

# LASSO ISOTone for High Dimensional Additive Isotonic Regression

Zhou Fang

Department of Statistics, University of Oxford  
and

Nicolai Meinshausen

Department of Statistics, University of Oxford

June 1, 2010

## Abstract

Additive isotonic regression attempts to determine the relationship between a multi-dimensional observation variable and a response, under the constraint that the estimate is the additive sum of univariate component effects that are monotonically increasing. In this article, we present a new method for such regression called LASSO Isotone (LISO). LISO adapts ideas from sparse linear modelling to additive isotonic regression. Thus, it is viable in many situations with high dimensional predictor variables, where selection of significant versus insignificant variables are required. We suggest an algorithm involving a modification of the backfitting algorithm CPAV. We give a numerical convergence result, and finally examine some of its properties through simulations. We also suggest some possible extensions that improve performance, and allow calculation to be carried out when the direction of the monotonicity is unknown.

*Keywords:* Nonparametric regression; Isotonic regression; Lasso

## 1 Introduction

We often seek to uncover or describe the dependence of a response on a large number of covariates. In many cases, parametric and in particular linear models may prove overly restrictive. Additive modelling, as described, for instance in Hastie and Tibshirani [1990], is well known to be an useful generalisation.

Suppose we have  $n$  observations available of the pair  $(X_i, Y_i)$ , where  $Y_i \in \mathbb{R}$  is a response variable, and  $X_i = (X_i^{(1)}, \dots, X_i^{(p)}) \in \mathbb{R}^p$  is a vector of covariates.

In additive modelling, we typically assume that the data is well approximated by a model of the form

$$Y_i = \sum_{k=1}^p f_k(X_i^{(k)}) + \varepsilon_i,$$

where  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$  is a random error term, assumed independent of the covariates and identically distributed with mean zero. For every covariate  $k = 1, \dots, p$ , each component fit  $f_k$  is chosen from a space of univariate functions  $\mathcal{F}_k$ . Usually, these spaces are constrained to be smooth in some suitable sense, and in fitting, we minimise the L2 norm of the error,

$$\frac{1}{2} \left\| Y - \sum_{k=1}^p f_k(X^{(k)}) \right\|^2 := \frac{1}{2} \sum_{i=1}^n \left( Y_i - \sum_{k=1}^p f_k(X_i^{(k)}) \right)^2,$$

under the constraint that  $f_k \in \mathcal{F}_k$ , for each  $k = 1, \dots, p$ . In the case that  $\varepsilon$  is assumed to be normal, this can be directly justified as maximising the likelihood.

Work on such methods of additive modelling have produced a profuse array of techniques and generalisations. In particular, Bacchetti [1989] suggested the additive isotonic model. With the additive isotonic model, we are interested in tackling the problem of conducting regression under the restriction that the regression function is of a pre-specified monotonicity with respect to each covariate. (Isotonic means the functions are increasing, though decreasing can be accommodated easily by reversing the signs of covariates.) Such restrictions may be sensible whenever there is subject knowledge about the possible influence or relationship between predictor and response variables. A broad survey of the subject may be found in Barlow et al. [1972]. It turns out that in the univariate case, the Pool Adjacent Violators Algorithm, as first suggested in Ayer et al. [1955], allows rapid calculation of a solution to the least squares problem using this restriction alone. By doing so, we retain only the ordinal information in the covariates, and hence obtain a result that is invariant under strictly monotone transformations of the data. In addition, the form of the regression, being simply a maximisation of the likelihood, means that apart from the monotonicity constraint, we do not put on any regularisation, or smoothing.

Bacchetti [1989] built on this, by generalizing to multiple covariates. Here, the regression function is considered to be a sum of univariate functions of specified monotonicity. Fitting is conducted via the cyclic pool adjacent violators (CPAV) algorithm, in the style of a backfitting procedure built around PAVA — that is, cycling over the covariates, the partial residuals using the remaining covariates are repeatedly fitted to the current one, until convergence. Later theoretical discussion from Mammen and Yu [2007] outlined some positive properties of this procedure.

Nevertheless, CPAV, like many types of additive modelling, can fail in the high dimensional case — for instance, once  $p > n$ . The particular problem is that the least squares criterion loses strictness of convexity when the number of covariates is large, since it becomes easy for allowed component fits in some covariates to combine in the training data so as to replicate component fits in unrelated covariates. It is hence impossible for the CPAV to distinguish between two radical different regression functions since they give the same fitted values on the training dataset. Some success might be achieved, though, if the solution sought is sparse, in the sense that most of the covariates have little or no effect on the response. If the identity of the significant variables were known, then, the CPAV could be conducted on a much smaller set of covariates. However, exhaustive search to identify this sparsity pattern would be rapidly prohibitive in terms of computational cost, scaling exponentially in the number of covariates.

In the context of parametric linear regression, it has emerged recently that such sparse regression problems can be dealt with by use of a L1-norm based penalty in the optimisation. This can resolve the identifiability problem and achieve good predictive accuracy. Tibshirani [1996], Donoho [2006], amongst others, have identified several significant empirical and theoretical results to support this ‘LASSO’ estimator, while Efron et al. [2004], Friedman et al. [2007] and others have invented fast algorithms for calculating both individual estimates and full LASSO solution paths.

Generalisation of the L1 penalisation principle to nonparametric regression can also lead to successful with additive modelling. For example, recent work on this subject includes SpAM [Ravikumar et al. 2007], which describes the application of the grouped LASSO to general smoothers, and high dimensional

additive modelling with smoothness penalties [Meier et al. 2009] which follows similar principles, using a spline basis.

In this paper, we propose the Lasso-Isotone (LISO) estimator. By modifying the additive isotonic model to include a LASSO-style penalty on the total variation of component fits, we hope to conduct isotonic regression in the sparse additive setting.

The LISO is similar to the degree 0 case of the LASSO knot selection of Osborne et al. [1998], which is also identical to the fused LASSO of Tibshirani et al. [2005], if we replace the covariate matrix with ordered Haar wavelet bases, and do not consider coefficient differences for coefficients corresponding to different covariates. It is also similar to the univariate problem considered by Mammen and van de Geer [1997]. In contrast to each of these procedures, however, we allow the additional imposition of a monotonicity constraint, producing an algorithm similar in complexity to the CPAV.

In section 2 we shall describe the LISO optimisation, and in section 3 we will discuss algorithms for computation for fairly large  $n$  and  $p$ . We will discuss the effect of the regularisation, and then in section 4 suggest some extensions. Finally, in section 5 we will explore its performance using some simulation studies. Proofs of theorems are left for the appendix.

## 2 The LASSO-ISOTone Optimisation

For now, let us assume without loss of generality that we are conducting regression constrained to monotonically increasing regression functions. Let us first define some terms.

Let  $Y \in \mathbb{R}^n$  be the response vector. Assume, subtracting by a constant intercept term if necessary, that  $\sum_{i=1}^n Y_i = 0$ .  $X = (X^{(1)}, \dots, X^{(p)}) \in \mathbb{R}^{n \times p}$  is the matrix of covariates.

For a specified  $X$ , for  $k = 1, \dots, p$ , let  $\mathcal{F}_k$  be the space of bounded, univariate, and monotonically increasing functions, that have expectation zero on the  $k$ -th covariate.  $-\mathcal{F}_k$  is then the same for monotonically decreasing functions.

$$\mathcal{F}_k := \left\{ f : \mathbb{R} \rightarrow \mathbb{R} \mid \sum_{i=1}^n f(X_i^{(k)}) = 0, \text{ and } \exists U, V \text{ s.t. } \forall a < b, U \leq f(a) \leq f(b) \leq V \right\}$$

Additive isotonic models involve sums of functions from these spaces. It is simple to observe that each  $\mathcal{F}_k$  is a convex half-space that is closed except at infinity, and so as a result the space of sums of these functions must also be convex and closed except at infinity.

**Definition 2.1.** We define the Lasso-Isotone (LISO) solution for a particular value of tuning parameter  $\lambda \geq 0$  as the minimiser  $\hat{f}_\lambda = \left( \hat{f}_{k,\lambda} \right)_{k=1}^p$ , with  $\hat{f}_{k,\lambda} \in \mathcal{F}_k \forall k$ , of the LISO loss

$$L_\lambda(f_1, \dots, f_p) := \frac{1}{2} \left\| Y - \sum_{k=1}^p f_k(X^{(k)}) \right\|^2 + \lambda \sum_{k=1}^p \Delta(f_k). \quad (2.1)$$

Here  $\Delta(f_k)$  denotes the total variation of  $f_k$ , which for  $f_k \in \mathcal{F}_k$  can be calculated as

$$\Delta(g) = \sup_{x \in \mathbb{R}} f_k(x) - \inf_{x \in \mathbb{R}} f_k(x).$$

As with the LASSO, the LISO objective function is the sum of a log-likelihood term and a penalty term. It is clear that the domain is convex and, considered in the space of allowed solutions, the objective itself is convex and bounded below. Indeed, outside a neighbourhood of the origin, both terms in the objective are increasing, so a bounded solution exists for all values of  $\lambda$ . However, the objective may not be strictly convex, so this solution may not be unique.

The log-likelihood term does not consider the values of  $f_k$  except at observed values of each covariate, while the total variation penalty term, assuming monotonicity, only takes account of the upper and lower bounds of the covariate-wise regression function — indeed, for optimality, these bounds must be attained at the extremal observed values of the appropriate covariate, with the solution flat beyond this region. Thus, given any one minimiser to  $L_\lambda$ , another fit with the same function values at observed covariate points, interpolating monotonically between them, will have the same value of  $L_\lambda$ , and so also be a LISO solution. This means that we can equivalently consider optimisation in the finite dimensional space of fitted values  $\hat{f}_k(X^{(k)})$ .

For simplicity, we will represent found LISO solution components by the corresponding right continuous step function with knots only at each observation. For the remainder of this paper, we shall consider uniqueness and equivalence in terms of having equal values at the observed  $X^{(k)}$ .

We have introduced a mean zero constraint on the fitted components for identifiability, since we can easily add a constant term to any component fit  $f_k$ , and deduct it from another component, and still arrive at the same final regression function. However, we will show later that this constraint arises naturally in the univariate case, where even without it being explicitly applied,

$$\sum_{i=1}^n \hat{f}(X_i) = \sum_{i=1}^n Y_i.$$

The total variation penalty shown here has been previously suggested for regression in Mammen and van de Geer [1997], though in that case, the focus was on smoothing of univariate functions, without a monotonicity constraint.

