STAT 625: Homework 1

Instruction: Submit a PDF report (scanned handwriting allowed) and a separate *executable* file for your code to Canvas. Part of this assignment is to implement Gaussian processes on real data, and the idea is to let you write your own code from scratch. This implies that you should not use existing GP toolboxes, although feel free to use packages/libraries for linear algebra, random number generators, commonly used distributions, etc. Collaborations among peers are always welcomed, but you need to write down your submission on your own.

Let $f \sim GP(0, k(x, x'))$ where $k(x, x')$ is the squared exponential kernel

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

1. Simulate and plot 5 realizations from the prior distribution of f over a grid of x's at each combination of $(\sigma_f, \ell) \in \{(1/2, 2), (1/4, 1/2), (1/2, 1/2)\}$.

2. Analyze the Canadian wages data using a Gaussian process for `logWage` $\sim$ `Age`. First standardarize the Age variable to have zero mean and unit variance. For each method below to address the parameters (σ, σ_f, ℓ), a common task is to report one single plot consisting of the posterior mean of f, 95% pointwise credible intervals, and 95% predictive intervals for new observations (all as a function of x ranging from -2 to 2.5).

   Recall that the marginal likelihood is

   $$\log p(\boldsymbol{y}|\boldsymbol{X}, \sigma, \sigma_f, \ell) = -\frac{1}{2}\boldsymbol{y}^T \left(K + \sigma^2 I\right)^{-1} \boldsymbol{y} - \frac{1}{2}\log\left|K + \sigma^2 I\right| - \frac{n}{2}\log(2\pi)$$

   (a) Empirical specification of (σ, σ_f, ℓ). Obtain your favorite nonparametric estimate of f and use residuals to estimate the model parameter σ, denoted by $\hat{\sigma}$. Then use subjective choices of $\sigma_f = 10$ and $\ell = 0.5$.

   (b) Empirical Bayes. Use the prior $\pi(\sigma^2) \propto \sigma^{-2}$ for σ^2, and reparameterize $\sigma_f^2 = \tau^2\sigma^2$. First obtain the marginal likelihood of (τ, ℓ) by integrating out σ^2 with respect to its prior distribution, which has a closed form expression. Then choose parameters (τ, ℓ) by maximizing the marginal likelihood.

   (c) Fully Bayes. Use the prior $\pi(\sigma^2) \propto \sigma^{-2}$ for σ^2, and use your favorite priors for σ_f and ℓ. Draw M posterior samples of (σ, σ_f, ℓ); M could be 1,000 or more, depending on whether the posterior samples mix well diagnosed visually from the trace plot. Plot the marginal posterior distributions of $(\log\sigma, \log\sigma_f, \log\ell)$.

   [Hint 1: Note that you only have three parameters and you can use any method you are comfortable with for the posterior calculation (for example, you may refer to the book *Bayesian Data Analysis* Chapter 10–13, particularly the slice sampling by Radford M. Neal with R code provided by this webpage).

   Hint 2: Conditional on each sample of (σ, σ_f, ℓ), you can obtain a sample of $f(\cdot)$ and predictive curve $y(\cdot)$. Aggregating all M posterior samples will give you the posterior mean and pointwise credible/predictive intervals.]

(d) Comment on the three approaches above in terms of computational complexity, posterior inference, and other high level perspectives.

3. Design a small simulation to demonstrate that your code can accurately estimate the ground truth by using empirical Bayes and fully Bayes approaches. In particular, first specify the true parameter values (σ_0^2, f_0) and three sample sizes that correspond to small, moderate, and large sample. Then generate data and run your code on simulated data. Summarize your findings using plots as in Problem 2 and a table as follows:

n	L_2 error of $\hat{f}$ using empirical Bayes	L_2 error of $\hat{f}$ using fully Bayes
small		
moderate		
large		

Here the Bayes estimator $\hat{f}$ is the posterior mean.