Methodology and Computing in Applied Probability (2022) 24:2633–2645
https://doi.org/10.1007/s11009-022-09937-2

Accelerating the Pool-Adjacent-Violators Algorithm for Isotonic Distributional Regression

Alexander Henzi[1] · **Alexandre Mösching**[2] · **Lutz Dümbgen**[1]

Received: 22 July 2021 / Revised: 3 December 2021 / Accepted: 20 January 2022 /
Published online: 31 March 2022

Abstract
In the context of estimating stochastically ordered distribution functions, the pool-adjacent-violators algorithm (PAVA) can be modified such that the computation times are reduced substantially. This is achieved by studying the dependence of antitonic weighted least squares fits on the response vector to be approximated.

Keywords Monotone regression · Sequential computation · Weighted least squares

AMS 2000 Subject Classifications 62G08 · 62G30 · 62-08

1 Introduction

Let $\mathcal{X}$ be a set equipped with a binary relation $\preceq$, for instance, some partial order. The general problem is as follows: For $m \geq 2$ pairs $(x_1, z_1), \ldots, (x_m, z_m) \in \mathcal{X} \times \mathbb{R}$ and weights $w_1, \ldots, w_m > 0$, let

$$A(\boldsymbol{z}) \ := \arg\min_{\boldsymbol{f} \in \mathbb{R}^m_{\downarrow, \boldsymbol{x}}} \sum_{j=1}^{m} w_j (z_j - f_j)^2, \tag{1}$$

where

$$\mathbb{R}^m_{\downarrow, \boldsymbol{x}} \ := \ \{\boldsymbol{f} \in \mathbb{R}^m : x_i \preceq x_j \text{ implies that } f_i \geq f_j\}.$$

✉ Alexander Henzi
alexander.henzi@stat.unibe.ch

Alexandre Mösching
alexandre.moesching@uni-goettingen.de

Lutz Dümbgen
duembgen@stat.unibe.ch

[1] University of Bern, Bern, Switzerland

[2] Georg-August-University of Göttingen, Göttingen, Germany

Suppose that $z^{(0)}, z^{(1)}, \dots, z^{(n)}$ are vectors in $\mathbb{R}^m$ such that for $1 \le t \le n$, the two vectors $z^{(t-1)}$ and $z^{(t)}$ differ only in a few components, and our task is to compute all antitonic (i.e. monotone decreasing) approximations $A(z^{(0)}), A(z^{(1)}), \dots, A(z^{(n)})$. We show that $A(z^{(t)})$ can be computed efficiently, provided we know already $A(z^{(t-1)})$. Briefly speaking, this is achieved by noticing that $A(z^{(t-1)})$ and $A(z^{(t)})$ share some identical components, and that the remaining components of $A(z^{(t)})$ can be determined directly from $A(z^{(t-1)})$ and $z^{(t)}$ with only a few operations.

The efficient computation of a sequence of antitonic approximations appears naturally in the context of isotonic distributional regression, see Henzi et al. (2021), Mösching and Dümbgen (2020) and Jordan et al. (2021). There, one observes random pairs (X_1, Y_1), $(X_2, Y_2), \dots, (X_n, Y_n)$ in $\mathcal{X} \times \mathbb{R}$ such that, conditional on $(X_i)_{i=1}^n$, the random variables $Y_1, Y_2, \dots, Y_n$ are independent with distribution functions $F_{X_1}, F_{X_2}, \dots, F_{X_n}$, where $(F_x)_{x \in \mathcal{X}}$ is an unknown family of distribution functions. Then the goal is to estimate the latter family under the sole assumption that $F_x \ge F_{x'}$ pointwise whenever $x \preceq x'$. This notion of ordering of distributions is known as stochastic ordering, or first order stochastic dominance. This isotonic distributional regression leads to the aforementioned least squares problem, where $x_1, \dots, x_m$ denote the different elements of $\{X_1, X_2, \dots, X_n\}$, and $z^{(t)}$ has components

$$z_j^{(t)} := w_j^{-1} \sum_{i \,:\, X_i = x_j} 1_{[Y_i \le Y_{(t)}]}$$

with $w_j := \#\{i \le n : X_i = x_j\}$, $Y_{(0)} := -\infty$ and $Y_{(t)}$ is the t-th order statistic of the sample $\{Y_1, Y_2, \dots, Y_n\}$.

Section 2 provides some facts about monotone least squares which are useful for the present task. For a complete account and derivations, we refer to Barlow et al. (1972) and Robertson et al. (1988). Then it is shown in Section 3 how to turn this into an efficient computation scheme in case of a total order $\preceq$. Finally, we discuss the specific application to isotonic distributional regression, and provide numerical experiments which show that computation times of the naive approach are decreased substantially with our procedure.

2 Some Facts About Antitonic Least Squares Estimation

Since the sum on the right hand side of (1) is a strictly convex and coercive function of $\boldsymbol{f} \in \mathbb{R}^m$, and since $\mathbb{R}^m_{\downarrow, \boldsymbol{x}}$ is a closed and convex set, $A(z)$ is well-defined. It possesses several well-known characterizations, two of which are particularly useful for our considerations.

