Least Squares estimation of two ordered monotone regression curves

Fadoua Balabdaoui, Kaspar Rufibach, Filippo Santambrogio

▶ **To cite this version:**

Fadoua Balabdaoui, Kaspar Rufibach, Filippo Santambrogio. Least Squares estimation of two ordered monotone regression curves. 2009. hal-00417281

HAL Id: hal-00417281
https://hal.science/hal-00417281

Preprint submitted on 15 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Least Squares estimation of two ordered monotone regression curves

running headline: ordered monotone regression

Fadoua Balabdaoui[(1,2)], Kaspar Rufibach[(3)] and Filippo Santambrogio[(1)]

[1] CEREMADE
Université de Paris-Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris CEDEX 16, France

[2] Universität Göttingen
Institut für Mathematische Stochastik
Goldschmidtstrasse 7
37077 Göttingen

[3] Universität Zürich
Institut für Sozial- und Präventivmedizin
Abteilung Biostatistik
Hirschengraben 84
8001 Zürich

fadoua@ceremade.dauphine.fr
kaspar.rufibach@ifspm.uzh.ch (corresponding author)
filippo@ceremade.dauphine.fr

Abstract

In this paper, we consider the problem of finding the Least Squares estimators of two isotonic regression curves g_1° and g_2° under the additional constraint that they are ordered; e.g., $g_1^\circ \leq g_2^\circ$. Given two sets of n data points $y_1, .., y_n$ and $z_1, \ldots, z_n$ observed at (the same) design points, the estimates of the true curves are obtained by minimizing the weighted Least Squares criterion $L_2(a,b) = \sum_{j=1}^n (y_j - a_j)^2 w_j^1 + \sum_{j=1}^n (z_j - b_j)^2 w_j^2$ over the class of pairs of vectors $(a,b) \in \mathbb{R}^n \times \mathbb{R}^n$ such that $a_1 \leq a_2 \leq ... \leq a_n$, $b_1 \leq b_2 \leq ... \leq b_n$, and $a_i \leq b_i, i = 1, ..., n$. The characterization of the estimators is established. To compute these estimators, we use an iterative projected subgradient algorithm, where the projection is performed with a "generalized" pool-adjacent-violaters algorithm (PAVA), a byproduct of this work. Then, we apply the estimation method to real data from mechanical engineering.

Keywords: least squares, monotone regression, pool-adjacent-violaters algorithm, shape constraint estimation, subgradient algorithm

1 Introduction and motivation

Estimating a monotone regression curve is one of the most classical estimation problems under shape restrictions, see e.g. Brunk (1958). A regression curve is said to be isotonic if it

is monotone nondecreasing. We chose in this paper to look at the class of isotonic regression functions. The simple transformation $g \to -g$ suffices for the results of this paper to carry over to the antitonic class.

Given n fixed points $x_1, \ldots, x_n$, assume that we observe y_i at x_i for $i = 1, \ldots, n$. When the points (x_i, y_i) are joined, the shape of the obtained graph can hint at the increasing monotonicity of the true regression curve, g° say, assuming the model $y_i = g^\circ(x_i) + \varepsilon_i$, with ε_i the unobserved errors. This shape restriction can also be a feature of the scientific problem at hand, and hence the need for estimating the true curve in the class of antitonic functions. We refer to Barlow et al. (1972) and Robertson et al. (1988) for examples. The weighted Least Squares estimate of g° in the class of isotonic functions taking y_i at x_i is the unique minimizer of the criterion

$$L(a) = \sum_{i=1}^{n} w_i (y_i - a_i)^2 \tag{1}$$

over the class of vectors $a \in \mathbb{R}^n$ such that $a_1 \leq a_2 .... \leq a_n$ where $w_1 > 0, w_2 > 0, \ldots, w_n > 0$ are given positive weights. In what follows, we will say that a vector $v \in \mathbb{R}^n$ is increasing or isotonic if $v_1 \leq \ldots \leq v_n$, and use the notation $v \leq w$ for $v, w \in \mathbb{R}^n$ if the inequality holds componentwise.

It is well known that the solution a^* of the Least Squares problem in (1) is given by the so-called min-max formula; i.e.,

$$a_i^* = \max_{s \leq i} \min_{t \geq i} Av(\{s, \ldots, t\}) \tag{2}$$

where $Av(\{s, \ldots, t\}) = \sum_{i=s}^{t} y_i w_i / \sum_{i=s}^{t} w_i$ (see e.g. Barlow et al., 1972).

van Eeden (1957a,b) has generalized this problem to incorporate known bounds on the regression function to estimate; i.e., she considered minimization of L under the constraint

$$a_L \leq a \leq a_U, \tag{3}$$

for two increasing vectors a_L and a_U. As in the classical setting, the solution of this problem admits also a min-max representation. The PAVA can be generalized to efficiently compute this solution and has been implemented in the `R` package `OrdMonReg` (Balabdaoui et al., 2009). Computation relies on a suitable functional M defined on the sets $A \subseteq \{1, ..., n\}$ which generalizes the function Av in (2). This functional for the bounded monotone regression in (3) is given by

$$M(A) = \Big(Av(A) \vee \max_A a_L \Big) \wedge \min_A a_U$$

where $\min_A v = \min_{i \in A} v_i$ and $\max_A v = \max_{i \in A} v_i$. Compare Barlow et al. (1972, p. 57), where a functional notation is used. However, in the latter reference no formal justification was given for the form of the functional M nor for the validity of (the modified version of) the PAVA, see discussion after Theorem 2.1.

Chakravarti (1989) discusses the bounded isotonic regression problem for the absolute value criterion function, yielding the bounded isotonic median regressor. Chakravarti (1989) proposes a PAVA-like algorithm as well, and establishes some connections to linear programming theory. Unbounded isotonic median regression was first considered by Robertson and Waltman (1968), who provided a min-max formula for the estimator and a PAVA-like algorithm to compute it. They also studied its consistency.

Now suppose that instead of having only one set of observations $y_1, \ldots, y_n$ at the design points $x_1, \ldots, x_n$, we are interested in analyzing two sets of data $y_1, \ldots, y_n$ and $z_1, \ldots, z_n$ observed at the same design points. Furthermore, if we have the information that the underlying true regression curves are increasing and ordered, it is natural to try to construct estimators that fulfill the same constraints.

The current paper presents a solution to this problem of estimating two isotonic regression curves under the additional constraint that they are ordered. This solution is the unique minimizer (a^*, b^*) of the criterion

$$L_2(a,b) \quad = \quad \sum_{i=1}^{n} w_i^1 (y_i - a_i)^2 + \sum_{i=1}^{n} w_i^2 (z_i - b_i)^2 \tag{4}$$

over the class of pairs of vectors $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$ such that a and b are increasing and $a \leq b$, with w^1 and w^2 given vectors of positive weights in $\mathbb{R}^n$.

The problem was motivated by an application from mechanical engineering. We will make use of experimental data obtained from dynamic material tests (see Shim and Mohr, 2009) to illustrate our estimation method. In engineering mechanics, it is common practice to determine the deformation resistance and strength of materials from uniaxial compression tests at different loading velocities. The experimental results are the so-called stress-strain curves (see Figure 1), and these may be used to determine the deformation resistance as a function of the applied deformation. The recorded signals contain substantial noise which is mostly due to variations in the loading velocity and electrical noise in the data acquisition system.

The data in this example consist of 1495 distinct pairs (x_i, y_i) and (x_i, z_i) where x_i is the measured strain, while y_i (gray curve) and z_i (black curve) correspond to the experimental stress results for two different loading velocities. The true regression curves are expected to be (a) monotone increasing as the stress is known to be an increasing function of the strain (for a given constant loading velocity), and (b) ordered as the deformation resistance typically increases as the loading velocity increases. In Section 3, we show the resulting estimates as well as a smoothed version thereof.

