Scandinavian Journal of Statistics
Theory and Applications

Scandinavian Journal of Statistics
doi: 10.1111/j.1467-9469.2010.00716.x


# A Method for Bayesian Monotonic Multiple Regression

OLLI SAARELA

*Department of Chronic Disease Prevention, National Institute for Health and Welfare*

ELJA ARJAS

*Department of Mathematics and Statistics, University of Helsinki and National Institute for Health and Welfare*

**ABSTRACT. When applicable, an assumed monotonicity property of the regression function w.r.t. covariates has a strong stabilizing effect on the estimates. Because of this, other parametric or structural assumptions may not be needed at all. Although monotonic regression in one dimension is well studied, the question remains whether one can find computationally feasible generalizations to multiple dimensions. Here, we propose a non-parametric monotonic regression model for one or more covariates and a Bayesian estimation procedure. The monotonic construction is based on marked point processes, where the random point locations and the associated marks (function levels) together form piecewise constant realizations of the regression surfaces. The actual inference is based on model-averaged results over the realizations. The monotonicity of the construction is enforced by partial ordering constraints, which allows it to asymptotically, with increasing density of support points, approximate the family of all monotonic bounded continuous functions.**

*Key words:* Bayesian non-parametric regression, marked point process, model-averaged inference, model selection, monotonic regression

## 1. Introduction

The assumption of monotonicity in regression modelling can be used in place of more restrictive parametric or structural assumptions such as linearity or additivity when there is adequate prior knowledge of the studied phenomenon to justify the monotonicity assumption, and enough data to allow for non-parametric modelling. Compared with smoothing methods where no assumptions are made on the shape of the estimated function, using the monotonicity postulate adds stability to the function estimate by making it less sensitive to random fluctuations and outliers in the data. In this article, we propose a monotonic non-parametric regression procedure for one or more covariates, incorporating the principle of Bayesian model averaging. For simplicity, in the following, we use only the term monotonic to describe the regression functions. Any isotonic function can be made monotonic by reversing some of the coordinate axes, so there is no real need to make a distinction between monotonic, isotonic and antitonic situations.

The traditional monotonic regression problem with one covariate is the restricted least squares optimization, where each datapoint has its own ordered level parameter. The computational method for solving this problem in one dimension is known as 'pool adjacent violators' (PAV; Ayer *et al.*, 1955; Barlow *et al.*, 1972; applied to time-to-event data by Ancukiewicz *et al.*, 2002). In the PAV algorithm, adjacent function values violating monotonicity are replaced by their mean until the function becomes non-decreasing. With several covariates present in the model, solving the monotonic regression problem is less straightforward, except in the special case of generalized additive models (GAM; see, e.g. Morton-Jones *et al.*, 2000; Cheng, 2008). In the general multivariate case, the optimization is carried out under

partial ordering constraints. Algorithmic solutions for this problem are considered by, for example, Beran & Dümbgen (2010). We do not further consider such approaches here, since constrained optimization is a separate problem to the fully probabilistic modelling aimed at in this article.

Another family of monotonic regression methods is based on splines of a different kind, either with fixed or random knot points. Ramsay (1988) discussed integrated splines where the spline is made monotone by using monotonic basis functions with non-negative coefficients. An estimation procedure for such splines in GAMs was presented by Tutz & Leitenstorfer (2007). If the spline basis functions are not monotone, monotonicity of the spline has to be achieved by placing constraints on the spline parameters. For B-splines, it is enough to order the adjacent basis coefficients. An estimation procedure for B-splines in the GAM framework was presented by Leitenstorfer & Tutz (2007). Brezger & Steiner (2008) considered penalized B-splines in GAMs, using a Bayesian approach where a roughness penalty was applied by postulating an autoregressive smoothing model for the spline coefficients. Cai & Dunson (2007) used a somewhat similar formulation, but allowing for multiple dependent outcome variables. Alternatives to splines are wavelets; a monotone penalized estimation method for these was presented by Antoniadis *et al.* (2007).

Continuing on purely Bayesian developments, Arjas & Gasbarra (1994) considered non-parametric modelling of a hazard rate in one dimension, based on piecewise constant random functions, and presented also a solution for estimating an increasing hazard rate. Holmes & Heard (2003) considered a univariate regression model, where the data are grouped according to the covariate of interest, with random cutpoints and with a level parameter for each group. A monotonic estimate for the regression function was obtained by sampling from the unconstrained model and by taking the part of the posterior sample where the level parameters were consistent with the monotonicity constraints. Although such an approach is computationally convenient, it is unlikely to be applicable in higher dimensions as the probability of obtaining such a consistent set of values gets quickly smaller when the number of covariates grows larger. Neelon & Dunson (2004) used a piecewise linear model with a built-in monotonicity constraint of non-negative slopes, with a large number of fixed cutpoints and with an additional autoregressive smoothing prior for the slope parameters. In addition, they placed a probability point mass at zero values of the slope parameters, thus allowing for flat regions in the regression function. A similar mixture distribution idea was used by Dunson (2005) for a count response, but instead using a piecewise constant model, where the monotonicity was realized through multiplicative increment parameters restricted to be greater or equal to one, with a probability point mass at one. Bornkamp & Ickstadt (2009) considered an approach where the non-parametric monotonic regression function was built as a mixture of parametric probability distribution functions, using a general random probability measure as the mixing distribution.

Recently, independently from this work, Bornkamp *et al.* (2010) presented a multivariate extension to Bornkamp & Ickstadt (2009), coupled with the density regression approach of Dunson *et al.* (2007). To date, this appears to be the only method presented for multivariate Bayesian monotonic regression. Bornkamp *et al.* (2010) note that finding computationally tractable necessary conditions ensuring monotonicity of multivariate functions is a non-trivial problem. Instead, they proceed with a sufficient condition, enforcing monotonicity by considering linear combinations of monotonic components with non-negative coefficients. By doing so, they lose some generality, however; for instance, adding up monotonic main effects and monotonic interaction terms does not define a general monotonic relationship (a point we further demonstrate below). In contrast, we aim at finding a construction which can asymptotically approximate the family of all monotonic multivariate regression functions when

the complexity of the construction is increased. To achieve this, as in the traditional monotonic regression problem, we enforce monotonicity of our construction directly in terms of partial ordering constraints which define monotonicity. However, instead of constrained optimization, the computation relies on integration using Markov chain Monte Carlo (MCMC) methods. The computation can then be carried out by making small local modifications on the regression surface. Hence, we only ever need to consider the partial ordering constraints locally, which greatly facilitates the computational task.

Our monotonic construction is based on a marked point process formulation, where the random point locations and the associated marks (function levels) together define a piecewise constant regression surface. Unlike the traditional approaches where a parameter is associated with each datapoint, our method is not directly tied to the observed datapoints. This allows the algorithm to favour parsimonious models irrespective of the number of observations. Another important property of the method is that it allows reductions into lower dimensional submodels when some of the covariates in the model are redundant, thus acting as a device in model selection. Being based on a full probability model, our approach provides a natural quantification of the uncertainty in the function estimate. The monotonic construction can accommodate any type of likelihood, and can also be plugged into a part of a larger probability model. This, for example, allows the packaging of covariates as in Arjas & Liu (1996), where covariates can be divided into blocks, some of which are modelled non-parametrically and some by using parametric functions.

The plan of the article is as follows. Section 2 defines the mathematical framework for the monotonic construction and discusses some examples of possible model parameterizations. Section 3 first presents an illustration with simulated data. In an illustration using real data, we apply the model for absolute risk estimation. The article concludes with a discussion in section 4.

