Pliable regression splines with an auxiliary variable

Jae-Kwon Oh

Department of Information Statistics
Chungbuk National University

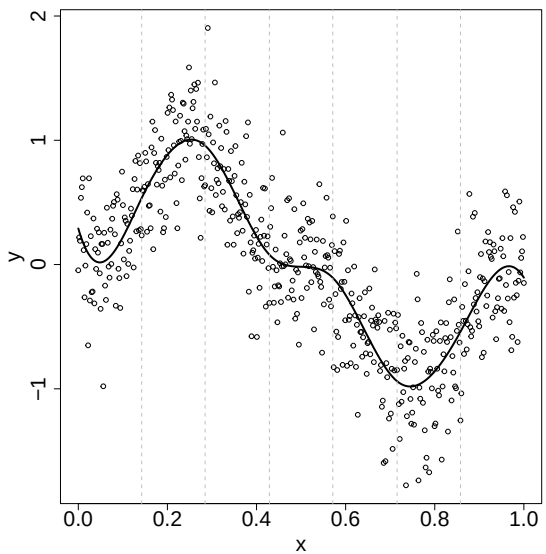
2021. 5. 26.

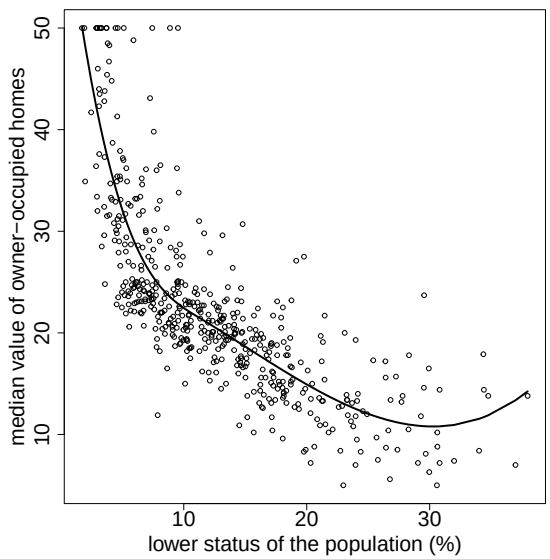
Contents

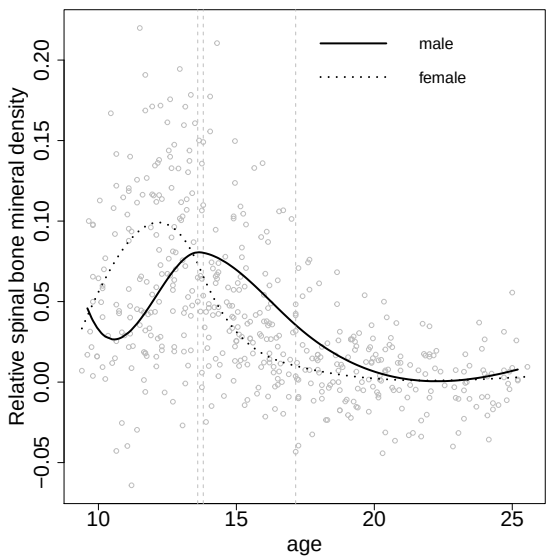
Introduction

Nonparametric function estimation

- 실 데이터에서 나타나는 다양한 현상을 함수로 모형화
- 함수를 무한차원의 모수공간에 속한다고 가정
- 관찰된 데이터를 기반으로 함수를 추정하는 것이 목표
- 대표적인 방법 : Kernel Density Estimation, Local Polynomial, Spline 등







Basis function methodology

- 유한한 데이터로 무한 차원의 모수를 추정하는 것은 불가능
- 연구 특성에 맞게 추정대상 함수 f의 형태를 제한
- 제한한 함수공간 $\mathcal{F}$를 생성하는 기저함수를 고려
- 추정대상 함수 f를 기저함수들의 선형결합으로 표현

$$f(x) = \sum_{j=1}^{J} \beta_j B_j(x), \quad \text{if } f \in \mathcal{F}$$

- $B_1, \ldots, B_j$는 함수공간 $\mathcal{F}$를 생성하는 기저함수
- 통계적 방법론들을 적용할 수 있다는 장점을 가진다. (ex. 최소제곱법, 가능도방법 등)

Spline

- 각 매듭 (knots) 구간에 조각별 다항식 (piecewise polynomial) 으로 정의된다.

