Spline Regression Modeling Using R – Methods and First Results

Matthias Schmid

Department of Medical Biometry, Informatics and Epidemiology
University of Bonn

on behalf of TG2 of the STRATOS Initiative

August 31, 2017

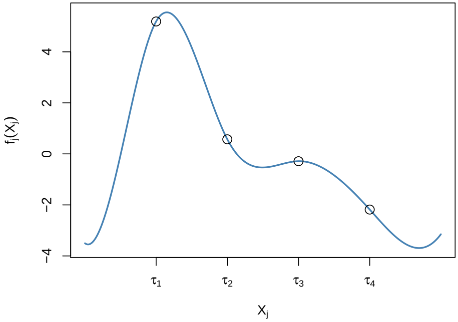
The Subject

- Fit a statistical model of the form $g(Y|X) = \beta_0 + f(X)$
  - p explanatory variables $X = (X_1, \ldots, X_p)$
  - f unknown, allowed to be nonlinear but should be interpretable
- Common specification: $f(X_1, \ldots, X_p) = f_1(X_1) + \ldots + f_p(X_p)$

  $\rightarrow$ Generalized additive models (GAMs)
- Splines are the most popular method to estimate $f_1, \ldots, f_p$
  - GAM books by Hastie/Tibshirani and Wood are hugely popular ($> 14,000$ and $> 6,000$ citations, respectively)

Definition of Splines

- Set of piecewise polynomials, each of degree d
  - Joined together at a set of knots $\tau_1, \ldots, \tau_K$
  - Continuous in value + sufficiently smooth at the knots



TG2 Talk at 2016 CEN Conference, Munich

- Review of spline implementations in R
- Conclusions:

  "Details of spline routines [...] are often not contained in [R] help files + may be difficult to retrieve from literature"

  ''Notable exception: **mgcv**"
- **mgcv** package (Wood, 2017) is arguably the most popular spline modeling package in R
- Accompanies the book "Generalized Additive Models – An Introduction with R" (Wood, 2017, 2nd edition)
- Book + articles referenced in **mgcv** help provide an excellent documentation of the implemented methods

Spline Implementations in **mgcv**

- Simulation study on spline implementations in **mgcv**
- Specification of the desired spline method is done via the `s` function (part of the formula argument that is passed to the `gam` function of **mgcv**)
- Popular types of splines:
  - Thin plate regression splines (argument `s(x, bs = "tp")`)
  - Penalized cubic regression splines (argument `s(x, bs = "cr")`)
  - P-splines (argument `s(x, bs = "ps")`)
- Here, we rely on **mgcv**'s default procedures for knot selection and smoothing parameter optimization

Thin Plate Regression Splines

- Low-rank approximation of thin plate splines
- Knot positions = data locations (with sub-sampling of data locations if n is large)
- Defaults in **mgcv**:
  - Degree 3
  - Estimation with integrated second-order derivative penalty
  - 9 coefficients per smooth term (null space dimension (= 2) plus 8 minus intercept)
  - Optimization of smoothing parameter via GCV

Penalized Cubic Regression Splines

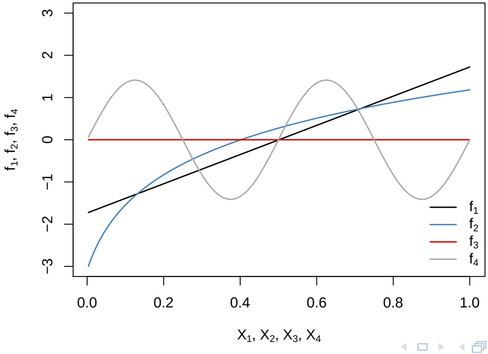
- Natural cubic splines with k knots, integrated second-order derivative penalty
- Based on cardinal spline basis (constructed such that j-th basis function is 1 at the j-th knot and 0 at the other knots, $1 \leq j \leq k$)
- Knots are placed evenly throughout the ordered covariate values
- Defaults in **mgcv**:
  - 10 knots per smooth term (9 coefficients: # knots minus intercept)
  - Optimization of smoothing parameter via GCV

P-Splines

- Polynomial splines, based on B-spline basis
- Integrated squared derivative penalty is approximated by an *m*-th order difference penalty
- Knots are placed evenly throughout the ordered covariate values
- Defaults in **mgcv**:
  - Cubic splines (degree 3) with second-order difference penalty
  - 6 inner knots and 2 boundary knots per smooth term
    (9 coefficients: # inner knots + degree 3 + 1 minus intercept)
  - Optimization of smoothing parameter via GCV

