Text: Chapter 5

Some disadvantages of basis expansions

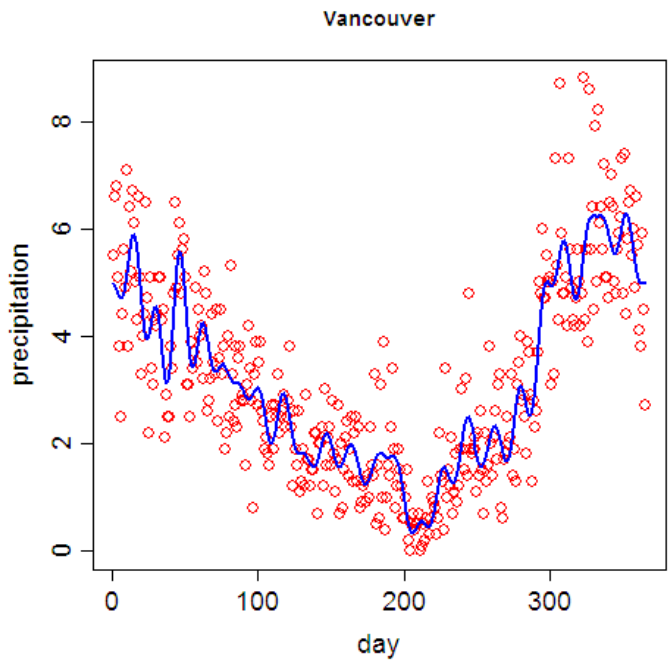
- Discrete choice of number of basis functions $\Rightarrow$ additional variability.
- Not necessarily easy to tailor to problems at hand

What do we mean by smoothness?

Some things are fairly clearly smooth:

- constants
- straight lines

What we really want to do is eliminate small "wiggles" in the data.



The D Operator

We use the notation that for a function $f(t)$,

$$Df(t) = \frac{d}{dt} f(t)$$

We can also define further derivatives in terms of powers of D:

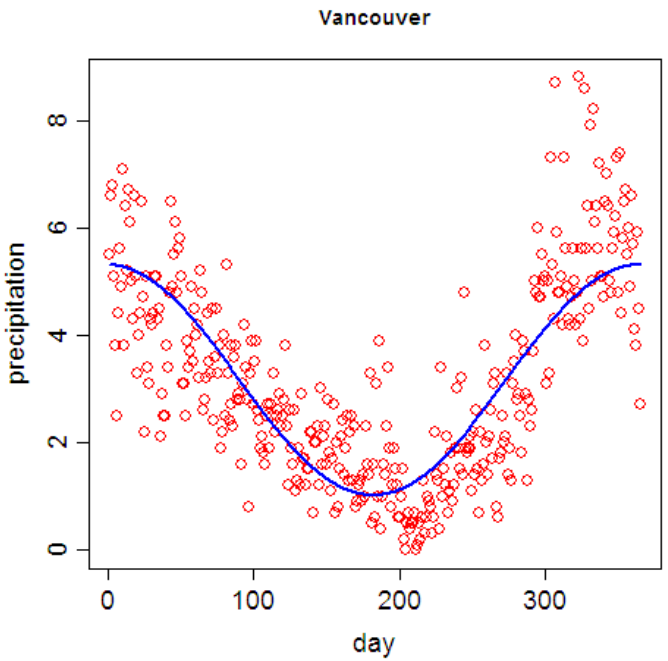
$$D^2 f(t) = \frac{d^2}{dt^2} f(t), \ldots, D^k f(t) = \frac{d^k}{dt^k} f(t), \ldots$$

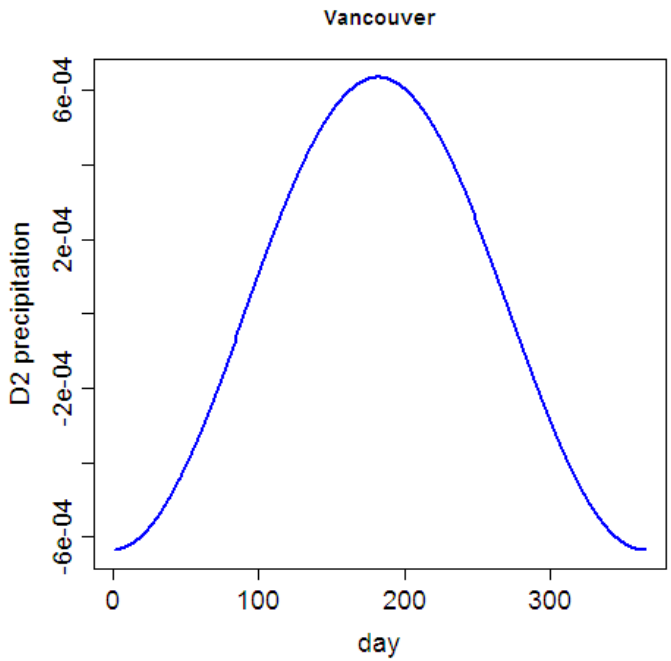
- $Df(t)$ is the instantaneous *slope* of $f(t)$; $D^2 f(t)$ is its *curvature*.
- We measure the size of the curvature for all of f by

$$J_2(f) = \int \left[D^2 f(t)\right]^2 dt$$

Curvature of Vancouver Precipitation

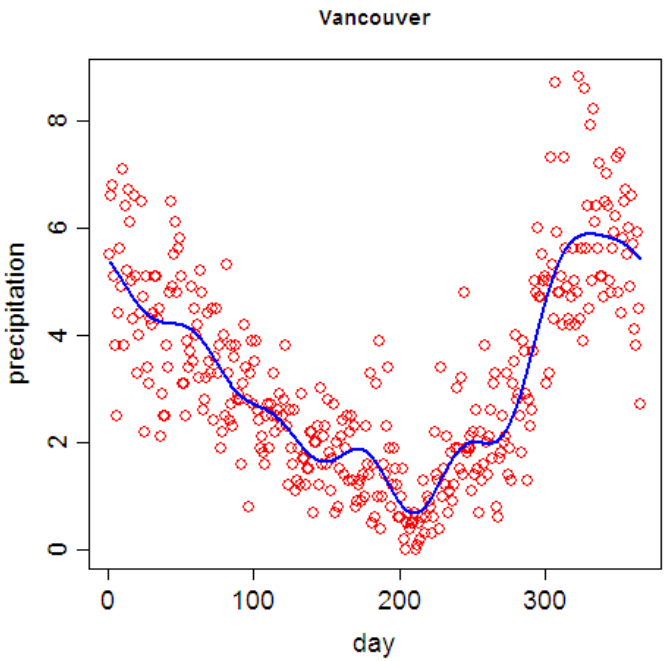
3 Fourier bases:

