STA 414/2104, Spring 2013 — Assignment #2

Due at the start of class on March 5. Please hand it in on 8 1/2 by 11 inch paper, stapled in the upper left, with no other packaging.

This assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. In particular, you should not leave any discussion with someone else with any written notes (either on paper or in electronic form).

In this assignment, you will apply a set of R functions for fitting Gaussian process models that I have provided to the task of modeling a time series of the number of deaths each day in Toronto over a period of four years, along with the average temperature each day. One thing this will illustrate is how Gaussian process models can use a variety of covariance functions, which can be chosen to suit the characteristics of the problem being solved. To keep the effort and computation time needed for this assignment from being too large, I have provided only one covariate (temperature), but many other covariates (eg, humidity, levels of air pollution) could be handled in this modeling framework.

The data file available from the course web page gives the number of deaths in Toronto for each day from January 1, 1992 to December 31, 1995, along with the average temperature that day (in degrees Celsius). The first line is a header with the variable names, and each later line begins with the date as a "row name". The data should be read with read.table using the head=TRUE option, and then converted to a matrix with as.matrix (since this speeds up access). You will use the first three years (that is, the first $366 + 365 + 365$ days) as training data, and predict for the last year (365 days). You should start by looking at plots of the number of deaths and the temperature versus time, and versus each other, to see what general characteristics the data have. (But don't hand in these plots.)

Since most deaths occur one at a time (rather than in groups, such as from car crashes with multiple fatalities), we might expect the number of deaths to have close to a Poisson distribution, with a mean that depends on covariates such as temperature and time of year. However, Gaussian processes most easily model responses that have a normal distribution given the covariates. Fortunately, there is a trick for treating Poisson distributed data as normally distributed — just take the square root. This works well as long as the observed numbers are always fairly large, as is the case with this data (the minimum number of deaths in a day is 23). We can then consider the data to be real-valued without losing much information. Seen this way, it is easy to show that the square root of the number of deaths will have approximately a normal distribution with a standard deviation of 1/2, regardless of what the mean given the covariates is. You should therefore transform the number of deaths by taking the square root, and treat this as the response variable to be modeled.

Both the data and common sense tell us that the number of deaths might depend on both the time of year and the temperature, and that the average number of deaths in a day might temporarily go up or down for a few days or weeks (for instance, if there is a major week-long sporting event, or an epidemic of flu, or a strike by garbage collectors). We will try to account for these possible effects using an additive model, expressed using a covariance function that is the sum of several terms. Each of these terms (except the constant term) will have a hyperparameter controlling its magnitude, since it is unclear *a priori* how important each will be.

The first term is a constant, which you should set to 10^2. This models the fact that we don't know the exact overall mean of the response variable (though from the data it is clearly somewhere around 7). Since we know that the response variable is never negative, we could do better than model its overall average level as having a prior mean of zero and standard deviation 10, but since we are fitting to 1096 data points, this would make very little difference in practice, so we won't bother.

The second term will model the effect of the season. You can create a variable with values 1, 2, 3, ... for successive observations, and then divide by 365.25 to get a variable that is the number of years from the start of the series. From this, you can compute two variables by taking the sine and cosine of this "year" variable times 2π. Call these variables s and c. The following term in the covariance function will be large for observations, i and j, that are at nearly the same time of year (but possibly in different years), and small for observations that are at very different times of year:

$$\eta_2^2 \exp(-((s_i - s_j)^2 + (c_i - c_j)^2)/2^2)$$

The division by 2^2 above is based on my common-sense judgement of how smooth seasonal effects are likely to be. The hyperparameter η_2 controls the magnitude of this seasonal effect.

The third and fourth terms will model temporary changes in the mean of the response variable. Since these might occur on time scales of around a week, or on somewhat longer time scales of a month or so, two terms are used, differing only in their length scales, both depending on the "year" covariate, r:

$$\eta_3^2 \exp(-|r_i - r_j|/0.1) \quad \text{and} \quad \eta_4^2 \exp(-|r_i - r_j|/0.02)$$

The hyperparameters η_3 and η_4 control the magnitudes of these temporary variations. The use of absolute differences rather than squared differences above produces functions that are not differentiable, which seems appropriate since such temporary influences may be due to erratic factors.

The fifth term in the covariance function should model the effect of temperature on the square root of the number of deaths. You should try two forms for this term. The first form is as follows:

$$\eta_5^2 \exp(-(t_i - t_j)^2/20^2)$$

where t_i and t_j are the temperatures for two observations. The scale factor of 20 reflects my common-sense judgement that a difference in temperature of 20 degrees Celsius could be large enough to have a significant effect. Alternatively, we might use a second form, that reflects the possibility that only extremes of temperature matter. This idea can be captured by looking at the cube of temperature, which compresses values near zero Celsius (a reasonable centre point) while magnifying extreme temperatures. This form is therefore as follows:

$$\eta_5^2 \exp(-(t_i^3 - t_j^3)^2/20000^2)$$

The scale factor of 20000 is again chosen on the basis of common sense knowledge of how big a temperature change might be needed to have a noticeable effect on death.

A noise variance term is also needed. In view of the argument above regarding the square roots of Poisson variables, we might fix the noise variance to 0.5^2, but for this assignment you should let it be a hyperparameter, η_1^2. You can then see whether η_1 is indeed estimated to be close to 0.5.

In a fully Bayesian approach, the hyperparameters $\eta_1, \ldots, \eta_5$ would be given prior distributions, and predictions would be based on the average over their posterior distributions. In this assignment, however, you will find the maximum marginal likelihood estimates for these hyperparameters. In a real application, some or all of the scale factors fixed above using "common sense" would also be estimated, since it's hard to choose the best values just from prior knowledge, but that would have increased the computation time required for the assignment. (In a real application, one would probably not optimize using the "nlm" function without provided it with gradient information, but I haven't covered such better optimization methods in class.)

You should find estimates for the hyperparameters using the R functions that I provide for this assignment on the course web page. Note that these are slightly different from the functions demonstrated in class. In particular, since all the hyperparameters are squared before use, I have eliminated the provision designed to stop them from becoming negative. Instead, the gp_find_hypers function just returns the absolute value of the estimates found.

You should find values for the hyperparameters using gp_find_hypers based on the first three years of data. You should then predict values for all four years of data. Only for the last year will these be real predictions (based however on knowing the actual temperatures during that year). For the first three years, these "predictions" will be for hypothetical different years that happen to have the same temperatures; this is still of interest in showing what the model thinks is the systematic part of the variation in the number of deaths.

Since there are two forms for the fifth term in the covariance function, you will actually estimate hyperparameters and make predictions twice, once for each form. The gp_find_hypers function provided prints the log marginal likelihood at the optimum found, which you can use to assess which of these two forms of the fifth term in the covariance function is more probable given the data.

You will need to choose suitable initial estimates for the hyperparameters, for gp_find_hypers to start with. You may need to try several values, until you get reasonable results. Passing an argument of print.level=2 to gp_find_hypers (which it passes on to nlm) will cause nlm to print what it is doing, which might help when choosing starting values.

You should hand in the R script you used to do all this, the estimates and marginal likelihoods you found, and a discussion of the results. For your discussion, you may want to output more information, or produce various plots, in order to gain as much insight as you can into what is going on with these models and with this data.