Nearly-Isotonic Regression

Ryan J. Tibshirani[*] Holger Hoefling[†] Robert Tibshirani[‡]

Abstract

We consider the problem of approximating a sequence of data points with a "nearly-isotonic", or nearly-monotone function. This is formulated as a convex optimization problem that yields a family of solutions, with one extreme member being the standard isotonic regression fit. We devise a simple algorithm to solve for the path of solutions, which can be viewed as a modified version of the well-known pool adjacent violators algorithm, and computes the entire path in $O(n \log n)$ operations, (n being the number of data points). In practice, the intermediate fits can be used to examine the assumption of monotonicity. Nearly-isotonic regression admits a nice property in terms of its degrees of freedom: at any point along the path, the number of joined pieces in the solution is an unbiased estimate of its degrees of freedom. We also extend the ideas to provide "nearly-convex" approximations.

Key words: *isotonic regression*, *pool adjacent violators*, *path algorithm*, *degrees of freedom*

1 Introduction

Isotonic regression solves the following problem: given a sequence of n data points $y_1, \ldots y_n$, how can we best summarize this by a monotone sequence $\beta_1, \ldots \beta_n$? Formally, the problem is to find

$$\hat{\beta}^{\text{iso}} = \underset{\beta \in \mathbb{R}^n}{\text{argmin}} \sum_{i=1}^{n} (y_i - \beta_i)^2 \text{ subject to } \beta_1 \leq \ldots \leq \beta_n. \tag{1}$$

Here and throughout we assume that "monotone" means monotone nondecreasing: an analogous problem statement exists for monotone nonincreasing.

A unique solution to problem (1) exists and can be obtained using the pool adjacent violators algorithm (PAVA) (Barlow et al. 1972). Roughly speaking, PAVA works as follows. We start with y_1 on the left. We move to the right until we encounter the first violation $y_i > y_{i+1}$. Then we replace this pair by their average, and back-average to the left as needed, to get monotonicity. We continue this process to the right, until finally we reach y_n. If skillfully implemented, PAVA has a computational complexity of $O(n)$ (Grotzinger & Witzgall 1984).

There is quite a large literature on isotonic regression. The aforementioned book (Barlow et al. 1972) is a classic reference, along with Robertson et al. (1988), although there have been many earlier references to the same idea (Brunk 1955, Ayer et al. 1955, Miles 1959, Bartholomew 1959*a*, Bartholomew 1959*b*). A recent paper by de Leeuw et al. (2009) gives a nice overview of the problem's history and computational aspects. Luss et al. (2010) propose a recursive partitioning algorithm that produces a series of less smooth isotonic fits, culminating in the usual isotonic regression.

In this paper we consider a different problem, that of approximating the data with a "nearly-isotonic" function. We formulate a convex optimization problem that yields a family of solutions, with the saturated fit at one end ($\hat{\beta} = y$) and the isotonic regression fit ($\hat{\beta} = \hat{\beta}^{\text{iso}}$) at the other. The

[*]Dept. of Statistics, Stanford University, ryantibs@stanford.edu
[†]Molecular Diagnostics, Novartis, hhoeflin@gmail.com
[‡]Depts. of Health, Research & Policy, and Statistics, Stanford University, tibs@stanford.edu

entire path of solutions can be computed in $O(n \log n)$ operations, using a simple algorithm that can be thought of as a modified version of PAVA. Furthermore, we show that at any point on the path, an unbiased estimate of the solution's degrees of freedom is achieved by simply counting the number of joined pieces.

2 Nearly-isotonic fitting

Given a sequence $y_1, \dots y_n$, we seek a nearly-monotone approximation, and so we consider the problem

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda \sum_{i=1}^{n-1} (\beta_i - \beta_{i+1})_+, \tag{2}$$

with x_+ indicating the positive part, $x_+ = x \cdot 1(x > 0)$. This is a convex optimization problem for each fixed $\lambda \geq 0$. The penalty term penalizes adjacent pairs that violate the monotonicity property, that is, having $\beta_i > \beta_{i+1}$. When $\lambda = 0$, the solution interpolates the data, and letting $\lambda \to \infty$, we obtain the solution to the isotonic regression problem (1). For intermediate values of λ we get nonmonotone solutions which trade off monotonicity with goodness-of-fit.

We have implicitly assumed in (2) that the data points are measured along an equally spaced grid. If this is not the case, then it would make sense to change the penalty to

$$\lambda \sum_{i=1}^{n-1} \frac{(\beta_i - \beta_{i+1})_+}{x_{i+1} - x_i},$$

where x_i denotes the value where the measurement y_i is taken. Also, we note that the criterion in (2) is closely related to the criterion of the fused lasso signal approximator (FLSA) (Friedman et al. 2007), which is

$$\frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=1}^{n-1} |\beta_i - \beta_{i+1}|.$$

The first term above encourages sparsity, while the second term penalizes the jumps in adjacent coefficients. In nearly-isotonic regression, we omit the sparsity term and penalize only nonmonotonicities.

Figure 1 shows a toy example with $n = 7$ data points. The solution $\hat{\beta}_\lambda$ is shown for four values of λ, with the bottom right panel showing the full isotonic regression. In fact, these four values of λ represent the knots in the entire path of solutions, with each coordinate $\hat{\beta}_{\lambda,i}$ a linear function of λ in between these knots. The arrows indicate places where the adjacent pairs $\hat{\beta}_{\lambda,i}, \hat{\beta}_{\lambda,i+1}$ are joined and set equal to a common value. It is not hard to believe that, in general, if two adjacent coordinates are joined at some value of λ, then they remain joined for all larger values of λ. This is indeed true, and it leads to a simple algorithm to compute the path of solutions, discussed in the next section.