We will show that minimizing L_2 is equivalent to minimizing another convex functional over the class of isotonic vectors $a \in \mathbb{R}^n$. By doing so, we reduce a two-curve problem under the constraints of monotonicity and ordering to a one-curve problem under the constraint of monotonicity and boundedness. Actually, we can even perform the minimization over

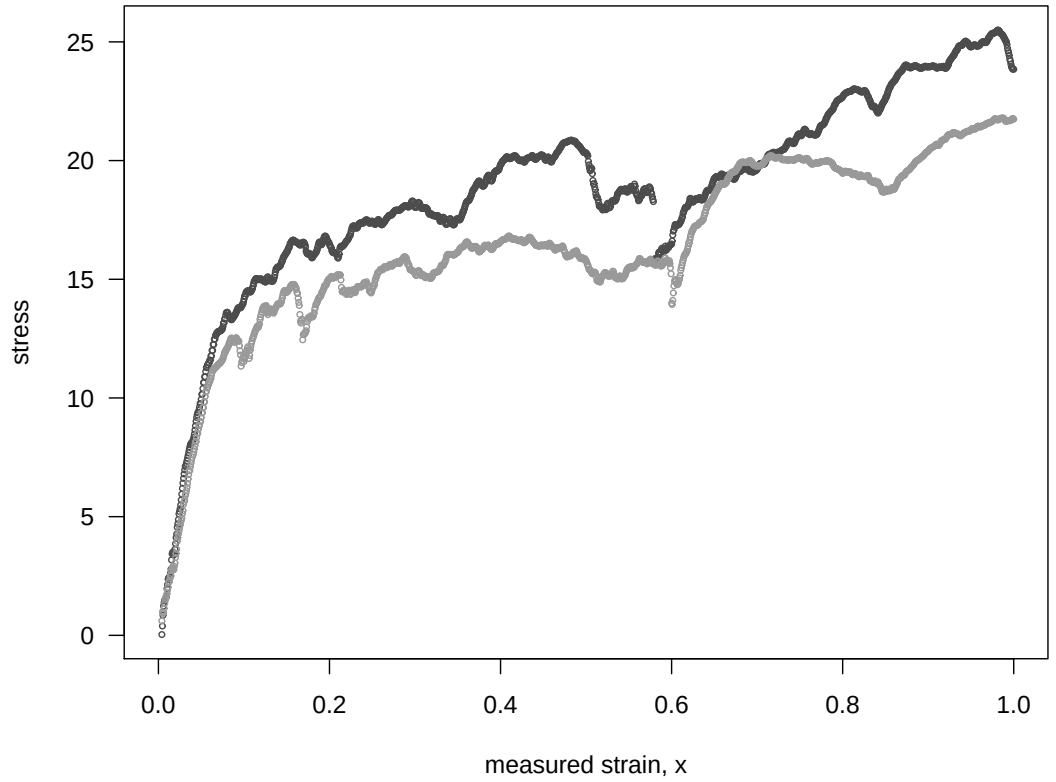

Figure 1: Original observations.

the class of isotonic vectors $(a_1, \ldots, a_{n-1})$ of dimension $n-1$ satisfying the constraint $a_1 \leq \ldots \leq a_{n-1} \leq a_n^*$ as we can explicitly determine a_n^* by a *generalized* min-max formula (see Proposition 2.3). The solution of this equivalent minimization problem, which gives the solution a^* (and also b^* because it is a function of a^*), is computed using a projected subgradient algorithm where the projection step is performed using a suitable generalization of the PAVA.

We would like to note that Brunk et al. (1966) considered a related problem, that of non-parametric Maximum likelihood estimation of two ordered cumulative distribution functions. In the same class of problems, Dykstra (1982) treated estimation of survival functions of two stochastically ordered random variables in the presence of censoring, which was extended by Feltz and Dykstra (1985) to $N \geq 2$ stochastically ordered random variables. The theoretical solution can be related to the well-known Kaplan-Meier estimator and can be computed using an iterative algorithmic procedure for $N \geq 3$ (see Feltz and Dykstra, 1985, p. 1016). The $\sqrt{n}-$ asymptotics of the estimators for $N = 2$, whether there is censoring or not, were established by Præstgaard and Huang (1996).

The paper is organized as follows. In Section 2, we give the characterization of the ordered isotonic estimates. We also provide the explicit form of the solution of the related bounded

isotonic regression problem where the upper of the two isotonic curves is assumed to be fully known.

In Section 3 we describe the projected subgradient algorithm that we use to compute the Least Squares estimators of the ordered isotonic regression curves, and apply the method to real data from mechanical engineering. The technical proofs are deferred to appendices A and B.

2 Estimation of two ordered isotonic regression curves

If the larger of the two isotonic curves was known, then there would of course be no need to estimate it. If we put $a_U = a^0$, the weighted Least Squares estimate a^* of the smaller isotonic curve is the minimizer of

$$L(a) \quad = \quad \sum_{i=1}^{n} w_i (y_i - a_i)^2,$$

where $w \in \mathbb{R}^n$ is a vector of given positive weights, and $a \in \mathcal{I}_n^{a^0}$, the class of isotonic vectors $a \in \mathbb{R}^n$ such that $a \leq a^0$ and $a^0 \in \mathbb{R}^n$. When the components of a^0 are all equal, the vector a^0 will be assimilated with the common value of its components as done in Proposition 3.4 below.

The notation $\mathcal{I}_n^w$ will be used again hereafter to denote the class of isotonic vectors $v \in \mathbb{R}^n$ such that $v \leq w$.

The statement of Barlow et al. (1972, p. 57) implies that if we define

$$M(A) = Av(A) \wedge \min_A a^0$$

for a subset $A \subseteq \{1, ..., n\}$, then the solution a^* can be computed using an appropriately modified version of the PAVA.

Theorem 2.1. *For $i = 1, \ldots, n$, we have*

$$a_i^* = \max_{s \leq i} \min_{t \geq i} M(\{s, \ldots, t\}) = \max_{s \leq i} \min_{t \geq i} \Big(Av(\{s, \ldots, t\}) \wedge a_s^0 \Big).$$

To keep this paper at a reasonable length, the proof of Theorem 2.1 is omitted. A short note containing a more thorough discussion of the one-curve problem and a proof of Theorem 2.1 can be obtained from the authors upon request. A general description of the modified PAVA and a proof that it works whenever the functional M satisfies the so-called *Averaging Property* can be found in Section 3.

We now return to the main subject of this paper. Theorem 2.1 is crucial for finding the Least Squares estimates of two ordered isotonic regression curves. In particular, the result will be used to develop an appropriate algorithm to compute the solution.

Let $y_1, ...y_n$ and $z_1, ..., z_n$ be the observed data from two unknown isotonic curves g_1° and g_2° such that $g_1^\circ \leq g_2^\circ$. Given two vectors in $\mathbb{R}^n$ of positive weights w^1 and w^2, we would like to minimize

$$L_2(a,b) \quad = \quad \sum_{i=1}^{n}(y_i - a_i)^2 w_i^1 + \sum_{i=1}^{n}(z_i - b_i)^2 w_i^2 \tag{5}$$

over the class of pairs of vectors $(a,b) \in \mathbb{R}^n \times \mathbb{R}^n$ such that a and b are isotonic and $a \leq b$. Call this class $\mathcal{I}_n$.

Existence and uniqueness of the solution. They follow from convexity and closedness of $\mathcal{I}_n$ and strict convexity of L_2.

Characterization of the solution. For completeness, we give the characterization of the solution of minimizing (5) over $\mathcal{I}_n$; i.e, a necessary and sufficient condition for $(a,b) \in \mathcal{I}_n$ to be equal to this solution. Let $i_1 < ... < i_k$ such that $i_1 = 1, i_k = n$ and

$$a_1^* = ... = a_{i_1}^* < a_{i_1+1}^* = ... = a_{i_2-1}^* < ... < a_{i_k}^* = ... = a_n^*.$$

We call $B_{i_j}^0$ (resp. $B_{i_j}^1$) a set of indices $\{i_j, ..., i_{j+1} - 1\}, j = 1, ..., k-1$ such that $a_{i_j}^* = b_{i_j}^*$ (resp. $a_{i_j}^* < b_{i_j}^*$). Similarly, let $l_1 < ... < l_r$ such that $l_1 = 1, l_r = n$ such that

$$b_1^* = ... = b_{l_1}^* < b_{l_1+1}^* = ... = b_{l_2-1}^* < ... < b_{l_k}^* = ... = b_n^*$$

and call $C_{l_j}^0$ (resp. $C_{l_j}^1$) a set of indices $\{l_j, ..., l_{j+1} - 1\}, j = 1, ..., r-1$ such that $b_{l_j}^* = a_{l_j}^*$ (resp. $b_{l_j}^* > a_{l_j}^*$).

