Biometrika (2016), **103**, 1, *pp*. 1–16
Printed in Great Britain Advance Access publication on 31 July 2016

Bayesian spatial monotonic multiple regression

BY C. ROHRBECK, D. A. COSTAIN
Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K.
c.rohrbeck@lancaster.ac.uk d.costain@lancaster.ac.uk

AND A. FRIGESSI
Department of Biostatistics, University of Oslo, PB 1122 Blindern, 0317 Oslo, Norway.
arnoldo.frigessi@medisin.uio.no

SUMMARY

We consider monotonic, multiple regression for contiguous regions. The regression functions vary regionally and may exhibit spatial structure. We develop Bayesian nonparametric methodology that permits estimation of both continuous and discontinuous functional shapes using marked point process and reversible jump Markov chain Monte Carlo techniques. Spatial dependence is incorporated by a flexible prior distribution which is tuned using cross-validation and Bayesian optimization. We derive the mean and variance of the prior induced by the marked point process approach. Asymptotic results show consistency of the estimated functions. Posterior realizations enable variable selection, the detection of discontinuities and prediction. In simulations and in an application to a Norwegian insurance data set, our methodology shows better performance than existing approaches.

Some key words: Cross-validation; Isotonic regression; Marked point process; Optimization; Reversible jump Markov chain Monte Carlo algorithm; Spatial dependence.

1. INTRODUCTION

Geospatial data are considered in forestry (Penttinen et al., 1992), epidemiology (Waller & Gotway, 2004) and other domains. Due to practicality or confidentiality concerns, locally aggregated data are common and are typically available on an irregular lattice. Statistical methods for such data aim to explore the association between a response and explanatory variables while accounting for spatial dependence in the model parameters. To introduce such dependence, a neighbourhood structure, often based upon the arrangement of the areal units on a map, is typically defined via an adjacency matrix.

Most modelling frameworks assume a common effect of the explanatory variables for all regions (Waller & Gotway, 2004; Wakefield, 2007). Spatial variation is then typically accommodated via a spatially structured random effect on the intercept. Some applications, however, need to allow for a spatially-varying regression function (Bell et al., 2004; Zhang & Shi, 2004; Cahill & Mulligan, 2007). Statistical methods for such scenarios are available for generalized linear (Fotheringham et al., 2003; Assunção, 2003; Scheel et al., 2013) and additive models (Congdon, 2006). However, these approaches are limited, as continuity of the regression function is assumed: abrupt changes in the regression surface are not captured unless they are explicitly included in the model. Neglecting such effects may result in a bias due to oversmoothing; see Bowman & Azzalini (1997) p.26.

Since continuity may be inappropriate, we replace it by monotonicity (Royston, 2000; Farah et al., 2013; Wilson et al., 2014). Whilst continuity cannot, in general, be verified, tests of monotonicity are available (Bowman et al., 1998; Ghosal et al., 2000; Scott et al., 2015). Based upon a number of observations for each region, we develop Bayesian nonparametric methodology which estimates the regional regression functions whilst exploiting any neighbourhood structure.

The estimation of a single monotonic function is usually called isotonic regression. Early publications discuss inference under monotonicity constraints (Ayer et al., 1955; Brunk, 1955; Barlow & Brunk, 1972) and solution algorithms are available (Brunk et al., 1957; Luss et al., 2012). Isotonic regression is further considered for additive (Bacchetti, 1989; Tutz & Leitenstorfer, 2007) and high-dimensional models (Fang & Meinshausen, 2012; Bergersen et al., 2014), in functional data analysis (Ramsay, 1998; Ramsay & Silverman, 2005) and Bayesian nonparametric modelling (Gelfand & Kuo, 1991; Shively et al., 2009; Saarela & Arjas, 2011; Lin & Dunson, 2014). In order to learn about potentially spatially structured monotonic regression functions, a dependence model for functions, possibly with discontinuities, is required.