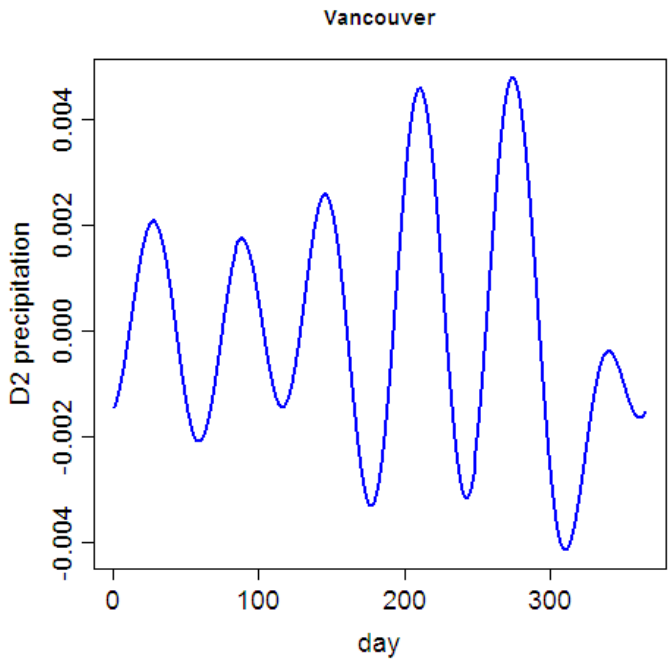




Curvature of Vancouver Precipitation

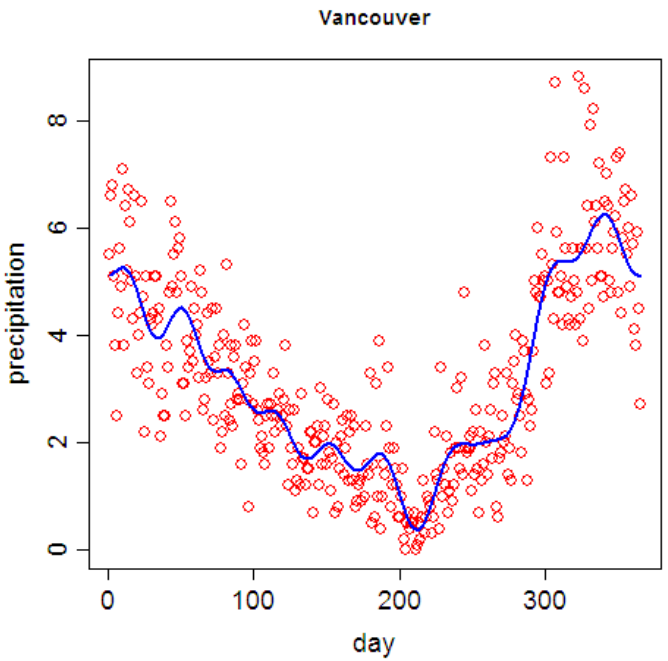
13 Fourier bases:

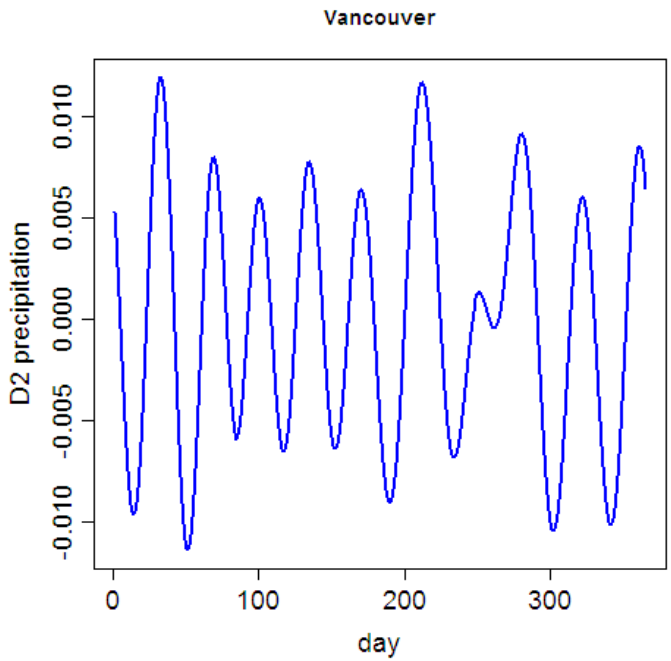




Curvature of Vancouver Precipitation

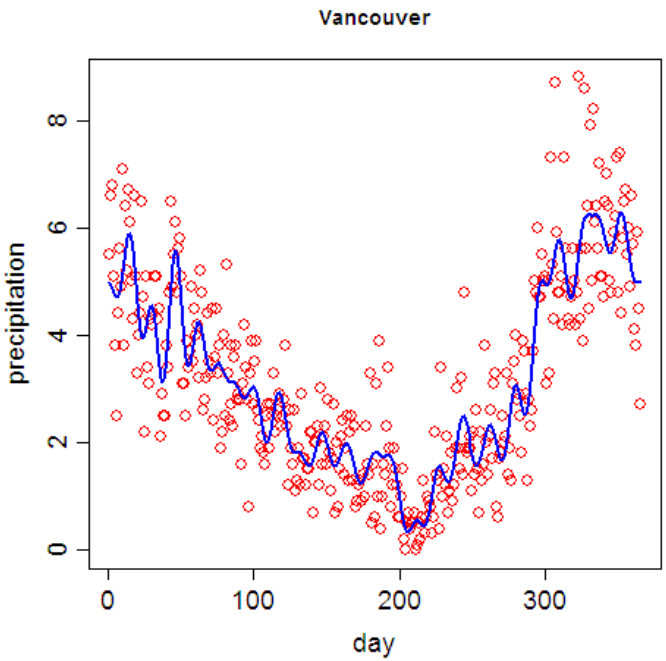
25 Fourier bases:

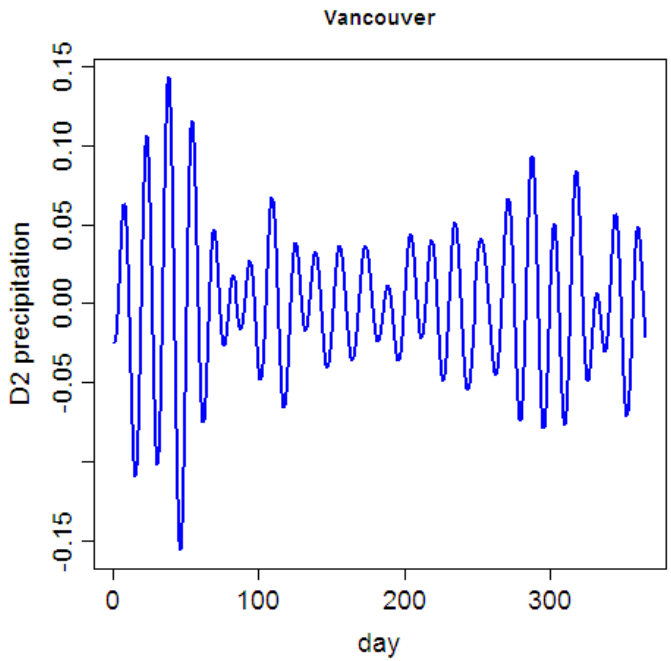




Curvature of Vancouver Precipitation

53 Fourier bases:





The Roughness of Derivatives

- Sometimes we may want to examine a derivative of $f(t)$, say $D^2 f(t)$.
- We then should consider the roughness of that derivative

