Generalized Additive Models

- GAMs in a nut shell
- A brief sketch of splines
- Estimating spline based GAMs
- Two important measures
- Consistency

- GAMs are regression models for a random variable Y from the exponential family (*Gaussian*, gamma, Bernoulli, categorial, exponential, beta, ...)
- Extension of a standard linear regression model that allows to model *non-linear functions*
- Tabular dataset: $[[\mathrm{x}^n, y^n]_{n=1}^N]^\top$ where $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$

General form of interpretable GAMs

$$\mathbb{E}[Y|x_1,\ldots,x_p] = \underbrace{\overbrace{f_{1,1}(x_1)+f_{1,2}(x_2)+\ldots}^{\text{univariate}}+\overbrace{f_{2,1}(x_1,x_2)+\ldots}^{\text{bivariate}}}_{\text{non-parametric}} + \underbrace{X\beta}_{\text{parametric}}$$

- $f(\cdot)$ called smoother (non-linear function)
  - non-parametric regression models
  - *splines*
  - deep neural networks
  - regression trees
- parametric part is typically used to model categorical variables
- $f(\cdot)$ and β are estimated from the data
- See [Wood, 2017, Hastie and Tibshirani, 1986, Hastie and Tibshirani, 1990, Wahba, 1990, Green and Silverman, 1993, Riezler and Hagmann, 2021].

- Well known technique from numerical mathematics for function interpolation
- Key Idea: Interpolation is done by piece-wise polynomial functions that connect smoothly at knots to model globally smooth functions

Definition: Spline

A function p:$[\tau_0, \tau_{n-1}) \mapsto \mathbb{R}$ that can be expressed by a polynomial with a degree of at most d for each sub-interval $[\tau_i, \tau_{i+1}]$ of a strictly increasing knot sequence $\tau := [\tau_i]_{i=0,\ldots,n-1}$ is called a piece-wise polynomial function or *spline* on τ of maximum degree d.

The spline space $S_{d,\tau}$

$S_{d,\tau}$ denotes the vector space of all $(d-1)$-times continuously differentiable splines on τ.

Truncated power function

$$(u)_+^d := \begin{cases} 0 & u < 0 \\ u^d & \text{otherwise} \end{cases} \quad \text{with} \quad d \in \mathbb{N}_0$$

Result

For every spline p on τ with maximum degree d exist a unique set of coefficients c_{ij} for $i = 0, \ldots, d$ and $j = 0, \ldots, (n-2)$ such that

$$p(x) = \sum_{j=0}^{n-2} \sum_{i=0}^{d} c_{ij} (x - \tau_j)_+^d$$

The most commonly used splines (natural splines, B-Spline, cubic splines, TP-splines, etc) differ mostly by the chosen base to represent $S_{d,\tau}$.

Functional minimization problem

Let $\mathfrak{H}$ be the class of twice differentiable univariate functions and assume N datapoints:

$$\min_{h \in \mathfrak{H}} \sum_{n=1}^{N} (y^n - h(x^n))^2 + \lambda \int (h''(x))^2 \, dx$$

where $\lambda \in \mathbb{R}^+$ and $\int (h''(x))^2 dx$ is a measure for the roughness of a function over its domain.

Solution: Natural cubic splines with knots at each input x^n

Idea for spline based GAMs

Fix a Basis for $S_{d,\tau}$, transform the input feature x by the base functions and estimate the $c_{i,j}$ from data

Matrix notation of a spline

$$f(\cdot) = \sum_{j=1}^{d} \beta_j b_j(\cdot) = \mathrm{b}(\cdot)\boldsymbol{\beta}$$

$$\text{where} \quad \mathrm{b}(\cdot) = [b_1(\cdot), b_2(\cdot), \ldots, b_d(\cdot)]$$

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_d]^\top$$

Penalized Least Squares Parameter Estimation

Penalized least squares objective

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^s}{\operatorname{argmin}} \left\| \mathrm{Y} - \mathrm{G}\boldsymbol{\beta} \right\|^2 + \sum_{k=1}^{p} \lambda_k \int (f_k''(x))^2 \, dx$$

where $s = \sum_{k=1}^{p} d_k$, $\lambda_k \in \mathbb{R}^+$ and G stores the base function values of the input features.

Useful fact about the roughness penalty

$$\int (f''(x))^2 \, dx = \boldsymbol{\beta}^\top \Omega \boldsymbol{\beta}$$

$$\text{where} \quad \Omega := [\int b_s''(x) b_t''(x) dx]_{s,t=1,\ldots,N}$$

PLSE objective (for one spline)

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^N} \left\|Y - G\boldsymbol{\beta}\right\|^2 + \lambda\boldsymbol{\beta}^\top\Omega\boldsymbol{\beta}$$