Our approach represents each monotonic regional function by a set of marked point processes. Potential spatial structure is modelled via a joint prior distribution, which is based upon a flexible discrepancy measure. The prior allows the functional dependence to be constant, increasing or decreasing with increasing function values. The Bayesian framework induces a consistent posterior (Barron et al., 1999; Walker & Hjort, 2001), and permits both smooth contours and discontinuities. To tune the prior, we combine cross-validation and Bayesian optimization. Realizations of the posterior are obtained by a reversible jump Markov chain Monte Carlo algorithm (Green, 1995) and enable variable selection, prediction and the detection of discontinuities.

2. MODELLING AND INFERENCE

2·1. *Likelihood and notation*

Consider K contiguous regions whose neighbourhood structure is given by an adjacency matrix or a lattice graph. Let $y_k \in \mathbb{R}$ and $x_k \in \mathbb{R}^m$ denote the response and explanatory variables, respectively, for region k ($k = 1, \ldots, K$). The likelihood is

$$f\left\{y_k \mid \lambda_k(x_k), \theta_k\right\}, \tag{1}$$

where $\lambda_k : \mathbb{R}^m \to [\delta_{\min}, \delta_{\max}]$ is the monotonic regression function for region k, for which $\lambda_k(x_k)$ is assumed to lie within the predefined interval $[\delta_{\min}, \delta_{\max}]$. Monotonicity is defined in terms of the partial Euclidean ordering $\preceq$: if $u \leq v$ component-wise, $\lambda_k(u) \leq \lambda_k(v)$, $u, v \in \mathbb{R}^m$. The vector θ_k denotes additional, potentially spatially varying, model parameters which are a priori independent of $\lambda_1, \ldots, \lambda_K$.

In what follows, we perform inference on λ_1 through λ_K while accounting for potential spatial structure in these functions. Each λ_k ($k = 1, \ldots, K$) is estimated over a closed set $X \subset \mathbb{R}^m$. In applications, X and $[\delta_{\min}, \delta_{\max}]$ may be defined in terms of the explanatory variables and responses, respectively. For instance, if $\lambda_k(x_k)$ in (1) is the mean response, $\delta_{\min}$ may be set to the minimum observed response across the K regions. In §2·2 to §2·4 we complete the Bayesian framework by defining a joint prior on $\lambda_1, \ldots, \lambda_K$ while §2·5 and §2·6 detail the estimation procedure.

2·2. *A spatial dependence model for the monotonic functions*

We wish to impose spatial structure on $\lambda_1, \ldots, \lambda_K$ and hence define a joint prior density that favours these functions to be similar. We set $\delta_{\min} = 0$ and write $\lambda_k(x)$ instead of $\lambda_k(x) - \delta_{\min}$ ($k = 1, \ldots, K$) below.

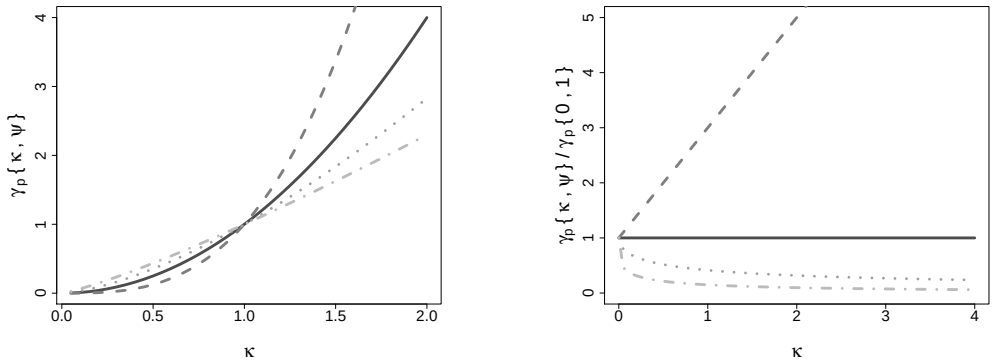

Fig. 1. Behaviour of $\gamma_p\{\kappa, \psi\}$ with respect to κ for $p = 1$ (——), $p = 2$ (– –), $p = 0 \cdot 5$ (·····) and $p = 0 \cdot 2$ (–·–·–), subject to $\psi = 0$ (left) and $\psi = \kappa + 1$ (right) being fixed.

