Additive Regression Splines with Total Variation and Nonnegative Garrote Penalties

Kwan-Young Bak

Department of Statistics

Korea University

Joint work with Jae-Hwan Jhong, Jae-Kyung Shin, and Ja-Yong Koo

Agenda

Introduction

Additive regression model

- Data: $\{(x_n, y_n)\}_{n=1}^N$, $x_n = (x_n^1, \ldots, x_n^M) \in [0,1]^M$, $y_n \in \mathbb{R}$.
- Additive regression model:

$$y_n = f(x_n) + \varepsilon_n = \mu + \sum_{m=1}^{M} f_m(x_n^m) + \varepsilon_n$$

  $\mu \in \mathbb{R}$: intercept term
  f_m: component function defined on $[0,1]$
  ε_n: error term with $\mathbb{E}(\varepsilon_n) = 0$.

- Overcome the curse of dimensionality.
- More interpretable than tensor-product regression surfaces.

Main contributions of the study

- Sparse representation via a two-stage procedure.
- Data-dependent knot selection via total variation penalization.
- Spatially adaptive estimator in the additive model.
- Variable selection via garrote procedure.
- Convergence rate and selection consistency.

Penalized additive regression spline estimator

Additive spline model

- Order r B-spline expansion:

$$\mathsf{s}_{\beta^m} = \sum_{j=1}^{J_m} \beta_j^m B_j^m$$

$\beta^m = (\beta_1^m, \ldots, \beta_{J_m}^m) \in \mathbb{R}^{J_m}$.

- Additive spline:

$$\mathsf{f}_\beta = \beta_0 + \sum_{m=1}^{M} \mathsf{s}_{\beta^m}$$

$\beta = (\beta_0, \beta^1, \ldots, \beta^M) \in \mathbb{R}^K$, $K = 1 + \sum_{m=1}^M J_m$.

Two stage estimation procedure

- First stage: smoothing for each component.

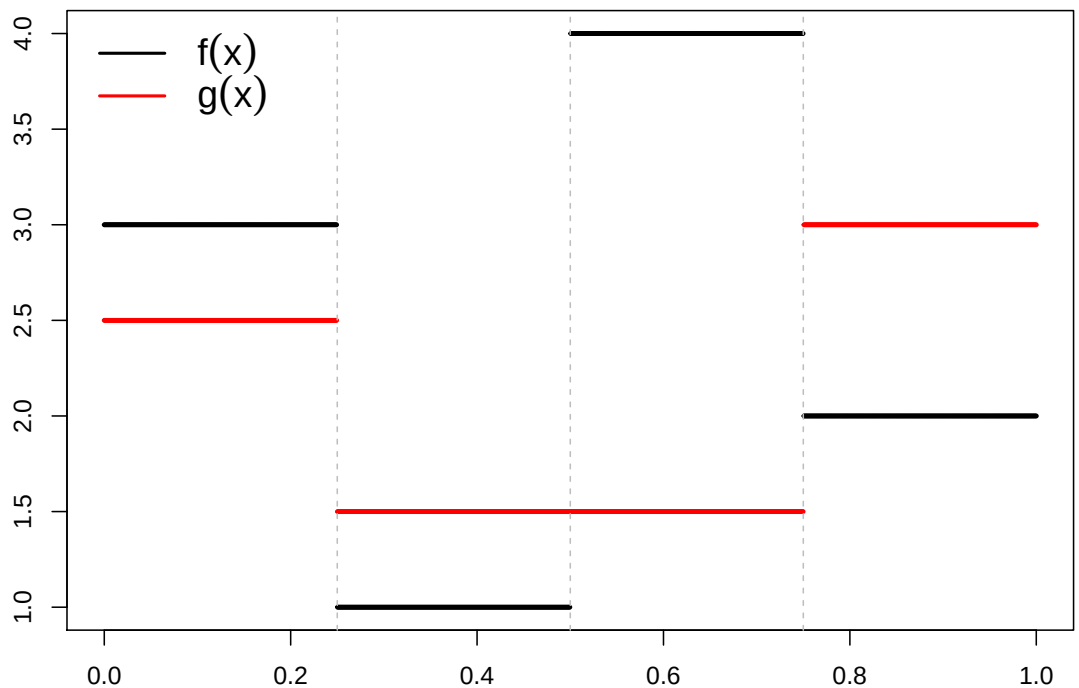
  Data-dependent knot selection using total variation penalization.

