribution over regression parameters, which we can use to obtain a posterior distribution over those regression parameters. But, what we calculated (and plotted) was a posterior distribution over *functions*. Similarly, we can think of our prior on β as specifying a prior distribution on the space of cubic functions. Evaluated at a finite number of input locations – as you did in Exercise 2 – this posterior distribution is multivariate Gaussian. This is in fact the definition of a Gaussian process: A distribution over functions, such that the marginal distribution evaluated at any finite set of points is multivariate Gaussian.

A priori, the covariance of f is given by

$$\text{cov}(x, x') = E[(f(x) - m(x))(f(x^T) - m(x^T))] = k(x, x')$$

. For this reason, our kernel k is often referred to as the covariance function (note, it is a function since we can evaluate it for any pairs x, x'). In the above example, where β had zero mean, the mean of f is zero; more generally, we will assume some mean function $m(x)$.

Rather than putting a prior distribution over β, we can specify a covariance function – remember that our covariance function can be written in terms of the prior covariance of β. For example, we might let

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

– this is known as a squared exponential covariance function, for obvious reasons. This prior encodes the following assumptions:

- The covariance between two datapoints decreases monotonically as the distance between them increases.
- The covariance function is stationary – it only depends on the distance between x and x', not their locations.
- Even more than being stationary, it is isotropic: It depends only on $|x - x'|$.

Exercise 4.3 *Let's explore the resulting distribution over functions. Write some code to sample from a Gaussian process prior with squared exponential covariance function, evaluated on a grid of 200 inputs between 0 and 100. For $\ell = 1$, sample 5 functions and plot them on the same plot. Repeat for $\ell = 0.1$ and $\ell = 10$. Why do we call ℓ the lengthscale of the kernel?*

Proof: The results are shown in Figure 4.2. With smaller lengthscales, the functions become more zigzag. With larger lengthscales, the functions become more flat.

■

Exercise 4.4 *Let $\mathbf{f}_* := f(\mathbf{X}_*)$ be the function f evaluated at test covariate locations $\mathbf{X}_*$. Derive the posterior distribution $p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y})$, where $\mathbf{y}$ and $\mathbf{X}$ comprise our training set. (You can start from the answer to Exercise 1 if you'd like).*

Proof: Based on the definition of Gaussian process,

$$\begin{bmatrix} f_* \\ f \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} k(X_*, X_*), k(X_*, X) \\ k(X, X_*), k(X, X) \end{bmatrix}),$$

$$\text{therefore, } \begin{bmatrix} f_x \\ y \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} k(X_*, X_*), k(X_*, X) \\ k(X, X_*), k(X, X) + \sigma^2 I \end{bmatrix}).$$

Therefore,

$$f_*|X_*, X, y \sim \mathcal{N}(k(X_*, X)(k(X, X) + \sigma^2 I)^{-1} y, k(X_*, X_*) - k(X_*, X)(k(X, X) + \sigma^2 I)^{-1} k(X, X_*)). \quad (4.2)$$

■

Exercise 4.5 *Return to the faithful dataset. Evaluate the posterior predictive distribution $p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y})$, for some reasonable choices of parameters (perhaps explore a few length scales if you're not sure what to pick), and plot the posterior mean plus a 95% credible interval on a grid of 200 inputs between 0 and 100, overlaying the actual data.*

Proof: The results are shown in Figure 4.3. Three different lengthscales have been tried. Larger lenthscales give smoother(flatter) predictions. ■