Theorem 2.2. *The pair $(a^*, b^*) \in \mathcal{I}_n$ is the minimizer of (5) if and only if*

$$\sum_{i=1}^{n}(a_i^* - y_i)(a_i^* - a_i)w_i^1 + \sum_{i=1}^{n}(b_i^* - z_i)(b_i^* - b_i)w_i^2 \quad \geq \quad 0, \ \ \forall\, (a,b) \in \mathcal{I}_n \tag{6}$$

$$\sum_{s \in \cup_j B_{i_j}^1}(a_s^* - y_s)a_s^* w_s^1 \quad = \quad 0, \text{ and} \tag{7}$$

$$\sum_{s \in \cup_j C_{l_j}^1}(b_s^* - z_s)b_s^* w_s^2 \quad = \quad 0. \tag{8}$$

Proof. See Appendix A.

An explicit formula in the sense of a min-max representation similar to (2) of (a^*, b^*) turned out be to hard to find. However, since a^* (resp. b^*) is also the minimizer of

$$\sum_{i=1}^{n}(a - y_i)^2 w_i^1 \quad \Big(\text{resp.} \sum_{i=1}^{n}(b - z_i)^2 w_i^2\Big)$$

over the class $\mathcal{I}_n^{b^*}$ (resp. the class of isotonic vectors $b \in \mathbb{R}^n$ such that $b \geq a^*$), Theorem 2.1 implies that

$$a_i^* = \max_{s \leq i} \min_{t \geq i} \left(Av_1(\{s, \ldots, t\}) \wedge b_s^* \right) \tag{9}$$

$$b_i^* = \max_{s \leq i} \min_{t \geq i} \left(Av_2(\{s, \ldots, t\}) \vee a_t^* \right) \tag{10}$$

for $i = 1, \ldots, n$, where

$$Av_1(A) = \frac{\sum_{i \in A} y_i w_i^1}{\sum_{i \in A} w_i^1}, \text{ and } Av_2(A) = \frac{\sum_{i \in A} z_i w_i^2}{\sum_{i \in A} w_i^2}$$

for $A \subseteq \{1, ..., n\}$.

Thus, the solution (a^*, b^*) is a fixed point of the operator $P : \mathcal{I}_n \to \mathcal{I}_n$ defined as

$$\begin{aligned} P((a,b)) &= (P_1(b), P_2(a)) \qquad (11) \\ &= \left(\max_{s \leq i} \min_{t \geq i} \left(Av_1(\{s, \ldots, t\}) \wedge b_s \right), \max_{s \leq i} \min_{t \geq i} \left(Av_2(\{s, \ldots, t\}) \vee a_t \right) \right). \end{aligned}$$

However, this fixed point problem does not admit a unique solution. Therefore, there is no guarantee that an algorithm based on the above min-max formulas yields the solution, except in the unrealistic and uninteresting case where the starting point of the algorithm is the solution itself. To see that P does not admit a unique fixed point, note that the minimizer of the criterion

$$\sum_{i=1}^{n} (a_i - y_i)^2 w_i^1 + B \sum_{i=1}^{n} (b_i - z_i)^2 w_i^2$$

is a fixed point of P for any $B > 0$. Therefore, a computational method based on starting from an initial candidate and then alternating between (9) and (10) cannot be successful. In parallel, we have invested a substantial effort in trying to get a closed form for the estimators. Although we did not succeed, we were able to obtain a closed form for a_1^* (and by symmetry for b_n^*).

Proposition 2.3. *We have that*

$$a_1^* = \min_{t \geq 1} Av_1(\{1, \ldots, t\}) \wedge \min_{t \geq t' \geq 1} \tilde{M}(\{1, \ldots, t\}, \{1, \ldots, t'\}) \tag{12}$$

where

$$\tilde{M}(A, B) = \frac{Av_1(A)(\sum_{i \in A} w_i^1) + Av_2(B)(\sum_{j \in B} w_j^2)}{\sum_{i \in A} w_i^1 + \sum_{j \in B} w_j^2}.$$

By symmetry, we also have that

$$b_n^* = \max_{t \leq n} Av_2(\{t, \ldots, n\}) \vee \max_{t \leq t' \leq n} \tilde{M}(\{t', \ldots, n\}, \{t, \ldots, n\}). \tag{13}$$

Some remarks are in order. On the one hand, the expressions obtained above indicate that the Least Squares estimator must depend, as expected, on the relative ratio of the weights w^1 and w^2. In particular, if $w^2 = 0$ (resp. $w^1 = 0$), the expression of a_1^* (resp. b_n^*) specializes to the well-known min-max formula in the classical Least Squares estimation of an (unbounded) isotonic curve. On the other hand, the expression of b_n^* is essential for our subgradient algorithm below.

Proof of Proposition 2.3. See Appendix A.

In the next section, we describe how we can make use of the min-max formula in (9) to compute the estimators using a projected subgradient algorithm. As mentioned above, we use in this algorithm the identity (13) given in the previous proposition.

3 Algorithms and Application to real data

In this section, we show that the bounded isotonic estimator can be computed using the well-known PAVA, or to be more precise a modified version of it. Recall that the bounded isotonic estimator in the one-curve problem is given by

$$a_i^* \quad = \quad \max_{s \leq i} \min_{t \geq i} M(\{s, \ldots, t\})$$

where $M(A) = Av(A) \vee \max_A a^0, A \subseteq\in \{1, ..., n\}$. That a^* can be computed using a PAVA is a consequence of a more general result: This computational fact is true provided that a functional M of sets $A \subseteq \{1, ..., n\}$ satisfies what is referred to as the *Averaging Property* , (see Chakravarti, 1989, p. 138), also called *Cauchy Mean Value Property* by Leurgans (1981, Section 1). See also Robertson et al. (1988, p. 390). Note that in the classical unconstrained monotone regression problem, the min-max expression of the Least Squares estimator follows from Theorem 2.8 in Barlow et al. (1972, p. 80).

3.1 Getting the min-max solution by the PAVA

First, let us describe how the PAVA works for some set functional M.

- At every step the current configuration is given by a subdivision of $\{1, ..., \}$ into k subsets $S_1 = \{1, \ldots, i_1\}$, $S_2 = \{i_1 + 1, \ldots, i_2\}, \ldots, S_k = \{i_{k-1} + 1, \ldots, n\}$ for some indices $1 = i_0 \leq i_1 < i_2 < \cdots < i_{k-1} < i_k = n$.
- The initial configuration is given by the finest subdivision; i.e., $I_j = \{j\}$.
- At every step we look at the values of M on the sets of the subdivision. A violation is noted each time there exists a value j such that $M(S_j) > M(S_{j+1})$. We consider the first violation (the one corresponding to the smallest j) and then merge the subsets S_j and S_{j+1} into one interval.

- Given a new subdivision (which has one subset less than the previous one), we look for possible violations.
- The algorithm stops when there are no violations left.

Since for any violation a merging is performed (thus reducing the number of subsets), it is clear that the algorithm stops after a finite number of iterations.

We require now the set functional M to satisfy the following property. See Leurgans (1981, Section 1), Robertson et al. (1988, p. 390) and Chakravarti (1989, p. 138).

Definition 3.1. *We say that the functional M satisfies the Averaging Property if for any sets A and B such that $A \cap B = \emptyset$ we have that*

$$\min\{M(A), M(B)\} \leq M(A \cup B) \leq \max\{M(A), M(B)\}.$$

If h and $w > 0$ are given vectors $\in \mathbb{R}^n$, then beside

$$A \mapsto Av(A) \quad = \quad \sum_{i \in A} w_i h_i / \sum_{i \in A} w_i,$$

the following examples of functions also satisfy the Averaging Property :

$$\begin{aligned}
A &\mapsto \left(Av(A) \vee \max_A h_i^1\right) \wedge \min_A h^0, \text{ with } h^0, h^1 \text{ two vectors} \in \mathbb{R}^n, \\
A &\mapsto \min_A h = \min_{i \in A} h_i, \\
A &\mapsto \operatorname{med}_A h = \arg\min_{m \in \mathbb{R}} \sum_{i \in A} |h_i - m| w_i \\
&\qquad \text{where the } \arg\min \text{ is taken to be the smallest } m \text{ in case non-uniqueness occurs}, \\
A &\mapsto \max_A h = \max_{i \in A} h_i.
\end{aligned}$$

Note that the maximum, the minimum and the sum of two functionals satisfying the Averaging Property satisfy the same property as well.