3 The path of solutions

In this section we derive an algorithm for computing the entire path of solutions for problem (2) as a function of λ. This algorithm is a kind of "modified" version of PAVA that starts not at the left end of the data, but rather joins adjacent points when needed, to produce the sequence of best near-isotonic approximations. In contrast, the sequence of intermediate fits in PAVA are not of special interest, only the final isotonic fit.

We begin by stating a lemma suggested at the end of the last section. Essentially, the lemma says that the pieces in the solution can only be joined together, not split apart, as λ increases. This fact is analogous to that for the FLSA (see Section 4 of Friedman et al. 2007), and the proof given in the Appendix, Section A.1.

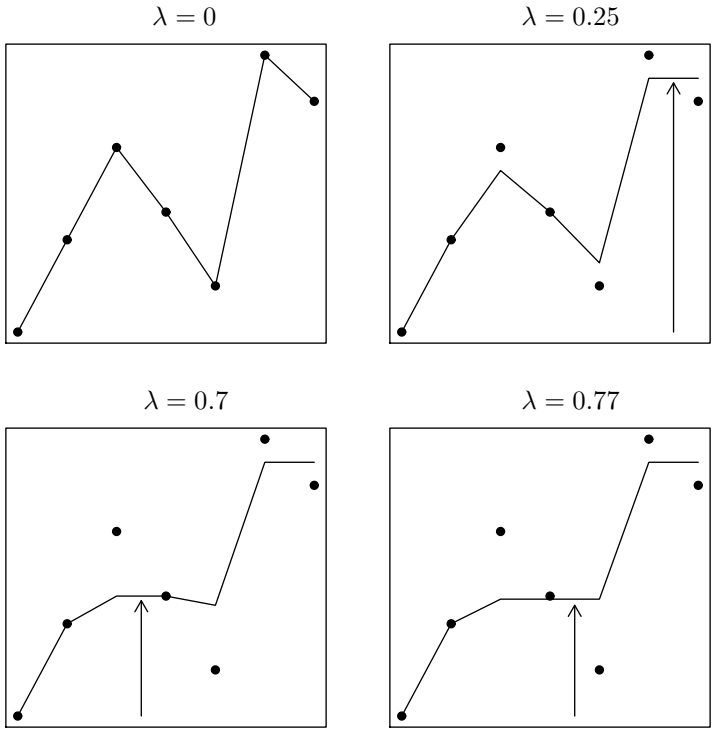

Figure 1: *Nearly-isotonic fits for toy example. An interpolating function is shown in the top left panel. There are three joining events (indicated by the arrows) shown in the remaining panels, with the usual isotonic regression appearing in the bottom right panel.*

Lemma 1. *Suppose that for some λ, we have two adjacent coordinates of the solution satisfying $\hat{\beta}_{\lambda,i} = \hat{\beta}_{\lambda,i+1}$. Then $\hat{\beta}_{\lambda_0,i} = \hat{\beta}_{\lambda_0,i+1}$ for all $\lambda_0 > \lambda$.*

This fact greatly simplifies the construction of the nearly-isotonic solution path, and our treatment is similar to that for the FLSA in Hoefling (2010). Suppose that at a parameter value λ, there are K_λ joined pieces, which we represent by groups of coordinates $A_1, \ldots A_{K_\lambda}$. Then we can rewrite the criterion in (2) as

$$\frac{1}{2}\sum_{i=1}^{K_\lambda}\sum_{j\in A_i}(y_j - \beta_{A_i})^2 + \lambda\sum_{i=1}^{K_\lambda - 1}(\beta_{A_i} - \beta_{A_{i+1}})_+. \tag{3}$$

To find the optimal levels $\hat{\beta}_{\lambda,A_i}$ of each group A_i, we examine the subgradient equations of (3):

$$-\sum_{j\in A_i} y_j + |A_i|\hat{\beta}_{\lambda,A_i} + \lambda(s_i - s_{i-1}) = 0 \quad \text{for } i = 1, \ldots K_\lambda. \tag{4}$$

Here $s_i = 1(\hat{\beta}_{\lambda,A_i} - \hat{\beta}_{\lambda,A_{i+1}} > 0)$, and for notational convenience, we let $s_0 = s_{K_\lambda} = 0$.

Now suppose that the groups $A_1, \ldots A_{K_\lambda}$ do not change for an interval of increasing λ. Then we can differentiate (4) with respect to λ, treating the groups A_i, and hence the values s_i, as constants. This yields

$$\frac{d\hat{\beta}_{\lambda,A_i}}{d\lambda} = \frac{s_{i-1} - s_i}{|A_i|}, \tag{5}$$

which is itself a constant, meaning that each $\hat{\beta}_{\lambda,A_i}$ is a (locally) linear function of λ.

As λ increases, equation (5) will continue to give the slope of the solution until the groups $A_1, \ldots A_{K_\lambda}$ change. According to Lemma 1, this can only happen when two groups merge, that is, their solution paths intersect. Using the slopes $m_i = d\hat{\beta}_{\lambda,A_i}/d\lambda$ derived in (5), and doing a little algebra, we find that groups A_i and A_{i+1} will merge at

$$t_{i,i+1} = \frac{\hat{\beta}_{\lambda,A_{i+1}} - \hat{\beta}_{\lambda,A_i}}{m_i - m_{i+1}} + \lambda, \tag{6}$$

for each $i = 1, \ldots K_\lambda - 1$. Therefore we can move all the way until the "next" value of λ

$$\lambda^* = \min_{i:\, t_{i,i+1} > \lambda} t_{i,i+1}, \tag{7}$$

and merge groups A_{i^*} and A_{i^*+1}, where

$$i^* = \operatorname*{argmin}_{i:\, t_{i,i+1} > \lambda} t_{i,i+1}. \tag{8}$$

Note that there may be more than one pair of groups that achieve this minimum, in which case we merge all of the appropriate groups. The minimizations (7) and (8) are only taken over the values of $t_{i,i+1}$ that are larger than λ. If none of the $t_{i,i+1}$ are larger than λ, then none of the existing groups will ever merge [in fact, the slopes in (5) are zero], so we know the solution for the rest of the path and the algorithm can terminate.

