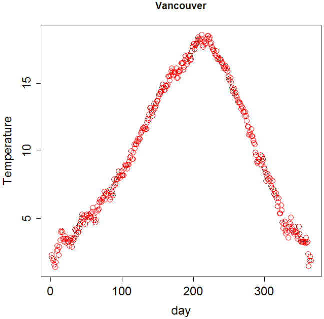
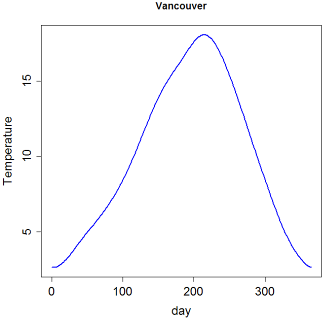
How do we get from

data to functions?





Least-Squares

Assume we have observations for a single curve

$$y_i = f(t_i) + \epsilon$$

and we want to estimate

$$f(t) = \sum_{j=1}^{J} c_j \phi_j(t)$$

Least-Squares

Assume we have observations for a single curve

$$y_i = f(t_i) + \epsilon$$

and we want to estimate

$$f(t) = \sum_{j=1}^{J} c_j \phi_j(t)$$

Minimize the sum of squared errors:

$$H = \sum_{i=1}^{n} (y_i - f(t_i))^2 = \sum_{i=1}^{n} (y_i - \mathbf{c}^T \Phi(t_i))^2$$

Least-Squares

Assume we have observations for a single curve

$$y_i = f(t_i) + \epsilon$$

and we want to estimate

$$f(t) = \sum_{j=1}^{J} c_j \phi_j(t)$$

Minimize the sum of squared errors:

$$H = \sum_{i=1}^{n} (y_i - f(t_i))^2 = \sum_{i=1}^{n} (y_i - \mathbf{c}^T \Phi(t_i))^2$$

This is just linear regression!

Linear Regression on Basis Functions

- If the n by J matrix $\mathbf{\Phi}$ contains the values $\phi_k(t_j)$, and $\mathbf{y}$ is the vector $(y_1, \ldots, y_n)$, we can write

$$H(\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})^T (\mathbf{y} - \mathbf{\Phi c})$$

- The error sum of squares is minimized by the *ordinary least squares estimate*

$$\hat{\mathbf{c}} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{y}$$

- Then we have the estimate

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

- n by J matrix $\mathbf{\Phi}$ contains the values $\phi_k(t_j)$
- The least squares estimate

$$\hat{\mathbf{c}} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{y}$$

- The estimate

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

- When we look at the values of $\hat{f}$ at the observation points we have

$$\hat{\mathbf{y}} = \mathbf{\Phi} \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{y} = S\mathbf{y}$$

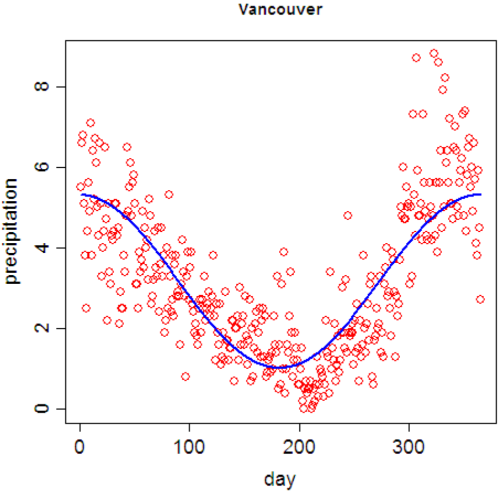
S is referred to as the *smoothing matrix*.

Choosing the Number of Basis Functions

- Small numbers of basis functions mean little flexibility
- Larger numbers of basis functions add flexibility, but may "overfit"
- For Monomial and Fourier bases, just add functions to the collection.
- Spline bases: adding knots or increasing the order changes the basis; but makes it more flexible.
- Spline bases: *changing* the knots may not help even if you add more of them; but this is unusual.