### 3 LISO Backfitting

Considering the representation of the LISO in terms of step functions, the LISO optimisation for a given dataset can be viewed as ordinary LASSO optimisation for a linear model, constrained to positive coefficients, using an expanded design matrix  $\tilde{X} \in \mathbb{R}^{n \times p(n-1)}$ , where  $\tilde{X} = (\tilde{X}^{(1)} \dots \tilde{X}^{(p)})$ . Each  $\tilde{X}^{(k)} \in \mathbb{R}^{n \times (n-1)}$ ,  $k = 1, \dots, p$  contains  $n - 1$  step functions in the  $k$ -th covariate, which form a basis for the vector  $f_k(X^{(k)})$ , and so isotonic functions in that covariate. The coefficients  $\beta$  optimised over then represent step sizes.

Such a construction is suggested in Osborne et al. [1998], amongst others. Under this re-parametrisation of the problem, existing LASSO algorithms for linear regression may be applied, with a modification to restrict solutions to non-negative values. In particular, the Least Angle Regression algorithm of Efron et al. [2004] is effective, since shortcuts exist for calculating the necessary correlations.

On the other hand, the high dimensionality of  $\tilde{X}$  means that standard methods become very costly in higher dimensions, both in terms of required computation, but especially in terms of the storage

requirements associated with very large matrices. Hence, we must consider more specialised algorithms for such cases. One such approach involves backfitting, and is workable due to the simple form of the solution when restricted to a single covariate.

### 3.1 Thresholded PAVA

In the  $p = 1$  case, it turns out that we have an exceptionally simple way to calculate the LISO estimate, which we will later use to establish a more general multivariate procedure.

With no LISO penalty (i.e.  $\lambda = 0$ ) and a single covariate, the LISO optimisation is equivalent to the standard univariate isotonic regression problem. In this case, the loglikelihood residual sum of squares term is strictly convex, and so, as a strictly convex optimisation on a convex set, a unique solution exists. Trivially, the solution must also be bounded. In fact, there exists, as described in Barlow et al. [1972] and attributed to Ayer et al. [1955], a fast algorithm for calculating the solution – the Pool Adjacent Violators Algorithm (PAVA).

Hence, defining  $\hat{f}_\lambda$  as the solution to optimisation (2.1) for  $\lambda$ , we have  $\hat{f}_0 = \hat{f}_{PAVA}$ . The following theorems describe the solutions for other values of  $\lambda$ :

**Theorem 1.** For  $A \leq B$ , denote by  $\hat{f}_{>A,<B}$  the Winsorized PAVA estimate

$$\hat{f}_{>A,<B}(x) := \begin{cases} A & \text{if } \hat{f}_{PAVA}(x) < A \\ B & \text{if } \hat{f}_{PAVA}(x) > B \\ \hat{f}_{PAVA}(x) & \text{otherwise.} \end{cases}$$

Then if  $p = 1$ , there exist thresholds  $A_\lambda \leq B_\lambda$  for each value of  $\lambda \geq 0$  such that the LASSO-Isotone solution is given by  $\hat{f}_\lambda \equiv \hat{f}_{>A_\lambda,<B_\lambda}$ .

**Theorem 2.** In Theorem 1, given  $\hat{f}_{PAVA}$ , the pair  $A_\lambda, B_\lambda$  (the optimal thresholding levels) are a piecewise linear, continuous and monotone (increasing for  $A_\lambda$ , decreasing for  $B_\lambda$ ) function of  $\lambda$ , for  $\lambda \geq 0$ .

Specifically, if

$$2\lambda \geq \sum_{i=1}^n |\hat{f}_{PAVA}(X_i) - \bar{Y}|, \quad (3.1)$$

then  $A_\lambda = B_\lambda = \bar{Y}$ .

Otherwise,  $A_\lambda, B_\lambda$  are the solutions to

$$\sum_{i=1}^n (\hat{f}_{PAVA}(X_i) - B_\lambda)_+ = \lambda \quad (3.2)$$

$$\sum_{i=1}^n (A_\lambda - \hat{f}_{PAVA}(X_i))_+ = \lambda. \quad (3.3)$$

**Corollary 3.** Let  $\pi$  be a permutation taking  $1, \dots, n$  to indices that put  $X$  in ascending order. Then if

$$\lambda \geq \max_m \left| \sum_{i=1}^m (Y_{\pi(i)} - \bar{Y}) \right| \quad (3.4)$$

$A_\lambda = B_\lambda = \bar{Y}$ .

*Remark 3.1.* The LHS of (3.2) and (3.3) specify the amount by which each threshold changes the sum of the fitted values on the appropriate side of the mean. Hence, we see that  $\sum_{i=1}^n \widehat{f}_\lambda(X_i) = \sum_{i=1}^n \widehat{f}_{PAVA}(X_i) = \sum_{i=1}^n Y_i$ , for all  $\lambda$ .

In other words, if  $Y$  has mean zero, then the mean zero constraint on the fit arises naturally, without having to be externally applied. If  $Y$  does not have mean zero, the solution is simply a shifted version of the fit for  $Y - \bar{Y}$ . This justifies deducting the mean of the response and dealing with it separately.

*Remark 3.2.* The PAVA algorithm itself can accommodate observation weights, as well as tied values in the covariates. In terms of the LISO, working with unequal observation weights demands that we work with weighted residual sums of squares. This does not affect Theorem 1, but for equations (3.3) and (3.2), weights should be introduced in the summation. Tied values should be also dealt with by merging the relevant steps, and weighting them according to the number of data points at that covariate observation.

### 3.2 Backfitting algorithm

In general, however, simple thresholding fails to solve the LISO optimisation in higher dimensions, due to correlations between steps in different covariate component functions. We can, however, extend the 1D algorithm to higher dimensions by applying it iteratively as a backfitting algorithm.

In other words, we define LISO-backfitting by the following steps:

---

**Algorithm 1** LISO-Backfitting

---

- 1: Set  $m = 0$ .
  - 2: Initialise component fits  $(f_1, \dots, f_p)$  as identically 0, or as the estimate for a different value of  $\lambda$ , storing these as the  $n \times p$  marginal fitted values.
  - 3: **repeat**
  - 4:    $f^m \leftarrow (f_1, \dots, f_p)$ .
  - 5:    $m \leftarrow m + 1$ .
  - 6:   **for**  $k = 1$  to  $p$  (or a random permutation) **do**
  - 7:     Recalculate residuals  $r_i \leftarrow Y_i - \sum_{k=1}^p f_k(X_i^{(k)})$ ,  $i = 1, \dots, n$ .
  - 8:     Refit conditional residual  $\left\{ r_i + f_k(X_i^{(k)}) \right\}_{i=1}^n$  using  $X^{(k)}$  by PAVA, producing  $\tilde{f}_k(X_i^{(k)})$ , for  $i = 1, \dots, n$ .
  - 9:     Calculate thresholds  $A_\lambda, B_\lambda$  from  $\lambda$  and  $\tilde{f}_k$  by Theorem 2.
  - 10:     Adjust component fit  $f_k(X_i^{(k)}) \leftarrow \tilde{f}_{k, >A_\lambda < B_\lambda}(X_i^{(k)})$ .
  - 11:   **end for**
  - 12: **until** sufficient convergence is achieved, through considering  $f^m$  and  $f^{m-1}$ .
  - 13: Interpolate  $f_k$  between the samples  $X_i^{(k)}$ .
- 

**Theorem 4.** For  $f^m = (f_1^m, \dots, f_p^m)$ , the sequence of states resulting from the LISO-backfitting algorithm,  $L_\lambda(f^m)$  converges to its global minimum with probability 1. Specifically, if there exists a unique solution to (2.1),  $f^m$  converges to it.

*Remark 3.3.* If there is no unique solution, the backfitting algorithm may not necessarily converge, though the LISO loss of each estimate will converge monotonically to the minimum. In addition, because the objective function is locally quadratic, as the change in the LISO loss converges to zero, the change in the estimate after each individual refitting cycle converges also to zero.

*Remark 3.4.* Moreover, defining  $X_{(i)}^{(k)}$  as the  $i$ -th smallest value of  $X^{(k)}$ , if a certain individual step in the final functional fit

$$f_k \left( X_{(i)}^{(k)} \right) - f_k \left( X_{(i-1)}^{(k)} \right)$$

has a value of zero in all solutions to the LISO minimisation, then, after a finite number of steps, all results from the algorithm must take that step exactly to zero.

This is because steps being estimated as zero in a LISO solution implies that the partial derivative of the LISO objective function  $L_\lambda$  in the above individual step direction is greater than zero when evaluated at this solution. The partial derivatives are continuous, so as the algorithm converges, the partial derivatives associated with zero steps eventually be above 0 and remain so. But then, this can only be the case following a thresholded PAVA calculation involving the covariate associated with that step if that single covariate optimisation takes the step exactly to zero.

Convergence of the algorithm can be checked for by a variety of methods. One of the simplest is to note that due to the nature of the repeated optimisation, the LISO loss will always decrease in each step, and we will converge towards the minimum. Hence, one viable stopping rule would be to cease calculating when the LISO loss of the current solution drops by too small an amount. Alternatively, we can exploit Remark 3.3, and monitor the change in the results in each cycle, stopping when this becomes small.

### 3.3 Choice of regularisation parameter

It will be always necessary to choose a tuning parameter  $\lambda$  to facilitate appropriate fitting. As with the LASSO, too high a tuning parameter will shrink the fits towards zero. Indeed, consideration of Corollary 3 shows that, with  $\bar{Y} = 0$ , and  $\pi^{(k)}$  defined as a permutation that puts the  $k$ -th covariate into ascending order, a choice of  $\lambda$  greater than

$$\max_{\substack{k=1, \dots, p, \\ m=1, \dots, n}} \left| \sum_{i=1}^m Y_{\pi^{(k)}(i)} \right|$$

will result in a zero fit in every thresholded PAVA step starting from zero, and hence a zero fit overall for the LISO.

Conversely, too small a value of  $\lambda$  will lead to improper fitting. This arises from two sources. Firstly, as with the LASSO, the noise term may flood the fit, as the level of thresholding is not sufficient to suppress correlations of the noise with the covariate step functions – the columns of  $\tilde{X}$ . Secondly,  $\lambda$  has a role in terms of fit complexity, with a small value of  $\lambda$  implying that the LISO, when restricted to the true covariates, would select more steps. This means a less sparse signal in the implied LASSO problem, so it becomes in turn more likely for selected columns of  $\tilde{X}$  to be correlated with columns belonging to irrelevant covariates, hence producing spurious fits in the other covariates.

