NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

Computational Statistics
MA-GY 6973
Fall 2021

Homework 2

Due: 5:00 PM EDT on Fri Oct 29, 2021

Instructions: The homework consists of written as well as programming exercises. Your homework is your own – while you are free to discuss the problems with other students, each of you must submit your own writeup, code, etc. There is a zero-tolerance policy regarding cheating and submission of identical work.

Submitting your homework consists of two parts:

1. A **single PDF** of your written work as well as *the output of your code* should be uploaded to Gradescope (which can be accessed via Brightspace). When uploading to Gradescope, please make sure to mark which pages contain the solutions to which problems. This saves everyone a lot of time when grading.

2. Any code that you write for the assignment must also be submitted via GitHub Classroom. You will receive an invitation URL for the assignment via email. It is *your responsibility* to setup a Github account and join the GitHub Classroom.

   The code you submit *must be easy* to execute. For example, if it is Matlab code in the file `prob1.m`, the code should run when I execute the command `matlab prob1` from the command line. If the code is in C or Fortran, for example, you must include a script which compiles and executes the code. If you do not submit code for a particular problem, you will not receive credit even if you submitted the output of your code.

---

1. Consider the following data:

x	y
0	-0.00799366
0.6981317	0.50389564
1.3962634	0.92653312
2.0943951	0.7628036
2.7925268	0.37189376
3.4906585	-0.1965461
4.1887902	-0.93017225
4.88692191	-1.04932639
5.58505361	-0.72417058
6.28318531	-0.07469816

   (a) (**10 points**) Assuming a model

$$y = f(x) + \epsilon \tag{0.1}$$

   where $f \sim \mathcal{GP}(m, k)$ and $\epsilon \sim \mathcal{N}(0, 0.01)$, write a code to perform Gaussian process regression on the data using a covariance kernel of $k(x, x') = \exp(-(x - x')^2)$. Make two plots of the expected value of the regression on $[0, 2\pi]$: one plot should assume a mean function of $m(x) = 0$, and the other one should assume a mean function of $m(x) = \sin(x)$.

(b) (**10 points**) If $\mathbf{C} = \mathbf{W}\mathbf{W}^T$ is an $n \times n$ covariance matrix, and $\mathbf{z}$ is an $n \times 1$ random vector whose entries are IID $\mathcal{N}(0, 1)$ random variables, then the $n \times 1$ vector

$$\mathbf{x} = \mathbf{W}\mathbf{z} \tag{0.2}$$

is a collection of normal random variables with covariance matrix $\mathbf{C}$. Use this fact to generate 20 draws from the posterior distribution computed in the previous part (for $m = 0$). Evaluate each of these draws at 100 equispaced points in $[0, 2\pi]$ and plot them. (You can use a built-in function in your programming language of choice to generate the IID $\mathcal{N}(0, 1)$ realizations.)

2. (**10 points**) Suppose that $X_1, \ldots, X_n$ are n random variables with zero mean and finite variances σ_i^2. Denote by $\mathbf{C}$ the associated $n \times n$ covariance matrix and by $\mathbf{R}$ the associated $n \times n$ correlation matrix. Suppose that $\mathbf{C}$ has unique eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. What is the relationship between the eigenvector of $\mathbf{C}$ corresponding to λ_1 and the eigenvector of $\mathbf{R}$ corresponding to its largest eigenvalue? Often when doing principal component analysis, the variables X_i are normalized to have unit variance. Given the above calculation, how is the principal component of the normalized variables related to the principal component of the un-normalized variables?

3. (**10 points**) Let $\mathbf{A}$ be an $m \times n$ matrix with $m > n$ and rank $k < n$. Let $\mathbf{A} = \mathbf{Q}\mathbf{R}$ be a QR factorization of $\mathbf{A}$ (note that $\mathbf{A}$ is rank deficient). Show how this factorization be used to find the least-squares solution

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2. \tag{0.3}$$

4. (**10 points**) Let $\mathbf{A}$ be an $m \times n$ matrix with $m < n$ and rank m. The equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ clearly has multiple solutions. Show how the singular value decomposition of $\mathbf{A}$ can be used to find the solution $\mathbf{x}_0$ that has minimum ℓ_2 norm among all solutions of the system.