

A nonparametric changepoint model for stratifying continuous variables under order restrictions and binary outcome

Georgia Salanti and Kurt Ulm Institute for Medical Statistics and Epidemiology, Munich, Germany

Modelling using monotonic regression can be a useful alternative to parametric approaches when optimal stratification for continuous predictors is of interest. This method is described here in the context of binary response. Within this framework we aim to address two points. First, we propose a method to enhance the parsimony of the model, by applying a reducing procedure based on a sequence of Fisher exact tests and a bootstrap method to select between full monotonic and reduced model. Secondly, we discuss the case of multiple predictors: an iterative algorithm (an extension of the *Pool Adjacent Violators Algorithm*) can be applied when more than one predictor variable is taken into account. The resulting model is a monotonic surface and can be applied alternatively to the additive monotonic models as described by Morton-Jones and colleagues when the explanatory variables are assumed to interact. The monotonic-surface model provides also a multivariate extension of the monotonic likelihood ratio test. This test is discussed here and an approach based on permutations to assess the p -value is proposed. Finally, we combine both ideas (reduced monotonic regression and monotonic-surface estimation) to a simple and easy to interpret model, which leads to a combination of the predictors in a few constant risk groups. Despite the fact that the proposed approach becomes somewhat cumbersome due to the lack of asymptotic methods to infer, it is attractive because of its simplicity and stability. An application will outline the benefit of using bivariate step functions in modelling.

1 Introduction

Categorizing continuous variables arises as an important issue in statistical analysis, particularly in studies concerning exposure–effect problems. Usually a single cutpoint is used, which is determined by the maximization of a test statistic.^{1,2} A binary split is simple to use and to interpret, but its simplicity is gained at the expense of throwing away a lot of information, increasing the bias and losing power. An optimal stratification of the predictor into more than two groups can often be more informative, especially if the shape of the dose–response relationship is of interest. Once the decision to categorize is taken, it is not obvious how many groups should be built and where the cutpoints should be placed. The pattern of the response, the underlying biological mechanism and the sample size should be taken into account. Equally spaced or equally sized cutpoints suggested by the sample size are used in practice. However, rather than grouping according to the distribution of the explanatory variable, a better strategy is to base the selection of the cutpoints on the outcome. If there is more than one explanatory variable, the application of a statistical model is necessary.

Address for correspondence: Georgia Salanti, Institute for Medical Statistics and Epidemiology, Klinikum Rechts der Isar, Ismaninger Strasse 22, 81675 Munich, Germany. E-mail: vorgia@web.de

Creating meaningful groups for the predictor variables regarding the outcome is desirable in many studies. A representative example is the MAK (Maximal Arbeitsplatz Konzentration) study.³ One of the goals of the statistical analysis has been to test whether the inhalable dust concentration in the workplace has adverse effects for the health of the workers. Apart from inhalable dust, additional parameters such as the time since first exposure and smoking habits need to be taken into account. The end point of the study has been chronic bronchitis. In the statistical analysis of this data, the proof of a dose–response relationship, that is, increasing risk with increasing dust concentration, was an important task in order to establish causality. In the case of evidence, the stratification of dust concentration into certain risk groups was of great interest. According to the established risk categories the MAK commission could take decisions to protect workers against the dust effects and to assess an overall threshold for dust concentration in the workplace.

On analysing the results of this study, parametric models such as the probit and logit as well as nonparametric models have been applied. A common assumption made behind the parametric models is that the relationship is linear either directly or after some transformation. This approach turns out often to be inadequate as it is too restrictive. More flexible models, such as the generalized additive one, fitted by smoothing splines or fractional polynomials, are useful but they are not always helpful when the establishment of a dose–response relationship needs to be proven. Moreover, these models make the assessment of cutpoints rather cumbersome. Usually the investigator needs to decide on the number and location of cutpoints on a subjective base, after screening the graph of the fitted model. Then, modelling using step functions provides a reasonable alternative. Classification trees are the most popular model in this situation.

Motivated from this special case we propose a model based on monotonic regression.^{4,5} This approach is adequate when the goals of the analysis are to prove a monotonic trend between a binary response and one or more predictor variables, and further to assess an optimal stratification of the explanatory variables for practical purposes. This method retains the monotonicity assumption but relaxes the linearity requirement, and results in fitting step functions without any *a priori* assumption about the location of the shifts.

The simplicity of monotonic regression has recently turned its use into a popular tool. Among the many benefits it provides, one of the most important is the monotonic test for trend. This test has attracted a lot of attention and its advantages compared to standard tests for trend have been outlined in several papers.^{6–9} Nevertheless, there is some controversy about it. The large sample approximation used does not always hold (as we will outline later) and one has to apply permutations to infer for the effect of the predictors.

An important improvement of the monotonic model has been reduced monotonic regression.¹⁰ This focuses on reducing the level sets in a one-dimensional monotonic regression by detecting cut points that do not correspond to an important change in the response. Up to now, reduced monotonic regression has been described only for one explanatory variable and continuous response variable.

Regarding multidimensional monotonic modelling, Bacchetti¹¹ and later Morton-Jones¹² proposed the additive monotonic regression. This model is an extension of the generalized additive model¹³ where the monotonic ‘smoother’ is used in the partial fit.

At this point, two important remarks need to be made regarding the monotonic framework. First, some problems arise when one deals with binary response problems. In this case, the consistency of the large sample approximation for the provided test for trend derived in ref. 14 is poor and the procedure proposed by Schell¹⁰ for reducing the model cannot be applied because it makes use of an F-test that requires normality in the response. Secondly, difficulties arise when more than two predictors are included in the analysis, for both the monotonic test and the reduced regression. The additive monotonic model offers a potential solution, but a lot of work remains to be done on introducing adequately an elimination algorithm in the model-fitting procedure.

In the present paper, restricted to the case of binary response, an alternative will be presented when modelling with multiple predictors. We propose to use a monotonic-surfaces model and we will present the corresponding multidimensional monotonic test for trend. After modifying the procedure of Schell for the case of binary outcome, we extend it to more than one predictor variable and we will use it to eliminate the ‘needless’ level sets of a monotonic-surfaces model. The obtained result is a combination of the predictors in constant risk groups.

