Basis Expansion and Regularization

Presented by

Allison Bruce,
Benjamin Han,
Francisco Pereira,
Yanjun Qi

Roadmap

- How to re-introduce non-linearity?
  - Basis expansions: transform the input space.
  - Piecewise regressions: connecting lines.
- Combine both we have the most general function to model anything, but
  - What basis functions to use? (cubic)
  - How to carve up the regions? (smoothing)
- The last automation: how to select smoothing parameters?

Adventure into Non-linearity-dom

- Linear models: $f(X) = \sum_{m=0}^{P} \beta_m X_m,\ X = (X_1, \ldots X_P)$
- Three ways to escape linearity
  - Basis expansions $f(X) = \sum_{m=1}^{M} \beta_m h_m(X),\ h_m : R^P \mapsto R$
  - Carve it up: Piecewise regressions.

$$f(X) = \sum_{m=1}^{R} f_m(X) I_m(X)$$

$$I_m : R^P \mapsto \{0,1\} \text{ indicator function for } [L_m, U_m)$$

  - Do both!

Basis Expansions

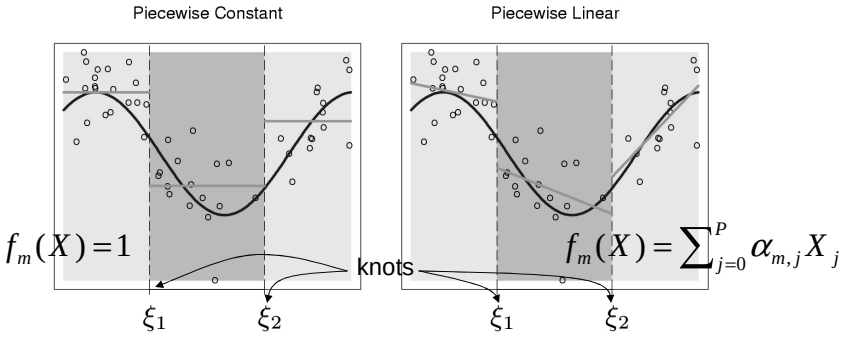
- Samples of h (basis functions)
  - $h_m(X) = X_m$: Vanilla linear models
  - $h_m(X) = X_j^2$ or $X_j X_k$: Higher order expansions – beware of exponential blow-up in # of parameters ($O(p^d)$ for degree d)
  - $h_m(X) = I(L_m \leq X_j < U_m)$: Indicator functions.
  - $h_m(X) = \log(X_j)$ or $\sqrt{X_j}$ or what have you.

Basis Expansions (cont'd)

- How do we select basis functions?
  - **Restriction methods**: pick a selection and stick to it.
  - **Selection methods**: adaptively select a set of functions which contribute most to the fit: CART, MARS, boosting, etc.
  - **Regularization methods**: use all functions we have but shrink the coefficients.
    (natural cubic splines are general enough)

Carve it up: Piecewise Regressions

- Basic idea: fit one function per region; i.e., fit $f(X) = \sum_{m=1}^{R} f_m(X) I_m(X)$

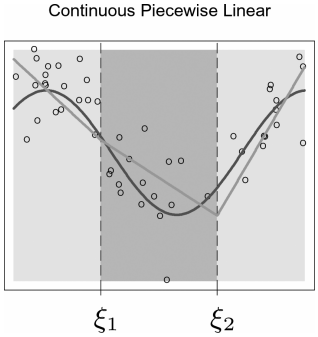


Carve it up: Piecewise Regressions (cont'd)

- Question (Weihao): In the varieties of spline methods, we have to pick up "knots". In the textbook, knots are usually selected uniformly or from quartiles. What are possible reasons to justify the selection, or are they just the best we can do without further information?
  - Answer: There're two ways mentioned in the book – the uniform way and the percentile way. The latter makes more sense since this gives the same "data support" to each region.

Join the Lines: Knot Constraints

- But we like continuities at the knots
  - Extrapolating boundary points can make use of the points in the adjacent regions
    (that's why we run into trouble in the boundary regions)
  - Looks nicer!