More precisely, in the noiseless case, if the true model function can be written exactly as the sum of step functions with, in the expanded design matrix  $\tilde{X}$ , corresponding column indices  $S$ , then correct recovery, given that LISO has fit non-zero fits to the true step functions, requires

$$\lambda \geq \tilde{X}_{S^c}^T \left( Y - \tilde{X}_S \left( \tilde{X}_S^T \tilde{X}_S \right)^{-1} \left( \tilde{X}_S^T Y - \lambda \right) \right) \quad (3.5)$$

$$= \tilde{X}_{S^c}^T \left( \tilde{X}_S \left( \tilde{X}_S^T \tilde{X}_S \right)^{-1} \lambda \right). \quad (3.6)$$

This is the Irrepresentable Condition of the LASSO, as detailed in Zhao and Yu [2006], Meinshausen and Bühlmann [2006], and it may fail if  $S$  is too large. With the LISO, then, the particular choice of  $\lambda$  itself influences the form the true covariates can take and so alters the criterion for Irrepresentability.

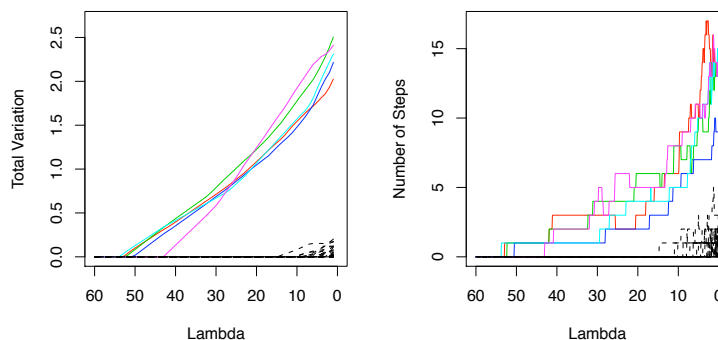


Figure 3.1: Effects of changing the regularisation parameter in the noiseless case.  $n = 100, p = 200$ . Each line represents how an individual covariate's estimate changes as  $\lambda$  varies, with the solid lines for the true covariates, while the dashed lines denote spurious fits on irrelevant variables.

These effects are illustrated in Figure 3.1, in which we have generated  $X$ , with  $n = 100, p=200$ , according to an uniform distribution, and produced  $Y$  as the sum of  $k = 5$  of the covariates. In other words,  $f$  is the sparse sum of linear functions. We give the full paths of fits in terms of, firstly, the total variation of fitted components  $\Delta(f_k)$ , and secondly the number of component steps in each covariate,

$$\left| \left\{ i : f_k \left( X_{(i)}^{(k)} \right) \neq f_k \left( X_{(i-1)}^{(k)} \right) \right\} \right|.$$

Of particular note is that, unlike the LASSO, even without noise, the size of the basis of step functions and the non-sparsity of the true signal means that as  $\lambda \rightarrow 0$ , we do not converge to the true sparsity pattern. However, with higher  $\lambda$ , the number of steps we choose diminishes rapidly, and as a result we can remove the spurious fits and simultaneously not mistakenly estimate the relevant covariates as zero.

In Figure 3.2, we add an independent normal noise component to  $Y$ , with variance chosen so that the signal to noise ratio,  $SNR = 5$ . In the new Total Variation plot, we see that the noise component has added additional noise fits in some of the irrelevant variables, and as in the LASSO these vanish for higher  $\lambda$ . Since the spurious fits vanish before the true covariate components do, we see that recovery of the true sparsity pattern is still possible in this case.

Now, in the above examples, we worked with the true sparsity pattern being assumed known. In real problems, we need to estimate the correct value of  $\lambda$  directly from the data. To do this, with the goal of

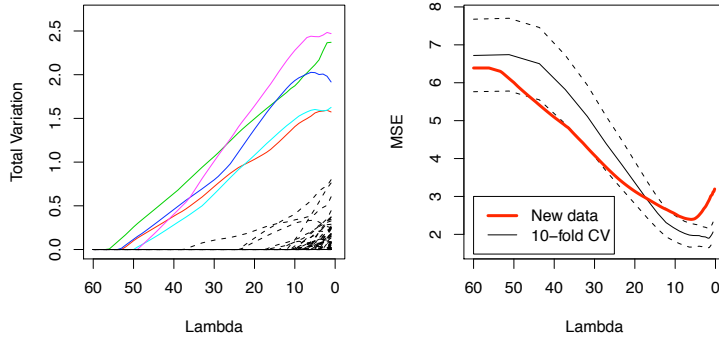


Figure 3.2: Effects of changing the regularisation parameter in the noisy case.  $n = 100, p = 200, SNR = 5$ . We show again in the first graph the total variation of each covariate estimate as  $\lambda$  alters, with solid lines for the truly important covariates, while the dashed lines denote spurious fits on irrelevant variables. The second graph shows the MSE from a 10-fold cross validation procedure with  $\pm 1$  s.d. in dashes, as well as the true MSE on a new set of data as the thick line.

recovering the correct sparsity pattern, is generally understood to be very difficult. (See e.g. Meinshausen and Bühlmann [2006] for some attempts.) However, as suggested in literature from Tibshirani [1996] onwards, cross validation is effective for minimising predictive error, and is illustrated by second graph of Figure 3.2. Here, we calculate CV error from a 10-fold cross validation. We may then take the  $\lambda$  that minimises the average mean squared error across the folds. If we desire a simpler model, we can, as is often suggested, take the largest  $\lambda$  that achieves a CV value within 1 s.d. of the minimum. Examining the thick line for the true predictive MSE shows that such a procedure, while not perfect, can give good results. In minimising predictive error, however, we do still fit some irrelevant covariates as non-zero, a phenomenon previously observed with the LASSO in Leng et al. [2006].

Now, unlike a LARS-like approach, LISO Backfitting will only give us the solution for an individual choice of  $\lambda$ . However, CV can still be practical, because coordinatewise minimisation can be very fast for sparse problems, something already observed for the normal LASSO [Friedman et al. 2007]. We can further reduce the computational cost by noting that LISO solutions for similar values of  $\lambda$  are likely to be similar, and hence use the result for one value of  $\lambda$  as a start point for the calculation for a nearby value of tuning parameter. This is especially effective if we order the  $\lambda$  values we need to calculate in decreasing order, since large  $\lambda$  solutions are more sparse and so faster to calculate.

## 4 Extensions and variations

A variety of extensions and variations of the basic LISO procedure may be proposed, that may offer improvements in some circumstances.

### 4.1 Bagged LISO

Bagging [Breiman 1996] may be used with the LISO, by aggregating the results of applying the LISO to a number of bootstrap samples through any of a variety of methods. This usually succeeds in smoothing

the observation, especially if we use smoothed bagging [Raviv and Intrator 1996]. However, this method is not reliably a great improvement in our empirical studies. Further, since the aggregated fit will produce a sparsity pattern involving a set of selected covariates that is the union of the selected covariates for each individual subsample calculation, we have that bagging will almost inevitably reduce the degree of sparsity in the fit, for any given degree of regularisation.

## 4.2 Adaptive LISO

A potential problem with the LISO is that it treats the constituent steps of each fit individually. In other words, there is no difference, in the eyes of the optimisation, between a fit that involves single step fits in a large number of covariates, and a single more complex fit in one covariate. As a result, the method may not achieve a great deal of sparsity in terms of covariates used, an issue we may want to rectify through making the algorithm in some sense recognise the natural grouping of steps in the step function basis.

Many existing solutions to this issue, such as Huang et al. [2009], involve explicitly or implicitly a Group LASSO [Yuan and Lin 2006] calculation to produce this grouping effect. Incorporating this into LISO is possible, though it may produce a greatly increased computational burden. Instead, we shall apply ideas from Zou [2006].

Consider the following two stage procedure – we first conduct an ordinary LISO optimisation, arriving at an initial fit  $(f_1^0, \dots, f_p^0)$ . Then, we conduct a second LISO procedure, this time introducing covariate weights  $w_1, \dots, w_p$  based on the first fit, and use the results of this as the output. We define the Adaptive LISO as the implementation of this, with  $w_k = 1/\Delta f_k^0$ , for  $k = 1, \dots, p$ .

---

### Algorithm 2 Adaptive LISO

---

- 1: Calculate initial fit  $f^0$  using LISO. (For instance, using Algorithm 1.)
- 2: Set  $w_k = 1/\Delta(f_k^0)$ , for  $k = 1, \dots, p$ .
- 3: Calculate, using e.g. Algorithm 1,

$$\arg \min_{f_1, \dots, f_p} \frac{1}{2} \left\| Y - \sum_{k=1}^p f_k(X^{(k)}) \right\|^2 + \sum_{k=1}^p w_k \Delta(f_k), \text{ with } f_k \in \mathcal{F}_k, k = 1, \dots, p.$$

- 4: If necessary, set  $f^0 = f$ , and repeat from Step 2.
- 

The analogy to the adaptive LASSO is that we apply a relaxation of the shrinkage for covariates with large fits in the initial calculation, and strengthen the shrinkage for covariates with small fits – indeed, omitting entirely from consideration covariates initially fitted as zero. Usually, more than one reweighted calculation is not required.

The Adaptive LISO encourages grouping of the underlying LASSO optimisation because large steps contribute to relaxation of other steps in the same covariate. In addition, it means that we in general require less regularisation of true fits in order to shrink irrelevant covariates to zero, through the concavity of the implied overall optimisation, to which we are essentially calculating a Local Linear Approximation [Zou and Li 2008]. We will also always enhance sparsity through this procedure – indeed, the fact that we reject straight away previously zero variables ensures the computational complexity of the method is

usually at most equal to that of repeating the original LISO procedure for each iteration.

It is, however, not clear what would be the best way to choose the tuning parameter introduced with each iteration of the process. We note that the discussants to Zou and Li [2008] have recommended a scheme based on individual prediction error minimising cross validation at every step, and our empirical studies suggest that this can pose significant improvements over the basic LISO. In our experiments, we also implement a variant of the adaptive procedure, LISO-SCAD, where instead the weights are calculated with an implied group-wise SCAD penalty. LISO-SCAD and LISO-Adaptive hence both fit under a broad group of possible LISO-LLA procedures.

### 4.3 Sign discovery and total variation penalty

Conventional isotonic regression focuses on the scenario where the monotonicity of the model function component in each covariate is known. However, this is not always realistic. Especially with large  $p$ , it may be the case that while we believe that the covariates contribute mostly in a monotonic way, we do not know, for at least some covariates, whether the covariate's component fit should be increasing, decreasing, or indeed even contribute non-monotonically. It is then perhaps reasonable to attempt to test for or estimate this monotonicity. This subject is dealt with by Bowman et al. [1998], amongst others, with a focus on the univariate case.

