Convex-constrained Sparse Additive Modeling and Its Extensions

Junming Yin[1] Yaoliang Yu[2]

[1]Eller College of Management, University of Arizona [2]School of Computer Science, University of Waterloo

Overview

- We consider how to incorporate certain shape constraints such as **convexity/concavity** and their extensions into sparse additive models

$$Y = m(X_1, X_2, \ldots, X_p) + \xi = \sum_{j=1}^{p} f_j(X_j) + \xi$$

  - Many functions that arise in practice tend to be convex/concave or monotonic (Groeneboom and Jongbloed 2014)
  - In high-dimensional setting, many covariates may not be relevant
- **Our contributions:**
  - A sparse convex additive model (SCAM) to estimate convex (and monotonic) component functions in high dimensional additive modeling
  - A sparse difference of convex additive model (SDCAM) to address potential robustness issue of SCAM, e.g., convex functions are mistakenly believed to be concave
  - An efficient backfitting algorithm with linear per-iteration complexity

Sparse Convex Additive Model (SCAM)

- Given a set of data samples $\{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}, i = 1, \ldots, n\}$, solve

$$\min_{\forall j, f_j \in \mathcal{C}_j} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} f_j(x_{ij})\Big)^2 + \lambda_s \sum_{j=1}^{p} \|f_j\|_2$$

  - $\mathcal{C}_j := \{f : [0,1] \to \mathbb{R} \mid \mathbb{E}(f(X_j)) = 0, f \text{ is convex}\}$
  - $\|f_j\|_2 := \sqrt{\mathbb{E}(f_j^2(X_j))}$ is the L_2 norm of component function f_j
- Can reduce to an equivalent finite-dimensional optimization problem:

$$\min_{\forall j, \tilde{\mathbf{z}}_j \in \mathcal{K}_j \cap \mathcal{H}} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} z_{ij}\Big)^2 + \lambda_s \sum_{j=1}^{p} \|\mathbf{z}_j\|_2$$

  - $\mathbf{z}_j \in \mathbb{R}^n$ are the component fits on the observed values: $z_{ij} = f_j(x_{ij}), i = 1 \ldots, n$
  - $\tilde{z}_{ij} = f_j(\tilde{x}_{ij})$ is a permuted version of $\mathbf{z}_j$, according to $\mathbf{x}_j$
  - $\mathcal{H} := \{\mathbf{z} \in \mathbb{R}^n : \sum_{i=1}^{n} z_i = 0\}$ is the empirical centering constraint
  - The convex cone

$$\mathcal{K}_j := \left\{\mathbf{z} \in \mathbb{R}^n : \frac{z_2 - z_1}{\tilde{x}_{2j} - \tilde{x}_{1j}} \le \cdots \le \frac{z_n - z_{n-1}}{\tilde{x}_{nj} - \tilde{x}_{n-1,j}}\right\}$$

  is sufficient and necessary to ensure f_j to be convex (Hildreth 1954).

Difference of Convex (DC) Functions

- What if we were wrong and the component function f_j is in fact *concave*?
- **Idea:** consider the class of difference of convex (DC) functions

$$\mathcal{DC}_j := \{f : [0,1] \to \mathbb{R} \mid f = f_1 - f_2, f_1 \in \mathcal{C}_j, f_2 \in \mathcal{C}_j\}$$

- **Fact:** most continuous functions (convex/smooth or not) in practice are DC
- Naively replacing $\mathcal{C}_j$ with $\mathcal{DC}_j$ in the constraint of SCAM will severely overfit to the data
- **Theorem:** For any sample $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\} \subseteq [0,1]^p \times \mathbb{R}$ such that $\mathbf{x}_i \neq \mathbf{x}_j$ for all $1 \le i \neq j \le n$, there always exists a multivariate DC function $f : [0,1]^p \to \mathbb{R}$ such that for all $i = 1, \ldots, n$, $f(\mathbf{x}_i) = y_i$.