The first characterization uses local weighted averages. Let us first introduce some notations. In this article, upper, lower and level sets are seen as subsets of $\{1, \dots, m\}$ inheriting the structure of $(\mathcal{X}, \preceq)$. More precisely, a set $U \subset \{1, \dots, m\}$ is an upper set if $i \in U$ and $x_i \preceq x_j$ imply that $j \in U$. A set $L \subset \{1, \dots, m\}$ is a lower set if $j \in L$ and $x_i \preceq x_j$ imply that $i \in L$. The families of all upper and all lower sets are denoted by $\mathcal{U}$ and $\mathcal{L}$, respectively. For a non-empty set $S \subset \{1, \dots, m\}$, its weight and the weighted average of z over S are respectively defined as

$$w_S := \sum_{j \in S} w_j \quad \text{and} \quad M_S(z) := w_S^{-1} \sum_{j \in S} w_j z_j.$$

Characterization I For any index $1 \le j \le m$,

$$A_j(z) = \min_{U\in\mathcal{U}:\, j\in U} \max_{L\in\mathcal{L}:\, j\in L} M_{U\cap L}(z) = \max_{L\in\mathcal{L}:\, j\in L} \min_{U\in\mathcal{U}:\, j\in U} M_{L\cap U}(z).$$

For all vectors $\boldsymbol{f} \in \mathbb{R}^m$, numbers $\xi \in \mathbb{R}$ and relations $\ltimes$ in $\{<, \leq, =, \geq, >\}$, let

$$[\boldsymbol{f} \ltimes \xi] := \{j \in \{1, \ldots, m\} : f_j \ltimes \xi\}.$$

For example, the family of sets $[\boldsymbol{f} = \xi]$ indexed by $\xi \in \{f_1, \ldots, f_m\}$ yields a partition of $\{1, \ldots, m\}$ such that two indices i and j belong to the same block if and only if $f_i = f_j$. In case of $\boldsymbol{f} \in \mathbb{R}^m_{\downarrow \boldsymbol{x}}$, $[\boldsymbol{f} < \xi]$ and $[\boldsymbol{f} \leq \xi]$ are upper sets, whereas $[\boldsymbol{f} > \xi]$ and $[\boldsymbol{f} \geq \xi]$ are lower sets.

Characterization II A vector $\boldsymbol{f} \in \mathbb{R}^m_{\downarrow \boldsymbol{x}}$ equals $A(z)$ if and only if for any number $\xi \in \{f_1, \ldots, f_m\}$,

$$M_{U\cap[\boldsymbol{f}=\xi]}(z) \geq \xi \quad \text{for } U \in \mathcal{U} \text{ such that } U \cap [\boldsymbol{f} = \xi] \neq \emptyset, \tag{2}$$

$$M_{L\cap[\boldsymbol{f}=\xi]}(z) \leq \xi \quad \text{for } L \in \mathcal{L} \text{ such that } L \cap [\boldsymbol{f} = \xi] \neq \emptyset. \tag{3}$$

In particular, $\xi = M_{[\boldsymbol{f}=\xi]}(z)$.

One possible reference for Characterizations I and II is Domínguez-Menchero and González-Rodríguez (2007). They treat the case of a quasi-order $\preceq$ and more general target functions $\sum_{j=1}^m h_j(f_j)$ to be minimized over $\boldsymbol{f} \in \mathbb{R}^m_{\downarrow \boldsymbol{x}}$. For the present setting with an arbitrary binary relation $\preceq$ and weighted least squares, a relatively short and self-contained derivation of these two characterizations is available from the authors upon request.

The next lemma summarizes some facts about changes in $A(z)$ if some components of z are increased.

Lemma 2.1 *Let $z, \tilde{z} \in \mathbb{R}^m$ such that $\tilde{z} \geq z$ component-wise. Then the following conclusions hold true for $\boldsymbol{f} := A(z)$, $\tilde{\boldsymbol{f}} := A(\tilde{z})$ and $K := \{k : \tilde{z}_k > z_k\}$:*

(i) $\boldsymbol{f} \leq \tilde{\boldsymbol{f}}$ *component-wise.*
(ii) $\tilde{f}_i = f_i$ *whenever* $f_i < \min_{k\in K} f_k$.
(iii) $\tilde{f}_i = f_i$ *whenever* $\tilde{f}_i > \max_{k\in K} \tilde{f}_k$.
(iv) $\tilde{f}_i = \tilde{f}_j$ *whenever* $f_i = f_j$ *and* $x_i, x_j \preceq x_k$ *for all* $k \in K$.

Figure 1 illustrates the statements of Lemma 2.1 on $\mathbb{R}^2$ equipped with the component-wise order in case of $K = \{j_o\}$. The colored areas show level sets of a hypothetical antitonic regression $\boldsymbol{f}$, and x_{j_o} is the point where $\tilde{z}_{j_o} > z_{j_o}$. By part (ii) of Lemma 2.1, we know that $\tilde{f}_i = f_i$ if $f_i < f_{j_o}$, so the values of $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$ are equal on the orange and yellow regions in the top right corner, which is indicated by saturated colors. Furthermore, when passing from z to $\tilde{z}$, the slightly transparent pink, blue and green level sets on the bottom left (including the point x_{j_o}) can only be merged, but never be split. This follows from part (iv) of Lemma 2.1. Finally, for all points in the faded pink, blue and green areas, there is no statement about the behavior of the antitonic regression when passing from z to $\tilde{z}$.

Proof of Lemma 2.1 Part (i) is a direct consequence of Characterization I.

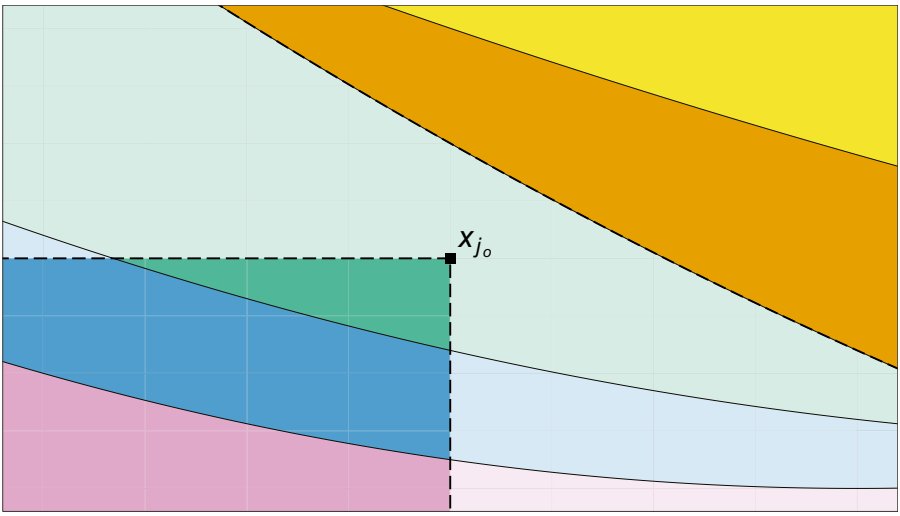

Fig. 1 Illustration of the statements of Lemma 2.1 on $\mathbb{R}^2$

As to part (ii), if $f_i < \min_{k\in K} f_k$, then $K \subset [\boldsymbol{f} > f_i]$, whence