Simulation Design

- Model: $Y = f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) + \epsilon$
- $f_1(X_1) = X_1,\ f_2(X_2) = \log(X_2 + 0.05),\ f_3(X_3) = 0,$ $f_4(X_4) = \sin(4 \cdot \pi \cdot X_4)$

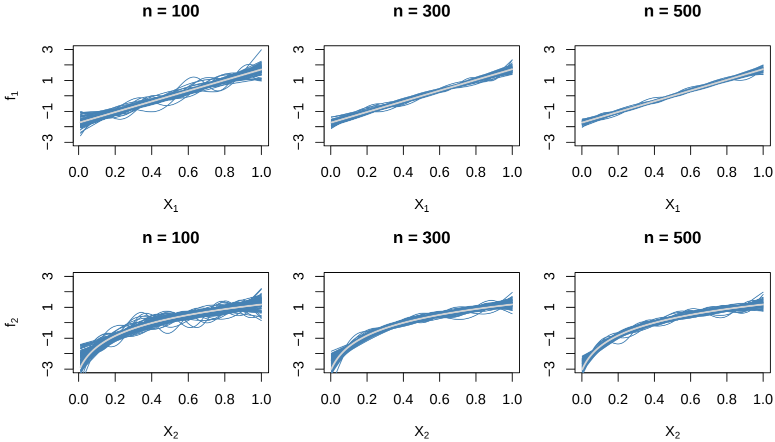


Simulation Design (2)

- 100 simulation runs with sample sizes $n = 100, 300, 500$
- Data values of X_1, X_2, X_3, X_4: independent permutations of $1/n, 2/n, \ldots, n/n$
- Use standardized values of $f_j(X_j)$, $j = 1, 2, 3, 4$
- $\epsilon \sim \mathcal{N}(\sigma^2)$
- σ^2 adjusted such that $R^2 = 0.75$
- For $n = 300$: Additionally investigate $R^2 = 0.25, 0.5$
- Run `gam` with tp, cr and ps implementations (using default procedures)
- Defaults in **mgcv** ensure that all spline bases have the same dimensionality
- Evaluation: covariate-wise mean squared error, $\int_{x_j} (f_j - \hat{f}_j)^2 dP_{x_j}$

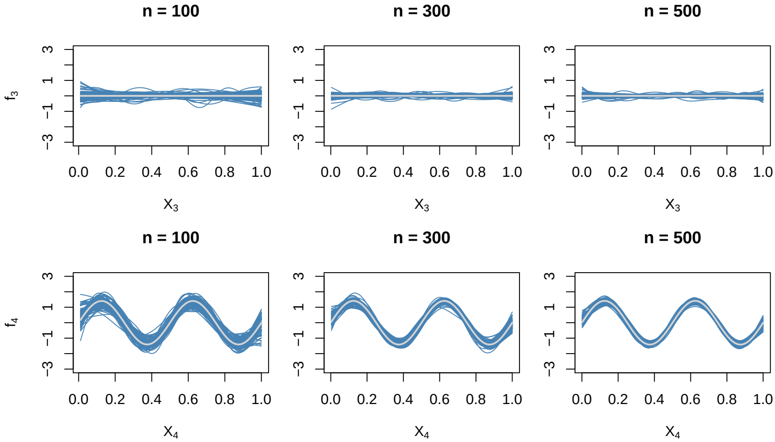
Estimates (1)

tp estimates of f_1 and f_2:



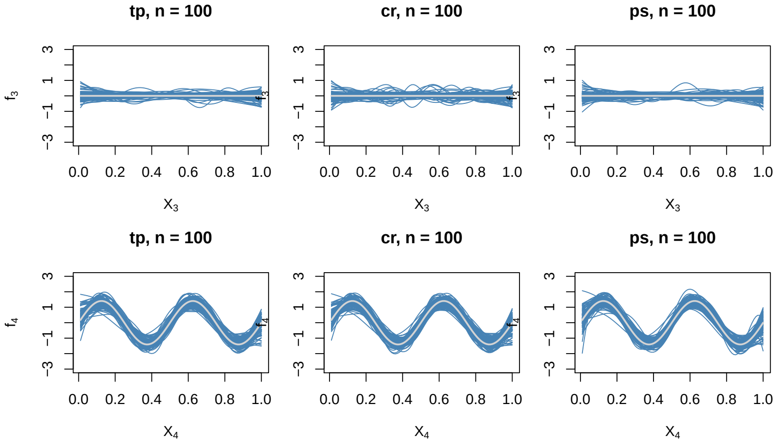
Estimates (2)

tp estimates of f_3 and f_4:



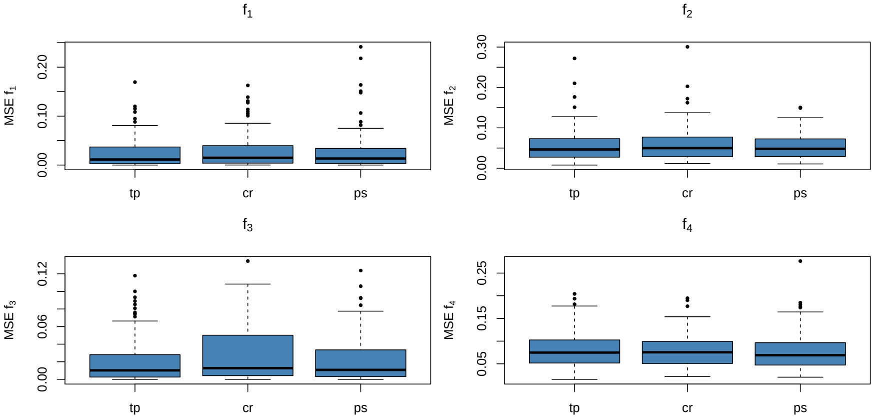
Estimates (3)

tp, cr and ps estimates of f_3 and f_4, $n = 100$:



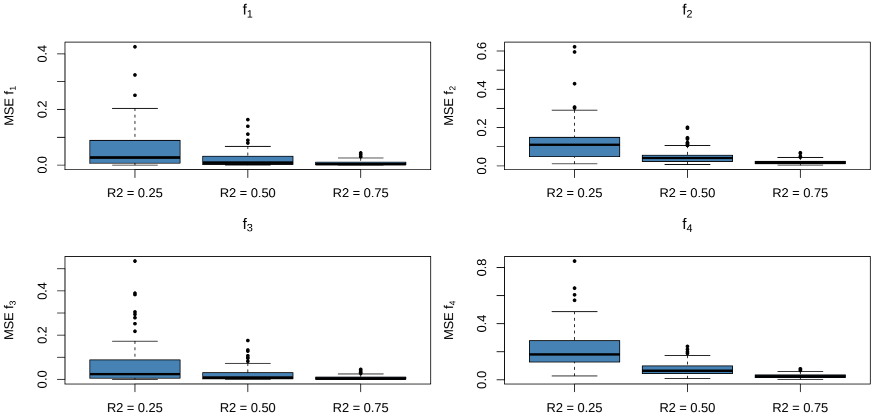
Model Performance (1)

MSE estimates obtained from tp, cr and ps, $n = 100$:



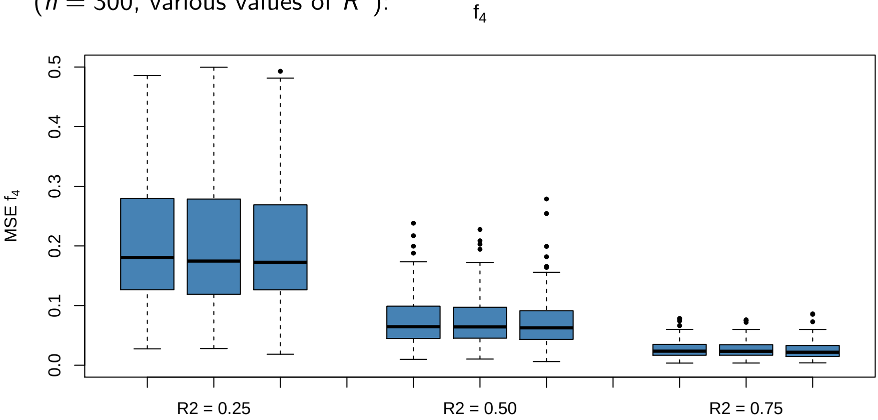
Model Performance (2)

MSE estimates obtained from tp, $n = 300$, various values of R^2:



Model Performance (3)

MSE estimates for f_4, as obtained from tp, cr and ps ($n = 300$, various values of R^2):



Summary of the Simulation Study

- Regression setting with reasonably large sample sizes
- Setting refers to "typical" predictor-response relationships, not too wiggly
- Uncorrelated predictors, no outliers in X

$\Rightarrow$ In this setting, **mgcv** defaults worked well

$\Rightarrow$ Differences between tp, cr and ps appear to be negligible

- Next steps: Correlated predictors, more noise variables, less smooth variable transformations