For the sake of clarity, we lay out the steps of the algorithm. The algorithm finds the set of critical points or knots $\{\lambda_1, \ldots \lambda_T\}$ at which the groups merge, and also computes the solution $\hat{\beta}_\lambda$ at each critical point.

Algorithm 1 (**Modified Pool Adjacent Violators**)**.**

- *Start with $\lambda = 0$, $K_\lambda = n$, and $A_i = \{i\}$ for each $i = 1, \ldots n$. The solution is $\hat{\beta}_{\lambda,i} = y_i$ for each i.*
- *Repeat:*
  1. *Compute the slopes m_i of each group, according to* (5)*.*
  2. *Compute the collision times $t_{i,i+1}$ of each pair of adjacent groups, according to* (6)*.*
  3. *If each $t_{i,i+1} \leq \lambda$, terminate.*
  4. *Compute the critical point λ^* as in* (7)*, and update the solution based on the slopes:*

$$\hat{\beta}_{\lambda^*,A_i} = \hat{\beta}_{\lambda,A_i} + m_i \cdot (\lambda^* - \lambda)$$

for each $i = 1, \ldots K_\lambda$. Merge the appropriate groups as in (8) *(so $K_{\lambda^*} = K_\lambda - 1$), and finally, set $\lambda = \lambda^*$.*

It is not difficult to verify that the path visited by the algorithm satisfies the subgradient equations (4) at each λ, and hence is indeed the solution path to nearly-isotonic regression (2). It is also not difficult to see that the total number of critical points (or the total number of iterations) is $T = n - K_\infty$, where K_∞ is the number of joined pieces in the full isotonic regression fit. Since the solution path is linear between the critical points, a practical implementation could, for example, store the solution at each of these points and then use linear interpolation to return the solution at any other λ. See Hoefling (2010) for a more efficient tree-based implementation; when applied to the current situation this gives a total running time of $O(T \log n) = O(n \log n)$.

Rewriting (4) gives us the following formula for the fitted value $\hat{\beta}_{\lambda,A_i}$:

$$\hat{\beta}_{\lambda,A_i} = \frac{\sum_{j\in A_i} y_j \;-\; \lambda(s_i - s_{i-1})}{|A_i|}, \tag{9}$$

where recall $s_i = 1(\hat{\beta}_{A_i} - \hat{\beta}_{A_{i+1}} > 0)$. Hence in any monotone increasing "stretch" of data points, the estimate is simply an average of y_i values, just like in PAVA. In general, there are two key differences between our approach and PAVA:

1. our algorithm chooses to join groups based on the tradeoff between goodness-of-fit with monotonicity, governed by the parameter λ, rather than sequentially enforcing monotonicity like PAVA;

2. our algorithm averages a shrunken version of the data points y_i, as in (9), so that nonmonotone pieces are pulled together (for example, see the top right panel in Figure 1).

4 Degrees of freedom

Suppose that the data vector y is drawn from the normal model

$$y \sim N(\mu, \sigma^2 I), \tag{10}$$

where I is the $n \times n$ identity matrix. The degrees of freedom of a fitted vector $\hat{y}$ of y can be defined as

$$\mathrm{df}(\hat{y}) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i). \tag{11}$$

This is discussed in Efron (1986) and Hastie et al. (2008), for example. For isotonic regression, it turns out that the number of joined pieces in $\hat{\beta}^{\mathrm{iso}}$ is an unbiased estimate of this quantity. This follows from Proposition 1 of Meyer & Woodroofe (2000), and was pointed out in Luss et al. (2010).

Fortuitously, this same property holds for the entire nearly-isotonic path. In particular, if K_λ is the number of joined pieces in $\hat{\beta}_\lambda$, the solution of (2) at λ, then

$$\mathrm{df}(\hat{\beta}_\lambda) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Cov}(\hat{\beta}_{\lambda,i}, y_i) = \mathrm{E}(K_\lambda). \tag{12}$$

A proof of this appears in the Appendix, Section A.2.

One might ask: why is this property useful? Degrees is freedom is a measure of model complexity: it describes the effective number of parameters that are used in the fit. In this sense, it is reassuring to know that the number of parameters used by nearly-isotonic regression is not drastically more than the number of joined pieces in the fit, but is actually equal to the number of joined pieces, on average. Moreover, an estimate of degrees of freedom allows us to use different model selection criteria like C_p, AIC, or BIC. The C_p statistic, for example, is given by

$$C_p(\lambda) = \sum_{i=1}^{n}(y_i - \hat{\beta}_{\lambda,i})^2 - n\sigma^2 + 2\sigma^2 \mathrm{df}(\hat{\beta}_\lambda), \tag{13}$$

and is an unbiased estimate of the true risk $\mathrm{E}\big[\sum_{i=1}^{n}(\mu_i - \hat{\beta}_{\lambda,i})^2\big]$. We can define $\hat{C}_p(\lambda)$ as

$$\hat{C}_p(\lambda) = \sum_{i=1}^{n}(y_i - \hat{\beta}_{\lambda,i})^2 - n\sigma^2 + 2\sigma^2 K_\lambda, \tag{14}$$

replacing $\mathrm{df}(\hat{\beta}_\lambda)$ in (13) with its unbiased estimate K_λ. The modified statistic $\hat{C}_p(\lambda)$ is still unbiased as an estimate of the true risk, which suggests choosing λ to minimize $\hat{C}_p(\lambda)$.

