Homework 4: Prediction over Time and Space

36-467, Fall 2018

Due at 6 pm on Thursday, 1 October 2020

Abstract

AGENDA: Practice with linear prediction of time series, especially over space. Practice with linear prediction of spatial data, especially over time.

The data set `wind` in the library `gstat` was introduced in Lecture 5. It consists of daily measurement of wind speed at twelve different locations across Ireland from 1961 through 1978. It is accompanied by the data set `wind.loc` which gives the full names and map coordinates for the stations. $X(r,t)$ will stand for the wind speed at station r on day t.

1. *Yesterday is like today*

   (a) (1) What is $\mathrm{Cov}\left[X(\mathrm{DUB},t), X(\mathrm{DUB},t-1)\right]$?

   (b) (1) What is $\mathrm{Var}\left[X(\mathrm{DUB},t)\right]$?

   (c) (3) What are the slope and intercept for the optimal linear predictor of $X(\mathrm{DUB},t)$ from $X(\mathrm{DUB},t-1)$? What should the variance of the prediction errors be? *Hint:* Lecture 7.

   (d) (5) Use `lm` to linearly regress $X(\mathrm{DUB},t)$ on $X(\mathrm{DUB},t-1)$. What are the coefficients it reports? What is the connection between this and what you did in problem 1c? *Hint:* What would `lm(wind$DUB[2:11] ~ wind$DUB[1:10])` do?

   (e) (5) For each day in the data set, calculate the actual prediction error for linearly predicting $X(\mathrm{DUB},t)$ from $X(\mathrm{DUB},t-1)$. What is the mean error? What is the variance of the errors? How well do these match what you calculated in problem 1c?

   (f) (3) Suppose we want to retrodict $X(\mathrm{DUB},t-1)$ from $X(\mathrm{DUB},t)$. What are the optimal linear coefficients?

2. *Yesterday is like tomorrow*

   (a) (3) What is the optimal slope for linearly predicting $X(\mathrm{DUB},t)$ from $X(\mathrm{DUB},t-2)$?

   (b) (5) Suppose we want to linearly predict $X(\mathrm{DUB},t)$ using both $X(\mathrm{DUB},t-1)$ and $X(\mathrm{DUB},t-2)$. Which variances and covariances we need to know in order to find the coefficients? What are all those variances and covariances? (You may want to report them in a table.)

(c) (5) What are the optimal coefficients for linearly predicting $X(\text{DUB}, t)$ using both $X(\text{DUB}, t-1)$ and $X(\text{DUB}, t-2)$? (There should be an intercept and two slopes.)

(d) (3) The relative difference in the two slopes from the last problem should be much larger than the relative difference in the corresponding covariances with $X(\text{DUB}, t)$, or the relative difference in the slopes in the univariate linear predictors. Why?

3. *Correlations over time*

(a) (3) Plot the autocorrelation function for $X(\text{DUB}, t)$ out to a lag of 800 days. Describe the shape of the function. Can you explain the location of the peaks and troughs of the function?

(b) (5) What is the slope for linearly predicting $X(\text{DUB}, t)$ from $X(\text{DUB}, t-365)$? What are the slopes for linearly predicting $X(\text{DUB}, t)$ from $X(\text{DUB}, t-1)$ and $X(\text{DUB}, t-365)$?

(c) (3) Plot the cross-correlation function between Dublin and Shannon out to a lag of 800. Describe the shape of the function. Can you explain the location of the peaks and troughs of the function?

4. *Our first spatial regression* Suppose we want to predict $X(\text{DUB}, t)$ using all the $X(r, t)$, for all $r \neq \text{DUB}$.

(a) (4) What are all the variances and covariances needed to find the optimal linear predictor? (Report your answer in the form of a table.)

(b) (4) What are the coefficients of the optimal linear predictor? (There should be twelve of them, so think carefully about how to report them.)

(c) (3) What, theoretically, should the variance of the prediction errors be?

(d) (3) For each data point, calculate the prediction error. What is the mean prediction error? The variance of the prediction errors?

5. *Our first spatiotemporal regression* Suppose we want to predict $X(\text{DUB}, t)$ using all the $X(r, t-1)$, for all $r \neq \text{DUB}$.

(a) (3) What are all the variances and covariances needed to find the optimal linear predictor? (Report your answer in the form of a table.)

(b) (3) What are the coefficients of the optimal linear predictor? How do they compare to the coefficients that you found in problem 4b

(c) (3) What, theoretically, should the variance of the prediction errors be?

(d) (3) For each data point, calculate the prediction error. What is the mean prediction error? The variance of the prediction errors?

6. *In- and out- of sample comparisons*

(a) (1) Make a table showing the root-mean-squared prediction errors for the linear predictors you fitted in problems 1c, 2c, 4b and 5b. Which predictor seems to do best?

(b) (4) Re-estimate the coefficients of each of those four models, using only data from the first nine years of the data set (i.e., 1961–1969). What are the old and new coefficients for each model?

(c) (3) Using the coefficients you found in problem 6b, calculate a prediction for each of the four models for January 1st, 1970. What are those predictions, and what are the prediction errors?

(d) (5) For each of the four models, calculate a prediction for each day in 1970–1978. What are the mean prediction errors of the four models? What are the root mean square prediction errors?

(e) (5) Why do the numbers in problem 6a differ from those in 6d? Which set of numbers gives a better idea of how well the different models predict?

7. (1) How much time did you spend on this problem set?

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. All plots and tables are generated using code embedded in the R Markdown and automatically re-calculated from the data. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant, without dangling or useless commands. All parts of all problems are answered with coherent sentences, and raw computer code or output are only shown when explicitly asked for.