Written Report – 6.419x Module 5

Name: ptoche

Long calculations and code snippets are in the appendix at the end of this report. Autograded sections are also included.

Part I – Ocean Flow

Problem 1a: Ocean Flow (autograded)

Provide the coordinates (in kilometers) of the point with smallest variation in speed flow (magnitude of the vector). Hint: Recall that technically the speed flow inland is zero.

You should first calculate the magnitude for each data point, $\sqrt{u^2 + v^2}$, then compute the variance across time. Some things to look out for:

- *Some elements do not have meaningful data (all zeros, generally land and the border of the map). Remove any locations with a variance of zero before finding the minimum.*
- *The zero-indexed element $(0, 0)$ corresponds to $(0km, 0km)$. The grid spacing is 3 km. Multiply your zero-indexed indices by 3 to get the location in kilometers.*
- *Libraries such as pandas will load these `.csv` files with the first index over rows, and the second index over columns. Double check the ordering of indices and ensure they correspond to the definition of the x and y axes.*

Please enter the x and y coordinates in kilometers.

Problem 1b (autograded)

Provide the coordinates (in kilometers) and the time stamp (in hours) of the point where the flow has its maximum x-axis velocity (the maximum signed value).

Something to look out for: One answer is in hours, as specified above, the zero-indexed element corresponds to $0hrs$. The time spacing is 3 hours. Multiply your zero-indexed time index by 3 to get the location in hours.

Please enter the time in hours. Please enter the x and y coordinates in kilometers.

Problem 1c (autograded)

Take the average of the velocity vector over all time and positions, so that you get an overall average velocity for the entire data set.

Note: You should average over land positions for this problem, do not attempt to remove them from the average. (You may want to consider in your own time how this might bias the result.)

Please enter the x (horizontal, u files) component of this velocity, in kilometers/hour. Please enter the y (vertical, v files) component of this velocity, in kilometers/hour.

Problem 2: Identifying long-range correlations (10 bonus points)

In this problem, we will try to identify areas in the Philippine Archipelago with long-range correlations. Your task is to identify two places on the map that are not immediately next to each other but still have some high correlation in their flows. Your response should be the map of the Archipelago with the two areas marked (e.g., circled). Explain how you found that these two areas have correlated flows.

A point is a particular location in the provided data set. We have measurements of the flow velocities in the x-axis and y-axis for each point at 100 *different times. To compute the correlation between two points, select two points and compute the correlation coefficient along the x-direction, or the y-direction, or both. That is a correlation between two vectors of dimension* 100.

Hints:

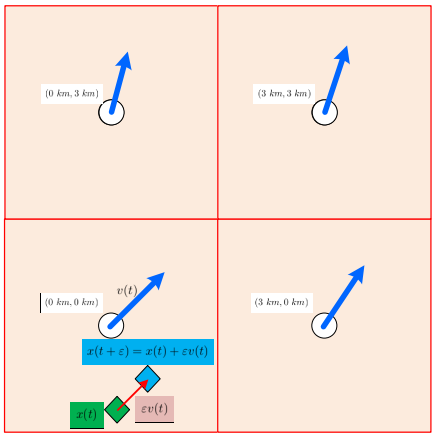
- *The provided data set is quite large, with a map of 555×504 points. Computing the correlation between every possible pair of points is computationally expensive. Instead of computing $(555 \times 504)^2 = 78,243,278,400$ correlations, you may randomly sample pairs of points.*
- *Since this is a relatively small area, there might be many correlations between the points, so set the threshold to define "high correlations" sufficiently large. If it is too small, you will find that most of the points are correlated. If it is too high, no pairs of points will be correlated.*
- *An area might be a single point correlated with other points. However, maybe there are clusters of points that are correlated with other clusters of points.*
- *Remember that correlation can be positive or negative. You are free to select whether you want to find positively correlated areas or negatively correlated areas.*
- *You are advised to find a correlation for each direction separately. You are free to select how to combine these two values. Maybe set the correlation between the two points as the maximum directional correlation. Average between the two directional correlations also work. Of course, the minimum as well.*

Rubric: *Award each of the following according to whether the requirement is met:*

- *(5 points): A map with the two points with correlations marked.*
- *(3 points): Provides an explanation of how the correlation was computed.*
- *(2 points): Provides a convincing commentary on why the two marked locations could be correlated.*

Problem 3: Simulating particle movement in flows (20 points)

In this problem, you are asked to build a simulator that can track a particle's movement on a time-varying flow. We assume that the velocity of a particle in the ocean, with certain coordinates, will be determined by the corresponding water flow velocity at those coordinates. Implement a procedure to track the position and movement of multiple particles as caused by the time-varying flow given in the data set. Explain the procedure, and show that it works by providing examples and plots.

