Lec 3. Moving Beyond Linearity: Basis Expansion.

- Transformation of $\vec{x}$ to correct non linearity.
  - Core idea: augment / replace X with additional variables.
    then use linear models in the new space of derived input features.

  - Denote $h_m(x): R^p \to R$, $m = 1, \cdots M$.

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(X) \quad , \quad \text{a linear expansion of } X.$$

  - $h_m(x) = X_m$, $m = 1, \cdots p$. original linear model.
  - $h_m(x) = X_j^2$ or $h_m(x) = X_j X_k$. polynomial (quadratic).
  - $h_m(x) = \log(X_j)$, $\sqrt{X_j}$.
  - $h_m(x) = I(L_m \le X_k \le U_m)$. indicator. (step function). piecewise constant.

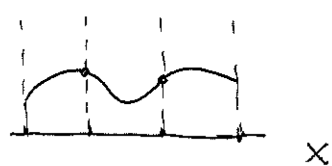
~~Assume X is one dimensional.~~

~~1. Polynomials.~~

  - Basis function: a family of functions or transformations that can be applied to X
    e.g. Fourier series / wavelets. regression splines.

Assume X is 1-dim.

1. Regression splines.



knots: the points where the coeff change

order: degree + 1

1.1 Piece-wise polynomials.

  - $f(x)$. divide the domain of X into continuous intervals. & represent f by a separate polynomial in each interval
  - These fixed knots splines $\overset{\text{called}}{\Rightarrow}$ regression splines.

~~1.2. constraints & splines.~~

  - More knots $\to$ more flexible. K knots, $\to$ K + 1 polynomials.
  - Face discontinous & overfitting.
    (Show Figure ISLR 7.3 & ESL 5.2.).

1.2. Constraints.

- continuous. [i.e. no jump at knots].

- First & second derivatives are continuous $\rightarrow$ smooth.
  [i.e. a cubic splines or higher order]. $M-2$ continuous derivatives.
  ↳ a cubic spline with K knots uses $K+4$ degree of freedoms.

- boundary constraints.
  Splines can have high variance at the outer range of x (see Fig. 7.4 ~~line~~. Confidence band wide).

  * A natural spline is a regression spline with additional boundary constraints. :
    the function need to be linear at the boundary. [may introduce bias]

- Other types of splines : B-splines, truncated power splines.

1.3 Choosing the number & location of the Knots.

2. Smoothing Splines.

- Put a knot at every data point, i.e. knots $= \{x_1, x_2, \cdots, x_n\}$. maximum set of knots
- The complexity of the fit is controlled by regularization. i.e., among all $f(x)$ with 2 continous derivatives. find. the one.

(smoothing). regularization para.

☆ $$RSS(f,\lambda) = \min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(t))^2\, dt.$$

↑ fit the data. ↑ penalize wiggliness

df n ↑ ↓ 2

$\lambda = 0$	interpolating spline
small	wiggly
large	smooth.
$\lambda \to \infty$	straight line

(intercept + slope)

- It can be shown ☆ has an explicit, finite-dimensional, unique minimizer.
  ↳ a natural cubic spline with knots at the unique values of x_i, $i = 1, \cdots, N$.
  $f(x) = \sum_{j=1}^{N} N_j(x)\,\theta_j$, $N_j(x)$ — an N-dim set of basis functions of natural spline

$\Rightarrow RSS(\theta, \lambda) = (y - N\theta)^T (y - N\theta) + \lambda \theta^T \Omega_N \theta.$ $\{N\}_{ij} = N_j(x_i)$

$\{\Omega_N\}_{jk} = \int N_j''(t)\, N_k''(t)\, dt.$

$\Rightarrow \hat{\theta} = (N^T N + \lambda \Omega_N)^{-1} N^T y$, a generalized ridge regression.

2.1. Degrees of Freedom & Smoother Matrices.

$\hat{f} = N(N^TN + \lambda \Omega_N)^{-1} N^T y.$ $\Rightarrow$ A smoothing spline is a linear smoother.