Theorem 3.2. *The final configuration obtained by the PAVA is such that the two following properties are satisfied.*

1. *The functional M is increasing on the sets of the subdivision.*
2. *If one of the sets $S_j = C \cup D$ is the disjoint union of two subsets $C = \{i_{j-1} + 1, \ldots, k\}$ and $D = \{k+1, \ldots, i_j\}$, then $M(C) > M(D)$; i.e., a finer subdivision would necessarily cause a violation.*

Proof. The fact that M is increasing on the final configuration is an easy consequence of the absence of violations (otherwise the algorithm would not have stopped).

As for the second part of the property, note that this is satisfied by the initial configuration (since no set is the disjoint union of two non-trivial subsets), as well as by any configuration that one could obtain after the first merging (since a merging occurs only because of a violation). Now we will use an inductive reasoning.

To this end, we have to check two situations: Suppose we merge two subsequent sets A and B and want to check whether there is a violation on C and D, with $A \cup B = C \cup D$. We are in one of the two following cases: either $A = A_1 \cup A_2$, $C = A_1$ and $D = A_2 \cup B$, or $B = B_1 \cup B_2$, $C = A \cup B_1$ and $D = B_2$ (the case $C = A$ and $D = B$ is trivial).

In the first case, if we suppose $M(D) \geq M(C)$, we get

$$M(A_2 \cup B) \geq M(A_1),\ M(A_2) < M(A_1),\ M(B) < M(A) = M(A_1 \cup A_2),$$

(the first inequality follows by assumption, the second by induction, and the third is true since A and B have been merged) and this is impossible since one would conclude that

$$\max\{M(A_2), M(B)\} \geq M(A_1) > M(A_2),$$

and hence $M(A) > M(B) \geq M(A_1) > M(A_2)$, which implies $M(A) > \max\{M(A_1), M(A_2)\}$, which contradicts the Averaging Property .

In the second case we would have

$$M(A \cup B_1) \leq M(B_2),\ M(B_2) < M(B_1),\ M(A) > M(B) = M(B_1 \cup B_2),$$

which implies

$$\min\{M(A), M(B_1)\} \leq M(B_2) < M(B_1),$$

and then $\min\{M(A), M(B_1)\} = M(A)$ and $M(A) \leq M(B_2) < M(B_1)$, which contradicts either $M(A) < M(B)$ or the Averaging Property . □

Theorem 3.3. *If $(S_j)_j$ is the partition obtained at the end of the PAVA described above, then $m_i = M(S_{j_i})$ such that $i \in S_{j_i}$ takes the same values given by the min-max formula for the index i.*

Proof. See Appendix A.

3.2 Preparing for a projected subgradient algorithm

The following proposition is crucial for computing the ordered isotonic estimators via a projected subgradient algorithm.

Proposition 3.4. *Let Ψ be the criterion*

$$\Psi(b_1, \ldots, b_{n-1}) \quad = \quad \sum_{i=1}^{n} \Big(\max_{s \leq i} (G_{s,i} \wedge b_s) - y_i \Big)^2 w_i^1 + \sum_{i=1}^{n-1} (b_i - z_i)^2 w_i^2 \qquad (14)$$

which is to be minimized on the convex set

$$\mathcal{I}_{n-1}^{b_n^*} = \{(b_1, \ldots, b_{n-1}) \in \mathbb{R}^{n-1} : \; b_1 \leq b_2 \leq \ldots \leq b_{n-1} \leq b_n^*\}$$

where

$$G_{s,i} = \min_{t \geq i} Av_1(\{s, \ldots, t\}) \;\; \textit{and} \;\; b_n = b_n^* \;\; \textit{in} \;\; G_{n,n} \wedge b_n, \text{(14)}.$$

The criterion Ψ is convex. Furthermore, its unique minimizer $(b_1^{**}, \ldots, b_{n-1}^{**})$ equals $(b_1^*, \ldots, b_{n-1}^*)$.

Proof. Let us write

$$\mathcal{I} = \mathcal{I}_n^\infty = \{a = (a_1, \ldots, a_n) \in \mathbb{R}^n : a_1 \leq \ldots \leq a_n\},$$

$$\mathcal{I}_n^* = \left\{b = (b_1, \ldots, b_n) : (b_1, \ldots, b_{n-1}) \in \mathcal{I}_{n-1}^{b_n^*} \text{ and } b_n = b_n^*\right\}$$

and consider

$$\mathcal{I}_n^b = \{a : a \in \mathcal{I} \text{ and } a \leq b\}$$

for $b \in \mathcal{I}_n^*$.

Now note that the min-max formula in (9) allows us to write

$$\begin{aligned}
&\sum_{j=1}^{n} \Big(\max_{s \leq j} (G_{s,j} \wedge b_s) - y_j \Big)^2 w_j^1 + \sum_{j=1}^{n-1} (b_j - z_j)^2 w_j^2 \\
&= \min_{a \in \mathcal{I}_n^b} \sum_{j=1}^{n} (a_j - y_j)^2 w_j^1 + \sum_{j=1}^{n-1} (b_j - z_j)^2 w_j^2.
\end{aligned}$$

Hence, we have for $b \in \mathcal{I}_n^*$

$$\begin{aligned}
\Psi(b_1, ..., b_{n-1}) &= \min_{a \in \mathcal{I}_n^b} \sum_{j=1}^{n} (a_j - y_j)^2 w_j^1 + \sum_{j=1}^{n-1} (b_j - z_j)^2 w_j^2 \\
&= \sum_{j=1}^{n} (\tilde{a}_j(b) - y_j)^2 w_j^1 + \sum_{j=1}^{n-1} (b_j - z_j)^2 w_j^2
\end{aligned}$$

where $\tilde{a}_j(b) = \max_{s \leq j} (G_{s,j} \wedge b_s)$ is the j-th component of the minimizer of the function $\sum_{j=1}^{n} (a_j - y_j)^2 w_j^1$ in $\mathcal{I}_n^b$. Let $\lambda \in [0, 1]$, and b and b' in $\mathcal{I}_n^*$. By definition of $\mathcal{I}_n^b$ and $\mathcal{I}_n^{b'}$, we have that

$$\lambda \, \tilde{a}(b) + (1 - \lambda) \, \tilde{a}(b') \leq \lambda \, b + (1 - \lambda) \, b'$$

and hence

$$\begin{aligned}
&\sum_{j=1}^{n} \Big(\tilde{a}_j(\lambda\, b + (1-\lambda)\, b') - y_j\Big)^2 w_j^1 \\
&\leq \sum_{j=1}^{n} \Big(\lambda\, \tilde{a}(b) + (1-\lambda)\, \tilde{a}(b') - y_j\Big)^2 w_j^1 \\
&\leq \lambda \sum_{j=1}^{n} \Big(\tilde{a}_j(b) - y_j\Big)^2 w_j^1 + (1-\lambda) \sum_{j=1}^{n} \Big(\tilde{a}_j(b') - y_j\Big)^2 w_j^1.
\end{aligned}$$

This shows convexity of the first term of Ψ. Convexity of Ψ now follows from convexity of the function $\sum_{j=1}^{n-1}(b_j - z_j)^2 w_j^2$ and the fact that the sum of two convex functions defined on the same domain is also convex. □

The idea behind considering the convex functional Ψ is to reduce the dimensionality of the problem as well as the number of constraints (from $3n-2$ to $n-1$ constraints). Once Ψ is minimized; i.e, the isotonic estimate b^* is computed, a^* can be obtained using the min-max formula given in (9). However, the convex functional Ψ is not continuously differentiable, hence the need for an optimization algorithm that uses the subgradient instead of the gradient as the latter is not defined everywhere.