$$D^2 \left[D^2 f(t) \right] = D^4 f(t)$$

- This means we measure

$$J_4[f] = \int \left[D^4 f(t) \right]^2 dt$$

- And the roughness of the mth derivative is

$$J_{m+2}[f] = \int \left[D^{m+2} f(t) \right]^2 dt$$

Penalized Squared Error

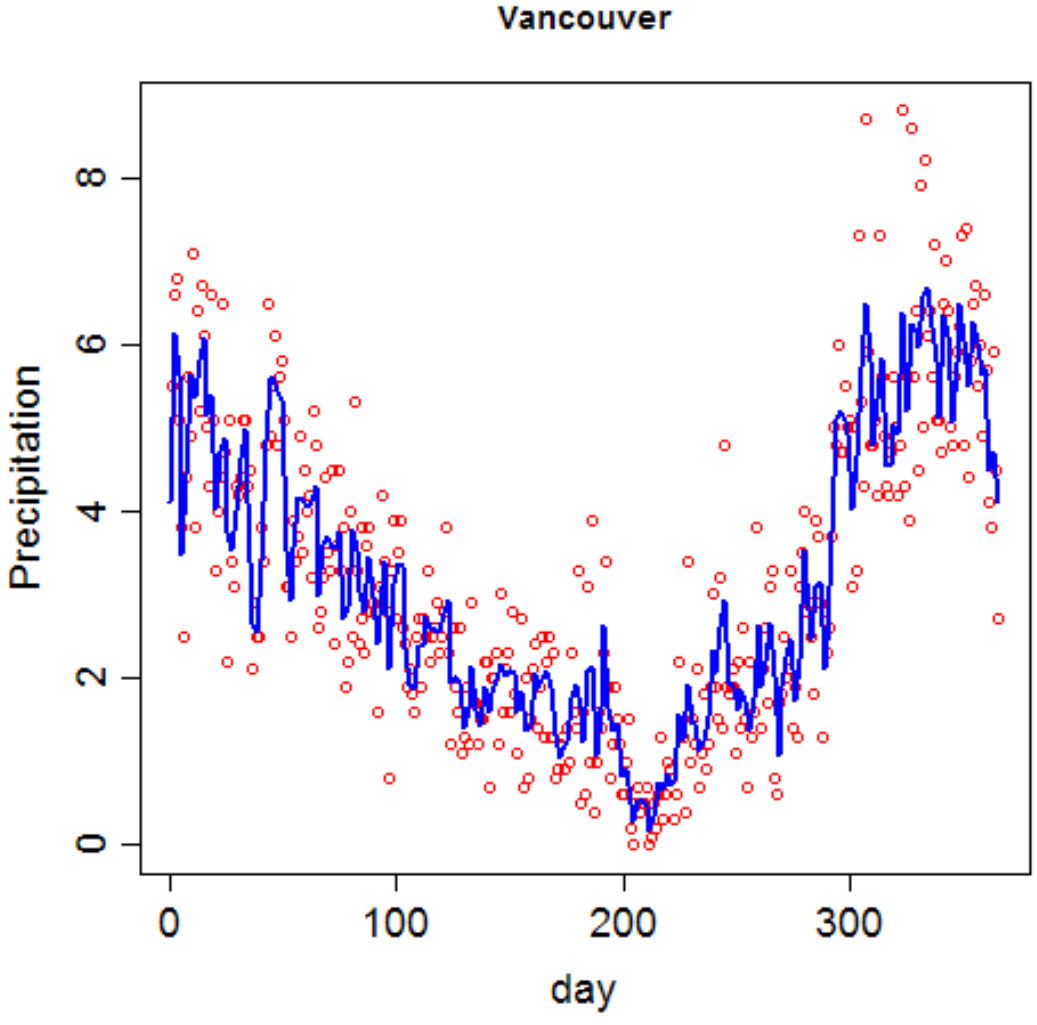
- We now have two competing desires: fit to data and smoothness.
- We will explicitly trade them off by minimizing *penalized squared error*:

$$PENSSE_\lambda(f) = [\mathbf{y} - f(\mathbf{t})]^T [\mathbf{y} - f(\mathbf{t})] + \lambda J_2[f]$$

- λ is a *smoothing parameter* measuring compromise between fit and smoothness.
- As λ increases, roughness is increasingly penalized and $f(t)$ will become linear.
- As λ decreases, the penalty is reduced and allows $f(t)$ to fit the data better.