The paper is structured as follows: in the methodology part, the basics about monotonic regression are revised and difficulties arising for the monotonic test for trend when the response is binary are discussed. Then, reduced monotonic regression is adapted for binary response and a limited simulation study is presented. Section 4 describes the monotonic-surfaces model and the implementation of the reducing procedure. In the application, we will analyse the MAK data and we will compare the obtained results to those taken from an additive monotonic model and a classification tree.

2 Methods

2.1 The monotonic method: estimation and test for trend

The only assumption underlying these models¹⁴ is monotonicity: either increasing (monotonic) or decreasing (antitonic) trend. Without loss of generality, we will consider only the monotonic case. Starting from the assumption that a monotonic dose–response relationship exists, a maximum likelihood estimator under order restrictions is assessed for the response. This estimator can be provided by several equivalent algorithms: either the Minimum Lower Sets Algorithm, the Maximum Upper Sets Algorithm or the Pool Adjacent Violators Algorithm (PAVA),¹⁴ which is used in the present paper.

Focusing on binary response the PAVA can be described as follows: consider the situation of N dose groups where the dose d_i , ($i = 1, \dots, N$) is in increasing order, n_i observations falling in the i th dose and the end point is the probability p_i of an event estimated by the observed proportions \hat{p}_i . We wish to have \hat{p}_i in nondecreasing order, given that $d_i \leq d_{i+1}$, ($i = 1, \dots, N$). If there is a violator somewhere such that $\hat{p}_i > \hat{p}_{i+1}$ for some i , then the isotonic estimator of both values needs to be found. That is

provided by their weighted mean $\hat{p}_{i,i+1}^* = (\hat{p}_i n_i + \hat{p}_{i+1} n_{i+1}) / (n_i + n_{i+1})$, where the weights are the number of observations per dose group. Now the elements $i, i + 1$ form a block – called the level set (LS) – containing $n_i + n_{i+1}$ observations. This process is repeated using the new probabilities and weights until an isotonic set of response probabilities is obtained. The algorithm assuming a decreasing trend is similar. The goodness of fit of the isotonic transformation described above is measured by the likelihood function, taking into account the number of level sets.

The isotonic framework has some advantages compared to parametric methods. No specific assumptions other than monotonicity are required for the form of the dose–response relationship. Nevertheless, the main advantage is the test for trend connected with isotonic regression. In search of such an adequate test, recall that many tests for trend, as for example the commonly used Cochran–Armitage test, give results that depend on the form in which the dose is used.⁹ However, isotonic regression not only provides one of the most reliable tests for trend,^{6,9} but is also expected to have increased power by setting the isotonic transformation of the response as the alternative hypothesis to the constant risk assumption H_0 as outlined in ref. 8. This test is known as the isotonic likelihood ratio test¹⁴ and follows a weighted X^2 distribution.

We define the following hypothesis:

- $H_0: p_1 = p_2 = \dots = p_k = p_0$ against the alternative
- $H_1: p_1 \leq p_2 \leq \dots \leq p_k$ with at least one strict inequality

where $p_0 = (\sum_{i=1}^N n_i p_i) / (\sum_{i=1}^N n_i)$. Now, let T_{01} be the statistic that tests H_0 against H_1 . This test has the form

$$T_{01} = D(\hat{p}_{H_0}) - D(\hat{p}_{H_1}) = 2 \sum_{i=1}^N \left[n_i \hat{p}_i \ln \left(\frac{\hat{p}_i^*}{\hat{p}_0} \right) + n_i (1 - \hat{p}_i) \ln \left(\frac{1 - \hat{p}_i^*}{1 - \hat{p}_0} \right) \right] \tag{1}$$

where the deviance $D(\hat{p}_{H_i})$ is the function $-2\log(\text{Likelihood})$ under the hypothesis H_i . Then, the large sample approximation of the distribution for T_{01} under H_0 is

$$P(T_{01} \geq c) = \sum_{l=2}^N P(l, N, w) P[X_{l-1}^2 \geq c] \tag{2}$$

where $P(l, N, w)$ (having $\sum_l P(l, N, w) = 1$) denote the probabilities that under H_0 and given N distinct dose levels, the isotonic regression will build l level sets. For a more detailed description of the weights $P(l, N, w)$ see section 2.4 in ref. 14.

However, this approximation does not always hold for a binary response. We consider the situation where $k = 8$ proportions are compared and the sample size in each group is 50. The response rate is assumed to vary by 5, 10 and 25%. Table 1 shows the theoretical critical values estimated from equation (2), and these assessed from 10 000 permutations. Simulations under different scenarios for k and sample size, showed similar results: the large sample approximation fails, especially for small response rates. Note that in this simulation study we assumed that the number of observations is equal in each group, which is unlikely to occur in practice. The

Table 1 Simulations (10000) under H_0 assumption (constant risk). Isotonic Likelihood Ratio (R) test: significance levels and 95% critical values for comparing $K=8$ dose groups and sample size 400 (50 in each dose group) when the response rate is 5, 10 and 25%

Nominal significance level (theoretical critical values)	Estimated significance level if the theoretical critical value is used (estimated critical values for the nominal levels)		
	$p_0 = 5\%$	$p_0 = 10\%$	$p_0 = 25\%$
0.05 (6.088)	0.067 (6.526)	0.055 (6.355)	0.052 (6.200)
0.01 (9.640)	0.015 (10.142)	0.013 (9.963)	0.011 (9.711)

calculation of the level probabilities $P(l, N, w)$ becomes very cumbersome when the weights in each dose level are unequal. Moreover, when more than one explanatory variable is taken into account the likelihood ratio test does not follow any known distribution. Thus, the approximation derived by Robertson and colleagues¹⁴ is not useful here.