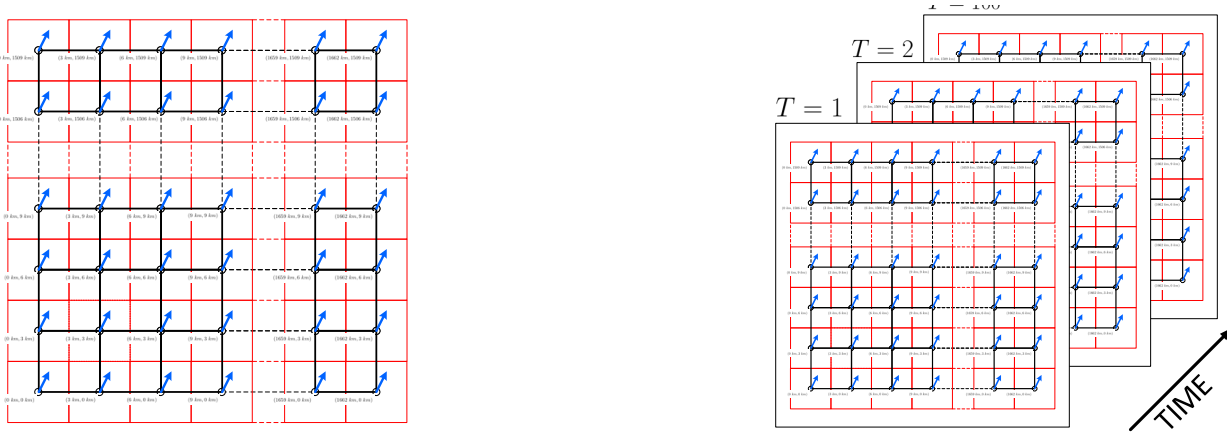


An explanation of simulating the movement of a particle over a flow.

An explanation of simulating the movement of a particle over a flow. The above figure shows a simple flow system with four points, shown as white circles. Each point corresponds to a physical location also shown in kilometers. Attached to each point is a flow data point, which is shown as a blue arrow. Recall that flow data is given in the x and y direction. It is assumed that a particle moving in one of the zones or boxes acquires the velocity given by the flow data. This example shows that at time t, a particle is located at a position $x(t)$, shown as a green diamond. After some time ϵ has passed, at time $t+\epsilon$, the particle is at a new location $x(t+\epsilon)$ shown as a blue diamond. The new location can be computed as the previous location $x(t)$ plus a displacement computed as the time-step ϵ multiplied by the velocity at that point $v(t)$. Note that the velocity is given in kilometers per hour. Thus, the location $x(t)$ must be given in the same distance units, e.g., kilometers, and the time-step ϵ should be given in hours.

Suggested approach: *The data provides a discretization of the ocean flow. The particles will, however, be moving on a continuous surface. For simplicity, let us assume that the surface is the plane $\mathbb{R}$. The data can be seen to provide flow information at integer points, namely at (m, n) for m and n integers. Divide the continuous surface into squares in such a way that each square contains a unique data point. One way to achieve this is to assign to every point in the surface the closest data point. For instance, given $(x, y) \in \mathbb{R}^2$, this consists of rounding both x and y to the closest integer. You may then suppose that each square has the same flow information as the data point it contains. Then take a particle at (x, y) in a particular square. The flow in the square will displace it at the corresponding velocity. Once the particle moves out of this square, it is governed by the flow information of the next square that it enters.*

The below figure shows a grid that you should have in mind when building the simulator. For each point, which corresponds to a physical location, there is a flow vector given, for each time step (recall there are 100 *time instances).*



A grid with the given data.