## 2. Mathematical framework

We are interested in non-parametric modelling of the association between a group of covariates $x_1, \ldots, x_p$ and a response variable $y$. A probability model for $y$ is defined as:

$$p(y \mid \lambda(x_1, \ldots, x_p), \theta). \tag{1}$$

Here, $\lambda : \mathbb{R}^p \to A$ is a single realization of a random monotonic function of the covariates $x_1, \ldots, x_p$, while $\theta$ includes possible other parameters. $A \subseteq \mathbb{R}$ is the set where the regression function is defined. In regression problems, monotonicity is conventionally defined in terms of the mean of the response variable. This would also be natural in the present case if the response variable is assumed to be Bernoulli, Gaussian or Poisson distributed. In these cases, $\lambda(x_1, \ldots, x_p)$ would represent the mean of the response variable given the covariates, with $A = [0, 1]$, $\mathbb{R}$ or $\mathbb{R}^+$, respectively. Examples with Gaussian and Bernoulli likelihoods are presented in sections 3.1 and 3.2, respectively. However, definition of $\lambda$ is by no means restricted to these special cases. For example, if the model for $y$ is defined using a monotonic link function, $\lambda$ can be chosen to replace the linear predictor in such a model. In the above Bernoulli case, a natural example would be the logit link with $A = \mathbb{R}$. In more complex probability models, $\lambda$ could also represent higher levels in hierarchical parameterizations. An example of this would be a situation where observed covariate values involve measurement error and $\lambda$ would be defined in terms of the true (latent) covariate values. For a further discussion on measurement error models under Bayesian framework, we refer to, for example, Gustafson (2003).

The aim is to consider Bayesian inference based on the posterior distribution of functions $\lambda$, given observed data on the response and the covariates. Our explicit construction of $\lambda$ postulates this function to have a piecewise constant form, each realization being defined by a finite number of random support points at which the function is assigned a random level. This results in variable dimensional modelling, where the main interest is in model-averaged inference rather than in any single realization of the random regression function. The Bayesian paradigm carries an implicit penalty for model complexity, which under vague priors usually results in a sparse characterization of the relationships between variables. This is because Bayesian model comparison can be understood as being based on the marginal probability of the observed data (integrated over the parameters), and a large model can accommodate a wider range of different observations (see, e.g. Dawid, 1984). Ultimately, the main interest may lie in predictive distributions of the response values rather than in the posterior distribution of $\lambda$ itself. The predictive distributions are naturally obtained as a side product of Bayesian estimation.

For notational simplicity, in the following we assume that $x_1, \ldots, x_p$ are each scaled to interval $[0, 1]$. Let now $\mathcal{I} \subseteq 2^{\{1,\ldots,p\}}$, where $2^{\{1,\ldots,p\}}$ denotes the set of all non-empty subsets of the covariates. We define the marked point processes

$$\Delta_i = \{(\xi_{ij}, \delta_{ij})\} \subset [0,1]^{|I_i|} \times A, \tag{2}$$

for all $I_i \in \mathcal{I}$, $i = 1, \ldots, |\mathcal{I}|$. (In the general case, the point processes can as well be defined in general spaces $S(\Delta_i) \times A$ depending on the observable ranges of the covariates.) Let $n(\Delta_i)$ denote the (random) number of points in a realization of the process $\Delta_i$. We define 'completed' points $\tilde{\xi}_{ij} = (\tilde{\xi}_{ij}^1, \ldots, \tilde{\xi}_{ij}^p) \in [0,1]^p$ by letting $\tilde{\xi}_{ij}^k = \xi_{ij}^k$ when $k \in I_i$, and $\tilde{\xi}_{ij}^k = 0$ when $k \in \{1, \ldots, p\} \setminus I_i$. In addition, we introduce a fixed point and the corresponding mark located at the origin: $\Delta_0 = (\xi_{01}, \delta_{01})$, where $\tilde{\xi}_{01} = (0, \ldots, 0)$. We can now define a natural partial ordering for the points in the set $\bigcup_{i=0}^{|\mathcal{I}|} \bigcup_{j=1}^{n(\Delta_i)} \tilde{\xi}_{ij}$, denoting $\tilde{\xi}_{ij} \preceq \tilde{\xi}_{rs}$ if $\tilde{\xi}_{ij} \in \prod_{k=1}^p [0, \tilde{\xi}_{rs}^k]$, and equivalently $\tilde{\xi}_{rs} \succeq \tilde{\xi}_{ij}$ if $\tilde{\xi}_{rs} \in \prod_{k=1}^p [\tilde{\xi}_{ij}^k, 1]$. (Here, $\prod_{k=1}^p [0, \tilde{\xi}_{rs}^k]$ and $\prod_{k=1}^p [\tilde{\xi}_{ij}^k, 1] \subseteq [0,1]^p$ denote Cartesian products, and can be viewed as, respectively, lower and upper corner sets associated with points $\tilde{\xi}_{rs}$ and $\tilde{\xi}_{ij}$.) The partial ordering of points is then combined with an ordering of the marks, by postulating that $\delta_{ij} \leq \delta_{rs}$ when $\tilde{\xi}_{ij} \preceq \tilde{\xi}_{rs}$. This means that, given points (and marks) in the lower and upper corner sets of a point $\xi_{ij}$, the mark $\delta_{ij}$ is restricted to interval $\delta_{ij} \in [\max\{\delta_{rs} : \tilde{\xi}_{rs} \preceq \tilde{\xi}_{ij}\}, \min\{\delta_{rs} : \tilde{\xi}_{rs} \succeq \tilde{\xi}_{ij}\}]$, where $(r,s) \neq (i,j)$.

We now construct a monotonic piecewise constant regression function $\lambda$ using such a point process formulation. First note that at an arbitrary point $x = (x_1, \ldots, x_p)$ the regression function $\lambda(x)$ is constrained to be in the interval $[\max\{\delta_{ij} : \tilde{\xi}_{ij} \preceq x\}, \min\{\delta_{ij} : \tilde{\xi}_{ij} \succeq x\}]$. With this in mind, we define $\lambda$ by

$$\lambda(x) = \max\{\delta_{ij} : \tilde{\xi}_{ij} \preceq x\}, \tag{3}$$

resulting in a piecewise constant realization of the regression function. Owing to the fixed point located at origin, the lower corner set $\{\xi_{ij} : \tilde{\xi}_{ij} \preceq x\} \neq \emptyset$ for all $x \in [0,1]^p$. The monotonicity follows from (3) by considering an arbitrary pair of points $x, x' \in [0,1]^p$, $x' \preceq x$:

$$\begin{aligned} x' \preceq x &\Rightarrow \{\xi_{ij} : \tilde{\xi}_{ij} \preceq x'\} \subseteq \{\xi_{ij} : \tilde{\xi}_{ij} \preceq x\} \\ &\Rightarrow \max\{\delta_{ij} : \tilde{\xi}_{ij} \preceq x'\} \leq \max\{\delta_{ij} : \tilde{\xi}_{ij} \preceq x\} \\ &\Rightarrow \lambda(x') \leq \lambda(x). \end{aligned}$$

Functions of the form (3) can approximate an arbitrary monotonic bounded continuous function $f$ when the number of support points is increased. This follows readily from uniform continuity of such functions. Because of that, for any $\varepsilon > 0$, we can introduce a grid of points

in $[0,1]^p$ such that, in each 'box' created by the grid, the values of $f$ differ by at most $\varepsilon$. The approximation then follows if we select the grid points as support points $\tilde{\xi}_{ij}$ of the construction (3), and let $(\tilde{\xi}_{ij}, \delta_{ij}) = (\tilde{\xi}_{ij}, f(\tilde{\xi}_{ij}))$ at these points. Thus, by increasing the density of points, the approximation can be made arbitrarily accurate. The example of section 3.1 provides a practical demonstration of this property.

