Published in final edited form as:

J Comput Graph Stat. 2016 ; 25(4): 1005–1025. doi:10.1080/10618600.2015.1073155.

Fused Lasso Additive Model

Ashley Petersen^{*}, Daniela Witten[†], and Noah Simon[‡]

Department of Biostatistics, University of Washington, Seattle WA 98195

Abstract

We consider the problem of predicting an outcome variable using p covariates that are measured on n independent observations, in a setting in which additive, flexible, and interpretable fits are desired. We propose the *fused lasso additive model* (FLAM), in which each additive function is estimated to be piecewise constant with a small number of adaptively-chosen knots. FLAM is the solution to a convex optimization problem, for which a simple algorithm with guaranteed convergence to a global optimum is provided. FLAM is shown to be consistent in high dimensions, and an unbiased estimator of its degrees of freedom is proposed. We evaluate the performance of FLAM in a simulation study and on two data sets. Supplemental materials are available online, and the R package *flam* is available on CRAN.

Keywords

additive model; feature selection; high-dimensional; non-parametric regression; piecewise constant; sparsity

1 Introduction

In this paper, we consider the task of predicting a response variable using p features measured on n independent observations. We restrict ourselves to the setting of *additive* models. Additive models have a number of attractive properties:

- Additive models are generally more interpretable than models involving interactions. In an additive model, associations between each covariate and the outcome can be described individually. In contrast, fits from a model involving interactions are more challenging to summarize.
- Additive models often win out over non-additive models in terms of generalization error: allowing interactions in the model typically reduces bias at

^{*} ajpete@uw.edu

[†] dwitten@uw.edu

[‡] nrsimon@uw.edu

Supplementary Material

R Code and Data: The supplementary material includes R code, which in conjunction with the R package *flam* available on CRAN, can be used to reproduce all of the results given in the paper. The happiness data set used in Section 6.1 is also included. The three gene expression data sets are publicly available from NCBI's GEO website using the accession numbers given in Section 6.2. The README file contains additional details. (DataCode.zip)

Appendix: The supplementary material includes the Appendix, which gives the proofs for all results given in Sections 4 and 7. (Appendix.pdf)

the price of increased variance. Especially in the high-dimensional setting, the restriction to additivity can be crucial in controlling the bias-variance trade-off.

Additive models typically offer either interpretability but limited flexibility (for instance, linear regression or a piecewise constant model with pre-specified knots) or flexibility but limited interpretability (for instance, an approach based on smoothing splines). In this paper, we propose a method that balances the trade-off between interpretability and flexibility, while also allowing for sparsity in high dimensions when $p > n$. It selects a subset of features to include in the model, and for these features it fits piecewise constant functions with knots that are chosen *adaptively* based on the data.

We now introduce some notation. We let $\mathbf{X}$ denote an $n \times p$ matrix, for which $\mathbf{x}_j$ is the j th column (feature), and for which the i th element (observation) is x_{ij} . When we consider the case of $p = 1$, we use $\mathbf{x}$ to denote the single feature, with i th element x_i . The response is an n -vector $\mathbf{y}$, with i th element y_i . To reference subvectors, we use $\mathbf{a}_{\mathcal{S}}$ to denote $\mathbf{a}$ with only the elements contained in the set $\mathcal{S}$. To reference submatrices, we use $\mathbf{A}_{\mathcal{S}}$ to denote $\mathbf{A}$ with only the columns contained in the set $\mathcal{S}$.

The rest of this paper is organized as follows. In Section 2, we review related work. In Sections 3 and 4, we propose our method, present an algorithm to implement it, and examine some of its properties. Sections 5 and 6 contain the results of a simulation study and the analyses of two data sets. We consider some extensions in Section 7, and we close with a discussion in Section 8. Proofs are in the Appendix.

2 Previous Work

Generalized additive models (GAM) provide a flexible and general framework for modeling a response in low dimensions ($n > p$). We assume $E[y_i | \mathbf{x}_i] = g\left(\sum_{j=1}^p f_j(x_{ij})\right)$, where g is a specified function and each f_j is an unknown function that we wish to estimate [Hastie and Tibshirani, 1986]. For now, we restrict our attention to the case when g is the identity function. There are a number of ways to estimate f_j — for example, we might use a smoothing or regression spline.

Flexible additive modeling in high dimensions has been an active area of research in recent years [Avalos et al., 2007, Bühlmann and Hothorn, 2007, Bühlmann and Yu, 2003, Huang et al., 2010, Lin and Zhang, 2006, Sardy and Tseng, 2004, Wood et al., 2014, Zhang et al., 2011]. Recently, Ravikumar et al. [2009] proposed a high-dimensional extension of GAM called sparse additive models (SpAM), which induces sparsity in the function estimates using a standardized group lasso penalty [Simon and Tibshirani, 2012]. SpAM solves the problem minimize

$$\underset{\beta_j \in \mathbb{R}^d, 1 \leq j \leq p}{\text{minimize}} \quad \frac{1}{2n} \left\| \mathbf{y} - \sum_{j=1}^p \Psi_j \beta_j \right\|_2^2 + \lambda \sum_{j=1}^p \sqrt{\frac{1}{n} \beta_j^T \Psi_j^T \Psi_j \beta_j}, \quad (1)$$

where $\beta_j = (\beta_{j1} \cdots \beta_{jd})^T$ is a d -vector of coefficients, $\lambda \geq 0$ is a tuning parameter, and $\Psi_j = [\psi_{j1} \cdots \psi_{jd}]$ is an $n \times d$ matrix of which the columns are the d basis functions used to model f_j . The penalty in (1) can be seen as the data-level version of the population-level penalty

$$\sqrt{\mathbb{E}[f_j^2(x)]} = \sqrt{\int f_j^2(x) dP(x)}.$$

Meier et al. [2009] modified (1) in order to obtain data-adaptive fits that can capture complex relationships if needed, but that otherwise are smooth. Their estimator is the solution to the optimization problem

$$\underset{\beta_j \in \mathbb{R}^d, 1 \leq j \leq p}{\text{minimize}} \quad \frac{1}{n} \left\| \mathbf{y} - \sum_{j=1}^p \Psi_j \beta_j \right\|_2^2 + \lambda_1 \sum_{j=1}^p \sqrt{\frac{1}{n} \beta_j^T \Psi_j^T \Psi_j \beta_j} + \lambda_2 \sum_{j=1}^p \sqrt{\beta_j^T \mathbf{W}_j \beta_j}, \quad (2)$$

where $\beta_j = (\beta_{j1} \cdots \beta_{jd})^T$ is a d -vector of coefficients to be estimated, $\lambda_1, \lambda_2 \geq 0$ are tuning parameters, $\Psi_j = [\psi_{j1} \cdots \psi_{jd}]$ is an $n \times d$ matrix of which the columns are the cubic B-spline basis vectors of the j th predictor, and $\mathbf{W}_j$ is a $d \times d$ matrix containing the inner products of the second derivatives of the cubic B-spline basis functions [de Boor, 1978]. The penalties in (2) can be seen as the data-level versions of the population-level penalties

$\sqrt{\int f_j^2(x) dP(x)}$ and $\sqrt{\int (f_j''(x))^2 dx}$. Other variations of this sparsity-smoothness penalty have also been proposed [Bühlmann and van de Geer, 2011, Meier et al., 2009].

Recently, Lou et al. [2014] proposed the sparse partially linear additive model, which allows some features to have a linear fit and others to have a non-linear fit. The linear features do not need to be chosen *a priori*.

3 The Fused Lasso Additive Model

The methods described in Section 2 rely on a pre-specified set of basis functions. This limits flexibility, since the basis functions must be chosen *a priori* rather than in a data-adaptive way, as well as interpretability, since for many choices of basis functions (e.g., natural cubic splines) the resulting fits are complex and non-monotonic without clear change points.