$$\begin{aligned}
\tilde{f}_i &= \max_{L\in\mathcal{L}:\, i\in L} \min_{U\in\mathcal{U}:\, i\in U} M_{L\cap U}(\tilde{\boldsymbol{z}}) && \text{(Char. I)}\\
&\le \max_{L\in\mathcal{L}:\, i\in L} M_{L\cap[\boldsymbol{f}\le f_i]}(\tilde{\boldsymbol{z}}) && (i \in [\boldsymbol{f} \le f_i] \in \mathcal{U})\\
&= \max_{L\in\mathcal{L}:\, i\in L} M_{L\cap[\boldsymbol{f}\le f_i]}(\boldsymbol{z}) && (K \cap [\boldsymbol{f} \le f_i] = \emptyset)\\
&= \max_{L\in\mathcal{L}:\, i\in L} \sum_{\xi\le f_i:\, L\cap[\boldsymbol{f}=\xi]\neq\emptyset} \frac{w_{L\cap[\boldsymbol{f}=\xi]}}{w_{L\cap[\boldsymbol{f}\le f_i]}} M_{L\cap[\boldsymbol{f}=\xi]}(\boldsymbol{z})\\
&\le \max_{L\in\mathcal{L}:\, i\in L} \sum_{\xi\le f_i:\, L\cap[\boldsymbol{f}=\xi]\neq\emptyset} \frac{w_{L\cap[\boldsymbol{f}=\xi]}}{w_{L\cap[\boldsymbol{f}\le f_i]}}\, \xi && \text{(Char. II)}\\
&\le f_i.
\end{aligned}$$

This inequality and part (i) show that $\tilde{f}_i = f_i$.
Part (iii) is proved analogously. If $\tilde{f}_i > \max_{k\in K} \tilde{f}_k$, then $K \subset [\tilde{\boldsymbol{f}} < \tilde{f}_i]$, whence

$$\begin{aligned}
f_i &= \min_{U\in\mathcal{U}:\, i\in U} \max_{L\in\mathcal{L}:\, i\in L} M_{U\cap L}(\boldsymbol{z}) && \text{(Char. I)}\\
&\ge \min_{U\in\mathcal{U}:\, i\in U} M_{U\cap[\tilde{\boldsymbol{f}}\ge\tilde{f}_i]}(\boldsymbol{z}) && (i \in [\tilde{\boldsymbol{f}} \ge \tilde{f}_i] \in \mathcal{L})\\
&= \min_{U\in\mathcal{U}:\, i\in U} M_{U\cap[\tilde{\boldsymbol{f}}\ge\tilde{f}_i]}(\tilde{\boldsymbol{z}}) && (K \cap [\tilde{\boldsymbol{f}} \ge \tilde{f}_i] = \emptyset)\\
&= \min_{U\in\mathcal{U}:\, i\in U} \sum_{\xi\ge\tilde{f}_i:\, U\cap[\tilde{\boldsymbol{f}}=\xi]\neq\emptyset} \frac{w_{U\cap[\tilde{\boldsymbol{f}}=\xi]}}{w_{U\cap[\tilde{\boldsymbol{f}}\ge\tilde{f}_i]}} M_{U\cap[\tilde{\boldsymbol{f}}=\xi]}(\tilde{\boldsymbol{z}})\\
&\ge \min_{U\in\mathcal{U}:\, i\in U} \sum_{\xi\ge\tilde{f}_i:\, U\cap[\tilde{\boldsymbol{f}}=\xi]\neq\emptyset} \frac{w_{U\cap[\tilde{\boldsymbol{f}}=\xi]}}{w_{U\cap[\tilde{\boldsymbol{f}}\le\tilde{f}_i]}}\, \xi && \text{(Char. II)}\\
&\ge \tilde{f}_i.
\end{aligned}$$

This inequality and part (i) show that $\tilde{f}_i = f_i$.

Part (iv) follows directly from parts (i) and (iii). Let i and j be different indices such that $f_i = f_j$ and $x_i, x_j \preceq x_k$ for all $k \in K$. It follows from $\tilde{\boldsymbol{f}} \in \mathbb{R}^m_{\downarrow \boldsymbol{x}}$ that $\tilde{f}_i, \tilde{f}_j \geq \max_{k \in K} \tilde{f}_k$. Consequently, if $\tilde{f}_j > \tilde{f}_i$, then $\tilde{f}_j > \max_{k \in K} \tilde{f}_k$, so parts (i) and (iii) would imply that

$$\tilde{f}_i \;\geq\; f_i = f_j \;=\; \tilde{f}_j,$$

contradicting $\tilde{f}_j > \tilde{f}_i$.

The Special Case of a Total Order If one replaces the binary relation $\preceq$ by a total order $\leq$ on $\mathcal{X}$, as for example in the case of the usual total order on a subset of $\mathbb{R}$, the conclusions of Lemma 2.1 take a simpler form. In case of a total order, we assume that the covariates are ordered as follows

$$x_1 \;\leq\; x_2 \;\leq\; \cdots \;\leq\; x_m,$$

so that $i \leq j$ implies that $x_i \leq x_j$, while $x_i < x_j$ implies that $i < j$.