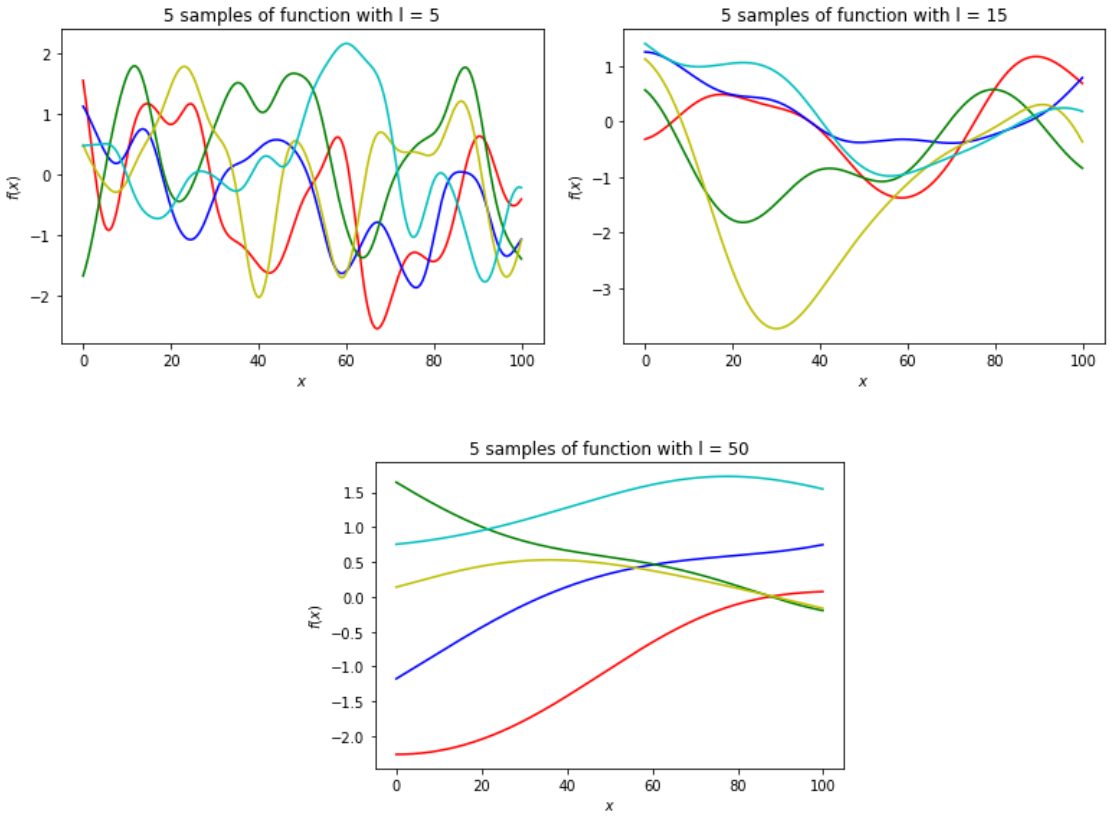


Figure 4.2: Comparison of Gaussian process samples with different lengthscales

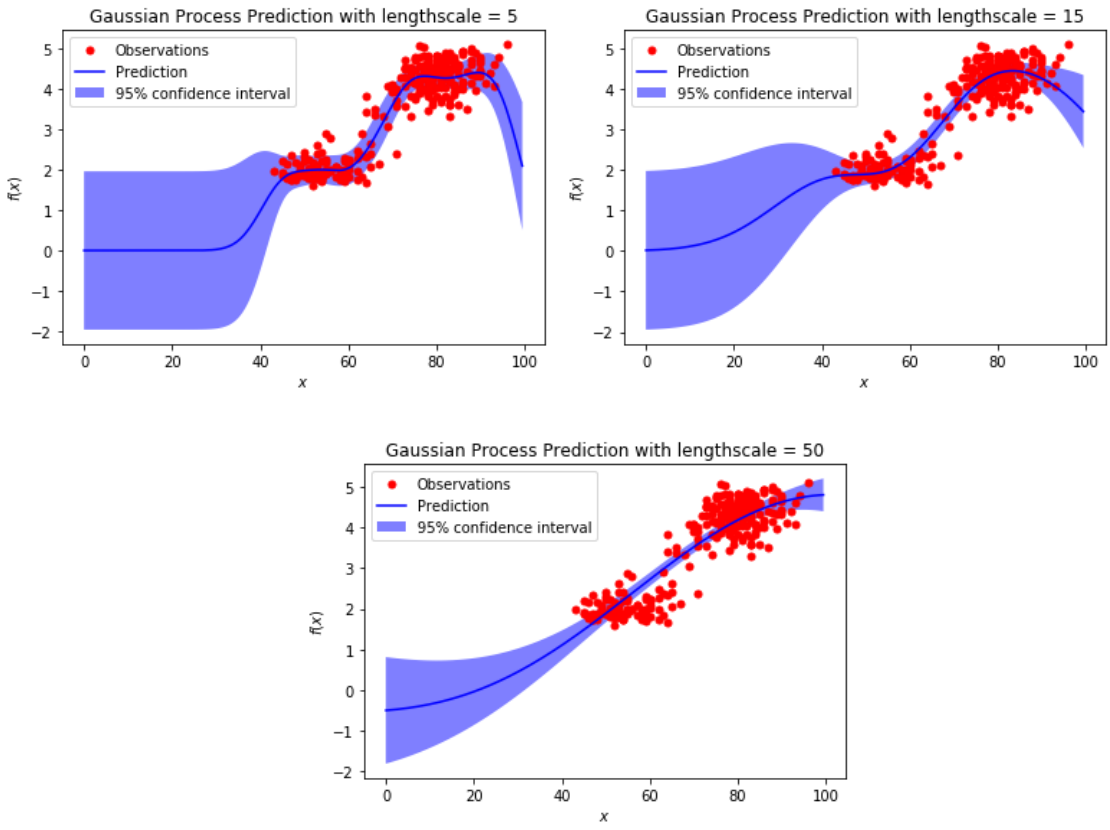


Figure 4.3: Comparison of Gaussian process predictions with different lengthscales

4.2 Parameter Estimation

As we saw in the previous section, the choice of hyperparameters (for the squared exponential case, the length scale ℓ) effects the properties of the resulting function. Rather than pick a specific value for the hyperparameter, we can specify the model in a hierarchical manner—just like we did in the linear case.