A model choice representing no prior information on the point locations $\xi_{ij}$ would be the homogeneous Poisson process, defined by the constant intensity parameters $\rho_i$. Many other choices, possibly incorporating prior information, would be possible. For instance, the model for the point locations could be defined as a repulsive point process, such as the Strauss process. In the case of the homogeneous Poisson process, if the priors for the intensity parameters are defined as $\rho_i \sim \text{Gamma}(a, b)$, the posterior distribution for these parameters, given the current realization of points, is $\rho_i \sim \text{Gamma}(a + n(\Delta_i), b + |S(\Delta_i)|)$. (In our examples, we have $S(\Delta_i) = [0,1]^{|I_i|}$, with $|S(\Delta_i)| = 1$.)

The marks $\delta_{ij}$ can be assumed (jointly) uniformly distributed in the space restricted by the partial ordering constraints imposed by the point locations and some appropriate bounded interval $[\delta_{\min}, \delta_{\max}] \subseteq A$ (see Appendix). Such a prior definition involves no additional parameters and makes no assumption on the shape of the regression function other than the monotonicity and the restriction to a bounded interval. It should be noted that, for this reason, it also does not provide any information for extrapolating the regression function outside the domain of the data, but this is the necessary price of non-parametric analysis. In principle, it would be possible to consider also other types of priors involving hierarchical parameterization, or to consider $\delta_{\min}$ and $\delta_{\max}$ as random variables. However, the homogeneous Poisson prior for the point locations with uniform distribution for the marks is a good benchmark, as the model formulation involves no smoothing component (in addition to the monotonicity postulate itself) and thus resembles the traditional univariate monotonic regression problem described by, for example, Schell & Singh (1997). The resulting regression surface estimate is completely driven by the data, thus being as close to non-parametric modelling as is possible with this construction. An MCMC algorithm for producing samples from the posterior distribution of $\lambda$, given the data, is described in the Appendix. The main idea here is to attempt birth, death and location change moves in each of the point processes in turn. In addition, individual marks are updated in turn given the set of other marks and the partial ordering defined by the point locations. This approach avoids the need to consider multiple ordered random variables and is thus very easy to implement in practice.

Another model which may be of interest, although it does not define general monotonicity, would be obtained by dropping the ordering constraints for $\delta_{ij}$, setting $A = \mathbb{R}^+$, and then postulating $\lambda(x)$ to have an additive form

$$\lambda(x) = \sum_{i=0}^{|\mathcal{I}|} \sum_{j=1}^{n(\Delta_i)} \mathbf{1}_{\{\tilde{\xi}_{ij} \preceq x\}} \delta_{ij}. \tag{4}$$

Such functions are readily seen to be monotonic, using a similar reasoning as with function (3), but they are limited in the shapes of regression surfaces they can represent. In the terminology used by Kong & Lee (2006, 2008), functions defined by (4) cannot represent antagonistic interactions. Model (4) is of interest, because when $\mathcal{I} = \{\{1\}, \ldots, \{p\}\}$ or $n(\Delta_i) = 0$, $i \in \mathcal{I} \backslash \{\{1\}, \ldots, \{p\}\}$, it reduces to a GAM type model. One possible parameterization of the model (4) could be introduced by following Arjas & Gasbarra (1994), and using a common exponential prior distribution for the marks of each point process.

It would not be too difficult to come up with continuous function realizations defined using realizations of the point processes (2), based on various interpolations between the partially ordered support points $(\tilde{\xi}_{ij}, \delta_{ij})$. However, generally, such functions will not be monotonic.

For instance, piecewise linear surfaces constructed from $(p+1)$-dimensional hyperplanes, each going through $p+1$ support points $(\tilde{\xi}_{ij}, \delta_{ij})$, could with a simple counterexample be shown to be non-monotonic. A monotonic continuous realization, defined at an arbitrary point $x$, could, however, be obtained by defining a rectangular grid based on the point coordinates, consisting of boxes $\prod_{k=1}^{p}[\max\{\tilde{\xi}_{ij}^k : \tilde{\xi}_{ij}^k \leq x_k, 0\}, \min\{\tilde{\xi}_{ij}^k : \tilde{\xi}_{ij}^k \geq x_k, 1\}] \subseteq [0,1]^p$, and defining $\lambda(x)$ to be the weighted average of the maximum mark values found in the lower corner set of each of the $2^p$ corners of the box containing $x$. The weights in this construction are taken to be inversely proportional to the volumes of the $2^p$ boxes bounded by the above box and the coordinates of $x$. However, this approach loses the computational efficiency of the piecewise constant construction, and more importantly, its ability to perform local updating moves. When we briefly experimented with such a continuous approximation, the result was a very poorly mixing chain. Thus, we proceed with the computationally convenient piecewise constant approximation (3).

## 3. Illustrations

### *3.1. Model checking*

In this section, we demonstrate the ability of the model (3) to represent a wide variety of regression surfaces, and to produce correct model fit in terms of residuals. Consider the estimation of three dimensional regression surfaces in the case of $p=2$ and $\mathcal{I}=\{\{1\},\{2\},\{1,2\}\}$. Let us suppose that we have $n$ observations of a response $y$ and covariates $x_1$ and $x_2$. The likelihood function of such data expresses the probability (density) of $y$ as a function of a random realization of the piecewise constant regression function (3), now evaluated at $x=(x_1, x_2)$. To simulate some data, we take $x_{k1}, x_{k2} \sim U[0,1]$ and $y_k \mid x_{k1}, x_{k2} \sim N(\mu(x_{k1}, x_{k2}), \sigma^2)$, $k=1,\ldots,1000$, varying the residual standard deviation $\sigma$. Priors were chosen as described before, with $a=b=0.1$, $\delta_{\min}=-1$ and $\delta_{\max}=2$. Six very different shapes for regression function $\mu$ were tried, all with values in the interval $[0,1]$. The considered functions were $\mu_1=\sqrt{x_1}$, $\mu_2=0.5x_1+0.5x_2$, $\mu_3=\min(x_1,x_2)$, $\mu_4=0.25x_1+0.25x_2+0.5\times 1_{\{x_1+x_2>1\}}$, $\mu_5=0.25x_1+0.25x_2+0.5\times 1_{\{\min(x_1,x_2)>0.5\}}$ and $\mu_6=1_{\{(x_1-1)^2+(x_2-1)^2<1\}}\times\sqrt{1-(x_1-1)^2-(x_2-1)^2}$. Likelihood (1) of the data $y=(y_1,\ldots,y_{1000})$ was defined as the product of independent likelihood contributions from normally distributed observations with mean $\lambda(x_{k1}, x_{k2})$ and standard deviation $\sigma$.

The samples from the posterior distribution of $\lambda$, $\sigma$ and $\rho$ were obtained from running the sampler, described in the Appendix, for 25,000 iterations after a burn-in period of 25,000 rounds, and saving every fifth state of the chain. For a comparison to model (3), we also fitted the model (4) to the same data, using the same homogeneous Poisson process prior for the point locations and the hierarchical $\delta_{ij} \mid v_i \sim \exp(v_i)$, $v_i \sim \text{Gamma}(0.1, 0.1)$ priors for the marks. Various statistics of model fit are presented in Tables 1 and 2. Here, posterior mean residuals and predicted values are compared with their true counterparts from the normal distribution used to simulate the data. The last three columns are model fit statistics proposed by Spiegelhalter *et al.* (2002).

