Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

LASSO Isotone for High-Dimensional Additive Isotonic Regression

Zhou FANG and Nicolai MEINSHAUSEN

Additive isotonic regression attempts to determine the relationship between a multidimensional 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 modeling to additive isotonic regression. Thus, it is viable in many situations with high-dimensional predictor variables, where selection of significant versus insignificant variables is 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.

Supplemental materials are available online for this article.

Key Words: Additive modelling; LISO; Monotonic regression; Nonparametric regression.

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

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 modeling, as described for instance by Hastie and Tibshirani (1990), is well known to be an useful generalization.

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

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

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

Zhou Fang is Phd Candidate (E-mail: *fang@stats.ox.ac.uk*) and Nicolai Meinshausen is Lecturer (E-mail: *meinshausen@stats.ox.ac.uk*), Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom.

© 2012 *American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America*
Journal of Computational and Graphical Statistics, Volume 21, Number 1, Pages 72–91
DOI: 10.1198/jcgs.2011.10095

where μ is a constant intercept term, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is a random error term, assumed independent of the covariates and identically distributed with mean zero. For every covariate $k = 1, \ldots, p$, 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 minimize the L2 norm of the error,

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

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

Work on such methods of additive modeling have produced a profuse array of techniques and generalizations. 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 prespecified monotonicity with respect to each covariate. 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 the book by Barlow et al. (1972). It turns out that in the univariate case, the pool adjacent violators algorithm (PAVA), as first suggested by 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 maximization of the likelihood, means that apart from the monotonicity constraint, we do not put on any regularization 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 modeling, 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 radically 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. Then, if the identity of the significant variables could be found, the CPAV could be conducted on this 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.

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

In the context of parametric linear regression, it has emerged recently that such sparse regression problems can be dealt with by use of an L1-norm based penalty in the optimization. This can solve the identifiability problem and achieve good predictive accuracy. Tibshirani (1996), Donoho (2006), among others, have identified several significant empirical and theoretical results to support this 'LASSO' estimator, while Osborne, Presnell, and Turlach (2000), Friedman et al. (2007), and others have invented fast algorithms for calculating both individual estimates and full LASSO solution paths.

Generalization of the L1 penalization principle to nonparametric regression can also lead to success with additive modeling (Avalos, Grandvalet, and Ambroise 2003). For example, recent work on this subject includes SpAM (Ravikumar et al. 2007), which describes the application of the group LASSO to general smoothers, and high-dimensional additive modeling with smoothness penalties (Meier, van de Geer, and Bühlmann 2009), which follows similar principles, using a spline basis.

In this article, 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. This article thus builds on the work of Ravikumar et al. (2007) and Bacchetti (1989).

The LISO is similar to the degree 0 case of the LASSO knot selection of Osborne, Presnell, and Turlach (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 optimization, and in Section 3 we will discuss algorithms for computation for fairly large n and p. We will discuss the effect of the regularization, and then in Section 4 suggest an important extension. Finally, in Section 5 we will explore its performance using some simulation studies. Proofs of theorems are left for the Appendix as part of online supplementary material.

2. THE LASSO-ISOTONE OPTIMIZATION

The term 'isotonic' means the functions are assumed to be increasing. However, an assumption of decreasing functions can be accommodated easily for any estimator by applying the algorithm using reversed sign observed covariates. The monotonically increasing function $\tilde{f}$ thus found can then be transformed to be a decreasing function estimate in the original covariates by

$$\widehat{f}(x) = \widetilde{f}(-x).$$

Hence, 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. $X = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^{n \times p}$ is the matrix of covariates.

