Convex Formulations for Flexible and Interpretable Non-Linear Modeling

Daniela Witten
Departments of Statistics & Biostatistics
University of Washington

June 22, 2016
Happiness Data

- Country-level data on 109 countries
- Outcome: happiness index from Cantril Scale
- Twelve predictors:
  - Log gross national income
  - Log scientific journal articles published
  - Percent satisfied with freedom of choice
  - Percent satisfied with job
  - . . . and more . . .
Additive Regression Models

- Response $Y$, features $X_1, \ldots, X_p$.
- $Y = f_1(X_1) + \cdots + f_p(X_p) + \epsilon$.
- How to estimate $f_1, \ldots, f_p$?
Additive Model Using Smoothing Splines\textsuperscript{1}

\textsuperscript{1}Hastie & Tibshirani, \textit{Generalized Additive Models}, 1990.
Recent Work on Additive Models

- A lot of recent papers consider the problem of fitting an additive model in high dimensions\(^2\).
- The additive functions, \(f_1, \ldots, f_p\), are typically assumed to be smooth. This leads to challenges in interpretation.

\(^2\)Ravikumar et al., JRSSB 2009; Huang et al., Annals of Statistics 2010; Zhang et al., JASA 2011; Lou et al., JCGS 2016; and more.
Fused Lasso Additive Model (FLAM)

We want a model that is

- **flexible**: can accommodate non-linear relationships.
- **interpretable**: no harder to interpret than linear regression.
Fused Lasso Additive Model (FLAM) (FLAM)

For $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$, fit the model

$$Y = f_1(X_1) + \ldots + f_p(X_p) + \epsilon,$$

where the fitted functions $\hat{f}_1, \ldots, \hat{f}_p$ are piecewise constant with a small number of adaptively-chosen knots.
What If We Only Had One Covariate?

- Estimating $f$ is equivalent to making predictions at each $x_i$.
- Call these predictions $\hat{\theta}$ where $\hat{\theta}_i = \hat{f}(x_i)$.
What If We Only Had One Covariate?

- Estimating $f$ is equivalent to making predictions at each $x_i$.
- Call these predictions $\hat{\theta}$ where $\hat{\theta}_i = \hat{f}(x_i)$.
What If We Only Had One Covariate?

- Estimating $f$ is equivalent to making predictions at each $x_i$.
- Call these predictions $\hat{\theta}$ where $\hat{\theta}_i = \hat{f}(x_i)$.
- Knots occur when $\hat{\theta}_i \neq \hat{\theta}_{i+1}$.
Optimization Problem With One Covariate

▶ Solve

\[
\min_{\theta \in \mathbb{R}^n} \frac{1}{2} \| y - \theta \|_2^2 + \lambda \| D \theta \|_1
\]

where

\[
D \theta = \begin{pmatrix}
1 & -1 & 0 & \cdots & 0 & 0 \\
0 & 1 & -1 & \cdots & 0 & 0 \\
\vdots & & & & & \\
0 & 0 & 0 & \cdots & 1 & -1
\end{pmatrix} \begin{pmatrix}
\theta_1 \\
\theta_2 \\
\vdots \\
\theta_{n-1} \\
\theta_n
\end{pmatrix} = \begin{pmatrix}
\theta_1 - \theta_2 \\
\theta_2 - \theta_3 \\
\vdots \\
\theta_{n-1} - \theta_n
\end{pmatrix}
\]

▶ The non-zero elements of \( D \theta \) correspond to knots.
▶ This is the fused lasso\(^3\).

\(^3\)Tibshirani et al., JRSSB 2005
Estimating $\theta$

\[
\min_{\theta \in \mathbb{R}^n} \frac{1}{2} \| y - \theta \|_2^2 + \lambda \| D\theta \|_1
\]
Estimating $\theta$

\[
\min_{\theta \in \mathbb{R}^n} \frac{1}{2} \| y - \theta \|_2^2 + \lambda \| D\theta \|_1
\]
Extension to Multiple Covariates

- **Single (Ordered) Covariate:**

\[
\min_{\theta \in \mathbb{R}^n} \frac{1}{2} \| y - \theta \|_2^2 + \lambda \| D\theta \|_1
\]

4Yuan and Lin, JRSSB 2007
Extension to Multiple Covariates

▶ Single (Ordered) Covariate:

\[
\min_{\theta \in \mathbb{R}^n} \frac{1}{2} \| y - \theta \|_2^2 + \lambda \| D\theta \|_1
\]

▶ Multiple Covariates:

\[
\min_{\theta_0 \in \mathbb{R}, \theta_j \in \mathbb{R}^n, 1 \leq j \leq p} \frac{1}{2} \left\| y - \sum_{j=1}^{p} \theta_j - \theta_0 1 \right\|_2^2 + \lambda \sum_{j=1}^{p} \| DP_j \theta_j \|_1
\]

where \( P_j \) orders \( x_j \) from least to greatest.