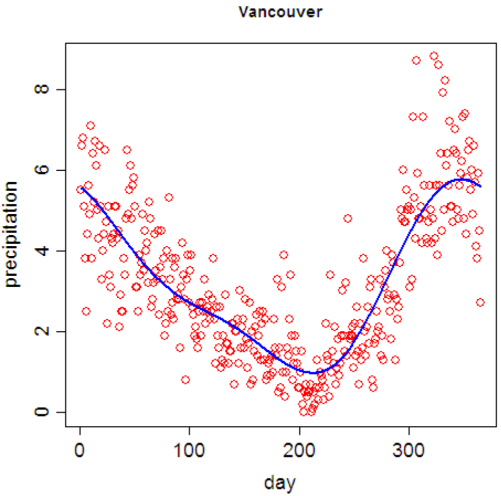
Choosing the Number of Basis Functions

Vancouver Precipitation: 3 Fourier Bases



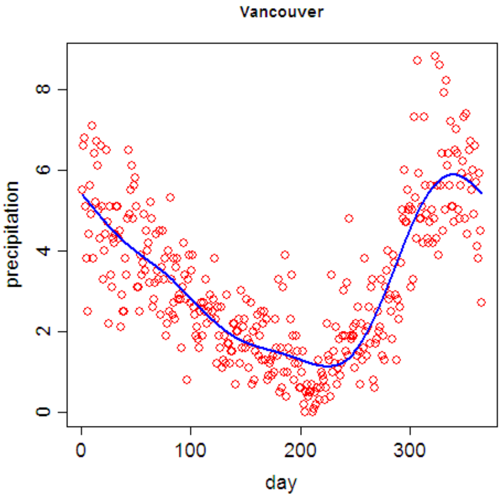
Choosing the Number of Basis Functions

Vancouver Precipitation: 5 Fourier Bases



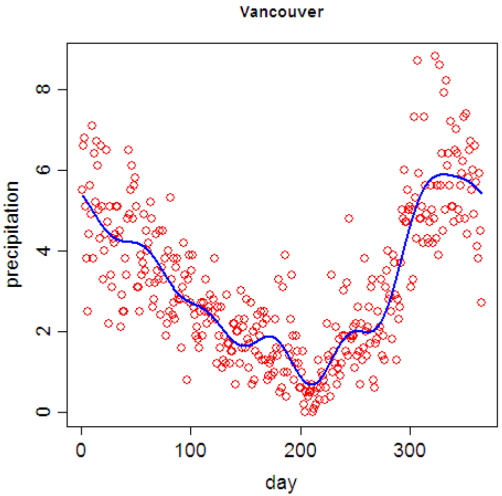
Choosing the Number of Basis Functions

Vancouver Precipitation: 7 Fourier Bases



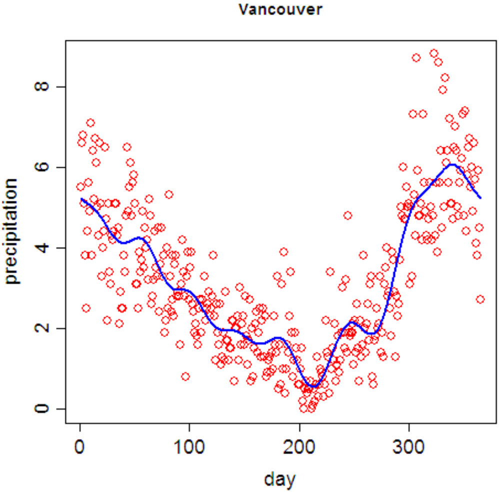
Choosing the Number of Basis Functions

Vancouver Precipitation: 13 Fourier Bases



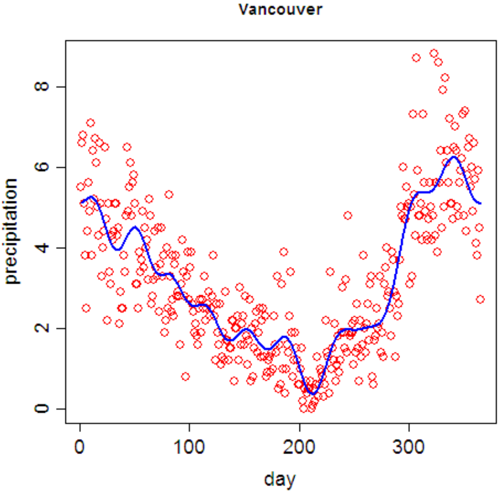
Choosing the Number of Basis Functions

Vancouver Precipitation: 19 Fourier Bases



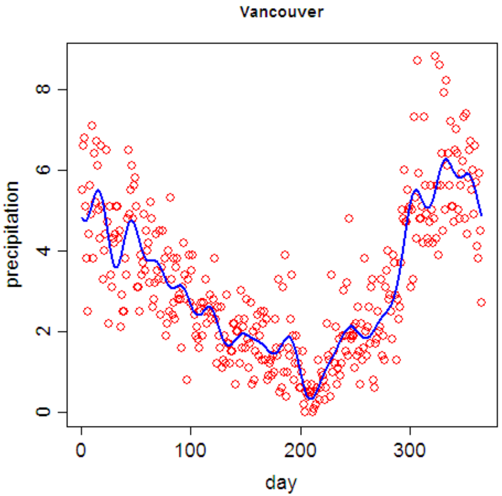
Choosing the Number of Basis Functions

Vancouver Precipitation: 25 Fourier Bases



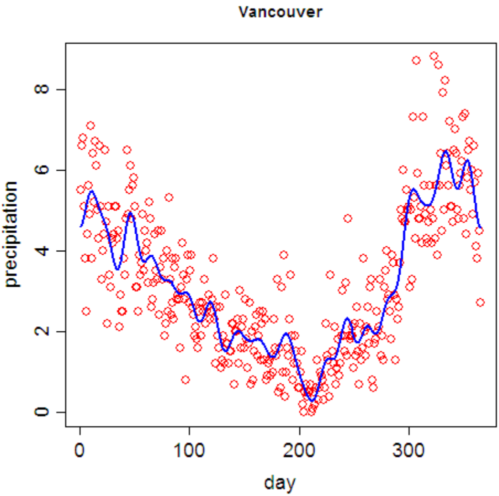
Choosing the Number of Basis Functions

Vancouver Precipitation: 31 Fourier Bases



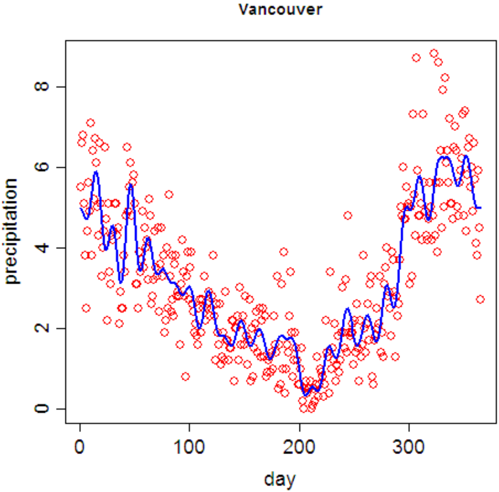
Choosing the Number of Basis Functions

Vancouver Precipitation: 41 Fourier Bases



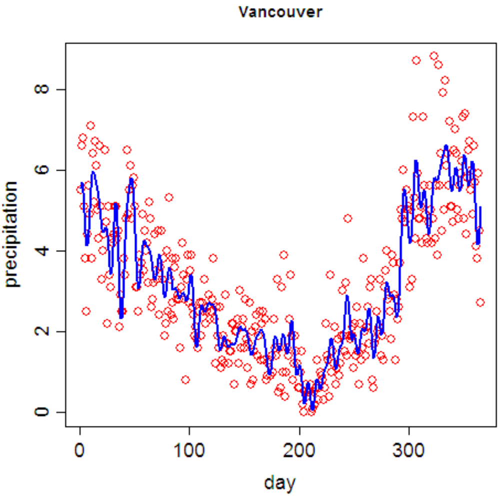
