STA561/CS571 — Fall 2013 **Homework 5** Due: October 28, 2013

Instructions: Please put all answers in a single PDF with your name and NetID and upload to SAKAI before class on the due date (there is a LaTeX template on the course web site for you to use). Definitely consider working in a group; please include the names of the people in your group and write up your solutions separately. If you look at any references (even wikipedia), cite them. If you happen to track the number of hours you spent on the homework, it would be great if you could put that at the top of your homework to give us an indication of how difficult it was.

Problem 1

PCA

Let $X = \Lambda Z + \epsilon$ where $X \in \mathbb{R}^{p \times n}$, $\Lambda \in \mathbb{R}^{p \times k}$, $Z \in \mathbb{R}^{k \times n}$, and $\epsilon \in \mathbb{R}^{p \times n}$. As in Factor analysis, assume that the entries of Z have standard normal distribution priors, and that ϵ_i follows a $\mathcal{N}_p(0, \Psi)$ distribution for $i \in \{1, \ldots, n\}$ where Ψ is diagonal. Unlike in the FA model, let each element of the Ψ matrix $\psi_j = \psi$ (i.e., all of the diagonal elements are the same). This is a model known as Probabilistic PCA.

Now, generate three different matrices X in the following way: set $n = p = 100$, set $k = 3$, and generate each element of $z_{i,k} \sim \mathcal{N}(0, 1)$, and similarly for each element of $\lambda_{k,j} \sim \mathcal{N}(0, 1)$. Generate matrix $X = \Lambda Z$ and then add on $\mathcal{N}(0, \psi)$ noise to each element, where $\psi = \{0.2, 2, 10\}$. You should have three matrices now, each 100×100, and each generated from a low dimensional subspace.

In this question, you will reconstruct the covariance of this matrix using eigenvalues and eigenvectors.

(a) Use the `eigen()` function in R to compute the eigenvalues and eigenvectors for the covariance of X (cov(X)) for each of the three matrices, and plot the normalized eigenvalues (turn in this plot). How does the distribution of the eigenvalues change as the amount of noise (ψ) in the original matrices increases?

(b) Compute the RMSE between $\mathsf{Cov}(X)$ and the matrix reconstruction using the first three eigenvectors:
`xeig$vectors[,1:3] %*% diag(xeig$values[1:3]) %*% t(xeig$vectors[,1:3])`,
or $\Phi\Omega\Phi^T$ for Φ the truncated matrix of eigenvalues and $\Omega = diag(\omega)$ for ω eigenvalues. How well do these eigenvectors recapitulate the original low dimensional data matrices? How does this change as the amount of noise increases?

(c) What is the effect of reconstructing these matrices using the first three eigenvalues? How might this be useful for a specific application?

Problem 2

String Kernel and Gaussian Processes.

In molecular biology, transcription factors (some proteins) typically recognize and bind to specific DNA regulatory sequences. The protein binding microarray (PBM) is a novel biotechnology that offers a quantitative way to measure the DNA binding specificities of any single transcription factor protein. In this task, we are trying to build a Gaussian processes regression model to predict the binding intensities of a transcription factor on a set of DNA sequences. In SAKAI, we have 1000 training sequences and their measured intensities (in the training_data_gp.txt file), and 100 test sequences and their measured intensities (in the test_data_gp.txt file).

(a) Write a program to implement the *string kernel* on these DNA sequences, using the lengths of substrings$|A| = \{1, 2, 3\}$, and compute the Gram Matrix K for the samples. (*Hint*: The alphabet of a DNA sequence is "A", "C", "G", "T".) For this limited set of underlying strings, you can do this in a brute force way, instead of using a suffix tree (i.e., enumerate all of the strings, and count the number of occurrences of each string in each DNA sequence. Then $\kappa(x_i, x_j)$ is a simple function of those vectors).

(b) Using Gaussian processes regression (use RBF kernel with the characteristic length scale $\sigma = 5$), predict the protein binding intensities for the test data set using the training data set.
Note: For questions a and b, write down the steps for building your string kernel and equations for your Gram Matrix and Gaussian processes regression, and paste your code at the end.

(c) Draw a scatter plot for your predicted intensities versus the measured intensities in the test data. Calculate RMSE for the predicted intensities versus measured intensities.

(d) How might you improve the prediction accuracy (*name three different ways*)?

Note: This problem is still an open and challenging area in genomic sciences.