We now present a permutation test for accepting or rejecting H_0 . Given an observed value for the isotonic likelihood ratio test T_{01}^{obs} , the outcome variable (the events for the case of binary response) is permuted while the predictor variable is kept constant. In each permutation $b = 1, \dots, B$, the change in the deviance applying equation (1) after isotonic regression is assessed. Then, the proportion of deviance improvements (denoted T_{01}^b) that exceed the observed value, provides the p -value of the test. This is given by

$$p\text{-value}_{B\text{-perm}} = \frac{\sum_{b=1}^B I_+(T_{01}^b - T_{01}^{obs}) + 1}{B + 1} \tag{3}$$

where

$$I_+(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

In equation (3) we have also taken into account the observed value T_{01}^{obs} . Following the same idea, one can construct pointwise confidence intervals for the estimates. The B isotonic estimates from the permuted data sets form a pointwise distribution of the data under H_0 . We construct a 95% CI by simply picking the 2.5% smallest and 2.5% largest estimations. The width of these confidence surfaces provides useful information additional to the test result; that is, in case of a nonsignificant result, examination of the intervals can help us to distinguish between imprecise estimates of \hat{p}_i for which the nonsignificant result may leave opened the possibility of an important variation in the true risk (wide confidence band), and estimates where statistical consistency of risk coincidence is true (narrow confidence band). In the case of a significant result, one can additionally assess confidence intervals by simulating under the assumption that the isotonic estimates are true, in order to estimate the adequacy of the transformation against any other possible shape. That would be equivalent to a test T_{12} where H_1 : the isotonic transformation is tested against any other possible shape H_2 : no restrictions for the p 's.

2.2 Reduced isotonic regression

In the previous section, we described how PAVA detects violators of the isotonicity assumption and builds level sets by amalgamating adjacent groups until there are no more violators. However, some of the resulting level sets can be pooled together, especially those with few elements or those whose estimated values do not differ much from their neighbours. Moreover, it has been shown that the use of isotonic regression overfits somewhat the data whereas a model with fewer level sets (and therefore fewer degrees of freedom) may fit better.¹⁰ Thus, once isotonic regression is fitted, we need to proceed with a backward elimination in order to improve the parsimony of the model.

In order to compute the eliminated isotonic regression, two steps have to be considered: first, which level sets can be pooled together and, second, when should the pooling procedure be stopped? Several methods can be applied to answer these questions. Schell and Singh¹⁰ propose an F-test when the response is continuous. For binary response Bacchetti¹¹ reduces the partial fitted functions in the additive isotonic model by comparing the change in the likelihood to a considered but *ad hoc* amount.

In contrast to Bacchetti we propose a reducing procedure for binary response based on Fisher's exact test for contingency tables: to identify the level sets that do not significantly differ, one has to look at all 2×2 tables for the adjacent level sets. The 'pairs' that are not proven to differ significantly are pooled together. The procedure ends when all pivotal tables give significant p -values. To clarify the elimination procedure, we will describe the backward algorithm used to reduce the degrees of freedom in a one-dimensional isotonic regression.

Let the isotonic regression summarize the dose in L constant risk groups and estimated proportions \hat{p}_i^* , $i = 1, \dots, L$ and n_l observations falling at the l th level set. The aim of the elimination procedure is to reduce the groups to S level sets ($S < L$) with respect to the outcome. The algorithm for elimination can be described as follows.

2.2.1 Algorithm for reduced isotonic regression

- 1) Construct all $L - 1$ contingency tables for the adjacent level sets and calculate $L - 1$ exact Fisher tests and their corresponding p -values.
- 2) If all p -values $< \epsilon^*$, where ϵ^* is a predefined significance level, then stop. Else, go to step 3.
- 3) From the set of level pairs resulting in a p -value $> \epsilon^*$ select the one with the greatest p -value and pool it. This reduces the number of level sets by one. Go to step 1.

Obviously the reduced isotonic regression depends on the choice of ϵ^* . For $\epsilon^* = 1$ the reduced isotonic regression is identical to the isotonic level sets whereas for $\epsilon^* = 0$ we get a single level set. The use of $\epsilon^* = 0.05$ in the backward elimination will not yield an overall 0.05-level test as usual i.e., if the H_0 assumption of constant risk holds, the elimination procedure will not yield a single level set with probability greater than 5%. This is not surprising since we base the elimination procedure on a maximal selected p -value and we face a multiple comparison problem.

Isotonic framework is poorly supported by asymptotic theory, especially in the case of binary response. We lack a theoretical solution, thus we will use simulations to assess the ϵ^* that will yield an overall significance level of 5%. All we need to do is to simulate by producing random noise data (no association between dose and response) and then

to assess in each data set the isotonic estimators and their reduced equivalents using $\epsilon^* = 0$. In each replication we retain the p -value from the last Fisher test when only two level sets remain to be pooled. The corrected ϵ^* is the 5% value from the distribution of all those ‘end’ p -values. A similar approach¹⁰ has been used in order to correct the significance level in the F-test used to reduce a continuous response regression.

Another crucial point in reduced isotonic regression is whether the reduced model or its parent isotonic model should be used. Up to now, no distribution theory is available for these models, so the AIC, BIC or the determination coefficient R^2 must be used to choose between simple and more complex models. Alternatively, one can apply a sort of parametric bootstrap.¹⁵ The term ‘parametric’ refers to the idea that the data set at hand is assumed to be extracted from a population whose distribution F is known, although here the underlying model (reduced isotonic regression) is not parametric. To be more precise we claim that under the assumption that the reduced model is the correct one, the reduced \hat{p}_i^* s are an estimator of F . Following the notation of Efron and Tibshirani,¹⁵ the measured function of interest for a data set x is $\theta(x) = D_{\text{reduced}}(x) - D_{\text{full}}(x)$ with D denoting the deviance. The procedure can be described as follows.

2.2.2 Parametric bootstrap for selecting model

- 1) Generate B simulated data sets x_j^* from F .
- 2) In each x_j^* assess the isotonic and the reduced model and the corresponding deviances.
- 3) Assess $\theta(x_j^*) = D_{\text{reduced}_j^*} - D_{\text{full}_j^*}$ for $j = 1, \dots, B$.
- 4) If the 95% interval of $\theta(x_j^*)$ s contains the observed value from the original sample $\theta(x_{\text{obs}}) = D_{\text{reduced}}^{\text{obs}} - D_{\text{full}}^{\text{obs}}$, prefer the reduced isotonic model to the full isotonic model, since the observed improvement in the fit for the full model can be expected by its higher number of level sets.