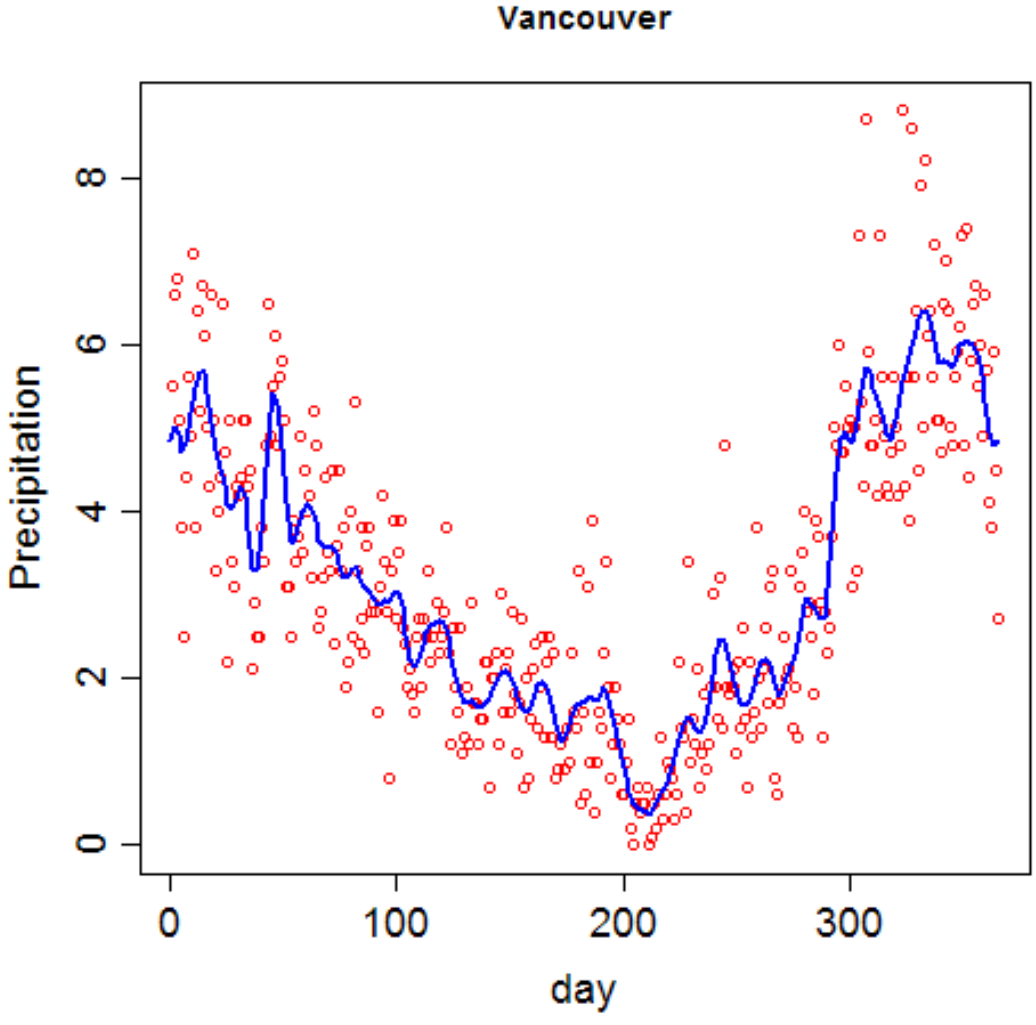
Experiments with the Vancouver Weather

$\log \lambda = -1$



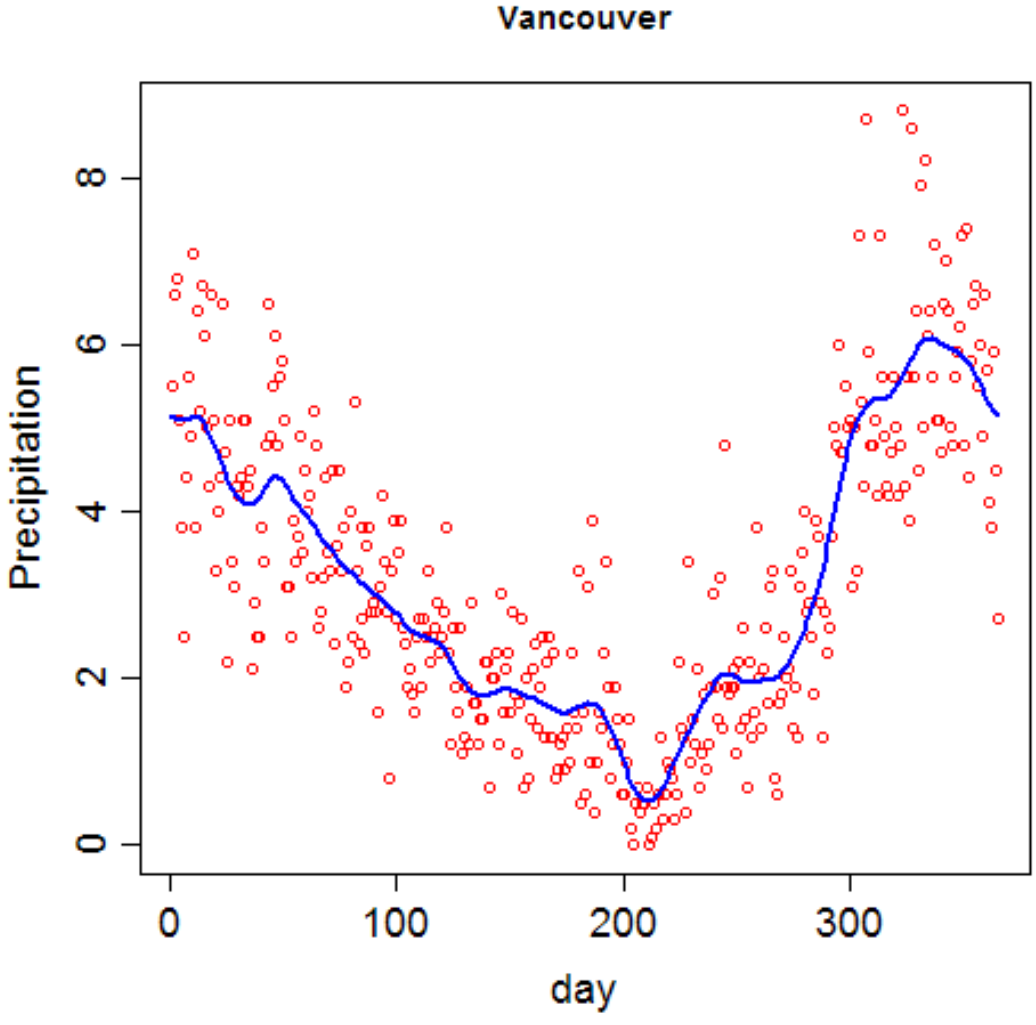
Experiments with the Vancouver Weather

$\log \lambda = 3$



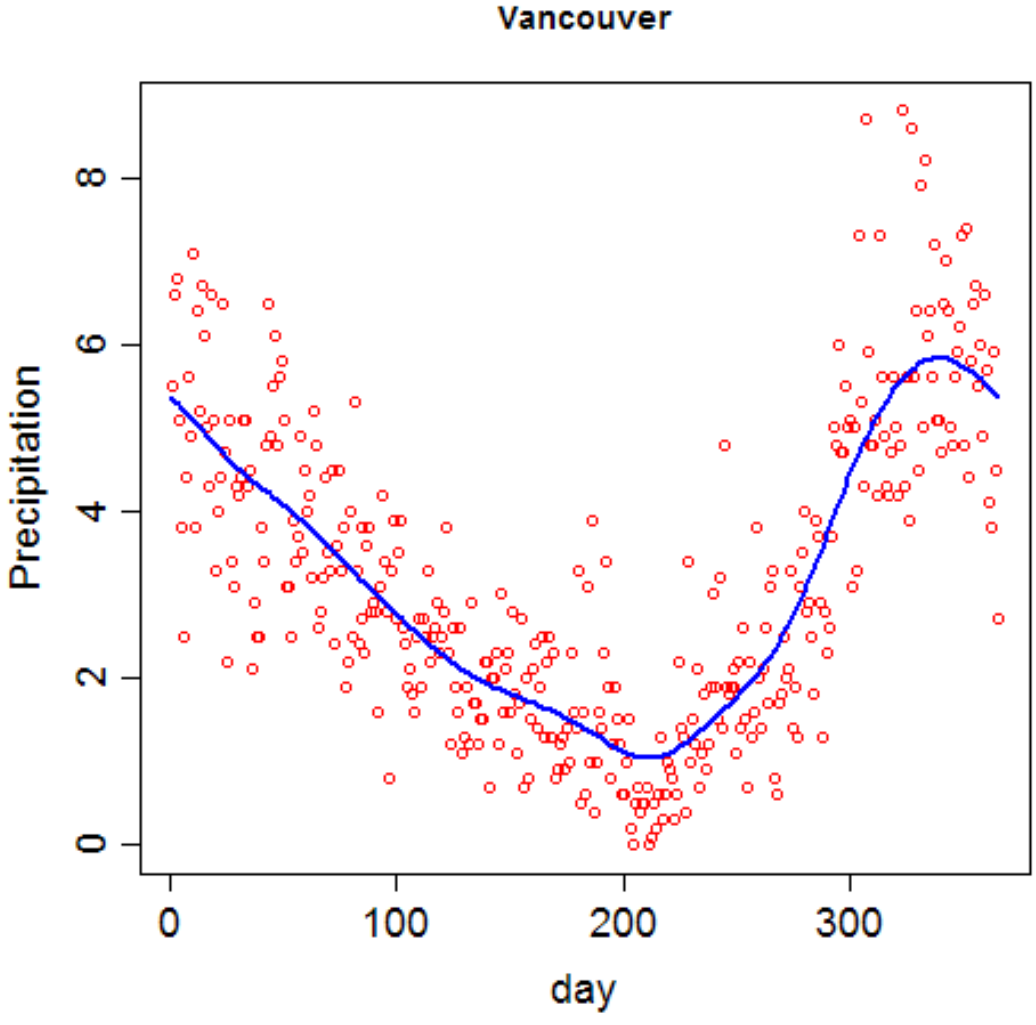
Experiments with the Vancouver Weather

$\log \lambda = 7$



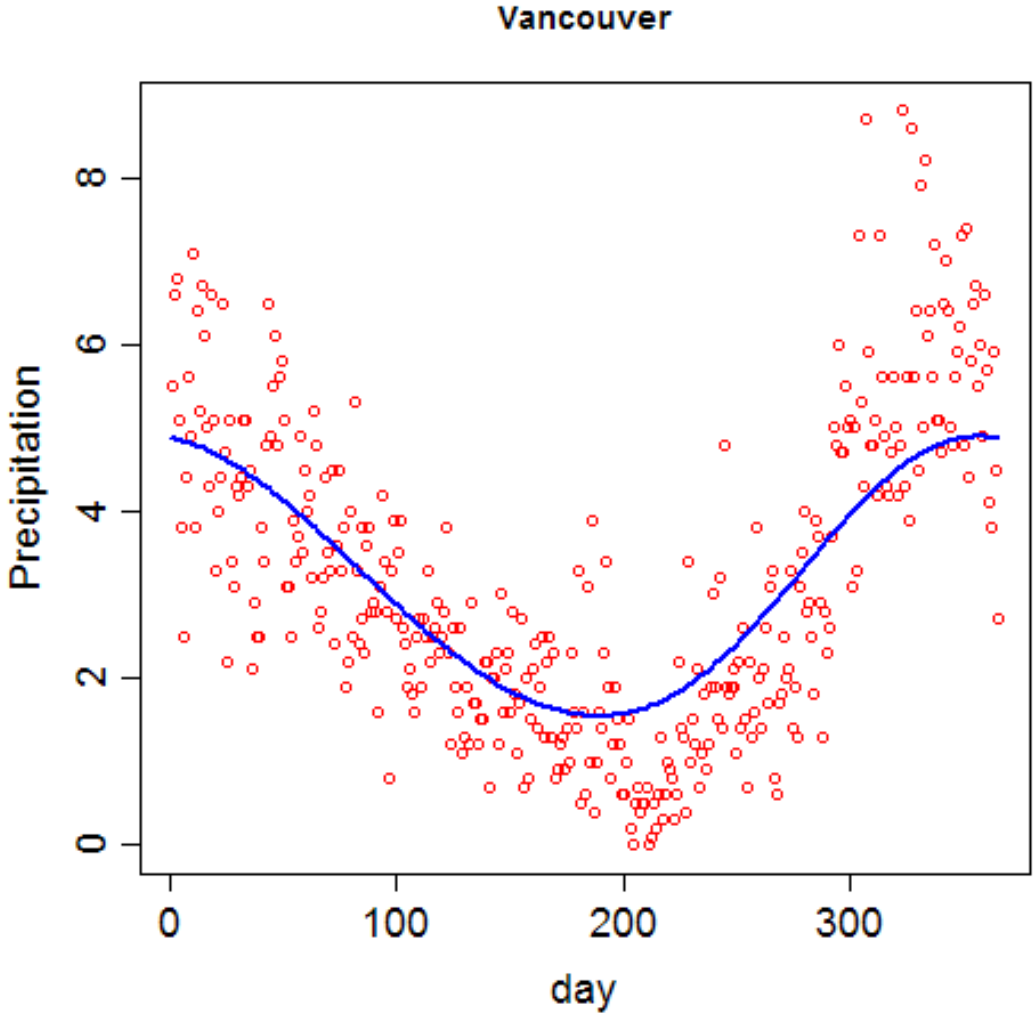
Experiments with the Vancouver Weather

$\log \lambda = 11$



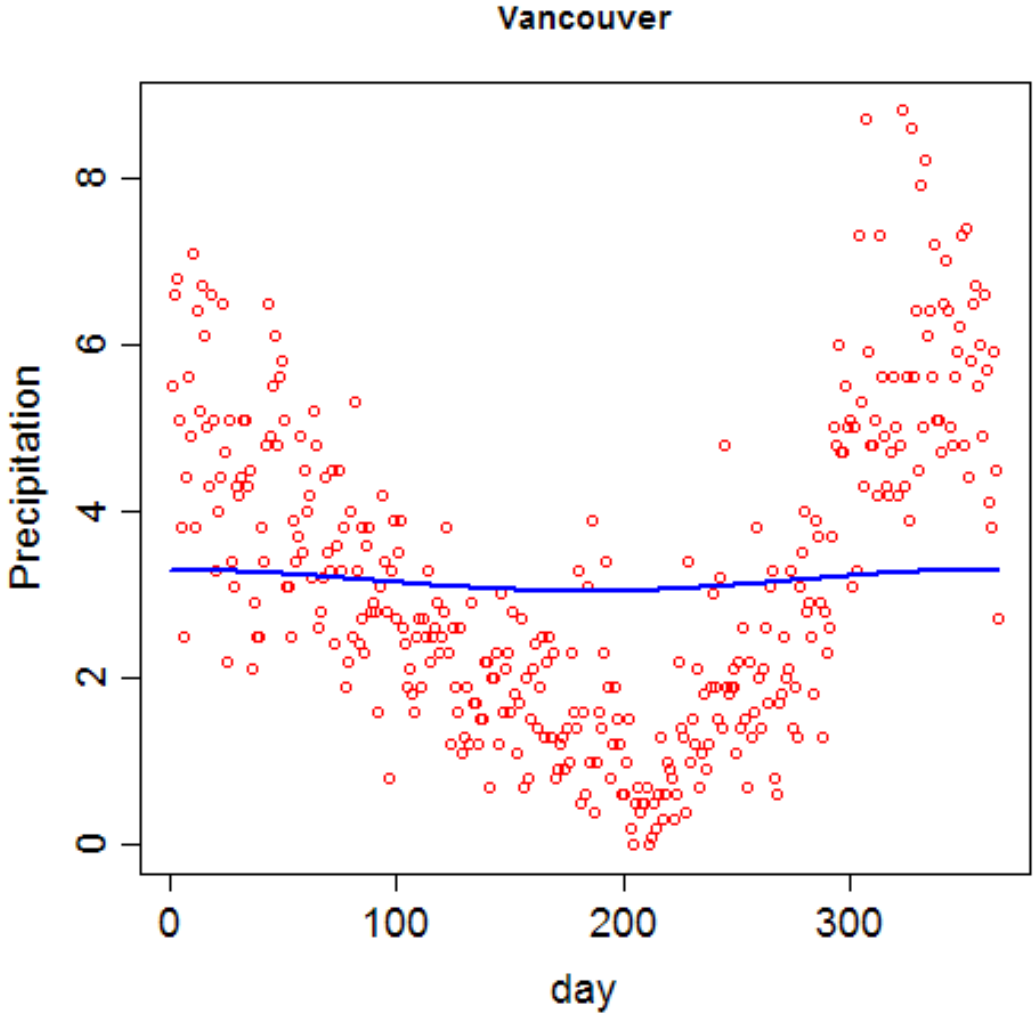
Experiments with the Vancouver Weather

$\log \lambda = 15$



Experiments with the Vancouver Weather

$\log \lambda = 19$



The Smoothing Spline Theorem

Consider the "usual" penalized squared error:

$$PENSSE_\lambda(x) = [\mathbf{y} - f(\mathbf{t})]^T [\mathbf{y} - f(\mathbf{t})] + \lambda \int \left[D^2 f(t)\right]^2 dt$$

- A remarkable theorem tells us that the function $f(t)$ that minimizes $PENSSE_\lambda(f)$ is
  - a spline function of order 4 (piecewise cubic)
  - with a knot at each sample point t_j
- this is often referred to simply as *cubic spline smoothing*.