Table 1 shows that, with all six monotonic regression functions, our method based on model (3) produces a good model fit in terms of standard deviation of the posterior mean residuals and the correlations between the true and the posterior mean residuals. With data simulated using the values $\sigma=0.5$ and $\sigma=0.1$, it does so with a moderate amount of support points used, with the distribution of the posterior mean residuals very close to the true residuals from the normal distribution (an example in Fig. 1). The smallest residual standard deviation $\sigma=0.01$ enables the estimation procedure to use a very high density of support points to approximate the regression surface. Perspective plots of the posterior mean estimates

Table 1. *Statistics of model fit; model* (3). *The columns correspond to the residual standard deviation* $\sigma$ *and regression function* $\mu$ *used in simulating the observations, posterior mean* $\sigma$, *standard deviation of posterior mean residuals, correlation between true and posterior mean residuals, correlation between true and posterior mean predicted responses, posterior mean number of support points used, deviance at posterior means, effective number of parameters and deviance information criterion*

| $\sigma$ | $\mu$ | $\bar{\sigma}$ | SD($\bar{\varepsilon}$) | cor($\varepsilon, \bar{\varepsilon}$) | cor($y, \bar{y}$) | $\sum_{i=1}^{\bar{3}} n(\Delta_i)$ | $D$ | $p_D$ | DIC |
|---|---|---|---|---|---|---|---|---|---|
| 0.5 | $\mu_1$ | 0.499 | 0.495 | 0.999 | 0.443 | 16.4 | 1430.8 | 14.7 | 1460.3 |
| | $\mu_2$ | 0.487 | 0.478 | 0.990 | 0.467 | 20.4 | 1360.0 | 37.2 | 1434.3 |
| | $\mu_3$ | 0.497 | 0.490 | 0.994 | 0.480 | 9.8 | 1409.6 | 29.5 | 1468.7 |
| | $\mu_4$ | 0.511 | 0.499 | 0.980 | 0.577 | 26.6 | 1449.0 | 44.9 | 1538.7 |
| | $\mu_5$ | 0.504 | 0.500 | 0.993 | 0.518 | 7.7 | 1450.3 | 15.5 | 1481.3 |
| | $\mu_6$ | 0.482 | 0.471 | 0.987 | 0.595 | 24.6 | 1333.8 | 44.7 | 1423.2 |
| 0.1 | $\mu_1$ | 0.099 | 0.098 | 0.995 | 0.921 | 57.7 | −1808.7 | 30.4 | −1747.9 |
| | $\mu_2$ | 0.094 | 0.088 | 0.963 | 0.924 | 88.2 | −2020.7 | 133.5 | −1753.8 |
| | $\mu_3$ | 0.099 | 0.095 | 0.979 | 0.930 | 17.4 | −1866.4 | 73.0 | −1720.4 |
| | $\mu_4$ | 0.100 | 0.096 | 0.964 | 0.963 | 57.4 | −1855.4 | 91.7 | −1672.0 |
| | $\mu_5$ | 0.099 | 0.096 | 0.984 | 0.953 | 32.2 | −1851.8 | 73.5 | −1704.8 |
| | $\mu_6$ | 0.096 | 0.087 | 0.955 | 0.969 | 99.4 | −2023.2 | 165.1 | −1693.1 |
| 0.01 | $\mu_1$ | 0.010 | 0.009 | 0.965 | 0.999 | 274.5 | −6533.3 | 114.2 | −6304.9 |
| | $\mu_2$ | 0.008 | 0.004 | 0.774 | 1.000 | 1137.9 | −7533.3 | 676.3 | −6180.7 |
| | $\mu_3$ | 0.010 | 0.008 | 0.881 | 0.999 | 68.9 | −6685.9 | 316.0 | −6053.9 |
| | $\mu_4$ | 0.009 | 0.007 | 0.850 | 1.000 | 410.3 | −7015.5 | 459.5 | −6096.6 |
| | $\mu_5$ | 0.009 | 0.006 | 0.870 | 1.000 | 461.0 | −7082.2 | 462.4 | −6157.5 |
| | $\mu_6$ | 0.010 | 0.007 | 0.693 | 1.000 | 648.6 | −6821.7 | 521.8 | −5778.1 |

Table 2. *Statistics of model fit; model* (4). *The columns correspond to the residual standard deviation* $\sigma$ *and regression function* $\mu$ *used in simulating the observations, posterior mean* $\sigma$, *standard deviation of posterior mean residuals, correlation between true and posterior mean residuals, correlation between true and posterior mean predicted responses, posterior mean number of support points used, deviance at posterior means, effective number of parameters and deviance information criterion*

| $\sigma$ | $\mu$ | $\bar{\sigma}$ | SD($\bar{\varepsilon}$) | cor($\varepsilon, \bar{\varepsilon}$) | cor($y, \bar{y}$) | $\sum_{i=1}^{\bar{3}} n(\Delta_i)$ | $D$ | $p_D$ | DIC |
|---|---|---|---|---|---|---|---|---|---|
| 0.5 | $\mu_1$ | 0.499 | 0.496 | 0.998 | 0.441 | 12.8 | 1433.0 | 11.8 | 1456.5 |
| | $\mu_2$ | 0.490 | 0.485 | 0.996 | 0.437 | 13.7 | 1391.9 | 19.0 | 1429.8 |
| | $\mu_3$ | 0.498 | 0.493 | 0.996 | 0.469 | 13.0 | 1422.0 | 19.5 | 1460.9 |
| | $\mu_4$ | 0.527 | 0.521 | 0.958 | 0.523 | 19.7 | 1533.6 | 21.3 | 1576.2 |
| | $\mu_5$ | 0.515 | 0.513 | 0.965 | 0.481 | 3.4 | 1500.0 | 10.0 | 1520.0 |
| | $\mu_6$ | 0.488 | 0.482 | 0.983 | 0.570 | 10.2 | 1377.1 | 26.3 | 1429.6 |
| 0.1 | $\mu_1$ | 0.099 | 0.098 | 0.993 | 0.921 | 35.9 | −1812.7 | 31.5 | −1749.6 |
| | $\mu_2$ | 0.097 | 0.095 | 0.986 | 0.911 | 42.4 | −1873.3 | 46.1 | −1781.1 |
| | $\mu_3$ | 0.099 | 0.096 | 0.979 | 0.928 | 21.9 | −1847.9 | 59.6 | −1728.7 |
| | $\mu_4$ | 0.167 | 0.163 | 0.606 | 0.888 | 28.3 | −793.8 | 53.1 | −687.5 |
| | $\mu_5$ | 0.159 | 0.157 | 0.613 | 0.867 | 12.1 | −864.3 | 28.0 | −808.3 |
| | $\mu_6$ | 0.111 | 0.107 | 0.847 | 0.952 | 17.3 | −1629.9 | 63.7 | −1502.4 |
| 0.01 | $\mu_1$ | 0.010 | 0.009 | 0.959 | 0.999 | 158.7 | −6561.7 | 134.1 | −6293.5 |
| | $\mu_2$ | 0.009 | 0.008 | 0.936 | 0.999 | 214.2 | −6698.5 | 190.0 | −6318.6 |
| | $\mu_3$ | 0.010 | 0.009 | 0.850 | 0.999 | 67.8 | −6563.7 | 210.3 | −6143.1 |
| | $\mu_4$ | 0.131 | 0.127 | 0.093 | 0.927 | 28.7 | −1285.1 | 56.7 | −1171.7 |
| | $\mu_5$ | 0.125 | 0.123 | 0.065 | 0.910 | 14.8 | −1347.4 | 33.4 | −1280.6 |
| | $\mu_6$ | 0.055 | 0.052 | 0.170 | 0.988 | 25.7 | −3066.5 | 99.1 | −2868.3 |

of the six different function shapes in this case are shown in Fig. 2. These values were evaluated at all points of a $50 \times 50$ square grid on $[0, 1]^2$. They illustrate the previously noted consistency in approximating continuous functions when the density of support points is increased, while the structure (3) can naturally handle the discontinuities in the cases $\mu_4$ and $\mu_5$. In the model fit, the support points are placed where they are most needed; for example,