Choosing the Number of Basis Functions

Vancouver Precipitation: 53 Fourier Bases



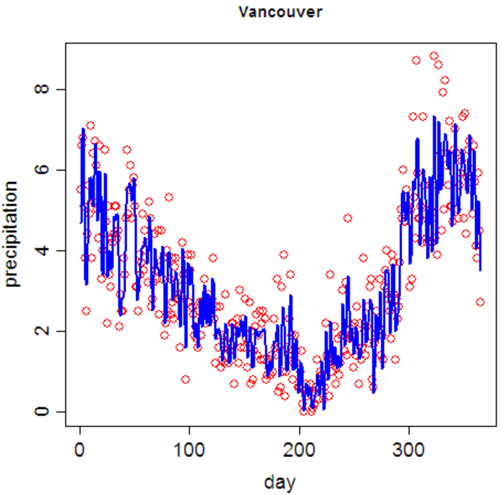
Choosing the Number of Basis Functions

Vancouver Precipitation: 105 Fourier Bases



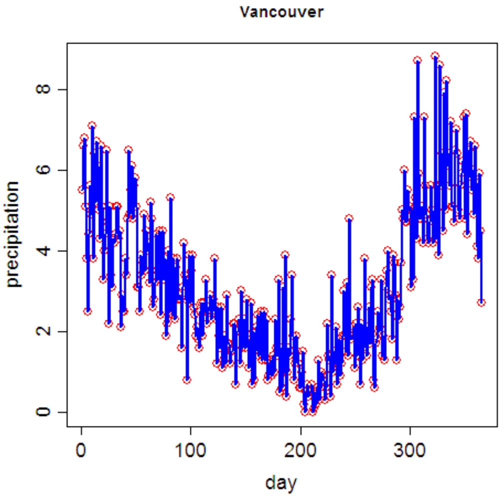
Choosing the Number of Basis Functions

Vancouver Precipitation: 207 Fourier Bases



Choosing the Number of Basis Functions

Vancouver Precipitation: 365 Fourier Bases



Choosing the Number of Basis Functions

Trade off:

- Too many basis functions over-fits the data and reflect errors of measurement
- Too few basis functions fails to capture interesting features of the curves.

Bias and Variance Tradeoff

- Express this trade-off in terms of
  - the *bias* of the estimate of $f(t)$:

$$\text{Bias}\left[\hat{f}(t)\right] = f(t) - \hat{f}(t)$$

  - the *sampling variance* of the estimate

$$\text{Var}\left[\hat{f}(t)\right] = E\left[\left\{\hat{f}(t) - E\hat{f}(t)\right\}^2\right]$$

- Too many basis functions means small bias but large sampling variance.