Computing the smoothing spline

- The theorem tells us that $f(t)$ is of the form

$$f(t) = \phi(t)^T \mathbf{c}$$

  where $\phi(t)$ is a vector of B-spline basis functions.
- The number of basis functions is $(n-2)+4 = n+2$ where n is the number of sampling points
- How do we calculate $\mathbf{c}$?

Calculating the Penalized Fit

When $f(t) = \Phi(t)\mathbf{c}$, we have that

$$\int [D^m f(t)]^2 \, dt = \int \mathbf{c}^T D^m \Phi(t) D^m \Phi(t)^T \mathbf{c} = \mathbf{c}^T R \mathbf{c}$$

R is known as the *penalty matrix*.

The penalized least squares estimate for $\mathbf{c}$ is now

$$\hat{\mathbf{c}} = \left[\mathbf{\Phi}^T \mathbf{\Phi} + \lambda R \right]^{-1} \mathbf{\Phi}^T \mathbf{y}$$

This is still a linear smoother:

$$\hat{\mathbf{y}} = \mathbf{\Phi} \left[\mathbf{\Phi}^T \mathbf{\Phi} + \lambda R \right]^{-1} \mathbf{\Phi}^T \mathbf{y}$$

Linear Smooths and Influence

A procedure is a linear smooth if

$$\hat{\mathbf{y}} = S\mathbf{y}$$

for some S. Eg

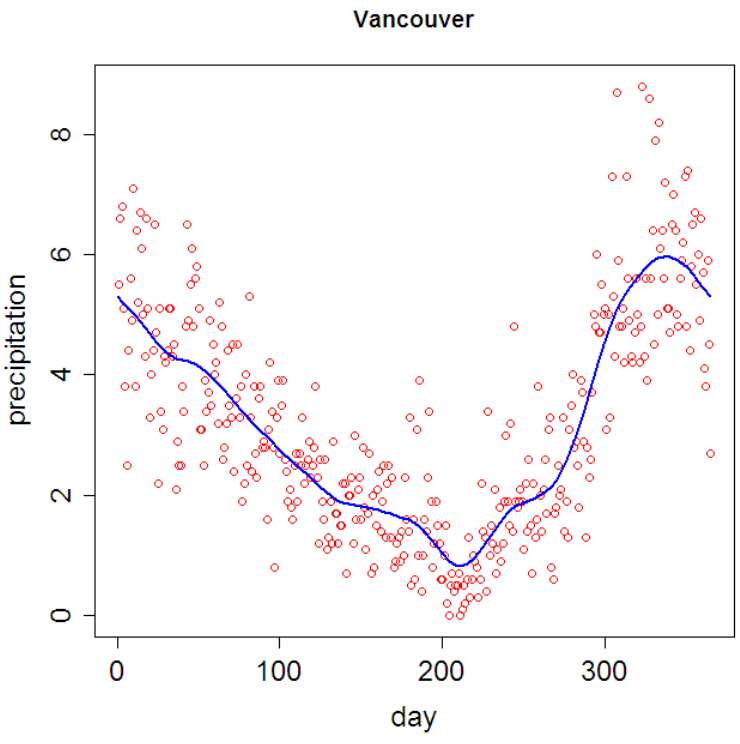
$$S = \mathbf{\Phi}\left[\mathbf{\Phi}^T\mathbf{\Phi} + \lambda R\right]^{-1}\mathbf{\Phi}^T$$

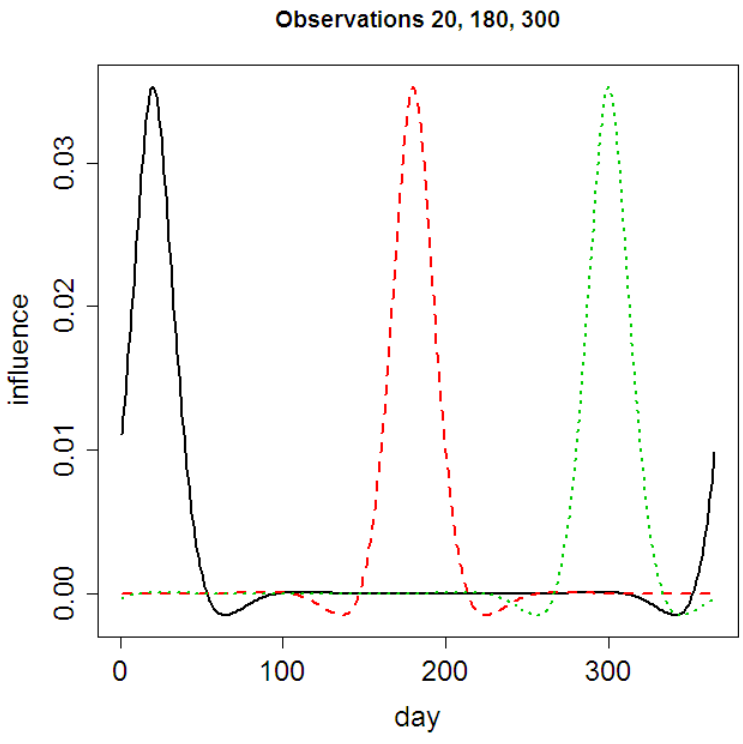
Computing the smoothing spline is also a linear operator

$$\hat{f}(t) = \Phi^T(t)\left[\mathbf{\Phi}^T\mathbf{\Phi} + \lambda R\right]^{-1}\mathbf{\Phi}^T\mathbf{y}$$

and we can calculate the influence of each y_i on the curve.

Influence in Vancouver Precipitation





Linear Smooths and Degrees of Freedom

- In least squares fitting, the degrees of freedom used to smooth the data is exactly J, the number of basis functions
- In penalized smoothing, we can have $J > n$.
- The smoothing penalty reduces the flexibility of the smooth (ie, we say we know something).
- The degrees of freedom are controlled by λ. A natural measure turns out to be

$$df(\lambda) = \text{trace}(S)$$

- Vancouver precipitation above was fit with 365 basis functions, $\lambda = 10^4$ resulting in $df = 12.92$.