Problem 3a (10 points)

Draw particle locations uniformly at random across the entire map — do not worry if some of them are placed on land. Simulate the particle trajectories for 300 *hours and provide a plot of the initial state, a plot of the final state, and two plots at intermediate states of the simulation. You may wish to draw colors for your particles in order to help distinguish them.*

Rubric: *Award each of the following according to whether the requirement is met:*

- *(3 points): Provides an explanation of the simulation algorithm, with equations for the evolution of the particle trajectory.*
- *(2 points): Provides a plot of the initial state of the simulation.*
- *(3 points): Provides two plots of intermediate states of the simulation.*
- *(2 points): Provides a plot of the final state of the simulation.*

Problem 3b (10 points)

A (toy) plane has crashed north of Palawan at $T = 0$*. The exact location is unknown, but data suggests that the location of the crash follows a Gaussian distribution with mean* $(100, 350)$ *(namely* $300km$*,* $1050km$*) with variance* σ^2*. The debris from the plane has been carried away by the ocean flow. You are about to lead a search expedition for the debris. Where would you expect the parts to be at* $48hrs$*,* $72hrs$*,* $120hrs$*? Study the problem by varying the variance of the Gaussian distribution. Either pick a few variance samples or sweep through the variances if desired.*

Rubric: *Award each of the following according to whether the requirement is met:*

- *(3 points): Provides plots showing the state of the simulation at the times:* $T = 48hrs$*,* $72hrs$*,* $120hrs$*. (Three plots required.)*
- *(3 points): Two or more additional choices of the variances were tried, and three plots of the state of the simulation at the above three times are provided. (Six additional plots required.)*
- *(4 points): Comments on where one should concentrate search activities based on the observed results.*

Part II – Estimating Flows with Gaussian Processes

Problem 4 (20 points)

In the previous exercises, we studied flow data. The data used was given for a set of physical locations and, for each location, over 100*-time instances separated by* 3 *hours. In the previous exercise, we simulated possible locations for the debris after a specific time, given some estimated location for the (toy) plane crash. However, because we only have data for* 100*-time instances separated by* 3 *hours, we can only run such simulation for a total of* 300 *hours. In this problem, we will perform a toy experiment where we can simulate larger time scales. We will model flows as a Gaussian process to estimate some unobserved variables.*

Change in the assumptions: *From here onward, we assume that the flow data available is given for measurements taken every* 3 *days (instead of* 3 *hours). This implies that we can potentially simulate the movement of debris over a period of nearly one year. On short time scales the points do not move too far. This is no longer true on longer time scales. None the less, we make the unrealistic assumption that ocean flows remain approximately the same over each three-day period.*

Problem 4a (10 points)

Pick a location of your liking from the map for which you are given flow data. Consider the two vectors containing the flow velocity for each direction (two vectors of dimension 100*). Find the parameters of the kernel function that best describes the data for each direction, treating the directions as independent. For each step, please clearly state your selections, your thought process, and design choices.*

Estimate the parameters of the kernel function:

- *Pick a kernel function. Please clearly explain you selection.*
- *Identify the parameters of the kernel function. For example, if you select the squared exponential kernel function, then its parameters are* $\theta = \{\sigma, l\}$*.*

$$K(z_i, z_j) = \sigma^2 \, exp\left(\frac{||z_i - z_j||^2}{l^2}\right)$$

- *Find a suitable search space for each of the parameters. For example, for the length-scale parameter* l*, you could consider a range such as* 7.2 *hours to* 360 *hours (*0.1 *to* 5 *time indices).*
- *Estimate a set of parameters via cross-validation. Each vector is of dimension* 100*. Pick a number* k *and split the data* k*-wise to define training and testing points. For instance, if* $k = 10$*, you will have ten partitions, with* 90 *points for training and* 10 *points for testing.*

For each possible set of parameters and each data partition, do the following:

- *Estimate the mean at each possible time instance, e.g. a moving average over the training data.*
- *Construct the covariance matrix for the selected kernel functions and the selected set of parameters.*