- Too few basis functions means small sampling variance but large bias.

Mean Squared Error

- Usually, we would really like to minimize *mean squared error*

$$\mathrm{MSE}\left[\hat{f}(t)\right] = E\left[\left\{\hat{f}(t) - f(t)\right\}^2\right]$$

- there is a simple relationship between MSE and bias/variance

$$\mathrm{MSE}\left[\hat{f}(t)\right] = \mathrm{Bias}^2\left[\hat{f}(t)\right] + \mathrm{Var}\left[\hat{f}(t)\right]$$

- This is expressed for each t, in general, we would like to minimize the *integrated* mean squared error:

$$\mathrm{IMSE}\left[\hat{f}(t)\right] = \int \mathrm{MSE}\left[\hat{f}(t)\right] dt$$

A Simulation

- Fit Vancouver precipitation by B-splines, to get $f(t_i)$
- Pretend this is the "truth"
- Calculate "errors"

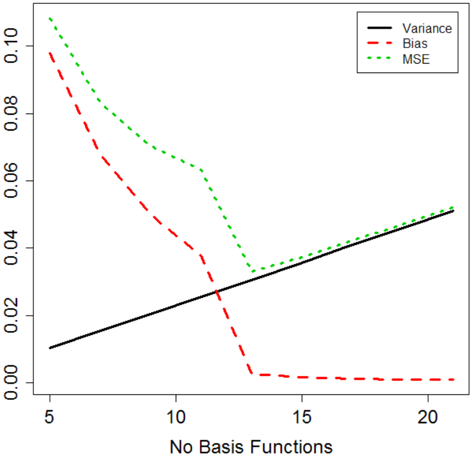
$$\epsilon_i = y_i - f(t_i)$$

- Create new "data" by randomly re-arranging the errors

$$y_i^* = f(t_i) + \epsilon_{i^*}$$

- Now fit the new data using a Fourier basis
- Repeat 1000 times; calculate bias and variance from sample.