First, we introduce a pairwise discrepancy measure for λ_k and $\lambda_{k'}$ $(k, k' = 1, \ldots, K;\ k \neq k')$. Such a measure should be minimal if and only if λ_k and $\lambda_{k'}$ are equal, and should increase with an increasing difference in these functions. A possible choice is the integrated squared distance

$$\int_X \{\lambda_k(x) - \lambda_{k'}(x)\}^2 \ \mathrm{d}x. \tag{2}$$

Sometimes, we may have prior knowledge that differences in the lower, or higher, function values of λ_k and $\lambda_{k'}$ should be downweighted, or avoided. For example, increased measurement error in higher values of the explanatory variables may be better handled through increased information borrowing. Thus, we replace $\{\lambda_k(x) - \lambda_{k'}(x)\}^2$ in (2) by

$$\gamma_p\{\lambda_k(x), \lambda_{k'}(x)\} = [\ \{\lambda_k(x)\}^p - \{\lambda_{k'}(x)\}^p\ ]\ \{\lambda_k(x) - \lambda_{k'}(x)\}, \qquad p > 0, \tag{3}$$

for which $p = 1$ yields the integrated squared distance. See the 2017 Lancaster University PhD thesis by C. Rohrbeck for a more general formulation. Expression (3) can also be interpreted as the squared distance with weight $[\{\lambda_k(x)\}^p - \{\lambda_{k'}(x)\}^p] \,/\, \{\lambda_k(x) - \lambda_{k'}(x)\}$.

Figure 1 illustrates the behaviour of $\gamma_p\{\lambda_k(x), \lambda_{k'}(x)\}$ at a fixed point $x \in \mathbb{R}^m$ for different settings of p. For brevity, let $\kappa = \lambda_k(x)$ and $\psi = \lambda_{k'}(x)$. The left panel shows that $\gamma_p\{\kappa, \psi\}$ increases with an increasing difference between κ and $\psi = 0$ for all settings of p. Hence, γ_p satisfies the desired properties stated above. Furthermore in the right panel, the fixed difference $\psi = \kappa + 1$ is penalized more for higher κ if $p > 1$, while being penalized less for $p < 1$. A constant penalty is induced for $p = 1$. As such, the parameter p allows the penalty for differences between λ_k and $\lambda_{k'}$ to vary with the function values.

The dependence model for the K-set $\lambda_1, \ldots, \lambda_K$ is then defined as a Gibbs measure (Georgii, 2011) with the discrepancy measure constructed in (2) and (3) as a pair-potential. Formally,

$$\pi(\lambda_1, \ldots, \lambda_K \mid \omega) \propto \prod_{1 \leq k < k' \leq K} \exp\left[-\omega\ d_{k,k'} \int_X \gamma_p\{\lambda_k(x), \lambda_{k'}(x)\}\ \mathrm{d}x\right], \qquad \omega \geq 0, \tag{4}$$

where the product is over all pairs of regions. The constant $d_{k,k'} \geq 0$ describes our prior belief concerning the degree of similarity of λ_k and $\lambda_{k'}$. In spatial statistics, we often set $d_{k,k'} = 1$ if the regions k and k' are adjacent and $d_{k,k'} = 0$ otherwise. Such a choice reduces the computational cost since the integral in (4) need only be evaluated for pairs of adjacent regions. The degree of dependence increases in ω, and $\omega = 0$ corresponds to $\lambda_1, \ldots, \lambda_K$ being independent. Sensitivity

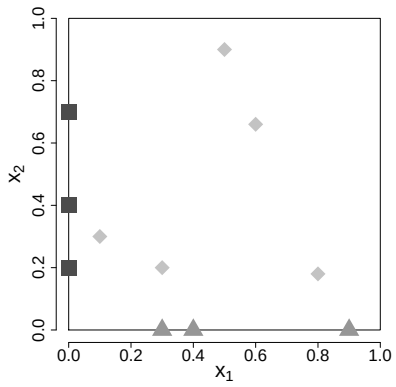

Fig. 2. Point locations to represent a step function λ on $X = [0,1]^2$ via a set of $I = 3$ marked point processes $(\Delta_1, \Delta_2, \Delta_3)$. The processes Δ_1 (▲), Δ_2 (■) and Δ_3 (◆) are defined on the sets $X_1 = [0,1] \times 0$, $X_2 = 0 \times [0,1]$ and $X_3 = (0,1] \times (0,1]$, respectively.

to choice of p is explored in §3·3. Expression (4) can be extended to regionally varying X, permitting borrowing of information for extrapolation; see the Supplementary Material.