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

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

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 1: We define the LASSO-Isotone (LISO) estimator for a particular value of tuning parameter $\lambda \ge 0$ as $\widehat{f}_\lambda(x) = \widehat{\mu} + \sum_{k=1}^{p} \widehat{f}_{k,\lambda}(x)$, where $\widehat{\mu}$, a constant, and $\widehat{f}_{k,\lambda} \in \mathcal{F}_k$ $\forall k$ together minimize the LISO loss

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

Here $\|\cdot\|$ denotes the empirical L2 norm, while $\Delta(f_k)$ denotes the total variation of f_k, which for $f_k \in \mathcal{F}_k$ can be calculated as

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

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. Under this constraint, it follows trivially that $\widehat{\mu}$ must equal the sample mean of the response, and hence (2.1) reduces to the case of optimizing over $f_1, \ldots, f_p$ with $\overline{Y} = 0$.

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 neighborhood of the origin, both terms in the objective are increasing, so a bounded solution exists for all values of λ. 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 minimizer to L_λ, another fit with the same function values at observed covariate points, interpolating monotonically between them, will have the same value of L_λ, and so also be a LISO solution. This means that we can equivalently consider optimization in the finite-dimensional space of fitted values $\widehat{f}_k(X^{(k)})$.

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

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 article, we shall consider uniqueness and equivalence in terms of having equal values at the observed $X^{(k)}$.

The total variation penalty shown here has been previously suggested for regression in the article by 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 optimization for a given dataset can be viewed as ordinary LASSO optimization for a linear model, constrained to positive coefficients, using an expanded design matrix $\widetilde{X} \in \mathbb{R}^{n \times p(n-1)}$, where $\widetilde{X} = (\widetilde{X}^{(1)} \ \cdots \ \widetilde{X}^{(p)})$. Each $\widetilde{X}^{(k)} \in \mathbb{R}^{n \times (n-1)}, k = 1, \ldots, p$, contains $n-1$ step functions in the kth covariate, which form a basis for the vector $f_k(X^{(k)})$, and so isotonic functions in that covariate. The coefficients β optimized over then represent step sizes.

Such a construction was suggested by Osborne, Presnell, and Turlach (1998), among others. Under this reparameterization of the problem, existing LASSO algorithms for linear regression may be applied, with a modification to restrict solutions to nonnegative values. In particular, the least angle regression algorithm of Osborne, Presnell, and Turlach (2000) and Efron et al. (2004) is effective, since short cuts exist for calculating the necessary correlations.

On the other hand, the high dimensionality of $\widetilde{X}$ 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 specialized 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 optimization is equivalent to the standard univariate isotonic regression problem. In this case, the log-likelihood residual sum of squares term is strictly convex, and so, as a strictly convex optimization on a convex set, a unique solution exists. Trivially, the solution must also be bounded. In fact, there exists, as described by 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 $\widehat{f_\lambda}$ as the solution to optimization (2.1) for λ, we have $\widehat{f_0} \equiv \widehat{f}_{\text{PAVA}}$. The following theorem describes the solutions for other values of λ:

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

Theorem 1. *Given* $\widehat{f}_{\text{PAVA}}$, *the LISO solution for* $\lambda \geq 0$, $p = 1$ *is the Winsorized PAVA fit*,

$$\widehat{f}_{>A_\lambda, <B_\lambda} := \max(\min(\widehat{f}_{\text{PAVA}}, B_\lambda), A_\lambda),$$

with A_λ, B_λ *thresholding constants that are piecewise linear, continuous, and monotone* (*increasing for* A_λ, *decreasing for* B_λ) *functions of* λ.

Specifically, if

$$2\lambda \geq \sum_{i=1}^{n} |\widehat{f}_{\text{PAVA}}(X_i) - \overline{Y}|, \tag{3.1}$$

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

Otherwise, A_λ, B_λ *are the solutions to*

$$\sum_{i=1}^{n} (\widehat{f}_{\text{PAVA}}(X_i) - B_\lambda)_+ = \lambda, \tag{3.2}$$

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

Corollary 1. *Let* π *be a permutation taking* $1, \ldots, n$ *to indices that put the covariate in ascending order. Then if*

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

we have that $\widehat{f}_\lambda \equiv \overline{Y}$.