In higher dimensions, the problem is more difficult. One possible heuristic is to choose signs by a preliminary correlation check with the response. However, correlation is not invariant under general monotonic transformations, and examples exist where covariates have positive marginal effects, but, due to correlations between the covariates, turn out to have negative contributions in the final model. Now, with the LISO, it is trivial to use the same LISO-backfitting method for calculation with relaxation, or selective relaxation of the monotonicity condition. In this case, the relaxed form is just minimising the residual sum of squares, penalised by the total variation of the fitted step function. In other words, we find the minimiser, with  $f_1, \dots, f_p$  being univariate functions that have empirical mean zero and follow the specified combination of monotonicity constraints, of

$$L_\lambda((f_k)_{k=1}^p) := \frac{1}{2} \left\| Y - \sum_{k=1}^p f_k(X^{(k)}) \right\|^2 + \lambda \sum_{k=1}^p \Delta(f_k), \quad (4.1)$$

where  $\Delta(f_k)$  is calculated using

$$\Delta(f_k) = \begin{cases} f_k(\max X^{(k)}) - f_k(\min X^{(k)}) & \text{if } f_k \text{ is monotonically increasing,} \\ f_k(\min X^{(k)}) - f_k(\max X^{(k)}) & \text{if } f_k \text{ is monotonically decreasing,} \\ \sum_{i=2}^n |f_k(X_{(i)}^{(k)}) - f_k(X_{(i-1)}^{(k)})| & \text{otherwise.} \end{cases}$$

One way to implement this is to include reversed versions of non-monotonic covariates in the calculation, (and hence fitting a monotonically decreasing function to them as well as a monotonically increasing function) and then combine the fits with their corresponding twins after the calculation is complete.

**Definition 4.1.** Let  $(f_k)_{k=1}^p$  be any set of right continuous step functions with knots in the  $k$ -th covariate at  $x_1^{(k)} < \dots < x_{n_k}^{(k)}$ , and for each  $k$ , mean zero when evaluated at these knots. For  $k = 1, \dots, p$ , define

$f_k^+$  inductively as the right continuous step function with knot values

$$f_k^+ \left( x_1^{(k)} \right) = C_1^{(k)},$$

and for  $j = 1, \dots, n_k - 1$ ,

$$f_k^+ \left( x_{j+1}^{(k)} \right) = \begin{cases} f_k^+ \left( x_j^{(k)} \right) + f_k \left( x_{j+1}^{(k)} \right) - f_k \left( x_j^{(k)} \right) & \text{if } f_k \left( x_{j+1}^{(k)} \right) > f_k \left( x_j^{(k)} \right) \\ f_k^+ \left( x_j^{(k)} \right) & \text{otherwise,} \end{cases}$$

with  $C_1^{(k)}$  chosen so that  $\sum_{i=1}^n f_k^+ \left( x_i^{(k)} \right) = 0$ .

Similarly, define  $f_k^-$  as

$$f_k^- \left( x_1^{(k)} \right) = C_2^{(k)},$$

and for  $j = 1, \dots, n_k - 1$ ,

$$f_k^- \left( x_{j+1}^{(k)} \right) = \begin{cases} f_k^- \left( x_j^{(k)} \right) + f_k \left( x_{j+1}^{(k)} \right) - f_k \left( x_j^{(k)} \right) & \text{if } f_k \left( x_{j+1}^{(k)} \right) < f_k \left( x_j^{(k)} \right) \\ f_k^- \left( x_j^{(k)} \right) & \text{otherwise,} \end{cases}$$

with  $C_2^{(k)}$  chosen so that  $\sum_{i=1}^n f_k^- \left( x_i^{(k)} \right) = 0$ .

Then, in the fully non-monotonic case, we have the following theorem:

**Theorem 5.** Let  $(\hat{g}_k, \hat{h}_k)_{k=1}^p$  be the minimiser, with  $g_k \in \mathcal{F}_k, h_k \in -\mathcal{F}_k, k = 1, \dots, p$  of

$$M_\lambda \left( (g_k, h_k)_{k=1}^p \right) := \frac{1}{2} \left\| Y - \sum_{k=1}^p \left( g_k \left( X^{(k)} \right) + h_k \left( X^{(k)} \right) \right) \right\|^2 + \lambda \sum_{k=1}^p \left( \Delta(g_k) + \Delta(h_k) \right), \quad (4.2)$$

Then there is a one-to-one correspondence between such minimisers and minimisers  $(\hat{f}_k)$  to (4.1). This correspondence is given by the decomposition above, so that  $\hat{g}_k = \hat{f}_k^+, \hat{h}_k = \hat{f}_k^-$ , and  $\hat{g}_k + \hat{h}_k = \hat{f}_k$ , for all  $k$ .

An alternative implementation to using the above can be found by replacing the PAVA thresholding step in Algorithm 1 with a local thresholding style algorithm [Mammen and van de Geer 1997]. This can be slower, however, due to the computational burden involved with dealing with a covariate that would be taken exactly to zero, compared to checking (3.1) in the former case.

Now, extending Theorem 5, the Adaptive LISO can provide an alternative way of dealing with the problem of sign discovery. Starting with an initial non-monotonic LISO fit,  $\tilde{f}_k, k = 1, \dots, p$  say, we can conduct a second non-monotonic LISO fit, with covariate weights as in the Adaptive LISO case – except that we treat the positive and negative component fits separately, with respect to the weights used.

Let  $\tilde{f}_k^+, \tilde{f}_k^-, k = 1, \dots, p$  be the decomposed version of the initial fit. The  $M_\lambda$  approach will give us this decomposition directly, while we can apply the decomposition procedure from Definition 4.1 to obtain the appropriate decomposition with the second implementation, or indeed an initial fit found by any other method. Then, setting  $w_k^+ = 1/\Delta(\tilde{f}_k^+), w_k^- = 1/\Delta(\tilde{f}_k^-)$  we find the LISO adaptive sign discovery solution to be simply  $\hat{g} + \hat{h}$ , where  $\hat{g}_k, -\hat{h}_k \in \mathcal{F}_k, k = 1, \dots, p$  are solutions of

$$\arg \min_{\substack{g_1, \dots, g_p \\ h_1, \dots, h_p}} \frac{1}{2} \left\| Y - \sum_{k=1}^p (g_k(X^{(k)}) + h_k(X^{(k)})) \right\|^2 + \lambda \sum_{k=1}^p (w_k^+ \Delta(g_k) + w_k^- \Delta(h_k)).$$

Thus, as well as the effect seen in the adaptive LISO, where we have strengthened shrinkage of small function fits towards zero, functions with small negative or positive components in the initial fit will be shrunk towards an monotonically increasing or decreasing function respectively.

## 5 Numerical results

We will present a series of numerical examples designed to illustrate the effectiveness of the LISO in handling additive isotone problems. The experiments are calculated in R, using a standard desktop workstation. The full path solutions are found using a LISO modification to the Lars algorithm [Efron et al. 2004], while the larger comparison studies and fits are conducted using an implementation of the backfitting algorithm, with a logarithmic grid for the tuning parameter.

### 5.1 Example LISO fits

The following examples, conducted on single datasets, illustrate the performance of the algorithm.

#### 5.1.1 Boston Housing dataset

The Boston Housing dataset, as detailed in Harrison and Rubinfeld [1978], is a dataset often used in the literature to test estimators – see e.g. Hastie et al. [2003]. The dataset comprises of  $n = 506$  observations of 13 covariates, plus one response variable, which is the median house prices at each observation location. The response is known to be censored at the value 50, while the covariates range from crime statistics to discrete variables like index of accessibility to highways. We use here the version included in the R MASS library, though we shall discard the indicator covariate `chas`, for ease of presentation. (Experiments suggest that this variable does not have a great effect on the response, in any case.)

As suggested in Ravikumar et al. [2007], we will test the selection accuracy of the model by adding  $U(0,1)$  irrelevant variables. We add 28, so that our final  $p = 40$ . Since signs are not known, we will apply the sign discovery version of the LISO from Section 4.3, by first conducting a non-monotonic total variation fit, and then a weighted second fit. Tuning parameters are chosen by two 10-fold cross validations.

Our selected model, finally, is

$$Y = \alpha + f_1(\text{crim}) + f_2(\text{nox}) + f_3(\text{rm}) + f_4(\text{dis}) + f_5(\text{tax}) + f_6(\text{ptratio}) + f_7(\text{lstat}) + \varepsilon.$$

The remaining covariates are judged to have an insignificant effect on the response, with zero regression fits.  $f_3$  was found to be monotonically increasing,  $f_1$  non-monotonic, and the remaining functions monotonically decreasing. The full results are shown in Figure 5.1.

We see in our experiments that for higher values of  $\lambda$ , we successfully remove all the irrelevant variables, and end up with only a small number of selected variables to explain the response. However, in the one step procedure, the amount of shrinkage required is often large. With cross validation as a

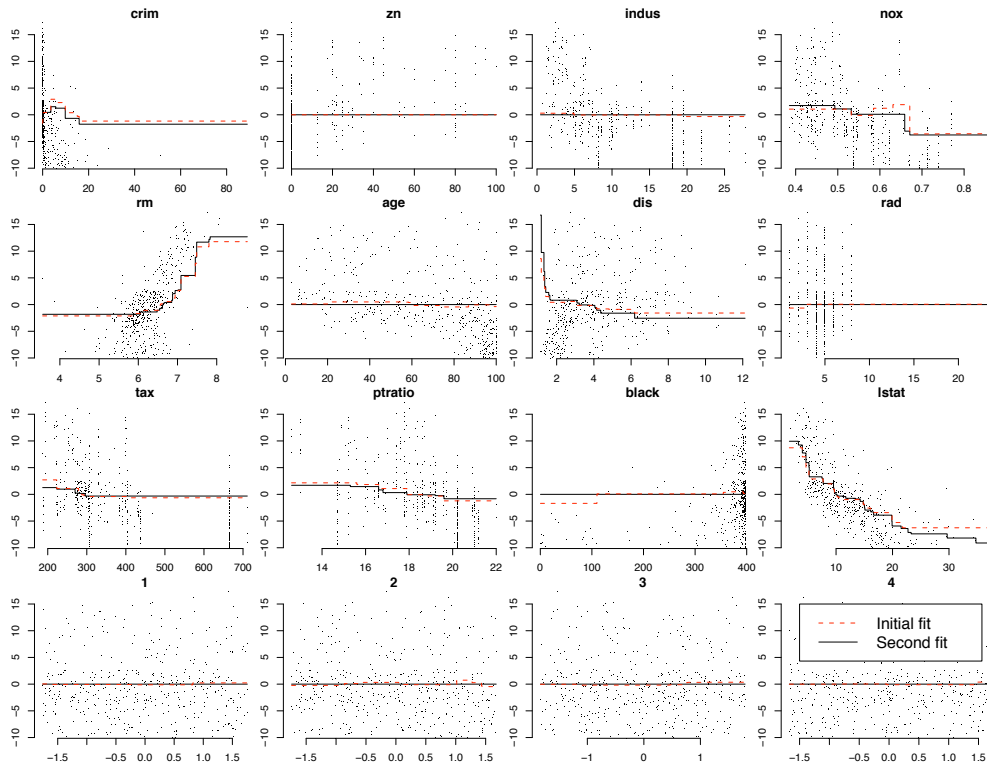


Figure 5.1: Fitted component functions on the Boston Housing dataset, for covariates originally present in the data plus four others. The dashed line shows the selected model after the first LISO step, while the solid black line shows the final result of the adaptive sign finding procedure. The single step fit produced additional non-zero fits in some of the artificial covariates, which are not shown, while the two step procedure fit all of them as zero.

criterion, we do choose a  $\lambda$  that involves some irrelevant variables as well, though these are in general small in magnitude. A second step greatly improves the model selection characteristics, as well as creating monotonicity which is often absent in the first step.