We now propose an approach to fit an additive model in which each function is estimated to be piecewise constant with a small number of knots. While this problem is easily solved when the knots are chosen *a priori*, our proposal allows the knots to be chosen adaptively.

Our method is very flexible due to the adaptively-chosen knots, and the piecewise constant fits are interpretable. Piecewise constant models have many practical applications, such as developing risk scoring systems for use in medical settings. Patients can obtain their risk scores by identifying which risk groups they fall in for each risk factor, and then tallying up their total risk score. An example is provided in Van Belle et al. [2012].

3.1 The Optimization Problem

To begin, we assume that we have a single feature $\mathbf{x}$ that is ordered, i.e., $x_1 < x_2 < \dots < x_n$. We wish to estimate an n -vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, where $E[y_i|x_i] = f(x_i) = \theta_i$. We obtain a piecewise constant estimate of $\boldsymbol{\theta}$ with a small number of knots by solving the problem

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{y} - \boldsymbol{\theta}\|_2^2 + \lambda \|\mathbf{D}\boldsymbol{\theta}\|_1, \quad (3)$$

where $\lambda \geq 0$ is a tuning parameter and $\mathbf{D}$ is the first difference matrix,

$$\mathbf{D} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}.$$

We let $\hat{\boldsymbol{\theta}}$ denote the solution to (3). The ℓ_1 penalty encourages $|\hat{\theta}_{i-1} - \hat{\theta}_i|$ to equal zero when λ is large. The non-zero elements of $|\hat{\theta}_{i-1} - \hat{\theta}_i|$ correspond to knots in $\hat{\boldsymbol{\theta}}$. Consequently, $\hat{\boldsymbol{\theta}}$ provides a piecewise constant fit to the data, with adaptively-chosen knots. The addition of an ℓ_1 penalty on $\boldsymbol{\theta}$ to (3) gives the fused lasso [Tibshirani et al., 2005]. Several algorithms for solving (3) have been proposed [Hoeftling, 2010, Johnson, 2013, Liu et al., 2010].

We now consider the model $E[y_i|x_i] = \sum_{j=1}^p f_j(x_{ij}) = \sum_{j=1}^p \theta_{ji}$. We assume that each $f_j(\cdot)$ is piecewise constant, mean-zero over the training observations, and we include an intercept θ_0 . Let $\mathbf{P}_j$ denote a permutation matrix that orders the elements of the vector $\mathbf{x}_j$ from least to greatest. We can then solve the problem

$$\underset{\theta_0 \in \mathbb{R}, \boldsymbol{\theta}_j \in \mathbb{R}^n, 1 \leq j \leq p}{\text{minimize}} \quad \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^p \boldsymbol{\theta}_j - \theta_0 \mathbf{1} \right\|_2^2 + \lambda \sum_{j=1}^p \|\mathbf{D}\mathbf{P}_j \boldsymbol{\theta}_j\|_1 \quad \text{subject to } \mathbf{1}^T \boldsymbol{\theta}_j = 0 \quad \forall j. \quad (4)$$

In high dimensions, we may wish to impose sparsity on the $\boldsymbol{\theta}_j$'s, so that a given feature is completely excluded from the model. For λ sufficiently large, the solution to (4) is sparse. However, this value of λ will tend to overshrink all of the estimates for $\boldsymbol{\theta}_j$, since the fused lasso penalty discourages large jumps in $\boldsymbol{\theta}_j$ (see Section 5 for related simulations). Therefore, we consider the modified optimization problem

$$\underset{\theta_0 \in \mathbb{R}, \boldsymbol{\theta}_j \in \mathbb{R}^n, 1 \leq j \leq p}{\text{minimize}} \quad \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^p \boldsymbol{\theta}_j - \theta_0 \mathbf{1} \right\|_2^2 + \alpha \lambda \sum_{j=1}^p \|\mathbf{D}\mathbf{P}_j \boldsymbol{\theta}_j\|_1 + (1-\alpha) \lambda \sum_{j=1}^p \|\boldsymbol{\theta}_j\|_2, \quad (5)$$

where $\lambda \geq 0$ and $0 \leq \alpha \leq 1$. Here, α provides a trade-off between encouraging $\hat{\theta}_j$ to be piecewise constant, and inducing sparsity on the entire vector $\hat{\theta}_j$ using the group lasso [Yuan and Lin, 2006]. We refer to the solution to (5) as the *fused lasso additive model* (FLAM).

We note that (4) involved constraints $\mathbf{1}^T \theta_j = 0$ for $j = 1, \dots, p$, these constraints are omitted from (5). This is because any solution to (5) will automatically satisfy these constraints. This can be seen by noting that if a given θ_j is uncentered, then the objective of (5) can be reduced by absorbing the mean of that θ_j into the intercept θ_0 .

Replacing $\|\theta_j\|_2$ in (5) with $\|\theta_j\|_1$ yields an extension of the fused lasso problem [Tibshirani et al., 2005] to the setting with multiple fits, $\theta_1, \dots, \theta_p$. However, to induce complete feature-wise sparsity, we use an ℓ_2 penalty rather than an ℓ_1 penalty in (5).

In (5), we are using the same tuning parameter, λ , for each covariate. In so doing, we are penalizing jumps at knots in each feature equally, and also encouraging sparsity of each feature equally. In certain settings, we may prefer to instead allow a different tuning parameter λ_j for each covariate. This necessitates only a minor modification to Algorithm 1, presented in Section 3.2. However, in practice, allowing a different tuning parameter for each covariate may lead to difficulties in tuning parameter selection.

In contrast to most other regularized regression approaches, such as the lasso, the FLAM proposal (5) is scale-invariant: rescaling the features does not affect the optimization problem or the fitted values.

We propose to select FLAM's tuning parameters, λ and α , via K -fold cross-validation. As an alternative, an approach for tuning parameter selection based on Akaike's information criterion [AIC; Akaike, 1973] or Bayesian information criterion [BIC; Schwarz, 1978] could be developed. Such an approach would require an estimator for the degrees of freedom of FLAM, which we will present in Section 4.1.

3.2 An Algorithm for FLAM

Problem 5 is convex, and so can be solved using a general-purpose interior point method. Here we develop a much faster backfitting algorithm using block coordinate descent to solve (5) [Tseng, 2001]. We cycle through the features and repeatedly perform a partial minimization in a single θ_j , holding all others fixed. The solution for the partial minimization is given in Corollary 3.1, which follows from a more general result presented in Section 7.3. This corollary allows us to solve the FLAM optimization problem in $\mathcal{O}(n)$ operations per feature per iteration, by leveraging an existing fused lasso solver that requires $\mathcal{O}(n)$ operations for an n -dimensional problem [Johnson, 2013].

Corollary 3.1—The solution to the optimization problem

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{y} - \theta\|_2^2 + \alpha \lambda \|D\theta\|_1 + (1 - \alpha) \lambda \|\theta\|_2 \quad (6)$$

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

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|y - \theta\|_2^2 + \alpha \lambda \|D\theta\|_1. \quad (7)$$

Corollary 3.1 leads directly to Algorithm 1, which can be made very efficient using warm starts and active sets. Algorithm 1 yields the global optimum to (5) [Tseng, 2001] for fixed values of FLAM's non-negative tuning parameters, λ and α .