Sparse Difference of Convex Additive Model (SDCAM)

- **Theorem:** (Roberts and Varberg 1973, c.f.): Let $f : [0,1] \to \mathbb{R}$ be a function with finite one-sided derivatives at 0 and 1. Then f is DC iff

$$\|f\|_{\mathcal{DC}} := \sup_{\mathcal{P}} \sum_{i=2}^{n-1} \left| \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} - \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \right| < \infty,$$

  where the supremum is taken over n and the partitions $\mathcal{P}$ of $[0,1]$.
- Based on such characterization, we estimate DC functions by solving

$$\min_{\forall j, f_j \in \mathcal{DC}_j} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} f_j(x_{ij})\Big)^2 + \sum_{j=1}^{p} (\lambda_d \|f_j\|_{\mathcal{DC}} + \lambda_s \|f_j\|_2)$$

- Can reduce to an equivalent finite-dimensional optimization problem:

$$\min_{\forall j, \tilde{\mathbf{z}}_j \in \mathcal{H}} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} z_{ij}\Big)^2 + \sum_{j=1}^{p} \big(\lambda_d \|\tilde{\mathbf{z}}_j\|_{\mathcal{DC}_j} + \lambda_s \|\mathbf{z}_j\|_2\big),$$

  where for $\mathbf{z} \in \mathbb{R}^n$ we define

$$\|\mathbf{z}\|_{\mathcal{DC}_j} := \sum_{i=2}^{n-1} \left| \frac{z_{i+1} - z_i}{\tilde{x}_{i+1,j} - \tilde{x}_{ij}} - \frac{z_i - z_{i-1}}{\tilde{x}_{ij} - \tilde{x}_{i-1,j}} \right|.$$

Modified Backfitting Algorithm

- In each iteration, fix all component fits except for one $\mathbf{z}_j$ and solve the resulting subproblem:

$$\min_{\tilde{\mathbf{z}}_j \in \mathcal{H}} \tfrac{1}{2}\|\tilde{\mathbf{r}}_j - \tilde{\mathbf{z}}_j\|_2^2 + \lambda_d \|\tilde{\mathbf{z}}_j\|_{\mathcal{DC}_j} + \lambda_s \|\tilde{\mathbf{z}}_j\|_2,$$

  where $\tilde{\mathbf{r}}_j \in \mathbb{R}^n$ is the partial residual that removes the contribution of $\tilde{\mathbf{z}}_j$.
- **Theorem:** The solution can be characterized as

$$\mathsf{P}_{\lambda_d\|\cdot\|_{\mathcal{DC}_j} + \lambda_s\|\cdot\|_2 + \mathcal{H}}(\tilde{\mathbf{r}}_j) = \mathsf{P}_{\lambda_s\|\cdot\|_2}\Big[\mathsf{P}_{\mathcal{H}}\big(\mathsf{P}_{\lambda_d\|\cdot\|_{\mathcal{DC}_j}}(\tilde{\mathbf{r}}_j)\big)\Big],$$

  where P_f is the proximal operator associated with a convex function f

$$\mathsf{P}_f(\mathbf{r}) = \operatorname*{argmin}_{\mathbf{z} \in \mathbb{R}^n} \frac{1}{2}\|\mathbf{z} - \mathbf{r}\|_2^2 + f(\mathbf{z}), \quad \forall \mathbf{r} \in \mathbb{R}^n.$$

  - $\mathsf{P}_{\mathcal{H}}$ amounts to subtracting the average
  - $\mathsf{P}_{\lambda_s\|\cdot\|_2}(\mathbf{r}) = \Big(1 - \frac{\lambda_s}{\|\mathbf{r}\|_2}\Big)_+ \mathbf{r}$ is the block soft thresholding operator
  - With a suitable change of variables, computing $\mathsf{P}_{\lambda_d\|\cdot\|_{\mathcal{DC}_j}}(\tilde{\mathbf{r}}_j)$ is equivalent to

$$\min_{s \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^{n-1}} \frac{1}{2}\left\| A_j \begin{pmatrix} s \\ \mathbf{w} \end{pmatrix} - \tilde{\mathbf{r}}_j \right\|_2^2 + \lambda_d \|\mathbf{w}\|_{\text{tv}},$$

