ESTIMATING BINARY MONOTONIC REGRESSION MODELS AND THEIR UNCERTAINTY BY INCORPORATING KERNEL SMOOTHERS

Oleg Sysoev

Division of Statistics
Department of Computer and Information Science
Linköping University
SE-58183 Linköping
(e-mail: oleg.sysoev@liu.se)

Abstract. Monotonic regression (MR) is a method that is used for fitting a model to multivariate data in which a response is increasing or decreasing with respect to several explanatory variables. Recently, MR methods were substantially improved, and a group of algorithms can now produce monotonic fits to multivariate datasets containing over a million observations. It is demonstrated here that the accuracy of the monotonic fits produced by these algorithms is unacceptable when the response variable is binary, and kernel smoothers are incorporated into these algorithms to increase the accuracy. A standard approach for estimating confidence limits for the models obtained by these algorithms is to use a resampling technique that applies the same modeling algorithm to each of the numerous bootstrap sets. However, the computational complexity of this approach becomes prohibitively large in the large-scale MR setting. Here, the standard resampling approach is modified to estimate confidence limits efficiently.

1 Introduction

Monotonic regression (MR) is a nonparametric method designed for applications in which the expected value of the response variable is assumed to be monotonic, i.e. increase or decrease with respect to the set of predictors $x \in \mathbb{R}^p$. Such applications are widespread in medicine, economics, engineering and other areas, see Acton et al. (1998), Ant-Sahalia and Duarte (2003), Ulm and Salanti (2003). More specifically, response function $f$ is increasing with respect to $x$ if for any pair $x_1, x_2 \in \mathbb{R}^p$ such that $x_1 \prec x_2$ the relation $f(x_1) \leq f(x_2)$ holds. Here, $x_1 \prec x_2$ implies $x_{1i} \leq x_{2i}$, for all $i = 1, \ldots, p$.

Most of applications generate a data set $D = \{(X_i, Y_i), X_i \in \mathbb{R}^p, Y_i \in \mathbb{R}^1, i = 1, \ldots, n\}$. The observed response values $Y_i$ are typically non-monotonic, and the objective of the MR is to solve

$$\min_{u_1, \ldots, u_n} \sum_{i=1}^{n} w_i (u_i - Y_i)^2$$

s.t. $u_i \leq u_j$ iff $X_i \prec X_j, i, j = 1, \ldots, n$.

Due to growth of computer capabilities, modern applications collect and store large amounts of data. To process such large data, efficient MR methods are needed. When $p = 1$, classical Pool-Adjacent-Violators (PAV) algorithm (Ayer et al. (1955)) is able to process quite large data due to its complexity $O(n)$. When $p \geq 2$, the methods able to produce an exact solution...
to (1) are too computationally expensive. Recently, so-called GPA V-type algorithms were developed that obtain approximate solutions to (1) of high accuracy for data containing over a million observations.

In some applications, the response is distributed as $y|x \sim \text{Bernoulli}(f(x))$ and $f(x)$ is assumed to be monotonic with respect to $x$. In this case, the optimal values of $f$ are usually found by maximum likelihood optimization, but ML these estimates coincide with values obtained by solving (1), as it is shown in Robertson el al. (1988), page 32. Therefore, MR algorithms able to solve (1) are also suitable for finding maximum likelihood estimates for data with binary response.

In section 3, it will be demonstrated that GPA V-type algorithms are able to process large-scale multivariate data with binary response, however, the accuracy of the fit is quite poor in this case. It will be illustrated that the accuracy of the algorithms is dependent on the order in which the observations are treated, and to find a proper order, kernel methods (Wand and Jones (1995)) are used here. Accordingly, kernel estimates are incorporated into the GPA V-type algorithms in a special way to obtain high-accuracy solutions. To select a correct bandwidth for a given kernel function, the rule of thumb (Wand and Jones (1995)) and the holdout method (Lachenbruch and Mickey (1968)) are applied.

It is of high importance to be able to estimate uncertainty (for example, in the form of confidence intervals) of the derived estimates when doing statistical inference. The methods based on analytical expressions under normality assumptions such as Korn (1982), Lee (1986) are unsatisfactory for the general case (Dilleen et al. (2003)). Accordingly, the bootstrap (Efron and Tibshirani (1993)) or other forms of resampling is a natural choice. In Dilleen et al. (2003), the percentile bootstrap was used to compute confidence intervals (CI)s for the expected response when data contained one predictor.

Some strategies are developed in this paper to enable an adequate bootstrap and jackknife estimation, and these strategies are further employed to compute the bootstrap percentile and the bootstrap BCa confidence intervals (Efron and Tibshirani (1993)).

