Homework 1: A Cherry-Blossom-Viewing Party

36-740/620, Fall 2026

Due at 6 pm on Thursday, 3 September 2026

> AGENDA: Practice with estimating trends by smoothing; getting a feel for the eigendecomposition of influence matrices and its implications; a glimpse of another de-trending technology.

Cherry trees flower in the spring, and the opening and blooming of the blossoms is very sensitive to temperature — full bloom is advanced by warm weather and held back by cold. Because of the important role of the cherry blossom in Japanese culture over the last 1200 years, it is has been possible to work out the date of full blossom at Kyoto going back to the early 800s AD[1]. Our first homework will analyze this data to look for trends.

The data file is at `http://www.stat.cmu.edu/~cshalizi/dst/26/data/kyoto-2026.csv`. It's the data collected by Prof. Yasuyuki Aono and collaborators, slightly reformatted to make it easier to work with in R. The three columns record (1) the year (in the Common Era [CE], a.k.a. AD), (2) the day of the year of full flowering, i.e., counting January 1 as day 1 and December 31st as day 365 or 366; (3) the date in the contemporary (Gregorian) calendar, in MMDD format (so March 15 would be 0315).[2]

1. Load the data and plot the day of flowering against the year.
   (a) (2) Add a horizontal line to the plot showing the mean day of flowering.
   (b) (3) Comment on any patterns you see in your plot.
2. *Moving average smoothing* Calculate a centered 5-year moving average of the day-of-flowering time series. That is, for each year t, calculate

$$\overline{x}_t = \frac{1}{5}\sum_{k=-2}^{k=2} x_{t+k}$$

   (See slides for Lecture 2, or Recipe 14.10 in *The R Cookbook*, for one way to do this.)

[1] There's lots more about all this in the slides for the first lecture.

[2] The full data set also includes information about data sources and the historical evidence (e.g., diary descriptions of cherry-blossom viewing parties vs. newspapers), which would be valuable for a more detailed examination of this data.

(a) (3) Carefully explain how your code handles NA values. *Hint:* Look at what happens around 1945. (Why 1945?)

(b) (5) Add 3 curves to your final plot from Problem 1: the (centered) 5-year moving average, an 11-year moving average, and a 51-year moving average. Make sure that the moving-average curves are clearly distinguished visually from the data, from the horizontal level you plotted before, and from each other. Comment on the plot.

(c) (5) Find the leave-one-out cross-validated MSE for the three moving averages; report the LOOCV scores in a table. Is the best-scoring smoother *visually* better than the other two? *Hint:* p. 16 of the handout / slide 44 of lecture 2.

3. *What do splines like?*

(a) (5) Use spline smoothing[3] to find the trend of flowering day vs. year, and then add it to the plot from Problem 1. Comment on the plot. How is it similar to the moving averages, and how does it differ?

(b) (3) Most spline fitting routines, including R's `smooth.spline()`, unfortunately do not store the smoother matrix $\mathbf{w}$. We can however recover the whole of $\mathbf{w}$ by re-fitting the spline on artificial data. The following code (also on the class website) will do it:

```
smoother.matrix <- function(a.spline, x) {
  n <- length(x)
  w <- matrix(0, nrow=n, ncol=n)
  for (i in 1:n) {
    y <- rep_len(0, n)  # Equivalent to rep(0, length.out=n) but faster
    y[i] <- 1
    w[,i] <- fitted(smooth.spline(x, y, lambda=a.spline$lambda))
  }
  return(w)
}
```

Pick any square matrix $\mathbf{w}$, and let $\mathbf{e}_j$ be the $n \times 1$ matrix which is 0 everywhere, except in row j, where it is 1. Explain why $\mathbf{w}\mathbf{e}_j$ gives the j^{th} column of $\mathbf{w}$. Explain how this relates to the provided code.

(c) (5) Run the code on the spline you fit to the whole data, and a suitable choice of `x`. (What's your choice, and why is it suitable?) Make a plot of all of the eigenvalues (not eigenvectors!) of the smoother matrix, in order. How many of them are > 0.95? How many are > 0.1? How many are > 0.01? Are any of them 0, to within numerical precision?