Algorithm 1

— Block Coordinate Descent for Fused Lasso Additive Model (Equation 5)

-
- 1 Initialize $\hat{\theta}_0 = 0$ and $\hat{\theta}_j = 0$ for all $j = 1, \dots, p$.
 - 2 For each $j = 1, \dots, p$, perform the following:
 - a. Compute the residual $r_j = y - \hat{\theta}_0 \mathbf{1} - \sum_{j' \neq j} \hat{\theta}_{j'}$.
 - b. Using an algorithm for the fused lasso (e.g., flsa on CRAN [Hoeftling, 2013]), solve

$$\hat{\theta}_j \leftarrow \underset{\theta_j \in \mathbb{R}^n}{\text{argmin}} \quad \frac{1}{2} \|r_j - \theta_j\|_2^2 + \alpha \lambda \|DP_j \theta_j\|_1.$$

- c. Compute the intercept, $\hat{\theta}_0 \leftarrow \hat{\theta}_0 + \text{mean}(\hat{\theta}_j)$, and center, $\hat{\theta}_j \leftarrow \hat{\theta}_j - \text{mean}(\hat{\theta}_j)$.
- d.

$$\hat{\theta}_j \leftarrow \left(1 - \frac{(1-\alpha)\lambda}{\|\hat{\theta}_j\|_2}\right)_+ \hat{\theta}_j$$

Soft-scale the estimate:

- 3 Repeat Step 2 until convergence of the objective of (5).
-

On a Macbook Pro with a 2.0 GHz Intel Sandy Bridge Core i7 processor, fitting FLAM using our R package `flam` with $n = 2000$ and $p = 200$ takes 21.6 seconds for a sequence of 50 λ values. Similarly for $n = 50,000$ and $p = 50$, the run time is 86.4 seconds. Further computational improvements could be achieved by implementing Algorithm 1 using an active set approach.

3.3 Connections to Other Methods

The fused lasso can be interpreted as ℓ_1 trend filtering with order $k = 0$ [Kim et al., 2009, Tibshirani, 2014]. Tibshirani [2014] showed that 0th order trend filtering is equivalent to 0th order locally adaptive regression splines [Mammen and van de Geer, 1997]. Therefore, we can interpret FLAM with $\alpha = 1$ as a multi-variable extension of locally adaptive regression splines. Indeed, we illustrate FLAM's local adaptivity, or ability to produce a fit that is highly variable in one portion of the domain and constant in another, in Section 5.

When $\alpha = 0$, FLAM is equivalent to SpAM with $\Psi_j = \mathbf{I} \in \mathbb{R}_{n \times n}$ in (1). However, this is an impractical special case in which the design matrix does not depend on the covariates.

4 Properties of FLAM

We define $\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}$ where $\bar{y} = \frac{1}{n}\mathbf{1}^T \mathbf{y}$ and $\mathbf{V} = [\mathbf{P}_1^T \mathbf{U} \cdots \mathbf{P}_p^T \mathbf{U}]$, where $\mathbf{U} \in \mathbb{R}_{n \times (n-1)}$ is the matrix obtained by centering the columns of the upper triangular matrix of 1's, and

$$\mathbf{U} = \begin{pmatrix} 2/3 & 1/3 \\ -1/3 & 1/3 \\ -1/3 & -2/3 \end{pmatrix}$$

removing the n th column. For example, when $n = 3$. The following lemma indicates that FLAM can be reparameterized in terms of the pairwise differences among the ordered elements of $\boldsymbol{\theta}_j$ (i.e., the elements of $\boldsymbol{\beta}_j = \mathbf{D}\mathbf{P}_j \boldsymbol{\theta}_j$).

Lemma 4.1—Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T \cdots \hat{\boldsymbol{\beta}}_p^T)^T$ be the solution to

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{(n-1)p}}{\text{minimize}} \quad \frac{1}{2} \|\tilde{\mathbf{y}} - \mathbf{V}\boldsymbol{\beta}\|_2^2 + \alpha \lambda \|\boldsymbol{\beta}\|_1 + (1-\alpha)\lambda \sum_{j=1}^p \|\mathbf{U}\boldsymbol{\beta}_j\|_2. \quad (8)$$

Then the solution $\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \dots, \hat{\boldsymbol{\theta}}_p$ to the optimization problem (5) is

$$\hat{\boldsymbol{\theta}}_0 = \bar{\mathbf{y}} \text{ and } \hat{\boldsymbol{\theta}}_j = \mathbf{P}_j^T \mathbf{U} \hat{\boldsymbol{\beta}}_j \text{ for } j=1, \dots, p.$$

The proof of Lemma 4.1 can be found in Appendix A.1. From (8), FLAM with $\alpha = 1$ is equivalent to solving a lasso problem. The reparametrization given in Lemma 4.1 will allow us to easily derive some properties of FLAM.

4.1 Degrees of Freedom for FLAM

Suppose that $\text{Var}(\mathbf{y}) = \sigma^2 \mathbf{I}$, and let $g(\mathbf{y}) = \hat{\mathbf{y}}$ denote the fit corresponding to some model-fitting procedure g . Then the degrees of freedom of g is defined as $\frac{1}{\sigma^2} \sum_{i=1}^n \text{Cov}(y_i, \hat{y}_i)$ [Efron, 1986, Hastie and Tibshirani, 1990]. We now consider a modified version of FLAM,

$$\underset{\boldsymbol{\theta}_0 \in \mathbb{R}, \boldsymbol{\theta}_j \in \mathbb{R}^n, 1 \leq j \leq p}{\text{minimize}} \quad \frac{1}{2} \left\| \mathbf{y} - \sum_{j=1}^p \boldsymbol{\theta}_j - \boldsymbol{\theta}_0 \mathbf{1} \right\|_2^2 + \alpha \lambda \sum_{j=1}^p \|\mathbf{D}\mathbf{P}_j \boldsymbol{\theta}_j\|_1 + (1-\alpha)\lambda \sum_{j=1}^p \|\boldsymbol{\theta}_j\|_2 + \frac{\gamma}{2} \sum_{j=1}^p \|\mathbf{D}\mathbf{P}_j \boldsymbol{\theta}_j\|_2^2,$$

(9)

where $\gamma > 0$ is a small constant; the additional term in (9) ensures strict convexity and uniqueness of the solution. Subject to reparameterization, (9) is equivalent to

$$\underset{\beta \in \mathbb{R}^{(n-1)p}}{\text{minimize}} \quad \frac{1}{2} \|\tilde{y} - V\beta\|_2^2 + \alpha \lambda \|\beta\|_1 + (1-\alpha)\lambda \sum_{j=1}^p \|U\beta_j\|_2 + \frac{\gamma}{2} \sum_{j=1}^p \|\beta_j\|_2^2. \quad (10)$$

We propose to estimate the degrees of freedom of FLAM as

$$\hat{df}_{FLAM} = \text{Tr} \left(V_{\mathcal{A}} [V_{\mathcal{A}}^T V_{\mathcal{A}} + (1-\alpha)\lambda S_2 + \gamma I]^{-1} V_{\mathcal{A}}^T \right) + 1, \quad (11)$$

where $\mathcal{A} = \{i \in \{1, \dots, (n-1)p\} : \hat{\beta}_i \neq 0\}$ where $\hat{\beta}_i$ denotes the i th element of $\hat{\beta}$, and S_2 is

block diagonal with the j th block equal to $\frac{U_{\mathcal{A}_j}^T U_{\mathcal{A}_j} \hat{\beta}_{j\mathcal{A}_j} \hat{\beta}_{j\mathcal{A}_j}^T U_{\mathcal{A}_j}^T U_{\mathcal{A}_j}}{\|U_{\mathcal{A}_j} \hat{\beta}_{j\mathcal{A}_j}\|_2^3}$ with $\mathcal{A}_j = \{i \in \{1, \dots, n-1\} : \hat{\beta}_{ji} \neq 0\}$. Note that S_2 is derived in Appendix A.2.

Proposition 4.2—Assume $y \sim MVN(\mu, \sigma^2 I)$. Then $\hat{df}_{FLAM}$ is an unbiased estimator of the degrees of freedom of FLAM.

A proof sketch of Proposition 4.2 can be found in Appendix A.2. When $\alpha = 1$ and $\gamma = 0$, (11) reduces to $\text{rank}(V_{\mathcal{A}}) + 1$, which agrees with the estimator proposed in Tibshirani and Taylor [2012]. Recall that $\hat{\beta}_j = DP_j \hat{\theta}_j$, so $\hat{\beta}_{ji}$ is the difference between $[P_j \hat{\theta}_j]_{i-1}$ and $[P_j \hat{\theta}_j]_i$. Thus non-zero elements of $\hat{\beta}$ correspond to knots in the estimated fits. When $V_{\mathcal{A}}$ is full rank (which only occurs when the number of knots is smaller than n), the following corollary provides a simple estimator for FLAM's degrees of freedom.

