Three Levels of Spline Models:

Understanding, Application and Beyond

Boyi Guo

Department of Biostatistics
University of Alabama at Birmingham

April 8th, 2021

Who Am I?

Who Am I?

- 4th-year Ph.D. student in BST @ UAB
- Dissertation: Bayesian high-dimensional additive models
- Background:
  - Balanced methodology & collaboration
  - Experienced R programmer & package creator
- Graduate in about 1 year, Looking for
  - Faculty postion in Biostat
  - Post-doc in methodology dev. on HD, causal inference

Overview

Overview

- Understanding
  - Spline Concepts
  - Regression Splines
- Application
  - Non-linear Effect Modifier
  - Non-proportional Hazard Models
  - Generalized Additive Mixed Model
- Beyond
  - Spline Surface
  - Smoothing Splines
  - Function Selection in High Dimension

Objectives

Objectives

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

Disclaimer

- Minimum level of theoretical justification
- No discussion on model fitting algorithms or software implementations

Understanding

Motivation

"It is extremely unlikely that the true (effect) function f(X) (on the outcome) 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
  - Assume all subjects within a group shares the same risk/effect
  - Loss of data fidelity
- 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
- A.k.a. non-parametric regression, semi-parametric regression, (generalized) additive model
- Can be easily incorporated in linear regression, generalized linear regression, Cox regression, as **regression splines**

Spline Components

- Order/degree of the polynomial function, m
  - Normally, $m = 3$, i.e. cubic spline is sufficient
- An increasing breakpoints sequence τ
  - a.k.a. knots, where the piece-wise functions joint
  - e.g. $k \equiv |\tau| = 5$, equally spaced
- Continuity conditions at knots, v
  - to control the smoothness between pieces
  - e.g. continuous at second derivative for cubic spline

Toy Example

A spline function of the variable X, $f(X)$, of order $m = 0$ with $k = 2$ knots $(\tau_1 = 1, \tau_2 = 5)$ and no continuity condition

$$f(X) = \begin{cases} 2, & X \leq 1 \\ 1.2, & 1 < X \leq 5 \\ 1.5, & X > 5 \end{cases}$$

Visual Demonstration

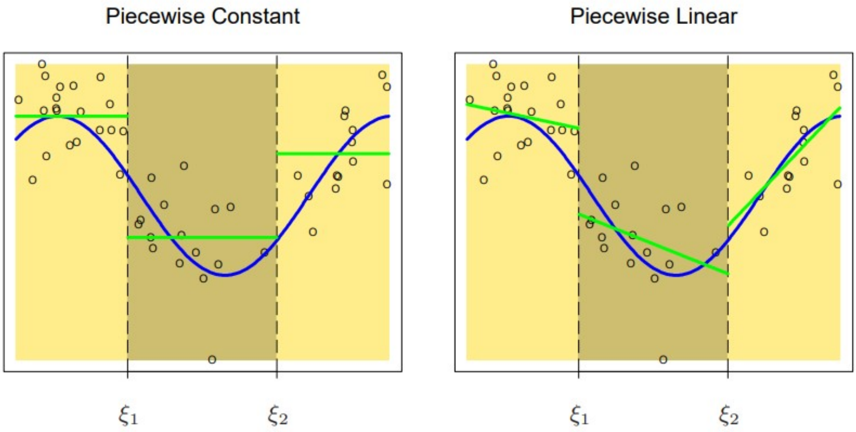


Figure from Hastie, Tibshirani, and Friedman (2009) PP.142

Cubic Spline

- Cubic polynomial in each piece-wise function, i.e. m = 3
  - E.g. truncated power bases with 3 knots at τ_1, τ_2, τ_3

$$
\begin{aligned}
f(X) &= \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X - \tau_1)_+^3 \\
&+ \beta_5 (X - \tau_2)_+^3 + \beta_6 (X - \tau_3)_+^3 \\
&= \beta^T \boldsymbol{B}(X)
\end{aligned}
$$