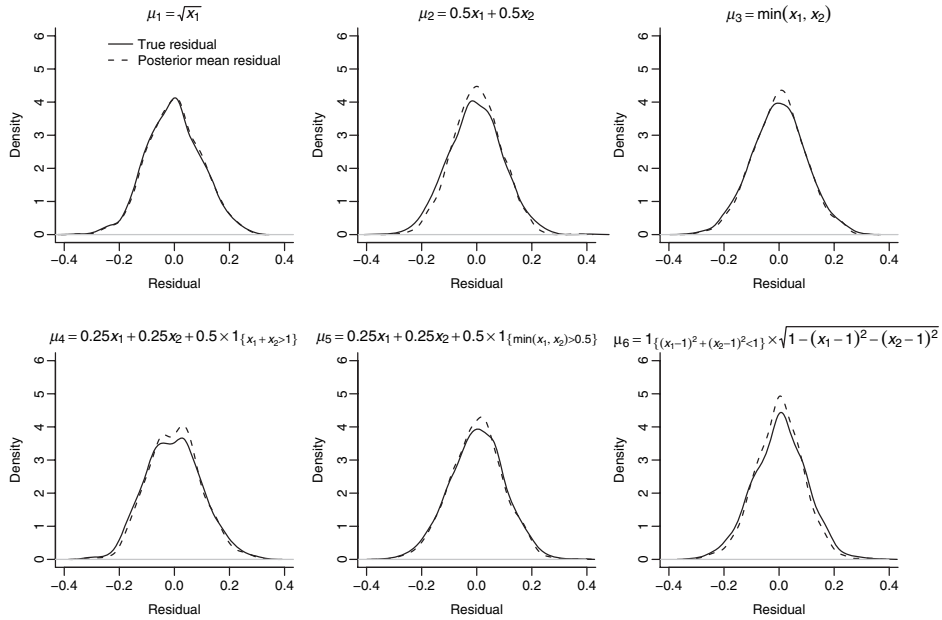


*Fig. 1.* Posterior mean residuals from model (3) compared with true residuals. Lines are density estimates. Observations simulated with $\sigma = 0.1$ and varying the regression function $\mu$.

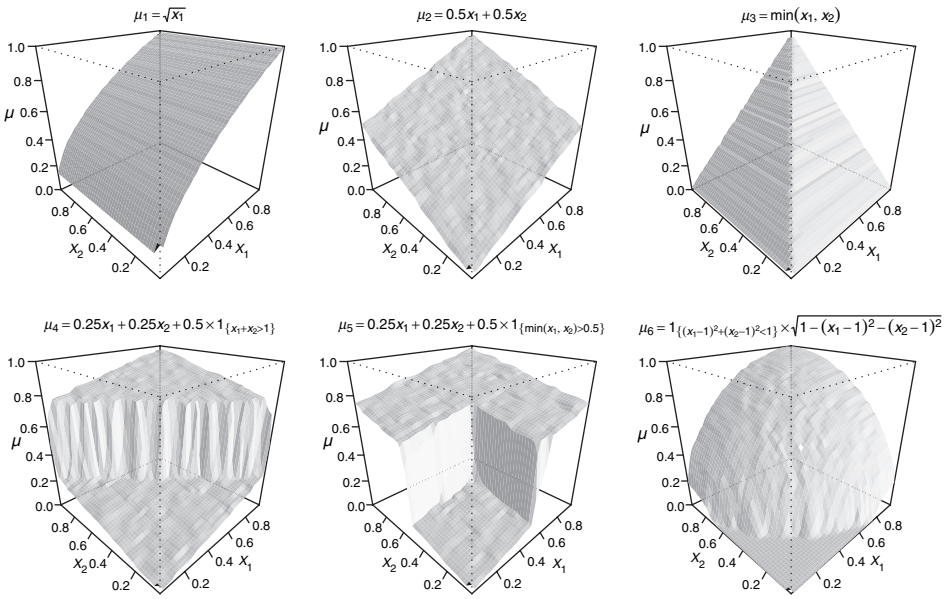


*Fig. 2.* Perspective plots for posterior mean regression surfaces from model (3). Observations simulated with $\sigma = 0.01$ and varying the regression function $\mu$.

the case $\mu_6$ shows that the estimation procedure has used no support points where the true regression surface is flat and a high density of points where the surface is rapidly increasing.

Table 2 shows that, in the case of the first three function shapes, the method using the additive model (4) produces results very similar to those obtained with the general model

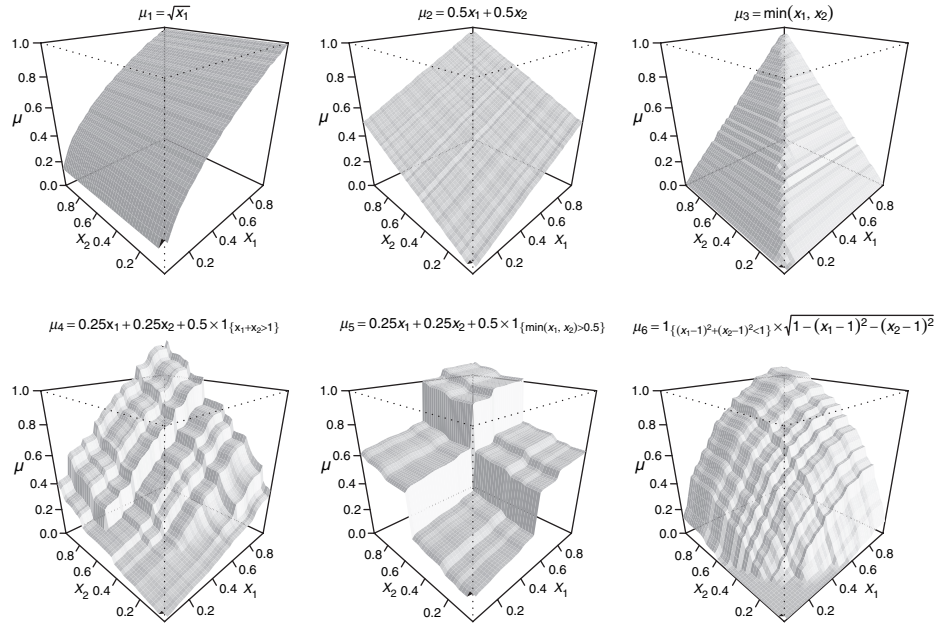


*Fig. 3.* Perspective plots for posterior mean regression surfaces from model (4). Observations simulated with $\sigma = 0.01$ and varying the regression function $\mu$.