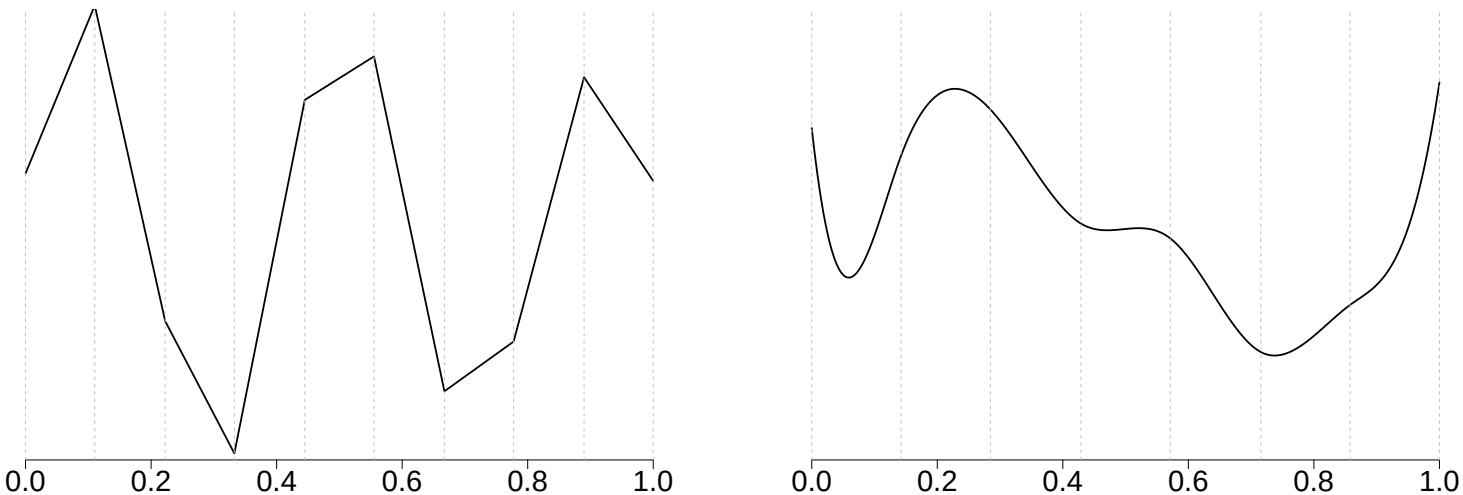


그림: Piecewise linear (left) and cubic (right) splines

Spline Basis

- 스플라인 기저로 스플라인 표현 가능

$$f(x) = \sum_{j=1}^{J} \beta_j B_j(x) \quad \text{for} \quad \beta \in \mathbb{R}^J$$

- 대표적인 기저함수
  - Truncated power spline basis
  - B-spline basis

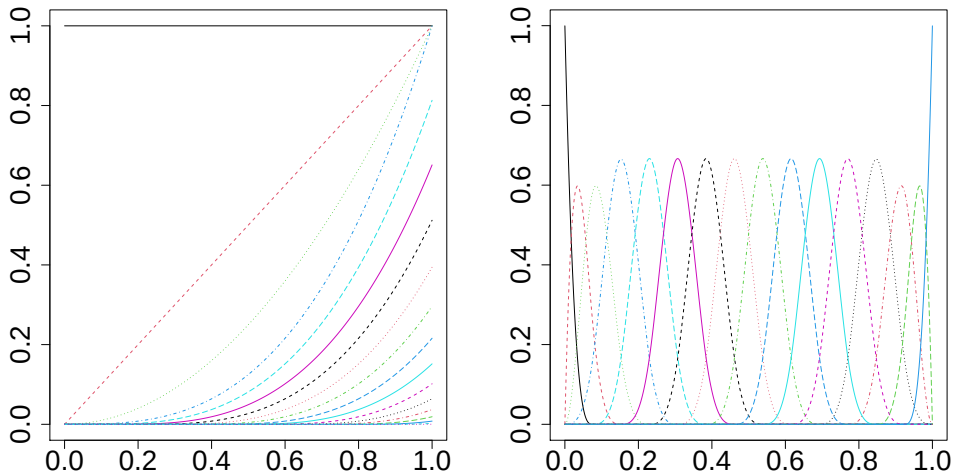


그림: Truncated power spline basis (left) and B-spline basis (right)