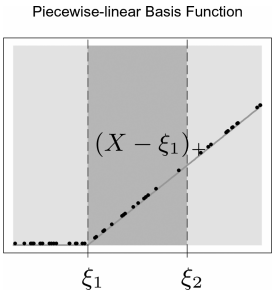
Join the Lines: Knot Constraints (cont'd)

- Knot constraints (linear with one-dimensional input)
  - $f(\xi_1^-) = f(\xi_1^+)$ implies $\beta_1 + \xi_1\beta_4 = \beta_2 + \xi_1\beta_5$ decreases df (degree of freedom) by 1
  - In practice we use

    $h_1(X) = 1, h_2(X) = X,$

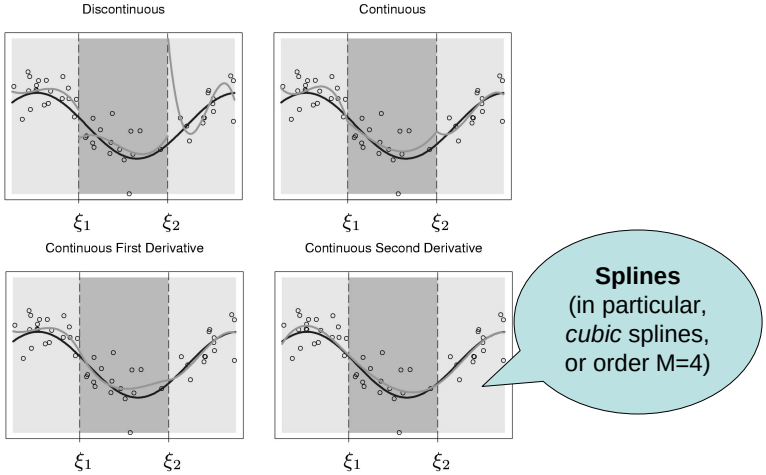
    $h_3(X) = (X - \xi_1)_+, h_4(X) = (X - \xi_2)_+$

    (verify the constraints!)



Spice it up: Higher-order Splines

- Continuity of derivatives



Spice it up: Higher-order Splines (cont'd)

- But why cubic?
  - It's the lowest degree where we still get *inflection* points!
- But why insist $C^2[a,b]$?
  - It pleases us (or our eyes).

Spice it up: Higher-order Splines (cont'd)

- Knot constraints (order M, k knots, with one-dimensional input)

M basis functions for the "basic form"

$$h_1(X)=1, h_2(X)=X, h_3(X)=X^2, \ldots, h_M(X)=X^{M-1}$$

$$h_{M+1}(X)=(X-\xi_1)_+^{M-1}, \ldots, h_{M+K}(X)=(X-\xi_K)_+^{M-1}$$

K basis functions for the knot constraints

$df = K + M$ (verify from another angle: $(K+1)\times M - K\times(M-1) = K+M$)

↑ # of region

Spice it up: Higher-order Splines (cont'd)

- Question (Ashish): How does this basis represent high degree continuity, nothing seems to be added from the basic connection connectivity, or it is something that just happens to be the case with basis of this form?
  - Answer: up to (M-2) order derivative of the M basis functions

    $$h_{M+i}(X) = (X - \xi_i)_+^{M-1}$$

    contain power of $(X - \xi_i)_+$ thus

    $$h_{M+i}^{(j)}(\xi_i) = 0,\ j = 0...M-2$$

    Another hint: $(K+1) \times M - K \times (M-1) = K + M$

    (what's the 2nd term on the left?)

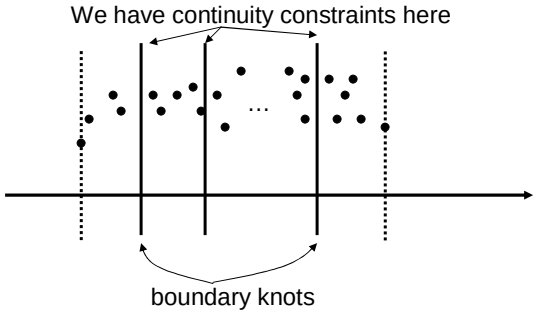
  - So the general form (for order M) $\in C^{M-2}[a,b]$

Join the Lines: Knot Constraints (cont'd)

