SpAM: Sparse Additive Models

Author: Pradeep Ravikumar, Han Liu,
John Lafferty and Larry Wasserman

Outline

- 1) The SpAM Optimization Problem
- 2) A backfitting algorithm
- 3) Properties of SpAM
- 4) Simulations showing the estimator's behavior

Outline

- 1) The SpAM Optimization Problem
- 2) A backfitting algorithm
- 3) Properties of SpAM
- 4) Simulations showing the estimator's behavior

The SpAM Optimization Problem

- How sparsity is achieved

  Recall the standard additive model optimization problem:

$$\min_{f_j \in \mathcal{H}_j,\, 1 \le j \le p} \mathbb{E}\left(Y - \sum_{j=1}^{p} f_j(X_j)\right)^2$$

The SpAM Optimization Problem

- How sparsity is achieved

  Some modif cation:

$$\min_{f_j \in \mathcal{H}_j,\, 1 \le j \le p} \mathbb{E}\left(Y - \sum_{j=1}^{p} f_j(X_j)\right)^2$$

$$f_j(\cdot) = \beta_j g_j(\cdot)$$

$$\min_{\beta \in \mathbb{R}^p,\, g_j \in \mathcal{H}_j} \mathbb{E}\left(Y - \sum_{j=1}^{p} \beta_j g_j(X_j)\right)^2$$

The SpAM Optimization Problem

- Consider following modification that imposes additional constraints:

$$\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} \quad \mathbb{E}\left(Y - \sum_{j=1}^{p} \beta_j g_j(X_j)\right)^2$$

$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq L$$

$$\mathbb{E}\left(g_j^2\right) = 1, \quad j = 1, \ldots, p$$

$$\mathbb{E}\left(g_j\right) = 0, \quad j = 1, \ldots, p$$

The SpAM Optimization Problem

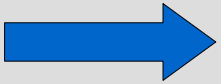
- L1 encourages sparsity of estimated beta

$$\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} \quad \mathbb{E}\left(Y - \sum_{j=1}^{p} \beta_j g_j(X_j)\right)^2$$

$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq L$$

$$\mathbb{E}\left(g_j^2\right) = 1, \quad j = 1, \ldots, p$$

$$\mathbb{E}\left(g_j\right) = 0, \quad j = 1, \ldots, p$$



$$\{\beta : \|\beta\|_1 \leq L\}$$

$$f(x) = \sum_{j=1}^{p} f_j(x_j) = \sum_{j=1}^{p} \beta_j g_j(x_j)$$

The SpAM Optimization Problem

- Drawback: the optimization problem is convex in and $\{g\}$ separately
- But not convex in and $\{g\}$ jointly.
- So consider a related optimization problem.

$$\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} \quad \mathbb{E}\left(Y - \sum_{j=1}^{p} \beta_j g_j(X_j)\right)^2$$

$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq L$$

$$\mathbb{E}\left(g_j^2\right) = 1, \quad j = 1, \ldots, p$$

$$\mathbb{E}\left(g_j\right) = 0, \quad j = 1, \ldots, p$$

- First call the original optimization problem as P

The SpAM Optimization Problem

- A related optimization problem, called Q:

$$\min_{f_j \in \mathcal{H}_j} \quad \mathbb{E}\left(Y - \sum_{j=1}^{p} f_j(X_j)\right)^2$$

$$\text{subject to} \quad \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(X_j))} \leq L$$

$$\mathbb{E}(f_j) = 0, \quad j = 1, \ldots, p.$$

This problem is convex in $\{f\}$ and the problem P and Q are equivalent

The SpAM Optimization Problem

- The problem P and Q are equivalent:

$$\left(\left\{\beta_j^*\right\}, \left\{g_j^*\right\}\right) \text{ optimizes } (P) \text{ implies } \left\{f_j^* = \beta_j^* g_j^*\right\} \text{ optimizes } (Q);$$

$$\left\{f_j^*\right\} \text{ optimizes } (Q) \text{ implies } \left(\left\{\beta_j^* = (\|f_j\|_2)^T\right\}, \left\{g_j^* = f_j^*/\|f_j^*\|_2\right\}\right) \text{ optimizes } (P).$$