Corollary 4.3—Assume $y \sim MVN(\mu, \sigma^2 I)$. Suppose that $V_{\mathcal{A}}$ is full rank, and let $\alpha = 1$ in (9). Then an unbiased estimate of the degrees of freedom of FLAM is one greater than the total number of knots across all estimated fits.

In 1000 replicate data sets, we compare the mean of (11) to the mean of

$$\frac{1}{\sigma^2} \sum_{i=1}^n (\hat{y}_i - \mu_i)(y_i - \mu_i), \quad (12)$$

which provides a Monte Carlo estimate of $\frac{1}{\sigma^2} \sum_{i=1}^n \text{Cov}(y_i, \hat{y}_i)$. Data are generated from the high-dimensional setting of scenario 1, described in Section 5. Results are in Figure 1(a).

We also propose an estimator for the degrees of freedom of SpAM (1). Defining $\Psi = [\Psi_1 \cdots \Psi_d]$ and $\beta = (\beta_1^T \cdots \beta_d^T)^T$, we estimate SpAM's degrees of freedom as

$$\hat{df}_{SpAM} = \text{Tr} \left(\Psi_{\mathcal{A}} \left[\Psi_{\mathcal{A}}^T \Psi_{\mathcal{A}} + \lambda D_2 \right]^{-1} \Psi_{\mathcal{A}}^T \right), \quad (13)$$

where $\mathcal{A} = \{i \in \{1, \dots, pd\} : \hat{\beta}_i \neq 0\}$ and D_2 is block diagonal with the j th block equal to

$$\frac{\Psi_{j\mathcal{A}_j}^T \Psi_{j\mathcal{A}_j} \hat{\beta}_{j\mathcal{A}_j} \hat{\beta}_{j\mathcal{A}_j}^T \Psi_{j\mathcal{A}_j}^T}{\|\Psi_{j\mathcal{A}_j} \hat{\beta}_{j\mathcal{A}_j}\|_2} - \frac{\Psi_{j\mathcal{A}_j}^T \Psi_{j\mathcal{A}_j} \hat{\beta}_{j\mathcal{A}_j} \hat{\beta}_{j\mathcal{A}_j}^T \Psi_{j\mathcal{A}_j}^T}{\|\Psi_{j\mathcal{A}_j} \hat{\beta}_{j\mathcal{A}_j}\|_2^3} \text{ with } \mathcal{A}_j = \{i \in \{1, \dots, d\} : \hat{\beta}_{ji} \neq 0\}.$$

Proposition 4.4—Assume $\mathbf{y} \sim \text{MVN}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$. Then $\hat{df}_{SpAM}$ is an unbiased estimator of the degrees of freedom of SpAM.

In contrast, Ravikumar et al. [2009] proposed

$$\sum_{j=1}^p \sum_{k=1}^d I(\hat{\beta}_{jk} \neq 0) \quad (14)$$

as an estimator for SpAM's degrees of freedom. Figure 1(b) compares the means of the two estimators (13) and (14) with the mean of (12) across 1000 replicate data sets generated according to the high-dimensional setting of scenario 1 (see Section 5). While (14) provides an intuitive estimate of model complexity, it is not an unbiased estimator of degrees of freedom as defined by $\frac{1}{\sigma^2} \sum_{i=1}^n \text{Cov}(y_i, \hat{y}_i)$. Our estimator, (13), is unbiased for this quantity.

4.2 Range of λ that Yields Complete Sparsity

We now consider the range of λ for which $\hat{\boldsymbol{\theta}}_j = \mathbf{0}$ for $j = 1, \dots, p$, for $\alpha = 1$ and $\alpha = 0$.

Lemma 4.5—Let $\mathcal{Y} = \mathbf{y} - \bar{y}\mathbf{1}$. If $\alpha = 1$, then the solution to (5) is completely sparse if and only if $\lambda \geq \|\mathbf{V}^T \mathcal{Y}\|_\infty$. If $\alpha = 0$, then the solution is completely sparse if and only if $\lambda \geq \|\mathcal{Y}\|_2$.

The proof of Lemma 4.5 can be found in Appendix A.3. In Lemma 4.5, note that $\|\mathbf{V}^T \mathcal{Y}\|_\infty = \max(g(\mathbf{P}_1 \mathcal{Y}), \dots, g(\mathbf{P}_p \mathcal{Y}))$, where $g(\mathbf{a}) = \max\{|a_1|, |a_1 + a_2|, \dots, |a_1 + a_2 + \dots + a_{n-1}|\}$. We now present a sufficient condition for the FLAM solution to be completely sparse, for any α .

Corollary 4.6—For any $\alpha \in [0, 1]$, if $\lambda \geq \min \left(\frac{\|\mathbf{V}^T \mathcal{Y}\|_\infty}{\alpha}, \frac{\|\mathcal{Y}\|_2}{1-\alpha} \right)$, then the solution to (5) is completely sparse.

The proof of Corollary 4.6 can be found in Appendix A.4. When selecting λ for FLAM, we need never consider a value larger than that in Corollary 4.6.

4.3 Prediction Consistency

In this section, we establish prediction consistency for FLAM. For simplicity, we assume $\mathbf{y}$ has mean zero in this subsection. The estimated prediction error compares the predicted outcome, $\hat{\mathbf{y}} = \sum_{j=1}^p \hat{\boldsymbol{\theta}}_j$, to the true mean outcome, $\mathbf{y}^0 = \sum_{j=1}^p \boldsymbol{\theta}_j^0$, where $\boldsymbol{\theta}_1^0, \dots, \boldsymbol{\theta}_p^0$ are the true

coefficient values, which are assumed to have mean zero. Lemma 4.7 provides a finite sample bound for the prediction error.

Lemma 4.7—Assume $y = \sum_{j=1}^p \theta_j^0 + \varepsilon$ with $\varepsilon \sim \text{MVN}(\mathbf{0}, \sigma^2 \mathbf{I})$. If $\lambda \geq 2\sigma \sqrt{\frac{\log((n-1)p)}{n}}$, then

$$\frac{1}{n} \left\| \sum_{j=1}^p (\hat{\theta}_j - \theta_j^0) \right\|_2^2 \leq 3\lambda \sum_{j=1}^p \left[\alpha \|D P_j \theta_j^0\|_1 + (1-\alpha) \|\theta_j^0\|_2 \right]$$

holds with probability at least $1 - \left(\frac{2}{(n-1)p} + \frac{1}{n} \right)$.

The proof of Lemma 4.7 can be found in Appendix A.5. Now, assume that $\theta_{ji}^0 = f_j(x_{ij})$ where f_j has bounded variation, and all elements of $\mathbf{x}_j \in [a, b]$ for some a and b . Assume also that the number of non-sparse functions is bounded, i.e., $\sum_{j=1}^p \|f_j\|_0 = K < \infty$. Together, these two assumptions imply that $\sum_{j=1}^p \|D P_j \theta_j^0\|_1 = O(1)$, and that $\sum_{j=1}^p \|\theta_j^0\|_2 = O(\sqrt{n})$. Thus, FLAM is prediction consistent provided that $(1 - \alpha) = \alpha(\log((n-1)p))^{-1/2}$, and $\lambda = 2\sigma \sqrt{\frac{\log((n-1)p)}{n}}$.

Our theory suggests that the necessary α to achieve prediction consistency converges to 1 as n and p grow. However, we have found in practice that $\alpha < 1$ is desirable in finite samples with a large number of covariates (see Section 5 for simulations).