Corollary 2.2 *Let $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathbb{R}^m$ such that $\boldsymbol{z} \leq \tilde{\boldsymbol{z}}$ component-wise. Then the following conclusions hold true for $\boldsymbol{f} := A(\boldsymbol{z})$ and $\tilde{\boldsymbol{f}} := A(\tilde{\boldsymbol{z}})$:*

(i) $\boldsymbol{f} \leq \tilde{\boldsymbol{f}}$ *component-wise.*

(ii) *Let $k \in \{1, \ldots, m-1\}$ such that $f_k > f_{k+1}$ and $(\tilde{z}_j)_{j>k} = (z_j)_{j>k}$. Then*

$$(\tilde{f}_j)_{j>k} \;=\; (f_j)_{j>k}.$$

(iii) *Let $k \in \{2, \ldots, m\}$ such that $\tilde{f}_{k-1} > \tilde{f}_k$ and $(\tilde{z}_j)_{j<k} = (z_j)_{j<k}$. Then*

$$(\tilde{f}_j)_{j<k} \;=\; (f_j)_{j<k}.$$

(iv) *Let $k \in \{2, \ldots, m\}$ such that $(\tilde{z}_j)_{j<k} = (z_j)_{j<k}$. Then*

$$\{j < k : \tilde{f}_j > \tilde{f}_{j+1}\} \;\subset\; \{j < k : f_j > f_{j+1}\}.$$

3 A Sequential Algorithm for Total Orders

Lemma 2.1 is potentially useful to accelerate algorithms for isotonic distributional regression with arbitrary partial orders, possibly in conjunction with the recursive partitioning algorithm by Luss and Rosset (2014), but this will require additional research. Now we focus on improvements of the well-known pool-adjacent-violators algorithm (PAVA) for a total order.

3.1 General Considerations

In what follows, we assume that $x_1 < \cdots < x_m$, so $\mathbb{R}^m_{\downarrow \boldsymbol{x}}$ coincides with $\mathbb{R}^m_{\downarrow} = \{\boldsymbol{f} \in \mathbb{R}^m : f_1 \geq \ldots \geq f_m\}$. To understand the different variants of the PAVA, let us recall two basic facts about $A(\boldsymbol{z})$. Let $\mathcal{P} = (P_1, \ldots, P_d)$ be a partition of $\{1, \ldots, m\}$ into blocks $P_s = \{b_{s-1} + 1, \ldots, b_s\}$,

where $0 = b_0 < b_1 < \cdots < b_d = m$, and let $\mathbb{R}^m_{\mathcal{P}}$ be the set of vectors $\boldsymbol{f} \in \mathbb{R}^m$ such that $f_i = f_j$ whenever i, j belong to the same block of $\mathcal{P}$.

Fact 1 Let $r_1 > \cdots > r_d$ be the sorted elements of $\{A_i(\boldsymbol{z}) : 1 \le i \le m\}$, and let $\mathcal{P}$ consist of the blocks $P_s = \{i : A_i(\boldsymbol{z}) = r_s\}$. Then $r_s = M_{P_s}(\boldsymbol{z})$ for $1 \le s \le d$.

Fact 2 Suppose that $A(\boldsymbol{z}) \in \mathbb{R}^m_{\mathcal{P}}$ for a given partition $\mathcal{P}$ with $d \ge 2$ blocks. If $s \in \{1, \ldots, d-1\}$ such that $M_{P_s}(\boldsymbol{z}) \le M_{P_{s+1}}(\boldsymbol{z})$, then $A_i(\boldsymbol{z})$ is constant in $i \in P_s \cup P_{s+1}$. That means, one may replace $\mathcal{P}$ with a coarser partition by pooling P_s and P_{s+1} and still, $A(\boldsymbol{z}) \in \mathbb{R}^m_{\mathcal{P}}$.

Fact 1 is a direct consequence of Characterization II. To verify Fact 2, suppose that $\boldsymbol{f} \in \mathbb{R}^m_{\downarrow} \cap \mathbb{R}^m_{\mathcal{P}}$ such that $f_i = r_s$ for $i \in P_s$, $f_i = r_{s+1}$ for $i \in P_{s+1}$, and $r_s > r_{s+1}$. Now we show that $\boldsymbol{f}$ cannot be equal to $A(\boldsymbol{z})$. For $t \ge 0$ let $\boldsymbol{f}(t) \in \mathbb{R}^m_{\mathcal{P}}$ be given by

$$f_i(t) \;=\; f_i - 1_{[i \in P_s]} t w_{P_s}^{-1} + 1_{[i \in P_{s+1}]} t w_{P_{s+1}}^{-1}.$$

Then $\boldsymbol{f}(0) = \boldsymbol{f}$, and $\boldsymbol{f}(t) \in \mathbb{R}^m_{\downarrow}$ if $t \le (r_s - r_{s+1})/(w_{P_{s+1}}^{-1} + w_{P_s}^{-1})$. But

$$\frac{d}{dt}\Big|_{t=0} \sum_{i=1}^m w_i (f_i(t) - z_i)^2 \;=\; 2(r_{s+1} - r_s) - 2\big(M_{P_{s+1}}(\boldsymbol{z}) - M_{P_s}(\boldsymbol{z})\big) \;<\; 0,$$

so for sufficiently small $t > 0$, $\boldsymbol{f}(t) \in \mathbb{R}^m_{\downarrow}$ and is superior to $\boldsymbol{f}(0)$. Hence $\boldsymbol{f} \ne A(\boldsymbol{z})$.

Facts 1 and 2 indicate already a general PAV strategy to compute $A(\boldsymbol{z})$. One starts with the finest partition $\mathcal{P} = (\{1\}, \ldots, \{m\})$. As long as $\mathcal{P}$ contains two neighboring blocks P_s and P_{s+1} such that $M_{P_s}(\boldsymbol{z}) \ge M_{P_{s+1}}(\boldsymbol{z})$, the partition $\mathcal{P}$ is coarsened by replacing P_s and P_{s+1} with the block $P_s \cup P_{s+1}$.

