AIMS CDT DATA, ESTIMATION AND INFERENCE LAB SHEET

October 8, 2018

1. The aim of this laboratory is to use Gaussian processes to fit and predict time series data derived from a weather sensor network.
2. Your demonstrator is ✉ Tim Rudner.
3. Course materials are available here.
4. To be submitted to ✉ Michael Osborne by *09:00 Monday 15th October*.

Data

Sotonmet makes weather sensor data from the Port of Southampton available to the public. It's useful for sailors and the port authorities, who need local and up-to-the-minute measurements of quantities like tide height and air temperature. The challenge is that the sensors, exposed to the elements, often fail to transmit readings, leading to missing data.

Our goal in this lab is to predict for *missing sensor measurements*. This data comes from a stormy period in 2007 with a number of patches of missing data. Point your browser here to grab the data. It should have columns

1. Update Date and Time (ISO)
2. Update Duration (ms)
3. Reading Date and Time (ISO)
4. Air pressure (mb)
5. Air temperature (C) – *A variable of interest*.
6. Tide height (m) – *Another variable of interest*.
7. Wind direction (deg)
8. Wind gust speed (kn)
9. Wind speed (kn)
10. True air temperature (C) – *Ground truth air temperature, against which you should compare your predictions.*
11. True tide height (m) – *Ground truth tide height, against which you should compare your predictions.*
12. Independent tide height prediction (m) – *These are some GP predictions prepared earlier for you to compare against, if you so choose.*
13. Independent tide height deviation (m) – *The standard deviation of the GP predictions above.*
14. Dependent tide height prediction (m) – *Another GP prediction built using three additional sensors not provided to you.*
15. Dependent tide height deviation (m) – *The standard deviation of the GP predictions above.*
16. Independent air temperature prediction (C) – *These are some GP predictions prepared earlier for you to compare against, if you so choose.*
17. Independent air temperature deviation (C) – *The standard deviation of the GP predictions above.*
18. Dependent air temperature prediction (C) – *Another GP prediction built using three additional sensors not provided to you.*

19. Dependent air temperature deviation (C) – *The standard deviation of the GP predictions above.*

Gaussian Processes

We'll cover Gaussian processes in the lectures; Mark Ebden also provides a nice primer.

I'd suggest that you start out with the exponentiated quadratic (also known as the squared exponential, RBF or Gaussian) covariance function. Once you get interested in trying more complicated covariance functions, you may wish to have a look at David Duvenaud's Kernel Cookbook.

You may find that conditioning errors begin to induce a deep sense of frustration and a questioning of life decisions. You can monitor the conditioning of your covariance matrix with the condition number, which you probably want to keep below (roughly) 10^{10}. To improve conditioning, you might wish to add in a small positive value (known as *jitter*) to the diagonal of your covariance matrix, just as you would if there was a bit more noise. You should also feel free to downsample the data, if desired, to ameliorate the problem.

Goals

Below is a non-exhaustive set of goals you may wish to pursue given this dataset. While reasonably ordered by priority, completing all goals would be ambitious: better to perform fuller analysis on a few problems than cursory analysis on all.

1. Load the data, defining the tide height readings to be y and the reading times to be t.
2. Write *your own* Gaussian process code to perform retrospective prediction for the missing readings. Hold covariance and mean function hyperparameters fixed to sensible values of your choice as a first step.
3. Compare against the ground truth tide heights using root-mean-square-error or the predictive log-likelihood, $\log p(\text{test data}|\text{training data})$.
4. Test some more sophisticated covariance and mean functions to try to improve performance.
5. Investigate alternate means of managing the hyperparameters, including maximum likelihood and maximum a-posteriori.
6. Produce code to allow for *sequential prediction:* that is, using only readings from prior to (and including) a time t to predict for the readings at t. This might seem a bit easy, as you have an observation at time t, but the noise in the observations means that it's still interesting to do.
7. Create plots that show such predictions for a fixed *lookahead* (the separation in time between the most recent reading and the time at which predictions are made in the sequential setting). That is, such plots should show predictions for $t+\text{lookahead}$ given data only prior to (and including) t, for all t.
8. Investigate the impact of increasing the lookahead. Results for this kind of experiment are plotted in Figure 1.
9. Repeat experiments with other readings available in the data, starting with air temperature.
10. Investigate the possibility of fusing multiple readings: e.g. using both air temperature and tide height within a single model to improve performance.

