

Homework 1

Due: Tuesday, February 27

Answer at least 4 questions. Submit any Matlab code you used (you may print and submit an HTML demonstration, but do not need to).

1. B-splines: to avoid worrying about boundary effects, consider knots placed at integer values on the whole real line $\tau_i = i$ for $i \in \mathbb{N}$.

Recall that B-splines are defined recursively as follows

$$B_{0,i}(t) = I(\tau_i \leq t < \tau_{i+1})$$

$$B_{p,i}(t) = \frac{t - \tau_i}{\tau_{i+p} - \tau_i} B_{p-1,i}(t) + \frac{\tau_{i+p+1} - t}{\tau_{i+p+1} - \tau_{i+1}} B_{p-1,i+1}(t)$$

Demonstrate the following:

- (a) $B_{p,i}(t) = 0$ for $t < \tau_i$ or $t > \tau_{i+p+1}$.
- (b) $B_{p,i}(t) \geq 0 \forall t$.
- (c) For t not a knot, $B_{p,i}(t)$ is a polynomial of degree p .
- (d) $B_{p,i}(t)$ is continuous and has $p-1$ continuous derivatives when $p \geq 1$.
- (e) $\sum_i B_{p,i}(t) = 1$.

Hint: use induction. For 1d first show

$$\frac{d}{dt} B_{i,p}(t) = \frac{p}{\tau_{i+p} - \tau_i} B_{i,p-1}(t) - \frac{p}{\tau_{i+p+1} - \tau_{i+1}} B_{i+1,p-1}(t) \quad (1)$$

2. Integration.

- (a) When we want to analyze the bias of an estimator, two different measures usually come up. The design bias is

$$\frac{1}{n} \sum_{i=1}^n \left(E\hat{f}(t_i) - f(t_i) \right)^2$$

while the integrated bias may be written as

$$\int_0^1 \left(E\hat{f}(t) - f(t) \right)^2 dt$$

Show that for $f \in W^1$ with $t_i = (2i - 1)/2n$, these two approximate each other to $o(n^{-1})$.

- (b) Suppose that the basis functions $\{\phi_i(t)\}_{i=1}^K$ are ortho-normal with respect to the design $t_i = (2i - 1)/2n$. That is

$$\frac{1}{n} \sum_{i=1}^n \phi_j(t_i) \phi_k(t_i) = I(j = k)$$

Demonstrate that the variance of the least squares estimate for $\hat{f}_K(t) = \sum_{i=1}^K b_i \phi_i(t)$ is $o(K/n)$.

Bonus: Show this is true whenever the ϕ_i are linearly independent with respect to the design.

- (c) Hence show that the optimal choice for K for minimizing the risk:

$$\frac{1}{n} \sum_{i=1}^n E(\hat{f}_K(t_i) - f(t_i))^2$$

depends only on the ability of the ϕ_i to approximate f in the \mathcal{L}^2 norm.

3. Estimating error variance. Let

$$y_{t_i} = f(t_i) + \epsilon_i$$

where $t_i = (2i - 1)/2n$. Assume that $f \in W^2$ and the ϵ_i have mean zero, finite variance σ^2 and finite fourth moment κ .

Estimating variance by

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(t_i))^2$$

may be biased when we don't know the degrees of freedom of \hat{f} . A simple regression-free estimate of the variance is

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_{i=2}^n (y_i - y_{i-1})^2$$

Assume that $f \in W^2$. What is the bias of this estimate? How could it be improved?

Show that $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$ tends to a non-degenerate distribution with mean zero as $n \rightarrow \infty$.

4. Montreal weather data. These data are available from the montreal.mat file on the class web-site. The data represent daily temperatures in Montreal averaged over 40 years.

- (a) Smooth the data with a Fourier basis using an GCV to select the number of basis functions. Does the result look reasonable?
 - (b) Perform the same procedure using equi-spaced cubic B-splines and using GCV to select the number of knots. Plot the estimated MSE curves for each.
 - (c) Calculate confidence intervals for your best estimates at the design points. How do they compare?
5. Smoothing operators. Perform the same calculations as above, but with second-derivative and harmonic acceleration penalties. (You can find harmonic acceleration penalties implemented in the weather example from the FDA package). Use a B-spline basis with 365 knots and let λ vary over integers on the log scale. How do these compare with the results above?
6. Miscellaneous Montreal experiments
- (a) There is a small "bump" in temperature around day 25 (this is the "January thaw" that locals talk about). Design a probe to test its reality. Are we being duped?
 - (b) Try smoothing the data with polynomial terms $1, t, t^2, t^3, \dots$. Stop at the minimum of GCV. How many terms do you need? What does the estimate do outside the interval $[0, 365]$?
7. Simulation studies. Incorporating model selection into estimates of variance is active area of research. How important is this effect?

To test this, design a response function on $[0, 1]$. It should have at least one peak and one valley, have non-zero derivative at the end points and may not be represented by a linear combination of B-splines and trigonometric functions. Include at least one discontinuity.

Set up a simulation to explore the effect with the following parameters

- You should explore the effect of the amount of data using designs with 20, 40, 60, 80 and 100 observations.
- For each design, generate data with a signal to noise ratios 0.25, 0.5, 1, 2 and 4. (The signal to noise ratio is the square integral of the response function (approximate) divided by the variance of the errors).
- Use a Fourier basis and use GCV to select the number of basis functions for each of 1000 simulations for each design and noise level.
- Calculate an estimate of the error variance by the usual formula

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum (y_i - \hat{f}(t_i))^2$$

where p is the number of basis functions selected. Calculate the variance estimate given in Question 3 as well (the function 'diff' may be useful).

For each simulation and each design and each noise level, record two estimates of variance and the number of basis functions selected. Report the bias, relative to the variance, of your estimates of variance. Report the mean number of bases selected by GCV; how does it scale with the size of your design?