All You Wanted To Know About Splines

Boyi Guo

Department of Biostatistics
University of Alabama at Birmingham

Feb 16, 2021

Who Am I?

Who Am I?

- 4th-year Ph.D. student in BST @ UAB
- Dissertation: Bayesian high-dimensional additive model
- Background:
  - B.S. in Computer Science & Stat @ UIUC
  - M.S. in Statistics (Analytic track) @ UIUC
  - Experienced R programmer (8-year)
  - "Ridiculously awesome" commented by a REGARDS collaborator

Overview

Overview

- Spline
  - Explanation & Demonstration
  - Implementation
  - Inferences
- Advance Topics
  - Penalized spline
  - Multivariate spline

Objectives

Objectives

- To review the basic concepts of spline
- To raise you awareness of other advanced spline applications

Spline

Motivation

"It is extremely unlikely that the true function f(X) is actually linear in X."

— Hastie, Tibshirani, and Friedman (2009) PP. 139

Previous Solutions:

- Variable categorization: e.g. using quartiles of a continuous variable in a model
  - Assuming all subjects within a group shares the same risk/effect
  - Different magnitude of effects around thresholds
- Polynomial regression:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m + \epsilon$$

  - Precision issues, e.g. x is blood pressure measure, and x^3 would be extremely large
  - Goodness of fit: deciding which order of polynomial term should be included

Spline

- A spline is a piece-wise function where each piece is a polynomial function of order m
- AKA semi-parametric regression, non-parametric regression, (generalized) additive model
- Can be easily incorporated in linear regression, generalized linear regression, Cox regression

Spline Components

- Order/degree of the polynomial function, m
  - Normally, $m = 3$ is sufficient
- Number of knots, k, & their placements
  - By default, equally spaced
- Basis functions:
  - different representations of the spline that have specific mathematical properties
  - e.g. b-splines basis, Gaussian radial basis, etc.

Spline Example

A spline of order 0 with 2 knots

$$\hat{y} = \begin{cases} 2, & x \leq 1 \\ 1.2, & 1 < x \leq 5 \\ 1.5, & x > 5 \end{cases}$$

Visual Demonstration

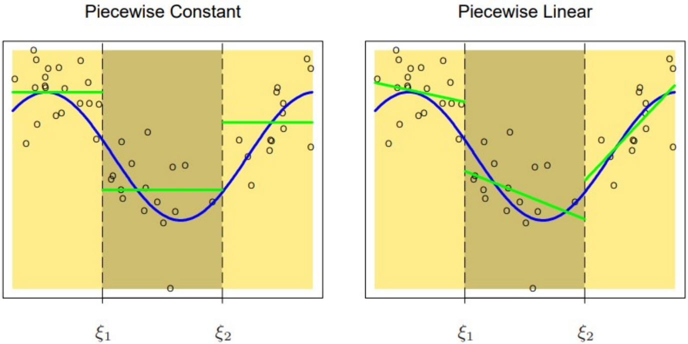

Figure from Hastie, Tibshirani, and Friedman (2009) PP.142

Natural Cubic Spline

- Cubic polynomial in each piece-wise function, i.e. m = 3
- Great mathematical properties
  - The smoothest possible interpolant
- Many different representations:
  - Restricted cubic spline
  - Cubic B-spline

Natural Cubic Spline

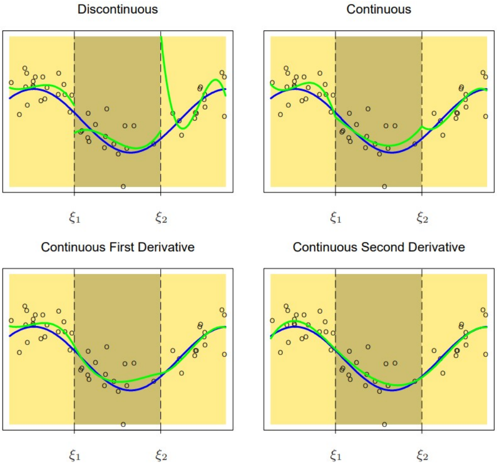

Figure from Hastie, Tibshirani, and Friedman (2009) PP.143

Implementation

Given a response Y and a variable X, implementing a cubic B-spline with 5 knots

```
library(splines)  # Package for b-spline

x_spline <- bs(x, degree = 3, # cubic polynomial
                  df = 8)    # 5 (df-degree) knots
glm(y ~ x_spline) # Fitting the spline model

# Equivalently
glm(y ~ bs(x, degree=3, df=8))
```

Variability Band

- A delicate statistical problem
- Most commonly used: point-wise 95% confidence interval
- Can be calculate using statistical contrasts

Hypothesis Testing

- Two hypothesis tests
  - If the non-linear terms are necessary
  - If the variable is necessary in the model
- Be careful when reading program manual

Advanced Topics

Penalized Spline

- Motivation:
  - To simplify the decision making about the knots
- Idea:
  - Set the number of knots to a really large value (k=25, 40, N)
  - Use variable selection methods, penalized models specifically, to decide the smoothness of the spline
- Complication & extension:
  - Complicated hypothesis testing and interval inference
  - Bayesian generalized additive model, generalized additive mixed model

Multivariate Splines

- Model the non-linear interaction between two variables
- Thin-plate regression splines, tensor product splines
- Application:
  - Loop, M. S., Howard, G., de Los Campos, G., Al-Hamdan, M. Z., Safford, M. M., Levitan, E. B., & McClure, L. A. (2017). Heat maps of hypertension, diabetes mellitus, and smoking in the continental United States. **Circulation: Cardiovascular Quality and Outcomes, 10(1), e003350.**

Conclusion

Conclusion

- Reviewed concepts of spline
- New insight of advanced spline models
- **Consult with statisticians when feeling not comfortable doing spline models**

Q & A

Q & A

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Loop, Matthew Shane, George Howard, Gustavo de Los Campos, Mohammad Z Al-Hamdan, Monika M Safford, Emily B Levitan, and Leslie A McClure. 2017. "Heat Maps of Hypertension, Diabetes Mellitus, and Smoking in the Continental United States." *Circulation: Cardiovascular Quality and Outcomes* 10 (1): e003350.