Standard PAVA Specifically, one works with three tuples: $\mathcal{P} = (P_1, \ldots, P_d)$ is a partition of $\{1, \ldots, b_d\}$ into blocks $P_s = \{b_{s-1} + 1, \ldots, b_s\}$, where $0 = b_0 < b_1 < \cdots < b_d$. The number b_d is running from 1 to m, and the number $d \ge 1$ changes during the algorithm, too. The tuples $\mathcal{W} = (W_1, \ldots, W_d)$ and $\mathcal{M} = (M_1, \ldots, M_d)$ contain the corresponding weights $W_s = w_{P_s}$ and weighted means $M_s = M_{P_s}(\boldsymbol{z})$. Before increasing b_d, the tuples $\mathcal{P}$, $\mathcal{W}$ and $\mathcal{M}$ describe the minimizer of $\sum_{i=1}^{b_d} w_i (f_i - z_i)^2$ over all $\boldsymbol{f} \in \mathbb{R}^{b_d}_{\downarrow}$. Here is the complete algorithm:

Initialization: We set $\mathcal{P} \leftarrow (\{1\})$, $\mathcal{W} \leftarrow (w_1)$, $\mathcal{M} \leftarrow (z_1)$, and $d \leftarrow 1$.
Induction step: If $b_d < m$, we add a new block by setting

$$\mathcal{P} \;\leftarrow\; (\mathcal{P}, \{b_d + 1\}), \quad \mathcal{W} \;\leftarrow\; (\mathcal{W}, w_{b_d+1}), \quad \mathcal{M} \;\leftarrow\; (\mathcal{M}, z_{b_d+1}),$$

and $d \leftarrow d + 1$. Then, while $d > 1$ and $M_{d-1} \le M_d$, we pool the "violators" P_{d-1} and P_d by setting

$$\begin{aligned}
\mathcal{P} &\leftarrow \big((P_j)_{j<d-1}, P_{d-1} \cup P_d\big), \\
\mathcal{M} &\leftarrow \Big((W_j)_{j<d-1}, \frac{W_{d-1} M_{d-1} + W_d M_d}{W_{d-1} + W_d}\Big), \\
\mathcal{W} &\leftarrow \big((W_j)_{j<d-1}, W_{d-1} + W_d\big),
\end{aligned}$$

and $d \leftarrow d-1$.
Finalization: Eventually, $\mathcal{P}$ is a partition of $\{1,\dots,m\}$ into blocks such that $M_1 > \cdots > M_d$ and

$$A_j(z) \;=\; M_s \quad \text{for } j \in P_s \text{ and } 1 \le s \le d.$$

Modified PAVA In our specific applications of the PAVA, we are dealing with vectors z containing larger blocks $\{a,\dots,b\}$ on which $i \mapsto z_i$ is constant. Indeed, in regression settings with continuously distributed covariates and responses, z will always be a $\{0,1\}$-valued vector. Then it is worthwhile to utilize fact 2 and modify the initialization as well as the very beginning of the induction step as follows:

For the initialization, we determine the largest index b_1 such that $z_1 = \cdots = z_{b_1}$ and the corresponding weight W_{P_1} with $P_1 = \{1,\dots,b_1\}$. Then we set $\mathcal{P} \leftarrow (P_1)$, $\mathcal{W} \leftarrow (w_{P_1})$ and $\mathcal{M} \leftarrow (z_{b_1})$, where $P_1 = \{1,\dots,b_1\}$.

At the beginning of the induction step, we determine the largest index $b_{d+1} > b_d$ such that $z_{b_d+1} = \cdots = z_{b_{d+1}}$ and the corresponding weight $W_{P_{d+1}}$ with $P_{d+1} = \{b_d+1,\dots,b_{d+1}\}$. Then we set $\mathcal{P} \leftarrow (\mathcal{P}, P_{d+1})$, $\mathcal{W} \leftarrow (\mathcal{W}, W_{P_{d+1}})$, $\mathcal{M} \leftarrow (\mathcal{M}, z_{b_{d+1}})$, and $d \leftarrow d+1$.

Abridged PAVA Suppose that we have computed $A(z)$ with corresponding tuples $\mathcal{P} = (P_1,\dots,P_d)$, $\mathcal{W} = (W_1,\dots,W_d)$ and $\mathcal{M} = (M_1,\dots,M_d)$ via the PAVA. Now let $\tilde{z} \in \mathbb{R}^m$ such that $\tilde{z}_{j_o} > z_{j_o}$ for one index $j_o \in \{1,\dots,m\}$, while $(\tilde{z}_j)_{j\neq j_o} = (z_j)_{j\neq j_o}$. Let $j_o \in P_{s_o}$ with $s_o \in \{1,\dots,d\}$. By parts (ii) and (iv) of Corollary 2.2, the partition corresponding to $A(\tilde{z})$ will be a coarsening of the partition with the following blocks:

$$P_s \;\text{ for } 1 \le s < s_o, \quad \{b_{s_o-1}+1,\dots,j_o\}, \quad \{j\} \;\text{ for } j_o < j \le b_{s_o}, \quad P_s \;\text{ for } s_o < s \le d.$$

Moreover, $A_i(\tilde{z}) = A_i(z)$ for $i > b_{s_o}$. This allows us to compute $A(\tilde{z})$ as follows, keeping copies of the auxiliary objects for $A(z)$ and indicating this with a superscript z:

Initialization: We determine $s_o \in \{1,\dots,d^z\}$ such that $j_o \in P^z_{s_o}$. Then we set

$$\begin{aligned}
\mathcal{P} &\leftarrow \big((P^z_s)_{s<s_o}, \{b^z_{s_o-1}+1,\dots,j_o\}\big),\\
\mathcal{M} &\leftarrow \big((M^z_s)_{s<s_o}, M_{P_{s_o}}(\tilde{z})\big),\\
\mathcal{W} &\leftarrow \big((W^z_s)_{s<s_o}, w_{P_{s_o}}\big)
\end{aligned}$$

and $d \leftarrow s_o$. While $d > 1$ and $M_{d-1} \le M_d$, we pool the violators P_{d-1} and P_d as in the induction step of PAVA. (This initialization is justified by part (iv) of Corollary 2.2.)