It is interesting to contrast our fit with the findings from using SpAM [Ravikumar et al. 2007]. Bearing in mind that our problem was in some sense more difficult, since we had 12 original covariates instead of 10 (**rad** and **zn** were not included in the SpAM study), and 28 artificial covariates instead of 20, our findings are largely similar. In addition to the covariates selected in SpAM, we add a fairly large effect from **nox**, and smaller effects in **dis** and **tax**. The most significant fits on **rm** and **lstat** are very similar, though the LISO fit is clearly less smooth. However, while almost all of the fits from SpAM exhibit non-monotonicity, the LISO fit we have found is mostly monotone, aside from the fit in **crim**.

The non-monotonicity found in **crim** may seem problematic, given the interpretation of that covariate as a crime rate. While, nevertheless, this is a characteristic present in the conditional residuals, perhaps it would be reasonable to impose a monotonicity constraint instead.

### 5.1.2 Artificial dataset

We are also interested in the success of LISO in correctly selecting variables for varying levels of  $n$  and  $p$ . We adopt the following setup – we generate pairs  $X \in \mathbb{R}^{n \times p}, Y \in \mathbb{R}^n$  by

$$X_{ij} \sim \text{Uniform}(0, 1)$$

$$Y_i = 2 \left( X_i^{(1)} \right)_+^2 + X_i^{(2)} + \text{sign} \left( X^{(3)} \right) \left| X_i^{(3)} \right|^{1/5} + 2I_{\{X_i^{(4)} > 0\}} + \varepsilon_i$$

with  $n = 1024, p = 1024$ , independent  $\varepsilon_i \sim N(0, 1)$ . The covariates are then centred and standardised to have mean zero and variance 1, and  $Y$  is centred to have mean zero.

For  $p' = 32, 64, 128, 256, 512, 1024, n' = 5, 10, 15, \dots$ , we then take as  $X', Y'$  subsets of  $X, Y$  corresponding to the first  $p'$  columns of  $X$ , and random samples without replacement of  $n'$  rows of  $X, Y$ . Hence we consider the problem of correctly finding 4 true variables, from amongst  $p'$  potential ones, based on  $n'$  observations. We quantify the success of LISO by looking at the proportion of 50 replications where the algorithm, for at least one value of  $\lambda$ , produces an estimate where the true covariates have at least one step while the other covariates are taken to zero. (We adopt this framework so as to reduce the additional noise from generating a complete new random dataset with each attempt.)

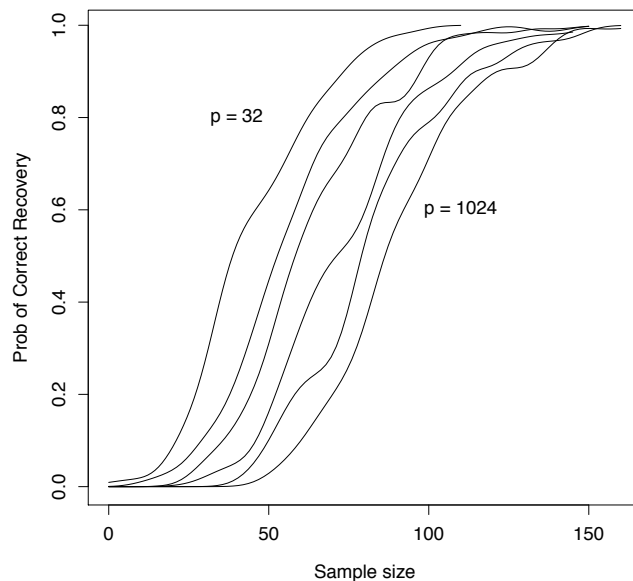


Figure 5.2: Probabilities of correct sparsity recovery with 4 true nonlinear but monotonic covariates,  $SNR = 4$ . Each line shows how the recovery probability changes as the sample size  $n$  changes for a single value of  $p$ , taking values  $2^5, \dots, 2^{10}$ .

Figure 5.2 gives these results. As we can see, as in a variety of LASSO-type algorithms [Wainwright 2006], there is a sharp threshold between success and failure in recovery of sparsity patterns as a function of  $n$ . Moreover, as we increase  $p$  exponentially, the required number of observations  $n$  increases much more slowly, thus implying that  $p \gg n$  recovery is possible.

Figure 5.3 gives an example of LISO fits arising from this simulation. The dashed lines shows the results of the LISO under the minimum regularisation required for correct sparsity recovery – note the high level of shrinkage required to shrink the other variables to zero. This shrinkage exhibits itself as

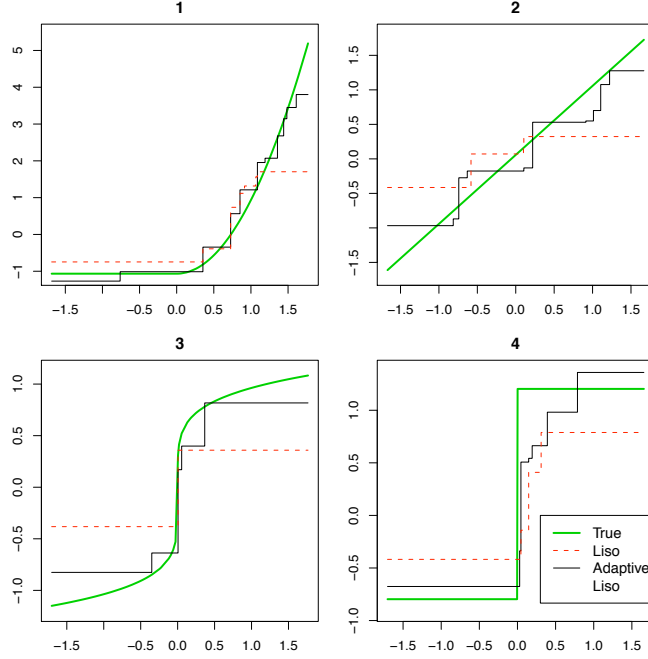


Figure 5.3: Example LISO covariate fits, for  $n = 180$ ,  $p = 1024$ . The true component functions are given by the thick line, while the dashed line gives the raw LISO fit for the smallest amount of regularisation required to bring spurious fits in irrelevant covariates to zero. The solid black line shows a fit made by the adaptive liso, using tuning parameters found by cross validation. The fitted and true model functions for all 1020 remaining covariates are all constant zero.

not only a thresholding on the ends of the component fits, which we have seen in the univariate case, a , but also an additional loss of complexity in the middle parts of each component fit. We can avoid these shrinkages by using this initial result to perform the Adaptive LISO, in the solid black line, thus greatly improving the fit while still keeping the correct sparsity pattern recovery. As an added bonus, we get good results here with the Adaptive LISO even without using knowledge of the true process that generated the data.

## 5.2 Comparison studies

We shall now compare LISO to a range of other procedures in some varying contexts. Varying  $f$  between scenarios, consider generating pairs  $X, Y$  by, for each repetition,

$$\begin{aligned} X_i^{(j)} &\sim \text{Uniform}(-1, 1), \quad i = 1, \dots, n, j = 1, \dots, p \\ \varepsilon_i &\sim N(0, 1), \quad i = 1, \dots, n \\ Y_i &= f(X_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n. \end{aligned}$$

100 repetitions were done of each combination of model and noise level, with  $\sigma$  chosen to give  $SNR = 1, 3$  or  $7$ , plus one further case where we have  $SNR = 3$  but  $X$  is instead generated to have stronger correlation between the covariates, as a rescaled (to the range  $(-1, 1)$ ) version of  $\Phi(Z)$ ,  $Z \sim N(0, \Sigma)$ ,

with  $\Sigma_{ij} = 2^{-|i-j|}$ .

For comparison, we will compare the performance of LISO and LISO-LLA (both Adaptive and SCAD), calculated using the backfitting algorithm, to

- Random Forests (RF), from Breiman [2001]. A tree based method using aggregation of trees generated using a large number of resamplings.
- Multiple Adaptive Regression Splines (MARS), from Friedman [1991], using the `earth` implementation in R. A method using greedy forward/backward selection with a hockeystick shaped basis. We use a version restricted to additive model fitting.
- Sparse Additive Models (SpAM), from Ravikumar et al. [2007]. A similar group LASSO based method using soft thresholding of component smoother fits.
- Sparsity Smoothness Penalty (SSP), from Meier et al. [2009]. A group LASSO based method using two penalties – a sparsity penalty and an explicit smoothness penalty.

For the choice of tuning parameter in all algorithms, we take the value that minimises the prediction error on a separate validation set of the same size as the training set. (Note that in the case of SSP, due to the slowness of finding two separate tuning parameters, we instead perform a small number of initial full validation runs for each scenario. We then plug in the averaged smoothness tuning parameter in all following runs, optimising for only the sparsity parameter.)

We record both the mean value across runs of the MSE on predicting a new test set (generated without noise), and, in brackets, the mean relative MSE, defined for the  $k$ -th algorithm on each individual run as

$$MSE_{\mathbf{Relative}}^k := \frac{MSE^k}{\min_{j=1,\dots,7} MSE^j}.$$

### 5.2.1 All components linear

In this case, we have the response being just a scaled sum of  $k = 5$  randomly chosen covariates, plus a noise term.  $n = 200$ ,  $p = 50$  overall. In the test set, the variance of the response (and hence the MSE of a constant prediction) was approximately 1.7.