Objective function and Estimator

- Objective function : Residual sum of squares

$$R(\beta) = \sum_{i=1}^{n} \left\{ y - \sum_{j=1}^{J} \beta_j B_j(x_i) \right\}^2$$

- Estimator of β

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^J}{\mathit{argmin}}\ R(\beta)$$

- Function estimator

$$\hat{f}(\cdot) = \sum_{j=1}^{J} \hat{\beta}_j B_j(\cdot)$$

Coordinate descent algorithm

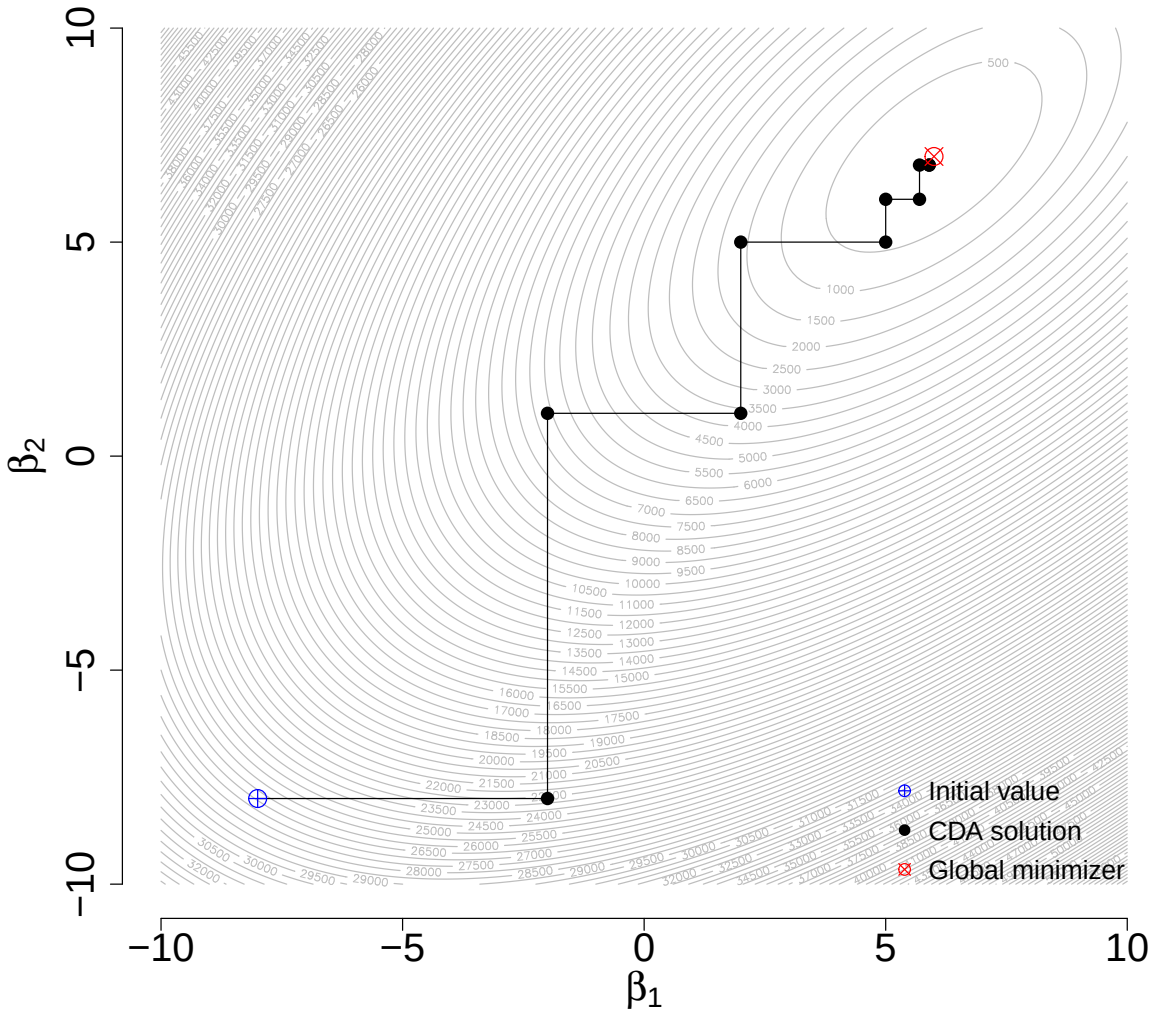
- 다차원 벡터에 대한 볼록 (오목) 함수 최소화 (최대화) 문제의 해를 구하는 과정
- 하나의 계수를 갱신할때, 나머지 계수들은 상수로 고정
  $\Rightarrow$ 해당 계수에 대한 일차원 함수로 간주
- 초기 계수 벡터

