Please submit your homework in pdf form, which can be either a scanned copy of your hand-written answers, or, computer generated documents (e.g., via word/latex etc). Please include computer calculated results, necessary scientific figures (clearly label your axes with proper units) along with description of the method, conclusions and discussion to the problem.

Model regression, Fourier Transform and Data Structure

1. Model Regression (10 pts)

(1) A polynomial function of degree 3 is given by:

$$y \equiv f(x) = 7 + 2(x - 0.2) - 3(x - 0.5)^2 - 6(x - 0.8)^3. \quad (1)$$

Use Monte Carlo method to generate $N = 30$ points to randomly (uniformly) sample $x \in [0, 2]$, and work out the expected y values at these x locations.

(2) Now let us assume a Gaussian error behavior that at each sampled location x_i, the error in y_i around $y_i^* = f(x_i)$ has a Gaussian standard deviation of $\sigma = 0.1$. Use Monte Carlo method to generate an error δ_{y_i} according to such a distribution for every data point $(x_i, y_i = y_i^* + \delta_{y_i})$.

(3) The 30 data points of (x_i, y_i) generated above now compose a measurement sample, with x being the independent variable and y the dependent variable. Suppose you have no idea about the true model through which this sample is generated. *The only information that you have is the Gaussian error behavior which has a standard deviation of* $\sigma = 0.1$ *(in reality, this can be a known measurement uncertainty).* Now play the linear regression game to fit the data, and using both (a) information criteria and (b) cross-validation method to find out the best model (and its parameters) that describes existing data.

(4) If changing $\sigma = 0.1$ to $\sigma = 1.0$ to generate the mock data and re-do the procedure above, do you get the same best model? Why?

(5) Using local regression with a Gaussian kernel to make prediction of y_j at new positions x_j based on existing data points $\{(x_i, y_i)\}$. Applying leave-one-out cross-validation to decide a reasonable bandwidth h. Then vary this h by a factor of 1/3 and 3, to see the model predictions.

In your final answer, please give: (1) the mathematical forms of your models (you are not limited to use the polynomial format); (2) the procedure and justification that you obtain the *best* model (and model parameters). Please also plot: (a) the polynomial distribution given by Eq. (1); (b) the 30 data points (x_i, y_i) generated by the Monte-Carlo method; (c) the best regression model that fits the data; (d) the model regression result from local regression with a Gaussian kernel.

2. Fourier signal searching (10 pts)

Sample a true signal function $y(t) = \sin(2t + 5) + 2\cos(t/3 - 7)$ using $N_{\rm D} = 80$ points that are randomly drawn from $t = 0$ to $t = 100$. Plot the sampled data on top of the input signal. By eye, can you spot any periodic feature? If the $N_{\rm D}$ sampling locations were evenly distributed, what is the highest and lowerest sampling frequencies according to the Nyquist-Shannon sampling theorem?

How would you decide whether there are any periodic features hidden in the $N_{\rm D}$ measurements? How would you further pick up the frequencies of the input signal?

Hint: use a finer and evenly-spaced time grid with $N_{\rm res} = 128$ between $t = 0$ and $t = 100$ to "re-sample" the existing measurements by adopting a window function on this finer grid, setting the window function to be 1 at the observed location and 0 elsewhere. Using DFT on this finer grid, calculate and plot power spectra density distributions of both the signal and the window function.

Then use $N_{\rm D} = 30$ to sample the original signal and repeat the same experiment, what do you see? Explain why.

3. Calculate correlation functions and search for structures (10 pts)

Please use Random Number Generator to generate three sets of bivariate Gaussian distributions on the $X - Y$ plane. The three components have standard deviations of 0.05, 0.15 and 0.3, and realized with 500, 800 and 1000 particles, respectively. The *centres* of these bivariate Gaussians are randomly located within a square region of 1×1.

(1) Introduce a $2D$ uniform mesh (of a reasonable size and spatial resolution) to cover the majority of the particle region, assign the particles to this mesh using the Cloud-In-Cell method. Make sure all three Gaussian components can be well resolved by your mesh grid. Through FFT calculate the clustering of your data points on scales larger than the mesh cell size. Plot both 1d correlation function $\xi(r)$ (in real space) and 1d power spectrum $P(k)$ (in frequencey space). (3 pt)

(2) Use the Pair-Estimation method with the Landy & Szalay estimator (1993) to calculate 1d two-point correlation function $\xi(r)$ of the density field. Compare the differences between this result and the FFT-based result from (1), plot them in the same figure. Explain what you see. (3 pt)

(3) Apply the Gaussian mixture model (GMM) method to estimate the underlying density field. Use Information Criteria or any cross-validation method, estimate how many Gaussian components best describe the particle distribution? Explain how you reach this number and label out the different GMM components with circles (ellipses) on top of the particle distribution. (4 pt)