- Second stage: sparse component selection.

  Variable selection using sparsity-inducing garrote procedure.

Total variation for constant function

$$\mathrm{TV}(f) = |3-1| + |1-4| + |4-2| = 7$$
$$\mathrm{TV}(g) = |2.5-1.5| + |1.5-3| = 2.5$$



First stage estimation

- Optimization:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{n=1}^{N} \left(y_n - \mathsf{f}_\beta(x_n) \right)^2 + \lambda_1 \sum_{m=1}^{M} \mathrm{TV}\left(\mathsf{s}_{\beta^m}^{(r-1)} \right) \right].$$

- First stage estimator:

$$\tilde{f} = \mathsf{f}_{\hat{\beta}} = \hat{\beta}_0 + \sum_{m=1}^{M} \mathsf{s}_{\hat{\beta}^m}.$$

Second stage estimation

- Shrinkage factor:

$$\hat{\tau} = \underset{(\tau_1,\ldots,\tau_M)\in\mathbb{R}_+^M}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \hat{\beta}_0 - \sum_{m=1}^{M} \tau_m \mathsf{s}_{\hat{\beta}^m}(x_n^m) \right)^2 + \lambda_2 \sum_{m=1}^{M} \tau_m.$$

- Penalized additive regression spline estimator:

$$\hat{f} = \hat{\beta}_0 + \sum_{m=1}^{M} \hat{f}_m,$$

where

$$\hat{f}_m = \hat{\tau}_m \mathsf{s}_{\hat{\beta}^m}.$$

Numerical study

Simulation setup

- Example functions: for $t \in [0, 1]$
  - $f_1(t) = 6\sin(3\cos(3t)\exp(2t))$
  - $f_2(t) = 5\sin((3.5t)^2)$
  - $f_3(t) = 20t^{2t}$
  - $f_4(t) = 6(0.1\sin(2\pi t) + 0.2\cos(2\pi t)) + 0.3\sin^2(2\pi t) + 0.4\cos^4(2\pi t) + 0.5\sin^3(2\pi t)$
- Number of variables: $M = 20$.
- Number of observations: $N = 200, 400, 600$.
- Data generation:

$$y_n = f_1(x_n^1) + f_2(x_n^2) + f_3(x_n^3) + f_4(x_n^4) + \varepsilon_n$$

$x_n^m \sim U(0, 1)$ and $\varepsilon_n \sim N(0, 1)$.

Performance measures

- Mean squared error:

$$\mathrm{MSE}(g) = \frac{1}{N}\sum_{n=1}^{N}(g(x_n) - f(x_n))^2.$$

- Mean absolute deviation:

$$\mathrm{MAD}(g) = \frac{1}{N}\sum_{n=1}^{N}|g(x_n) - f(x_n)|.$$

- Maximum deviation:

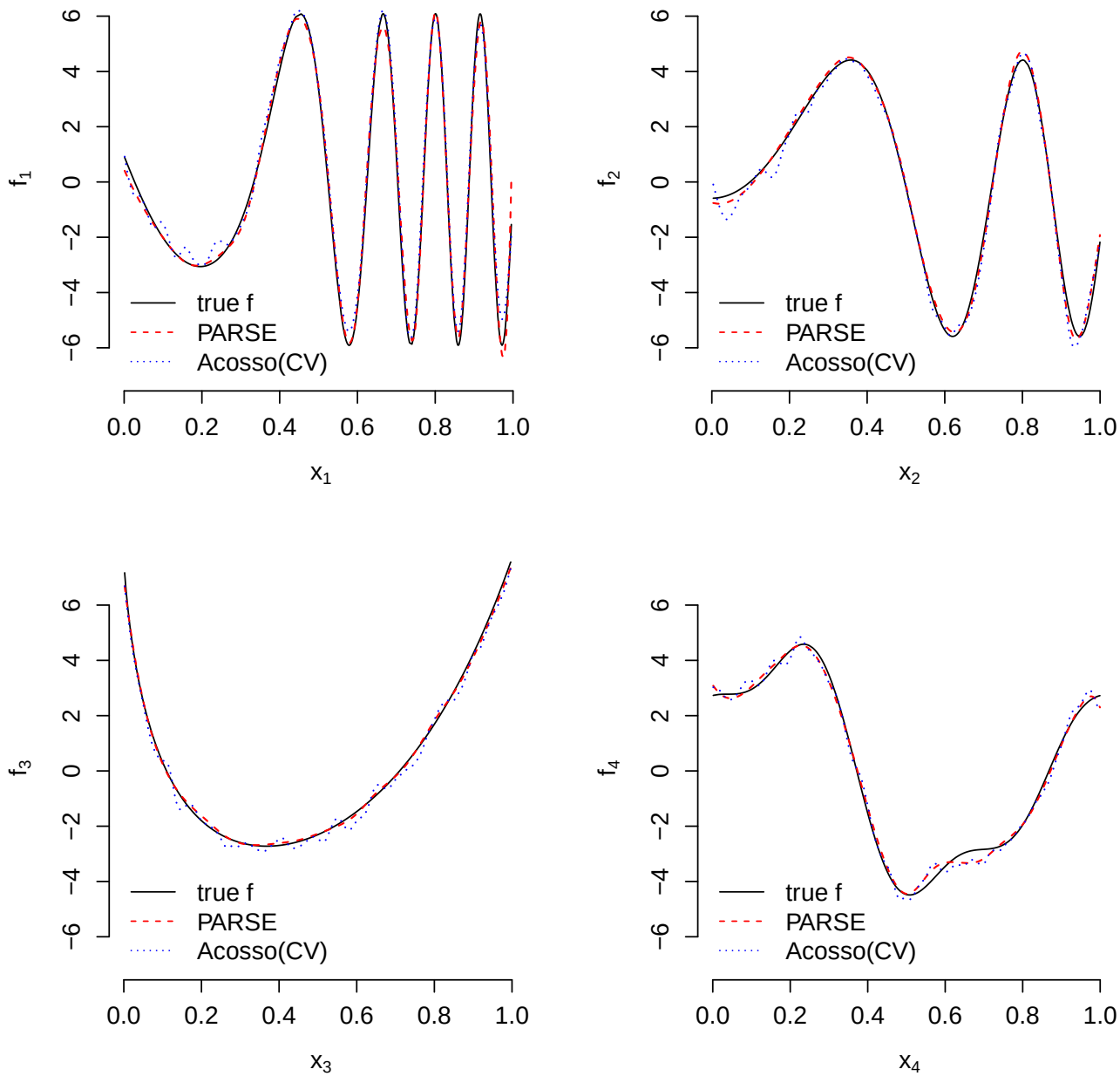
$$\mathrm{MXDV}(g) = \max_{1\le n\le N}|g(x_n) - f(x_n)|.$$

- AverR(AverI): average number of (ir)relevant variables selected.

