Homework # 6

Reading

- Attached you will find an article by Persi Diaconis, *The MCMC Revolution* published in 2009 at the height of the MCMC euphoria. The article touches on many of the ideas we have discussed and provides some interesting examples.

1. Attached you will find the files `TimeSeries2.csv`, `TimeSeries3.csv`, `TimeSeries4.csv`. Each of these files contains a 1000×20 matrix constructed in the same way as the time series dataset from hw 4. The difference is that `TimeSeries2.csv` contains only two base times series, while `TimeSeries3.csv` and `TimeSeries4.csv` contain 3 and 4 base time series, respectively.

   (a) Before doing any computations, decide what might be the best dimension K to use for a PCA in approximating each of these three datasets. Explain your reasoning. (No wrong answer here. I just want you to think things through.)

   (b) For each of the three data files do the following.

      i. Compute the eigenvalues of the covariance matrix. Then, plot the fraction of the dataset's variance captured by a K-dimensional PCA for $K = 1, 2, \ldots, 20$. Discuss how the fraction of variance captured as K varies reflects the number of base time series.

      ii. Use a 2-d PCA to reduce the dimensionality of the data and then produce a plot of the 1000 data points. (In a 2-d PCA each $X^{(i)}$ corresponds to a $(c_1^{(i)}, c_2^{(i)})$. Plot the $c^{(i)}$). Do the data points separate into the appropriate number of clusters?

      iii. For the 2-d PCA, compute the correlation between $c_1^{(i)}$ and $c_2^{(i)}$ for $i = 1, 2, \ldots, 1000$. Then compute the variance of $c_1^{(i)}$ and the variance of $c_2^{(i)}$ and relate to the eigenvalues you computed in (a).

   In answering (i)-(ii), do not use R or Python's pca function(s). Instead compute the PCA yourself. You can call R's **eigen** function and the Python equivalent.

2. Suppose we roll 100 dice and the sum of the die rolls is 450. Use an MCMC approach to determine (i) the expected number of 6's rolled (ii) the expected number of 1's rolled and (iii) the probability that we roll less than thirty 1's. Do all of these using a single run of a Markov chain. Be sure to include a burn-in time and run you chain for a sufficient length of time.