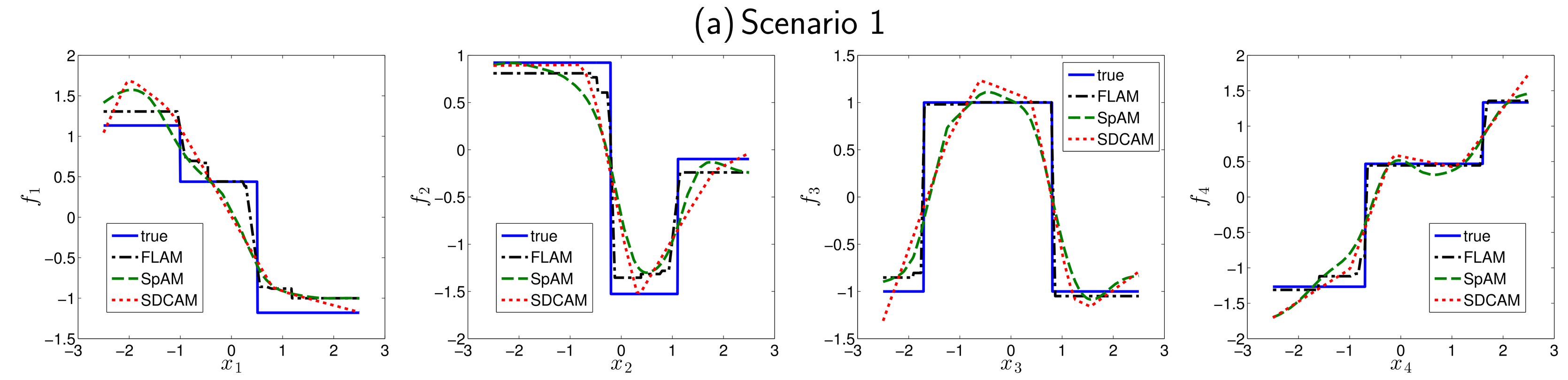
  for certain lower triangular matrix A_j.
- Using the linear-time algorithm in (Davies and Kovac 2001) to compute the proximal operator $\mathsf{P}_{\lambda_d\|\cdot\|_{\text{tv}}}$, we are able to compute $\mathsf{P}_{\lambda_d\|\cdot\|_{\mathcal{DC}_j}}(\mathbf{r}_j)$ iteratively using the accelerated proximal gradient algorithm