For example, in the squared exponential setting, we could specify our model as

$$\begin{aligned}
\ell^2 \sim & \text{Inv-Gamma}(a_\ell, b_\ell) \\
\alpha^2 \sim & \text{Inv-Gamma}(a_\alpha, b_\alpha) \\
\sigma^2 \sim & \text{Inv-Gamma}(a_\sigma, b_\sigma) \\
k(x, x') = & \alpha^2 \exp\left\{-\frac{1}{2\ell^2}|x - x'|^2\right\} + \sigma^2 \delta_{x-x'} \\
y|X \sim & N(0, \tilde{K})
\end{aligned}$$

where K is the covariance function evaluated at the input locations X. Note that we have integrated out f and placed our prior directly on y, incorporating the Gaussian likelihood into the covariance. We can then infer the posterior distribution over ℓ using Bayes' Law:

$$p(\ell|y, X) = \frac{p(y|X, \ell)p(\ell)}{\int_0^\infty p(y|X, \ell)p(\ell)d\ell}$$

Unfortunately, we typically do not have an analytical form for this posterior, so we must resort to either optimization, or MCMC-based inference.

4.2.1 Optimization

In practice, a common approach is to find the ML estimate for the hyperparameters. Let's assume a generic setting, where the log likelihood is parametrized by some vector of parameters θ. The log likelihood is given by

$$logp(y|X, \theta) = -\frac{1}{2}y^T K^{-1} y - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi$$

Taking partial derivatives, we see that

$$\begin{aligned}
\frac{\partial}{d\partial\theta_j}\log p(y|X, \theta) = & \frac{1}{2}y^T K^{-1}\frac{\partial K}{\partial\theta_j}K^{-1}y - \frac{1}{2}\text{tr}\left(K^{-1}\frac{\partial K}{\partial\theta_j}\right) \\
= & \frac{1}{2}\text{tr}\left((\alpha\alpha^T - K^{-1})\frac{\delta K}{\delta\theta_j}\right)
\end{aligned}$$

where $\alpha = K^{-1}y$. We can use these partial derivatives to find the ML estimate of θ, using a gradient-based optimization method

Exercise 4.6 *Calculate the appropriate derivatives for the one-dimensional, squared exponential case used for the* `faithful` *dataset. Use these gradient to find the optimizing value of ℓ^2, α^2 and σ^2. Plot the resulting fit.*

Proof: We only need to compute $\frac{\partial K}{\partial \theta_j}$:

$$\frac{\partial K_{ij}}{\partial l^2} = \alpha^2 \exp\{-\frac{(x_i - x_j)^2}{2l^2}\}\frac{(x_i - x_j)^2}{2l^4}$$

$$\frac{\partial K_{ij}}{\partial \alpha^2} = \exp\{-\frac{(x_i - x_j)^2}{2l^2}\}$$

$$\frac{\partial K_{ij}}{\partial \sigma^2} = 1_{\{x_i = x_j\}}.$$

With the gradient computed, we can maximize the likelihood with respect to parameters using blackbox optimization tools. Starting from $l^2 = 1$, $\alpha^2 = 1$, $\sigma^2 = 1$, I got $l^2 = 4.82$, $\alpha^2 = 7.27$, $\sigma^2 = 0.15$. Starting from another place, I got another value meaning that there are a lot of local minimums. I plotted the prediction for $l^2 = 4.82$, $\alpha^2 = 7.27$, $\sigma^2 = 0.15$ in Figure 4.4.