Bias and Variance from Simulation



Cross-Validation

One method of choosing a model:

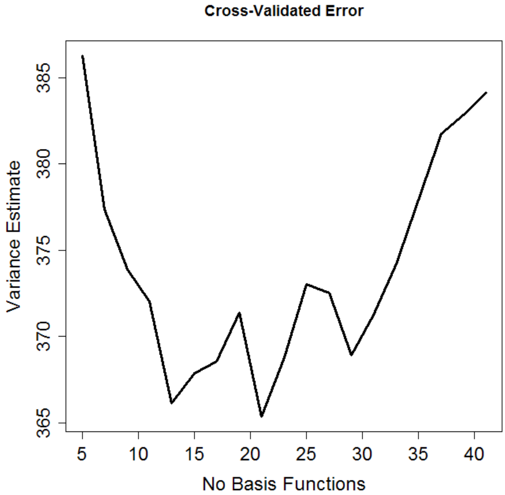
- leave out one observation (t_i, y_i)
- estimate $\hat{f}_{-i}(t)$ from remaining data
- measure $y_i - \hat{f}_{-i}(t_i)$
- Choose K to minimize the *ordinary cross-validation* score:

$$\text{OCV}(K) = \sum \left(y_i - \hat{f}_{-i}(t_i) \right)^2$$

- for a linear smooth $\hat{y} = Sy$,

$$\text{OCV}(K) = \sum \frac{\left(y_i - \hat{f}(t_i) \right)^2}{(1 - s_{ii})^2}$$

Cross Validation for Vancouver Precipitation



Estimating the Residual Covariance

- If we assume the standard model, then

$$\mathrm{Var}[\mathbf{y}] = \sigma^2 I$$

- An unbiassed estimate is

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

Sampling Variance of the Curve

- We know that $\mathbf{c} = C\mathbf{y}$ for $C = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T$
- Then under the standard model

$$\text{Var}\left[\mathbf{c}\right] = \sigma^2 CC^T = \sigma^2\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}$$

- Then the sample variance of $\hat{f}(t)$ is

$$\text{Var}\left[\hat{f}(t)\right] = \sigma^2\Phi(t)^T\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\Phi(t)$$

- And the variance-covariance matrix of the fitted values is

$$\text{Var}\left[\hat{\mathbf{f}}\right] = \sigma^2\mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T$$

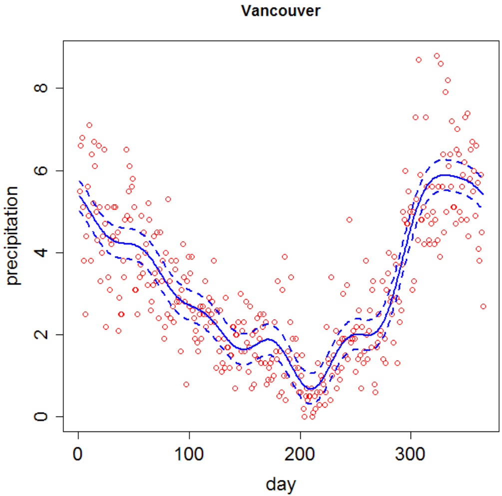
Pointwise Confidence Bands

- For each point we calculate lower and upper bands for $\hat{f}(t)$ by

$$\hat{f}(t) \pm 2\sqrt{\mathrm{Var}\left[\hat{f}(t)\right]}$$

- These bands are not confidence bands for the entire curve, but only for the value of the curve at a fixed point.
- Ignores bias in the estimated curve
- Provide an impression of how well the curve is estimated.

Fitted Vancouver Precipitation Data with 13 Fourier Bases



Summary

- Fitting smooth curves is just linear regression using basis functions as independent variables.
- Trade-off between bias and variance in choosing the number of basis functions
- Cross-validation is one way to quantitatively find the best number of basis functions
- Confidence intervals can be calculated using the standard model, but these should be treated with care
- We will see next time that there are better ways to control bias and variance.