$$\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_J)$$

- Update β_j

$$\tilde{\beta}_j \leftarrow \underset{\beta_j \in \mathbb{R}}{\textit{argmin}}\ R(\tilde{\beta}_1, \ldots, \tilde{\beta}_{j-1}, \beta_j, \tilde{\beta}_{j+1}, \ldots, \tilde{\beta}_J)$$

Coordinate descent algorithm



Knot selection

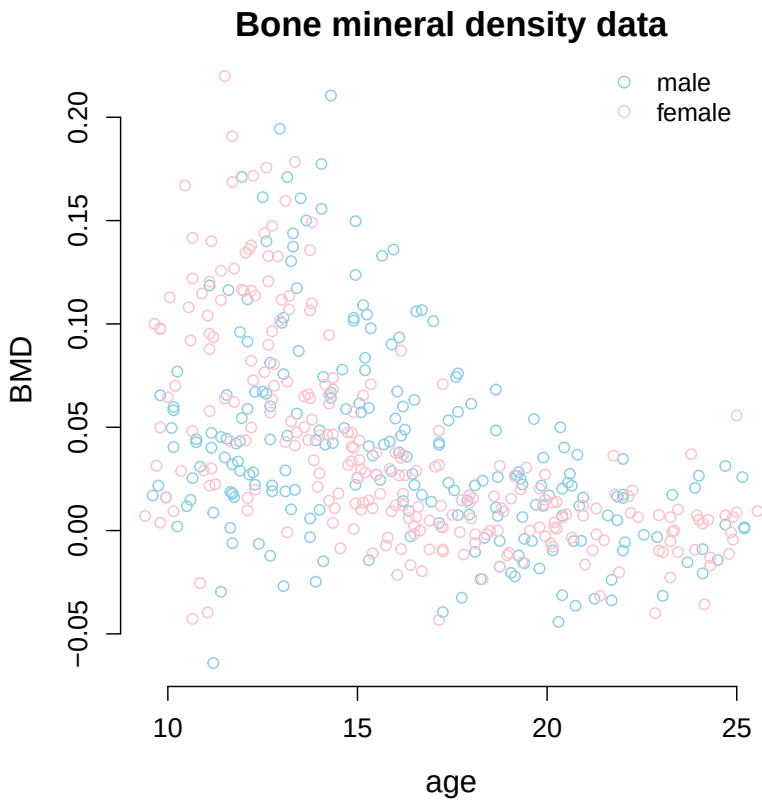
- 스플라인에서의 매듭 선택은 회귀분석에서의 변수 선택과 동일하다.
- 변수 선택 방법들을 적용가능하다.
- 변수 선택 방법중 Stepwise selection 을 도입
- 모델 평가 기준 (criterion)으로 Bayesian Information Criterion (BIC) 를 사용

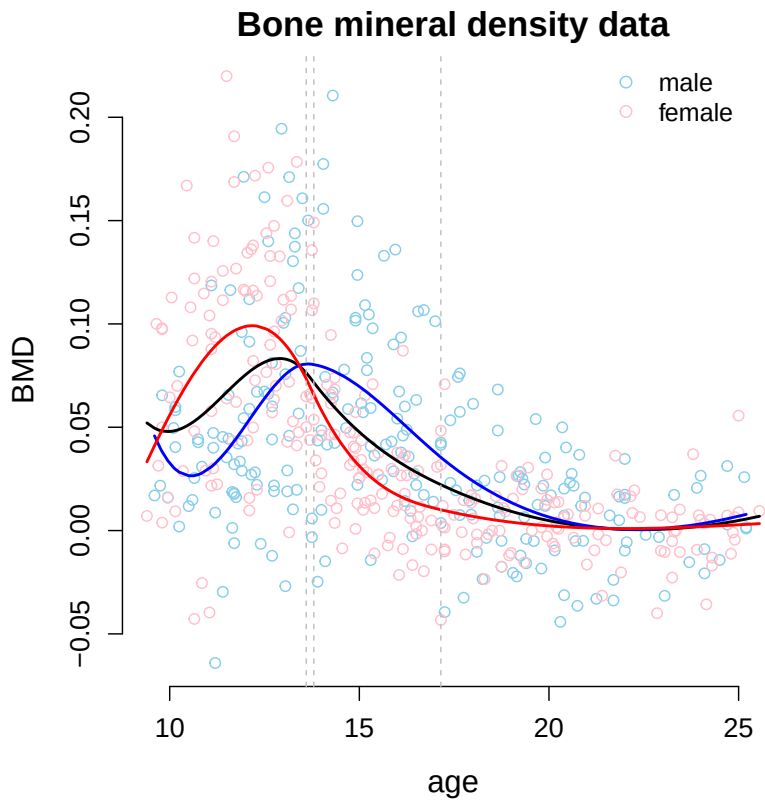
$$BIC = n \log\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathrm{f}(x_i))^2\right) + p\log(n)$$