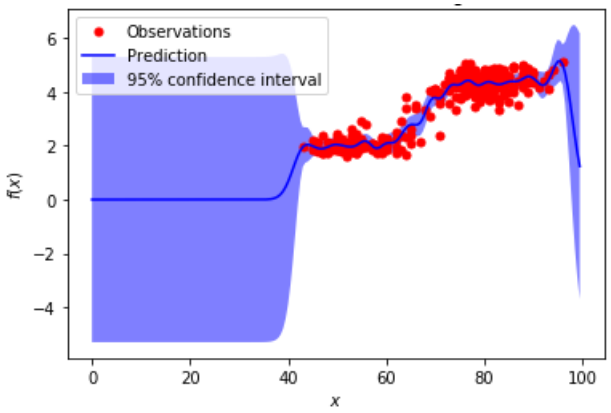


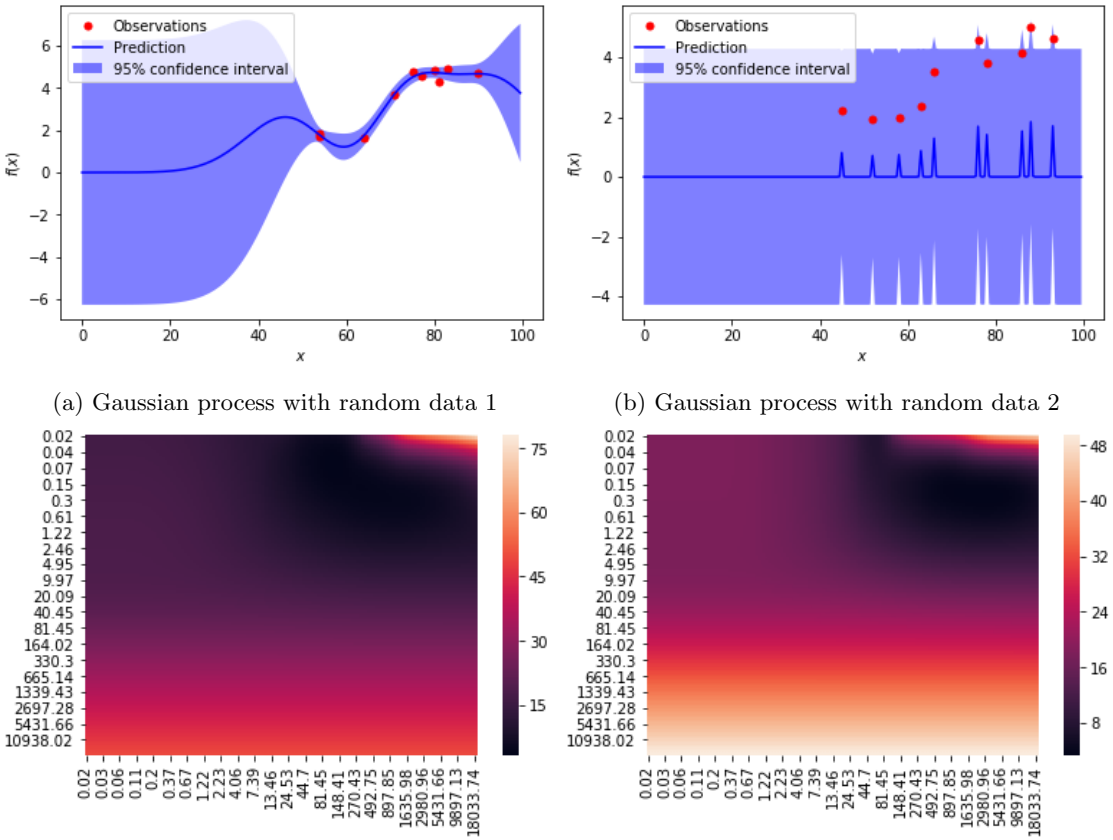
Figure 4.4: Comparison of Gaussian process predictions with estimated parameters

■

Exercise 4.7 *Repeat the previous exercise, but this time only use the first 10 data points from the faithful dataset. Repeat the optimization several times, using different initializations/random seeds. You will likely see widely different results – sometimes ℓ is big, sometimes σ^2 is big. Why is this? Discuss why this is a problem here, but wasn't in the previous setting. You may find it helpful to look at the corresponding scatter plot, or plot the log likelihood for certain values of σ^2 and ℓ.*

Proof: The results are shown in Figure 4.5. The two plots are Gaussian process predictions with parameters trained with two different data sets. The first plot (Figure 4.5a) corresponds to the parameters $[l^2, \alpha^2, \sigma^2] = [1.28 \cdot 10^2, 1.02 \cdot 10^1, 6.52 \cdot 10^{-2}]$. The second plot (Figure 4.5b) corresponds to the parameters $[l^2, \alpha^2, \sigma^2] = [0.029, 4.77, 8.19]$.

One of the reasons for the sensitivity of parameters is that there are a lot of local minimals or saddle points of the likelihood function. We can plot the likelihood function of l^2, σ^2 for two data sets respectively. In Figure 4.5c, the negative log-likelihood of $l^2(x-axis), \sigma^2(y-axis)$(with $\alpha^2 = 10$) for first data is plotted. As is shown in the plot, there is no obvious global minimum which leads to the problem of sensitivity. Similar result is shown in Figure 4.5d where the negative log-likelihood of $l^2(x - axis), \sigma^2(y - axis)$(with $\alpha^2 = 5$) for second data is plotted. ■



(a) Gaussian process with random data 1

(b) Gaussian process with random data 2

(c) Likelihood with random data 1

(d) Likelihood with random data 2

Figure 4.5: Comparison of Gaussian process predictions with different training data

4.2.2 MCMC

Optimization is typically pretty quick, which is why it is commonly used in practice. However, we have no guarantee that our optimization surface is convex. An alternative approach is to sample from the posterior distribution over our hyperparameters.