Note that K_λ is a step function with respect to λ, with steps at the critical points $\{\lambda_1, \ldots \lambda_T\}$. Meanwhile, the residual sum of squares $\sum_{i=1}^n (y_i - \hat{\beta}_{\lambda,i})^2$ is monotone nondecreasing for λ in between critical points (this is proved in the Appendix, Section A.3). This means that $\hat{C}_p(\lambda)$ must achieve its minimum at some $\lambda_k \in \{\lambda_1, \ldots \lambda_T\}$, and hence Algorithm 1 can be used to simultaneously compute the path of solutions and select a value of the tuning parameter. We provide an example of this in the next section.

5 Global warming data

Here we look at data on annual temperature anomalies from 1856 to 1999, relative to the 1961-1990 mean, studied in Wu et al. (2001). The data, along with an interpolating function, are shown in the top left of Figure 2. Temperature seems to increase monotonically with possible decreases around 1900 and 1950. The entire nearly-isotonic path has 139 knots, four of which are shown in Figure 2, with the usual isotonic regression shown in the bottom right panel.

In order to try to determine whether the nonmonotonicity was real, we carried out 10-fold cross-validation for λ, on a grid of 20 equally spaced values. This produced the CV error curve on the left-hand side of Figure 3. The curve and its standard error bars indicate that the isotonic regression fit (the largest value of λ on the right) has significantly worse error than the near-isotonic fits. The one standard error rule chooses the value $\lambda = 0.66$. We also computed the modified C_p statistic, $\hat{C}_p$, given in (14). This was evaluated at each of the critical points $\{\lambda_1, \ldots \lambda_T\}$ (in the current example we also had to use an estimate of σ^2, which was derived from the full isotonic regression fit). The $\hat{C}_p$ curve is shown on the right-hand side of Figure 3, and is minimized at $\lambda = 0.44$. This corresponds to the top right panel in Figure 2, and is a less regularized model than that chosen by cross-validation.

6 Nearly-convex fitting

We can apply an idea similar to nearly-isotonic fitting to obtain a "nearly-convex" approximation to a sequence of data. Namely, given $y_1, \ldots y_n$, this is

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\mathrm{argmin}} \ \frac{1}{2} \sum_{i=1}^n (y_i - \beta_i)^2 + \lambda \sum_{i=1}^{n-2} (\beta_i - 2\beta_{i+1} + \beta_{i+2})_+. \tag{15}$$

The quantity $-\beta_i + 2\beta_{i+1} - \beta_{i+2}$ can be viewed as an approximation to the second derivative at β_{i+1}, and so in a sense, the penalty in problem (15) encourages the sequence β to have a negative second derivative. As another way of looking at the penalty, note that the ith term is positive when $\beta_{i+1} < (\beta_i + \beta_{i+2})/2$, that is, β_{i+1} lies below the line through its neighbors. For a sufficiently large value of λ, problem (15) produces the best convex fit to the data. For smaller values of λ it yields a path of nearly-convex approximations that lie close to the data.

The path algorithm for this problem is more complicated than that for nearly-isotonic regression. As λ increases, some of the differences $\beta_i - 2\beta_{i+1} + \beta_{i+2}$ that are negative (and hence violate convexity), are set to zero. This is the analogue of the joining events that occur in nearly-isotonic regression. However such differences that are zero for one value of λ can become nonzero again for $\lambda_0 > \lambda$, unlike the more orderly behavior guaranteed by Lemma 1 for nearly-isotonic regression. Despite this, a path algorithm can be derived, by letting $\alpha_i = \beta_i - \beta_{i+1}$ for $i = 1, \ldots n-1$ and $\alpha_n = \beta_n$, rewriting problem (15) as

$$\hat{\alpha}_\lambda = \underset{\alpha \in \mathbb{R}^n}{\mathrm{argmin}} \ \frac{1}{2} \sum_{i=1}^n \Big(y_i - \sum_{j=i}^n \alpha_j \Big)^2 + \lambda \sum_{i=1}^{n-2} (\alpha_i - \alpha_{i+1})_+,$$