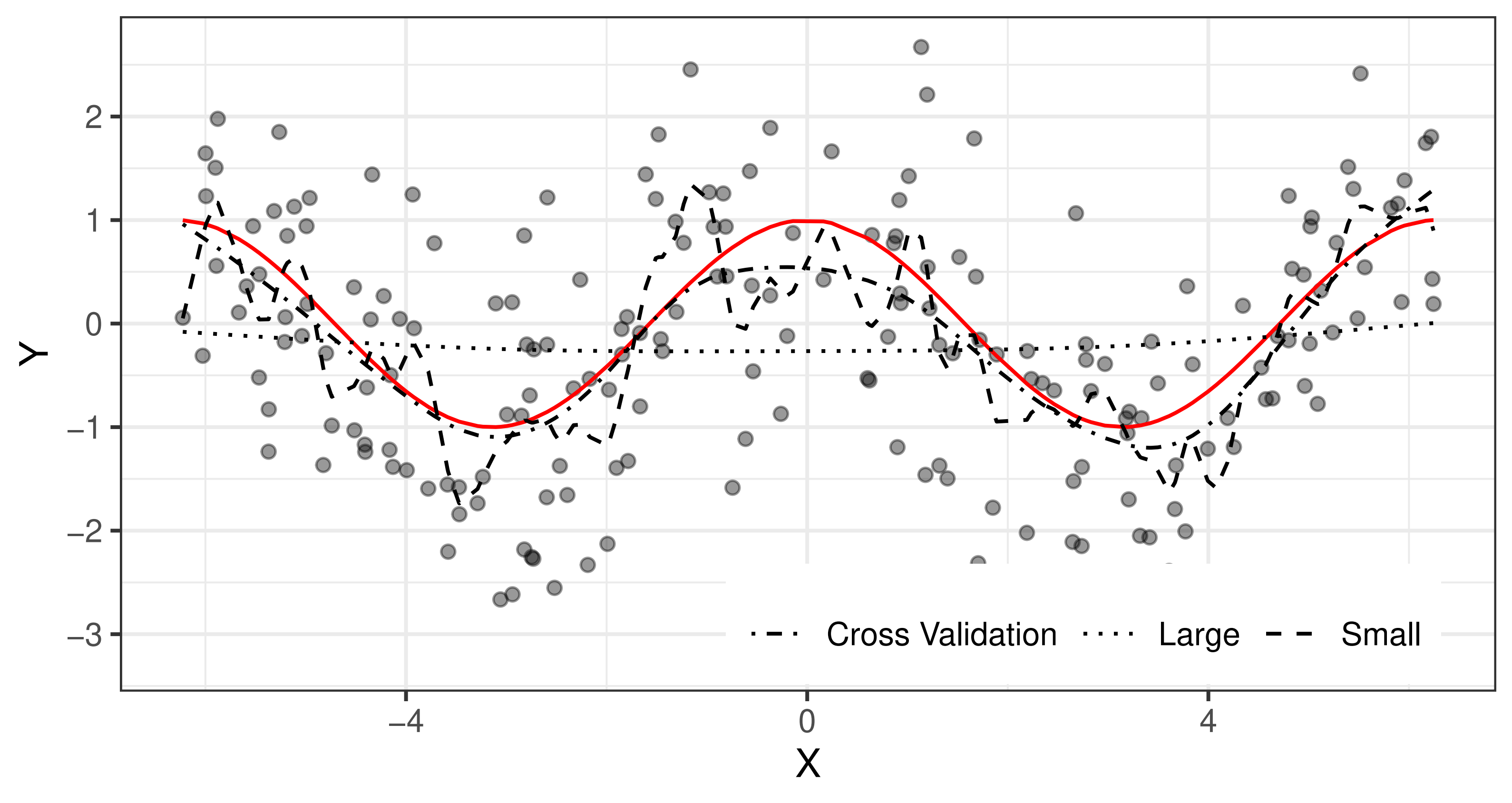
REMARK: Note similarity to OLS objective

Estimators

$$\hat{\boldsymbol{\beta}} = (G^\top G + \lambda\Omega)^{-1} G^\top y$$

Thus, the estimated smoother is:

$$\hat{f}(\cdot) = b(\cdot)(G^\top G + \lambda\Omega)^{-1} G^\top y$$



Estimating λ

- cross validation [Wood, 2017]
- marginal likelihood estimation in tandem with β [Wood et al., 2016]

Definition: Consistency

Let $M :=\{p_\theta : \theta \in \Theta\}$ be a parametric statistical model where $\theta \mapsto p_\theta$ is injective. Further, let $p_{\theta_0} \in M$ denote the true model of the data generating process for a dataset $D = \{(\mathrm{x}^n, y^n)\}_{n=1}^N$. Then an estimator θ_N is called *consistent* iff for all $\epsilon > 0$ holds

$$P\left(|\theta_N - \theta_0| > \epsilon\right) \xrightarrow{N\to\infty} 0.$$

Consistency has been shown for spline based GAMs by [Heckman, 1986].

Measuring Model Fit: Deviance

A likelihood based measure of model fit

Difference between the log-likelihood $\ell(\mu)$ of a model μ and the largest possible log-likelihood ℓ^*

$$D^*_\mu := 2(\ell^* - \ell(\mu))$$

ℓ^* corresponds to the likelihood of a model that perfectly reproduces the targets

Deviance explained

$$D^2(\mu) = 1 - \frac{D^*_\mu}{D_{\mu_0}} \in [0, 1]$$

where μ_0 denotes the intercept only model