(d) (4) Following the examples in the slides for Lecture 2, make a plot of the first six eigenvectors of the smoother matrix. Describe the kinds of patterns which these eigenvectors capture.

[3] If you're using R, you might find the examples in the slides and notes from Lecture 2 helpful.

(e) (3) Make a similar plot for the last six eigenvectors. Describe the kinds of patterns they capture.

(f) (5) What sorts of patterns will show up in the fitted values of the spline (= the estimate of the trend)? What sorts of patterns will show up in its residuals (= the estimate of the fluctuations)?

(g) (2) The example from Lecture 2 for the previous problem plots the entries in the eigenvectors against their position ("index") in the vector. Here, different positions correspond to different calendar years. Redo the plot from problem 3d so the eigenvectors are plotted against the year. Does this change your description of any of the patterns?

(h) (6) Make a plot of the smoother matrix. The x and y axes should be years, and the z axis should be the corresponding entry in $\mathbf{w}$, shown using color, contour lines, or a 3D plot. (The solutions will use the `contour` function in R, but you don't have to.) Explain how this plot relates to the idea that the spline is doing a kind of local averaging.

4. *Yule-Slutsky* In this problem, assume that $X(t) = \mu(t) + \epsilon(t)$, that $\hat{\mu}(t_i) = \sum_{j=1}^{n} w_{ij} X(t_j)$, and that $\hat{\epsilon}(t) = X(t) - \hat{\mu}(t)$. As proved in the notes, $\mathrm{Var}[\hat{\epsilon}] = (\mathbf{I} - \mathbf{w})\mathrm{Var}[\epsilon](\mathbf{I} - \mathbf{w})^T$.

   (a) (5) Suppose that $\mathrm{Var}[\epsilon] = \sigma^2 \mathbf{I}$. Make a plot of $\frac{1}{\sigma^2}\mathrm{Var}[\hat{\epsilon}]$ for the spline smoother, using the $\mathbf{w}$ matrix you already found. As in Problem 3h, the x and y axis should be years, and the z axis should be the (scaled) covariance, shown using color, contour lines, height, etc., as you prefer. Describe any patterns you see.

   (b) (6) Now suppose that $\mathrm{Cov}[\epsilon_i, \epsilon_j] = \sigma^2 0.9^{|t_i - t_j|}$. Calculate and plot $\frac{1}{\sigma^2}\mathrm{Var}[\hat{\epsilon}]$, using the same smoothing matrix for the spline you already found. Describe any patterns you see, contrasting it with both the plot in Problem 4a, and the shape of the $\mathrm{Var}[\epsilon]$ matrix.

5. *Detrending by differencing* We have focused on detrending by smoothing, where we first estimate the trend, and then subtract it off. An alternative procedure is to remove trends by taking *differences* between nearby observations. This problem explores how this works, in a situation where we have one coordinate, so our data can be written $X(t) = \mu(t) + \epsilon(t)$. Assume t is discrete, so it can be $1, 2, \ldots$, and that there are no missing values.

   Define $\Delta(t)$ as $X(t) - X(t-1)$.

   (a) (4) Write $\Delta(t)$ in terms of the μ's and ϵ's.

   (b) (4) Find the expectation and variance of $\Delta(t)$ in terms of μ, and the variance and covariance of ϵ.

   (c) (5) Explain why $\Delta(t)$ can be said to be "detrended", if μ changes slowly.

   (d) (4) Find an expression for $\mathrm{Cov}[\Delta(t), \Delta(t+1)]$.

(e) (5) *Yule-Slutsky again* Assume $\mathrm{Var}[\epsilon(t)] = \sigma^2$, $\mathrm{Cov}[\epsilon(t), \epsilon(s)] = 0$ (unless $t = s$). What is $\mathrm{Cov}[\Delta(t), \Delta(t+1)]$?

(f) (5) *Back to Kyoto* Calculate the difference between each year's flowering-day and the previous year's. (If the previous year's value is NA, the difference should be NA — why?) Plot the difference variable against time. Comment on any patterns you see.

6. (1) How long 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. 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. Any code displayed in the write-up is relevant, without dangling or unused fragments, and is indented, commented, and uses meaningful names. All parts of all problems are answered with coherent sentences, and raw computer code or output are only shown when explicitly asked for.