Problem Set #1

MACS 30250, Dr. Evans
Due Monday, May. 4 at 1:30pm

1. **1D kernel density estimator (5 points).** The COVIDincubation.txt file is a comma-delimited data file that includes 59 observations on individuals in China who tested positive for the COVID-19 virus. The data contain the following three variables: `gender` ("M" or "F"), `age` (integers between 10 and 70), and `symp_days` (days until symptomatic, float). This is a subset of the variables in the dataset used by Men et al. (2020). The `symp_days` variable represents the incubation period for each individual, or the number of days until symptoms were manifest.

   (a) Create three histograms, each of `symp_days` (Incubation period, days to symptomatic). The first one is the overall histogram. Let each histgram have 15 bins over the range of days from 2 to 15 (the maximum in the data). In the `matplotlib.pyplot.hist()` call, set the histogram to density `density=True`. Let the first histogram be for all the data. Let the second histogram be for individuals of `age`$\leq$ 40, and let the third histogram be for individuals `age`$>$ 40.

   (b) Fit a Gaussian KDE as an approximation ot the incubation period for each of the three subsets of data from part (a). Use the `GridSearchCV` and `LeaveOneOut` methods as in the VanderPlas notebook to choose an optimal bandwith, and report your optimal bandwidths for the three KDEs. For your grid search, use 500 exponentially spaced bandwidths between 0.1 and 10 using the code: `bandwidths = 10 ** np.linspace(-1, 1, 500)`. Plot each of the KDE distributions in one plot with a legend that shows which is which. This figure should look like the right panel of Figure 2 in Men et al. (2020).

   (c) What does this tell you about COVID-19 incubation periods of young versus old individuals?

2. **2D kernel density estimator (5 points).** This exercise uses two data files: `BQ_ind_data.txt` and `BQ_probmat.txt`. The first data file (`BQ_ind_data.txt`) contains 70,000 observations on two variables: `age` and `income_pctl`. These represent individual data where each observation is an individual that represents a person who received an equal bequest. So age and income groups that received more bequests will be represented by more individuals (observations). The second data file (`BQ_probmat.txt`) contains the empirical histogram information. It is a 73 $\times$ 7 matrix of percentages representing the values of a two-dimensional histogram of the percent of the U.S. population that receives all the bequests (inheritances) by a recipient's age (ages 18 to 90, rows) and by

a recipient's lifetime income group (7 categories, columns). The seven lifetime income groups are percentiles. Let $prcntl_j$ be the percent of the population in lifetime income group j. The lifetime income groups in the $J = 7$ columns of the BQ_probmat.txt data are the following.

$$\boldsymbol{prcntl} = \big[0.25, 0.25, 0.20, 0.10, 0.10, 0.09, 0.01\big], \quad \text{such that} \quad \sum_{j=1}^{7} prcntl_j = 1$$

You can read this file into memory using the `numpy.loadtxt` function.

```
bq_data = np.loadtxt('BQ_probmat.txt', delimiter=',')
```

So the $[11, 5]$-th element of the `bq_data` matrix represents the percent of total bequests (inheritances) received by age-28 and lifetime income group $j = 5$ (80th to 90th percentile of lifetime income).

(a) Read in the BQ_probmat.txt data as a 73×7 `NumPy` array. Plot the 2D empirical histogram of these data as a 3D surface plot with age and income group on the x-axis and y-axis and the histogram density on the z-axis using a 3D surface plot tool (not a 3D bar histogram tool). Make sure that the axes are labeled correctly. And make sure that your 3D histogram is presented from a perspective that allows a viewer to see that data (don't let the data be hidden by a poor angle of the plot.)

(b) Use the BQ_ind_data.txt data to fit a bivariate Gaussian kernel density estimator to the data using the `scipy.stats.gaussian_kde` method. Choose a bandwidth parameter λ that you think is best. Justify your choice of that parameter. Your justification should have to do with the tradeoff between overfitting (too low a value) and underfitting (too high a value). Plot the surface of your chosen kernel density estimator. Make sure that the axes are labeled correctly. And make sure that your 3D histogram is presented from a perspective that allows a viewer to see that data. What is the estimated density for bequest recipients who are age 61 in the 6th lifetime income category ($j = 6$, 90th to 99th percentile).

References

Men, Ke, Xia Wang, Yihao Li, Guanwei Zhang, JingjingHu, Yanyan Gao, and Henry Han, "Estimate the Incubation Period of Coronavirus 2019 (COVID-19)," February 2020. Unpublished.