Assessment

CDT students must submit code and an informal two-page report (inclusive of plots) describing their results to ✉ mosb@robots.ox.ac.uk. Please *include '[AIMS CDT: DEI]' in the subject field of your email.* Material will be assessed on a satisfactory/non-satisfactory scale, according to the degree to which they demonstrate insight into probabilistic inference with Gaussian processes. We're also interested in seeing your experiments with things that didn't end up working out: accurate results are not the only marker of success.

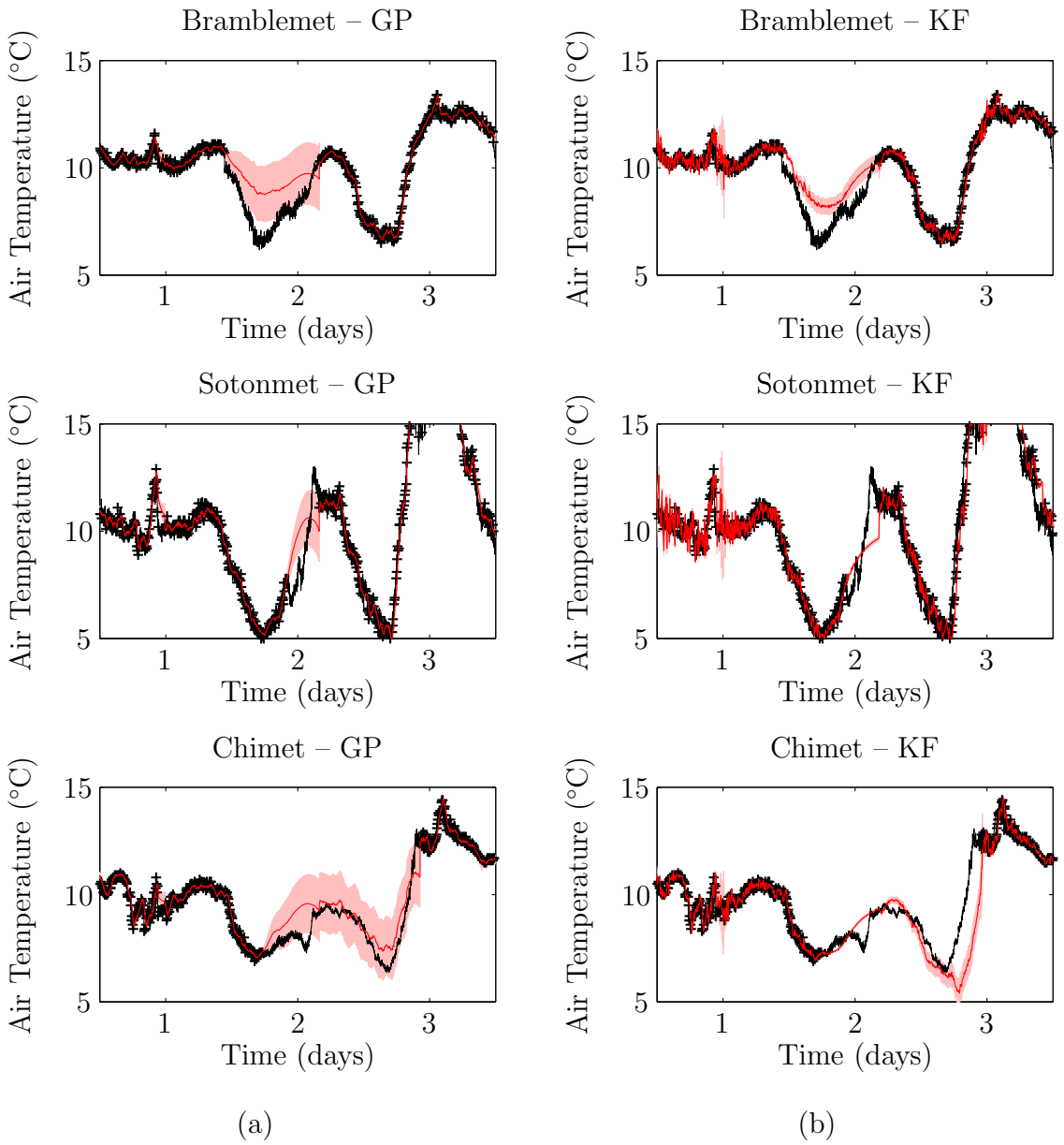

Figure 1: 5 minute lookahead prediction of tide height data for (a) a multi-output Gaussian process and (b) a Kalman filter.