Chi-Square Exercises

March 7, 2014

Problem 1

Show that if $x \sim N(\mu, \Sigma)$, where Σ is a $n \times n$ positive definite matrix, then

$$(x-\mu)^T \Sigma^{-1} (x-\mu) \sim \text{Chisquare}(n).$$

Problem 2

Plot the pdf of the χ^2 distribution for $\nu = 1,\ 2,\ 3,\ 10$. (Wolfram Alpha makes it dead easy.) Explain intuitively why χ^2 is distributed like it is in each case.

Problem 3

The goal of this problem is to get you comfortable with handling the χ^2 distribution, both mentally and in code. The final products should be three graphs, an overall χ^2 value, and an overall p-value.

1. First, define parameters d and *num_points* in your code for the dimension and the number of sample points. Keep these parameterized so you can play around with them later. Start with values $d = 2$ and $num\_points = 200$.
2. Generate random μ and Σ of dimension d.
3. Generate *num_points* random d-dimensional multivariate normal data points with your μ and Σ.
4. Calculate the sample mean and covariance matrix of your sampled data.
5. Calculate two arrays of χ^2 values, each with one χ^2 value for each data point. Calculate the first array using the true values of μ and Σ, and calculate the second array using the calculated sample values $\bar{x}$ and $\bar{\Sigma}$.
6. Calculate and print an overall χ^2 value for the data, as well as its corresponding p-value, using both the arrays of χ^2 values.
7. Plot a 2-d scatter plot of the first two dimensions of your sample data.
8. Plot two normed overlaying histograms of the values in your two χ^2 arrays. If you are using python, the command you want is plt.hist(array, num_bins, normed=True, histtype='step'). On top of these histograms plot the theoretical pdf of the appropriate χ^2 distribution.
9. Plot the pdf of the χ^2 distribution pertaining to the whole data set. Plot a vertical line at the calculated χ^2.
10. Now try lots of different combinations of d and *num_points*. Try all combinations of $d = 2, 3, 5, 10, 20$ and $num\_points = 20, 40, 200, 500, 2000$.

Problem 4

1. A dataset consisting of 1010 points in $\mathbb{R}^5$ is available on the wiki (see today's page) or as a python notebook resource ('data/mv_chi.txt'). Fit a multivariate Gaussian to this data.

2. 1000 of the points in this dataset were drawn from a multivariate Gaussian. Ten of the points are imposters and were not. Compute the χ^2 statistic for each point and use the result of Problem 1 to assign p-values to the points. Identify a set of points that you believe are imposters by applying the false discovery rate prescription with an α of 10%.