Journal of Computational Mathematics, Vol. 24, No. 6, 2006, 771–790.

DATA PREORDERING IN GENERALIZED PAV ALGORITHM FOR MONOTONIC REGRESSION [*1)]

Oleg Burdakov, Anders Grimvall and Oleg Sysoev
(*Department of Mathematics, Linköping University, SE-58183 Linköping, Sweden*)

Abstract

Monotonic regression (MR) is a least distance problem with monotonicity constraints induced by a partially ordered data set of observations. In our recent publication [In Ser. *Nonconvex Optimization and Its Applications*, Springer-Verlag, (2006) **83**, pp. 25-33], the Pool-Adjacent-Violators algorithm (PAV) was generalized from completely to partially ordered data sets (posets). The new algorithm, called GPAV, is characterized by the very low computational complexity, which is of second order in the number of observations. It treats the observations in a consecutive order, and it can follow any arbitrarily chosen topological order of the poset of observations. The GPAV algorithm produces a sufficiently accurate solution to the MR problem, but the accuracy depends on the chosen topological order. Here we prove that there exists a topological order for which the resulted GPAV solution is optimal. Furthermore, we present results of extensive numerical experiments, from which we draw conclusions about the most and the least preferable topological orders.

Mathematics subject classification: 90C20, 90C35, 62G08, 65C60.
Key words: Quadratic programming, Large scale optimization, Least distance problem, Monotonic regression, Partially ordered data set, Pool-adjacent-violators algorithm.

1. Introduction

Consider the monotonic regression (MR) problem for a partially ordered data set of n observations. We denote the vector of observed values by $Y \in R^n$. To express the partial order, we use a directed acyclic graph $G(N, E)$, where $N = \{1, 2, \ldots, n\}$ is a set of nodes and E is a set of edges. Each node is associated with one observation, and each edge is associated with one monotonicity relation as described below. In the MR problem, we must find among all vectors $u \in R^n$, which preserve the monotonicity of the partially ordered data set, the one closest to Y in the least-squares sense. It can be formulated as follows. Given Y, $G(N, E)$ and a strictly positive vector of weights $w \in R^n$, find the vector of fitted values $u^* \in R^n$ that solves the problem:

$$\begin{array}{ll} \min & \sum_{i=1}^{n} w_i (u_i - Y_i)^2 \\ \text{s.t.} & u_i \le u_j \quad \forall (i,j) \in E \end{array} \tag{1.1}$$

Let $\varphi(u)$ and C denote, respectively, the objective function and the feasible region of this problem. Note that (1.1) can be viewed as a problem of minimizing the weighted distance from the vector Y to the convex cone C. Since it is a strictly convex quadratic programming problem, there exists a unique optimal solution u^*. An example of problem (1.1) is considered at the end of Section 2.

[*] Received January 20, 2006.
[1)] This work was supported by the Swedish Research Council.

The MR problem has important statistical applications (see [2, 22]) in physics, chemistry, medicine, biology, environmental science etc. It is present also in operations research (production planning, inventory control etc.) and signal processing. These problems can often be regarded as monotonic data fitting problems (see Section 4). The most challenging of the applied MR problems are characterized by a very large value for n. For such large-scale problems, it is of great practical importance to develop algorithms whose complexity does not rise too rapidly with n.

It is easy to solve problem (1.1) when the constraints have the simple form

$$u_1 \leq u_2 \leq \ldots \leq u_n, \tag{1.2}$$

i.e. when the associated graph $G(N, E)$ is a path. For this special case of complete order, the most efficient and the most widely used algorithm is the Pool-Adjacent-Violators (PAV) algorithm [1, 17, 14]. Its computational complexity is $O(n)$ [11].

The conventional quadratic programming algorithms (see [20]) can be used for solving the general MR problem (1.1) only in the case of small and moderate values of n, up to few hundred. There are some algorithms [2, 22] especially developed for solving this problem. The minimum lower set algorithm [4, 5] is known to be the first algorithm of this kind. It was shown in [3] that this algorithm, being applied to problem (1.1) with the simple constraints (1.2), is of complexity $O(n^2)$, which is worse than the complexity of the PAV algorithm. The existing optimization-based [3, 16, 23] and statistical [18, 19, 24] MR algorithms have either too high computational complexity or too low accuracy of their approximate solutions. The best known complexity of the optimization-based algorithms is $O(n^4)$ [16, 23], which is prohibitive for large n. Perhaps, the most widely used algorithms for solving large-scale applied MR problems are based on simple averaging techniques. They can be easily implemented and have a relatively low computational burden, but the quality of their approximations to u^* is very case-dependent and furthermore, these approximations can be too far from optimal.

In [6, 7], we introduced a new MR algorithm, which can be viewed as a generalization of the Pool-Adjacent-Violators algorithm from the case of a completely ordered (1.2) to partially ordered data set of observations. We call it the GPAV algorithm. It combines both low computational complexity $O(n^2)$ and high accuracy. These properties extend the capabilities of the existing tools to solving very large-scale MR problems. The efficiency of the GPAV algorithm has been demonstrated in [6, 7] on large-scale test problems with obvious superiority over the simple averaging techniques [18, 19, 24]. In [13], it has been used advantageously for solving applied MR problems.

Our algorithm treats the nodes N or, equivalently, the observations in a consecutive order. Any topological order [9] of N is acceptable, but the accuracy of the resulted solution depends on the choice. In this paper, we focus on studying the effect of topological sort on the quality of the solution produced by the GPAV algorithm.

In Section 2, the GPAV algorithm is reformulated for the case when the nodes N are topologically ordered. We prove in Section 3 that, for any MR problem (1.1), there exists a topological order assuring that the GPAV algorithm produces the exact solution to this problem. In Section 4, test problems based on the monotonic data fitting are described. We introduce a wide variety of topological sorts and study their effect on the performance of the GPAV algorithm. Results of extensive numerical experiments are presented and discussed in this section. In Section 5, we draw conclusions about the most and the least preferable topological sorts.