(3). However, as expected, when the residual standard deviation of the data is decreased, the additive model fails miserably in its attempts to reproduce the non-additive function shapes (Fig. 3). (It should be noted that the function $\mu_3 = \min(x_1, x_2)$ is also non-additive, but can be approximated by model (4) by placing support points with constant-sized marks on the diagonal of the $(x_1, x_2)$ plane. However, such a shape could not be reproduced with a GAM type model.) It is interesting to note that in the additive case $\mu_2$, with $\sigma = 0.01$, the method based on the general model (3) provides a clearly better model fit in terms of the deviance $D$. However, because of increased complexity of this model, it loses slightly to model (4) in terms of deviance information criterion (DIC). It is well known that Bayesian model comparison consistently finds the 'true' model when such a model is included in the space of candidate models. Now, the general monotonic model obviously cannot select an additive model, but it nevertheless is able to produce a good approximation, in terms of DIC, by 'overfitting' at the cost of greater model complexity. These differences are only visible in the case of the smallest value of the error variance; in the other cases, the estimated residuals from the general monotonic model correspond very closely to their true counterparts. To check the effect of sample size, we estimated all combinations in Tables 1 and 2 also with $n = 500$ (results not shown). The conclusions on the model fit and the comparison between the general and additive models remained essentially unchanged.

### *3.2. Prediction*

In this example, we consider data from two population-based cohorts, FINRISK 87 (FR87) and FINRISK 92 (FR92). These cohorts are a part of the National FINRISK Study (Salomaa *et al.*, 1996; Vartiainen *et al.*, 2000) and are included in the MORGAM project, an international pooling of cardiovascular cohorts (Evans *et al.*, 2005). The cohorts were recruited as random samples from geographical populations between ages 24 and 64 and have

been followed up for various disease endpoints and all-cause mortality. Here, we consider the task of estimating 15-year risk of death because of any cause of men from North Karelia and Kuopio regions in eastern Finland, based on a few risk factors recorded at the start of the follow-up. Another application of the proposed method in predictive inference is presented in Arjas & Saarela (2010).

After omissions because of missing covariate data, we included 1995 men from FR87 (of whom 320 had died during the first 15 years of follow-up) as a training set and 1389 men from FR92 (173 deaths during the first 15 years of follow-up) as a validation set. Mortality rate in Finland had decreased between these two studies and thus risk estimates derived using the earlier cohort are not necessarily well calibrated in the later cohort. However, here we concentrate on the discriminative ability of the prediction model, and for this comparison, the later cohort can well be used as the validation set.

We consider the task of estimating absolute risk of death based on age at the start of the follow-up ($x_1$), smoking as self-reported average number of cigarettes smoked per day ($x_2$) and waist-to-hip ratio (WHR; $x_3$). Of the FR87 and FR92 men, 65.9 and 67.3 per cent, respectively, were non-smokers and for them the number of cigarettes was taken to be zero. We use WHR as an indicator of obesity, instead of the more common body mass index, because its association with mortality can more reasonably be assumed as monotonic (Bigaard *et al.*, 2004). We compared three alternative nested models, $\mathcal{I}=\{\{1\}\}$, $\mathcal{I}=\{\{1\},\{2\},\{1,2\}\}$ and $\mathcal{I}=\{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$, corresponding to models including age only, age and smoking, and age, smoking and WHR, respectively. The aim was to demonstrate that the proposed method performs sensibly when the dimension increases, in the sense that the model fit in the training set has to improve when more covariates are added to the model. We fitted the monotonic regression models to data in the training set and then used the resulting posterior predictive probabilities as risk estimates for the validation set. The priors and the estimation procedure were as described in the previous section, using the Bernoulli likelihood for the death event ($y$), with $\delta_{\min}=0$ and $\delta_{\max}=1$ and the covariates $x_1$, $x_2$ and $x_3$ rank transformed and scaled to interval $[0,1]$.

Figure 4A shows the model fit for the three alternative models in the training set. Because of the wide age range of the study cohorts, age alone is a very strong predictor. Adding smoking to the model clearly improves the fit, whereas the effect of a further addition of WHR

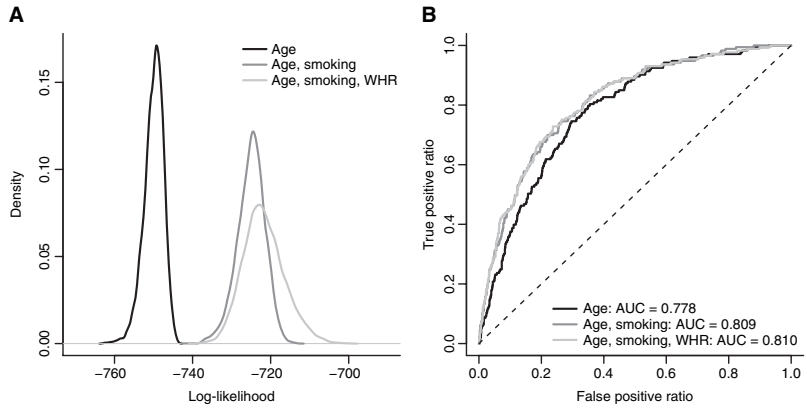


*Fig. 4.* Comparison between three nested risk models. (A) Model fit in training set. (B) Model discrimination in validation set. Lines in (A) are density estimates and lines in (B) are receiver operating characteristics (ROC) curves.

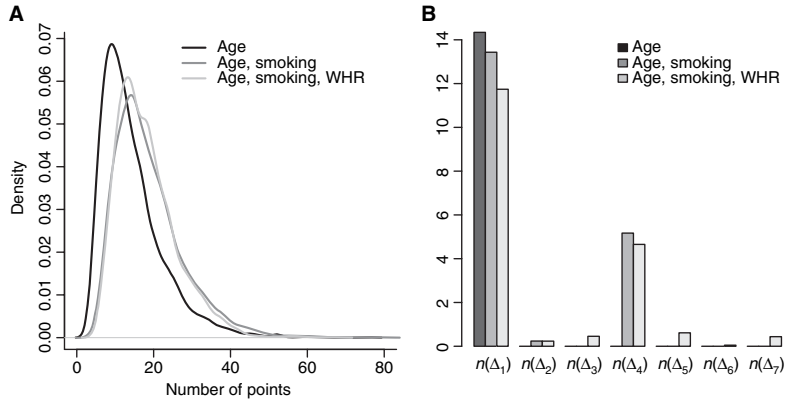


*Fig. 5.* Number of support points used in fitting the risk models. (A) Total number of points. (B) Mean number of points by process. Lines are density estimates and bars are posterior means.

is more modest. However, in both cases the mean model fit improves compared with the baseline model. Figure 4B shows the performance of the prediction model in the validation set, evaluated by using receiver operating characteristic curves and areas under the curves (AUC). Using only age achieved 77.8 per cent discrimination. When smoking was added, AUC improved by 3.1 percentage points to 80.9 per cent (jackknife standard error for the difference 0.7 per cent), but the further addition of WHR led to only a slight increase (0.1 per cent, SE 0.1 per cent). As could be expected, accounting for the effect of smoking improved the accuracy of the risk estimates particularly at the lower levels of absolute risk, that is, among the younger men in the cohort. Figure 5 shows the number of support points used in the model fit. As age is by far the strongest of the three predictors, most of these points were used on the one-dimensional point process on the age axis. When smoking was added to the model, some support points were also placed on the two-dimensional process on the age–smoking plane, to reflect the increased risk of smokers. In contrast, the distribution of the points remained almost unchanged when WHR was added to the model. Figure 6 shows the risk estimates in the validation set, derived as posterior means from the model with age and smoking. A notable feature of this figure is the rapid increase in the risk for the smokers after age 50.