Remark 1: 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. For (3.3) and (3.2), then, 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 optimization in higher dimensions, due to correlations between steps in different covariates. 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, see Algorithm 1.

Step 8 is performed to avoid accumulation of calculation errors, but may be avoided to reduce computation time.

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

Algorithm 1 LISO-Backfitting.

Take $\widehat{\mu} = \overline{Y}$, $Y \Leftarrow Y - \widehat{\mu}$.
Set $m = 0$.
Initialize component fits $(f_1, \ldots, f_p)$ as identically 0, or as the estimate for a different value of λ, storing these as the $n \times p$ marginal fitted values.
repeat
  $f^m \Leftarrow (f_1, \ldots, f_p)$.
  $m \Leftarrow m + 1$.
  for $k = 1$ to p (or a random permutation) **do**
    Recalculate residuals $r_i \Leftarrow Y_i - \sum_{k=1}^{p} f_k(X_i^{(k)})$, $i = 1, \ldots, n$.
    Refit conditional residual $\{r_i + f_k(X_i^{(k)})\}_{i=1}^n$ using $X^{(k)}$ by PAVA, producing $\widetilde{f}_k(X_i^{(k)})$, for $i = 1, \ldots, n$.
    Calculate thresholds A_λ, B_λ from λ and $\widetilde{f}_k$ by Theorem 1.
    Adjust component fit $f_k(X_i^{(k)}) \Leftarrow \widetilde{f}_{k,>A_\lambda<B_\lambda}(X_i^{(k)})$.
  end for
until sufficient convergence is achieved, through considering f^m and f^{m-1}.
Interpolate f_k between the samples $X_i^{(k)}$ and construct $\widehat{f}_\lambda$.

Theorem 2. *For $f^m = (\mu, f_1^m, \ldots, f_p^m)$, 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 2: A proof of this is available via a theorem in the article by Tseng (2001). However, a simpler alternative proof is available in the case of random permutations, and for completeness is given in the Appendix.

Remark 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 4: Moreover, defining $X_{(i)}^{(k)}$ as the ith smallest value of $X^{(k)}$, if a certain individual step in the final functional fit

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

has a value of zero in all solutions to the LISO minimization, 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_λ 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

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

PAVA calculation involving the covariate associated with that step if that single covariate optimization 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 optimization, the LISO loss will always decrease in each step, and we will converge toward 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, and monitor the change in the results in each cycle, stopping when this becomes small.

3.3 CHOICE OF REGULARIZATION PARAMETER

It will be always necessary to choose a tuning parameter λ to facilitate appropriate fitting. As with the LASSO, too high a tuning parameter will shrink the fits toward zero. Indeed, consideration of Corollary 1 shows that, with $\overline{Y} = 0$, and $\pi^{(k)}$ defined as a permutation that puts the kth covariate into ascending order, a choice of λ greater than

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

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 λ will also lead to improper fitting. This arises from two sources. First, 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 $\widetilde{X}$. Second, λ has a role in terms of fit complexity, with a small value of λ 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 $\widetilde{X}$ to be correlated with columns belonging to irrelevant covariates, hence producing spurious fits in the other covariates.

These effects are illustrated in Figure 1, in which we have generated X, with $n = 100$, $p = 200$, according to a uniform distribution, and produced Y as the sum of $k = 5$ of the

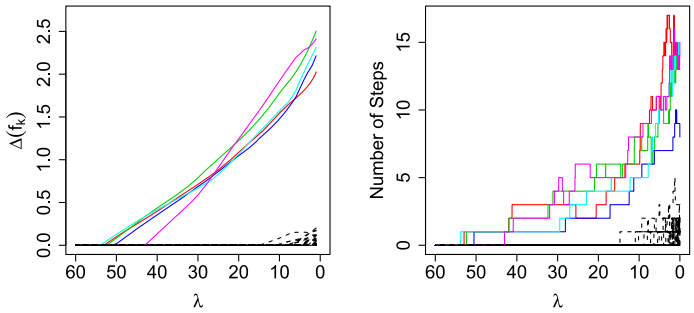

Figure 1. Effects of changing the regularization parameter in the noiseless case. $n = 100$, $p = 200$. Each line represents how an individual covariate's estimate changes as λ varies, with the solid lines for the true covariates, while the dashed lines denote spurious fits on irrelevant variables. The online version of this figure is in color.