The elimination procedure can be implemented to more sophisticated monotonic models, as for example the additive isotonic model. In ref. 11, an additive model for binary response is fitted and the need to reduce the degrees of freedom in the monotonic partial fitted functions is discussed. However, the elimination is accomplished by comparing the loss in the fit to an arbitrary amount. The reduced monotonic regression could be used instead, although that would potentially increase the computational complexity. It is intuitively simpler to combine the elimination procedure with a monotonic-surfaces model, as will be described in section 4.

3 Simulation study

3.1 Estimation of ϵ^*

3.1.1 Design A simulation study is conducted to explore the parameters that can influence ϵ^* . We simulate under different values for sample size ($N = 100, 200, 300, 600, 900$) and positive response rate ($p = 0.02, 0.05, 0.10, 0.15, 0.25$). The desired significance level has been set to the nominal value $\alpha = 0.05$. We applied the algorithm described in the previous section to 5000 samples from random noise data with

Table 2 Estimation of ϵ^* based on 5000 simulations for different sample sizes and response rates. The overall significance level was 5%

p_0	N				
	100	200	300	600	900
0.02	0.0398	0.0218	0.0120	0.0117	0.0093
0.05	0.0188	0.0129	0.0097	0.0071	0.0066
0.10	0.0126	0.0089	0.0074	0.0064	0.0053
0.15	0.0103	0.0080	0.0074	0.0059	0.0057
0.25	0.0101	0.0077	0.0077	0.0055	0.0043

predictor $X_i \sim U[0, 1]$ and response $Y_i \sim B(p)$. The exact significance level has been estimated through permutations as described in the previous chapter. The results are depicted in Table 2.

3.1.2 Results The estimated ϵ^* decreases as long as the sample size and the response rate increase. While the decrease is sharp for small p , it flattens out with greater response rate. For this simulation study, we assumed that the predictor variable X has no duplicates. The ϵ^* is slightly greater if there are some, and it becomes clearly greater if the variable X is categorical. For example, with sample size 100 and event rate 10% the estimated ϵ^* is 0.0227 when X is in four categories, i.e., about double the value when X is used as continuous (tabulated value 0.0126 in Table 2). Moreover, when the iterative algorithm for isotonic matrix data is used instead of PAVA (see section 4), the values in Table 2 no longer hold. Thus, we do not find it useful to estimate an approximate formula for ϵ^* , although that could potentially facilitate the elimination procedure. Instead, we propose to apply simulations to estimate it for every data set at hand.

3.2 Comparison of isotonic and reduced isotonic regression

3.2.1 Design We performed a limited simulation study to explore the benefits of using reduced isotonic regression instead of the full isotonic model. Two criteria were used: first, the *number of level sets* LS as a measure of model complexity (LS_{red} and LS_{full}); and secondly, as a measure of the model fit, while several criteria are possible, we used the *coefficient of determination* \bar{R}^2 as defined in ref. 16 for binary response models:

$$\bar{R}^2 = \frac{R_{LR}^2}{R_{max}^2} = \frac{1 - e^{-(LR/n)}}{1 - e^{-(D_0/n)}} \tag{4}$$

where LR is the difference in the deviance between reduced and full model and D_0 the deviance for the null model. \bar{R}^2 measures the ‘variation explained by the model’. The better model is the one with greater \bar{R}^2 and less complexity, i.e., less LS . Reduce isotonic regression decreases the model complexity, but it is also expected to reduce \bar{R}^2 . With this simulation study we want to estimate if the decrease in the complexity is worth the loss of fit.

Three parameters have been studied: regression shape, sample size, and R_{\max}^2 . We consider again a predictor variable $X \sim U[0, 1]$. Four regression lines have been analysed: a) linear LIN: $\text{logit}(p) = aX$; b) quadratic QUA: $\text{logit}(p) = aX^2$; c) hockey-stick HOK: $\text{logit}(p) = c + aX \cdot I_{\{X > \text{median}(X)\}}$; and d) step function STE: $\text{logit}(p) = c + a \cdot I_{\{X > \text{median}(X)\}}$, where $I_{\{\text{condition}\}}$ is an index that takes the value 1 if condition is satisfied and 0 otherwise. We simulate these functions under sample size $N = 100, 300, 500$. In each shape the parameter a has been determined such that the maximum coefficient of determination would be $R_{\max}^2 = 0.3, 0.5, 0.7$. That is actually equivalent to different assumptions about the positive response probability (about $p_0 = 4, 11, \text{ and } 29\%$ respectively).

3.2.2 Results Regarding the complexity of the model, the number of LS_{full} increased with sample size and R_{\max}^2 (range of mean value: 3.23–14.34). The same trend was observed for the number of the reduced LS_{red} , but the variation was not very important (range of mean value 1.23–3.88). The elimination procedure reduces the number of level sets to about one third of the starting isotonic level sets. It is important to note that fraction $LS_{\text{red}}/LS_{\text{full}}$ becomes smaller with increasing sample size.

Figure 1 presents the results regarding the change in \bar{R}^2 . On the x -axis is the sample size and on the y -axis is the relative ‘loss’ of fit $[= (\bar{R}_{\text{full}}^2 - \bar{R}_{\text{red}}^2)/\bar{R}_{\text{full}}^2]$ when the reduced model is used. Recall that we wish to have as similar \bar{R}_{full}^2 and \bar{R}_{red}^2 as possible, that is, small loss of fit.

The difference between the coefficients of determination for the two models becomes smaller with increasing sample size. However, the influence of the maximum value of the coefficient, or the response probability is very important. While for smaller R_{\max}^2 the isotonic model has considerably better fit than the reduced model, its advantage is not important when $R_{\max}^2 = 0.7$ (the reduced model reduces the coefficient \bar{R}^2 only by 7%). Regarding the different underlying shapes, the linear regression clearly presents the worst tolerance on reducing the model, whereas the results of HOK and STE were the best for every R_{\max}^2 .