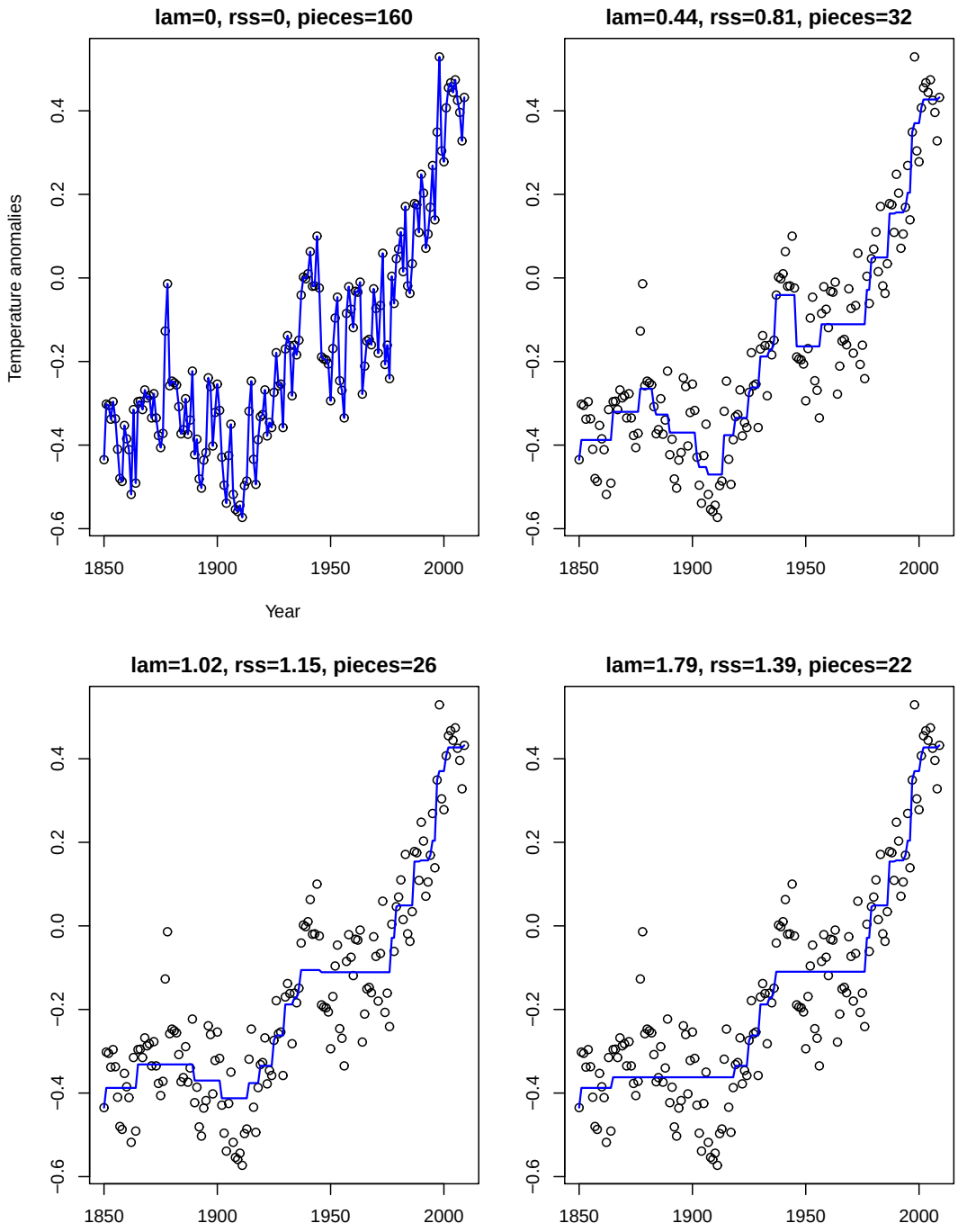

Figure 2: *Global warming data: temperature anomalies by year. An interpolating function is shown in the top left, along with the fit for three larger values of* λ*, and the usual isotonic regression shown in the bottom right panel. The title above each panel gives the value of* λ*, the residual sum of squares* $\sum(y_i - \hat{\beta}_{\lambda,i})^2$*, and the number of joined pieces* K_λ*.*

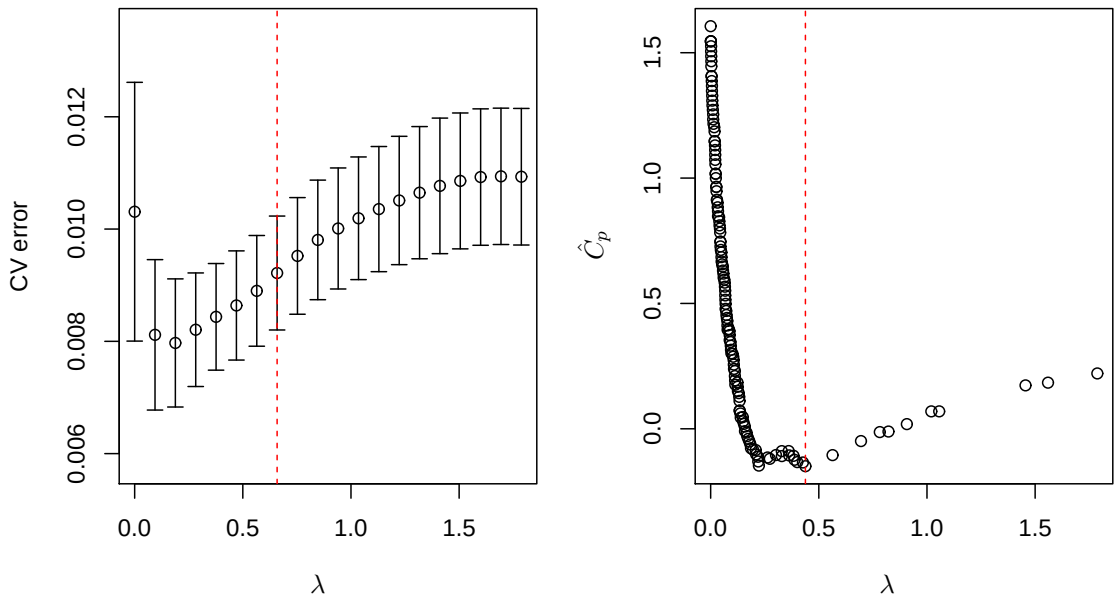

Figure 3: *Left panel: CV error curve for the global warming data, with* ± 1 *standard error bars. The vertical dashed line indicates the value of* λ *chosen by the one standard error rule. Right panel:* $\hat{C}_p$ *curve for the same dataset, with the vertical dashed line marking the minimizing value of* λ*.*

and using the tools developed in Hoefling (2010). Figure 4 shows a toy example.

7 Discussion

We have proposed a method for fitting a path of nearly-isotonic approximations to a sequence of data that culminates in the usual isotonic regression fit. We proved that the degrees of freedom at each point in the path is simply the average number of joined pieces in the fitted function. We have also extended the method to find nearly-convex approximations.