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

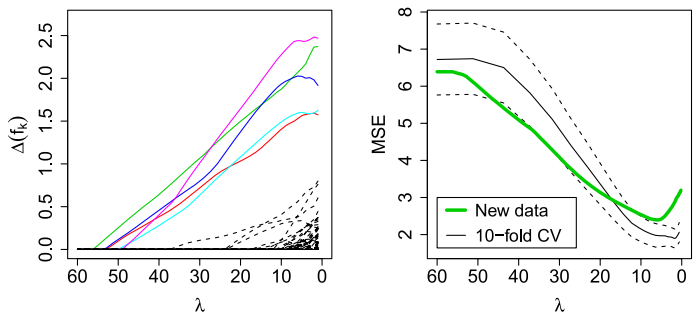

Figure 2. Effects of changing the regularization parameter in the noisy case. $n = 100$, $p = 200$, SNR $= 5$. We show again in the first graph the total variation of each covariate estimate as λ alters, with solid lines for the truly important covariates, while the dashed lines denote spurious fits on irrelevant variables. The second graph shows the MSE from a 10-fold cross-validation procedure with ± 1 s.d. in dashes, as well as the true MSE on a new set of data as the thick line. The online version of this figure is in color.

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

covariates. In other words, f is the sparse sum of linear functions. We give the full paths of fits in terms of, first, the total variation of fitted components $\Delta(f_k)$, and second, the number of component steps in each covariate,

$$\bigl|\bigl\{i : f_k\bigl(X^{(k)}_{(i)}\bigr) \neq f_k\bigl(X^{(k)}_{(i-1)}\bigr)\bigr\}\bigr|.$$

Of particular note is that, unlike the LASSO, even without noise, the size of the basis of step functions and the nonsparsity of the true signal mean that as $\lambda \to 0$, we do not converge to the true sparsity pattern. However, with higher λ, 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 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 λ. 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 λ directly from the data. To do this, with the goal of 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) onward, cross-validation is effective for minimizing predictive error, and is illustrated by the second graph of Figure 2. Here, we calculate CV error from a 10-fold cross-validation. We may then take the λ that minimizes the average mean squared error across the folds. If we desire a simpler model, we can, as is often suggested, take the largest λ 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 minimizing predictive error, however, we do still fit some irrelevant covariates as nonzero, a phenomenon previously observed with the LASSO in the article by Leng, Lin, and Wahba (2006).

Now, unlike a LARS-like approach, LISO backfitting will only give us the solution for an individual choice of λ. However, CV can still be practical, because coordinate-wise minimization 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 λ are likely to be similar, and hence use the result for one value of λ as a start point for the calculation for a nearby value of tuning parameter. This is especially effective if we order the λ values we need to calculate in decreasing order, since large λ solutions are more sparse and so faster to calculate.

4. ADAPTIVE LISO

A variety of extensions and variations of the basic LISO procedure may be proposed, that may offer improvements in some circumstances. For instance, 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. However, this method is not reliably a great improvement and will almost inevitably reduce the degree of sparsity in the fit, for any given degree of regularization.

More valuable is to observe that 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 optimization, 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 recognize the natural grouping of steps in the step function basis.