$:= S_\lambda y.$

$\uparrow$ smoother matrix.

Define the effective degrees of freedom

$df_\lambda = \text{trace}(S_\lambda).$ Large $\rightarrow$ flexible model; small $\rightarrow$ rigid model

$\downarrow$

related to λ. how. (see. previous table)

* Since S_λ. symmetric (& + semi-definite)

Rewrite $S_\lambda = \underbrace{N(N^TN)^{-1/2}}_{U} [I + \lambda \underbrace{(N^TN)^{-\frac{1}{2}} \Omega_N (N^TN)^{-\frac{1}{2}}}_{:= A}]^{-1} (N^TN)^{-\frac{1}{2}} N^T$

A ind of λ. symmetric. ~~semi~~ positive semi definite.

$= U [I + \lambda A]^{-1} U^T$

By. spectrum decomposition. $A = V \,\text{diag}(v_1, \cdots v_N) V^T$

$VV^T = I$

$= (UV)\,\text{diag}\left(\frac{1}{1+\lambda v_j}\right)(UV)^T$

$\Rightarrow df_\lambda = \sum_{j=1}^{N} \frac{1}{1+\lambda v_j}$ monotone $\downarrow$ of λ. $\Rightarrow$ 1-1 relationship. no local minima. no reversals. tuning stable. GCV. well-behaved.

2.2. Automatic selection of the smoothing para.

a. Fixing the Df.

b. Bias - Variance Tradeoff

$Y = f(x) + \varepsilon$ $\quad \hat{f} = S_\lambda y.$

$\varepsilon \sim N(0, \sigma^2)$

$\text{Cov}(\hat{f}) = S_\lambda \text{Cov}(y) S_\lambda^T$

$= \sigma^2 S_\lambda S_\lambda^T$

$\text{Bias}(\hat{f}) = f - E(\hat{f}). = f - S_\lambda f$

$= (I - S_\lambda) f.$

From Statistical Decisicion theorey.

$\text{EPE}(\hat{f}_\lambda) = E[Y - \hat{f}_\lambda(x)]^2$

$= E[(Y - f) + (f - \hat{f}_\lambda(x))]^2$

$= \sigma^2 + E[(f - \hat{f})]^2 = E[f - E(\hat{f}(x)) + E[\hat{f}(x)) - \hat{f}(x)]^2.$

$= E[\text{Bias}^2(\hat{f}(x)) + \text{Var}(\hat{f}(x))] = \text{MSE}(\hat{f}_\lambda).$

*. Don't know the true function. No acess to EPE. and need an estimate.

⇒ K-fold CV. Generalized CV (GCV)

(Multivariate Case).

3. Multidimensional Splines. & Generalized Additive Models.

Many vars. $y = f(x_1, x_2, \cdots x_p) + \varepsilon$. A full general f.
- very flexible.
- statistically impossible.

3.1. Multidimensional splines.

$X = (X_1, X_2)$. Tensor - product Splines
- use splines in each var.
- Capture interaction.
- Explode in dim $M_1 M_2$ para.

basis $g_{jk}(X) = h_{1j}(X_1)\, h_{2k}(X_2)$. $j = 1, \cdots M_1$, $k = 1, \cdots, M_2$.

$$\Rightarrow g(X) = \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} \theta_{jk}\, g_{jk}(X).$$

When to use: spatial data.; surface; small p, large n.

3.2. Generalized Additive Models

$$E(Y|X) = \beta_0 + f_1(X_1) + \cdots + f_p(X_p).$$

f_j - univariate smooth function.

Pros:
- Allow non-linearity.
- Because of additive nature, can examine the effect of each X_i. (interpretability)
- smoothness can be summarized by df.

$$df_{total} = 1 + \sum_j df_j$$

Cons: Restricted to "Additive".
(But could add interaction term).

4. Moving to Dictionaries.

Assumption: sparse. expansion. → regularization & sparsity.

Fixed bases. small. predefined set of functions.

Dictionary: more flexible. by using large or overcomplete collection of functions.