There are other ways that this work could be generalized. For two-dimensional data on a equally spaced $m \times n$ grid, we can modify the nearly-isotonic regression criterion to

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{i,j}-\beta_{i,j})^2+\lambda\Big[\sum_{i=1}^{m}\sum_{j=1}^{n-1}(\beta_{i,j}-\beta_{i,j+1})_+ + \sum_{i=1}^{m-1}\sum_{j=1}^{n}(\beta_{i,j}-\beta_{i+1,j})_+\Big],$$

which encourages monotonicity in both the horizontal (i index) and vertical (j index) directions. In another generalization, we could control the overall smoothness of the fitted function, in addition to its montonicity. For this, we could use the criterion

$$\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_i)^2+\lambda_1\sum_{i=1}^{n-1}(\beta_i-\beta_{i+1})_+ + \lambda_2\sum_{i=1}^{n-1}|\beta_i-\beta_{i+1}|,$$

where λ_1 and λ_2 are both tuning parameters. For example, letting $\lambda_1 \to \infty$ gives a fully monotone fit, whose smoothness can be controlled by varying λ_2. In this case ($\lambda_1 \to \infty$), the choice $\lambda_2 = 0$ gives the

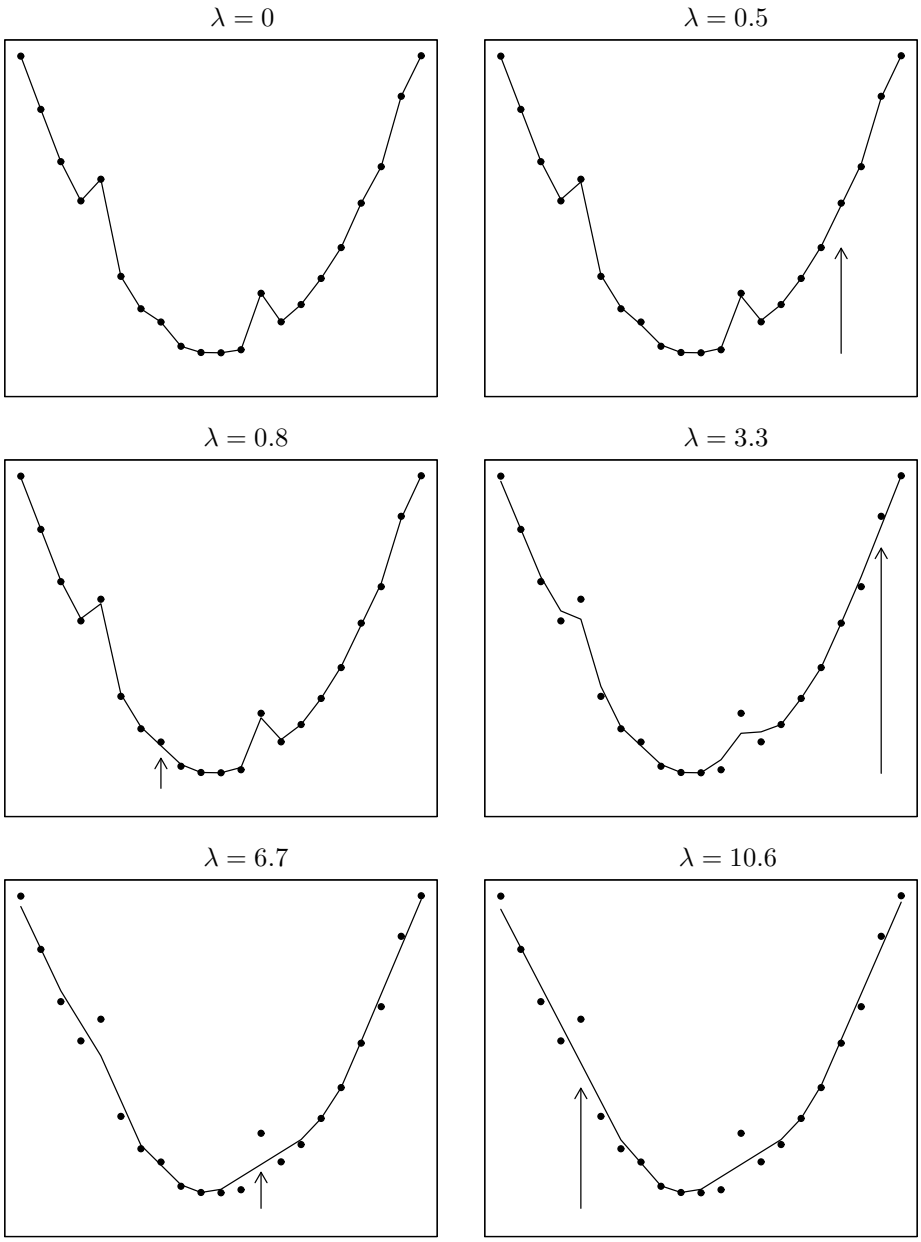

Figure 4: *Nearly-convex fits for a toy example. An interpolating function is shown in the top left. The other panels show members of the nearly-convex path where a nonconvex stretch (indicated by the arrow) is flattened into a linear segment. The best convex fit is shown in the bottom right panel.*

standard isotonic regression, while values larger than zero yield smoother monotone approximations. This provides a different way of achieving the same goal as that discussed in Luss et al. (2010).

An R package "neariso" that implements our path algorithm for nearly-isotonic regression will be made available on the CRAN website (R Development Core Team 2008).

Acknowledgements

The authors would like to thank Jonathan Taylor for his insights on the dual problem. The authors also thank Saharon Rosset and his coauthors for sending us a preprint of their related work. The third author was supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

A Appendix

A.1 Proof of Lemma 1

Similar to (4), but disregarding group structure, we can rewrite the subgradient equations of problem (2) as

$$-y_i + \hat{\beta}_{\lambda,i} + \lambda(s_i - s_{i-1}) = 0 \quad \text{for } i = 1, \ldots n, \tag{16}$$

where $s_i = 1(\hat{\beta}_{\lambda,i} - \hat{\beta}_{\lambda,i+1} > 0)$ if $\hat{\beta}_{\lambda,i} - \hat{\beta}_{\lambda,i+1}$ is > 0 or < 0, and otherwise $s_i \in [0,1]$. Following an argument of Friedman et al. (2007), suppose that we have a stretch of joined coordinates $\hat{\beta}_{\lambda,j} = \ldots = \hat{\beta}_{\lambda,j+k}$. We show that, as λ increases, this group of remains intact until it merges with a neighboring group.