5 Simulations

We compare the performance of FLAM to two competitors: GAM using smoothing splines (implemented with the R package gam [Hastie, 2013]), and SpAM with basis vectors corresponding to a natural cubic spline with $d - 1$ non-boundary knots at equally spaced quantiles of $\mathbf{x}_j$ (implemented with the R package SAM [Zhao et al., 2014]). Data are

generated according to $y_i = \sum_{j=1}^p f_j(x_{ij}) + \varepsilon_i$ with $\varepsilon_i \stackrel{iid}{\sim} N(0, 1)$, $x_{ij} \stackrel{iid}{\sim} \text{Uniform}[-2.5, 2.5]$, and $p = 4$. We consider four scenarios, displayed in Figure 2:

Scenario 1: Each function f_j is piecewise constant (Figure 2(a)).

Scenario 2: Each function f_j is smooth (Figure 2(b)). These are the exact functions used for the simulations in the original SpAM paper [Ravikumar et al., 2009].

Scenario 3: Two of the f_j are piecewise constant, and the other two f_j are smooth (Figure 2(c)). This is a compromise between scenarios 1 and 2.

Scenario 4: Each function f_j is constant in some areas of its domain and highly variable in others (Figure 2(d)).

All functions are constructed such that $\int_{-2.5}^{2.5} f_j = 0$ and $\int_{-2.5}^{2.5} f_j^2 = 1$. We refer to scenarios 1–4 as the *low-dimensional setting*. Additionally, we refer to the same scenarios with the

addition of 96 noise functions (i.e., $f_5, \dots, f_{100} = 0$) as the *high-dimensional setting*. We note that GAM can only be applied in the low-dimensional setting ($n > p$).

For each scenario, we generate training, test, and validation sets, each with $n = 100$.

Functions are fit on the training set, and mean squared error (MSE; $\frac{1}{n} \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$) is evaluated on the test set. For FLAM, we fix $\alpha = 0.5, 0.75$, or 1 and consider a range of λ . For GAM, we consider a range of degrees of freedom for each smoothing spline, from 1 (just a linear fit) to n/p . For SpAM, we fix $d = 3, 6$, or 10 and consider a range of λ .

We evaluate each method's performance as a function of its degrees of freedom. For GAM, the total degrees of freedom is p multiplied by the degrees of freedom for each covariate's smoothing spline, plus one degree of freedom for the intercept. For FLAM and SpAM, the degrees of freedom are estimated using (11) and (13), respectively.

Figure 3 displays the test set MSE versus total degrees of freedom for the three methods. FLAM achieves the lowest test set MSE across all scenarios except in scenario 2 where all f_j are smooth. GAM performs comparably to SpAM in scenario 3 without noise functions. As expected, FLAM with $\alpha = 1$ outperforms FLAM with $\alpha < 1$ in the scenarios without noise functions, as no additional sparsity is needed. In general, $\alpha < 1$ is preferred in the scenarios with noise functions, with the exception of scenario 4. We discuss this discrepancy below.

Additionally, we summarize performance for the *optimal* tuning parameter, defined as the tuning parameter corresponding to the minimum test set MSE. We calculate the validation set MSE for the training set fit corresponding to the optimal tuning parameter, as well as the

parameter fit ($\sum_{j=1}^p \|\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_j\|_2^2$), sparsity, and degrees of freedom (Table 1). Once again, FLAM performs best in all scenarios except when all f_j are smooth (scenario 2), with the best performance corresponding to FLAM with $\alpha = 1$ in the low-dimensional setting and $\alpha = 0.75$ in the high-dimensional setting. In scenario 4 with noise functions, FLAM with $\alpha = 1$ is able to achieve comparable sparsity to FLAM with $\alpha < 1$. This explains the optimal performance of FLAM with $\alpha = 1$ in this setting (Figure 3).

A strength of FLAM is its local adaptivity, or ability to produce a fit that is highly variable in one portion of the domain and constant in another. We can see this qualitatively by examining the function fits for scenario 4 with 96 noise functions. In Figure 4, we plot the fits corresponding to the the optimal tuning parameter (as defined above) for the truly non-zero functions. In general, FLAM more adeptly fits both the constant and highly variable regions of the functions, relative to SpAM. SpAM's local adaptivity is limited due to the types of penalties imposed in (1) — λ encourages the entire $\hat{f}_j$ to be zero, while d controls the amount of flexibility in each $\hat{f}_j$. Having fewer basis functions (i.e., small d) results in less variable function fits, while a large d can produce highly variable function fits. However, the amount of variability cannot be varied greatly over the domain of the function fit, unless basis functions are specifically chosen for this purpose *a priori*.

6 Data Application

6.1 Predictors of a Country's Happiness

We now consider whether wealth is associated with happiness, by estimating the conditional relationships between a country-level happiness index and gross national income, as well as 11 other country-level predictors. The happiness index is the average of Cantril Scale [Cantril, 1965] responses of approximately 3000 residents in each country obtained in Gallup World Polls from 2010–2012, publicly available from the United Nations (UN) 2013 World Happiness Report [Helliwell et al., 2013]. The predictors are publicly available through the UN Human Development Reports and the World Bank Development Indicators [United Nations Development Programme, 2012, World Bank Group, 2012]. They are from 2012 data or the closest year prior.

We consider 10 splits of the 109 countries with complete data into training and test sets. We compare FLAM to GAM with an identity link and smoothing splines using the R package `gam` [Hastie, 2013]. Tuning parameters are chosen using 10-fold CV in the training set. The estimated fits are shown in Figure 5. Both methods provide a large improvement in average test set MSE across 10 splits of the data compared to the intercept-only model (FLAM: 0.371; GAM: 0.308; intercept-only: 1.19). FLAM's estimated fits are both intuitive and fairly similar across the different splits of data. Conditional on the other predictors, FLAM estimates that increased gross national income is associated with increased happiness, up to a certain level of income. Beyond that, happiness is constant. We were quite surprised to see that GAM finds a *negative* conditional association between a country's happiness index and the number of scientific journal publications. Reassuringly, FLAM found no such association.

6.2 Classification based on Gene Expression

In this section, we apply FLAM with logistic loss (to be discussed in Section 7), in order to perform classification using gene expression measurements. The data sets we consider are:

1. *Autism* [Alter et al., 2011]: 1498 gene expression measurements from peripheral blood lymphocytes sampled from 82 children with autism and 60 controls, publicly available from GEO at accession number GDS4431 [Barrett et al., 2007].
2. *Lung S* [Spira et al., 2007]: 22,283 gene expression measurements from large airway epithelial cells sampled from 97 smokers with lung cancer and 90 smokers without lung cancer, available from GEO at accession number GDS2771.
3. *Lung NS* [Lu et al., 2010]: 54,675 gene expression measurements from 60 pairs of tumor and adjacent normal lung tissue from non-smoking women with non-small cell lung carcinoma, available from GEO at accession number GDS3837.

We consider only the 2000 genes with the largest variance in the *Lung S* and *Lung NS* data sets. We compare the performances of FLAM to SpAM and ℓ_1 -penalized logistic regression over 30 splits of the data into training and test sets, after standardizing each gene to have

mean zero and variance one in the training set. We choose the tuning parameters using 10-fold CV in the training set and calculate the misclassification rate in the test set.

Test error and sparsity (the percent of genes not used in the classifier) are shown in Figure 6. FLAM has the same or better predictive performance on average as SpAM, but uses a less sparse classifier. However, lasso's performance is comparable to FLAM and SpAM, which indicates that the sample size may be too small to successfully model non-linear relationships in these three data sets.

For one split of the Lung S data, Figure 6 displays the estimated fits from FLAM, SpAM, and lasso for six genes. These six genes were selected because they were among the 15 with the highest-variance fits for both FLAM and SpAM. Since the genes estimated to have a non-zero relationship with the response differed for each method, the conditional fits shown for a particular gene are not directly comparable across methods.