Simulation Study

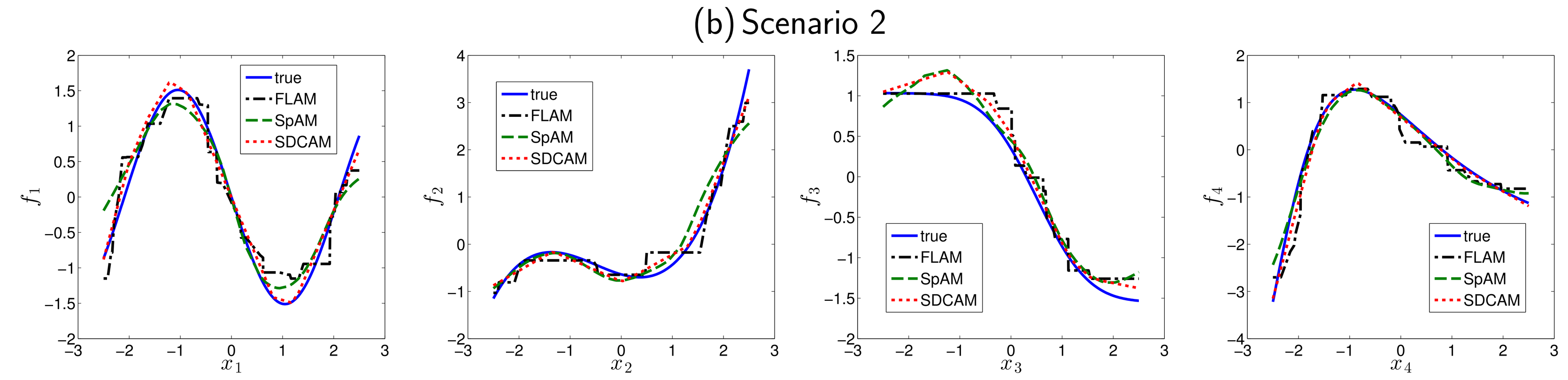
- $n = 100$, $p = 1000$
- $y_i = \sum_{j=1}^{4} f_j(x_{ij}) + \epsilon_i$
- $x_{ij} \sim \mathrm{Uniform}(-2.5, 2.5)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with SNR = 3 or 5
- Competing algorithms: SpAM (Ravikumar et al. 2008) and FLAM (Petersen, Witten, and Simon 2016)
- Scenario 1: all component functions are piecewise constant
- Scenario 2: all component functions are smooth
- Scenario 3: f_2 is smooth and the rest of them are piecewise linear

method	precision	recall	model size	MSE
Scenario 1, SNR = 3				
SpAM	0.87 (0.19)	0.99 (0.04)	4.93 (1.86)	1.15 (0.39)
FLAM	0.82 (0.21)	1.00 (0.00)	5.48 (2.60)	0.85 (0.24)
SDCAM	0.86 (0.17)	0.99 (0.05)	4.78 (1.14)	1.32 (0.41)
Scenario 1, SNR = 5				
SpAM	0.92 (0.14)	1.00 (0.00)	4.54 (1.14)	0.72 (0.20)
FLAM	0.78 (0.22)	1.00 (0.00)	5.73 (2.27)	0.43 (0.13)
SDCAM	0.93 (0.13)	1.00 (0.00)	4.41 (0.96)	0.75 (0.18)
Scenario 2, SNR = 3				
SpAM	0.87 (0.19)	0.99 (0.03)	4.91 (1.52)	0.89 (0.36)
FLAM	0.80 (0.20)	1.00 (0.00)	5.48 (2.14)	0.95 (0.29)
SDCAM	0.93 (0.12)	1.00 (0.00)	4.38 (0.78)	0.64 (0.14)
Scenario 2, SNR = 5				
SpAM	0.92 (0.14)	1.00 (0.00)	4.51 (1.17)	0.47 (0.17)
FLAM	0.76 (0.19)	1.00 (0.00)	5.70 (1.74)	0.49 (0.16)
SDCAM	0.98 (0.06)	1.00 (0.00)	4.10 (0.30)	0.25 (0.04)
Scenario 3, SNR = 3				
SpAM	0.92 (0.16)	1.00 (0.00)	4.57 (1.39)	0.47 (0.10)
FLAM	0.86 (0.16)	1.00 (0.00)	4.84 (1.17)	0.55 (0.11)
SDCAM	0.97 (0.08)	1.00 (0.00)	4.15 (0.41)	0.41 (0.06)
Scenario 3, SNR = 5				
SpAM	0.98 (0.07)	1.00 (0.00)	4.11 (0.34)	0.21 (0.04)
FLAM	0.77 (0.19)	1.00 (0.00)	5.62 (1.85)	0.26 (0.06)
SDCAM	0.99 (0.05)	1.00 (0.00)	4.06 (0.24)	0.15 (0.02)