Simulation result

		MSE	MAD	MXDV	AverR	AverI
$N = 200$	PARSE	**2.487(0.120)**	**1.188(0.104)**	**4.621(0.145)**	4.00	0.00
	Cosso(CV)	8.600(0.135)	2.314(0.180)	7.755(0.105)	4.00	2.51
	Cosso(BIC)	9.731(0.159)	2.454(0.193)	8.208(0.107)	3.98	0.12
	Acosso(CV)	5.876(0.102)	1.878(0.153)	6.515(0.085)	4.00	2.55
	Acosso(BIC)	6.650(0.125)	1.980(0.165)	6.819(0.080)	3.99	0.10
$N = 400$	PARSE	**0.324(0.010)**	**0.436(0.037)**	**1.884(0.046)**	4.00	0.00
	Cosso(CV)	3.750(0.024)	1.42(0.132)	5.628(0.065)	4.00	0.77
	Cosso(BIC)	3.871(0.023)	1.430(0.135)	5.657(0.063)	4.00	0.00
	Acosso(CV)	0.986(0.015)	0.748(0.065)	3.197(0.058)	4.00	0.45
	Acosso(BIC)	0.988(0.015)	0.746(0.066)	3.206(0.058)	4.00	0.00
$N = 600$	PARSE	**0.193(0.005)**	**0.337(0.028)**	**1.570(0.037)**	4.00	0.00
	Cosso(CV)	1.364(0.013)	0.841(0.081)	3.985(0.058)	4.00	0.29
	Cosso(BIC)	1.371(0.013)	0.840(0.082)	3.993(0.059)	4.00	0.00
	Acosso(CV)	0.330(0.014)	0.456(0.035)	1.604(0.040)	4.00	1.14
	Acosso(BIC)	0.430(0.016)	0.500(0.042)	2.377(0.057)	4.00	0.00

Estimated functional components



Theoretical results

Goodness-of-fit measure and function class

- L_2 norm of a function on $[0,1]$:

$$\|h\|_{m,2} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} h^2(x_n^m)}$$

- L_2 norm of a function on $[0,1]^M$:

$$\|H\|_2 = \sqrt{\frac{1}{N}\sum_{n=1}^{N} H^2(x_n)}$$

- Smoothness assumption on components:

$$f_m \in \left\{ h\colon h \text{ is of class } \mathcal{C}^r,\ \sup_{u\in\mathcal{I}} |h(u)| \le Q \right\}.$$

Oracle inequality

Theorem 1
Let

$$\lambda_1 = \sqrt{\frac{\log\left((2MJ+2)/\delta\right)}{M_1 J^{2r-1} N}}, \tag{1}$$

where $0 < \delta < 1$ *is a user-specified parameter. Then, with a probability at least* $1-\delta$*, for all* $\gamma > 1$ *and all* $\beta \in \mathcal{B}$*, we have*

$$\|f - \tilde{f}\|_2^2 \leq \frac{\gamma+1}{\gamma-1}\|f - \mathsf{f}_\beta\|_2^2 + \frac{2\gamma}{\gamma-1} M_2 J^{2r} \lambda_1^2.$$

Convergence rate for the first stage estimator

Theorem 2
Let λ_1 *be given by* (1) *with* $\delta = N^{-2}$. *Then, we have, with a probability at least* $1 - N^{-2}$,

$$\|f - \tilde{f}\|_2^2 \leq M_3 \left(\frac{N}{\log N} \right)^{-2r/(2r+1)}.$$

Selection consistency

Theorem 3
Suppose that λ_2 *is chosen such that* $\lambda_2/N \to 0$ *and* $(\log N)^{\frac{r}{2r+1}} N^{\frac{r+1}{2r+1}}/\lambda_2 \to 0$ *as* $N \to \infty$. *Then, we have*

$$\mathbb{P}\left(\hat{I} = I^0\right) \to 1$$

and

$$\sup_{1 \le m \le M} \|f_m - \hat{f}_m\|_{m,2}^2 \le M_4 \left(\frac{\lambda_2}{N}\right)^2.$$

Convergence rate for the second stage estimator

Theorem 4
Let

$$\lambda_2 = M_5 N^{\frac{r+1}{2r+1}} (\log N)^{\kappa} \quad for \quad \kappa > \frac{r}{2r+1}.$$

Then, we have

$$\sup_{1 \le m \le M} \|f_m - \hat{f}_m\|_{m,2}^2 \le M_6 N^{-\frac{2r}{2r+1}} (\log N)^{2\kappa},$$

where $M_6 = M_4 M_5^2$.

Discussion

Discussion

- Smoothing and component selection via two-stage penalization.
- Nonasymptotic oracle inequality for the total variation estimator.
- Component selection consistency in the garrote procedure.
- Circumventing the dimensionality issue.
- Optimal convergence rate (within a logarithmic factor).