\(^4\text{Yuan and Lin, JRSSB 2007}\)
Extension to Multiple Covariates

- **Single (Ordered) Covariate:**

  \[
  \min_{\theta \in \mathbb{R}^n} \frac{1}{2} \left\| y - \theta \right\|_2^2 + \lambda \left\| D\theta \right\|_1
  \]

- **Multiple Covariates:**

  \[
  \min_{\theta_0 \in \mathbb{R}, \theta_j \in \mathbb{R}^n, 1 \leq j \leq p} \frac{1}{2} \left\| y - \sum_{j=1}^{p} \theta_j - \theta_0 \mathbf{1} \right\|_2^2 + \lambda \sum_{j=1}^{p} \left\| DP_j \theta_j \right\|_1
  \]

  where \( P_j \) orders \( x_j \) from least to greatest.

- **Add in Sparsity:**

  \[
  \min_{\theta_0 \in \mathbb{R}, \theta_j \in \mathbb{R}^n} \frac{1}{2} \left\| y - \sum_{j=1}^{p} \theta_j - \theta_0 \mathbf{1} \right\|_2^2 + \alpha \lambda \sum_{j=1}^{p} \left\| DP_j \theta_j \right\|_1 + (1 - \alpha) \lambda \sum_{j=1}^{p} \left\| \theta_j \right\|_2
  \]

\[4\text{Yuan and Lin, JRSSB 2007}\]
Fused Lasso Additive Model (FLAM)  
Convex Regression with Interpretable Sharp Partitions (CRISP)

Block Coordinate Descent Algorithm\(^5\) for FLAM

- Initialize \( \hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_p \).
- Iterate for \( j = 1, \ldots, p \) until convergence:
  - Compute the residual \( r_j = y - \hat{\theta}_0 1 - \sum_{j' \neq j} \hat{\theta}_{j'} \).
  - Solve

\[
\text{minimize} \left\{ \frac{1}{2} \| r_j - \theta_j \|^2 + \alpha \lambda \| DP_j \|_1 + (1 - \alpha) \lambda \| \theta_j \|_2 \right\}.
\]

Proposition

The solution to the optimization problem

$$\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \| \mathbf{y} - \theta \|^2 + \alpha \lambda \| D\theta \|_1 + (1 - \alpha) \lambda \| \theta \|_2 \right\}$$

is $\left( 1 - \frac{(1-\alpha)\lambda}{\| \hat{\theta} \|_2} \right) + \hat{\theta}$, where $(u)_+ = \max(u, 0)$ and $\hat{\theta}$ is the solution to

$$\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \| \mathbf{y} - \theta \|^2 + \alpha \lambda \| D\theta \|_1 \right\}.$$
Block Coordinate Descent Algorithm\(^7\) for FLAM

- Initialize \(\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_p\).
- Iterate for \(j = 1, \ldots, p\) until convergence:
  - Compute the residual \(r_j = y - \hat{\theta}_0 1 - \sum_{j' \neq j} \hat{\theta}_{j'}\).
  - Use an algorithm for the fused lasso\(^6\) to solve
    \[
    \minimize_{\theta_j \in \mathbb{R}^n} \left\{ \frac{1}{2} \| r_j - \theta_j \|^2 + \alpha \lambda \| D P_j \|_1 \right\}.
    \]
  - Compute the intercept, \(\hat{\theta}_0 \leftarrow \hat{\theta}_0 + \text{mean}(\hat{\theta}_j)\), and center, \(\hat{\theta}_j \leftarrow \hat{\theta}_j - \text{mean}(\hat{\theta}_j)\).
  - Soft-scale the estimate: \(\hat{\theta}_j \leftarrow \left(1 - \frac{(1 - \alpha) \lambda}{\| \hat{\theta}_j \|_2}\right) + \hat{\theta}_j\).