Exercise 4.8 *Since the posterior is non-conjugate, we can't use a Gibbs sampler. We won't go into the details of appropriate sampling methods since this isn't an MCMC course, but we will explore using black-box samplers. In the R folder, there are three files:* `faithful_data.R`, `gp_regression.stan` *and* `run_gp_regression.R`. *Use these to sample from the model and produce 95% credible intervals for α, ℓ and σ, and 95% predictive intervals for t. Go through the code and make sure you understand what is going on.*

Proof: The plots are shown in Figure 4.6. With the default prior setting (normal prior for parameters) we can get some reasonable fitting as in 4.6a. And the corresponding histograms pf posterior samples is shown in 4.6c. However, if we set the prior as the non-informative prior, then the fitting does not make sense as in 4.6b. ■

Exercise 4.9 *Let's now look at a dataset with multiple predictors. Download the dataset weather.csv – this contains latitude and longitude data for 147 weather stations, plus a response "temperature", which is the difference between the forecasted and actual temperature for each station.*

How should we extend our kernel to multiple dimensions? (There is more than one option here). Should we use the same lengthscale for latitude and longitude? Construct an appropriate parametrized kernel, and learn the parameters either via optimization or using MCMC by editing the Stan code (Note: If you go for the stan code, you will need to implement your new kernel).

Using an appropriate visualization tool, plot the mean function (try imshow or contourf in matlab or matplotlib (for python), or image or filled.contour for R).

Proof: A reasonable choice of kernel function for multivariate case is

$$k(x, x') = \alpha^2 \exp\{-\sum_{i=1}^{d} \frac{1}{2l_i^2}|x_i - x_i'|^2\} + \sigma^2 \delta_{x=x'}.$$

Similarly, we can run MCMC to get the posterior samples for the parameters. The result is shown in Figure 4.7. The prediction is shown in Figure 4.8.

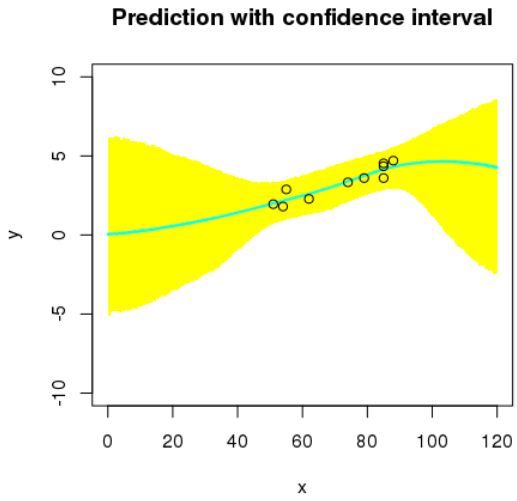
■

4.3 Beyond regression: non-conjugate likelihoods

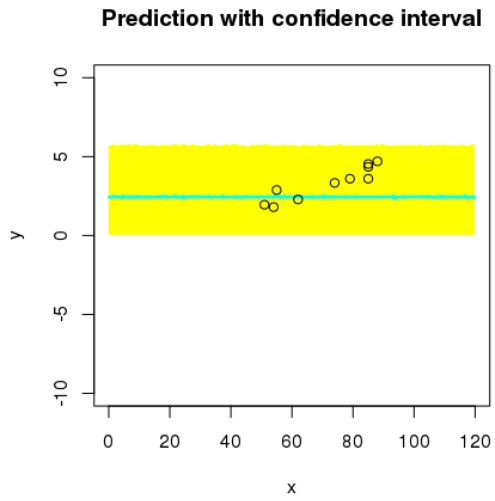
So far, we've focused on Gaussian processes in a regression context. We can however use them as the basis of a non-Gaussian regression... using exactly the same techniques as we used for the regression setting! For example, in Section 3, we dealt with non-Gaussian data by transforming our regression output:

$$y_i|\beta, x_i \sim f(g^{-1}(x_i^T\beta))$$

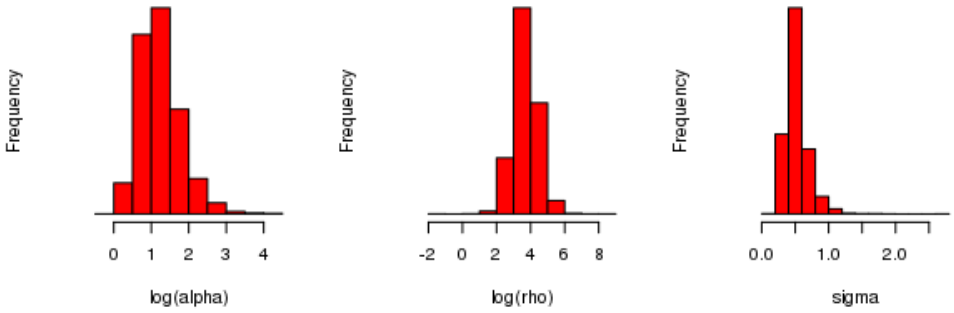
where f is an appropriate likelihood model (e.g. Bernoulli for binary data, Poisson for count data) and g^{-1} was a function that maps the real-valued $x_i^T\beta$ to an appropriate space for that likelihood.