3.3 A projected subgradient algorithm to compute $b_1^*, \ldots, b_{n-1}^*$

To minimize the non-smooth convex function Ψ we use a projected subgradient algorithm. Since the gradient does not exist on the entire domain of the function, one has to resort to computation of a subgradient, the analogue of the gradient at points where the latter does not exist. As opposed to classical methods developed for minimizing smooth functions, the procedure of searching for the direction of descent and steplengths is entirely different. The classical reference for subgradient algorithms is Shor (1985). Boyd et al. (2003) provides a nice summary of the topic, including the projected variant. Note that a recent application in statistics of the subgradient algorithms gives now the possibility to compute the log-concave density estimator in high dimensions; see Cule et al. (2008).

The main steps of the algorithm. Now recall that the functional Ψ should be minimized over the $(n-1)-$ dimensional convex set $\mathcal{I}_{n-1}^{b_n^*}$ given in Proposition 3.4. Of course, this is the same as minimizing Ψ over the $n-$ dimensional convex set $\{(b_1, \ldots, b_n) \mid b_1 \leq \ldots \leq b_{n-1}\}$, starting with an initial vector $(b_1^{(0)}, \ldots, b_n^{(0)})$ such that $b_n^{(0)} = b_n^*$ and constraining the $n-$th component of the sub-gradient of Ψ to be equal to 0.

Given a steplength τ_k, the new iterate $b^{k+1} = (b_1^k, \ldots, b_n^k)$ at the $k-$th iteration of a subgradient algorithm is given by

$$v_{k+1} \;=\; b_k - \tau_k D_k,$$

where D_k is the subgradient calculated at the previous iterate; i.e., $D_k = \tilde{\nabla}\Psi(v_k)$ (see Appendix B). However, it may happen that v_{k+1} is not admissible; i.e. $(b_1^{k+1}, \ldots, b_{n-1}^{k+1})$ does not belong to $\mathcal{I}_{n-1}^{b_n^*}$. When this occurs, an $\mathbb{L}_2$ projection of this iterate onto $\mathcal{I}_{n-1}^{b_n^*}$ is performed. This is equivalent to finding the minimizer of

$$\sum_{i=1}^{n}(a_i - b_i^{k+1})^2$$

over the set $\mathcal{I}_n^{b_n^*}$. The latter problem can be solved using the generalized PAVA for bounded isotonic regression as described above.

The computation of the subgradient D_k is described in detail in Appendix B. As for the steplength τ_k, we start the algorithm with a constant steplength. Once a pre-specified number of iterations has been reached we switch to

$$\tau_{k+1} \quad = \quad (h_k^{0.1}\|D_k\|_2)^{-1}$$

where $\gamma_k := h_k^{-0.1}$ is such that $0 \leq \gamma_k \to 0$ as $k \to \infty$ and $\sum_{k=1}^{\infty}\gamma_k = \infty$. Here, $\|\cdot\|_2$ denotes the $\mathbb{L}_2$-norm of a vector in $\mathbb{R}^n$. This combination of constant and non-summable diminishing steplength showed a good performance in our implementation of the algorithm over other classical choices of $(\gamma_k)_k$. Furthermore, convergence is ensured by the following theorem.

Theorem 3.5. ***(Boyd et al. (2003))*** *A subgradient algorithm complemented with least-square projection and using non-summable diminishing steplength yields for any $\eta > 0$ after $k = k(\eta)$ iterations a vector $b^k := (b_1^k, \ldots, b_n^k)$ such that*

$$\min_{i=1,\ldots,k}\Psi(b^i) - \Psi(b^*) \quad \leq \quad \eta,$$

where $b^* = (b_1^*, \ldots, b_n^*)$ is the vector given in Proposition 3.4.

The proof can be found in Boyd et al. (2003) by combining their arguments in Sections 2 and 3. Note that in our implementation we do not keep track of the iterate that yielded the minimal value of Ψ, since we apply a problem-motivated stopping criterion that guarantees us to have reached an iterate that is sufficiently close to $b^* = (b_1^*, \ldots, b_n^*)$.

Choice of stopping rule. Since in subgradient algorithms the convex target functional does not necessarily monotonically decrease with increasing number of iterations, the choice of a suitable stopping criterion is delicate. However, in our specific setting we use the fact that (a^*, b^*) is a fixed point of the operator P defined in (11) where $a^* = P_1(b^*)$; the solution of (1) with upper bound b^*. This motivates iterating the algorithm until the difference of entries of the two vectors b^k and $b_\#^k$ where

$$b_\#^k = P_2 \circ P_1(b^k)$$

is below a pre-specified positive constant δ.

The implementation. The projected subgradient algorithm for the two curve problem as well as the generalized PAVA computing the solution for one curve under the constraints (3) were implemented in `R` (R Development Core Team, 2008). The corresponding package `OrdMonReg` Balabdaoui et al. (2009) is available on CRAN. Note that the data analyzed in Section 3.4 is made available as a dataset in `OrdMonReg`.

To conclude this section on the algorithmic aspects of our work, we would like to mention the work by Beran and Dümbgen (2009) who propose an active set algorithm which can be tailored to solve the problem given in (5) for an arbitrary number of ordered monotone curves. However, Beran and Dümbgen (2009) do not provide an analysis of the structure of the estimated curves such as characterizations and rather put their emphasis on the algorithmic developments of the problem.

3.4 Real data example from mechanical engineering

We would like to estimate the stress-strain curves based on the available experimental data for two different velocity levels (see Figure 1). The expected curves have to be isotonic and ordered. The data consist of 1495 pairs (x_i, y_i) and (y_i, z_i). The values of the measured strain of the material (on the x-axis), are actually defined as $(-)$ the logarithm of the ratio of the current over the initial specimen length. The values are positive and take the maximal value 1, which corresponds to a maximum shortening of 63%.

Furthermore, since the stress measurements for different velocities are not performed exactly at the same strain, the values of the stress have been interpolated at equally spaced values of the strain. As pointed out by a referee, this will induce correlation between the strain data. Even if the strain measurement were not interpolated, having correlated stress measurements is rather inevitable in this particular application because of the data processing procedures associated with the measurement technique (see Shim and Mohr, 2009). The estimation method is however still applicable. When studying statistical properties of the isotonic estimators such as consistency and convergence, the correlation between the data should be of course taken into account.

In such problems, practitioners usually fit parametric models using a trial and error approach in an attempt to capture monotonicity of the stress-strain curves as well as their ordering. The methods used are rather arbitrary and can also be time consuming, hence the need for an alternative estimation approach. Our main goal is to provide those practitioners with a rigorous way for estimating the ordered stress-strain curves.

In Figure 2 (upper plot) we provide the original data (black and gray dots) and the proposed ordered isotonic estimates a^* and b^* as described above. Being step functions, the estimated isotonic curves are non-smooth, a well known drawback of isotonic regression, see among others Wright (1978) and Mukerjee (1988). The latter author pioneered the combination of isotonization followed by kernel smoothing. A thorough asymptotic analysis of

the smoothed isotonized and the isotonic smooth estimators was given by Mammen (1991). Mukerjee (1988, p. 743) shows that monotonicity of the regression function is preserved by the smoothing operation if the used kernel is log-concave. Thus, we define our smoothed ordered monotone estimators by

$$\tilde{a}_h^*(x) \;=\; \frac{\sum_{i=1}^n K_h(x-t)a_i^*}{\sum_{i=1}^n K_h(x-x_i)} \qquad \tilde{b}_h^*(x) \;=\; \frac{\sum_{i=1}^n K_h(x-t)b_i^*}{\sum_{i=1}^n K_h(x-x_i)}$$

for $0 \le x \le 1$. For simplicity, we used the kernel $K_h(x) = \phi(x/h)$ where ϕ is the density function of a standard normal distribution which is clearly log-concave. Figure 2 (lower plot) depicts the smoothed isotonic estimates. We set the bandwidth to $h = 0.1n^{-1/5} \approx 0.023$.

Motivated by estimation of stress-strain curves, an application from mechanical engineering, we consider in this paper weighted Least Squares estimators in the problem of estimating two ordered isotonic regression curves. We provide characterizations of the solution and describe a projected subgradient algorithm which can be used to compute this solution. As a by-product, we show how an adaptation of the well-known PAVA can be used to compute min-max estimators for any set functional satisfying the Averaging Property.

Acknowledgements. The first author would like to thank Cécile Durot for some interesting discussions around the subject. We also thank JongMin Shim for having made the data available to us.