Pliable regression splines

Bone mineral density data

- 청소년과 젊은 성인들의 골밀도 (bone mineral density) 데이터





Nonparametric regression model

- Data :

$$(x_1, z_1, y_1), \ldots, (x_n, z_n, y_n)$$

- Model :

$$y_i = f(x_i, z_i) + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, n,$$

where $y_i \in \mathbb{R}$, $x_i \in [0, 1]$, $z_i = (z_{i1}, \ldots, z_{iK}) \in \mathbb{R}^K$, $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) > 0$.

- Goal :

Estimate f based on the given data.

Objective function

- Pliable Spline Estimator (PSE) :

  For $x \in [0,1]$ and the binary vector $z = (z_1, \ldots, z_K)$ with length K, define

$$f(x, z; \theta) = \sum_{j=1}^{J} \beta_j B_j(x) + \sum_{j=1}^{J} \sum_{k=1}^{K} \gamma_{jk} z_k B_j(x),$$

  where $\theta = (\beta, \gamma_1, \ldots, \gamma_J)$ is a coefficient vector with
  $\beta = (\beta_1, \ldots, \beta_J) \in \mathbb{R}^J$ and $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jK}) \in \mathbb{R}^K$ for $j = 1, \ldots, J$.

- Residual sum of squares objective function :

$$R(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i, z_i; \theta))^2$$

Coordinate descent algorithm

- Univariate objective function of β_j and γ_{jk} :

$$r_j(\beta_j) = R(\tilde{\beta}^{(-j)}, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_J)$$

and

$$r_{jk}(\gamma_{jk}) = R\left(\tilde{\beta}, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_{j-1}, \tilde{\gamma}_j^{(-k)}, \tilde{\gamma}_{j+1}, \ldots, \tilde{\gamma}_J\right)$$

- Coordinate-wise update :

$$\tilde{\beta}_j \leftarrow \underset{\beta_j \in \mathbb{R}}{\operatorname{argmin}}\, r_j(\beta_j) \quad \text{and} \quad \tilde{\gamma}_{jk} \leftarrow \underset{\gamma_{jk} \in \mathbb{R}}{\operatorname{argmin}}\, r_{jk}(\gamma_{jk}).$$

Update $\theta = (\beta_j, \gamma_j)$ by CDA

- The quadratic form of β_j

$$r_j(\beta_j) = \frac{\sum_{i=1}^n B_j^2(x_i)}{2n} \left(\beta_j - \frac{\sum_{i=1}^n y_{ij} B_j(x_i)}{\sum_{i=1}^n B_j^2(x_i)} \right)^2 + (\text{terms independent for } \beta_j)$$

- Update β_j

$$\tilde{\beta}_j \leftarrow \frac{\sum_{i=1}^n y_{ij} B_j(x_i)}{\sum_{i=1}^n B_j^2(x_i)} \quad \text{for} \quad j = 1, \ldots, J.$$