Induction step: If $j_o < b^z_{s_o}$, we run the induction step of PAVA for b_d running from j_o+1 to $b^z_{s_o}$ with $\tilde{z}$ in place of z.
Finalization: If $b^z_{s_o} < m$, we set

$$\begin{aligned}
\mathcal{P} &\leftarrow \big(\mathcal{P}, (P^z_s)_{s_o<s\le d^z}\big),\\
\mathcal{M} &\leftarrow \big(\mathcal{M}, (M^z_s)_{s_o<s\le d^z}\big),\\
\mathcal{W} &\leftarrow \big(\mathcal{W}, (W^z_s)_{s_o<s\le d^z}\big)
\end{aligned}$$

and $d \leftarrow d + d^z - s_o$. The new pair $(\mathcal{P}, \mathcal{M})$ yields the vector $A(\tilde{z})$. This finalization is justified by part (ii) of Corollary 2.2.

Computational Complexity It directly follows from the algorithmic description that when $A(z)$ is available, the abridged PAVA for computing $A(\tilde{z})$ requires not more operations than the standard PAVA. Its computational complexity is therefore at most of order $O(m)$ if $x_1, \dots, x_m$ are already sorted. More precisely, the number of averaging operations in the abridged PAVA is bounded from above by $d^z + (b^z_{s_o} - b^z_{s_o-1})$, where d^z is the partition size of the antitonic regression $A(z)$ and $b^z_{s_o} - b^z_{s_o-1}$ is the number of elements in the set $P^z_{s_o}$ containing the index j_o where the value of z changes. In many practical applications this number is much smaller than m, but in the worst case it may equal exactly m; for example, let $w_i = 1$ and $z_i = m - i$ for $i = 1, \dots, m$, $j_o = m$, and $\tilde{z}_m = m^2$.

Numerical Example We illustrate the previous procedures with two vectors $z, \tilde{z} \in \mathbb{R}^9$ and $\boldsymbol{w} = (1)_{j=1}^9$. Table 1 shows the main steps of the PAVA for z. The first line shows the components of z, the other lines contain the current candidate for $(f_j)_{j=1}^{b_d}$, where $\boldsymbol{f} = A(z)$ eventually, and the current partition $\mathcal{P}$ is indicated by extra vertical bars. Table 2 shows the abridged PAVA for two different vectors $\tilde{z}$.

3.2 Application to Isotonic Distributional Regression

Now we consider a regression framework similar to the one discussed in Mösching and Dümbgen (2020), Henzi et al. (2021) and Jordan et al. (2021). We observe pairs (X_1, Y_1), $(X_2, Y_2), \dots, (X_n, Y_n)$ consisting of numbers $X_i \in \mathcal{X}$ (covariate) and $Y_i \in \mathbb{R}$ (response), where $\mathcal{X}$ is a given real interval. Conditional on $(X_i)_{i=1}^n$, the observations $Y_1, Y_2, \dots, Y_n$ are viewed as independent random variables such that for $x \in \mathcal{X}$ and $y \in \mathbb{R}$,

$$I\!P(Y_i \le y) \;=\; F_x(y) \quad \text{if } X_i = x.$$

Table 1 Running the PAVA for a vector z

z	1	3	2	0	-1	1	1/2	-1	1	
$b_d = 1$	1									d = 1
b_d=2	1	3								d = 2
	2	2								d = 1
b_d=3	2	2	2							d = 2
	2	2	2							d = 1
$b_d = 4$	2	2	2	0						d = 2
$b_d = 5$	2	2	2	0	-1					d = 3
$b_d = 6$	2	2	2	0	-1	1				d = 4
	2	2	2	0	0	0				d = 3
	2	2	2	0	0	0				d = 2
$b_d = 7$	2	2	2	0	0	0	1/2			d = 3
	2	2	2	1/8	1/8	1/8	1/8			d = 2
$b_d = 8$	2	2	2	1/8	1/8	1/8	1/8	-1		d = 3
$b_d = 9$	2	2	2	1/8	1/8	1/8	1/8	-1	1	d = 4
	2	2	2	1/8	1/8	1/8	1/8	0	0	d = 3

Table 2 Running the abridged PAVA for two vectors $\tilde{z} \approx z$

z	1	3	2	0	-1	1	1/2	-1	1	
A(z)	2	2	2	1/8	1/8	1/8	1/8	0	0	
$\tilde{z}$	1	3	2	0	**1**	1	1/2	-1	1	
$b_d = 5$	2	2	2	1/2	1/2					d = 2
$b_d = 6$	2	2	2	1/2	1/2	1				d = 3
	2	2	2	2/3	2/3	2/3				d = 2
$b_d = 7$	2	2	2	2/3	2/3	2/3	1/2			d = 3
$b_d = 9$	2	2	2	2/3	2/3	2/3	1/2	0	0	d = 4
z	1	3	2	0	-1	1	1/2	-1	1	
A(z)	2	2	2	1/8	1/8	1/8	1/8	0	0	
$\tilde{z}$	1	3	2	**2**	-1	1	1/2	-1	1	
$b_d = 4$	2	2	2	2						d = 2
	2	2	2	2						d = 1
$b_d = 5$	2	2	2	2	-1					d = 2
$b_d = 6$	2	2	2	2	-1	1				d = 3
	2	2	2	2	0	0				d = 2
$b_d = 7$	2	2	2	2	0	0	1/2			d = 3
	2	2	2	2	1/6	1/6	1/6			d = 2
$b_d = 9$	2	2	2	2	1/6	1/6	1/6	0	0	d = 3

Here $(F_x)_{x\in\mathcal{X}}$ is an unknown family of distribution functions. We only assume that $F_x(y)$ is non-increasing in $x \in \mathcal{X}$ for any fixed $y \in \mathbb{R}$. That means, the family $(F_x)_{x\in\mathcal{X}}$ is increasing with respect to stochastic order.