- *Compute the conditional mean and variance of the testing data points, given the mean and variance of the training data points and the data itself. Recall that:*

$$\mu_{X_1|X_2} = \mu_1 + \sigma_{12}(\sigma_{22} + \tau I)^{-1}(x_2 - \mu_2)$$
$$\sigma_{X_1|X_2} = \sigma_{11} - \sigma_{12}(\sigma_{22} + \tau I)^{-1}\sigma_{22}$$

  In this case, X_1 are the velocities for the testing data-points; X_2 are the velocities for the training data-points; μ_1 are the mean velocities for the testing data-points; μ_2 are the mean velocities for the training data-points; σ_{22} is the covariance of the training data-points; σ_{11} is the covariance of the testing data-points; σ_{12} is the cross-covariance; x_2 are the observed velocities for the training data-points; τ is the variance of the noise in the observations. You can pick $\tau = 0.001$.

- *Compute the log-likelihood performance for the selected parameters. You are free to select to compute it only on the testing data, or on the complete vector.*

  For each possible set of parameters, you will then have the performance for each of the partitions of your data. Find the parameters that maximize performance. Save the computed cost/performance metric for each choice of parameters, then create a plot over your search space that shows this metric.

  ***Rubric:** Award each of the following according to whether the requirement is met:*

- *(1 point): States the choice of kernel function and provides a justification for this choice.*
- *(1 point): Identifies the parameters of the kernel function.*
- *(1 point): Explicitly states the search space for each kernel parameter.*
- *(1 point): Explicitly states the number of folds (k) for the cross-validation.*
- *(3 points): Provides the optimal kernel parameters from the search.*
- *(3 points): Provides a plot of the computed cost/performance metric over the search space for the kernel parameters.*

Problem 4b (5 points)

Run the process described for the location in (a) for at least three more locations in the map. What do you observe? Which of your kernel parameters show patterns? Which do not?

***Rubric:** Award each of the following according to whether the requirement is met:*

- *(3 points): Provides the optimal kernel values for three new location that are different from the location in Problem 4a.*
- *(2 points): For each kernel parameter, states if a pattern was observed.*

Problem 4c (5 points)

We have suggested one particular value for τ. Consider other possible values and comment on the effects such parameter has on the estimated parameters and on performance. Try at least two values different from that used in Problem 4a.

Rubric: *Award each of the following according to whether the requirement is met:*

- *(1 point): Provides the optimal kernel values for at least two new choices of τ.*
- *(2 points): A plot showing the cost/optimization target is provided for the search space, for each choice of τ.*
- *(2 points): Comments on whether these results differ from those found in Problem 4a, and on whether results from the choices of τ in the problem differ from each other.*

Problem 4d (10 bonus points)

Currently, most of the commonly used languages like Python, R, Matlab, etc., have pre-installed libraries for Gaussian processes. Use one library of your choice, maybe the language or environment you like the most, and compare the obtained results. Did you get the same parameters as in problem 4a? If not, why are they different? Elaborate on your answer.

Rubric: *Award each of the following according to whether the requirement is met:*

- *(2 points): Provides the optimal kernel parameters as found through the software library.*
- *(2 points): Provides details on the library used.*
- *(2 points): Comments on whether these results differ from those found in Problem 4a.*
- *(2 points): The results are the same, or, the results are different and an explanation is provided.*
- *(2 points): A plot showing the cost/optimization target is provided for the search space, or a plot comparing the predictions generated (in problem 5) if the results are different.*

Problem 5: Estimating unobserved flow data. (15 points)

In the previous problem, we have found a good set of parameters to model the sequence of velocities at one location as a Gaussian process. Recall that we have assumed the 100 *observations came at a rate of one every three days. We now select a smaller time step and interpolate the flow at some unobserved points.*

You are given flow information every three days. Pick some time stamps in-between each observation for which to estimate the flow. For example, you want flows every day, so there will be two unknown points between two observations. You could pick only one, or more than two. Make your choice and explain why.

Compute the conditional distribution (mean and covariance) at the time locations selected in Problem 4a. Use the same kernel parameters and the same location. For the initial estimate of the mean at the unknown time locations, you can use zero, use the average of all the observations, or take the average of the two closest observed points.