2·3. *Marked point process prior*

We specify an individual prior model for $\lambda_k : X \to [\delta_{\min}, \delta_{\max}]$ $(k = 1, \ldots, K)$ and drop the index k in the rest of this subsection for brevity. Prior distributions proposed in the literature include an ordered Dirichlet process (Gelfand & Kuo, 1991) and a constrained spline (Shively et al., 2009). Our prior is similar to that of Saarela & Arjas (2011): λ is postulated to be a non-decreasing step function with $\lambda(x) \in [\delta_{\min}, \delta_{\max}]$; any monotonic, bounded function can be approximated to a desired accuracy by increasing the number of steps.

The location and height of the steps of λ define a marked point process on X. Following Saarela & Arjas (2011), we represent λ via a set of I marked point processes, $\Delta = (\Delta_1, \ldots, \Delta_I)$, where Δ_i $(i = 1, \ldots, I)$ is on a set X_i with $\bigcup_{i=1}^{I} X_i = X$. Here, we define $X_1, \ldots, X_I$ based on the non-empty subsets of $\{1, \ldots, m\}$. For example, if $m = 2$ we choose $I = 3$ and have separate processes Δ_1 and Δ_2 for each of the two explanatory variables, x_1 and x_2, respectively, and one process Δ_3 for both components, (x_1, x_2), jointly. Figure 2 provides an example for $X = [0,1]^2$. The benefits of this representation are discussed later in this subsection.

We now formalize the representation of λ via Δ and denote

$$\Delta_i = \{(\xi_{i,j}, \delta_{i,j}) \in X_i \times [\delta_{\min}, \delta_{\max}] \ : \ j = 1, \ldots, n_i\}, \qquad i = 1, \ldots, I. \tag{5}$$

Here, $\xi_{i,j}$ and $\delta_{i,j}$ refer to a point location and associated mark, respectively, and n_i is the number of points in Δ_i. Monotonicity is imposed by constraining the marks: if $\xi_{i,j} \preceq \xi_{i',j'}$, $\delta_{i,j} \leq \delta_{i',j'}$ $(i, i' = 1, \ldots, I;\ j = 1, \ldots, n_i;\ j' = 1, \ldots, n_{i'})$. The value $\lambda(x)$ is then defined as the largest mark $\delta_{i,j}$ such that x imposes a monotonicity constraint on the associated point location $\xi_{i,j}$. Formally,

$$\lambda(x) = \max_{i,j} \{\delta_{i,j} \ : \ \xi_{i,j} \preceq x\}. \tag{6}$$

Representing λ via the set $(\Delta_1, \ldots, \Delta_I)$ facilitates variable selection. Let $X = [0,1]^m$ and suppose that the explanatory variable x_1 is redundant. Hence, λ is constant with increasing val-

ues of x_1, that is, $\lambda(x) = \lambda\{x + (\epsilon, 0, \ldots, 0)\}$ $(x \in X;\ \epsilon > 0)$. As we represent λ via a marked point process, the redundancy of x_1 implies that the point locations are in the set $0 \times [0,1]^{m-1}$. For instance, if $m = 2$, all points then lie on the line $x_1 = 0$ in Fig. 2. As such, the processes Δ_1 and Δ_3 contain no points. Consequently, n_i $(i = 1, \ldots, I)$ provides an indicator of the redundancy of explanatory variables.