These findings, together with the results in model complexity reduction, enable us to conclude that when R_{\max}^2 is at least 0.5 and the investigator believes that the regression line is segmented, reduced isotonic regression controls quite successfully the trade-off between model complexity and fit.

4 The isotonic-surfaces model

The multiple regression setting: consider we have N observations on a dependent binary variable Y denoted by $\mathbf{p} = (p_1, p_2, \dots, p_N)^T$ measured at N designed vectors $\mathbf{d}^i = (d_{i1}, d_{i2}, \dots, d_{iP})$, assuming P predictor variables $D_j, j = 1, \dots, P$. We want to model the dependence of Y on D_1, \dots, D_N having two principal goals: to describe, for learning more about the process that produces the outcome, and to infer, i.e., assess the relative contribution of each variable to the response probability.

Additive isotonic models start from the assumption that the risk (response) does not decrease as long as any of the predictors increases, and extend generalized additive models¹³ by letting isotonic transformation act in the partial fitted functions. The local

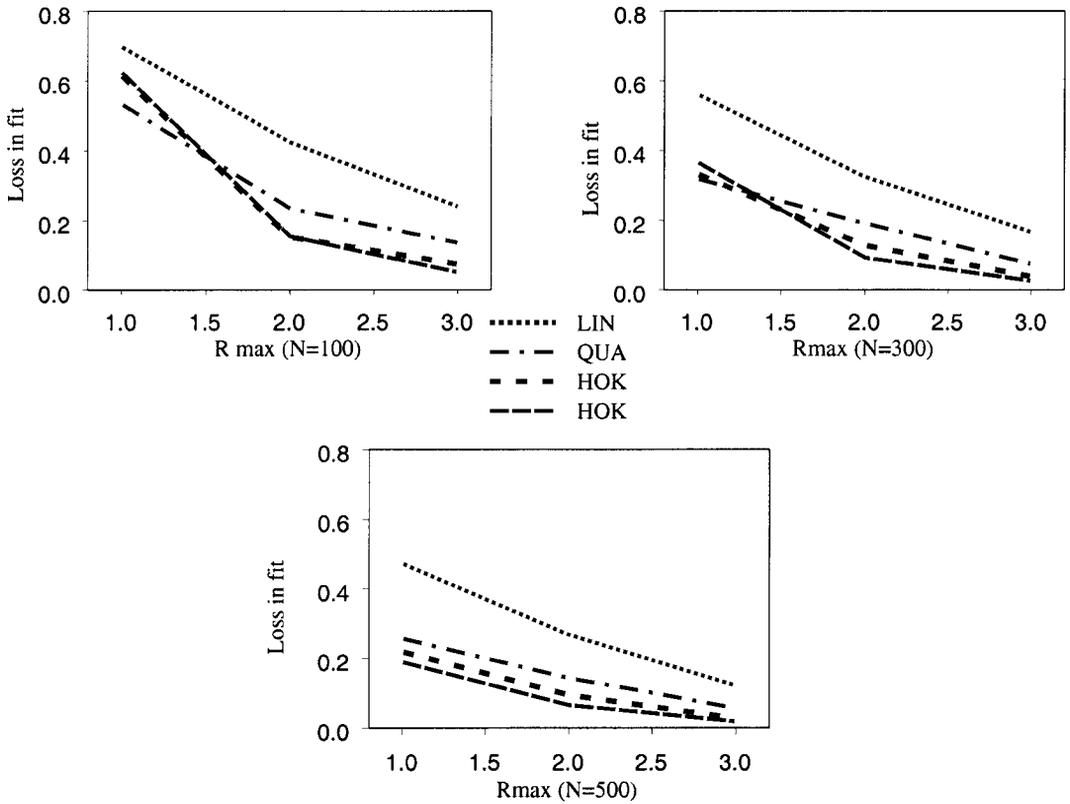


Figure 1 Results from simulation study: comparing monotonic regression and reduced monotonic regression regarding the relative loss in the fit as function of the regression shape

scoring algorithm usually used in generalized additive models is replaced here by PAVA and the contribution to the risk of each isotonic variable is a nondecreasing step function. The additive isotonic model takes the form:

$$b(p_i) = \sum_{j=1}^P \tilde{\phi}^*(\mathbf{d}^i), \tag{5}$$

where b is a link function and $\tilde{\phi}^*$ denotes a P -dimensional isotonic function $\phi^* = (\phi_1^*, \phi_2^*, \dots, \phi_P^*)$. Each ϕ_j^* is estimated by PAVA, which replaces the local scoring algorithm commonly used for estimation in generalized additive models. The estimation proceeds via the backfitting algorithm.

An alternative approach can be motivated by the fact that the isotonic procedure can be thought of as a ‘scatterplot smoother’. This consideration raises the question: How can isotonic regression be extended to ‘smooth’ a plot when one sets restrictions over multiple axes? In other words, following the same logical pattern as in the univariate case, how can we produce isotonic estimates in a three- (or even higher) dimensional

plot? The isotonic ‘smoothing’ through PAVA and the reducing procedure can be extended to more than one dimension (p predictors), applying an iterative algorithm. The idea is to fit a model of the form

$$p_i = \phi^*(d_{i1}, d_{i2}, \dots, d_{ip}) \tag{6}$$

where ϕ^* is the isotonic transformation and d_{ij} the i th observation of the j th predictor. Consider, for example, the case of two explanatory variables and imagine the data in a form of a matrix M . In the (i, j) cell falls the outcome \hat{p}_{ij} of the individual being in the i th category of the first variable and the j th category of the second one. Note that a matrix is isotonic with respect to the partial order if and only if the elements \hat{p}_{ij} of M fulfil the restriction $\hat{p}_{ij} \leq \hat{p}_{kl}$ for $i \leq k$ and $j \leq l$. The algorithm to assess the isotonic estimators works as follows.

4.1 The isotonic-surfaces algorithm

Step 1: Let M^{*1} denote the isotonic regression of M over rows. Let $R^1 = (M^{*1} - M)$ be the first set of row increments.