Plot your predictions, clearly showing:

- *The predicted means.*
- *The predicted standard deviation as a* 3σ *band (three standard deviations above and below the mean).*
- *The observed data points.*

Rubric: *Award each of the following according to whether the requirement is met:*

- *(2 points): Clearly states the choice of time-stamps at which to create predictions, and states why the choice was made.*
- *(2 points): Clearly states the method by which the prior means were chosen.*
- *(2 points): Provides a plot with a prediction for the horizontal velocity component at the chosen location.*
- *(2 points): Provides a plot with a prediction for the vertical velocity component at the chosen location.*
- *(3 points): Both plots have a labelled prediction for the mean for all of the time-stamps chosen.*
- *(3 points): Both plots have a labelled* 3σ *band around the predicted mean for all of the time-stamps chosen.*
- *(1 point): Both plots have the observations included.*

Problem 6: A longer time-scale simulation. (29 points)

In the previous problems, we learned to model the flow at one location as a Gaussian process. We now estimate the flow at any point in time at that particular location using the kernel function parameters. At a certain point in time, the flow can be computed as the realization of a multivariate Gaussian random variable with parameters given by the conditional distributions given the flow data. At this point, you will simulate a particle moving according to the flow data and use the estimates for times between the original timestamps. Ideally, one would have to estimate the parameters of the flow at every point in the map. However, running 504×555 *parameter selection models is too computationally intensive, so instead use the kernel parameters estimated in Problem 4b.*

Problem 6a (15 points)

Modify the simulator that you built in Problem 3.3 to use this new flow estimated flow information. Note that with this new change, you will be able to simulate the flow of particles for 300 days! Regarding data, originally, we have 100 measurements per point, now with this approach, let us say you use the estimates for two extra points to get one flow data per day, so in total, you should have at your disposal 300 descriptions of flow per location.

Repeat Problem 3b. This time you will be simulating flows for 300 days. This allows some debris to arrive on land. Where are some possible places along the coast where one could find debris? Pick some σ *of your choice and simulate the movement of particles with initial location sampled from the bivariate Gaussian. Evolve the location of the particles. Some times particles trajectories will terminate on the shore. Continue to keep track of such particles. These points are likely where you could find the debris.*

Provide a plot that includes your initial, final, and at least one intermediate state of your simulation. For the final state, clearly mark one location on land where you would search for debris. Also mark one location over the ocean where you would search for debris. Provide a brief justification for both choices.

Try at least one other value of σ *and recreate the three plots. Comment if your conclusions change.*

Rubric: *Award each of the following according to whether the requirement is met:*

- *(2 points): Provides a plot with the initial state of the simulation.*
- *(2 points): Provides a plot with an intermediate state of the simulation.*
- *(2 points): Provides a plot with the final state of the simulation.*
- *(2 points): Marks a location on the coast of the final state of the simulation where one should search for debris and provides a justification.*
- *(2 points): Marks a location over the ocean of the final state of the simulation where one should search for debris and provides a justification.*
- *(5 points): Provides three plots (initial, intermediate, final) for one other choice of* σ*, and comments on results (either to state why conclusions should change or why they should not).*

Problem 6b (14 points)

Thanks to your efforts, most parts of the (toy) plane were found, either inland or in the sea. As a final stage, you are tasked with locating three new monitoring stations on the coast. The purpose of these stations is to monitor general ocean debris. Using the tools you have build in this homework, propose the location of such three new stations. Simulate the trajectories of as many particles as you want, initialized at random locations uniformly distributed on the map. This is essentially a repeat of Problem 3 (a) with the new simulation; this time, remove particles that start on land so that they do not confuse your conclusions.

Many of the particles will end up on the coast. A good location for a monitoring station will be areas where many of such particles land on the coast.

Provide a plot that includes your initial, final, and at least one intermediate state of your simulation. For the final state, clearly mark where you would place your three monitoring stations. Provide a justification for why you chose these locations.

***Rubric:** Award each of the following according to whether the requirement is met:*

- *(2 points): Provides a plot with the initial state of the simulation, there should be no particles on land.*
- *(2 points): Provides a plot with an intermediate state of the simulation.*
- *(2 points): Provides a plot with the final state of the simulation.*
- *(4 points): Marks three locations on the final state of the simulation where monitoring stations should be placed.*
- *(4 points): Provides a convincing explanation for choosing these locations.*