Recap

High-dimensional additive models

$$Y_i = \mu + \sum_{j=1}^{p} f_j^0(X_i^{(j)}) + \varepsilon_i \;\; (i = 1, \ldots, n;\; p \gg n)$$

$$f_j^0 : \mathbb{R} \to \mathbb{R} \text{ smooth}, \; \mathbb{E}[f_j^0(X_i^{(j)})] \equiv 0 \; \forall \, j$$

aim: estimator such that either $\hat{f}_j(.) \equiv 0$ or $\hat{f}_j(.)$ is not the zero function
$\leadsto$ Group Lasso type problem

but: sparsity and smoothness penalty (SpS) is better than Group-sparsity (with respect to basis expansion) alone

Sparsity Smoothness (SpS) penalty with smoothing spline

$$\hat{f}_1, \ldots, \hat{f}_p = \mathrm{argmin}_{f_1,\ldots,f_p \in \mathcal{F}} \big(\| Y - \sum_{j=1}^{p} f_j \|_n^2 + \lambda_1 \sum_{j=1}^{p} \| f_j \|_n + \lambda_2 \sum_{j=1}^{p} I(f_j) \big)$$

$$\| f_j \|_n^2 = n^{-1} \sum_{i=1}^{n} | f_j(X_i^{(j)}) |^2, \; I(f_j)^2 = \int f_j''(x)^2 dx$$

where $\mathcal{F}$ = Sobolev space of functions that are continuously differentiable with square integrable second derivatives

$$\hat{f}_1, \ldots, \hat{f}_p = \mathrm{argmin}_{f_1,\ldots,f_p \in \mathcal{F}}\big(\|Y - \sum_{j=1}^{p} f_j\|_n^2 + \lambda_1 \sum_{j=1}^{p} \|f_j\|_n + \lambda_2 \sum_{j=1}^{p} I(f_j)\big)$$

is a parametric problem of dimension $d \approx 3pn$, parametrized by natural cubic splines with basis functions encoded in a matrix

$$H_{n \times d} = (H_1, \ldots, H_p),\ (H_j)_{n \times \approx 3n} \text{ matrix}$$

and integrated squared second derivatives encoded in a matrix

$$(W_j)_{k,\ell} = \int h''_{j,k}(x) h''_{j,\ell}(x) dx$$

$\rightsquigarrow$

$$\hat{\beta} = \mathrm{argmin}_{\beta} \left(\|Y - H\beta\|_2^2/n + \lambda_1 \sum_{j=1}^{p} \sqrt{\beta_j^T H_j^T H_j \beta_j / n} + \lambda_2 \sum_{j=1}^{p} \sqrt{\beta_j^T W_j \beta_j} \right)$$

if the problem is sparse and smooth:
only a few $X^{(j)}$'s influence Y (only a few non-zero f_j^0) and the non-zero f_j^0 are smooth
$\rightsquigarrow$ one can often afford to model and fit additive functions in high dimensions

reason:

- dimensionality is of order $\dim = O(pn)$
  $\log(\dim)/n = O((\log(p) + \log(n))/n)$ which is still small
- sparsity and smoothness then lead to: if each f_j^0 is twice continuously differentiable

$$\|\hat{f} - f^0\|_2^2/n = O_P(\underbrace{\text{sparsity}}_{\text{no. of non-zero } f_j^0} \sqrt{\log(p)}\, n^{-4/5})$$

(cf. Ch. 8.4 in Bühlmann & van de Geer (2011))