Step 2: Let M denote the isotonic regression over columns of $M + R^1$. Call $C^1 = M^{**1} - (M + R^1)$ the first set of column increments.

Step 3: At the beginning of the n th cycle M^{*n} is obtained by isotonizing $M + C^{n-1}$ over rows. The n th set of row increments is defined by $R^n = M^{*n} - (M + C^{n-1})$. Next, obtain M^{**n} by isotonizing $M + R^n$ over columns.

Both M^{*n} and M^{**n} converge to the isotonic regression M^{***} , with respect to the partial order. The result of a two-dimensional isotonic regression can be visualized as a surface that is nondecreasing as long as any of the predictors increases. The algorithm combines both the explanatory variables in l constant risk groups (the level sets), and therefore each step in the response variable corresponds to a specified bivariate group for the predictors.

In theory, the algorithm for isotonic surfaces can be extended to more than two variables. For a third factor in τ -ordered levels the result would be a sequence of τ -isotonic surfaces, each of them lying above the previous one or touching each other. However, in practice, if more than three isotonic predictors need to be included in the model, the use of this approach is not recommended due to its great computational complexity.

A main problem arising from this algorithm is that the convergence is not guaranteed in the case where the data contains many zero-weighted cells. Therefore the predictor variables need to be in preselected groups. Even if those groups are many, very thin and selected objectively (for example, using quantiles) we suspect that their choice can affect the results somewhat because of the decrease in the number of the candidate changepoint locations. However, this procedure captures interactions between the explanatory variables, a feature that the additive isotonic model described in ref. 12 does not provide.

The significance of any predictor included in the model is assessed again by the likelihood ratio test. There is no known large sample approximation for its distribution, so once more permutations are used to calculate the p -value of the overall fit and

conditional permutations for the effect of each variable included in the model adjusted for the other predictors. The term ‘conditional’ for the two-dimensional case refers to the following: given the observed marginal distribution of the events to the level sets of one predictor A to be true (the likelihood estimated, say, at the columns), we assess the probability to have the observed distribution at the cells (overall likelihood).

The permutation procedure for partial significance can be summarized as follows: In each response p_i corresponds to the vector $\mathbf{d}^i = (d_{i1}, d_{i2}, \dots, d_{iP})$. To test the effect of the j th predictor adjusted for the remaining $P - 1$ predictors, we split vector \mathbf{d}^i after the j th variable D_j and then combine $(p_i, d_{i1}, \dots, d_{i,j-1}, d_{i,j+1}, \dots, d_{iP})$ and $d_{i,j}$ randomly. In each combination the isotonic regression is fitted and the corresponding test T_{01} is computed. To reject then the H_0 assumption, the 95th quantile of the empirical distribution of the deviance is compared to the observed T_{01} value. Of course one can test all predictors at once if so desired by randomly combining Y_i to x_i and to follow the same procedure as described above. As in the univariate case, one can construct confidence surfaces for the estimates, simulating under H_0 or H_1 .

The algorithm described in section 2.2 for reducing the number of level sets, can also be applied in an isotonic surface. Each level set is compared to its neighbouring using the Fisher test. Those who do not differ are pooled together until the estimated ϵ^* level is achieved. The estimation of ϵ^* and the method to choose between isotonic model and reduced model are as described for the univariate case. Note that assuming a ‘multiple’ order restriction and applying the iterative algorithm, one obtains more isotonic level sets than applying PAVA over one dimension. Therefore, the estimated ϵ^* for a two-dimensional (or higher) elimination procedure is dramatically smaller than those in Table 2 depending on the number of restrictions. That is simply because ϵ^* decreases while the number of comparisons increases, and one has more LS (and thus more pivotal tables) when multiple predictors are taken into account.

Note that we outlined multivariate monotonic methodology considering only increasing trend can be extended for a given mixture of isotonic and antitonic explanatory variables. The algorithms described above with the corresponding tests are implemented in S language and are available from the first author on request.

5 Application

The data used to exemplify the proposed methods are taken from a study conducted by the German Research Foundation.³ The aim of the study was to investigate the influence of dust concentration (in mg/m^3) in the working area on chronic bronchitic reaction. The time since first exposure (in years) was also taken into account, and a two-dimensional model (6) was fitted. The amount of dust was categorized in 17 quantiles and time in 10 quantiles. As noted in the methodology part, this decision can effect the results. It would be more adequate to construct more than 17 quantiles, but the data are not that precise.

We depict the result of isotonic regression in Figure 2. The permutations procedure was applied to perform an overall test for the model and conditional permutations as described in section 4 were used to assess the statistical significance of the effect of dust given the effect of time. For the overall test the p -value for the observed value

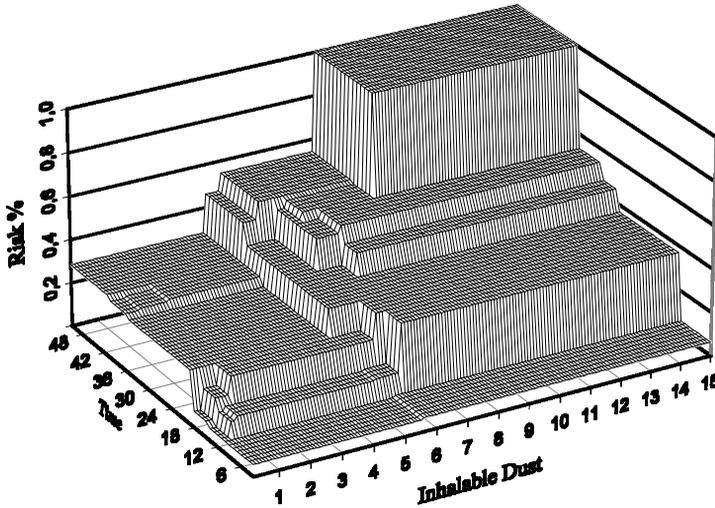


Figure 2 The isotonic-surfaces model (40 level sets)

$T_{01} = 98.3$ (see Table 1) was less than 0.001 based on 1000 permutations. The conditional test for the improvement in the fit after entering the dust in the model ($1008.65 - 959.87 = 48.78$) results in a p -value = 0.002. On fitting the reduced model, we used simulations to assess the ϵ^* -level. For this purpose we produced 5000 random permutations of the response variable as if it was independent on the explanatory variables (170 pairs). In each permutation we fitted the isotonic and reduced isotonic regression. Then to get the ϵ^* that leads to a 0.05-significance test, we picked the 250th smallest p -value when only two level sets remain. The estimated ϵ^* was 0.00038. Figure 3 presents the reduced model.