Alternative Definitions of Roughness

- $D^2 f(t)$ is only one way to measure the roughness of f.
- If we were interested in $D^2 f(t)$, we might think of penalizing $D^4 f(t) \Rightarrow$ cubic polynomials are not rough.
- What about the weather data? we know it's periodic, and not very different from a sinusoid.
- The *Harmonic acceleration* of f is

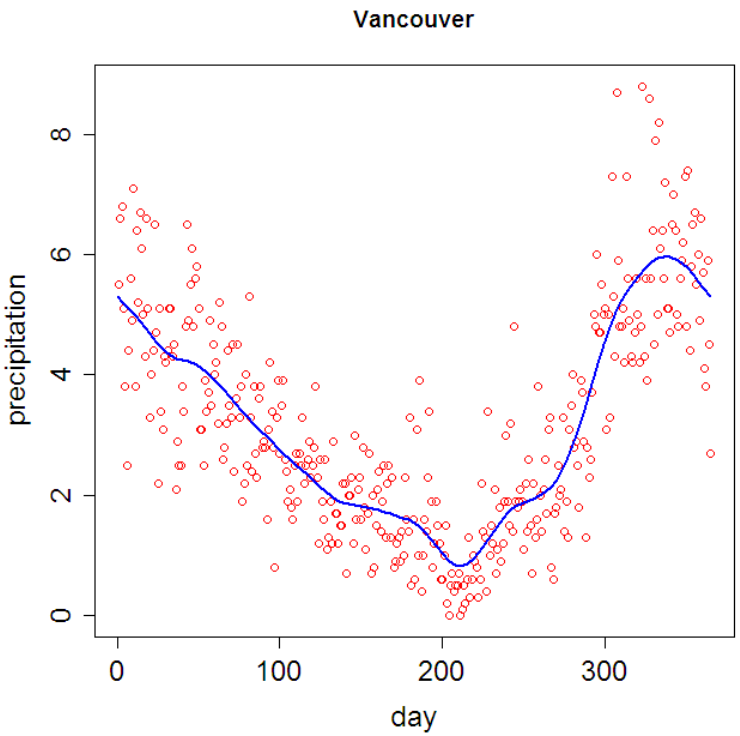
$$Lf = \omega^2 Df + D^3 f$$

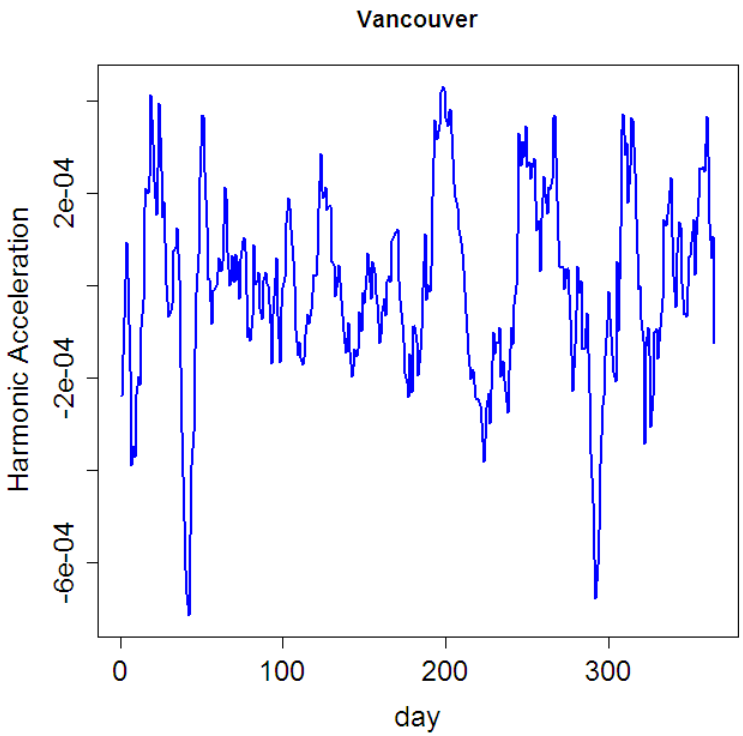
and $L\cos(\omega t) = 0 = L\sin(\omega t)$.

- We can measure departures from a sinusoid by

$$J_L(f) = \int [Lf(t)]^2 \, dt$$

Harmonic Acceleration in Vancouver Precipitation





A Very General Notion

We can be even more general and allow roughness penalties to use any *linear differential operator*

$$Lf(t) = \sum_{k=1}^{K} \alpha_k(t) D^k f(t)$$

Then f is "smooth" if $Lf(t) = 0$.

We will see later on that we can even ask the data to tell us what should be smooth.

However, we will rarely need to use anything so sophisticated.

Basis Expansions and Roughness

There is always a function f minimizing

$$\text{PENSSE}_L(f) = (\mathbf{y} - f(\mathbf{t}))^T(\mathbf{y} - f(\mathbf{t})) + \lambda \int [Lf(t)]^2 \, dt$$

However, it can be annoying (and expensive) to calculate explicitly.

Instead, we simply use a basis expansion with

$$R = \int L\Phi(t) L\Phi(t)^T dt$$

This is given by formulae for simple cases, but may need to be evaluated numerically.

Although the handwriting data potentially needs 1403 cubic B-splines, 50 does a very nice job.

There are mathematical results to back this up.

Choosing the Smoothing Parameter

There are a number of data-driven methods for choosing smoothing parameters.

- Ordinary Cross Validation: leave one point out and see how well you can predict it. For any linear smooth

$$\mathrm{OCV}(\lambda) = \frac{1}{n}\sum \frac{(y_i - f(t_i))^2}{(1 - S_{ii})^2}$$

  where S is the smoothing matrix.
- Generalized cross validation (see next slide)
- Various information criteria (not used here)
- ReML estimation

Generalized Cross Validation

$$\text{GCV}(\lambda) = \frac{n^{-1}\sum(y_i - f(t_i))^2}{[n^{-1}\text{trace}(I - S)]^2}$$

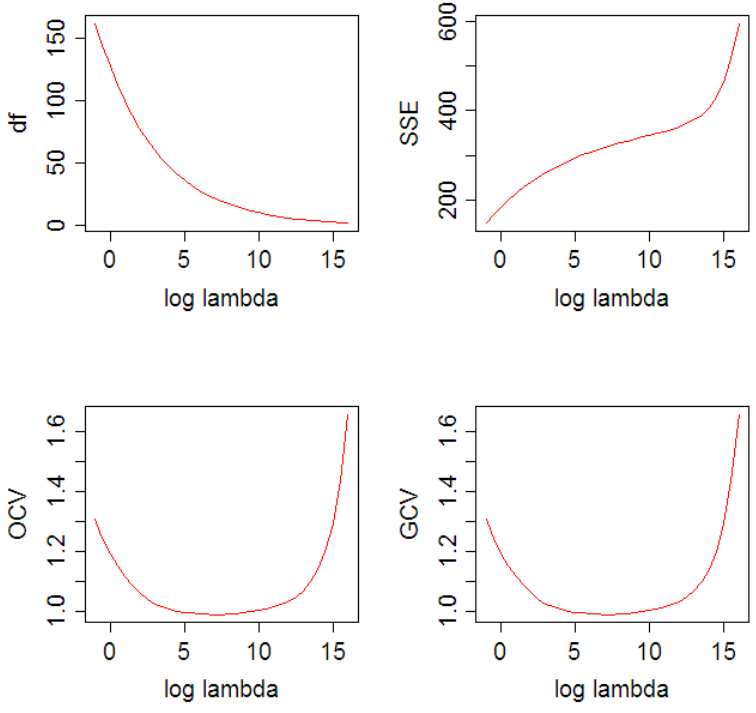
- Doesn't actually generalize anything
- We can re-write this as