Let $x_1 < x_2 < \cdots < x_m$ be the elements of $\{X_1, X_2, \ldots, X_n\}$, and let

$$w_j \ := \ \#\{i : \ X_i = x_j\}, \quad 1 \le j \le m.$$

Then one can estimate $\boldsymbol{F}(y) \ := (F_{x_j}(y))_{j=1}^m$ by

$$\widehat{\boldsymbol{F}}(y) \ := \ A(z(y)),$$

where $z(y)$ has components

$$z_j(y) \ := \ w_j^{-1} \sum_{i:\, X_i = x_j} 1_{[Y_i \le y]}, \quad 1 \le j \le m.$$

Suppose we have rearranged the observations such that $Y_1 \le Y_2 \le \cdots \le Y_n$. Let $z^{(0)} := \mathbf{0}$ and

$$z^{(t)} \ := \ \Big(w_j^{-1} \sum_{i \le t:\, X_i = x_j} 1_{[Y_i \le Y_t]} \Big)_{j=1}^m$$

for $1 \le t \le n$. Note that $z^{(t-1)}$ and $z^{(t)}$ differ in precisely one component, and that

$$z(y) \ = \ \begin{cases} z^{(0)} & \text{if } y < Y_1, \\ z^{(t)} & \text{if } Y_t \le y < Y_{t+1}, \ 1 \le t < n, \\ z^{(n)} & \text{if } y \ge Y_n. \end{cases}$$

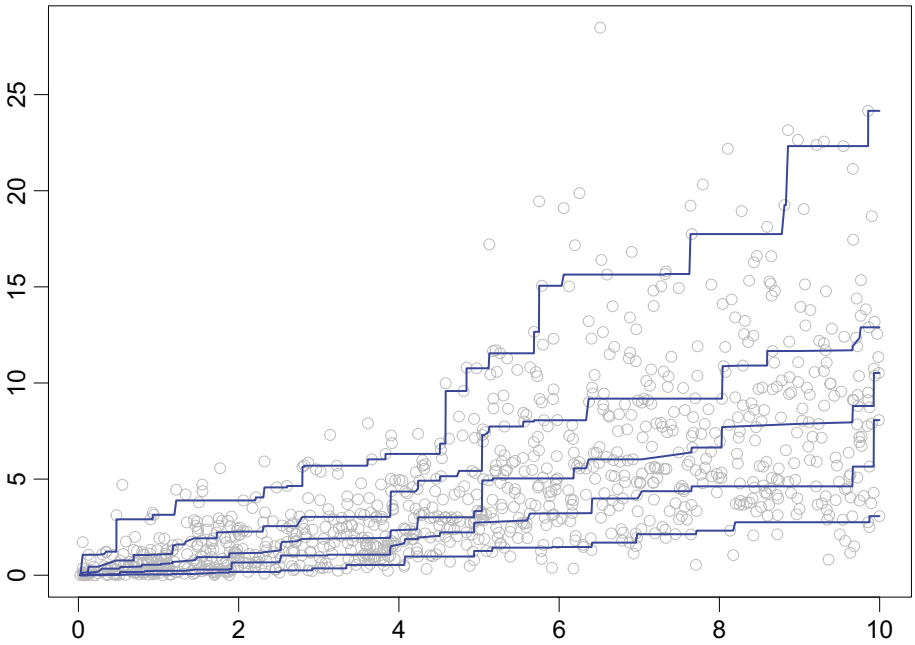

Fig. 2 A data set with estimated quantile curves

Thus it suffices to compute $A(z^{(t)})$ for $t = 0, 1, \ldots, n$. But $A(z^{(0)}) = \mathbf{0}$, $A(z^{(n)}) = \mathbf{1}$, and for $1 \le t < n$, one may apply the abridged PAVA to the vectors $z := z^{(t-1)}$ and $\tilde{z} := z^{(t)}$. This leads to an efficient algorithm to compute all vectors $A(z^{(t)})$, $0 \le t \le n$, if implemented properly.

Numerical Experiment 1 We generated data sets with $n = 1000$ independent observation pairs (X_i, Y_i), $1 \le i \le n$, where X_i is uniformly distributed on [0, 10] while $\mathcal{L}(Y_i \,|\, X_i = x)$ is the gamma distribution with shape parameter $\sqrt{x}$ and scale parameter $2 + (x-5)/\sqrt{2 + (x-5)^2}$. Figure 2 shows one such data set. In addition, one sees estimated β-quantile curves for levels $\beta \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, resulting from the estimator $\widehat{\boldsymbol{F}}$.

Now we simulated 1000 such data sets and measured the times T_1, T_2, T_3 for computing the estimator $\widehat{\boldsymbol{F}}$ via the standard, the modified and the abridged PAVA, respectively. Table 3 reports the sample means and standard deviations of these computation times in the 1000 simulations. In addition, one sees the averages and standard deviations of the ratios T_i/T_j, for $1 \le i < j \le 3$. It turned out that using the modified instead of the standard

Table 3 Computation times in seconds and ratios of running times

Variant of PAVA	mean (sd) of T_j		mean (sd) of T_1/T_j		mean (sd) of T_2/T_3	
Standard T_1	6.0394	(1.5257)				
Modified T_2	1.7482	(0.4224)	3.4618	(0.3816)		
Abridged T_3	0.2080	(0.1052)	30.8308	(6.1209)	8.9012	(1.4469)