Many existing solutions to this issue, such as that of Huang, Horowitz, and Wei (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 optimization, arriving at an initial fit $(\mu, f_1^0, \ldots, f_p^0)$. Then, we conduct a second LISO procedure, this time introducing covariate weights $w_1, \ldots, w_p$ 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 $k = 1, \ldots, p$. See Algorithm 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 optimization because large steps contribute to relaxation of other steps in the same covariate. In addition, it means that we in general require less regularization of true fits in order to shrink irrelevant covariates to zero, through the concavity of the implied overall optimization, to which we are essentially calculating a Local Linear Approximation (Zou and Li 2008).

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

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 $k = 1, \ldots, p$.
3: Calculate, using, for example, Algorithm 1,

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

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

We will also always enhance sparsity through this procedure; indeed, the fact that we reject straightaway 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 the article by Zou and Li (2008) have recommended a scheme based on individual prediction error minimizing 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.

Remark 5: An additional application for the Adaptive LISO is in the case where the direction or the presence of monotonicity for the model function component in each covariate is not known. One possible heuristic of dealing with this situation 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, it is a well-known fact (Itô 1993) that functions of bounded variation have a unique Jordan decomposition

$$f \equiv f^+ + f^-$$

into monotonically increasing and decreasing functions, such that $\Delta(f) = \Delta(f^+) + \Delta(f^-)$. It follows that the monotonicity-relaxed form of the LISO, the total variation penalized estimate, can be dealt with by Algorithm 1 by including as 'covariates' both the original covariates and sign-reversed versions of themselves. The original covariate is used to estimate f^+ while the reversed covariate finds f^-. These can then be combined to give an estimate. In this scheme, LISO can be thought of as setting the penalty on the decreasing component to infinity, and that on the increasing component to a finite quantity. For a related approach, see the work of Tibshirani, Hoefling, and Tibshirani (2010).

A possible variation to the LISO-Adaptive scheme above, then, is to conduct first a non-monotonic total variation penalized fit, consider the fit in terms of its increasing and

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

decreasing components, and then compute separate weights to be placed on the increasing and decreasing components in a second-stage fit. We see in this case the same effect seen in the adaptive LISO, where we have strengthened shrinkage of small function fits toward zero compared to the total variation penalized fit. However, in addition, fits found to be monotonic in the first stage will remain monotonic in the second stage, while functions with small negative or positive components in the initial fit will be shrunk toward a 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 (Osborne, Presnell, and Turlach 2000), 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 by Harrison and Rubinfeld (1978), is a dataset often used in the literature to test estimators; see, for example, Hastie, Tibshirani, and Friedman (2003). The dataset comprises $n = 506$ 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 by 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 Remark 5, 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