The association defined in (6) results in a mapping between the spaces of step functions and marked point processes with monotonicity constraints. Thus, we can define a prior for λ via one for Δ. A priori, the number $N = \sum_{i=1}^{I} n_i$ of steps representing λ is geometrically distributed with probability $1/\eta$ $(\eta > 1)$ and $N = 0$ corresponds to $\lambda = \delta_{\min}$ being constant. This choice promotes model parsimony and favours λ to have few steps. Given N, the vector $(n_1, \ldots, n_I)$ is uniformly distributed over the set of possibilities of allocating N points to the I processes. For Δ_i $(i = 1, \ldots, I)$, the location $\xi_{i,j}$ $(j = 1, \ldots, n_i)$ is uniformly distributed on X_i. The marks $\{\delta_{i,j} :\ j = 1, \ldots, n_i;\ i = 1, \ldots, I\}$ are uniformly distributed on $[\delta_{\min}, \delta_{\max}]$, subject to the monotonicity constraints imposed by the locations in $\Delta_1, \ldots, \Delta_I$. Using this hierarchical structure, we obtain the prior density

$$\phi(\Delta \mid \eta) = \pi(\{\delta_{i,j}\} \mid \{\xi_{i,j}\}) \left\{ \prod_{i=1}^{I} \prod_{j=1}^{n_i} \pi(\xi_{i,j}) \right\} \pi(n_1, \ldots, n_I \mid N)\, \pi(N \mid \eta)\,; \tag{7}$$

further details are provided in the Supplementary Material.

The density $\phi(\Delta \mid \eta)$ induces a density on the space of step functions, $\tilde{\phi}(\lambda \mid \eta)$, which can be characterized as follows:

PROPOSITION 1. *Let* $X = [0,1]$, $\delta_{\min} = 0$ *and* $\delta_{\max} = 1$. *Then the distribution with density* $\tilde{\phi}(\lambda \mid \eta)$ *has*

$$E\{\lambda(x) \mid \eta\} = x \sum_{n=1}^{\infty} \left\{ \frac{1}{\eta} \left(1 - \frac{1}{\eta}\right)^n \frac{n}{n+1} \right\} = x \left(1 - \frac{\log \eta}{\eta - 1}\right),$$

$$\textit{var}\{\lambda(x) \mid \eta\} = \sum_{n=1}^{\infty} \left\{ \frac{1}{\eta} \left(1 - \frac{1}{\eta}\right)^n \frac{nx(2 - x + nx)}{(n+1)(n+2)} \right\} - E\{\lambda(x) \mid \eta\}^2.$$

Hence, the expectation is a linear function whose slope depends on η. See the Supplementary Material for the the proof of Proposition 1.

This Bayesian framework has one small limitation. If, for instance, $X = [0,1]$, $\lambda(0) = \delta_{\min}$ almost surely. To address this, we define $\lambda(x) = \mu + \varphi(x)$, where $\varphi : X \to [\delta_{\min}, \delta_{\max}]$ is monotonic and $\mu \in \mathbb{R}$, and with priors $\tilde{\phi}(\varphi \mid \eta)$ and $\pi(\mu)$, respectively. A second approach is presented in the Supplementary Material.

2·4. *Combining the spatial dependence model and marked point process prior*

We now impose a spatial structure on the K sets of marked point processes $\Delta_1, \ldots, \Delta_K$, $\Delta_k = (\Delta_{k,1}, \ldots, \Delta_{k,I})$ $(k = 1, \ldots, K)$, by combining $\phi(\Delta_k \mid \eta)$ in (7) with $\pi(\lambda_1, \ldots, \lambda_K \mid \omega)$ in (4). The joint prior $\pi(\Delta_1, \ldots, \Delta_K \mid \omega, \eta)$ is then proportional to

$$\prod_{1 \le k < k' \le K} \exp\left[-\omega\, d_{k,k'} \int_X \gamma_p \left\{ \tilde{\lambda}_k(x), \tilde{\lambda}_{k'}(x) \right\} \mathrm{d}x \right] \times \prod_{k=1}^{K} \phi(\Delta_k \mid \eta)\,, \tag{8}$$

where $\tilde{\lambda}_k$ and $\tilde{\lambda}_{k'}$ are the step functions represented by Δ_k and $\Delta_{k'}$, respectively. Since $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_K$ are step functions, the integral in (4) simplifies to a sum and can be computed efficiently. The full conditional prior density for Δ_k in (8) converges to (7) as $\omega \to 0$. Further,

$\pi(\Delta_1, \ldots, \Delta_K \mid \omega, \eta)$ is proper because $\pi(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_K \mid \omega)$ lies within $(0, 1]$ and $\phi(\Delta_k \mid \eta)$ is a proper density function.