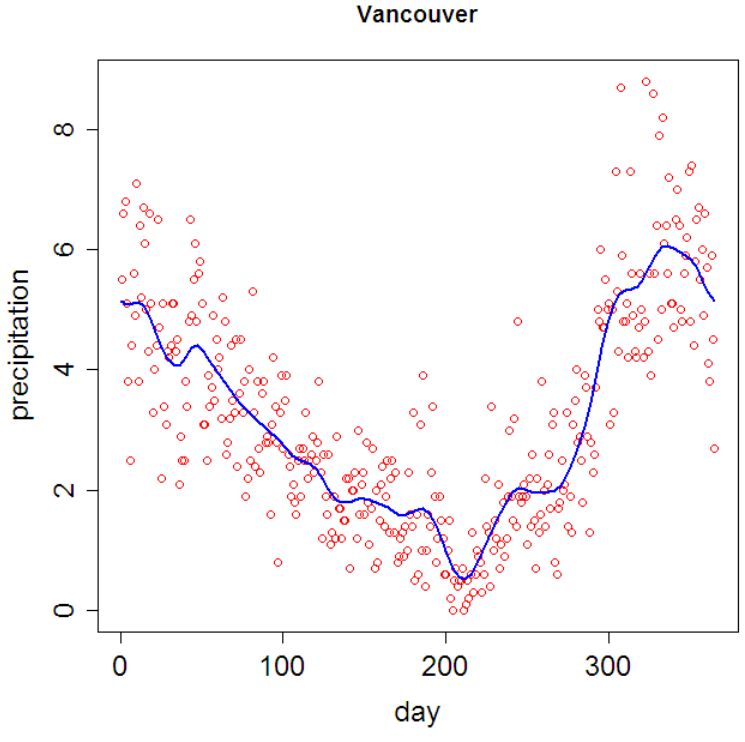
$$\text{GCV}(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{\text{SSE}}{n - df(\lambda)}\right)$$

- SSE is discounted for degrees of freedom like usual, but we then also make a discount for minimizing over λ.
- GCV smooths more than OCV; even then, it may need to be tweaked a little to produce pleasing results.

Various Statistics on Vancouver Precipitation

Optimal $\lambda = 1096.633$





Confidence Intervals for Linear Probes

Recall

$$\hat{\mathbf{y}} = S\mathbf{y}, \text{where } \; S = \mathbf{\Phi}\left[\mathbf{\Phi}^T\mathbf{\Phi} + \lambda R\right]^{-1}\mathbf{\Phi}^T$$

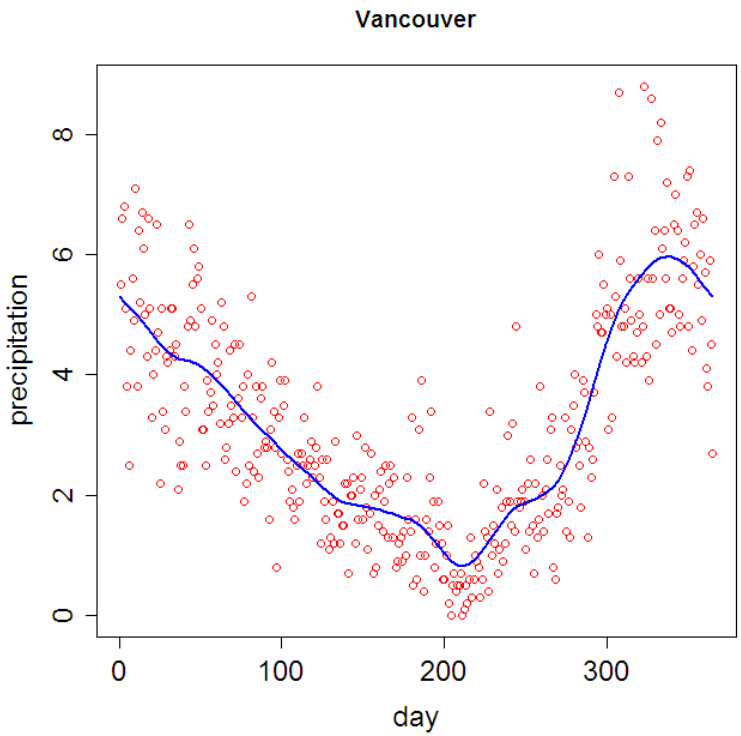
Then

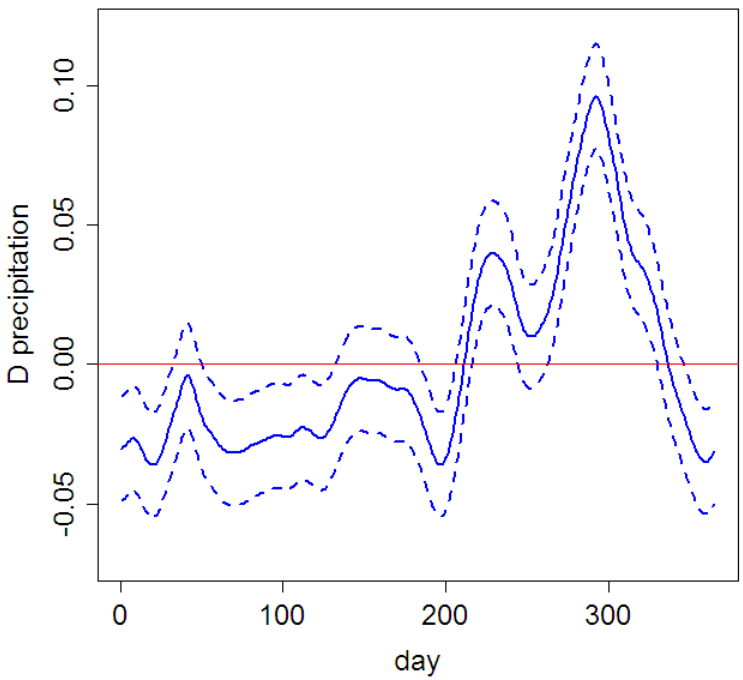
$$\text{Var}(\hat{\mathbf{y}}) = \sigma^2 SS^T$$

Confidence Intervals for a Derivative

We have

$$Df(t) = D\Phi(t)C\mathbf{y} \rightarrow \text{Var}[Df(t)] = D\Phi(t)C\text{Var}[\mathbf{y}]C^T D\Phi(t)^T$$





Summary

1. Defined "smoothness" in terms of derivatives of f.
2. We can use smoothness as a way to regularize the estimate of f.
3. Use of λ to trade-off smoothness and fit.
4. GCV as a means of choosing λ
5. Linear smooths and linear probes gives a means of providing confidence for features of f.