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

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

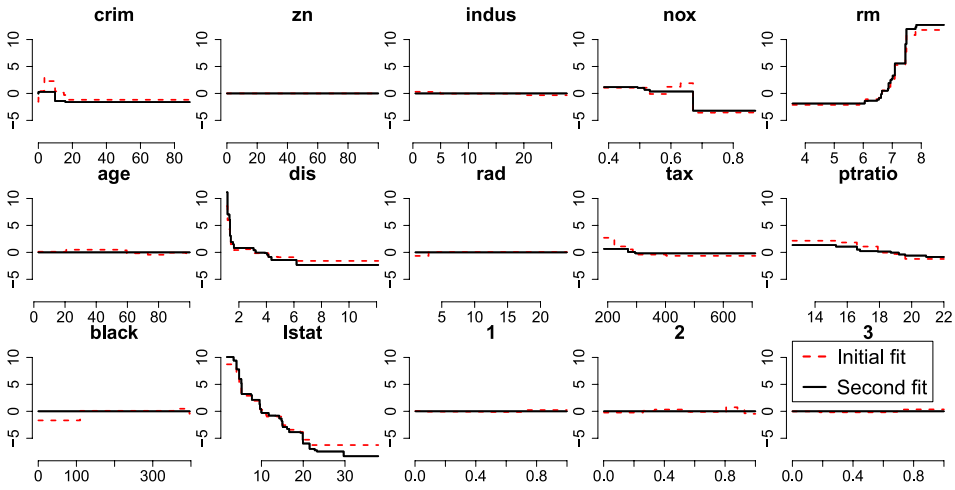

Figure 3. Fitted component functions on the Boston Housing dataset, for covariates originally present in the data plus three others. The dashed line shows the selected model after the first LISO step, while the solid black line shows the final result of the adaptive sign finding procedure. The single-step fit produced additional nonzero fits in some of the artificial covariates, which are not shown, while the two-step procedure fit all of them as zero. The online version of this figure is in color.

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 slightly non-monotonic, and the remaining functions monotonically decreasing. The full results are shown in Figure 3.

We see in our experiments that for higher values of λ, 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 criterion, we choose a λ 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—observe especially the case for `nox`.

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 monotonic, aside from a small step fit near 0 in `crim`.

The non-monotonicity found in `crim` may seem problematic, given the interpretation of that covariate as a crime rate. However, the small increasing step found near $\texttt{crim} = 0$ might be reasonable if such areas are qualitatively different from others. On the other hand, perhaps it would be reasonable to directly impose a monotonicity constraint instead.

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

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

$$X_{ij} \sim \text{Uniform}(-1, 1),$$
$$Y_i = 2\big(X_i^{(1)}\big)_+^2 + X_i^{(2)} + \text{sign}\big(X^{(3)}\big)\big|X_i^{(3)}\big|^{1/5} + 2I_{\{X_i^{(4)}>0\}} + \varepsilon_i,$$

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

For $p' = 32, 64, 128, 256, 512, 1024$, $n' = 5, 10, 15, \ldots$, 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 four true variables, from among 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 λ, 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 4 gives these results. As we can see, as in a variety of LASSO-type algorithms (Wainwright 2009), 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 gives an example of LISO fits arising from this simulation. As can be seen from the marginal plot of the data points, with such a small amount of data, it can be very difficult to find the true model function. The dashed lines show the results of the LISO under the minimum regularization required for correct sparsity recovery—note the high level of shrinkage required to shrink the other variables to zero. This shrinkage exhibits itself as not only a thresholding on the ends of the component fits, which we have seen in

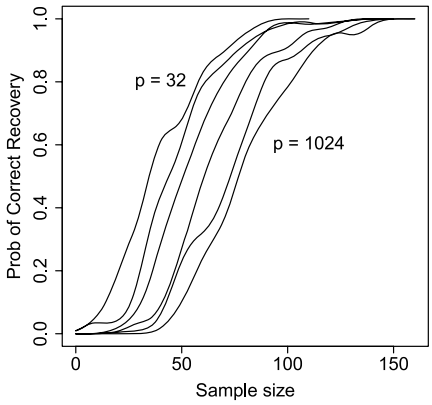

Figure 4. Probabilities of correct sparsity recovery with four true nonlinear but monotonic covariates, SNR = 4. Each line shows how the recovery probability changes as the sample size n changes for a single value of p, taking values $2^5, \ldots, 2^{10}$.