Algorithm	SNR = 7	SNR = 3	SNR = 1	SNR = 3, Correlated
LISO	0.113 (4.70)	0.186 (3.33)	0.358 (2.41)	0.203 (3.43)
LISO-Adaptive	0.070 (2.94)	0.118 (2.18)	0.242 (1.62)	0.134 (2.27)
LISO-SCAD	0.113 (4.71)	0.186 (3.33)	0.437 (3.00)	0.202 (3.41)
SpAM	0.082 (3.29)	0.149 (2.57)	0.346 (2.24)	0.159 (2.59)
SSP	0.026 (1.00)	0.061 (1.00)	0.167 (1.02)	0.065 (1.00)
RF	0.286 (11.97)	0.319 (5.85)	0.504 (3.36)	0.361 (6.21)
MARS	0.146 (6.28)	0.354 (6.72)	1.027 (6.91)	0.417 (7.19)

Because of the sparsity and additivity in the data, all LASSO-like methods do better than RF, a pattern that continues in all of these simulation studies. Indeed, due to the random selection of covariates in the RF algorithm, the presence of spurious covariates seems to produce a phenomenon of

excess shrinkage, which can be clearly see in plots of fitted values versus response values. Using the scaling corrections provided in the R implementation improves things, but not to a great extent. MARS, similarly, has difficulty in finding the correct variables. With such large  $p$ , the set of possible hockey stick bases MARS has to search through is very large, and hence the underlying greedy stepwise selection component of the algorithm is in general unsuccessful at handling this problem.

Amongst the LASSO-like methods, perhaps unsurprisingly, the SSP method performs by far the best, owing to the large degree of smoothness in the true model function. LISO-Adaptive is second best, however, beating SpAM even though it does not have an internal smoothing effect. The basic LISO method itself underperforms, perhaps because it does not strongly enforce sparsity amongst the original covariates.

Unexpectedly, LISO-SCAD performs fairly equivalently to the LISO itself in this and all following simulations. A likely explanation is that for sufficient regularisation to take place to take spurious covariates to zero, the penalty function is such that the solution lies mostly on the part of the penalty where it is identical to the original total variation penalty.

The introduction of a moderate amount of correlation does not greatly affect the performance of any of the algorithms.

### 5.2.2 Mixed powers

In this case, the response has a more complex relation to the covariates:

$$Y_i = \sum_{k=1}^5 f_k(X_i^{(a_k)}) + \sigma \varepsilon_i$$

$$f_1(x) = \text{sign}(x + C_1) |x + C_1|^{0.2}$$

$$f_2(x) = \text{sign}(x + C_2) |x + C_2|^{0.3}$$

$$f_3(x) = \text{sign}(x + C_3) |x + C_3|^{0.4}$$

$$f_4(x) = \text{sign}(x + C_4) |x + C_4|^{0.8}$$

$$f_5(x) = x + C_5$$

In this case, we have again  $n = 200, p = 50$ .  $C_1, \dots, C_5$  are small shifts, randomly generated as  $Uniform(-1/4, 1/4)$ , and  $a_1, \dots, a_5$  are covariates randomly chosen without replacement. In the test set, the variance of the response was approximately 2.6.

Algorithm	SNR = 7	SNR = 3	SNR = 1	SNR = 3, Correlated
LISO	0.128 (1.49)	0.230 (1.50)	0.459 (1.41)	0.255 (1.50)
LISO-Adaptive	0.088 (1.01)	0.160 (1.00)	0.352 (1.06)	0.177 (1.01)
LISO-SCAD	0.128 (1.49)	0.229 (1.49)	0.587 (1.82)	0.254 (1.50)
SpAM	0.157 (1.83)	0.267 (1.75)	0.539 (1.68)	0.285 (1.69)
SSP	0.126 (1.47)	0.226 (1.49)	0.429 (1.33)	0.252 (1.51)
RF	0.358 (4.21)	0.450 (2.96)	0.721 (2.26)	0.495 (2.96)
MARS	0.319 (3.78)	0.678 (4.54)	1.936 (6.32)	0.783 (4.71)

With the new, non-linear model function, the LISO and LISO-SCAD now perform equally as well as the SSP, while the adaptive LISO performs significantly better, being the best in almost all runs. All four methods outperform SpAM, and greatly outperform RF and MARS.

In this case, the explanation is that for fractional powers, the component functions are relatively flat in the extremes of the covariate range, with most of the variation occurring in the middle of the range. SpAM and SSP are unable to capture the sharp transition point of the small root functions without introducing inappropriate variability at the ends of the fit, and hence both perform significantly worse than previously. The LISO based methods, however, do not explicitly smooth the fit and only threshold the extremes. Being thus adapted to this sort of function, they actually improve their performance in proportional terms relative to the variance of the test set.

### 5.2.3 Mixed powers, large $p$

In this scenario, our model is the same as before, save that we have many more spurious covariates, resulting in  $n = 200, p = 200$ . The variance of the test response is unchanged at approximately 2.6.

Algorithm	SNR = 7	SNR = 3	SNR = 1	SNR = 3, Correlated
LISO	0.166 (1.89)	0.283 (1.86)	0.638 (1.78)	0.286 (1.84)
LISO-Adaptive	0.090 (1.00)	0.156 (1.00)	0.384 (1.01)	0.160 (1.01)
LISO-SCAD	0.169 (1.93)	0.292 (1.91)	0.935 (2.71)	0.296 (1.90)
SpAM	0.201 (2.32)	0.329 (2.17)	0.779 (2.21)	0.331 (2.14)
SSP	0.156 (1.78)	0.274 (1.80)	0.604 (1.73)	0.274 (1.78)
RF	0.504 (5.86)	0.588 (3.86)	0.992 (2.84)	0.593 (3.84)
MARS	0.805 (9.27)	1.704 (11.49)	4.707 (13.84)	1.763 (11.60)

In this case, LISO preserves its superiority. Due to the effect of high dimensionality, all algorithms see their performance decline - except the adaptive LISO, which has an increased MSE of less than 3% in the low noise case. This is due to the adaptive step, which retains a very sparse fit, picking the relevant variables even as the number of predictors grows.

## 6 Discussion

We have presented here a method of extending ideas from LASSO on linear models to the framework of non-parametric estimation of isotonic functions. We have found that in many contexts, it inherits the

behaviour of the LASSO in that it allows sparse estimation in high dimensions. By using our backfitting procedure, we have also shown empirically that it can be very competitive with many current methods, both in terms of computational time and memory requirements, and in terms of predictive accuracy. The precise criteria that govern its success would require further work, and it would be interesting to see if similar LASSO-style oracle results apply.

In addition, we find that an LLA/adaptive scheme is highly effective and efficient at improving the algorithm in a two step approach, producing sparser results and very high predictive accuracy. Further adaptations allow the LISO method to be used when monotonicity is assumed but the direction of the monotonicity is not known. To the authors' knowledge, this has not been attempted previously in this type of problem, and it would be interesting to see if LLA and similar concave penalty procedures can produce effective replacements for the group LASSO in the underlying calculation of non-parametric LASSO generalisations.

## A Appendix: Proofs of theorems

### Proof of Theorem 1

Our methodology is to show that adding boundary constraints to constrained or unconstrained isotonic regression problems result in unique solutions that are simply Winsorised PAVA estimates, and then demonstrate a method of constructing any LISO solution, in the univariate case, through boundary constraints. We prove first the following lemma, which provides an induction step in our eventual argument:

**Lemma 1.** *Suppose for  $A \leq B$ ,  $X = (X_1, \dots, X_n)$ ,  $Y = (Y_1, \dots, Y_n)$ ,  $X_i \in \mathbb{R}$  for all  $i$ , the Winsorised PAVA,*

$$f_1(x) = \widehat{f}_{>A, <B}(x) := \begin{cases} A & \text{if } \widehat{f}_{PAVA}(x) < A \\ B & \text{if } \widehat{f}_{PAVA}(x) > B \\ \widehat{f}_{PAVA}(x) & \text{otherwise.} \end{cases}$$

*solves the boundary constrained isotonic regression problem,*

$$\min_f \|Y - f(X)\|^2 \quad \text{such that } f \text{ monotone, } A \leq f(x) \leq B, \quad \forall x. \quad (\text{A.1})$$

*Then for  $A \leq A' \leq B' \leq B$ ,  $f_2 \equiv \widehat{f}_{>A', <B'}$  solves the further constrained isotonic regression problem*

$$\min_f \|Y - f(X)\|^2 \quad \text{such that } f \text{ monotone, } A' \leq f(x) \leq B', \quad \forall x. \quad (\text{A.2})$$

*Further, this solution is unique, in terms of its fitted values  $f(X_1), \dots, f(X_n)$ .*

*Proof of Lemma 1.* It suffices to prove the lemma for the case of  $A \leq A' \leq B' = B$ , since the argument for  $A = A' \leq B' \leq B$  is identical, and we can proceed to the full Lemma by adding the top and bottom constraints one by one.

Note that, specifying  $f$  through the fitted values  $f(X_1), \dots, f(X_n)$ ,  $\|Y - f(X)\|^2$  is strictly convex when considered as a function of  $f$ , and the constraints give a convex feasible set. Hence, for any combination of  $A, B$ , solutions must exist and be unique at  $X_1, \dots, X_n$ .

Therefore, let  $g$  be the solution of the optimisation (A.2), for  $A \leq A' \leq B' = B$ . Suppose for contradiction that  $g \neq f_2 \equiv \widehat{f}_{>A', <B'}$ .

Let  $u_f, u_g$  be the points where the functions  $f_2, g$  respectively exceed  $A'$ .

$$\begin{aligned} u_f &= \inf \{X_i \text{ s.t. } f_2(X_i) > A'\} \\ u_g &= \inf \{X_i \text{ s.t. } g(X_i) > A'\}. \end{aligned}$$

Then we have two cases.

(a) If  $u_f \leq u_g$ , then consider a new function  $\tilde{f}$  where

$$\tilde{f}(x) = \begin{cases} f_1(x) & \text{if } x < u_f \\ g(x) & \text{if } x \geq u_f. \end{cases} \quad (\text{A.3})$$

$\tilde{f}$  would be an increasing function satisfying the conditions of (A.1). Because  $g \neq f_2$ , and  $g$  and  $f_2$  are both equal to  $A'$  when restricted to  $\{x : x < u_f\}$ , it must be the case that  $g \neq f_2 \equiv f_1$  when restricted to  $\{x : x \geq u_f\}$ . Therefore,  $\tilde{f} \neq f_1$ . The residual sum of squares is then, applying (A.2) optimality of  $g$ ,

$$\begin{aligned} \|Y - \tilde{f}(X)\|^2 &= \sum_{X_i < u_f} (Y_i - f_1(X_i))^2 + \sum_{X_i \geq u_f} (Y_i - g(X_i))^2 \\ &= \sum_{X_i < u_f} (Y_i - f_1(X_i))^2 + \|Y - g(X)\|^2 - \sum_{X_i < u_f} (Y_i - A')^2 \\ &\leq \sum_{X_i < u_f} (Y_i - f_1(X_i))^2 + \|Y - f_2(X)\|^2 - \sum_{X_i < u_f} (Y_i - A')^2 \\ &= \sum_{X_i < u_f} (Y_i - f_1(X_i))^2 + \sum_{X_i \geq u_f} (Y_i - f_1(X_i))^2 \\ &= \|Y - f_1(X)\|^2. \end{aligned}$$

Therefore,  $\tilde{f}$  is optimal for (A.1). This contradicts uniqueness and (A.1) optimality of  $f$ .