- Continuous at second derivative
  - The smoothest possible interpolant
- Alternative representation
  - B-spline bases for stable computation
- Natural cubic spline for linearity beyond boundary knots ($f^{'''}(X) = 0$)

Cubic Spline

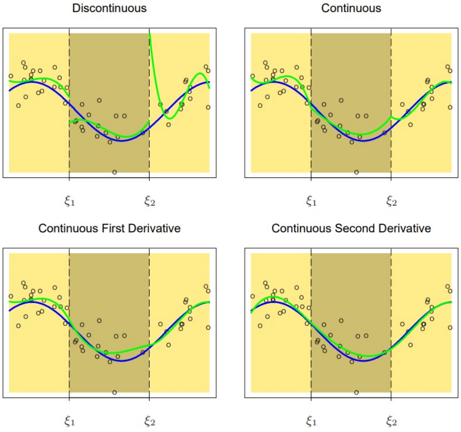


Figure from Hastie, Tibshirani, and Friedman (2009) PP.143

Regression Splines

Given the matrix form of the spline function $f(X) = \beta^T \boldsymbol{B}(X)$,

- Linear regression:

$$y_i \sim N(\beta^T \boldsymbol{B}(X_i) + \beta_{cov}^T \boldsymbol{Z}_i, \sigma^2)$$

- Generalized linear regression:

$$E(y_i) = g^{-1}(\beta^T \boldsymbol{B}(X_i) + \beta_{cov}^T \boldsymbol{Z}_i), Y_i \sim EF$$

- Cox regression:

$$h(t_i) = h_0(t_i) exp(\beta^T \boldsymbol{B}(X_i) + \beta_{cov}^T \boldsymbol{Z}_i)$$

