How do we get from data to functions?

**Least-Squares**

Assume we have observations for a single curve

\[ y_i = x(t_i) + \epsilon \]

and we want to estimate

\[ x(t) \approx \sum_{j=1}^{j} c_j \phi_j(t) \]

Minimize the sum of squared errors:

\[ SSE = \sum_{i=1}^{n} (y_i - x(t_i))^2 = \sum_{i=1}^{n} (y_i - c^T \Phi(t_i))^2 \]

This is just linear regression!

**Linear Regression on Basis Functions**

- If the \( N \) by \( K \) matrix \( \Phi \) contains the values \( \phi_k(t_j) \), and \( y \) is the vector \( (y_1, \ldots, y_N) \), we can write

\[ SSE(c) = (y - \Phi c)^T (y - \Phi c) \]

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

\[ \hat{c} = (\Phi^T \Phi)^{-1} \Phi^T y \]

- Then we have the estimate

\[ \hat{x}(t) = \Phi(t) \hat{c} = \Phi(t) \left( \Phi^T \Phi \right)^{-1} \Phi^T y \]

**The Standard Model for Residual Distribution**

- least squares is optimal for residuals that are independently and identically normal with mean 0 and variance \( \sigma \).
- That is

\[ E[y] = \Phi c \] and \( \text{Var}[y] = \sigma^2 I \)
- Call this the **standard model** for the distribution of residuals.
The standard model for residual distribution

- least squares is optimal for residuals that are independently and identically normal with mean 0 and variance $\sigma$.
- That is
  \[ E_y = \Phi c \] and \[ \text{Var}[y] = \sigma^2 I \]
- Call this the standard model for the distribution of residuals.

Weighted Least Squares

The standard model is often overly simplistic
- $\text{Var}[y]$ may vary with observation time
- The residuals may be correlated.

The first of these can be compensated for by weighting the observations

\[ WMSE[x] = \sum w_i (y_i - x(t_i))^2 \]

Set $W$ to have $w_i$ on the diagonal, we get

\[ \hat{x}(t) = \Phi(t) \hat{c} = \Phi(t) \left( \Phi^T W \Phi \right)^{-1} \Phi^T W y \]

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

\[ \hat{y} = \Phi \left( \Phi^T W \Phi \right)^{-1} \Phi^T W y = Sy \]

$S$ 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.

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: 53 Fourier Bases
Choosing the Number of Basis Functions

Vancouver Precipitation: 105 Fourier Bases

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.

Vancouver Precipitation: 207 Fourier Bases

Vancouver Precipitation: 365 Fourier Bases
Bias and Variance Tradeoff

- Express this trade-off in terms of:
  - the bias of the estimate of \( x(t) \):
    \[
    \text{Bias} [\hat{x}(t)] = x(t) - E\hat{x}(t)
    \]
  - the sampling variance of the estimate
    \[
    \text{Var} [\hat{x}(t)] = E \left( \hat{x}(t) - E\hat{x}(t) \right)^2
    \]
- 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
  \[
  \text{MSE} [\hat{x}(t)] = E \left( \hat{x}(t) - \hat{x}(t) \right)^2
  \]
- There is a simple relationship between MSE and bias/variance
  \[
  \text{MSE} [\hat{x}(t)] = \text{Bias}^2 [\hat{x}(t)] + \text{Var} [\hat{x}(t)]
  \]
- This is expressed for each \( t \), in general, we would like to minimize the integrated mean squared error:
  \[
  \text{IMSE} [\hat{x}(t)] = \int \text{MSE} [\hat{x}(t)] \, dt
  \]

A Simulation

- Fit Vancouver precipitation by B-splines, to get \( x(t_i) \)
- Pretend this is the "truth"
- Calculate “errors”
  \[
  \epsilon_i = y_i - x(t_i)
  \]
- Create new “data” by randomly re-arranging the errors
  \[
  y_i^* = x(t_i) + \epsilon_i^*
  \]
- 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 \(\hat{x}_{-i}(t)\) from remaining data
- measure \(y_i - \hat{x}_{-i}(t)\)
- Choose \(K\) to minimize the ordinary cross-validation score:

\[
OCV[\hat{x}] = \sum (y_i - \hat{x}_{-i}(t_i))^2
\]

for a linear smooth \(\hat{y} = S y\),

\[
OCV[\hat{x}] = \sum \frac{(y_i - \hat{x}(t_i))^2}{(1 - s_{ii})^2}
\]

Estimating the Residual Covariance

- If we assume the standard model, then

\[
\text{Var}[y] = \sigma^2 I
\]

- An unbiased estimate is

\[
\hat{\sigma}^2 = \frac{1}{N - K} \text{MSSE}
\]

- Can be more sophisticated if residuals are correlated (will ignore here).

Cross Validation for Vancouver Precipitation

Sampling Variance of the Curve

- We know that \(c = Cy\) for \(C = (\Phi^T W \Phi)^{-1} \Phi^T W\)
- Then under the standard model

\[
\text{Var}[c] = \sigma^2 C I C^T
\]

- More generally, if \(\text{Var}[y] = \Sigma\), we have

\[
\text{Var}[c] = \sigma^2 C \Sigma C^T
\]

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

\[
\text{Var}[\hat{y}(t)] = \Phi(t)^T C \Sigma C^T \Phi(t)
\]

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

\[
\text{Var}[\hat{y}] = \Phi C \Sigma C^T \Phi^T
\]
Pointwise Confidence Bands

- For each point we calculate lower and upper bands for $\hat{y}(t)$ by
  \[ \hat{y}(t) \pm 2\sqrt{\text{Var}[\hat{y}(t)]} \]
- 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

Montreal Temperature Data

Montrealers talk of a "January Thaw": is it real?

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.