2 MR ALGORITHMS FOR LARGE-SCALE PROBLEMS

GPAV-type algorithms are able to compute a high-accuracy solution to large-scale MR problem. GPAV algorithm (Burdakov et al. (2006b)) has complexity $O(n^3)$ if the MR fit is computed from observations and $O(n^2)$ if an adjacency graph (discussed below) is given instead of the observational data. In practice, GPAV algorithm can process data with over 10,000 observations if the adjacency graph is not given and over 100,000 observations if this graph is given. GPAVR algorithm (Burdakov et al. (2009)) has approximately three times higher computational time than the GPAV but demonstrates a higher precision. Segmentation-based algorithm has a worst-case complexity $O(n^3)$ but in practice it can handle data with over a million observations.

The GPAV-type algorithms make use of the following concepts. An adjacency graph $G$ can be created from data set $D$ as follows: a node $i$ in the graph is created for each observation $(X_i, Y_i)$, and there is an edge $(i, j)$ if and only if $X_i \prec X_j$. There may exist redundant edges in the graph, i.e. edges $(i, k)$ such that there exist edges $(i, j)$ and $(j, k)$. Such edges are normally removed from $G$. Node $i$ is called an immediate predecessor of node $j$ if there exists edge $(i, j)$
in $G$. Node $i$ is a predecessor of $j$ if there is a directed path from $i$ to $j$ in $G$. A topological order $t = \{t_1, \ldots, t_n\}$ is such a permutation of \{1, \ldots, n\} that all predecessors $t_i$ of node $t_j$ are listed before $t_j$ in list $t$, i.e. $i < j$. There are normally a plenty of different topological orders for a given graph, see Cormen et al. (2001).

The GPAV algorithm treats the observations sequentially in the order specified by a selected topological order $t$. At each step, observation $i$ is merged with some predecessors into a set (block). Finally, each node belongs to some block, and the fitted value for a node is computed as the weighted average of the response of the observations that belong to the corresponding block.

The segmentation-based algorithm (Sysoev et al. (2011)) splits the data into non-overlapping segments containing the same amounts of observations, and then the GPAV algorithm is run for each segment to produce a local monotonic response. The blocks obtained in all segments are treated as new observational data, and the GPAV run on this new graph computes a global monotonic response.

3 PROBLEMS WITH FITTING A BINARY RESPONSE

The accuracy of the GPAV-type algorithms may dramatically depend on the choice of the topological order used for treating the observations. Different topological orders are examined in Burdakov et al. (2006a). In Sysoev et al. (2011), it is argued that the topological order that follows the trend of the expected response $f$ is a good choice, and the trend-following order introduced in that paper employs the observed response values to compute a topological order that is hoped to be similar to the one calculated from the expected response values. However, when the response is Bernoulli distributed, the observed response values cannot be used as an approximation of the expected response values, because the observed values are binary and no trend is present in them. The preliminary experiments reveal that the trend-following order fails to produce an MR fit of a reasonable quality.

It is also investigated here how other topological orders influence the response obtained by the GPAV. Figure 1 demonstrates that the structure of the fitted response is dramatically dependent on the topological order. The obtained blocks are aligned along direction defined by the chosen topological order. Shortly speaking, this behavior can be explained by the fact that there is a high chance of merging of block $B_{t_i}$ with block $B_{t_{i+1}}$ due to large variation in the response, and thus several blocks $B_{t_i}, \ldots, B_{t_{i+k}}$ are merged into a single block which obviously will be aligned along the direction defined by $t$.

4 COMBINING GPAV AND KERNEL SMOOTHERS

Kernel smoothers is an attractive alternative for finding estimated response values that are close to the expected response values, even if the observed response is binary. Here, Nadaraya-Watson kernel estimator was used in which the bandwidth was estimated by the holdout method and the rule of thumb. In the holdout method, 50% of the data were used for fitting the kernel smoother and 50% of data for the validation, i.e. for selecting the optimal bandwidth. The reason for choosing the holdout method instead of cross-validation is that the data being considered are enough big. Another approach is to use the fact that the kernel smoother
Figure 1. Estimated risk of death in CHD ($P$) determined by the segmentation-based algorithm. Topological order is by increase of first cholesterol, then age(1stComp). Exact solution is obtained for a discretized CHD set (IBCR)

for the binary response is a ratio of kernel density estimates, and therefore the rule of thumb can be applied, see Wand et al. (1995), page 111. The derived kernel estimates are used for computing the topological order that approximately follows the estimated response $f$.