---

\(^6\)Hoefling, JCGS 2010; Johnson, JCGS 2013

\(^7\)Tseng, Journal of Optimization Theory and Applications 2001
Simulation Results: Piecewise Constant Functions

- $n = 100$, $p = 100$, four non-noise features.
- Comparison to SpAM\textsuperscript{8}, for which basis functions are natural cubic splines with $d - 1$ non-boundary knots.

\textsuperscript{8}Sparse Additive Models of Ravikumar et al., JRSSB 2009
Simulation Results: Piecewise Constant Functions

- $n = 100$, $p = 100$, four non-noise features.
- Comparison to SpAM\(^8\), for which basis functions are natural cubic splines with $d - 1$ non-boundary knots.

\(^8\)Sparse Additive Models of Ravikumar et al., JRSSB 2009
Simulation Results: Piecewise Constant Functions

- $n = 100$, $p = 100$, four non-noise features.
- Comparison to SpAM\textsuperscript{8}, for which basis functions are natural cubic splines with $d - 1$ non-boundary knots.

\textsuperscript{8}Sparse Additive Models of Ravikumar et al., JRSSB 2009
Simulation Results: Smooth Functions

- $n = 100$, $p = 100$, four non-noise features.
- Comparison to SpAM, for which basis functions are natural cubic splines with $d - 1$ non-boundary knots.
Simulation Results: Smooth Functions

- $n = 100$, $p = 100$, four non-noise features.
- Comparison to SpAM, for which basis functions are natural cubic splines with $d - 1$ non-boundary knots.
Simulation Results: Smooth Functions

- $n = 100$, $p = 100$, four non-noise features.
- Comparison to SpAM, for which basis functions are natural cubic splines with $d - 1$ non-boundary knots.
Simulation Results: Smooth Functions

- **FLAM** and **SpAM** fits.
- FLAM is better able to adapt its flexibility to the data!
- So it can perform better even when $f_1, \ldots, f_p$ are smooth.
Using FLAM to Predict Happiness

- Life expectancy at birth (years)
- Log scientific journal articles published

Graphs showing the relationship between life expectancy and the change in mean happiness index, and the relationship between log scientific journal articles and the change in mean happiness index.
**Motivation:** Can we retain the interpretability of FLAM, without the restriction to additivity?

**Idea:** Fit a regression model with a “block” structure.
Example of CRISP
Example of CRISP
Example of CRISP

Data With Grid

CRISP Fit
Example of CRISP

Mean Model

CRISP Fit
Set-Up And Notation

- \( Y = f(X_1, X_2) + \epsilon. \)
- Assume that \( f \) is constant within each bin of a \( q \times q \) grid.
- Can summarize \( f \) using \( M \in \mathbb{R}^{q \times q} \): \( Y = \Omega(M, X_1, X_2) + \epsilon. \)

- Estimate \( q^2 \) elements of \( M \) while encouraging adjacent rows and columns to be equal, leading to a **block structure**.
CRISP’s Optimization Problem

\[
\min_{M \in \mathbb{R}^{q \times q}} \quad \frac{1}{2} \sum_{i=1}^{n} (y_i - \Omega(M, x_{1i}, x_{2i}))^2 \\
+ \lambda \sum_{i=1}^{q-1} \left[ \| M_i - M_{(i+1)} \|_2 + \| M_i - M_{(i+1)} \|_2 \right]
\]

- Encourage \( M \) to **fit the observed data**
- Encourage adjacent **rows** and **columns** of \( M \) to be equal
Encourage Adjacent Rows & Columns of $\mathbf{M}$ to be Equal

Data With Grid

CRISP Fit

Together, $\lambda \geq 0$ and $q \in \{2, \ldots, n\}$ control granularity of fit.
ADMM Algorithm

Straightforward application of ADMM\(^9\).

\(^9\)Boyd et al., Foundations & Trends in Machine Learning 2011
"Block" Mean Model\textsuperscript{10}

\textsuperscript{10}Comparisons to Thin-Plate Splines (TPS; Duchon 1977) and CART (Breiman et al. 1984).
“Smooth” Mean Model
“Smooth” Mean Model
“Smooth” Mean Model
“Smooth” Mean Model
“Smooth” Mean Model
"Smooth" Mean Model
Summary

**FLAM ($p > n$):**
- Flexible: adaptive selection of covariates and knots
- Interpretable: simple piecewise constant fits

**CRISP ($p \ll n$):**
- Flexible: adaptive partitioning of covariate space
- Interpretable: “block” structure of pairwise fits
For More Information . . .