SpAM: Sparse Additive Models

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

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

- How sparsity is achieved

Some modification:

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

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

$$\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} \mathbb{E} \left( Y - \sum_{j=1}^{p} \beta_j g_j(X_j) \right)^2$$
Consider following modification that imposes additional constraints:

\[
\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} \mathbb{E} \left( Y - \sum_{j=1}^{p} \beta_j g_j(X_j) \right)^2 \\
\text{subject to} \\
\sum_{j=1}^{p} |\beta_j| \leq L \\
\mathbb{E} (g_j^2) = 1, \ j = 1, \ldots, p \\
\mathbb{E} (g_j) = 0, \ j = 1, \ldots, p
\]
L1 encourages sparsity of estimated beta

\[
\begin{align*}
\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} & \quad \mathbb{E} \left( Y - \sum_{j=1}^{p} \beta_j g_j(X_j) \right)^2 \\
\text{subject to} & \quad \sum_{j=1}^{p} |\beta_j| \leq L \\
& \quad \mathbb{E} \left( g_j^2 \right) = 1, \quad j = 1, \ldots, p \\
& \quad \mathbb{E} \left( g_j \right) = 0, \quad j = 1, \ldots, p
\end{align*}
\]

\[
\{ \beta : \|\beta\|_1 \leq L \}
\]

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

\[
\begin{align*}
\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} & \quad \mathbb{E} \left( Y - \sum_{j=1}^{p} \beta_j g_j(X_j) \right)^2 \\
\text{subject to} & \quad \sum_{j=1}^{p} |\beta_j| \leq L \\
& \quad \mathbb{E} \left( g_j^2 \right) = 1, \quad j = 1, \ldots, p \\
& \quad \mathbb{E} \left( g_j \right) = 0, \quad j = 1, \ldots, p
\end{align*}
\]

- First call the original optimization problem as P
A related optimization problem, called Q:

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

This problem is convex in \( \{f\} \) and the problem P and Q are equivalent.
The problem P and Q are equivalent:

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

- 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 provide 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 \( \pi_{12}C \) onto first two components (L2)
The projection \( \pi_{13}C \) 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 (smoothness)

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 $Q$

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

\[
\mathcal{L}(f, \lambda, \mu) = \frac{1}{2} \mathbb{E} \left( Y - \sum_{j=1}^{p} f_j(X_j) \right)^2 + \lambda \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(X_j))} + \sum_{j} \mu_j \mathbb{E}(f_j).
\]
A backfitting algorithm

- Define the $j^{th}$ 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:

$$ v_j = f_j / \sqrt{\mathbb{E}(f_j^2)} $$
A backfitting algorithm

- Conditioning on $X_j$, the Frechet derivative becomes:

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

- Letting $P_j = \mathbb{E}[ R_j | X_j ]$ 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: $v_j = f_j / \sqrt{\mathbb{E}(f_j^2)}$)
A backfitting algorithm

- This form

\[ \left( 1 + \frac{\lambda}{\sqrt{\mathbb{E}(f^2_j)}} \right) f_j = P_j \quad \text{if} \quad \mathbb{E}(P^2_j) > \lambda \]

implies

\[ \left( 1 + \frac{\lambda}{\sqrt{\mathbb{E}(f^2_j)}} \right) \sqrt{\mathbb{E}(f^2_j)} = \sqrt{\mathbb{E}(P^2_j)} \]

or

\[ \sqrt{\mathbb{E}(f^2_j)} = \sqrt{\mathbb{E}(P^2_j)} - \lambda. \]
A backfitting algorithm

- From the form,

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

or

\[
\sqrt{\mathbb{E}(f_j^2)} = \sqrt{\mathbb{E}(P_j^2)} - \lambda.
\]

- we arrive the following soft-thresholding update

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

$S_j$ is a linear smoother, such as a local linear or kernel smoother
Two terms are needed to be estimated

2) A simple but biased estimate for the denominator:

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

\[
\hat{s}_j = \frac{1}{\sqrt{n}} \| \hat{P}_j \|_2 = \sqrt{\text{mean}(\hat{P}_j^2)}.
\]
A backfitting algorithm

- We derived the SpAM backfitting algorithm

<table>
<thead>
<tr>
<th>Input: Data ((X_i, Y_i)), regularization parameter (\lambda).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialize (f_j = f_j^{(0)}), for (j = 1, \ldots, p).</td>
</tr>
<tr>
<td>Iterate until convergence:</td>
</tr>
<tr>
<td>For each (j = 1, \ldots, p):</td>
</tr>
<tr>
<td>Compute the residual: (R_j = Y - \sum_{k \neq j} f_k(X_k));</td>
</tr>
<tr>
<td>Estimate the projection (P_j = \mathbb{E}[R_j</td>
</tr>
<tr>
<td>Estimate the norm (s_j = \sqrt{\mathbb{E}[P_j]^2}) using, for example, (15) or (35);</td>
</tr>
<tr>
<td>Soft-threshold: (f_j = \left[1 - \frac{\lambda}{\hat{s}<em>j}\right]</em>+ \hat{P}_j);</td>
</tr>
<tr>
<td>Center: (f_j \leftarrow f_j - \text{mean}(f_j)).</td>
</tr>
<tr>
<td>Output: Component functions (f_j) and estimator (\hat{m}(X_i) = \sum_j f_j(X_{ij})).</td>
</tr>
</tbody>
</table>

**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 Persistent

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 Persistent

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 persistent relative to a class of functions \(M_n\) if

\[
R(\hat{m}_n) - R(m_n^*) \xrightarrow{P} 0
\]

where:

\[
m_n^* = \arg\min_{f \in M_n} R(f)
\]
2) SpAM is Sparsistent

Def: the support of \( \beta \) to be the location of the nonzero elements:

\[
\text{supp}(\beta) = \{ j : j \neq 0 \}
\]

Then the estimate of \( \beta \) is sparsistent if

\[
P(\text{supp}(\hat{\beta}) = \text{supp}(\beta)) \to 1
\]
Outline

1) The SpAM Optimization Problem
2) A backfitting algorithm
3) Properties of SpAM
4) Simulations showing the estimator's behavior
A simulations dataset
sample size n=150, generated from a 200 dimensional additive model. (196 irrelevant dimensions).
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!