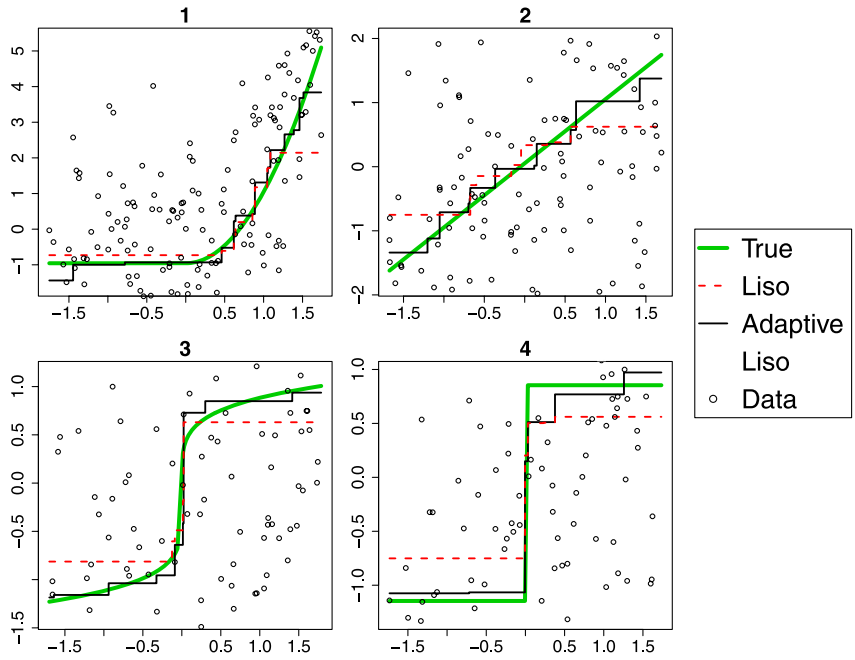

Figure 5. Example LISO covariate fits, for $n = 180$, $p = 1024$. The true component functions are given by the thick line, while the dashed line gives the raw LISO fit for the smallest amount of regularization required to bring spurious fits in irrelevant covariates to zero. The solid black line shows 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. The circles are marginal plots of the data points. The online version of this figure is in color.

the univariate case, but also an additional loss of complexity in the middle parts of each component fit. The Adaptive LISO, as the solid black line, avoids this shrinkage and thus greatly improves the fit while still keeping the correct sparsity pattern recovery. As an added bonus, we get good sparsity recovery results here with the Adaptive LISO using merely the cross-validated tuning parameter values.

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,

$$X_i^{(j)} \sim \text{Uniform}(-1, 1), \qquad i = 1, \ldots, n,\ j = 1, \ldots, p,$$
$$\varepsilon_i \sim N(0, 1), \qquad i = 1, \ldots, n,$$
$$Y_i = f(X_i) + \sigma\varepsilon_i, \qquad i = 1, \ldots, n.$$

One hundred repetitions were done of each combination of model and noise level, with σ 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 $\Sigma_{ij} = 2^{-|i-j|}$.

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

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 hockey-stick-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, van de Geer, and Bühlmann (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 minimizes 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, optimizing 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 kth algorithm on each individual run as

$$\mathrm{MSE}^k_{\mathrm{Relative}} := \frac{\mathrm{MSE}^k}{\min_{j=1,\ldots,7} \mathrm{MSE}^j}.$$

We show in bold text the best performing estimator in each scenario.

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. See Table 1.

Table 1. Test set MSE (and relative MSE) for linear scenario.

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 seen 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.

Among 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 among 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 regularization 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:

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

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

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

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

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

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

In this case, we have again $n = 200$, $p = 50$. $C_1, \ldots, C_5$ are small shifts, randomly generated as Uniform$(-1/4, 1/4)$, and $a_1, \ldots, a_5$ are covariates randomly chosen without replacement. In the test set, the variance of the response was approximately 2.6. See Table 2.

With the new, nonlinear 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

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

Table 2. Test set MSE (and relative MSE) for mixed powers scenario.

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)

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. See Table 3.

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 nonparametric estimation of isotonic functions. We have found that in

Table 3. Test set MSE (and relative MSE) for large p mixed powers scenario.

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)

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

many contexts, it inherits the behavior 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 a 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 nonparametric LASSO generalizations.

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012

SUPPLEMENTARY MATERIALS

R-package for LISO: R-package LISO containing code to perform the methods described in the article. (liso_0.2.tar.gz, GNU zipped tar file)

Proofs of theorems: Pdf document containing appendix with proofs of theorems in the article. (lisoappendix.pdf, PDF document)

ACKNOWLEDGMENTS

Zhou Fang acknowledges support from the EPSRC and Nicolai Meinshausen from the Leverhulme Trust. Both thank the editor and the anonymous referees and the associate editor for very helpful comments that helped improve an earlier version of the manuscript.