The change in the deviance between isotonic and reduced model is 16.58 (Table 3). The number of level sets has been reduced from 40 (isotonic) to 3 (reduced isotonic). The cutpoints for dust in the reduced model were at concentrations 0.9, 4.5, 5.5 and 5.8 mg/m³. We conduct a last step in order to compare these two models: simulating (1000 simulations) under the assumption that the reduced model is the correct one, we conclude that such a large change in the likelihood as the observed could have occurred with probability p -value = 0.632 and we choose the more parsimonious model. Thus, applying reduced isotonic regression we found a useful stratification for both variables time and dust, by combining them in three groups of higher and lower risk.

The additive isotonic model was also applied to the data. The model has deviance 999.94 and summarizes the dust in three groups (cutpoints: 4.5 and 7.4 mg/m³), and time in 4 groups, that is, a total of 12 level sets (Figure 4). In Table 4 we compare the three models: isotonic, reduced isotonic and additive isotonic. Reduced isotonic regression controls better the trade-off between fit and model complexity. Recall that

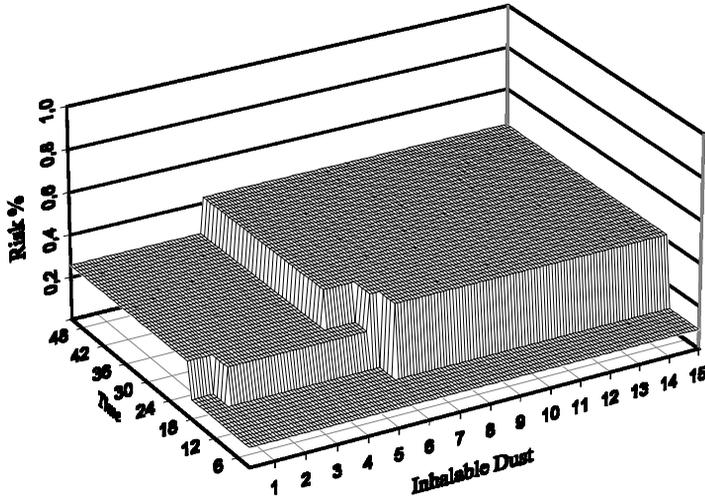


Figure 3 The reduced-monotonic model (three level sets)

our proposal is rather adequate for stratification and thus it would be relevant to compare the different models using ROC curves. The area under the curve was 0.658 for additive model, 0.690 for the isotonic surface and 0.677 for the reduced surface.

In light of this consideration we also applied a classification tree to the data. The results are presented in Table 5 and they correspond to a final tree with three terminal nodes, which have been selected after cross validation. The predictors are combined in three groups. The result is roughly similar to the reduced-surface model result. Note that the group with the higher risk in the classification tree is defined by $\text{time} \geq 16.5$ and $\text{total dust} \geq 4.8$. This high risk group is also to be seen in the reduced-surface model (Figure 3) where the estimated proportion is 0.42. The dust cutpoint of about 5 mg/m^3 is also present in the additive model.

Table 3 The deviance of monotonic and reduced models

Model	Deviance
H_0	1058.17
Monotonic (time)	1008.65
Monotonic (dust)	1025.65
Monotonic (dust and time)	959.87
Reduced monotonic	976.45

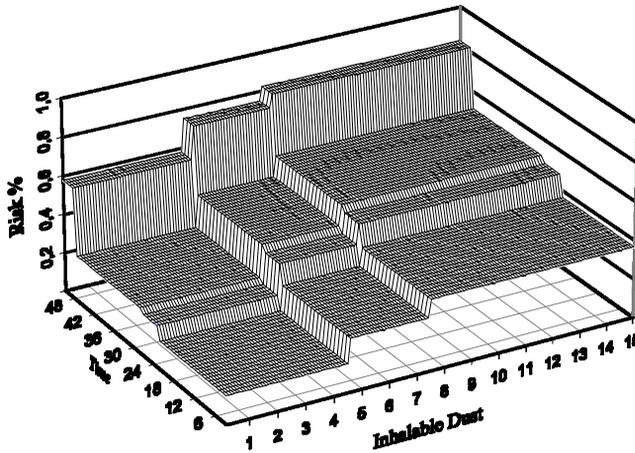


Figure 4 The additive-monotonic model (12 level sets, seven parameters)

Table 4 Several criteria to compare monotonic surface, reduced surface and additive monotonic

Model	Deviance	df	AIC	BIC	R^2
Monotonic surface	959.87	40	1039.87	1096.36	0.148
Reduced surface	976.45	3	982.45	986.69	0.085
Additive monotonic	999.94	7	1013.94	1023.83	0.061

6 Discussion

The analysis of a dose–response relationship is an important tool in medicine as well as in epidemiology. The proof of an association is one of the criteria needed in order to establish causality. Several methods are available to model association. In recent papers very flexible ways have been developed, such as fractional polynomials or smoothing splines.

Table 5 The final classification tree in numbers

Node	Covariate	Deviance	Deviance reduction	n	Proportion
1 (root)	Time < 16.5	1058.0	43.7	920	
2*	Time ≥ 16.5	169.50		243	0.111
3	Dust < 4.8	844.80	23.8	677	
4*	T < 16.5 and D < 4.8	481.90		429	0.249
5*	T < 16.5 and D ≥ 4.8	339.10		248	0.4315

Although much debated by many statisticians, the use of step functions in modelling can prove to be very useful in many settings. Step functions are easy to interpret and they fit changepoint models that summarize the predictors in such a way as to define groups of constant response. While step function models are attractive, it is not straightforward how one should select the cutpoints. As a useful alternative, monotonic regression can be applied which results in a monotonic step function. The method selects certain intervals with equal response under the constraint of monotonicity, and the dose–response relationship is of course not smooth. However, the main advantage is the immediate use for a practical purpose. In medicine there are only a limited number of treatment options available; for example, in treating high blood pressure the physician often needs to categorize the patients in several groups regarding the severity of hypertension. Monotonic models are adequate when the goal of the analysis is to establish a dose–response relationship and to find an optimal stratification for the explanatory variables.