- The optimization Problem Q is convex
- It encourages sparsity is not intuitive

The SpAM Optimization Problem

- It encourages sparsity is not intuitive
- Consider an example to provie some insight

$$C = \left\{ (f_{11}, f_{12}, f_{21}, f_{22})^T \in \mathbb{R}^4 : \sqrt{f_{11}^2 + f_{12}^2} + \sqrt{f_{21}^2 + f_{22}^2} \leq L \right\}$$

The projection π12*C onto* first two components (L2)
The projection π13*C onto* first third components (L1)

The SpAM Optimization Problem

$$C = \left\{ (f_{11}, f_{12}, f_{21}, f_{22})^T \in \mathbb{R}^4 : \sqrt{f_{11}^2 + f_{12}^2} + \sqrt{f_{21}^2 + f_{22}^2} \leq L \right\}$$

$$\sum_j \| f_j \|_2 \leq L$$

- Act as an L1 constraint across components(sparsity)
- Act as an L2 constraint within components (smoothnes s)

In case $\{f\}$ is linear

$$(f_j(x_{1j}), \ldots, f(x_{nj})) = \beta_j(x_{1j}, \ldots, x_{nj})$$

The optimization problem reduces to the lasso

Outline

- 1) The SpAM Optimization Problem
- 2) A backfitting algorithm
- 3) Properties of SpAM
- 4) Simulations showing the estimator's behavior

A backfitting algorithm

- A coordinate descent algorithm
- Write the Lagrangian for the optimization Q

$$\begin{aligned} \min_{f_j \in \mathcal{H}_j} \quad & \mathbb{E}\left(Y - \textstyle\sum_{j=1}^{p} f_j(X_j)\right)^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(X_j))} \le L \\ & \mathbb{E}(f_j) = 0, \ j = 1, \ldots, p. \end{aligned}$$

$$\mathcal{L}(f, \lambda, \mu) = \frac{1}{2}\mathbb{E}\left(Y - \textstyle\sum_{j=1}^{p} f_j(X_j)\right)^2 + \lambda \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(X_j))} + \sum_j \mu_j \mathbb{E}(f_j).$$

A backfitting algorithm

- Define the jth Residual:

$$R_j = Y - \sum_{k \neq j} f_k(X_k)$$

- Minimizing the Lagrangian as a function of f_j *is expressed in terms of Frechet derivative as:*

$$\delta\mathcal{L}(f, \lambda, \mu; \delta f_j) = \mathbb{E}\left[(f_j - R_j + \lambda v_j)\delta f_j\right] = 0$$

where: $v_j = f_j \Big/ \sqrt{\mathbb{E}(f_j^2)}$

A backfitting algorithm

- Conditioning on Xj, the Frechet derivative becomes:

$$f_j + \lambda v_j = \mathbb{E}(R_j \mid X_j)$$

- Letting $P_j = E[R_j \mid X_j]$ denote the projection of the residual onto H_j, the solution satisfies

$$\left(1 + \frac{\lambda}{\sqrt{\mathbb{E}(f_j^2)}}\right) f_j = P_j \text{ if } \mathbb{E}(P_j^2) > \lambda$$

(Recall that : $v_j = f_j \Big/ \sqrt{\mathbb{E}(f_j^2)}$)

A backfitting algorithm

- This form

$$\left(1+\frac{\lambda}{\sqrt{\mathbb{E}(f_j^2)}}\right) f_j = P_j \text{ if } \mathbb{E}(P_j^2) > \lambda$$

implies

$$\left(1+\frac{\lambda}{\sqrt{\mathbb{E}(f_j^2)}}\right) \sqrt{\mathbb{E}(f_j^2)} = \sqrt{\mathbb{E}(P_j^2)}$$

or

$$\sqrt{\mathbb{E}(f_j^2)} = \sqrt{\mathbb{E}(P_j^2)} - \lambda$$

A backfitting algorithm

- From the form,

$$\left(1+\frac{\lambda}{\sqrt{\mathbb{E}(f_j^2)}}\right)\sqrt{\mathbb{E}(f_j^2)}=\sqrt{\mathbb{E}(P_j^2)} \quad \text{or} \quad \sqrt{\mathbb{E}(f_j^2)}=\sqrt{\mathbb{E}(P_j^2)}-\lambda$$