[Received May 2010. Revised December 2010.]

REFERENCES

Avalos, M., Grandvalet, Y., and Ambroise, C. (2003), "Regularization Methods for Additive Models," in *Proceedings of the 5th International Symposium on Intelligent Data Analysis*, Berlin: Springer, pp. 509–520. [74]

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. [73,76]

Bacchetti, P. (1989), "Additive Isotonic Models," *Journal of the American Statistical Association*, 84, 289–294. [73,74]

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*, Hoboken, NJ: Wiley. [73,76]

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. [81]

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32. [87]

Donoho, D. (2006), "For Most Large Underdetermined Systems of Linear Equations the Minimal l1-Norm Solution Is Also the Sparsest Solution," *Communications on Pure and Applied Mathematics*, 59, 797–829. [74]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [76]

Friedman, J. (1991), "Multiple Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–141. [87]

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332. [74,81]

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. [83]

Hastie, T. J., and Tibshirani, R. J. (1990), "Additive Models," in *Generalized Additive Models. Monographs on Statistics and Applied Probability*, London: Chapman & Hall, chapter 4, pp. 82–95. [72]

Hastie, T., Tibshirani, R., and Friedman, J. (2003), *The Elements of Statistical Learning*, Berlin: Springer. [83]

Huang, J., Horowitz, J. L., and Wei, F. (2009), "Variable Selection in Nonparametric Additive Models," technical report, The University of Iowa. [81]

Itô, K. (1993), "Functions of Bounded Variation," in *Encyclopedic Dictionary of Mathematics*, Cambridge, MA: MIT Press, chapter 166, pp. 642–643. [82]

Leng, C., Lin, Y., and Wahba, G. (2006), "A Note on the Lasso and Related Procedures in Model Selection," *Statistica Sinica*, 16, 1273–1284. [80]

Mammen, E., and van de Geer, S. (1997), "Locally Adaptive Regression Splines," *The Annals of Statistics*, 25, 387–413. [74,76]

Mammen, E., and Yu, K. (2007), "Additive Isotone Regression," in *Asymptotics: Particles, Processes and Inverse Problems. IMS Lecture Notes—Monograph Series*, Vol. 55, Beachwood, OH: IMS, pp. 179–195. [73]

Meier, L., van de Geer, S., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [74,87]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [80]

Osborne, M. R., Presnell, B., and Turlach, B. A. (1998), "Knot Selection for Regression Splines via the Lasso," in *Dimension Reduction, Computational Complexity, and Information, Proceedings of the 30th Symposium on the Interface, Interface 98*, ed. S. Weisberg, Fairfax Station, VA: Interface Foundation of North America, pp. 44–49. [74,76]

——— (2000), "A New Approach to Variable Selection in Least Squares Problems," *IMA Journal of Numerical Analysis*, 20, 389–404. [74,76,83]

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2007), "Spam: Sparse Additive Models," in *Advances in Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press. [74,83,84,87]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [74,80]

Tibshirani, R. J., Hoefling, H., and Tibshirani, R. (2010), "Nearly-Isotonic Regression," unpublished manuscript, Stanford University, Palo Alto, CA. [82]

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, Ser. B*, 67, 91–108. [74]

Tseng, P. (2001), "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of Optimization Theory and Applications*, 109, 475–494. [78]

Wainwright, M. (2009), "Sharp Thresholds for High-Dimensional and Noisy Recovery of Sparsity," *IEEE Transactions on Information Theory*, 55, 2183–2201. [85]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society, Ser. B*, 68, 49–67. [81]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [81]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Methods," *The Annals of Statistics*, 36, 1509–1533. [81,82]

Downloaded by [Nicolai Meinshausen] at 01:57 26 April 2012