An important point is that, at the value λ, both s_{j-1} and s_{j+k} are neccessarily in $\{0,1\}$. As λ increases, s_{j-1} and s_{j+k} will be constant as long as the group of coordinates $\{j, \ldots j+k\}$ doesn't merge with an adjacent one. Hence we consider a subset of the subgradient equations (16) corresponding to $i = j, \ldots j+k$, and show that as λ increases, it has a solution with $\hat{\beta}_{\lambda,j} = \ldots = \hat{\beta}_{\lambda,j+k}$ and $s_j, \ldots s_{j+k-1} \in [0,1]$. This is sufficient to prove the lemma.

Taking pairwise differences between equations, and using the fact that $\hat{\beta}_{\lambda,j} = \ldots = \hat{\beta}_{\lambda,j+k}$, we get

$$As = \frac{1}{\lambda} Dy + c,$$

where

$$A = \begin{bmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ \ldots & & & & & & \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{bmatrix}, \quad D = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ & & & \ldots & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix},$$

and $s = (s_j, \ldots s_{j+k-1})$, $y = (y_j, \ldots y_{j+k})$, and $c = (s_{j-1}, 0, \ldots s_{j+k})$. But A is invertible, so we can again rewrite this as

$$s = \frac{1}{\lambda} A^{-1} Dy + A^{-1} c. \tag{17}$$

We only need to check that s will remain in $[0,1]$ coordinatewise. As λ increases, the first term above only gets smaller in magnitude. Therefore, if $A^{-1}c$ is in $[0,1]$ coordinatewise, then the right-hand side of (17) will stay in $[0,1]$ for increasing λ, completing the proof. As A is tridiagonal, its inverse has an explicit form (Schlegel 1970). All we need to know is that $(A^{-1})_{i1} = (n-i+1)/(n+1)$ and $(A^{-1})_{in} = i/(n+1)$, and then one can check directly that $A^{-1}c$ is coordinatewise in $[0,1]$ for all three possibilities for c.

A.2 Degrees of freedom proof

In many cases, the most straightforward calculation for degrees of freedom comes from Stein's formula (Stein 1981), which states that for a function $g(y) = \big(g_1(y), \dots g_n(y)\big)$ of y

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Cov}(g_i, y_i) = \mathrm{E}\Big[\sum_{i=1}^{n}\frac{\partial g_i}{\partial y_i}\Big].$$

But Stein's result only holds when g is continuous and almost differentiable.

Recall that if A_i, $i = 1 \dots K_\lambda$ are the joined pieces in the solution $\hat{\beta}_\lambda$, then we can express the fitted value of each joined piece as

$$\hat{\beta}_{\lambda,A_i} = \frac{\sum_{j\in A_i} y_j \;-\; \lambda(s_i - s_{i-1})}{|A_i|},$$

where $s_i = 1(\hat{\beta}_{\lambda,A_i} - \hat{\beta}_{\lambda,A_{i+1}} > 0)$. Suppose we take the derivative of the above expression with respect to y_j for $j \in A_i$, and simply treat the A_i's and s_i's as constants. Doing so would give $\partial\hat{\beta}_{\lambda,j}/\partial y_j = 1/|A_i|$, and assuming that we can apply Stein's formula to $\hat{\beta}_\lambda$,

$$\mathrm{df}(\hat{\beta}_\lambda) = \mathrm{E}\Big[\sum_{i=1}^{K_\lambda}\sum_{j\in A_i}\frac{1}{|A_i|}\Big] = \mathrm{E}(K_\lambda). \tag{18}$$

The following lemma states that $\hat{\beta}_\lambda$ is continuous and almost differentiable, and that for almost every y the groups A_i and values s_i, $i = 1, \dots K_\lambda$, are locally constant. These conditions allow us to precisely conclude (18).

Lemma 2. *For fixed λ:*

(i) the solution $\hat{\beta}_\lambda$ is continuous and almost differentiable as a function of y;

(ii) for almost every y, there is a ball B containing y such that K_λ and A_i, s_i, $i = 1, \dots K_\lambda$ are the same for any $y_0 \in B$.

Proof. The easiest proof comes from looking at the dual of problem (2). Letting D be the $(n-1)\times n$ matrix

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ & & & \dots & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix},$$

we can rewrite minimization (2) as

$$\hat{\beta}_\lambda = \operatorname*{argmin}_{\beta\in\mathbb{R}^n}\frac{1}{2}\|y-\beta\|_2^2 + \lambda 1^T(D\beta)_+. \tag{19}$$

Here we use x_+ for the positive part applied coordinate-wise to a vector x. Convex analysis tells us that the dual of (19) is

$$\hat{u}_\lambda = \operatorname*{argmin}_{u\in\mathbb{R}^{n-1}}\frac{1}{2}\|y - D^T u\|_2^2 \;\text{ subject to }\; 0 \le u \le \lambda. \tag{20}$$

The above inequality represents a coordinate-wise inequality, $0 \le u_i \le \lambda$ for each $i = 1, \dots n-1$. The solutions of (19) and (20) are related by

$$\hat{\beta}_\lambda = y - D^T\hat{u}_\lambda.$$

We will write $\hat{\beta}_\lambda(y)$ and $\hat{u}_\lambda(y)$ to emphasize that these are functions of the data y. Studying (20), the fit $D^T\hat{u}_\lambda(y)$ is the projection of y onto the convex set $\mathcal{P}_\lambda = \{D^T u : 0 \leq u \leq \lambda\}$. But a projection onto a convex set is a contraction, and hence $\hat{\beta}_\lambda(y)$ is Lipschitz with constant 2, since

$$\|\hat{\beta}_\lambda(y) - \hat{\beta}_\lambda(y_0)\|_2 \;\leq\; \|y - y_0\|_2 + \|D^T\hat{u}_\lambda(y) - D^T\hat{u}_\lambda(y_0)\|_2 \;\leq\; 2\|y - y_0\|_2.$$

This immediately means that $\hat{\beta}_\lambda(y)$ is continuous and almost differentiable [for example, see Theorem 2 in Section 3.2 of Evans & Gariepy (1992)], which establishes part (i) of the lemma.