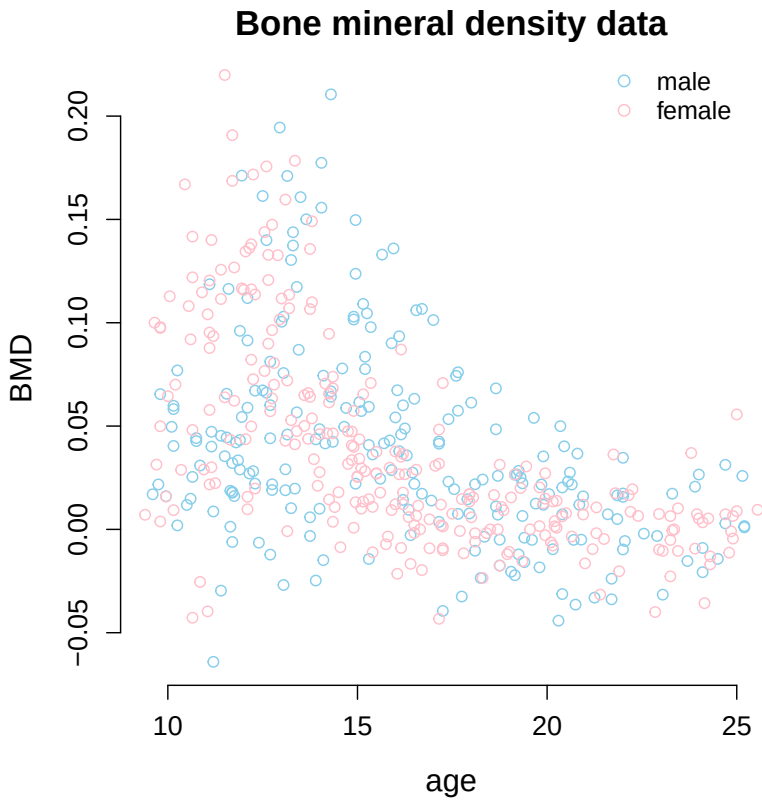
- Similarly, r_{jk} can be expressed as quadratic form of γ_{jk}
- Update γ_{jk}

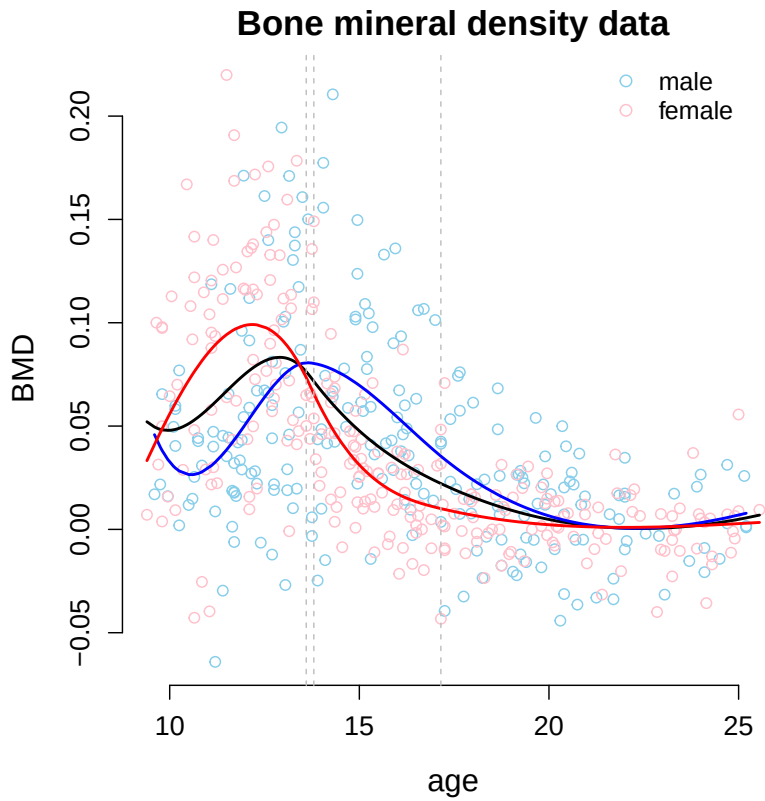
$$\tilde{\gamma}_{jk} \leftarrow \frac{\sum_{i=1}^n y_{ijk} z_{ik} B_j(x_i)}{\sum_{i=1}^n z_{ik}^2 B_j^2(x_i)} \quad \text{for} \quad j = 1, \ldots, J, \ k = 1, \ldots, K.$$