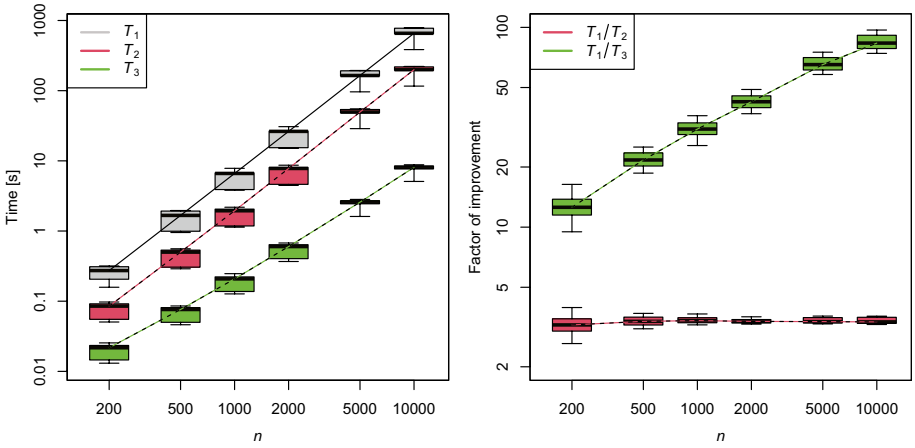

Fig. 3 Boxplots of computation times and ratios of running times for varying sample sizes. The whiskers indicate the 10% and 90% sample quantiles. The other elements of the boxplots are standard. A logarithmic scale was used for both axes

PAVA reduced the computation time by a factor of 3.46 already. Using the abridged PAVA yielded a further improvement by a factor of 8.90.

Figure 3 displays the result of simulation experiments for sample sizes ranging from 200 to 10 000, where the data were generated using the procedure mentioned earlier. The simulations indicate that the improvement due to using modified instead of standard PAVA is almost constant in n, whereas the improvement due to abridged instead of modified PAVA increases with n. Presumably, the complexity of the abridged PAVA for computing the isotonic distributional regression remains quadratic in n. But our numerical experiments show that the constant is substantially smaller than the one resulting from applying the usual PAVA with complexity $O(n)$ for $n-1$ different levels of the response.

Numerical Experiment 2 The goal of this experiment is to study the influence of the strength of the monotone association between X and Y on the efficiency gain of the abridged PAVA for isotonic distributional regression. The gains of abridged PAVA are expected to be milder when Y is independent of X, and to become larger as the monotone association strengthens. The reason behind it is that, while the standard PAVA proceeds independently of the stochastic order, the abridged PAVA relies on the index j_o indicating the component increasing in $z(t-1)$ and on the nature of the partition corresponding to $A(z(t-1))$, at a certain state $t \in \{1, \dots, n\}$ of the procedure. If the monotone association is weak, then the partition corresponding to $A(z(t-1))$ tends to contain fewer blocks in total and relatively large blocks in the middle of $\{1, \dots, n\}$. If the index j_o happens to lie in a block containing many indices to the right of j_o, even the abridged PAVA will have to inspect all of these.

To demonstrate this claim, we simulated n independent bivariate Gaussian random vectors $(X, Y)^\top$ with correlation $Corr(X, Y) = \rho \geq 0$. Note that the respective means and variances of X and Y have no influence on the results of the experiment. Indeed, the running times are invariant under strictly isotonic transformations of X and of Y. In particular, the simulations for $\rho = 0$ cover all situations in which X and Y are stochastically independent with continuous distribution functions. The stochastic order between $\mathcal{L}(Y|X = x_1)$ and $\mathcal{L}(Y|X = x_2)$ for $x_1 < x_2$

Table 4 Means (and standard deviations) of the factor of improvement T_3/T_1 for different correlation values ρ between X and Y and sample sizes n

n	$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
200	6.5337	(2.3496)	10.7581	(3.6704)	13.5695	(4.6390)
500	8.3029	(2.6393)	18.7010	(5.7806)	26.1813	(8.0763)
1 000	9.1351	(3.1800)	27.6290	(7.5007)	41.4116	(11.2161)
2 000	9.7559	(3.3532)	39.3180	(10.0337)	62.8382	(16.3293)
5 000	10.7495	(4.0525)	62.4600	(18.2002)	108.4198	(31.1414)
10 000	12.5190	(5.6193)	91.9084	(33.4657)	168.5030	(58.7712)

becomes stronger as the correlation $\rho \in [0, 1)$ increases, from an equality in distribution when $\rho = 0$ to a deterministic ordering when ρ approaches 1. Now, for sample sizes n ranging from 200 to 10 000 and for each correlation $\rho \in \{0, 0.5, 0.9\}$, the mean and standard deviation of the time ratio T_3/T_1 were estimated from 1 000 repetitions. The results are summarized in Table 4. As expected, the efficiency gain is smallest for $\rho = 0$. But even then, it is larger than 6 for $n \geq 200$ and larger than 9 for $n \geq 1\,000$.

Acknowledgements The authors are grateful to a reviewer for constructive comments.

Funding Information Open access funding provided by University of Bern. This work was supported by Swiss National Science Foundation.

Availability of Data and Material Not applicable.

Code availability R code is available at https://github.com/AlexanderHenzi/abridgedPava.

Declarations

Conflict of Interest Not applicable.

References

Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD (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

Domínguez-Menchero JS, González-Rodríguez G (2007) Analyzing an extension of the isotonic regression problem. Metrika 66:19–30

Henzi A, Ziegel JF, Gneiting T (2021) Isotonic distributional regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 83:963–993

Jordan AI, Mühlemann A, Ziegel JF (2021) Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals. Annals of the Institute of Statistical Mathematics to appear

Luss R, Rosset S (2014) Generalized isotonic regression. J Comput Graph Statist 23 192–210. https://doi.org/10.1080/10618600.2012.741550

Mösching A, Dümbgen L (2020) Monotone least squares and isotonic quantiles. Electron J Stat 14 24–49. https://doi.org/10.1214/19-EJS1659

Robertson T, Wright FT, Dykstra RL (1988) Order restricted statistical inference. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Ltd., Chichester