As for part (ii), it turns out that the number of pieces K_λ, as well as the sets A_i and values s_i, $i = 1, \ldots K_\lambda$, are determined entirely by the face of the polytope $\mathcal{P}_\lambda$ onto which y projects. For almost every $y \in \mathbb{R}^n$, there is a ball around y whose interior projects to a single face. The points that don't share this property necessarily lie on a ray that emanates orthogonally from one of the corners of $\mathcal{P}_\lambda$, but the union of these rays has measure zero. For more details, the reader can refer to Section 10 of Tibshirani & Taylor (2010) or Section 2 of Meyer & Woodroofe (2000).

□

A.3 Proof that $\sum_{i=1}^n (y_i - \hat{\beta}_{\lambda,i})$ is monotone nondecreasing for λ in between critical points

Consider the function $f(\lambda) = \sum_{i=1}^n (y_i - \hat{\beta}_{\lambda,i})^2$, where $\hat{\beta}_\lambda$ is the solution to nearly-isotonic regression (2). We show that $f(\lambda)$ is nondecreasing for λ in between any two critical points in the solution path. First define, for $j \in A_i$, $r_j = s_i = 1(\hat{\beta}_{\lambda,A_i} - \hat{\beta}_{\lambda,A_{i+1}} > 0)$. Then we can rewrite (9) as

$$\hat{\beta}_{\lambda,A_i} = \frac{\sum_{j\in A_i} y_j - \lambda \sum_{j\in A_i}(r_j - r_{j-1})}{|A_i|},$$

because the second sum above is telescoping. Therefore we can write the whole solution in vector notation as

$$\hat{\beta}_\lambda = P(y - \lambda D r), \tag{21}$$

where P is the matrix that averages within each group $A_1, \ldots A_{K_\lambda}$, and D is the matrix that gives adjacent differences. As λ varies, equation (21) continues to give the solution as long as the groups A_i and signs s_i do not change, which happens at the critical points. We have

$$y - \hat{\beta}_\lambda = (I - P)y + \lambda P D r,$$

and using the fact that P is a projection matrix,

$$\|y - \hat{\beta}_\lambda\|_2^2 = y^T(I - P)y + \lambda^2 r^T D^T P D r.$$

The coefficient multiplying λ^2 is nonnegative, which completes the proof.

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. & Silverman, E. (1955), 'An empirical distribution function for sampling with incomplete information', *Annals of Mathematical Statistics* **26**(4), 641–647.

Barlow, R. E., Bartholomew, D., Bremner, J. M. & Brunk, H. D. (1972), *Statistical inference under order restrictions; the theory and application of isotonic regression*, Wiley, New York.

Bartholomew, D. J. (1959*a*), 'A test for homogeneity for ordered alternatives', *Biometrika* **46**(1), 36–48.

Bartholomew, D. J. (1959*b*), 'A test for homogeneity for ordered alternatives II', *Biometrika* **46**(3), 328–335.

Brunk, H. D. (1955), 'Maximum likelihood estimates of monotone parameters', *Annals of Mathematical Statistics* **26**(4), 607–616.

de Leeuw, J., Hornik, K. & Mair, P. (2009), 'Isotone optimization in R: Pool-adjacent-violators (PAVA) and active set methods', *Journal of Statistical Software* **32**(5), 1–24.

Efron, B. (1986), 'How biased is the apparent error rate of a prediction rule?', *Journal of the American Statistical Association* **81**(394), 461–470.

Evans, L. & Gariepy, R. (1992), *Measure theory and fine properties of functions*, CRC Press, Boca Raton.

Friedman, J., Hastie, T., Hoefling, H. & Tibshirani, R. (2007), 'Pathwise coordinate optimization', *Annals of Applied Statistics* **1**(2), 302–332.

Grotzinger, S. J. & Witzgall, C. (1984), 'Projections onto simplices', *Applied Mathematics and Optimization* **12**(1), 247–270.

Hastie, T., Tibshirani, R. & Friedman, J. (2008), *The Elements of Statistical Learning; Data Mining, Inference and Prediction (2nd edition)*, Springer Verlag, New York.

Hoefling, H. (2010), A path algorithm for the fused lasso signal approximator. Unpublished.
URL: *http://www.holgerhoefling.com/Articles/FusedLasso.pdf*

Luss, R., Rosset, S. & Shahar, M. (2010), Isotonic recursive partioning. Submitted.

Meyer, M. & Woodroofe, M. (2000), 'On the degrees of freedom in shape-restricted regression', *Annals of Statistics* **28**(4), 1083–1104.

Miles, R. E. (1959), 'The complete amalgamation into blocks, by weighted means, of a finite set of real numbers', *Biometrika* **46**(3), 317–327.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Robertson, T., Wright, F. T. & Dykstra, R. L. (1988), *Order restricted statistical inference*, Wiley, New York.

Schlegel, P. (1970), 'The explicit inverse of a tridiagonal matrix', *Mathematics of Computation* **24**(111), 665–665.

Stein, C. (1981), 'Estimation of the mean of a multivariate normal distribution', *Annals of Statistics* **9**(6), 1135–1151.

Tibshirani, R. J. & Taylor, J. (2011), 'The solution path of the generalized lasso', *Annals of Statistics* **39**(3), 1335–1371.

Wu, W., Woodroofe, M. & Mentz, G. (2001), 'Isotonic regression; another look at the changepoint problem', *Biometrika* **88**(3), 794–804.