A Proofs

Proof of Theorem 2.2. Suppose that (a^*, b^*) is the solution. For $\epsilon \in (0,1)$, and $(a,b) \in \mathcal{I}_n$ consider the pair $(a^\epsilon, b^\epsilon) \in \mathbb{R}^n \times \mathbb{R}^n$ defined as

$$\begin{aligned} a^\epsilon &= a^* + \epsilon(a - a^*) \\ b^\epsilon &= b^* + \epsilon(b - b^*). \end{aligned}$$

For $i \le j \in \{1, ..., n\}$, we have

$$\begin{aligned} a_j^\epsilon - a_i^\epsilon &= (1-\epsilon)(a_j^* - a_i^*) + \epsilon(a_j - a_i) \ge 0 \\ b_j^\epsilon - b_i^\epsilon &= (1-\epsilon)(b_j^* - b_i^*) + \epsilon(b_j - b_i) \ge 0. \end{aligned}$$

Also, for $i \in \{1, ..., n\}$ we have

$$a_i^\epsilon - b_i^\epsilon \;=\; (1-\epsilon)(a_i^* - b_i^*) + \epsilon(a_i - b_i) \le 0.$$

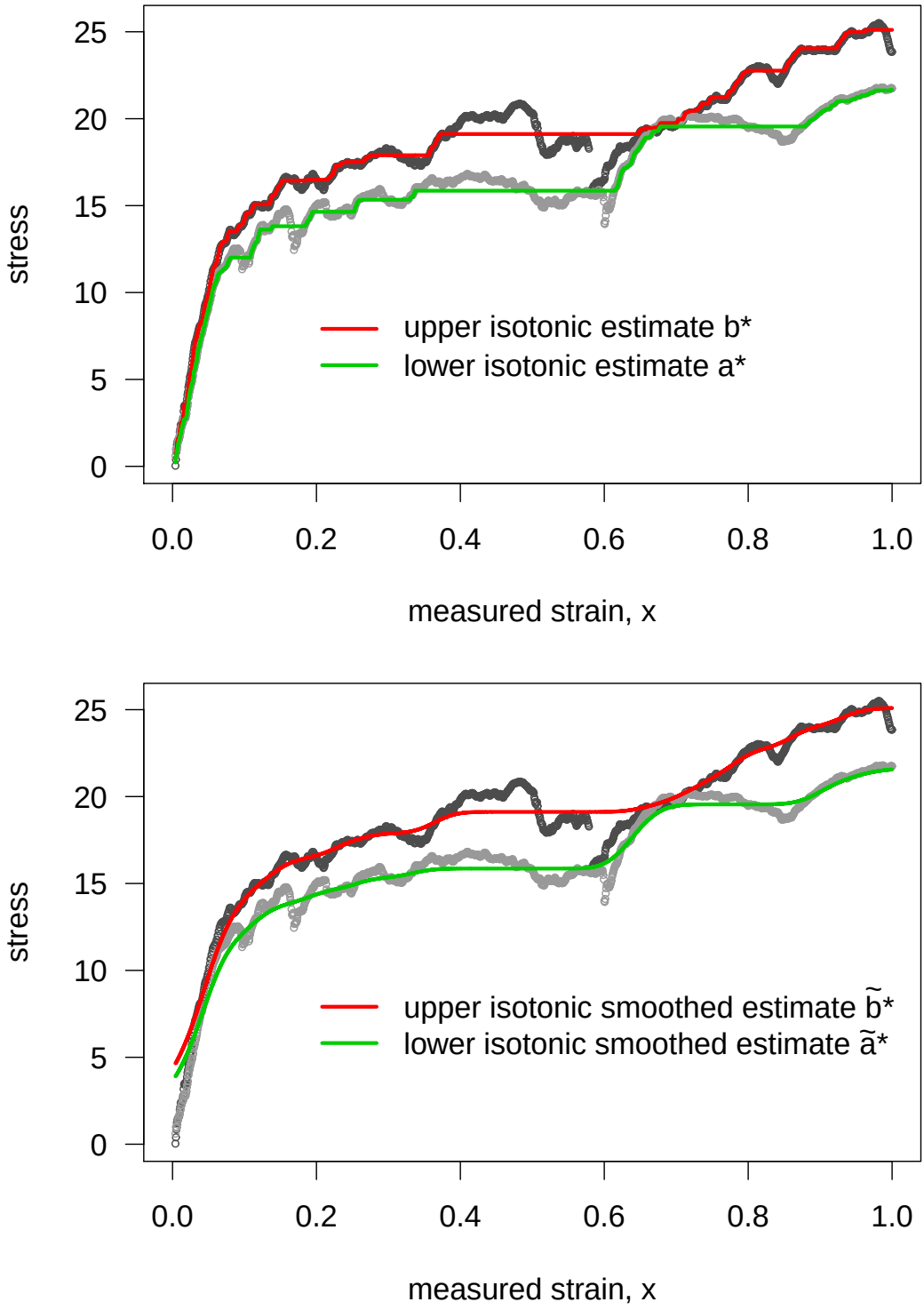

Figure 2: Original observations, isotonic and isotonic smoothed estimates.

Hence, $(a^\epsilon, b^\epsilon) \in \mathcal{I}_n$, and

$$\begin{aligned} 0 &\leq \lim_{\epsilon \searrow 0} \frac{1}{\epsilon}(L_2(a^\epsilon, b^\epsilon) - L_2(a^*, b^*)) \\ &= \sum_{i=1}^n (a_i^* - y_i)(a_i - a_i^*)w_i^1 + \sum_{i=1}^n (b_i^* - z_i)(b_i - b_i^*)w_i^2 \end{aligned}$$

yielding the inequality in (6).

Now consider the vectors a^ϵ and b^ϵ such that for $l = 1, ..., n$

$$\begin{aligned} a_l^\epsilon &= a_l^* + \epsilon\, a_l^*\, 1_{l \in B_{i_j}^1} \\ b_l^\epsilon &= b_l^* \end{aligned}$$

Let $r \leq s \in \{1, ..., n\}$. If $r \notin B_{i_j}^1$ and $s \notin B_{i_j}^1$, then $a_s^\epsilon - a_r^\epsilon = a_s^* - a_r^* \geq 0$. If $r \in B_{i_j}^1$ and $s \notin B_{i_j}^1$, then $a_s^* > a_r^*$ and $a_s^\epsilon - a_r^\epsilon = a_s^* - a_r^* + \epsilon a_s^* > 0$ for $|\epsilon|$ small enough. The same reasoning applies if $r \notin B_{i_j}^1$ and $s \in B_{i_j}^1$. Finally, if $r, s \in B_{i_j}^1$, then $a_s^\epsilon - a_r^\epsilon = 0$.

Now, for $r \in \{1, ..., n\}$, we have $a_r^\epsilon = a_r^* \leq b_r^*$ if $r \notin B_{i_j}^1$. Otherwise, $a_r^\epsilon = a_r^*(1+\epsilon) < b_r^*$ if $|\epsilon|$ is small enough. Hence, $(a^\epsilon, b^\epsilon) \in \mathcal{I}_n$, and

$$\begin{aligned} 0 &= \lim_{\epsilon \searrow 0} \frac{1}{\epsilon}(L_2(a^\epsilon, b^\epsilon) - L_2(a^*, b^*)) \\ &= \sum_{r=1}^n (a_r^* - y_r) 1_{r \in B_{i_j}^1}\, a_r^* w_r^1. \end{aligned}$$

Summing up over all the sets $B_{i_j}^1$ yields the identity in (7). We can prove very similarly the identity in (8).