(b) If  $u_f > u_g$ , then if we define  $\tilde{f}$  this time as

$$\tilde{f}(x) = \begin{cases} f_1(x) & \text{if } x < u_g \\ g(x) & \text{if } x \geq u_g, \end{cases} \quad (\text{A.4})$$

we obtain another increasing function satisfying the conditions of (A.1). As before, because we have assumed that  $g \neq f_2$ , and yet  $g \equiv f_2 \equiv A'$  for  $\{x : x < u_g\}$ ,  $g \neq f_2 \equiv f_1$  when restricted to  $\{x : x \geq u_g\}$ . This means that  $\tilde{f} \neq f_1$ , so unique optimality of  $f_1$  versus  $\tilde{f}$  means that

$$\begin{aligned} \sum_{X_i \geq u_g} (Y_i - f_1(X_i))^2 &= \|Y - f_1(X)\|^2 - \sum_{X_i < u_g} (Y_i - f_1(X_i))^2 \\ &< \|Y - \tilde{f}(X)\|^2 - \sum_{X_i < u_g} (Y_i - f_1(X_i))^2 \\ &= \sum_{X_i \geq u_g} (Y_i - g(X_i))^2. \end{aligned}$$

In addition, because  $u_f > u_g$ , setting  $\delta = (g(u_g) - A') / (g(u_g) - f_1(u_g))$  makes  $\tilde{g}$ , defined as

$$\tilde{g}(x) = \begin{cases} A' & \text{if } x < u_g \\ (1 - \delta)g(x) + \delta f_1(x) & \text{if } x \geq u_g, \end{cases} \quad (\text{A.5})$$

an increasing function satisfying the conditions of (A.2). By definition,  $g(u_g) > A'$  and  $f_1(u_g) < A'$ , so  $\delta \in (0, 1)$ , implying that  $\tilde{g}$  is a nontrivial convex combination of  $g$  and  $f_1$ , when restricted to  $\{x : x \geq u_g\}$ .

But by convexity,

$$\begin{aligned} \|Y - \tilde{g}(X)\|^2 &= \sum_{X_i < u_g} (Y_i - A')^2 + \sum_{X_i \geq u_g} (Y_i - (1 - \delta)g(X_i) - \delta f_1(X_i))^2 \\ &< \sum_{X_i < u_g} (Y_i - A')^2 + \sum_{X_i \geq u_g} (Y_i - g(X_i))^2 = \|Y - g(X)\|^2. \end{aligned}$$

This contradicts optimality of  $g$ .

Therefore,  $g \equiv f_2$  is the unique solution to (A.2).  $\square$

*Proof of Theorem 1.* We can prove Theorem 1 as a simple corollary.

When  $\lambda = 0$ , our objective function  $L_\lambda$  is strictly convex and quadratic, and indeed is the same as the PAVA optimisation. Hence, an unique optimal solution exists, and is given by  $\hat{f}_0 = \hat{f}_{PAVA}$ . Set  $A_0 = -\infty, B_0 = \infty$ . Then  $\hat{f}_0 \equiv \hat{f}_{>A_0, <B_0}$  is feasible for, and so, must also solve the constrained optimisation (A.1), with constraints at infinity.

For  $\lambda > 0$ ,  $L_\lambda$  and the domain we maximise it in are both still convex, with  $L_\lambda$  strictly convex and increasing away from the origin outside of a neighbourhood. Therefore, an unique bounded solution  $\hat{f}_\lambda$  must exist. Set  $A_\lambda, B_\lambda$  to be the upper and lower bounds of this solution.

$$A_\lambda = \min_i \hat{f}_\lambda(X_i), \quad B_\lambda = \max_i \hat{f}_\lambda(X_i).$$

Then consider the solution to the constrained isotonic least squares problem

$$\tilde{f}_\lambda = \arg \min_f \|Y - f(X)\|^2 \quad \text{such that } f \text{ monotone, } A_\lambda \leq f(x) \leq B_\lambda, \quad \forall x. \quad (\text{A.6})$$

$\Delta(\tilde{f}_\lambda) \leq B_\lambda - A_\lambda = \Delta(\hat{f}_\lambda)$ , and since  $\hat{f}_\lambda$  is feasible for A.6,  $\|Y - \tilde{f}_\lambda(X)\|^2 \leq \|Y - \hat{f}_\lambda(X)\|^2$ .

Therefore,  $\tilde{f}_\lambda$  is optimal for (2.1). Hence, by uniqueness,  $\tilde{f}_\lambda \equiv \hat{f}_\lambda$ , so for suitable  $A_\lambda, B_\lambda$  it suffices to solve the bound constrained least squares optimisation (A.6), to find the LISO solution.

But by Lemma 1, the solution of (A.6) for  $> A_\lambda < B_\lambda$  is just the Winsorised PAVA solution  $\hat{f}_{>A_\lambda, <B_\lambda}$ , so  $\hat{f}_\lambda \equiv \tilde{f}_\lambda \equiv \hat{f}_{>A_\lambda, <B_\lambda}$ .  $\square$

## Proof of Theorem 2

*Proof.* For  $\lambda = 0$ ,  $\hat{f}_0 \equiv \hat{f}_{PAVA}$ , so choosing  $A_0 = \min(\hat{f}_{PAVA}(x))$ ,  $B_0 = \max(\hat{f}_{PAVA}(x))$  is clearly a optimum for (2.1) that satisfies (3.2) and (3.3). Assume therefore  $\lambda > 0$ .

Now, from Barlow et al. [1972], the unregularised PAVA solution  $\hat{f}_{PAVA}$ , considered as a right continuous step function, is an example of a regressogram. In other words, there exists a partition into disjoint intervals of  $\mathbb{R}$ ,  $P_1, \dots, P_m$ , with the value of  $\hat{f}_{PAVA}(x)$  on each interval being the mean of the observed  $Y$  for  $X$  falling within the interval.

$$\widehat{f}_{PAVA}(x) = \frac{\sum_{i=1}^n Y_i I_{\{X_i \in P_j\}}}{\sum_{i=1}^n I_{\{X_i \in P_j\}}} = \widehat{f}_{PAVA}(P_j) \quad \text{for all } x \in P_j$$

Using  $\widehat{f}_0 \equiv \widehat{f}_{PAVA}$ , the LASSO criterion for the thresholded function  $\widehat{f}_{>A, <B}$  then can be written as,

$$\begin{aligned} L_\lambda(\widehat{f}_{>A, <B}) &= \frac{1}{2} \left\| Y - \widehat{f}_{>A, <B} \right\|^2 + \lambda(B - A) \\ &= \frac{1}{2} \sum_{i=1}^n \left( (Y_i - \widehat{f}_0(X_i))^2 I_{\{\widehat{f}_0(X_i) \in [A, B]\}} + (A - Y_i)^2 I_{\{\widehat{f}_0(X_i) < A\}} + (Y_i - B)^2 I_{\{\widehat{f}_0(X_i) > B\}} \right) \\ &\quad + \lambda(B - A) \\ &= \frac{1}{2} \left\| Y - \widehat{f}_0 \right\|^2 + \frac{1}{2} \sum_{i=1}^n \left( (A - \widehat{f}_0(X_i))^2 + 2(A - \widehat{f}_0(X_i))(\widehat{f}_0(X_i) - Y_i) \right) I_{\{\widehat{f}_0(X_i) < A\}} \\ &\quad + \frac{1}{2} \sum_{i=1}^n \left( (\widehat{f}_0(X_i) - B)^2 + 2(\widehat{f}_0(X_i) - B)(Y_i - \widehat{f}_0(X_i)) \right) I_{\{\widehat{f}_0(X_i) > B\}} + \lambda(B - A) \\ &= \frac{1}{2} \left\| Y - \widehat{f}_0 \right\|^2 + \frac{1}{2} \sum_{i=1}^n \left( (A - \widehat{f}_0(X_i))_+^2 + (\widehat{f}_0(X_i) - B)_+^2 \right) + \lambda(B - A) \\ &\quad + \sum_{j=1}^m \sum_{i=1}^n \left( (\widehat{f}_0(X_i) - B)(Y_i - \widehat{f}_0(X_i)) I_{\{\widehat{f}_0(X_i) > B\}} \right) I_{\{X_i \in P_j\}} \\ &\quad + \sum_{j=1}^m \sum_{i=1}^n \left( (A - \widehat{f}_0(X_i))(\widehat{f}_0(X_i) - Y_i) I_{\{\widehat{f}_0(X_i) < A\}} \right) I_{\{X_i \in P_j\}} \\ &= \frac{1}{2} \left\| Y - \widehat{f}_0 \right\|^2 + \frac{1}{2} \sum_{i=1}^n \left( (A - \widehat{f}_0(X_i))_+^2 + (\widehat{f}_0(X_i) - B)_+^2 \right) + \lambda(B - A) \\ &\quad + \sum_{j=1}^m (\widehat{f}_0(P_j) - B)_+ \sum_{i=1}^n \left( (Y_i - \widehat{f}_0(X_i)) \right) I_{\{X_i \in P_j\}} \\ &\quad + \sum_{j=1}^m (A - \widehat{f}_0(P_j))_+ \sum_{i=1}^n \left( (\widehat{f}_0(X_i) - Y_i) \right) I_{\{X_i \in P_j\}} \\ &= \frac{1}{2} \left\| Y - \widehat{f}_0 \right\|^2 + \frac{1}{2} \sum_{i=1}^n \left( (A - \widehat{f}_0(X_i))_+^2 + (\widehat{f}_0(X_i) - B)_+^2 \right) + \lambda(B - A). \end{aligned}$$

We seek a minimum to this with  $A \leq B$ . Differentiating in  $A$  and  $B$ , and setting equal to zero, gives,

$$\sum_{i=1}^n (\widehat{f}_0(X_i) - B)_+ = \lambda \tag{A.7}$$

$$\sum_{i=1}^n (A - \widehat{f}_0(X_i))_+ = \lambda. \tag{A.8}$$

Note that the left hand side in both cases is a piecewise linear, continuous and monotone (indeed, decreasing in  $B$  and increasing in  $A$ ) function of the threshold, equalling zero for  $A = A_0, B = B_0$ . Further, since the PAVA is a regressogram,  $\sum_{i=1}^n \widehat{f}_0(X_i) = \sum_{i=1}^n Y_i$ , so

$$\sum_{i=1}^n (\widehat{f}_0(X_i) - \bar{Y})_+ = \sum_{i=1}^n (\bar{Y} - \widehat{f}_0(X_i))_+ = \frac{1}{2} \sum_{i=1}^n |\widehat{f}_{PAVA}(X_i) - \bar{Y}|.$$

We therefore have two cases. If

$$2\lambda \leq \sum_{i=1}^n |\widehat{f}_{PAVA}(X_i) - \bar{Y}|,$$

then we can find solutions to (A.7) and (A.8) for which  $A_\lambda \leq \bar{Y} \leq B_\lambda$ , producing a minimiser for (2.1).