7 Extensions of FLAM

We now consider the general optimization problem minimize

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^{np}}{\text{minimize}} \quad \ell(\boldsymbol{\theta}) + \lambda \sum_{j=1}^p Q_j(\boldsymbol{\theta}_j), \quad (15)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T \cdots \boldsymbol{\theta}_p^T)^T$, $\ell: \mathbb{R}^{np} \rightarrow \mathbb{R}$ is a differentiable, convex loss function with Lipschitz continuous gradient $\dot{\ell}$ and $Q_j(\cdot)$ is a convex penalty function. We denote the Hessian of ℓ as $\ddot{\ell}$. Thus far we have considered (15) for squared error loss, i.e.,

$\ell(\boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{y} - \sum_{j=1}^p \boldsymbol{\theta}_j - \boldsymbol{\theta}_0 \mathbf{1}\|_2^2$, and $Q_j(\boldsymbol{\theta}_j) = \alpha \|\mathbf{D}\mathbf{P}_j \boldsymbol{\theta}_j\|_1 + (1 - \alpha) \|\boldsymbol{\theta}_j\|_2$ for $\alpha \in [0, 1]$. In this section, we discuss extensions of FLAM to (1) other losses $\ell(\boldsymbol{\theta})$ and (2) other penalties Q_j .

7.1 A General Algorithm

Generalized gradient descent (GGD) can be used to solve (15) [Beck and Teboulle, 2009]. That is, (15) can be solved by choosing an initial $\hat{\boldsymbol{\theta}}^0$ and continually updating

$$\hat{\boldsymbol{\theta}}^k = \underset{\boldsymbol{\theta} \in \mathbb{R}^{np}}{\text{argmin}} \quad \frac{L}{2} \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{k-1}\|^2 + \frac{1}{L} \dot{\ell}(\hat{\boldsymbol{\theta}}^{k-1})^2 + \lambda \sum_{j=1}^p Q_j(\boldsymbol{\theta}_j) \quad (16)$$

until convergence of the objective of (15), where $L \in \mathbb{R}$ is such that $\ddot{\ell}(\cdot) \preceq L\mathbf{I}$. Equation 16 is separable in $\boldsymbol{\theta}_j$, so the features can be updated in parallel during each iteration of GGD (in contrast to coordinate descent, in which the features are updated sequentially). In the special case of (15) given in (5), GGD provides an alternative to Algorithm 1. Details are omitted in the interest of brevity.

7.2 Generalized FLAM

We now consider the model $E[y_i|x_i]=g\left(\sum_{j=1}^p f_j(x_{ij})\right)$, where $g(\cdot)$ is a specified function. For instance, in the case of a binary response, we can consider the mean model

$E[y_i|x_i]=\text{expit}(\theta_0+\sum_{j=1}^p \theta_{ji})$, define $\theta=(\theta_0 \mathbf{1}^T \quad \theta_1^T \quad \dots \quad \theta_p^T)^T$, and take the loss to be logistic,

$$\ell(\theta)=-\mathbf{y}^T((\mathbf{1}_{p+1}^T \otimes \mathbf{I}_n)\theta)+\mathbf{1}^T \log(1+\exp((\mathbf{1}_{p+1}^T \otimes \mathbf{I}_n)\theta)).$$

We then solve (15) by continually updating (16), which amounts to the updates

$$\begin{aligned} \hat{\theta}_0^k &= \hat{\theta}_0^{k-1} - \frac{4}{n(p+1)} \left[\text{expit} \left(\hat{\theta}_0^{k-1} \mathbf{1} + \sum_{j=1}^p \hat{\theta}_j^{k-1} \right) - \mathbf{y} \right]^T \mathbf{1} \\ \hat{\theta}_j^k &= \underset{\theta_j \in \mathbb{R}^n}{\text{argmin}} \quad \frac{p+1}{8} \|\theta_j - \hat{\theta}_j^{k-1}\|_1 + \frac{4}{p+1} \left[\text{expit} \left(\hat{\theta}_0^{k-1} \mathbf{1} + \sum_{j=1}^p \hat{\theta}_j^{k-1} \right) - \mathbf{y} \right]^2_2 + \lambda Q_j(\theta_j) \end{aligned} \quad (17)$$

for $j = 1, \dots, p$. The solution of (17) follows from Corollary 3.1 when $Q_j(\theta_j) = \alpha \|\mathbf{D}\mathbf{P}_j \theta_j\|_1 + (1-\alpha) \|\theta_j\|_2$ for $\alpha \in [0, 1]$. Derivations of results in this subsection can be found in Appendix A.6 and simulation results for FLAM with logistic loss can be found in Appendix A.7. We now consider (17) with a more general form of Q_j .

7.3 FLAM with an Alternative Penalty

Thus far, we have seen that (15) can be solved by repeatedly solving a problem of the form (16). When $Q_j(\theta_j) = \alpha \|\mathbf{D}\mathbf{P}_j \theta_j\|_1 + (1-\alpha) \|\theta_j\|_2$ for $\alpha \in [0, 1]$, the solution to (16) follows from Corollary 3.1. Lemma 7.1 generalizes this Corollary to other forms of the penalty Q_j .

Lemma 7.1—For any norm $\|\cdot\|$, and any matrix $\mathbf{B}$ with n columns, the solution to

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{y} - \theta\|_2^2 + \alpha \lambda \|\mathbf{B}\theta\| + (1-\alpha) \lambda \|\theta\|_2 \quad (18)$$

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

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{y} - \theta\|_2^2 + \alpha \lambda \|\mathbf{B}\theta\|. \quad (19)$$

The proof of Lemma 7.1 can be found in Appendix A.8.

8 Discussion

We have presented the fused lasso additive model, a flexible yet interpretable framework for prediction, for which the estimated fits are piecewise constant with data-adaptive knots.

While the ℓ_1 penalty in (5) limits the number of knots in the fits, it also shrinks the magnitude of jumps where knots do occur. However, the resulting shrinkage can easily be addressed by debiasing the fit. That is, FLAM can be used to identify the knots; then the piecewise constant model can be refit using standard linear regression with the appropriate basis functions for the known knots.

While FLAM's piecewise constant fits are very flexible, they are not well-suited to a setting in which the true fits have large slopes. Piecewise linear fits are better-suited to this type of relationship. The problem of estimating piecewise trends of any order between a single predictor and response has been previously explored [Kim et al., 2009, Ramdas and Tibshirani, 2014, Tibshirani, 2014]. We leave the extension of FLAM to this setting to future work. Performing inference using FLAM is also left to future work.

The R package `flam` is available on CRAN. The R package Shiny [Inc., 2014] was used to develop interactive web applications demonstrating the performance of FLAM on simulated (ajpete.shinyapps.io/FLAM) and user-uploaded data (ajpete.shinyapps.io/FLAMdata).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank an associate editor and two referees for helpful comments. D.W. was supported by NIH Grant DP5OD009145, NSF CAREER Award DMS-1252624, and an Alfred P. Sloan Foundation Research Fellowship. N.S. was supported by NIH Grant DP5OD019820.

References

- Akaike, Hirotugu. Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory. 1973:267–281.
- Alter, Mark D., Kharkar, Rutwik, Ramsey, Keri E., Craig, David W., Melmed, Raun D., Grebe, Theresa A., Bay, R Curtis, Ober-Reynolds, Sharman, Kirwan, Janet, Jones, Josh J., et al. Autism and increased paternal age related changes in global levels of gene expression regulation. PLOS ONE. 2011; 6(2):e16715. [PubMed: 21379579]
- Avalos, Marta, Grandvalet, Yves, Ambroise, Christophe. Parsimonious additive models. Computational Statistics & Data Analysis. 2007; 51(6):2851–2870.
- Barrett, Tanya, Troup, Dennis B., Wilhite, Stephen E., Ledoux, Pierre, Rudnev, Dmitry, Evangelista, Carlos, Kim, Irene F., Soboleva, Alexandra, Tomashevsky, Maxim, Edgar, Ron. NCBI GEO: mining tens of millions of expression profiles — database and tools update. Nucleic Acids Research. 2007; 35(suppl 1):D760–D765. [PubMed: 17099226]
- Beck, Amir, Teboulle, Marc. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences. 2009; 2(1):183–202.
- Bühlmann, Peter, Hothorn, Torsten. Boosting algorithms: Regularization, prediction and model fitting. Statistical Science. 2007:477–505.