- Question (Joy): It is said "a more direct way" to add the constraint is to use two additional basis functions h3 and h4. It is not straightforward to me how to choose such functions, esp. in p119, the cubic spline case, can h5 and h6 be (X-ksi)^2?
  - Answer: See the previous slide. And no – to be in C^2[a,b] we need to make the constraint bases 'disappear' up to the 2nd order derivative at the knots, so we have to have $(X - \xi_i)_+^3$ instead of $(X - \xi_i)_+^2$

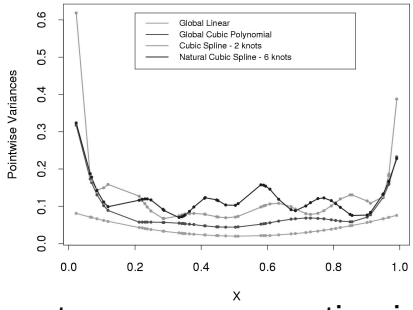
The Fringe of Reality

- It's perfect, until we enter the two boundary regions
  - No continuity constraints naturally emerge there!



The Fringe of Reality (cont'd)

- Symptoms: great variances on the sides



- Solution: act more conservative in the boundary regions

The Fringe of Reality (cont'd)

- Question (Jie): P121, Figure 5.3, are the bumps for the curves of cubic spline and natural cubic spline related to the positions of the knots? Any interpretation to the bumps (why and when they happen)?

Natural Splines

- "Functions are linear beyond the boundary knots"
  - Each boundary region needs 2 parameters

    $$df = (K-1)\times M + 2\times 2 = K - M + 4$$

  - For M=4

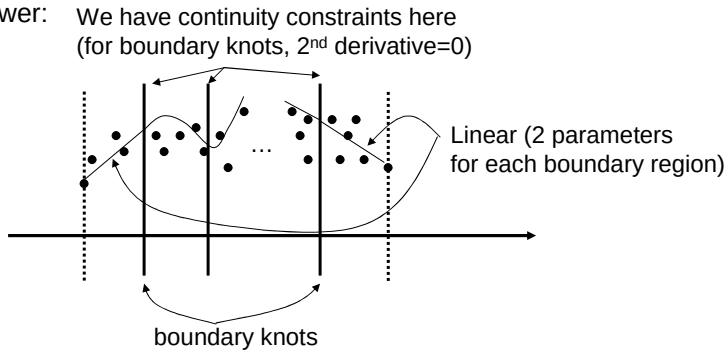
    $$N_1(X) = 1, N_2(X) = X, N_{k+2}(X) = d_k(X) - d_{K-1}(X)$$

    $$d_k(X) = \frac{(X-\xi_k)_+^3 - (X-\xi_K)_+^3}{\xi_K - \xi_k}$$

    $$df = K$$

Natural Splines (cont'd)

- Question (Jie): Could you give a specific form on natural boundary conditions for natural cubic splines (i.e. linear beyond the boundary knots)? What are the formula for the constraints?
  - Answer:



Example: South African Heart Disease

- Problem: predicting the posterior probability of having the disease

$$\text{logit}[\Pr(chd \mid X)] = \theta_0 + \sum_{i=1}^{6} h_i(X_i)^T \theta_i$$

$$X = (X_1,...,X_6) = (X_{sbp}, X_{tobacco}, X_{ldl}, X_{famhist}, X_{obesity}, X_{age})$$

Example: South African Heart Disease (cont'd)

- Suspecting non-linearity in each $h_i(X_i)$, we fit each one with natural splines with df=4
  - $h_i(X_i) = \sum_{j=1}^{4} h_{i,j}(X_i)\theta_{i,j}$ (except for famhist, which only has one h)
  - For each data point X, we have $\left(1, h_{1,1}(X_1),...,h_{1,4}(X_1),...,h_{6,1}(X_6),...,h_{6,4}(X_6)\right)^T_{(df=1+\sum_{i=1}^{6} df_i)\times 1}$ putting N observations together we have $H_{N\times df}$, we can then use the ordinary linear regression method

Example: South African Heart Disease (cont'd)

- Term selection: using AIC (Ch. 7) to drop terms and keep the ones giving smallest AIC
- Result: retain all the terms
  - compared to 4.4.2 where `sbp` and `obesity` are dropped, using logistic regression