- we arrive the following soft-thresholding update

$$f_j=\left[1-\frac{\lambda}{\sqrt{\mathbb{E}(P_j^2)}}\right]_+ P_j$$

A backfitting algorithm

$$f_j = \left[1 - \frac{\lambda}{\sqrt{\mathbb{E}(P_j^2)}} \right]_+ P_j$$

- Two terms are needed to be estimated
- 1) As in standard backfitting, the projection

  $P_j = E [R_j | X_j]$ is estimated by a smooth of the residuals

$$\widehat{P}_j = S_j R_j$$

  S_j is a linear smoother, such as a local linear or kernel s moother

A backfitting algorithm

$$f_j = \left[1 - \frac{\lambda}{\sqrt{\mathbb{E}(P_j^2)}}\right]_+ P_j$$

- Two terms are needed to be estimated
- 2) A simple but biased estimate for the denominator:

$$\widehat{s}_j = \frac{1}{\sqrt{n}} \| \widehat{P}_j \|_2 = \sqrt{\text{mean}(\widehat{P}_j^2)}.$$

A backfitting algorithm

- We derived the SpAM backfitting algorithm

Input: Data (X_i, Y_i), regularization parameter λ.

Initialize $f_j = f_j^{(0)}$, for $j = 1, \ldots, p$.

Iterate until convergence:

For each $j = 1, \ldots, p$:

Compute the residual: $R_j = Y - \sum_{k \neq j} f_k(X_k)$;

Estimate the projection $P_j = \mathbb{E}[R_j \mid X_j]$ by smoothing: $\widehat{P}_j = S_j R_j$;

Estimate the norm $s_j = \sqrt{\mathbb{E}[P_j]^2}$ using, for example, (15) or (35);

Soft-threshold: $f_j = \left[1 - \frac{\lambda}{\widehat{s}_j}\right]_+ \widehat{P}_j$;

Center: $f_j \leftarrow f_j - \text{mean}(f_j)$.

Output: Component functions f_j and estimator $\widehat{m}(X_i) = \sum_j f_j(X_{ij})$.

Figure 1: THE SPAM BACKFITTING ALGORITHM

Outline

- 1) The SpAM Optimization Problem
- 2) A backfitting algorithm
- 3) Properties of SpAM
- 4) Simulations showing the estimator's behavior

Properties of SpAM

- 1) SpAM is Presistent
- 2) SpAM is Sparsistent

Properties of SpAM

- 1) SpAM is Presistent

  Presistence comes from shortening “Predictive consistency”

  Def: Let (X,Y) be a new pair of data and the predictive risk when predicting Y with f(X)

$$R(f) = \mathbb{E}(Y - f(X))^2$$

Properties of SpAM

- 1) SpAM is Presistent

  Def: Let (X,Y) be a new pair of data and the predictive risk when predicting Y with f(X)

$$R(f) = \mathbb{E}(Y - f(X))^2$$

  we say an estimator is presistent relative to a class of functions M_n if

$$R(\widehat{m}_n) - R(m_n^*) \xrightarrow{P} 0$$

  where: $m_n^* = \operatorname{argmin}_{f \in \mathcal{M}_n} R(f)$

The SpAM Optimization Problem

- 2) SpAM is Sparsistent

  Def: the support of to be the location of the nonzero elements:

  supp( ) = { j: $_j$!=0 }

  Then the estimate of is sparsistent if

$$\mathbb{P}(\mathrm{supp}(\hat{\beta}) = \mathrm{supp}(\beta)) \to 1$$

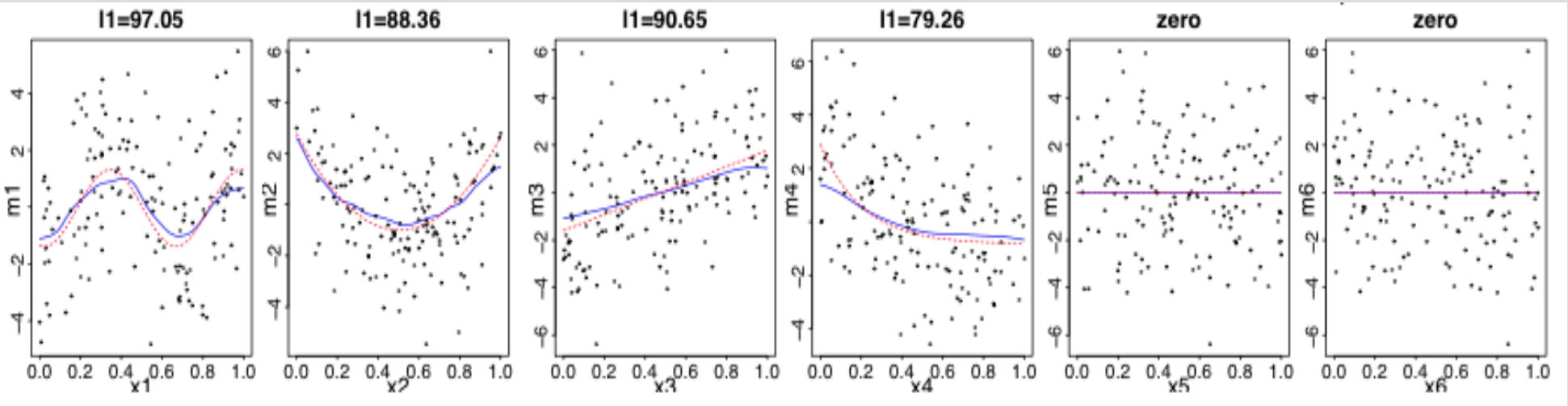
Outline

- 1) The SpAM Optimization Problem
- 2) A backfitting algorithm
- 3) Properties of SpAM
- 4) Simulations showing the estimator's behavior

Experiments

- A simulations dataset

  sample size n=150, generated from a 200 dimens-ional additive model. (196 irrelevant dimentsions).



Experiments

- Boston Housing

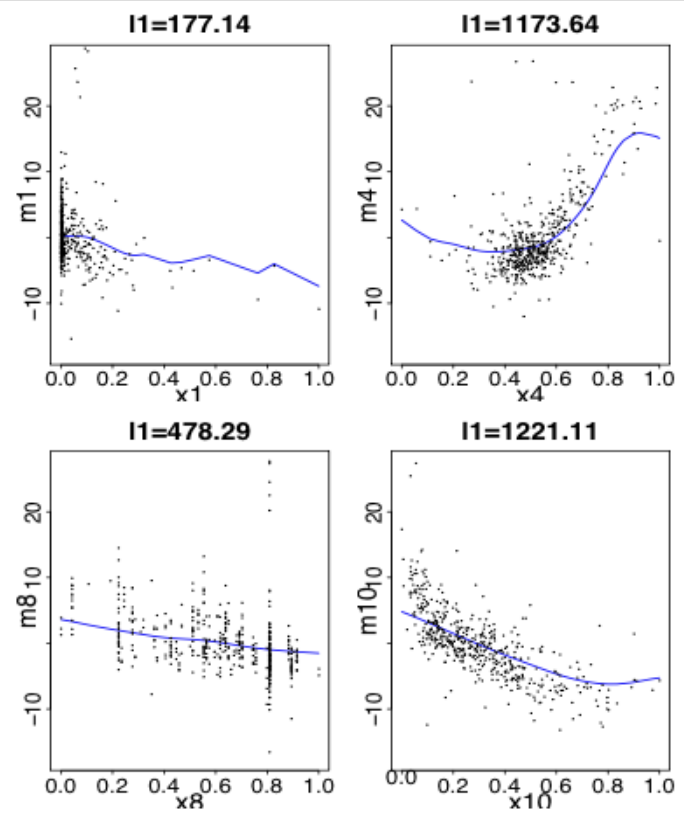
(506 observations with 10 covariates)

Then added 20 irrelevant variables:

1) 10 for randomly drawn from uniform(0,1)

2) 10 for random permutation of the original ten covariates

- Result shows the SpAM correctly zeros out both types of irrelevant variables, identifies 6 nonzero components



- Thank you!