Homework Assignment 1

Due in class, Thursday September 17th

SDS 383C Statistical Modeling I

1. **Jeffreys prior:** In class we read about MLE and MAP estimators. In this exercise we will also derive the posterior mean using a flat Jeffreys prior on the mean parameter. The posterior mean obtained using an improper prior is also known as the generalized Bayes estimator under square risk.

   Let $X_1, \ldots, X_n \sim N(\mu, I_{p\times p})$ where $\mu = (\mu_1, \ldots, \mu_p)^T$ and $I_{p\times p}$ is the p dimensional identity matrix. Each data point X_i is a vector of length p. Facts : If $Z_i \sim N(\mu_i, 1)$ for $i = 1, \ldots, p$, then $\sum_{i=1}^p Z_i^2$ has a non-central Chi-square distribution with p degrees of freedom and non-centrality parameter $\sum_{i=1}^p \mu_i^2$. The mean and variance of this distribution are given by $\|\mu\|^2 + p$ and $4\|\mu\|^2 + 2p$, where $\|\mu\|^2 = \sum_{i=1}^p \mu_i^2$.

   (a) Prove that the Jeffreys prior on μ is a flat prior. Recall that for a multidimensional parameter, the Jeffreys prior is proportional to $\sqrt{|I_n(\theta)|}$, where $|M|$ is the determinant of matrix M. The determinant of the identity matrix is 1.

   (b) Find the posterior distribution of μ under the flat prior.

   (c) We want to estimate $\theta := \|\mu\|^2$. Find the posterior mean of θ. Denote this by $\tilde{\theta}$.

   (d) The usual frequentist estimate is $\hat{\theta} = \|\bar{X}_n\|^2 - p/n$. Show that for any n, as $p \to \infty$, $\frac{E[(\theta-\tilde{\theta})^2]}{E[(\theta-\hat{\theta})^2]} \to \infty$, where the expectation is taken over the data. *You can use the fact the famous bias variance decomposition of the mean squared error:* $E[(\hat{\theta} - \theta)^2] = (E[\hat{\theta}] - \theta)^2 + var(\hat{\theta})$

   (e) Using $n = 20$, $p = 2000$, simulate 1000 datapoints from a $N(\mathbf{0}, I_{p\times p})$ distribution. Plot the histograms of $\hat{\theta}$ and $\tilde{\theta}$. What do your results suggest? *The R command for generating from a multivariate gaussian is "mvrnorm", whereas in matlab this command is "mvnrnd". The command for creating a histogram is "hist" for both.*

2. **Maximum Likelihood Estimates:** While the MLE enjoys interesting theoretical properties, its not necessary that it exists or is unique. We illustrate this with the two distributions specified below. For these questions, you must provide a proof.

   (a) Let $X_1, \ldots, X_n \sim Uniform([\theta, \theta + 1])$. What is the MLE? Is it unique? Does it exist?

   (b) $X_1, \ldots, X_n \sim Uniform([0, \theta))$. What is the MLE? Is it unique? Does it exist?

   (c) Let $X_1, \ldots, X_n \sim N(\mu, 1)$. Let $\theta := e^\mu$. Create a data set with $\mu = 5$ with a hundred observations.

      i. Use the delta method to get the variance of the estimator and a 95% confidence interval of θ.

ii. Compare the above with the corresponding quantities estimates using the parametric bootstrap.

3. In class we saw that the beta distribution is a conjugate prior to the binomial. As it turns out, the conjugate prior to multinomial is the Dirichlet distribution on the k-simplex. The density is given by:

$$f(x|\alpha) = \frac{\Gamma(\sum_{i=1}^{k+1} \alpha_i)}{\prod_{i=1}^{k+1} \Gamma(\alpha_i)} \prod_{i=1}^{k+1} x_i^{\alpha_i - 1}$$

where $x_i > 0$, $\sum_i^{k+1} x_i = 1$ and $\alpha_i \geq 0$.

(a) Prove that $E[\log x_i] = \Psi(\alpha_i) - \Psi(\sum_{i=1}^{k+1} \alpha_i)$,
where $\Psi(\alpha) = d \log \Gamma(\alpha)/d\alpha$ is the digamma function. *Hint: you can use the fact that if $X \sim Beta(\alpha, \beta)$, then $E[\log X] = \Psi(\alpha) - \Psi(\alpha + \beta)$.*

(b) For n data-points $\{x^{(i)}, i = 1, \ldots, n\}$ generated from $f(x|\alpha)$, show that the MLE $\hat{\alpha}$ satisfies

$$\overline{\log x_i} = \Psi(\hat{\alpha}_i) - \Psi(\sum_{i=1}^{k+1} \hat{\alpha}_i),$$

where $\overline{\log x_i} = \sum_j \log x_i^{(j)}/n$ is the average computed from data. Since the MLE cannot be computed in closed form, we will use numerical methods to find the MLE. Simple algorithm for doing that is

$$\Psi(\alpha_i^{new}) = \overline{\log x_i} + \Psi(\sum_{i=1}^{k+1} \alpha_i^{old})$$

This will require inverting the digamma function. For a simple method of doing so look at `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf` (appendix C).

(c) Use the dataset dir1.txt for a Dirichlet 2-dimensional simplex. It can be found next to the homework link on your instructor's course web page. Now do the following:

i. Give a scatter plot of the data.

ii. Compute the MLE $\hat{\alpha}$, and plot the log-likelihood as a function of iteration. Briefly give a description of the algorithm you use. You can use the built in R/matlab code for digamma, gamma or related functions.

iii. Give a scatter plot of the data together with a contour plot of the Dirichlet distribution with parameters $\hat{\alpha}$ that you have computed.