In the paper we used sometimes the term ‘smoother’ for the monotonic procedure. This may be excessive, since the result is far from being smooth because of the presence of ‘flat spots’ in an increasing regression. With smoothing we refer here more to the fact that ‘monotonic regression is connected to conditional expectation and this conditioning is referred to as a smoothing process: the values of the variables are regressed and replaced in the conditioning process by constant values, which is a smoothing operation’.¹⁴ In light of this consideration, the monotonic smoother has a global nature but results in locally flat averaging.

Two multivariate methods are available: the additive-monotonic model and the monotonic-surfaces model. The additive model, ignoring any interaction, can be used for continuous variables. The monotonic-surfaces model does allow interaction, but it cannot handle more than three variables. Owing to the number of possible combinations between the explanatory variables, the continuous ones have to enter the model as ordinal, but the result may depend on the starting cutpoints. We propose to use quantiles to categorize the amount of dose, in order to ensure convergence for the iterative algorithm and to minimize bias that can occur from ‘pretty-cuts’ (as, for example, by selecting round exposure levels). This approach could be combined to the additive model: the isotonic surface can be applied to deviance residuals in order to model interactions and obtain an hierarchical model.

An important advantage of the monotonic framework is the monotonic likelihood ratio test. The univariate version of the monotonic test [equations (1) and (2)] has recently gained much attention in applications, despite the fact that its asymptotic distribution does not hold for binary response. The reason for its popularity is that the monotonic test results independently of the quantification of the dose and does not imply the linearity assumption, unlike the widely used Cochran–Armitage test or the Mantel-extension test for trend. Mancuso *et al.*⁸ showed that setting the monotonic transformation as alternative to the constant risk assumption results in a test with increased power.

Several proposals have been made so far regarding multivariate tests for trend. The Mantel-extension test can be modified to handle more than one categorical variable, but it is expected to present the same drawbacks as in the univariate case. The logistic regression offers an alternative, but the linearity assumption remains a constraint. The

T-contrast test and Dosemeci–Benichou test⁶ can be extended to more than one variable, but as mentioned in their paper many details remain to be sorted out. Regarding monotonic regression we propose a test based on the monotonic-surface model to deal with more than one explanatory variable. It is a likelihood ratio test where the critical value is computed using permutations, and can be overall (testing all variables included in the model) or partial (assessing the influence of a variable adjusted for the others). The main problem remains the restriction to a maximum of three predictors to be used. Another equivalent approach can be accomplished applying additive-monotonic models and thereafter the same overall and partial permutation procedure as before. A comparative study of those multivariate tests could provide useful information. However, we believe the main characteristics of a test are the same, independently of the number of variables taken into account.

The reduced monotonic regression for binary response has also been introduced. The extension of the reduced method for two-dimensional monotonic-surface models is straightforward. The reducing procedure focuses on finding a subset of the cutpoints resulting from the monotonic regression, by selecting those that correspond to a significant increase in the risk. The model becomes more parsimonious but the selection of the best model (comparison of the monotonic model to the reduced one) should be based on simulations. That, obviously, is the main drawback arising from the reduced monotonic framework in binary response: the lack of an appropriate approximation for the distribution of the likelihood ratio test statistic.

References

- 1 Galloway S, Clark G. Practical p -value adjustment for optimally selected cutpoints. *Statistics in Medicine* 1996; **15**: 103–12.
- 2 Lausen B, Scumacher M. Maximally selected rank statistics. *Biometrics* 1992; **48**: 73–85.
- 3 DFG. *List of MAK and BAT values 2002*, Report no. 38, Weinheim: Wiley, 2002.
- 4 Ulm K. Nonparametric analysis of dose-response relationships. *Annals New York Academy of Sciences* 1999; **895**: 223–31.
- 5 Salanti G, Ulm K. Modelling under order restrictions. Discussion paper No. 265 Sonderforschungsbereich 386 der Ludwigs-Maximilian-University Munich, 2001.
- 6 Leuraud K, Benichou J. A comparison of several methods to test for the existence of a monotonic dose-response relationship in clinical and epidemiological studies. *Statistics in Medicine* 2001; **20**: 3335–51.
- 7 Ulm K, Dannegger F, Becker U. Tests for trends in binary response. Discussion paper 115 Sonderforschungsbereich 386 der Ludwigs-Maximilian-University Munich, 1998.
- 8 Mancuso J, Ahn H, Chen J. Order-restricted dose-related trend tests. *Statistics in Medicine* 2001; **20**: 2305–18.
- 9 Chuang-Stein C, Agresti A. Tutorial in biostatistics: A review of tests for detecting a monotone dose-response relationship with ordinal response data. *Statistics in Medicine* 1997; **26**: 2599–618.
- 10 Schell MJ, Singh B. The reduced monotonic regression method. *Journal of the American Statistical Association* 1997; **92**: 128–35.
- 11 Bacchetti P. Additive monotonic models. *Journal of the American Statistical Association* 1989; **84**: 289–94.
- 12 Morton-Jones T, Diggle P, Parker L. Additive monotonic regression models in epidemiology. *Statistics in Medicine* 2000; **19**: 849–60.
- 13 Hastie T, Tibshirani R. Generalized additive models for medical research. *Statistical Methods in Medical Research* 1995; **4**: 187–96.
- 14 Robertson T, Wright FT, Dykstra RL. *Order restricted statistical inference*. New York: John Wiley, 1988.
- 15 Efron B, Tibshirani R. *An introduction to the bootstrap*. Chapman & Hall, 1993.
- 16 Nagelkerke N. A note on a general definition of the coefficient of determination. *Biometrika* 1991; **78**: 691–92.