(a) Gaussian process with normal prior



(b) Gaussian process with non-informative prior



(c) Histograms of posterior samples for parameters with normal prior

Figure 4.6: Parameter selection with MCMC

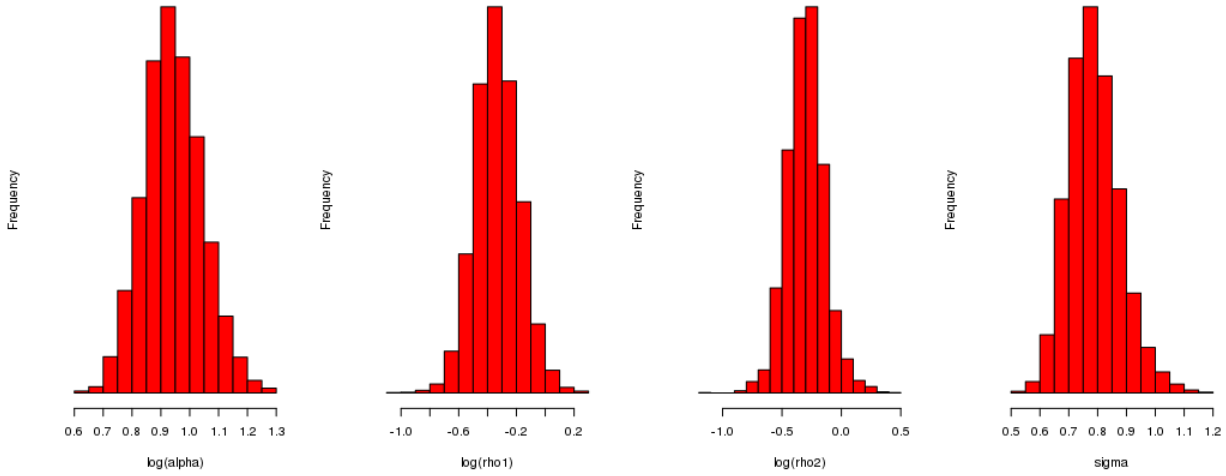


Figure 4.7: Parameter selection with MCMC

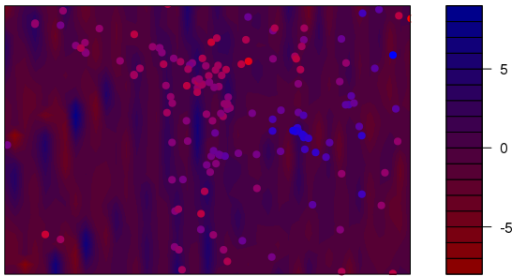


Figure 4.8: Heatmap of prediction

We can do exactly the same here, by letting

$$f \sim \text{GP}(0, k)$$
$$y_i \sim f(g^{-1}(f(x_i)))$$

Let's start by considering a binary example. In the GLM setting, we looked at both probit and logit regression. We can use the same approaches here!

Exercise 4.10 *Describe (including pseudo-code) how we could implement probit Gaussian process regression, using an auxiliary variable method analogous to that used in Exercise 3.1.*

Proof: Augmented probit Gaussian process regression:

$$y_i = 1_{\{z_i > 0\}},$$
$$z \sim N(f, \sigma^2 I),$$
$$f \sim GP(0, K(\dot{;})).$$

The Gibbs sampler would be:

$$z_i|- \propto 1_{\{y_i=0\}} 1_{\{z_i \leq 0\}} N(f_i, \sigma^2) + 1_{\{y_i=1\}} 1_{\{z_i>0\}} N(f_i, \sigma^2),$$
$$f|- \sim N((K^{-1} + \frac{I}{\sigma^2})^{-1} z, (K^{-1} + \frac{I}{\sigma^2})^{-1}).$$

So Gibbs sampler would iterate between sampling from univariate truncated normal and a multivariate normal. ■

For the logit case, we can again use a Laplace approximation to approximate our posterior. In the GLM setting, we used the Laplace approximation to approximate the posterior over β. Here, we will work directly with our function f evaluated at our training locations, and approximate $p(f|X, y, \theta)$, where θ are the parameters of our covariance function.

Let $P^*(f) \propto p(f|X, y, \theta)$ be our unnormalized posterior, so that $\log P^*(f) = \log p(y|f) + logp(f|X) = \log p(y|f) - \frac{1}{2} f^T K^{-1} f - \frac{1}{2} \log |K| + \text{const}$.

