Software for Fitting and Using Gaussian Processes

This assignment guides you through use of some `R` packages for fitting GPs, diagnostics, prediction, and sensitivity analysis. Hence, it builds on your expertise in the underlying math (Assignment 1) and writing your own code (Assignment 2). The packages we illustrate here can work with more complex problems, and familiarity with them will allow you to carry out many of the methods you have seen on the course.

All parts of this assignment are based on a sensitivity analysis of an ocean general circulation model applied to the North Atlantic Ocean (Gough and Welch, 1994). The computer model had seven inputs, and six outputs were analyzed.

To simplify this assignment, we will only consider one output called DOWN, the number of downwelling points. Gough and Welch found that statistical modelling of this output was less accurate than for the others; hence, it provides a challenge for us. Let's see if we can do better.

The input and output data are at the course web site as `x.txt` (the 7 inputs) and `y.txt` (the DOWN output). The data have been cleaned: NAs from failed computer model runs have been removed. Also the inputs have all been rescaled to $[0, 1]$, because some of the packages expect this. The data can be read by `read.table` in `R`, but note that some of the packages used here expect data in the form of a `matrix`, so be sure to convert the `data.frame` from `read.table`.

1. First, we fit a GP using three packages: `GPFit`, `mlegp`, and `DiceKriging`. You will need to install them.

   We start with a GP model with a constant mean and the squared-exponential (Gaussian) correlation function. Fit the GP using all three packages. (If a package does not allow the squared-exponential, use the power-exponential; it makes no practical difference here.)

   (a) Make a table showing estimates of the following parameters for all three packages: the constant mean, the correlation parameters, and σ^2. Note that the packages parameterize the correlation parameters in different ways! So convert your estimates to the θ scale we used in the September 16 class. That way you can easily compare the fits from the packages and identify any that need persistence on your part to find the MLE, i.e, you will sometimes have to change default settings in the MLE optimizer to get the packages to work here. Comment on the three fits.

   (b) Do the estimates of the mean and σ^2 make sense?

   (c) Make a table showing the magnitudes of the (leave-one-out) cross-validation errors. Report the root mean square error and the absolute error. Not all the packages provide these quantities easily; if not, don't bother.

2. For this question and the remainder of the assignment, we will use `DiceKriging` only. Continue using the model with a constant mean and squared-exponential correlation.

   (a) Make diagnostic plots like those on Slide 16 of the September 16 class. Comment on the prediction accuracy of the GP model. Look at the plot of the standardized cross-validation residuals versus cross-validation predictions. Does this plot suggest any problems with the statistical model? Why or why not?

(b) None of the packages easily generates main-effect and joint-effect plots like the slides shown in the Visualization class of September 30. But `DiceKriging` can generate related numerical sensitivity summaries: see Section 4.5 of Roustant et al. (2012) for explanation of how to interface with the `sensitivity` package. (Actually, any GP modelling packages with a `predict` function can use this method.)

Compute sensitivity measures for the seven inputs. How well do the measures agree with those provided by Gough and Welch?

3. Simple plotting of the data shows very strong trend with respect to the input BACK. We will now try to improve the statistical model by changing the mean function to

$$\beta_0 + \beta_1 \text{BACK}.$$

The correlation-function family will still be the squared-exponential, and we will continue to use `DiceKriging`.

Re-fit the GP. Compare with the fit of question 1 by answering the following questions.

(a) Compare the log likelihoods found. (There are many reasons why the asymptotic likelihood theory is suspect here, but use it as an informal guide to a "big" change.)

(b) What is the value of $\hat{\beta}$ for BACK? Explain whether this makes sense.

(c) Look at the estimates of σ^2 and the correlation parameters. Have they changed substantially? Speculate on why or why not.

(d) Is there evidence that prediction accuracy has improved?

(e) Have the sensitivity measures changed?

4. Summarize all your findings. Does anything you found change the scientific conclusions of Gough and Welch?

5. Is there anything else you noticed that's worth reporting? (Don't worry about leaving this blank.)

References

Gough, W. A. and Welch, W. J. (1994), "Parameter Space Exploration of an Ocean General Circulation Model Using an Isopycnal Mixing Parameterization," *Journal of Marine Research*, 52, 773–796.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012), "DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization," *Journal of Statistical Software*, 51, 1–55.