Otherwise, (A.7) and (A.8) require that  $A > \bar{Y}$  and  $B < \bar{Y}$ . Hence, there are no solutions to our minimisation with  $A < B$ .

For a solution on the boundary, we require  $A = B$ , so

$$L_\lambda(f_{>A, <B}) = \frac{1}{2} \sum_{i=1}^n (Y_i - A)^2.$$

This is clearly minimised by  $A_\lambda = B_\lambda = \bar{Y}$ .

Corollary 3 follows from the fact that from properties of the PAVA,

$$\arg \max_m \left| \sum_{i=1}^m (Y_{\pi(i)} - \bar{Y}) \right| = \max \left\{ m : \hat{f}_{PAVA}(X_{\pi(m)}) < 0 \right\}.$$

□

## Proof of Theorem 4

*Proof.* If we go through the covariates in a pre-determined order, then we can apply a theorem proved in Tseng [2001]. However, the proof is simplified in the case where we go through the variables in a random, independent order with each iteration, which we will show now.

Because we are doing repeated minimisations, for all  $m$ ,  $L_\lambda(f^{m+1}) \leq L_\lambda(f^m)$ . Moreover,  $L_\lambda$  is bounded below by 0. Therefore,  $L_\lambda(f^m)$  must converge monotonically in probability as  $m$  increases.

Now, choose any  $\delta_1 > 0$ . We define  $A_m$  as the event that there exists  $k \in \{1, \dots, p\}$  and  $g^m : \mathbb{R} \rightarrow \mathbb{R}$  such that

$$L_\lambda(f_1^m, \dots, f_p^m) - L_\lambda(f_1^m, \dots, f_{k-1}^m, g^m, f_{k+1}^m, \dots, f_p^m) \geq \delta_1,$$

the event that we can improve on the current fit by at least  $\delta_1$  by changing only one component estimate.

Now, with the  $m + 1$ th iteration, we have an equal probability of picking any one of  $p$  covariates as the first fitted of our new backfitting cycle, and hence

$$\text{Prob}(L_\lambda(f^m) - L_\lambda(f^{m+1}) \geq \delta_1) \geq \text{Prob}(A_m)/p.$$

But  $L_\lambda(f^m)$  converges in probability, so  $\text{Prob}(A_m) \rightarrow 0$  for all  $\delta_1 > 0$ .

$L_\lambda$  is continuously differentiable in the interior of  $\oplus \mathcal{F}_k$ . The set  $f \in \oplus \mathcal{F}_k$  such that  $L_\lambda(f) \leq L_\lambda(f^0)$  is closed, and compact. Therefore, the above implies that for any  $\eta > 0$ ,  $\delta_2 > 0$ , there exists  $M_1$  such that for all  $m > M_1$ , the subdifferential of  $L_\lambda$  at  $f^m$  contains a plane that is within  $\delta_2$  of zero with probability at least  $1 - \eta$ .

Therefore, by considering sufficiently small values of  $\delta_2$ , this implies that for all  $\eta > 0$ ,  $\delta_3 > 0$ , there exists  $M_2$  such that for all  $m > M_2$ , there exists with probability at least  $1 - \eta$ , another sum of functions  $\tilde{f}^m \in \oplus \mathcal{F}_k$  satisfying  $\|\tilde{f}^m - f^m\| < \delta_3$ , at which  $L_\lambda$  contains the zero plane in its subdifferential.

Since  $L_\lambda$  is convex,  $\tilde{f}^m$  is a global minimiser of  $L_\lambda$ . Taking  $\eta, \delta_3$  to zero, we see by continuity of  $L_\lambda$  that  $L_\lambda(f^m)$  converges in probability to the global minimum.

In addition, this implies that if the global minimum is unique, and so equals  $\tilde{f}^m \quad \forall m$ ,  $f^m$  must converge to it.

□

## Proof of Theorem 5

*Proof.* For simplicity, consider only the univariate case. Assume further for simplicity of notation, by permuting the observations if necessary, that the covariate is sorted in that  $X_1 < \dots < X_n$ , for  $i = 1, \dots, n-1$ .

Let  $\mathcal{G}$  be the set of pairs of monotonically increasing and monotonically decreasing functions  $g, h$ , with mean zero, with the constraint that in each interval at most one function of the two changes. Hence, with ordered covariate observations, either  $g(X_{i+1}) = g(X_i)$  or  $h(X_{i+1}) = h(X_i)$ .

Observe that for any right continuous mean zero step function  $f$ , any pair of monotonically increasing and monotonically decreasing mean zero functions  $g, h$  satisfying  $g + h \equiv f$  can only minimise  $\Delta g + \Delta h$  among such functions if  $(g, h) \in \mathcal{G}$ . Otherwise, if the constraint is broken for  $i$ , defining

$$\begin{aligned}\tilde{g}(x) &= g(x) - \min(g(X_{i+1}) - g(X_i), h(X_i) - h(X_{i+1})) I_{\{x > X_{i+1}\}} \\ \tilde{h}(x) &= h(x) + \min(g(X_{i+1}) - g(X_i), h(X_i) - h(X_{i+1})) I_{\{x > X_{i+1}\}},\end{aligned}$$

gives  $\tilde{g} + \tilde{h} \equiv g + h$ , and  $\Delta(\tilde{g}) + \Delta(\tilde{h}) = \Delta(g) + \Delta(h) - 2 \min(g(X_{i+1}) - g(X_i), h(X_i) - h(X_{i+1}))$ .

Therefore solutions to (4.2) must lie within  $\mathcal{G}$ .

Now, it is trivial to see that the decomposition in Definition 4.1 maps right continuous step functions with mean zero to pairs in  $\mathcal{G}$ . Two such step functions have the same decomposition if and only if they are equal at all knot points, and so, are for our purposes equivalent. Summation is an inverse with these spaces, since any such step function can be constructed as the sum of its decomposed functions, and we can produce any element in  $\mathcal{G}$  by from a right continuous mean zero step function by decomposing its sum. Thus, Definition 4.1 gives us a one to one map between the feasible set of (4.1) and a set that contains all solutions of (4.2).

Let  $f$  be any right continuous step function with mean zero, and let  $(f^+, f^-) \in \mathcal{G}$  be its decomposition. By construction,

$$\begin{aligned}\Delta(f^+) + \Delta(f^-) &= \sum_{i=1}^{n-1} (f(X_{i+1}) - f(X_i)) I_{\{f(X_{i+1}) - f(X_i) > 0\}} + (f(X_i) - f(X_{i+1})) I_{\{f(X_{i+1}) - f(X_i) < 0\}} \\ &= \sum_{i=1}^{n-1} |f(X_{i+1}) - f(X_i)| = \Delta(f).\end{aligned}$$

Hence,

$$\begin{aligned}M_\lambda(f^+, f^-) &= \frac{1}{2} \|Y - f^+(X) - f^-(X)\|^2 + \lambda(\Delta(f^+) + \Delta(f^-)) \\ &= \frac{1}{2} \|Y - f(X)\|^2 + \lambda\Delta(f) = L_\lambda(f).\end{aligned}$$

Therefore, minimising  $L_\lambda$  means the corresponding decomposed functions must minimise  $M_\lambda$ , and vice versa. Hence, the decomposition/summation transformations give a one to one map between solutions to (4.2) and (2.1).

The case of multiple covariates can be dealt with by applying the some argument to each covariate in turn.  $\square$

## References

- Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26:641–647.
- Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association*, 84:289–294.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions: the theory and application of isotonic regression*. Wiley.
- Bowman, A. W., Jones, M. C., and Gijbels, I. (1998). Testing monotonicity of regression. *Journal of Computational and Graphical Statistics*, 7:489–500.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
- Donoho, D. (2006). For most large underdetermined systems of linear equations the minimal  $l_1$ -norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59:797–829.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.
- Friedman, J. (1991). Multiple adaptive regression splines. *Annals of Statistics*, 19:1–141.
- Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 1:302–332.
- Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.
- Hastie, T., Tibshirani, R., and Friedman, J. (2003). *The Elements of Statistical Learning*. Springer.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, chapter 4, pages 82–95. Monographs on statistics and applied probability. Chapman and Hall.
- Huang, J., Horowitz, J. L., and Wei, F. (2009). Variable selection in nonparametric additive models. Technical report, The University of Iowa.
- Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16:1273–1284.
- Mammen, E. and van de Geer, S. (1997). Locally adaptive regression splines. *Annals of Statistics*, 25:387–413.
- Mammen, E. and Yu, K. (2007). Additive isotone regression. In *Asymptotics: Particles, Processes and Inverse Problems*, volume 55 of *IMS Lecture Notes - Monograph Series*, pages 179–195.

- Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821.
- Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462.
- Osborne, M. R., Presnell, B., and Turlach, B. A. (1998). Knot selection for regression splines via the lasso. In Weisberg, S., editor, *Dimension Reduction, Computational Complexity, and Information*, Proceedings of the 30th Symposium on the Interface, Interface 98, pages 44–49. Interface Foundation of North America.
- Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2007). Spam: Sparse additive models. *Advances in Neural Information Processing Systems (NIPS)*.
- Raviv, Y. and Intrator, N. (1996). Bootstrapping with noise: An effective regularization technique. In *Connection Science*, volume 8.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288.
- Tibshirani, R., Sanders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B*, 67:91–108.
- Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475–494.
- Wainwright, M. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. Technical report, UC Berkley.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68:49–67.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.
- Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood methods. *Annals of Statistics*, 36:1509–1533.