Exercise 4.11 *Derive the Hessian of* $\log P^*(f)$*, when* $y_i \sim Bernoulli\left(\frac{1}{1+e^{-f_i}}\right)$

Proof:

$$\log P^*(f) = \sum_{i=1}^{n} -\log(1 + e^{-f_i}) y_i - (1 - y_i) \log(1 + e^{f_i}) - \frac{1}{2} f^T K^{-1} f + const,$$

$$\frac{\partial \log^* f}{\partial f_k} = y_k - \frac{e^{f_k}}{1 + e^{f_k}} - [K^{-1} f]_k,$$

$$H = \begin{bmatrix} -\frac{e^{f_1}}{(1+e^{f_1})^2} & & \\ & \cdots & \\ & & -\frac{e^{f_n}}{(1+e^{f_n})^2} \end{bmatrix} - K^{-1}.$$

■

Exercise 4.12 *The dataset iris.csv contains details of 150 flowers from three species. Pick two of them (your choice) as your regression dataset. Find the MAP of f, for some reasonable choice of hyperparameters and squared exponential kernel. Visualize the corresponding class probabilities on a series of 2d plots.*

Proof: The results are shown in Figure 4.9. ■

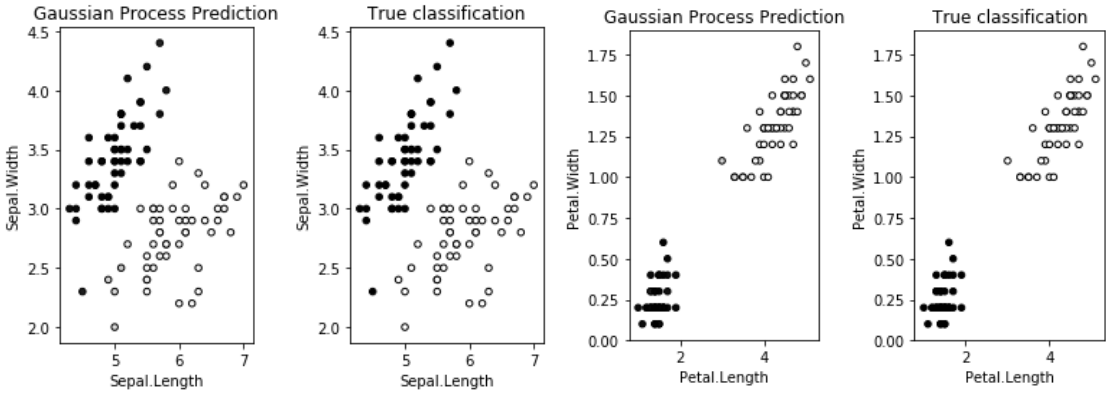


Figure 4.9: Comparison of Gaussian process prediction and true classification

Exercise 4.13 *Evaluate the Hessian using the same dataset, and visualize uncertainty in your plots from the previous exercise (e.g. by creating a contour plot of the marginal standard deviations).*

Proof: Using Laplacian approximation, we get an approximation for the posterior $f|X, y$, which is $N(f|\hat{f}, (K+W^{-1})^{-1})$. With this approximation, we can estimate the variance of our prediction f_* for x_*:

$$V[f_*|X, y, x_*] = k(x_*, x_*) - k_*^T (K + W^{-1})^{-1} k_*,$$

(where $W = -\frac{\partial^2 \log p(y|f)}{\partial f \partial f^T}$.)

The result is shown in Figure 4.10.

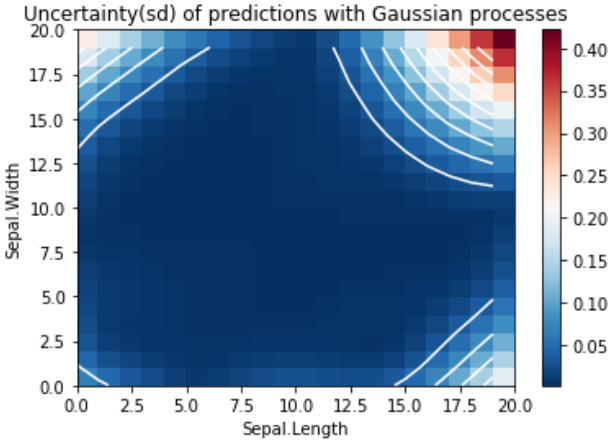


Figure 4.10: Heatmap of uncertainty

■

Exercise 4.14 *In a multi-class setting, an appropriate likelihood is the multinomial, which is parametrized by a simplex-valued vector* $\pi = (\pi_1, \ldots, \pi_K.)$. *We can map a real-valued vector* y *to the simplex using the softmax transformation:*

$$\pi_i = \frac{e^{y_i}}{\sum_{j=1}^{K} e^{y_k}}$$

To use this transformation, we will have to have one Gaussian process for each class.[1] *Practically, we can think of this as using a single Gaussian process, but with a block-diagonal covariance matrix with* K $N \times N$ *blocks. Using a Laplace approximation to the posterior distribution, repeat the previous three exercises using all three types of iris.*