- Bühlmann, Peter, van de Geer, Sara. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer; 2011.
- Bühlmann, Peter, Yu, Bin. Boosting with the l_2 loss: Regression and classification. Journal of the American Statistical Association. 2003; 98(462):324–339.
- Cantril, Hadley. The Pattern of Human Concerns. Rutgers University Press; 1965.
- de Boor, Carl. A practical guide to splines. Mathematics of Computation. 1978
- Efron, Bradley. How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association. 1986; 81(394):461–470.
- Hastie, Trevor. gam: Generalized Additive Models. 2013. <http://CRAN.R-project.org/package=gam>. R package version 1.09
- Hastie, Trevor, Tibshirani, Robert. Generalized additive models. Statistical Science. 1986:297–310.
- Hastie, Trevor J., Tibshirani, Robert J. Generalized Additive Models. Vol. 43. CRC Press; 1990.
- Helliwell, John F., Layard, Richard, Sachs, Jeffrey. Emirates Competitiveness Council. World happiness report 2013. Sustainable Development Solutions Network; 2013.
- Hoefling, Holger. A path algorithm for the fused lasso signal approximator. Journal of Computational and Graphical Statistics. 2010; 19(4):984–1006.
- Hoefling, Holger. flsa: Path algorithm for the general Fused Lasso Signal Approximator. 2013. <http://CRAN.R-project.org/package=flsa>. R package version 1.05
- Huang, Jian, Horowitz, Joel L., Wei, Fengrong. Variable selection in nonparametric additive models. The Annals of Statistics. 2010; 38(4):2282. [PubMed: 21127739]
- RStudio Inc. shiny: Web Application Framework for R. 2014. <http://CRAN.R-project.org/package=shiny>. R package version 0.9.1
- Johnson, Nicholas A. A dynamic programming algorithm for the fused lasso and l_0 -segmentation. Journal of Computational and Graphical Statistics. 2013; 22(2):246–260.
- Kim, Seung-Jean, Koh, Kwangmoo, Boyd, Stephen, Gorinevsky, Dmitry. l_1 trend filtering. SIAM Review. 2009; 51(2):339–360.
- Lin, Yi, Zhang, Hao Helen. Component selection and smoothing in smoothing spline analysis of variance models. The Annals of Statistics. 2006; 34(5):2272–2297.
- Liu, Jun, Yuan, Lei, Ye, Jieping. An efficient algorithm for a class of fused lasso problems. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM; 2010. p. 323–332.
- Lou, Yin, Bien, Jacob, Caruana, Rich, Gehrke, Johannes. Sparse partially linear additive models. 2014 arXiv preprint arXiv:1407.4729.
- Lu, Tzu-Pin, Tsai, Mong-Hsun, Lee, Jang-Ming, Hsu, Chung-Ping, Chen, Pei-Chun, Lin, Chung-Wu, Shih, Jin-Yuan, Yang, Pan-Chyr, Hsiao, Chuhsing Kate, Lai, Liang-Chuan, et al. Identification of a novel biomarker, sema5a, for non-small cell lung carcinoma in nonsmoking women. Cancer Epidemiology Biomarkers & Prevention. 2010; 19(10):2590–2597.
- Mammen, Enno, van de Geer, Sara. Locally adaptive regression splines. The Annals of Statistics. 1997; 25(1):387–413.
- Meier, Lukas, van de Geer, Sara, Bühlmann, Peter. High-dimensional additive modeling. The Annals of Statistics. 2009; 37(6B):3779–3821.
- Ramdas, Aaditya, Tibshirani, Ryan J. Fast and flexible admm algorithms for trend filtering. 2014 arXiv preprint arXiv:1406.2082.
- Ravikumar, Pradeep, Lafferty, John, Liu, Han, Wasserman, Larry. Sparse additive models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009; 71(5):1009–1030.
- Sardy, Sylvain, Tseng, Paul. Amlet, ramlet, and gamlet: Automatic nonlinear fitting of additive models, robust and generalized, with wavelets. Journal of Computational and Graphical Statistics. 2004; 13(2):283–309.
- Schwarz, Gideon. Estimating the dimension of a model. The Annals of Statistics. 1978; 6(2):461–464.
- Simon, Noah, Tibshirani, Robert. Standardization and the group lasso penalty. Statistica Sinica. 2012; 22(3):983. [PubMed: 26257503]
- Spira, Avrum, Beane, Jennifer E., Shah, Vishal, Steiling, Katrina, Liu, Gang, Schembri, Frank, Gilman, Sean, Dumas, Yves-Martine, Calner, Paul, Sebastiani, Paola, et al. Airway epithelial gene

expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*. 2007; 13(3):361–366.

- Stein, Charles M. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*. 1981:1135–1151.
- Tibshirani, Robert, Saunders, Michael, Rosset, Saharon, Zhu, Ji, Knight, Keith. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67(1):91–108.
- Tibshirani, Ryan J. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*. 2014; 42(1):285–323. <http://dx.doi.org/10.1214/13-AOS1189>. DOI: 10.1214/13-AOS1189
- Tibshirani, Ryan J., Taylor, Jonathan. Degrees of freedom in lasso problems. *The Annals of Statistics*. 2012; 40(2):1198–1232.
- Tseng, Paul. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*. 2001; 109(3):475–494.
- United Nations Development Programme. Human Development Indicators. United Nations Publications; 2012.
- Van Belle, Vanya MCA., Calster, Ben Van, Timmerman, Dirk, Bourne, Tom, Bottomley, Cecilia, Valentin, Lil, Neven, Patrick, Van Huffel, Sabine, Suykens, Johan AK., Boyd, Stephen. A mathematical model for interpretable clinical decision support with applications in gynecology. *PLOS ONE*. 2012; 7(3):e34312. [PubMed: 22479598]
- Wood, Simon N., Goude, Yannig, Shaw, Simon. Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2014
- World Bank Group. World Development Indicators 2012. World Bank Publications; 2012.
- Yuan, Ming, Lin, Yi. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006; 68(1):49–67.
- Zhang, Hao Helen, Cheng, Guang, Liu, Yufeng. Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association*. 2011; 106(495)
- Zhao, Tuo, Li, Xingguo, Liu, Han, Roeder, Kathryn. SAM: Sparse Additive Modelling. 2014. URL <http://CRAN.R-project.org/package=SAM>. R package version 1.0.5

Figure 1.

In (a), we compare the degrees of freedom of FLAM calculated using (12) (y-axis) to the unbiased estimator (11) (x-axis) for $\alpha = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$; each value of α is indicated by an (overlapping) colored line. In (b), we compare the degrees of freedom for SpAM with $d = 3$ calculated using (12) (y-axis) to the estimators (13) (—) and (14) (—) (x-axis). In both plots, the solid lines are obtained by varying λ for FLAM or SpAM. The black dotted lines indicate $y = x$.

Figure 2.
Functions used to generate data in: (a) scenario 1, (b) scenario 2, (c) scenario 3, and (d) scenario 4.

Figure 3.

Mean test set MSE plotted by degrees of freedom for: (a) scenario 1, (b) scenario 1 with noise functions, (c) scenario 2, (d) scenario 2 with noise functions, (e) scenario 3, (f) scenario 3 with noise functions, (g) scenario 4, and (h) scenario 4 with noise functions. Shaded bands indicate point-wise 95% confidence intervals over the 100 replicate data sets. GAM is only applicable in the low-dimensional setting.