Conversely, suppose that $(a^*, b^*) \in \mathcal{I}_n$ satisfies the inequality in (6). For any $(a, b) \in \mathcal{I}_n$, we have

$$\begin{aligned} L_2(a, b) - L_2(a^*, b^*) &= \frac{1}{2}\sum_{i=1}^n (a_i - a_i^*)^2 w_i^1 + \frac{1}{2}\sum_{i=1}^n (b_i - b_i^*)^2 w_i^2 \\ &\quad + \sum_{i=1}^n (a_i^* - y_i)(a_i - a_i^*)w_i^1 \\ &\quad + \sum_{i=1}^n (b_i^* - z_i)(b_i - b_i^*)w_i^2 \\ &\geq 0. \end{aligned}$$

We conclude that (a^*, b^*) is the solution of the minimization problem. □

Proof of Proposition 2.3. Let $\epsilon > 0$ and consider $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$\begin{aligned} a_i &= a_i^* - \epsilon\, 1_{i \in \{1,..,t\}},\ t \in \{1, ...n\} \\ b_i &= b_i^* \end{aligned}$$

for $i = 1, ..., n$. For small ϵ, $(a, b) \in \mathcal{I}_n$. Using the characterization in Theorem 2.2, it follows that

$$\sum_{j=1}^{t}(a_j^* - y_j)w_j^1 \leq 0$$

implying that

$$\sum_{j=1}^{t}(a_1^* - y_j)w_j^1 \leq 0, \text{ for } t \in \{1, ...n\}$$

or equivalently

$$a_1^* \leq \min_{t\geq 1} Av_1(\{1, \ldots, t\}).$$

Now, consider $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$ such that

$$\begin{aligned} a_j &= a_j^* - \epsilon 1_{j\in\{1,...,t\}}, \ t \in \{1, ..., n\} \\ b_j &= b_j^* - \epsilon 1_{j\in\{1,...,t'\}}, \ 1 \leq t' \leq t \end{aligned}$$

for $j = 1, ..., n$, with $\epsilon > 0$. For small ϵ, we have that $(a, b) \in \mathcal{I}_2$, and hence

$$\sum_{j=1}^{t}(a_j^* - y_j)w_j^1 + \sum_{j=1}^{t}(b_j^* - z_j)w_j^2 \leq 0.$$

It follows that

$$\sum_{j=1}^{t}(a_1^* - y_j)w_j^1 + \sum_{j=1}^{t'}(a_1^* - z_j)w_j^2 \geq 0,$$

that is

$$a_1^* \leq \min_{1\leq t'\leq t\leq n} \tilde{M}(\{1, \ldots, t\}, \{1, \ldots, t'\}).$$

We conclude that

$$a_1^* \leq \min_{t\geq 1} Av_1(\{1, \ldots, t\}) \wedge \min_{t\geq t'\geq 1} \tilde{M}(\{1, \ldots, t\}, \{1, \ldots, t'\}).$$

Now if $a_1^* < b_1^*$, let $i_1\{1, ..., n\}$ be such that $a_1^* = \ldots = a_{i_1}^*$. Then (a, b) is such that

$$\begin{aligned} a_j &= a_j^* + \epsilon \, 1_{j\in\{1,...,i_1\}} \\ b_j &= b_j^* \end{aligned}$$

for $j = 1, ..., n$ is in $\mathcal{I}_n$ when $|\epsilon|$ is small enough. It follows that

$$Av_1(\{1, \ldots, i_1\}) = a_1^*.$$

If $a_1^* = b_1^*$, and i_1' and i_1'' are such that $a_1^* = ... = a_{i_1'}^*$ and $b_1^* = ... = b_{i_1''}^*$, then (a, b) such that

$$\begin{aligned} a_j &= a_j^* + \epsilon\, 1_{j\in\{1,...,i_1'\}} \\ b_j &= b_j^* + \epsilon\, 1_{j\in\{1,...,i_1''\}} \end{aligned}$$

for $j = 1, ..., n$ is in $\mathcal{I}_n$ for $|\epsilon|$ small enough. Hence,

$$a_1^* = \tilde{M}(\{1, \ldots, i_1'\}, \{1, \ldots, i_1''\}).$$

(note that $i_1'' \leq i_1'$). Therefore,

$$a_1^* = \min_{t\geq 1} Av_1(\{1, \ldots, t\}) \wedge \max_{t\geq t'\geq 1} \tilde{M}(\{1, \ldots, t\}, \{1, \ldots, t'\}).$$

The expression of b_1^* follows easily by replacing respectively y_i and z_i by $-z_{n-i+1}$ and $-y_{n-i+1}$ for $i = 1, \ldots, n$. □

Proof of Theorem 3.3. Consider $a \in \mathbb{R}^n$ given by

$$a_i = \max_{s\leq i} \min_{t\geq i} M(\{s, \ldots, t\})$$

and also the subdivision into subsets $S_j = \{i_{j-1} + 1, \ldots, i_j\}$ obtained by the PAVA. Let us denote by G^- (resp. G^+) the grid set of indices which correspond to points at the beginning (resp. end) of those subsets; i.e. of the form $i_j + 1$ (resp. i_j).

We obviously have

$$a_i \leq \max_{s\leq i} \min_{t\geq i,\, t\in G^+} M(\{s, \ldots, t\}).$$

Then, consider $s \notin G^-$. This means that we have a set $\{s, \ldots, t\}$ of the form $B \cup C$, C being a union of subsets in the subdivision and B a right subset of a set of the partition of the form $A \cup B$. We want to prove that $M(\{s, \ldots, t\}) = M(B \cup C)$ is either smaller than $M(C)$ or $M(A \cup B \cup C)$. Suppose this is not the case. Then we would have

$$M(B \cup C) > M(C),\ M(B \cup C) > M(A \cup B \cup C),\ M(A) > M(B),$$

where the last inequality is implied by the second property in Theorem 3.2. Yet, the second inequality, together with the Averaging Property , implies that $M(A) < M(B \cup C)$. In the end we get

$$M(B \cup C) > M(C),\ M(B \cup C) > M(A) > M(B),$$

which contradicts the Averaging Property .

We conclude that $M(\{s, \ldots, t\})$ is smaller than the value of M at a set which is a union of sets of the subdivision; i.e. either $A \cup B \cup C$ or C itself. But on sets of this kind it is obvious, by the Averaging Property , that M is smaller than the value m_t, since this is the

maximal value of M on the intervals composing such a set (this is a consequence of M being increasing). Hence, $M(\{x_s, \dots, x_t\}) \leq m_t$, implying that

$$a_i \leq \max_{s \leq i} \min_{t \geq i,\, t \in G^+} m_t = m_i.$$

The opposite inequality is obtained exactly in a symmetric way (first take $s \in G^-$, then prove that $M(\{x_s, \dots, x_t\})$ is larger than the value of M on a union of intervals). □

B Computing the subgradient

Computing the subgradient of Ψ on a dense set. Consider the set

$$\begin{aligned} D \;=\; \Big\{ & b = (b_1, \dots, b_{n-1}) \in \mathbb{R}^{n-1} : b_i \neq b_j \;\forall\, i \neq j, \\ & \text{and } b_{i'} \neq G_{s,j'} \;\forall\, 1 \leq i' \leq n-1, 1 \leq s \leq n-1, 1 \leq j' \leq n \Big\}. \end{aligned}$$

We denote by $(e^1, \dots, e^{n-1})$ the canonical basis of $\mathbb{R}^{n-1}$. The set D is a dense open subset of $\mathbb{R}^{n-1}$ where the function Ψ is differentiable. Actually, for a fixed $b \in D$, in the explicit formula for Ψ there is no ex-aequo (up to possible equalities between the $G_{i,s}$ terms). The same will be true in a neighborhood of b. For each value of $i \in \{1, \dots, n\}$, we define the function

$$\Psi_i = \Big(\max_{s \leq i} (G_{s,i} \wedge b_s) - y_i \Big)^2 w_i^1.$$

Let us first consider $i \in \{1, \dots, n-1\}$. We define $\{s_{i_1}, \dots, s_{i_k}\}$ to be the set of indices s where $\max_{s \leq i}(G_{s,i} \wedge b_s)$ is attained.

If $k = 1$, then $G_{s_{i_1},i} \wedge b_{s_1} > G_{s,i} \wedge b_s$ for all $s \in \{1, \dots, i\} \setminus \{s_{i_1}\}$. This implies that the same strict inequalities will be true in a neighborhood of b and hence there are two cases: either the function is locally constant or the square of an affine function. Hence,

- If $b_{s_{i_1}} > G_{s_{i_1},i}$, then $\nabla \Psi_i(b) = 0$.
- If $b_{s_{i_1}} < G_{s_{i_1},i}$, then $\nabla \Psi_i(b) = 2\Big((G_{s_{i_1},i} \wedge b_{s_{i_1}}) - y_i \Big)\, w_i^1\, e^{s_{i_1}}$.

