Computational Statistics and Data Analysis 55 (2011) 2463–2476

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A segmentation-based algorithm for large-scale partially ordered monotonic regression

O. Sysoev [a,*], O. Burdakov [b], A. Grimvall [a]

[a] Department of Computer and Information Sciences, Linköping University, SE-58183, Linköping, Sweden

[b] Department of Mathematics, Linköping University, SE-58183, Linköping, Sweden

ARTICLE INFO

Article history:
Received 15 February 2010
Received in revised form 7 March 2011
Accepted 7 March 2011
Available online 13 March 2011

Keywords:
Quadratic programming
Large-scale optimization
Least distance problem
Monotonic regression
Partially ordered data set
Pool-adjacent-violators algorithm

ABSTRACT

Monotonic regression (MR) is an efficient tool for estimating functions that are monotonic with respect to input variables. A fast and highly accurate approximate algorithm called the GPAV was recently developed for efficient solving large-scale multivariate MR problems. When such problems are too large, the GPAV becomes too demanding in terms of computational time and memory. An approach, that extends the application area of the GPAV to encompass much larger MR problems, is presented. It is based on segmentation of a large-scale MR problem into a set of moderate-scale MR problems, each solved by the GPAV. The major contribution is the development of a computationally efficient strategy that produces a monotonic response using the local solutions. A theoretically motivated trend-following technique is introduced to ensure higher accuracy of the solution. The presented results of extensive simulations on very large data sets demonstrate the high efficiency of the new algorithm.

1. Introduction

The monotonic regression (MR) problem often originates from applications in which observations are composed of a vector of p explanatory variables $x \in R^p$ and a response variable $y \in R$. The values of x and y for each observation $i = 1, \ldots, n$ are denoted as X_i and Y_i, respectively, and they are collected in the *input data set* $D = \{(X_i, Y_i),\ i = 1, \ldots, n\}$, which is frequently used as an input to applied MR problems.

It is assumed that the unknown true response function $f(x)$ is monotonic, i.e., increasing with respect to some variables and decreasing with respect to others. For simplicity, we assume that f is isotonic, i.e. that it increases with respect to each component of $x = (x_1, \ldots, x_p)$. This means that

$$f(x') \le f(x''), \quad \forall x', x'' \in R^p \text{ such that } x' \prec x'',$$

where $x' \prec x''$ means that $x'_i \le x''_i$ for each $i = 1, \ldots, p$. In real-life applications, the data sets are seldom monotonic, and hence $X_i \prec X_j$ does not necessarily imply $Y_i \le Y_j$. This is due to some observation errors, $y = f(x) + \varepsilon$.

Here, we focus on the case of $p > 1$, for which the set of observed explanatory variables $X_1, \ldots, X_n$ is usually *partially ordered*. This means that some pairs (X_i, X_j) may not be comparable, i.e. neither $X_i \prec X_j$ nor $X_j \prec X_i$ holds. The original function $f(x)$ is not available and cannot be restored from the data set D, and it is even impossible to restore the function values $f(X_i)$ for $i = 1, \ldots, n$. Nevertheless, by using the knowledge of monotonicity and data set D, these function values can be approximated by solving the corresponding MR problem whose solution $f^* \in R^n$ is an approximation of the vector

* Corresponding author. Tel.: +46 13282785.
E-mail addresses: olsys@ida.liu.se (O. Sysoev), olbur@mai.liu.se (O. Burdakov), angri@ida.liu.se (A. Grimvall).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.03.001

$(f(X_1), \ldots, f(X_n))$. Moreover, the components of f^* are consistent with the monotonicity and are as close as possible to the observed response values Y_i. In the weighted L_2-norm, the MR problem is formulated as

$$\begin{aligned} &\min \quad \sum_{i=1}^{n} w_i (f_i - Y_i)^2 \\ &\text{s.t.} \quad f_i \le f_j \quad \text{iff } X_i \prec X_j,\ i, j = 1, \ldots, n, \end{aligned} \tag{1}$$

where $w_k > 0,\ k = 1, \ldots, n$ are given values of weights.