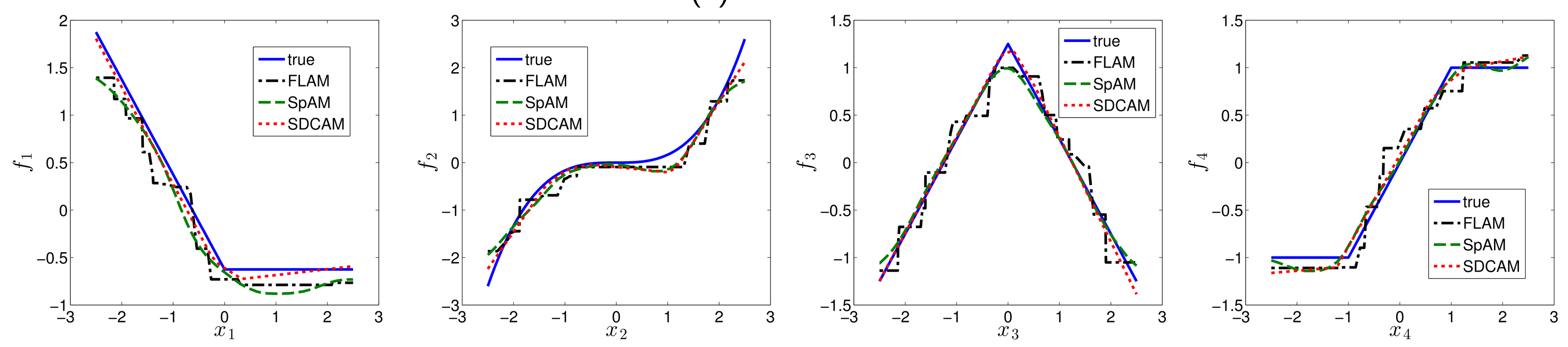
(a) Scenario 1



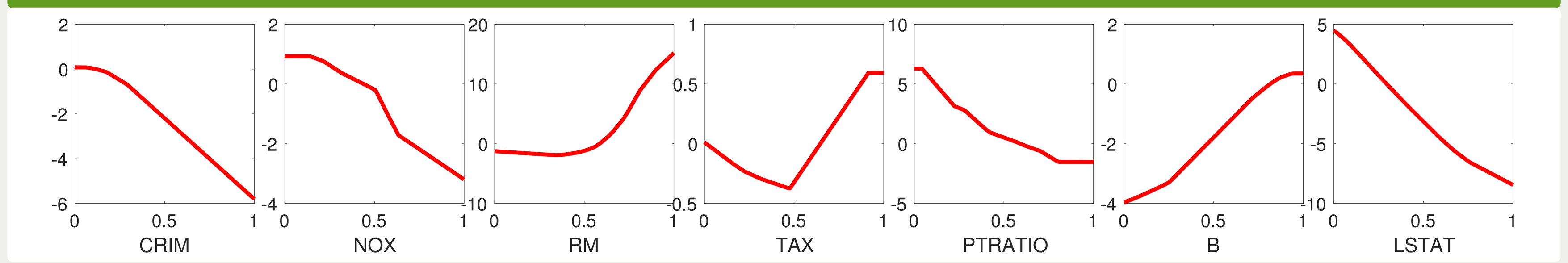
(b) Scenario 2



(c) Scenario 3



Boston Housing Data



References

- C. Hildreth (1954). "Point estimates of ordinates of concave functions". In: *Journal of the American Statistical Association* 49.267, pp. 598–619
- A. W. Roberts and D. E. Varberg (1973). *Convex Functions*. Academic Press
- P. L. Davies and A. Kovac (2001). "Local extremes, runs, strings and multiresolution". In: *The Annals of Statistics* 29.1, pp. 1–65
- P. Ravikumar et al. (2008). "SpAM: Sparse Additive Models". In: *NIPS*
- P. Groeneboom and G. Jongbloed (2014). *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press
- A. Petersen, D. Witten, and N. Simon (2016). "Fused lasso additive model". In: *Journal of Computational and Graphical Statistics* 25.4, pp. 1005–1025