Now if $k \geq 2$, then this implies that only $G_{s_{i_j},i}, j = 1, \dots, k$ can be equal (by definition of the set D), and hence the function is locally constant. Therefore, $\nabla \Psi_i(b) = 0$.

For $i = n$, the calculation also requires distinction between the cases $k = 1$ and $k \geq 2$. Thus, if $k = 1$ and the maximum $\max_{s \leq n}(G_{s,n} \wedge b_s)$ is attained at $s_{i_1} \neq n$, then

- If $b_{s_{i_1}} > G_{s_{i_1},n}$, then $\nabla \Psi_i(b) = 0$.
- If $b_{s_{i_1}} < G_{s_{i_1},n}$, then $\nabla \Psi_n(b) = 2\Big((G_{s_{i_1},n} \wedge b_{s_{i_1}}) - y_n \Big)\, w_n^1\, e^{s_{i_1}}$.

If $k = 1$ and $s_{i_1} = n$ (in this case $b_n = b_n^*$ is known) or $k \geq 2$, then $\nabla\Psi_n(b) = 0$. Now the gradient $\nabla\Psi(b)$ is given by

$$\nabla\Psi(b) = \sum_{i=1}^{n} \nabla\Psi_i(b) + 2\sum_{i=1}^{n-1}(b_i - z_i)w_i^2 e^i.$$

Calculating the subgradient of Ψ at any point. Take now any point $b \in \mathbb{R}^{n-1}$ which does not necessarily belong to D. We want to approximate b by points of D in the perspective of using the following property: If Ψ is convex, $p_\varepsilon \to p$, $\gamma_\varepsilon \to \gamma$ as $\epsilon \to 0$, and $\gamma_\varepsilon \in \partial\Psi(p_\varepsilon)$, then $\gamma \in \partial\Psi(p)$. This is useful when we only want to find one element of the subdifferential at a given point and we already know the gradients at nearby points.

We use the following approximation:

$$b_\varepsilon = b + \varepsilon u, \text{ where } u = (1, 2, \ldots, i, \ldots, n-1).$$

We claim that b_ε may belong to the complement of D for a finite number of values ε at most. Indeed, for any pair (i, j) with $i \neq j$, the equality $b_i + i\varepsilon = b_j + j\varepsilon$ is satisfied for a unique value of ε, and for any i, i' and s, the same thing holds true for the equality $G_{i,s} = b_{i'} + \varepsilon i'$. Hence, there exists $\varepsilon_0 > 0$ such that for $\varepsilon \in]0, \varepsilon_0[$, we have $b_\varepsilon \in D$, where the expression of the gradient is fully known by our calculations above.

We can act as follows: Take b and fix $i \leq n-1$. For any $s \leq i$, determine which one is minimal among $G_{i,s}$ and b_s. In case of equality, priority will be given to $G_{i,s}$ since in the approximation with b_ε, the value of $G_{i,s}$ would be smaller than $b_s + \epsilon s$. This way we classify the indices in two categories: The G-type and b-type. Next, look at all the indices $s_1, \ldots, s_k$ realizing the minimum of $G_{i,s} \vee b_s$. If among $s_1, \ldots, s_k$ there are some which are of the b-type, this would imply that in the approximation with b_ε, those indices will yield even a higher value for $G_{i,s_j} \vee (b_{s_j} + \varepsilon s_j)$. In particular the maximal one will correspond to the largest b-type index since it is the one where the coordinate is increased the most in the approximation. Due to the fact that b_n^* is fixed, we adopt, for $i = n$, the convention that the index $s = n$ is of the G-type when $G_{n,n} \wedge b_n^*$ is maximal. Thus, we can define the vector

$$\tilde{\nabla}\Psi_i(b) \quad = \quad 2((G_{s_{i_m},i} \wedge b_{s_{i_m}}) - y_i)\; w_i^1\; e^{s_{i_m}} \text{ or } 0,$$

where the index s_{i_m} is the largest index of b-type such that $G_{i,s} \wedge b_s$ is maximal (note that s_{i_m} is always $\leq n-1$). If no such index exists (i.e. if the maximal ones are all of G-type), then this is the case where the vector equals 0. Now consider

$$\tilde{\nabla}\Psi(b) \quad = \quad \sum_{i=1}^{n} \tilde{\nabla}\Psi_i(b) + 2\sum_{i=1}^{n-1}(b_i - z_i)\; w_i^2\; e^i.$$

This vector belongs to $\partial\Psi(b)$ by approximation and closedness of the subdifferential.

Note that we would have obtained another element of the subdifferential if we had fixed a different order of priority on the coordinates of b; for instance the first index instead of the last one (if $u = (1, 2, \ldots, i, \ldots n-1)$ was replaced with $(n-1, \ldots, 2, 1)$). We could also have decreased (instead of increased) the components, thus giving priority to b_s instead of $G_{i,s}$ in the minimum $G_{i,s} \wedge b_s$. In that case, we would have obtained 0 for the subgradient of Ψ_i as soon as one of the components realizing the maximum was of the G-type.

References

BALABDAOUI, F., RUFIBACH, K. and SANTAMBROGIO, F. (2009). *OrdMonReg: Compute least squares estimates of one bounded or two ordered isotonic regression curves*. R package version 1.0.1.

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*. John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

BERAN, R. and DÜMBGEN, L. (2009). Least squares and shrinkage estimation under bimonotonicity constraints. *Statistics and Computing, to appear* .

BOYD, S., XIAO, L. and MUTAPCIR, A. (2003). Subgradient methods. Lecture Notes, Stanford University.
URL `http://www.stanford.edu/class/ee392o/subgrad_method.pdf`

BRUNK, H. D. (1958). On the estimation of parameters restricted by inequalities. *Ann. Math. Statist.* **29** 437–454.

BRUNK, H. D., FRANCK, W. E., HANSON, D. L. and HOGG, R. V. (1966). Maximum likelihood estimation of the distributions of two stochastically ordered random variables. *J. Amer. Statist. Assoc.* **61** 1067–1080.

CHAKRAVARTI, N. (1989). Bounded isotonic median regression. *Comput. Statist. Data Anal.* **8** 135–142.

CULE, M., SAMWORTH, R. and STEWART, M. (2008). Maximum likelihood estimation of a multidimensional log-concave density.
URL `http://www.citebase.org/abstract?id=oai:arXiv.org:0804.3989`

DYKSTRA, R. L. (1982). Maximum likelihood estimation of the survival functions of stochastically ordered random variables. *J. Amer. Statist. Assoc.* **77** 621–628.

FELTZ, C. J. and DYKSTRA, R. L. (1985). Maximum likelihood estimation of the survival functions of N stochastically ordered random variables. *J. Amer. Statist. Assoc.* **80** 1012–1019.

LEURGANS, S. (1981). The Cauchy mean value property and linear functions of order statistics. *Ann. Statist.* **9** 905–908.

MAMMEN, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19** 724–740.

MUKERJEE, H. (1988). Monotone nonparameteric regression. *Ann. Statist.* **16** 741–750.

PRÆSTGAARD, J. T. and HUANG, J. (1996). Asymptotic theory for nonparametric estimation of survival curves under order restrictions. *Ann. Statist.* **24** 1679–1716.

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. and WALTMAN, P. (1968). On estimating monotone parameters. *Ann. Math. Statist* **39** 1030–1039.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Ltd., Chichester.

SHIM, J. and MOHR, D. (2009). Using split hopkinson pressure bars to perform large strain compression tests on polyurea at low, intermediate and high strain rates. *International Journal of Impact Engineering* **36** 1116 – 1127.

SHOR, N. (1985). *Minimization Methods for Non-Differentiable Functions*. Springer, Berlin.

VAN EEDEN, C. (1957a). Maximum likelihood estimation of partially or completely ordered parameters. I. *Nederl. Akad. Wetensch. Proc. Ser. A.* **60** = *Indag. Math.* **19** 128–136.

VAN EEDEN, C. (1957b). Maximum likelihood estimation of partially or completely ordered parameters. II. *Nederl. Akad. Wetensch. Proc. Ser. A.* **60** = *Indag. Math.* **19** 201–211.

WRIGHT, F. T. (1978). Estimating strictly increasing regression functions. *Journal of the American Statistical Association* **73** 636–639.
URL `http://www.jstor.org/stable/2286615`