Proof: Let us consider C classes, and n training points. Then, the latent vector f is :

$$\mathbf{f} = (f_1^1, ..., f_n^1, f_1^2, ..., f_n^2, ..., f_1^C, ..., f_n^C).$$

Let $\mathbf{y}$ be a vector of the same size as $\mathbf{f}$ which for each $i = 1, ..., n$ has an an entry of 1 for the class which is the label for x_i. Let π_i^c denote output of the softmax at training point i, i.e.,

$$p(y_i^c | f_i) = \frac{\exp(f_i^c)}{\sum \exp(f_i^{c'})}.$$

Then the unnormalized posterior is

$$\Phi(\mathbf{f}) = constant + -\frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} + y^T\mathbf{f} - \sum \log(\sum \exp f_i^c).$$

$$\nabla\Phi = -K^{-1}\mathbf{f} + \mathbf{y} - \pi,$$

where π is a vector of the same length as $\mathbf{f}$ with entries π_i^c.

$$\nabla\nabla\Phi = -K^{-1}\mathbf{f} - W,$$

where $W = \mathrm{d}iag(\pi) - \Pi\Pi^T$, and Π is a $Cn \times n$ matrix obtained by stacking vertically the diagonal matrices $\mathrm{d}iag(\pi^{\mathbf{c}})$.

Running the iris data, we get the prediction on training data with Gaussian process in Figure 4.11. As is shown in the plot, Gaussian process gives reasonable predictions. The first third is corrected predicted as class 1 and so on. Also, we can make predictions for new data. As is shown in Figure 4.12, the x-axis is Petal.Length and the y-axis Sepal.Width. These two variables are chosen for demonstration. Also, the uncertainty of prediction is shown and the result is consistent with the binary cases.

■

[1] Technically, we could use $K - 1$ GPs plus a constant reference value for the third class, but let's use K for now.

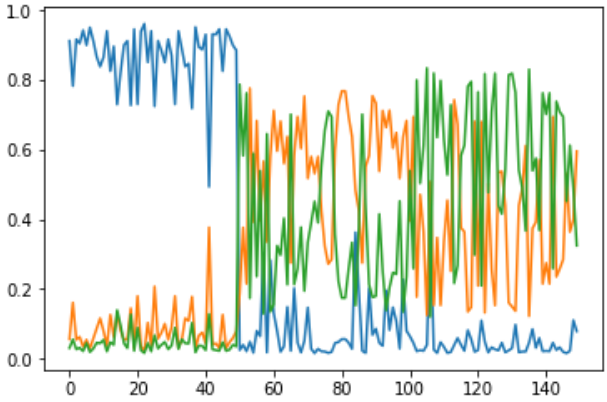


Figure 4.11: Gaussian process prediction on training data

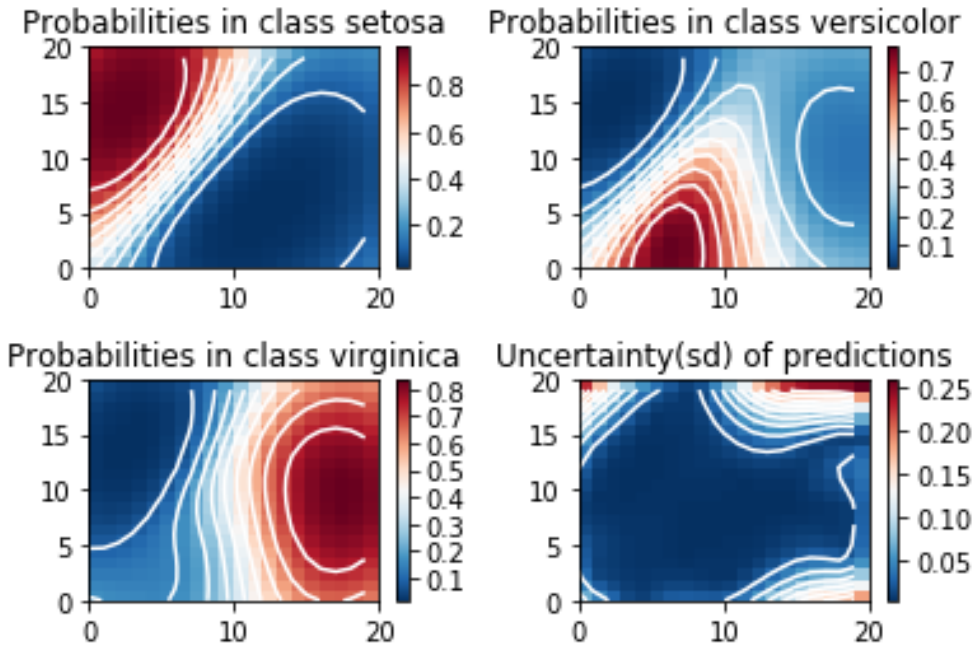


Figure 4.12: Heatmap of prediction for three classes and heatmap of uncertainty