## 4. Discussion

In areas such as epidemiology, strong parametric assumptions are often imposed on the form of the regression function describing covariate effects. This is done as a modelling convention, sometimes without real support from contextual substantive arguments, evidence coming from earlier studies or careful diagnostics afterwards. In contrast, completely non-parametric estimation of a regression surface may give a too complex ('noisy') picture of the true functional relationship as small disturbances in the data affect the function estimate. However, often there are reasonable grounds to assume a monotonic relationship between a covariate and the considered response.

In this article, we have formulated a mathematical construction for multidimensional non-parametric monotonic regression and proposed a Bayesian estimation procedure. The proposed method does not pretend to be able to break the curse of dimensionality; when moving into higher dimensions, the data quickly get sparse, and thus large amounts of data

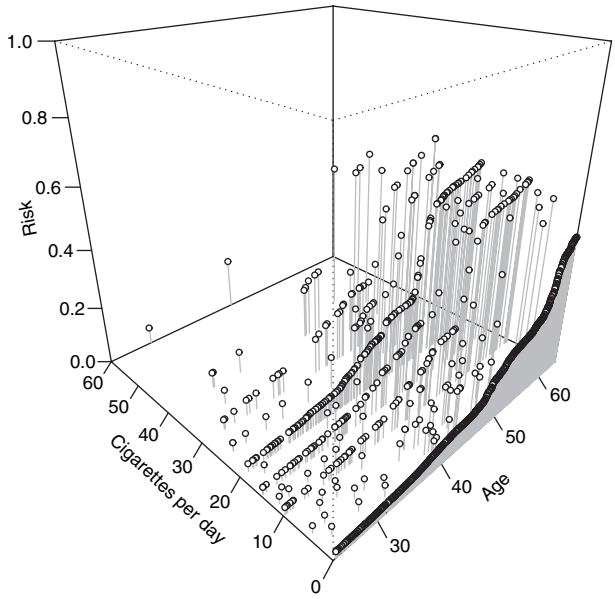


*Fig. 6.* Estimated 15-year risks of death for the validation set derived from the model with age and number of cigarettes per day.

and/or strong associations are still needed to detect truly multidimensional relationships. This is the necessary limitation of multidimensional regression without additional parametric or structural assumptions. In contrast, GAMs do not suffer from high dimensionality problems because of their assumption of independent covariate effects. For this reason, we introduced also an additive alternative based on the same point process formulation [formula (4)]. However, this readily illustrates the problem in such a construction; assuming an additive structure where the main effects and interactions should have monotonic shapes already implies that additional constraints have been imposed on the otherwise general monotone multiple regression function.

Some readers may find our use of a piecewise constant representation of multivariate functions as overly simplistic, feeling that some greater degree of smoothness should be required. Indeed, we do not contend that piecewise constant functions are 'true', or that they would be, in the versions they come up in the estimation algorithm, very close to such true functions in some suitable metric. Greater smoothness would be more important if we were seeking point estimates of the regression function, whereas here our main interest is in posterior distributions and the corresponding predictive probabilities, which involve integrations over the space of random functions, and in our case, possibly even over models in different dimensions. Thus, the space of piecewise constant functions should be primarily viewed as a skeleton over which suitable numerical approximations of such integrals can be found in a computationally efficient manner. Moreover, as we saw in section 3.1, the piecewise constant model is able to provide a correct fit in terms of residuals.

In this article, we presented the proposed method in its most basic form with as few modelling assumptions as possible. Further work is needed to find out whether the method presented here would benefit from further hierarchical parameterizations, to help in problems in which the dimension would be much higher than in the examples that were considered here.

**Acknowledgements**

The authors would like to thank Dr Juha Karvanen and Prof. Veikko Salomaa of the National Institute for Health and Welfare for helpful comments and for permission to use the FINRISK data in our illustration, respectively.

## References

Ancukiewicz, M., Finkelstein, D. M. & Schoenfeld, D. A. (2002). Modelling the relationship between continuous covariates and clinical events using isotonic regression. *Stat. Med.* **22**, 3151–3159.

Antoniadis, A., Bigot, J. & Gijbels, I. (2007). Penalized wavelet monotone regression. *Statist. Probab. Lett.* **77**, 1608–1621.

Arjas, E. & Gasbarra, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statist. Sinica* **4**, 505–524.

Arjas, E. & Liu, L. (1996). Non-parametric Bayesian approach to hazard regression: a case-study with a large number of missing covariate values. *Stat. Med.* **15**, 1757–1770.

Arjas, E. & Saarela, O. (2010). Optimal dynamic regimes: presenting a case for predictive inference. *Int. J. Biostat*. **6**. DOI: 10.2202/1557-4679.1204

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26**, 641–647.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. (1972). *Statistical inference under order restrictions*. Wiley, London.

Beran, R. & Dümbgen, L. (2010). Least squares and shrinkage estimation under bimonotonicity constraints. *Statist. Comput.* **20**, 177–189.

Bigaard, J., Frederiksen, K., Tjønneland, A., Thomsen, B. L., Overvad, K., Heitmann, B. L. & Sørensen, T. I. A. (2004). Waist and hip circumferences and all-cause mortality: usefulness of the waist-to-hip ratio? *Int. J. Obesity* **28**, 741–747.

Bornkamp, B. & Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose–response analysis. *Biometrics* **65**, 198–205.

Bornkamp, B., Ickstadt, K. & Dunson, D. (2010). Stochastically ordered multiple regression. *Biostatistics* **11**, 419–431.

Brezger, A. & Steiner, W. J. (2008). Monotonic regression based on Bayesian P-splines: an application to estimating price response functions from store-level scanner data. *J. Bus. Econom. Statist.* **26**, 90–104.

Cai, B. & Dunson, D. B. (2007). Bayesian multivariate isotonic regression splines: applications to carcinogenicity studies. *J. Amer. Statist. Assoc.* **102**, 1158–1171.

Cheng, G. (2008). Semiparametric additive isotonic regression. *J. Statist. Plann. Inference* **139**, 1980–1991.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *J. Roy. Statist. Soc. Ser. A* **147**, 278–292.

Dunson, D. B. (2005). Bayesian semiparametric isotonic regression for count data. *J. Amer. Statist. Assoc.* **100**, 618–627.

Dunson, D. B., Pillai, N. & Park, J.-H. (2007). Bayesian density regression. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **69**, 163–183.

Evans, A., Salomaa, V., Kulathinal, S., Asplund, K., Cambien, F., Ferrario, M., Perola, M., Peltonen, L., Shields, D., Tunstall-Pedoe, H. & Kuulasmaa, K. (2005). MORGAM (an international pooling of cardiovascular cohorts). *Int. J. Epidemiol.* **34**, 21–27.

Gelfand, A. E. & Kottas, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *J. Comput. Graph. Statist.* **11**, 289–305.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Chapman & Hall/CRC, Boca Raton, FL.

Holmes, C. C. & Heard, N. A. (2003). Generalized monotonic regression using random change points. *Stat. Med.* **22**, 623–638.

Kong, M. & Lee, J. J. (2006). A generalized response surface model with varying relative potency for assessing drug interaction. *Biometrics* **62**, 986–995.

Kong, M. & Lee, J. J. (2008). A semiparametric response surface model for assessing drug interaction. *Biometrics* **64**, 396–405.