Model fitting and diagnostic remain the same

Software Implementation

Two-step procedure

- Create the ‘design’ matrix of the spline function $B(X)$
- Fit the preferred model including $B(X)$ as covariates / predictors

```
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
  - Confidence about spline functions VS point estimates
- Most commonly used: 95% point-wise confidence interval
- Can be calculate using statistical contrasts for regression splines

Hypothesis Testing

- Two hypothesis tests
  - If the non-linear terms are necessary:

$$H_0 : \beta_2 = \beta_3 = \cdots = 0$$

  - If the variable is necessary in the model

$$H_0 : f(x) = 0$$

- Be careful when reading program manual

Rule of Thumb

- Cubic splines for smooth interpolant
  - B-spline for computation stability
  - 3-5 equally spaced knots
- Transform variables with extreme values for computational stability
  - e.g. prefer $f(\log(X))$ over $f(X)$ when modeling CRP
- Examine outlier's effect on statistically significant non-linear relationship
- Survival Model
  - Knots are decided by equal number of events in each group
  - Defer to Sleeper and Harrington (1990) for practical guidance

Application

Varying Coefficient

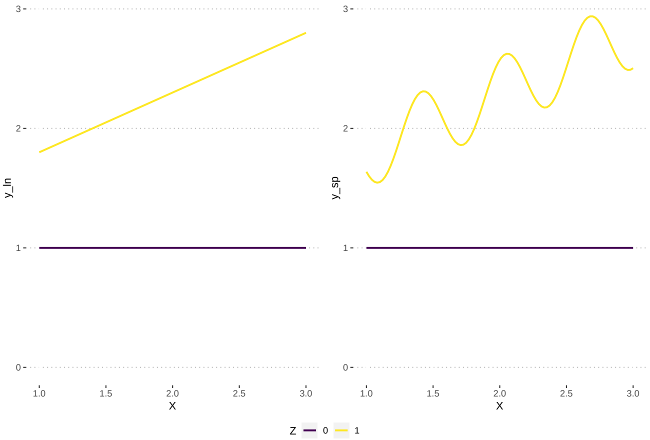
To model a non-constant effect of the variable Z as a function of another variable X

$$E(y) = f(X)Z,$$

where $f(X)$ is the varying coefficient of Z

- Example: statistical interaction βXZ where $f(X) = \beta X$
- What if the slope of the effect are not constant across the domain of X?

Non-linear Effect Modification



Non-linear Effect Modification

$$E(y) = f(X) + f'(X)Z = \beta^T_{Z=0}B(X) + \beta^T_{Z=1}B(X) * Z$$

- $f(X)$ models the effect of X when $Z = 0$
- $f'(X)Z$ models the modifying effect of Z at different values of X
- $f'(X)$ is the varying coefficient of Z, using a non-linear function, for non-constant slope.

Non-linear Effect Modification

- Assumptions of consideration
  - Should $f(X)$ be linear or non-linear?
  - Should $f(X)$ use the same bases as $f'(X)$?
  - Should $f(X)$ be the same level of complexity as $f'(X)$?

Non-proportional Hazard

- Cox PH model assumes proportional hazards, i.e. the hazard/effect of a variable X is independent to time
- Using Time-varying coefficients to model the non-proportional hazards

$$h(t) = h_0(t)exp(f(t)X)$$

- Defer to Gray (1992) and references therein

Mixed Model

To model the non-linear fixed effect while considering random effects

- Good for longitudinal studies or multi-center studies
- Easy to implement: to include your design matrix of $\boldsymbol{B}(X)$ in the fixed effect
- `gamm` in R-package `mgcv`

Beyond

Spline Surface

- Model the non-linear interaction between two continuous variables
- Thin-plate splines, tensor product splines
  - Thin-plate spline is scale-sensitive
    - Recommended when variables are on the same scale
  - Tensor product spline is scale-invariant
- Dealing with *over smoothing across boundary*
  - Soap film smoothing
- 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.**

Smoothing 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

Objective Functions

Given a spline model $y \sim N(f(X), \sigma^2)$

- Regression spline

$$\arg\min_{\beta} \sum_{i=1}^{n} \{y_i - \beta^T B(X_i)\}^2$$

- Smoothing spline

$$\arg\min_{\beta} \sum_{i=1}^{n} \{y_i - \beta^T B(X_i)\}^2 + \lambda \int f''(X)^2 dx$$

- λ is a tuning parameter, selected via (generalized) cross-validation

Statistical Complications

- Estimated degree of freedom due to shrinkage
  - Harder to conduct hypothesis testing, and calculate CI
- More decisions when modeling effect modification
  - Same smoothness for the spline functions?
  - If the same, how to estimate the smoothness

Function Selection

- Question of interest
  - If a variable X has effect on the outcome Y
  - High-dimensional data analysis, e.g. EHR, Genomics
- Solutions
  - Step-wise function selection
    - Locally optimal solution
    - Not feasible for high-dimensional analysis
  - Group penalized models
    - Biased estimation
    - Global penalization vs local penalization
  - Bayesian hierarchical models
    - Robust estimation
    - Slow...

Conclusion

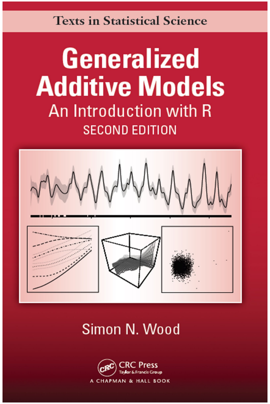
Conclusion

- Reviewed concepts of spline
- New insight of advanced spline models
- Same set of variables can lead to many models with different assumptions
  - Fit many models and compare
  - Explore the inconsistency
- Balance between interpolation and prediction
  - "Black box" models for improved prediction
- **Consult with statisticians when not comfortable dealing spline models**

Great Book

Wood, S. N. (2017). Generalized additive models: an introduction with R. CRC press.

- Chapter 7 for examples



Q & A

Reference

Reference

Gray, Robert J. 1992. "Flexible Methods for Analyzing Survival Data Using Splines, with Applications to Breast Cancer Prognosis." *Journal of the American Statistical Association* 87 (420): 942–51. https://doi.org/10.1080/01621459.1992.10476248.

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.

Sleeper, Lynn A., and David P. Harrington. 1990. "Regression Splines in the Cox Model with Application to Covariate Effects in Liver Disease." *Journal of the American Statistical Association* 85 (412): 941–49.