The MR problem has numerous applications in operations research, statistics, biology, signal processing, and other areas; see Barlow et al. (1972), Robertson et al. (1988) and Oh and Dong (2011). The most challenging of the practical problems in this context are characterized by a large n value. Various techniques for solving problem (1) have been developed in recent decades. The pioneering work of Ayer et al. (1955) led to introduction of the Pool-Adjacent-Violators (PAV) algorithm. This algorithm is used to solve a special case of the MR problem, typically for $p = 1$, where the observations are completely (linearly) ordered, i.e., either $X_i \prec X_j$ or $X_j \prec X_i$ holds for each pair of observations (X_i, X_j). A profound framework for MR theory was developed by Barlow et al. (1972) and Robertson et al. (1988). At present, the most widely known exact algorithms for solving the general MR problems with a partially ordered input data set D include the following: the minimum lower set algorithm of Brunk (1955), the min–max algorithm described by Lee (1983), the network-based algorithm of Maxwell and Muchstadt (1985), and the IBC algorithm introduced by Block et al. (1994). Unfortunately, these exact algorithms can only solve problems that contain a relatively small number of observations, and thus they cannot provide satisfactory results within a reasonable amount of time when addressing medium- or large-scale problems.

In our recent papers (Burdakov et al., 2006a,b), we presented a generalized PAV algorithm called the GPAV, with which high-accuracy solutions of large-scale multivariate MR problems can be obtained for a partially ordered D. Here, we present a segmentation-based algorithm (designated SB), which extends the area of its applications to much larger MR problems, that are too demanding for the original GPAV and all other well-known MR algorithms in terms of computational time and memory. The SB splits a large-scale MR problem into a number of medium-size problems and solves them using the GPAV. The fitted values are monotonic only locally within each segment, but the monotonicity may be violated on the boundary of the neighbor segments. The SB offers a special, computationally efficient strategy that produces an overall monotonic fit on the basis of the local solutions.

The remainder of this paper is organized as follows. In Section 2, we introduce and discuss an alternative formulation of problem (1). The SB algorithm is presented in Section 3, and in Section 4 we consider a trend-following order aimed at making this algorithm more efficient. In Section 5, we justify the correctness of the SB and study the theoretical properties of this algorithm. In Section 6, the worst-case complexity of the SB is estimated. In Section 7, we analyze the computational performance of the SB and compare it with performance of the GPAV. Section 8 contains conclusions and comments.

2. Alternative formulation and the GPAV algorithm

The problem formulation (1) is natural for most practical applications. In computational science, as well as in our previous papers (Burdakov et al., 2006a,b), an alternative formulation is used in which a graph $G = G\langle N, E\rangle$ with a set of nodes N and a set of edges E is supposed to be given. Each node $i \in N = \{1, \ldots, n\}$ is associated with the observation i, and each edge $(i, j) \in E$ is associated with the relation $X_i \prec X_j$. The graph is obviously acyclic.

Here, we use the following definitions and notations. A node $i \in N$ is a *predecessor* of node $j \in N$ if there is a directed path in the graph from i to j. A *block* is a connected subset $B \subset N$ such that if there is a directed path between two nodes in B, then all the nodes in the path belong to B. The MR problem admits the following graph formulation:

$$\begin{aligned} &\min \quad \sum_{i \in N} w_i (f_i - Y_i)^2 \\ &\text{s.t.} \quad f_i \le f_j \quad \text{for all } (i, j) \in E, \end{aligned} \tag{2}$$

or, equivalently,

$$\begin{aligned} &\min \quad \sum_{i \in N} w_i (f_i - Y_i)^2 \\ &\text{s.t.} \quad f_i \le f_j \quad \text{for all } (i, j) : A_{ij} = 1, \end{aligned} \tag{3}$$

where A is $n \times n$ *adjacency matrix* (Cormen et al., 2001) with the components

$$A_{ij} = \begin{cases} 1, & \text{if } (i, j) \in E, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, for any given partially ordered input data set D, one can construct a corresponding acyclic graph G (or equivalently an adjacency matrix A). In other words, formulation (1) implies (2) for some G, or it implies (3) for some A. It is not difficult to show that the reverse is also true, i.e. given an acyclic graph G and an associated observed response Y, it is possible