Figure 4.

We compare the fits of FLAM (left) and SpAM (right) with the true underlying functions (—) generated according to scenario 4. In each panel, 25 curves are shown. Each curve corresponds to a validation set fit with the optimal tuning parameter value, over one simulated data set.

Figure 5.

Conditional associations between the happiness index for a country and twelve country-level predictors were estimated using FLAM (piecewise constant fits, —) and GAM (smooth fits, —). Ten fits for each method were obtained by repeatedly splitting the data into training and test sets. The gray bar at the bottom of each plot indicates the distribution of that predictor.

Figure 6.

We compare the classification performance of FLAM (■) to SpAM (●) and lasso (▲) in terms of (a) test set error and (b) sparsity for three gene expression data sets. Both plots show mean estimates with 95% confidence intervals, which are calculated using 30 splits of the data into training and test sets. In (c)–(h), for six genes we show the fits from FLAM (—), SpAM (---), and lasso (---), which were estimated from one split of the Lung S data. The gray bar at the bottom of each plot indicates the distribution of that predictor.

Table 1

Results on the validation data, with the tuning parameter value chosen based on test set MSE. Mean (standard error) across 100 replicate data sets are shown. Parameter fit is defined as $\sum_{j=1}^p \|\theta_j - \hat{\theta}_j\|_2^2$. Note that GAM can only be applied in the low-dimensional setting. In the high-dimensional setting, we report the false negative and false positive rates, which are the average proportions of $\hat{f}_1, \dots, \hat{f}_4$ and $\hat{f}_5, \dots, \hat{f}_{100}$, respectively, estimated to be non-zero. All $\hat{f}_j$ were non-zero for all methods in the low-dimensional setting.

Low-dimensional				High-dimensional				
	MSE	Parameter fit	Degrees of freedom	MSE	Parameter fit	Degrees of freedom	False positive rate	False negative rate
Scenario 1								
FLAM, $\alpha = 0.5$	1.73 (0.03)	78.0 (2.0)	52.1 (1.2)	2.12 (0.05)	110.0 (3.1)	63.0 (1.7)	0.17 (0.01)	0 (0)
FLAM, $\alpha = 0.75$	1.53 (0.03)	57.5 (1.6)	39.9 (1.0)	1.92 (0.04)	91.3 (2.6)	53.6 (1.5)	0.20 (0.01)	0 (0)
FLAM, $\alpha = 1$	1.44 (0.02)	48.5 (1.3)	33.9 (0.8)	2.30 (0.05)	129.8 (3.0)	56.0 (1.8)	0.30 (0.01)	0 (0)
GAM	1.67 (0.03)	65.0 (1.3)	32.1 (0.9)					
SpAM, $d = 3$	2.05 (0.03)	111.3 (1.8)	12.3 (0.1)	2.60 (0.05)	160.5 (2.7)	35.7 (1.2)	0.19 (0.01)	0 (0)
SpAM, $d = 6$	1.82 (0.03)	109.3 (3.0)	23.0 (0.2)	2.59 (0.05)	172.2 (3.3)	51.0 (1.5)	0.20 (0.01)	0 (0)
SpAM, $d = 10$	1.85 (0.03)	133.8 (3.9)	35.8 (0.3)	2.91 (0.06)	210.1 (4.3)	63.2 (1.3)	0.20 (0.01)	0 (0)
Scenario 2								
FLAM, $\alpha = 0.5$	1.65 (0.03)	68.8 (1.9)	58.0 (1.0)	2.07 (0.04)	109.8 (2.9)	69.7 (1.7)	0.18 (0.01)	0 (0)
FLAM, $\alpha = 0.75$	1.52 (0.03)	55.8 (1.4)	51.9 (0.9)	2.10 (0.04)	112.7 (3.0)	66.2 (1.6)	0.24 (0.01)	0 (0)
FLAM, $\alpha = 1$	1.49 (0.02)	53.7 (1.4)	51.5 (0.9)	2.91 (0.06)	193.1 (4.3)	61.6 (2.1)	0.33 (0.01)	0 (0)
GAM	1.20 (0.02)	24.2 (0.7)	21.5 (0.4)					
SpAM, $d = 3$	1.23 (0.02)	33.2 (1.1)	12.5 (0.1)	1.60 (0.03)	69.8 (2.1)	35.0 (1.2)	0.19 (0.01)	0 (0)
SpAM, $d = 6$	1.30 (0.02)	53.6 (2.2)	23.6 (0.1)	1.89 (0.03)	108.6 (3.1)	54.1 (1.4)	0.21 (0.01)	0 (0)
SpAM, $d = 10$	1.54 (0.03)	83.2 (2.7)	36.8 (0.2)	2.54 (0.05)	172.0 (4.0)	67.4 (1.5)	0.21 (0.01)	0 (0)
Scenario 3								
FLAM, $\alpha = 0.5$	1.65 (0.03)	68.7 (1.7)	52.6 (1.0)	1.98 (0.04)	97.2 (2.7)	61.3 (1.5)	0.16 (0.01)	0 (0)
FLAM, $\alpha = 0.75$	1.47 (0.02)	51.5 (1.4)	43.2 (1.1)	1.83 (0.04)	83.5 (2.3)	54.7 (1.5)	0.20 (0.01)	0 (0)
FLAM, $\alpha = 1$	1.38 (0.02)	44.4 (1.1)	37.0 (0.8)	2.10 (0.04)	110.7 (2.7)	52.2 (1.8)	0.27 (0.01)	0 (0)
GAM	1.45 (0.02)	45.0 (1.0)	25.3 (0.7)					
SpAM, $d = 3$	1.63 (0.03)	69.5 (1.2)	12.4 (0.1)	2.09 (0.04)	113.2 (2.5)	34.0 (1.1)	0.19 (0.01)	0 (0)

<i>Low-dimensional</i>						<i>High-dimensional</i>					
	MSE	Parameter fit	Degrees of freedom	MSE	Parameter fit	Degrees of freedom	False positive rate	False negative rate			
SpAM, $d = 6$	1.54 (0.02)	72.6 (2.5)	23.3 (0.2)	2.22 (0.04)	134.5 (3.1)	50.8 (1.5)	0.20 (0.01)	0 (0)			
SpAM, $d = 10$	1.68 (0.03)	101.9 (3.2)	36.3 (0.3)	2.73 (0.05)	187.9 (4.1)	64.9 (1.6)	0.20 (0.01)	0 (0)			
<i>Scenario 4</i>											
FLAM, $\alpha = 0.5$	1.87 (0.03)	91.0 (1.9)	55.8 (1.3)	2.38 (0.05)	137.6 (3.7)	60.8 (1.9)	0.17 (0.01)	0 (0)			
FLAM, $\alpha = 0.75$	1.69 (0.03)	73.9 (1.5)	45.1 (1.3)	2.16 (0.04)	117.3 (3.1)	46.0 (1.5)	0.18 (0.01)	0 (0)			
FLAM, $\alpha = 1$	1.62 (0.03)	67.5 (1.4)	39.1 (1.5)	2.13 (0.04)	114.4 (3.0)	40.1 (1.3)	0.19 (0.01)	0 (0)			
GAM	1.88 (0.03)	81.5 (1.4)	31.2 (1.6)								
SpAM, $d = 3$	2.14 (0.04)	120.3 (2.0)	12.3 (0.1)	2.78 (0.05)	179.8 (3.4)	31.8 (0.9)	0.17 (0.01)	0 (0)			
SpAM, $d = 6$	2.02 (0.03)	118.7 (2.6)	22.6 (0.2)	2.84 (0.05)	193.5 (4.0)	49.4 (1.4)	0.20 (0.01)	0 (0)			
SpAM, $d = 10$	2.17 (0.04)	155.0 (3.8)	35.0 (0.3)	3.29 (0.06)	243.4 (4.4)	59.8 (1.7)	0.19 (0.01)	0.02 (0.01)			