Algorithm 1 K-order

Given $D=\{(X_i, Y_i) : X_i \in \mathbb{R}^p, Y_i \in \mathbb{R}^1, i = 1, \ldots, n\}$
1. Relabel observations $(X_i, Y_i)$ in such way that $I = \{1, 2, \ldots, n\}$ is a topological order.
2. Compute $\hat{f}_i$ by using a kernel smoother.
3. Sort $\hat{f}_1, \ldots, \hat{f}_n$ by increase of $\hat{f}_i$ and obtain $\hat{f}_{i_1}, \hat{f}_{i_n}$. Let $L = \{i_1, \ldots, i_n\}$
4. For $i=1$ to $n$
   - Find $j$: $i_j = i$. Compute $P$ as a set of immediate predecessors of $i$
   - Compute $S = \{k : i_k = j, j \in P\}$, let $m = \max(k)$
   - Move $i_j$ to the position between $i_k$ and $i_k+1$ in the list $L$

It can be shown that Algorithm 1 produces a topological order, and therefore can be used by the GPAV and segmentation-based algorithms.

5 CONFIDENCE INTERVALS

A problem appears when CIs need to be computed for large data: since MR should repeatedly be fitted to a number of bootstrap samples (normally around 1000), the computational time becomes prohibitively large. An even greater computational time is required if the jackknife computations are involved (for example, in the bootstrap BCa approach), since $n$ MR models
should be estimated. To overcome this problem, the segmentation-based algorithm can be ap-
plied in which information obtained by fitting data set \( D \) is used when the MR models are fit
to the bootstrap samples. It is actually possible to use the adjacency graphs created for each
segment of the original data instead of computing new adjacency graphs for the bootstrap
samples. Since computations of the adjacency graphs require the largest portion of the CPU
time in GPAV algorithm, the computations are reduced dramatically. The segmentation strat-
egy makes it possible to compute the jackknife even for large-scale data: instead of removing
an observation and then fitting the MR model to the remaining data, a segment can be found
that contains the observation to be removed, and the MR model is reestimated for this seg-
ment only. Often, this results in a global monotonic response for the jackknifed sample. In the
cases when the response is not monotonic due to monotonicity violations between segments,
a monotization step is needed, and this requires a little time because the data being pro-
cessed is almost monotonic. In addition, if bootstrap BCa intervals need to be estimated, the
computational time and memory can greatly be reduced if the estimation is done in parallel
for all observations, and such strategy was implemented here.

6 TEST RUNS

In this paper, artificial data are generated to examine the performance of the GPAV with in-
corporated kernel smoothing. In addition, a data on coronary heart disease (Engel and Malm-
strom (1969)) is used for illustrative purposes. This data includes an indicator of death and
information about age and cholesterol level of more than 40,000 women that participated in
a mass health screening survey.

Artificial data includes the data from the logistic model
\[
f(x) = \frac{1}{1 + e^{-\beta_1 x_1}}
\]
and typical
classification data: \( Y = 1 \) if \( X_2 > 1 + \epsilon \) and \( Y = 0 \) otherwise, where \( \epsilon \) is normally distributed
error. Different values of \( n \) from 2000 to 10000 (which are typical segment sizes), \( p \) equal to
2 or 3, normally or uniformly distributed \( x \) are tested.

To measure the efficiency of the kernel approach, the objective function \( d \) in (1) is com-
puted and the average deviation of \( \delta_{\text{GPAV}} - \delta_{\text{K-order}} \) is analysed. The results of the simulations
show that incorporating the kernel smoothers provides 8-15 % improvement in the accuracy
compared to the GPAV using number of predecessors as topological order, and this amount
grows with sample size \( n \) and dimension \( p \). Simulations for smaller data sizes \( n = 2000 \)
demonstrate that the GPAV using K-order deviates only about 1% from the exact solution to
the MR problem, and the high accuracy is also confirmed by Figure 1. It appears also that
the rule of thumb leads to only slightly worse results than the holdout method.

Different types of confidence intervals are obtained by applying the segmentation-based
algorithm using algorithm 1. It appears that if the bootstrap percentile or the bootstrap BCa is
used, different resampling methods lead to similar confidence bounds. However, the results
differ much between the bootstrap percentile and the bootstrap BCa as it can be expected.

7 CONCLUSIONS AND DISCUSSION

In this paper, a new method is developed that incorporates kernel smoothers into MR
algorithms. It is demonstrated that this method provides a significant improvement in ac-
accuracy of the large-scale MR algorithms for binary data. The bandwidth selection problem is investigated, and it appears that the holdout method leads to slightly more accurate MR estimates than rule of thumb. Finally, several approaches are suggested to make computing bootstrap-based CIs for the large-scale data feasible.

It can also be noted that this paper is devoted to binary response, but algorithm 1 will also work for continuous response. Comparison of the performance of this algorithm and the GPAV employing other topological orders for the case of continuous response can be a direction for further research.

REFERENCES


DILLEEN, M., HEIMANN, G., HIRSCH (2003), I.: Non-parametric estimators of a monotonic dose-response curve and bootstrap confidence intervals. Statistics in Medicine, 22, 869–882


SYSOEV, O., BURDAKOV, O., GRINVALL, A. (2011): A Segmentation-Based Algorithm for Large-Scale Monotonic Regression Problems. Computational Statistics and Data Analysis, 55, 2463–2476