Leitenstorfer, F. & Tutz, G. (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics* **8**, 654–673.
Møller, J. & Waagepetersen, R. P. (2004). *Statistical inference and simulation for spatial point processes*. Chapman & Hall/CRC, Boca Raton, FL.
Morton-Jones, T., Diggle, P., Parker, L., Dickinson, H. O. & Binks, K. (2000). Additive isotonic regression models in epidemiology. *Stat. Med.* **19**, 849–859.
Neal, R. M. (2000). Markov Chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9**, 249–265.
Neelon, B. & Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60**, 398–406.
Ramsay, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3**, 425–441.
Salomaa, V., Miettinen, H., Kuulasmaa, K., Niemelä, M., Ketonen, M., Vuorenmaa, T., Lehto, S., Palomäki, P., Mähönen, M., Immonen-Räihä, P., Arstila, M., Kaarsalo, E., Mustaniemi, H., Torppa, J., Tuomilehto, J., Puska, P. & Pyorälä, K. (1996). Decline of coronary heart disease mortality in Finland during 1983 to 1992: roles of incidence, recurrence, and case-fatality. The FINMONICA MI Register Study. *Circulation* **94**, 3130–3137.
Schell, M. J. & Singh, B. (1997). The reduced monotonic regression method. *J. Amer. Statist. Assoc.* **92**, 128–135.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **64**, 583–639.
Tutz, G. & Leitenstorfer, F. (2007). Generalized smooth monotonic regression in additive modeling. *J. Comput. Graph. Statist.* **16**, 165–188.
Vartiainen, E., Jousilahti, P., Alfthan, G., Sundvall, J., Pietinen, P. & Puska, P. (2000). Cardiovascular risk factor changes in Finland, 1972–1997. *Int. J. Epidemiol.* **29**, 49–56.

*Received February 2010, in final form August 2010*

Olli Saarela, Department of Chronic Disease Prevention, National Institute for Health and Welfare, P.O. Box 30 (Mannerheimintie 166), 00271 Helsinki, Finland.
E-mail: olli.saarela@thl.fi

## Appendix: Details of the MCMC algorithm

As a result of the random number of points in a realization, densities for point processes are usually defined w.r.t. another point process, rather than w.r.t. the Lebesgue measure (cf. Møller & Waagepetersen, 2004, pp. 24–25). Inference for point processes defined through an unnormalized density proceeds most conveniently utilizing simulation using MCMC techniques, since the Metropolis–Hastings ratio will depend on the densities only through the Papangelou conditional intensity (Møller & Waagepetersen, 2004, p. 83, 114). In the present application, we use point processes to define a large (uncountable) model space with finite realizations, and the inference is based on the random mixture over the realizations. In this sense, our point process approach resembles the use of the well-known Dirichlet process mixture model in non-parametric Bayesian inference (e.g. Neal, 2000; Gelfand & Kottas, 2002).

Point process realizations are updated by proposing reversible jump type modifications to existing point configurations (Green, 1995). Here, we consider four types of moves; birth, death, location change and simultaneous death/birth. The outline of the algorithm goes as follows: all point processes start from the empty point configuration. One of the processes $\Delta_i$, $i=1,\ldots,|\mathcal{I}|$ is selected with equal probabilities for updating. If the point configuration is empty, a birth step is attempted at probability 0.5, and with probability 0.5, the process stays at the empty state. From non-empty point configurations, birth and death moves are attempted with equal probability. In addition, the marks $\delta_{ij}$ are updated one at a time, conditionally given all the other marks and the partial ordering constraints imposed by the point locations. The joint prior of the marks given the partial ordering constraints was chosen as the uniform distribution with density

$$\frac{N!}{N^*}\left(\frac{1}{\delta_{\max}-\delta_{\min}}\right)^N,$$

where $N=\sum_{i=0}^{|\mathcal{I}|} n(\Delta_i)$ is the total number of points, $N!$ is the total number of possible ordered permutations and $N^*$ is the number of these permutations, which fulfil the partial ordering constraints. A simple proposal distribution for the mark $\delta_{ij}$ is given by the uniform distribution

$$\delta_{ij} \sim U[\max\{\delta_{rs} : \tilde{\xi}_{rs} \preceq \tilde{\xi}_{ij}, \delta_{\min}\}, \min\{\delta_{rs} : \tilde{\xi}_{rs} \succeq \tilde{\xi}_{ij}, \delta_{\max}\}],$$

where $(r,s) \neq (i,j)$. This is in fact equivalent to the full conditional (prior) distribution of $\delta_{ij}$ and thus the proposal and prior distributions cancel out from the Metropolis–Hastings ratio, leaving only the likelihood ratio. As the interval $[\delta_{\min}, \delta_{\max}]$ may contribute to the proposals in the boundary areas of the data, it should be chosen narrow enough to give reasonable proposals. However, it should be wide enough to not be able to 'flatten' the resulting function estimate. In the binary response case (see section 3.2) using the interval $[0,1]$ directly worked well. In a continuous response case, the interval can be varied as a sensitivity analysis if no prior information exists.

At the birth step, first a location $\xi_{ij}^*$ for the point is selected from the uniform distribution in $S(\Delta_i)$. Then, given the partial ordering constraints set up by the existing point configuration, a mark for the new point is drawn from $U[\max\{\delta_{rs} : \tilde{\xi}_{rs} \preceq \tilde{\xi}_{ij}^*, \delta_{\min}\}, \min\{\delta_{rs} : \tilde{\xi}_{rs} \succeq \tilde{\xi}_{ij}^*, \delta_{\max}\}]$. The proposal is accepted with probability

$$\min\left\{1, \frac{p(y \mid \lambda^*(x_1,\ldots,x_p))\rho_i |S(\Delta_i)|}{p(y \mid \lambda(x_1,\ldots,x_p))(n(\Delta_i)+1)}\right\},$$

where $\lambda^*$ is the regression surface after the addition of the new point and the associated mark.

In a death step, one of the points of the process to be updated is first selected randomly. The removal is accepted with probability

$$\min\left\{1, \frac{p(y \mid \lambda^*(x_1,\ldots,x_p))n(\Delta_i)}{p(y \mid \lambda(x_1,\ldots,x_p))\rho_i |S(\Delta_i)|}\right\}.$$

The position change moves are local in such a way that they do not break the existing partial ordering. Here, an existing point $(i,j)$ to be moved is first selected randomly. A new location for the point is drawn from uniform distribution in $\prod_{k=1}^{p}[\max\{\tilde{\xi}_{rs}^k : \tilde{\xi}_{rs}^k \leq \tilde{\xi}_{ij}^k, 0\}, \min\{\tilde{\xi}_{rs}^k : \tilde{\xi}_{rs}^k \geq \tilde{\xi}_{ij}^k, 1\}]$, where $(r,s) \neq (i,j)$. The Metropolis–Hastings ratio for the position change move involves only the likelihood ratio. This type of step will not move the location of the selected point beyond its closest neighbours. To make larger changes, a combined death/birth proposal is needed. This goes by first drawing randomly one of the processes (say $i$), and from that process drawing randomly a point to be removed. A process (say $i'$) for the new point is then selected randomly, with location and mark for the proposed point drawn from the same distributions as in the birth step. The joint proposal is accepted with probability

$$\min\left\{1, \frac{p(y \mid \lambda^*(x_1,\ldots,x_p))\rho_{i'} n(\Delta_i)|S(\Delta_{i'})|}{p(y \mid \lambda(x_1,\ldots,x_p))\rho_i (n(\Delta_{i'})+1)|S(\Delta_i)|}\right\}.$$

For efficient computation, it is important that the regression surface is not built from scratch each time a proposal is made and likelihood is evaluated. Rather, the marks associated with the regression surface at each datapoint should be tracked and only the associations affected by the proposal changed. The algorithm for carrying out the computation in the examples of section 3 was implemented in (ANSI) C and is available from the corresponding author.