Numerical studies

Bone mineral density data

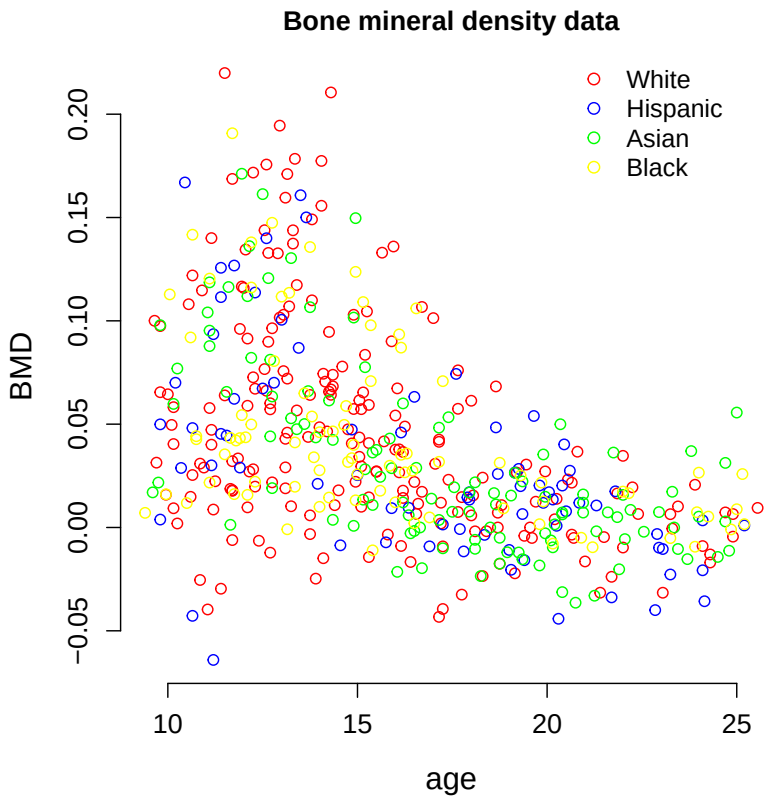
- 성별에 따라 골밀도의 분포가 다름

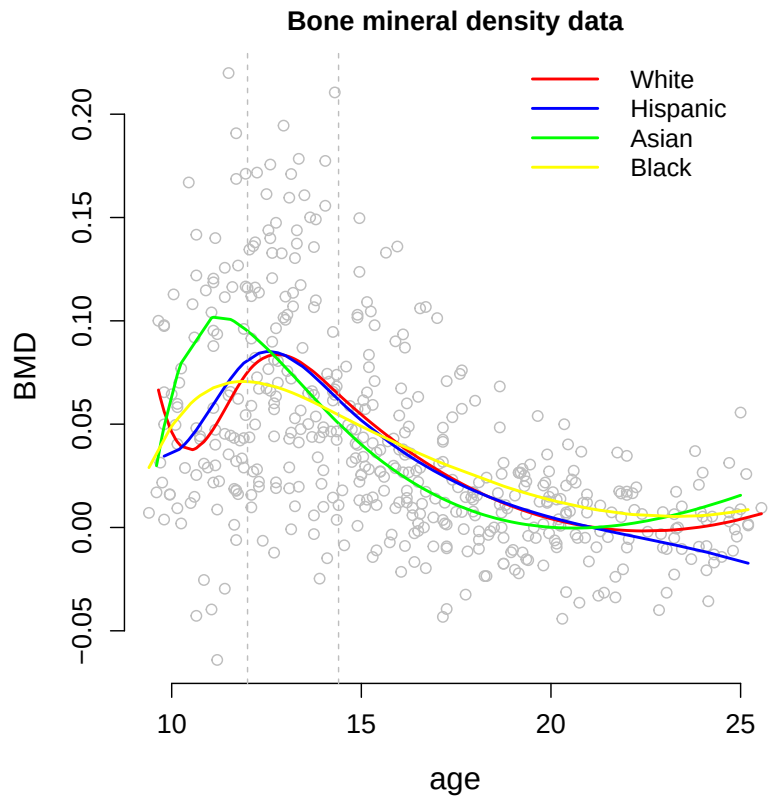


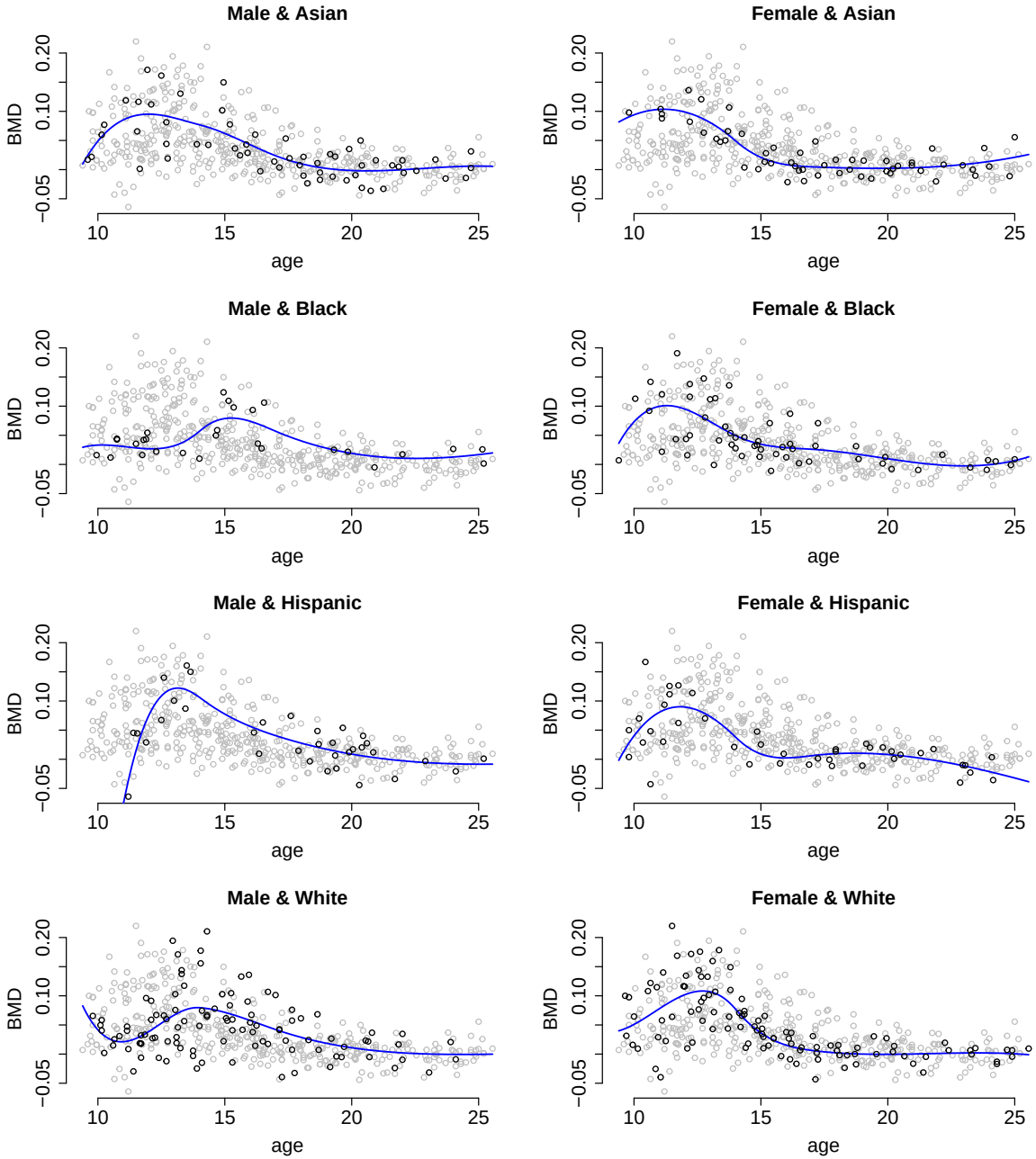


Bone mineral density data

- 인종에 따라 골밀도의 분포가 다름






Male & Asian
Female & Asian
Male & Black
Female & Black
Male & Hispanic
Female & Hispanic
Male & White
Female & White
BMD
−0.05
0.10
0.20
10
15
20
25
age

Conclusion

Conclusion

- 비모수 함수 추정에서 B-spline을 사용하여 예측변수 외에 보조변수 (auxiliary variable) 가 추가된 경우에 적합한 Pliable Spline Estimator 제시
- Coordinate descent algorithm 으로 목적함수를 최소화
- 매듭 선택으로 단계적 선택법을 도입으로 최적의 knots 선택
- 시뮬레이션과 실 데이터 분석을 통해 제안한 모델의 성능 확인