The likelihood (1) and prior (8) specify a posterior distribution for $\Delta_1, \ldots, \Delta_K$ with density

$$\pi(\Delta_1, \ldots, \Delta_K \mid \mathcal{D}, \omega, \eta) \propto \left[\prod_{k=1}^{K} \prod_{t=1}^{T_k} f\left\{ y_{k,t} \mid \tilde{\lambda}_k(x_{k,t}), \theta_k \right\} \right] \times \pi(\Delta_1, \ldots, \Delta_K \mid \omega, \eta), \quad (9)$$

where $\mathcal{D}$ denotes the data and T_k is the number of observations for region k ($k = 1, \ldots, K$).

An estimator should be consistent. In Bayesian nonparametrics, consistency is often considered in terms of the Hellinger distance. Let (λ_k, θ_k) denote the true model parameters for region k ($k = 1, \ldots, K$) and let G_k be the distribution of the explanatory variables, $x_k \sim G_k$. Following Walker & Hjort (2001), we denote the Hellinger distance between the densities with parameters (λ_k, θ_k) and $(\tilde{\lambda}_k, \tilde{\theta}_k)$ by

$$H_k\left(\tilde{\lambda}_k, \tilde{\theta}_k\right) = \left(1 - \int\int \left[f\left\{ y \mid \tilde{\lambda}_k(x_k), \tilde{\theta}_k \right\} f\left\{ y \mid \lambda_k(x_k), \theta_k \right\} \right]^{1/2} \mathrm{d}y \, G_k(\mathrm{d}x_k) \right)^{1/2}. \quad (10)$$

Let $\Lambda = (\lambda_1, \ldots, \lambda_K)$ and $\Theta = (\theta_1, \ldots, \theta_K)$. We then define a neighbourhood $U_\epsilon(\Lambda, \Theta)$ around the truth (Λ, Θ) with respect to $H_1, \ldots, H_K$ in (10) with

$$U_\epsilon(\Lambda, \Theta) = \left\{ \left(\tilde{\Lambda}, \tilde{\Theta}\right) \; : \; H_k\left(\tilde{\lambda}_k, \tilde{\theta}_k\right) \leq \epsilon, \; k = 1, \ldots, K \right\}, \qquad \epsilon > 0.$$

Here, $U_\epsilon(\Lambda, \Theta)$ contains only step functions and is non-empty because we can approximate λ_k by a step function to any degree of accuracy. In the following, we focus on $f\{y_k \mid \lambda_k(x_k), \theta_k\}$ being the normal density function with mean $\lambda_k(x_k)$ and variance θ_k, but the theory can be generalized and holds for all examples in this paper.

THEOREM 1. *Let G_k ($k = 1, \ldots, K$) be absolutely continuous and assign positive mass to any non-degenerate subset of X. Further, let the prior $\pi(\tilde{\Theta})$ put positive mass on any neighbourhood of Θ. Then, for $\lambda_1, \ldots, \lambda_K : X \to [\delta_{\min}, \delta_{\max}]$ monotonic and continuous, and $\epsilon > 0$, $\tilde{\Pi}\{U_\epsilon^c(\Lambda, \Theta) \mid \mathcal{D}, \omega, \eta\} \to 0$ almost surely as $\min_{k=1,\ldots,K} T_k \to \infty$. Here, $U_\epsilon^c(\Lambda, \Theta)$ is the complement of $U_\epsilon(\Lambda, \Theta)$ and $\tilde{\Pi}$ denotes the posterior distribution induced by the likelihood (1), and the priors $\pi(\tilde{\Theta})$ and $\pi(\Delta_1, \ldots, \Delta_K \mid \omega, \eta)$.*

Hence, the posterior distribution concentrates around the K true functions as the number of data points becomes large, conditional on appropriate boundaries $\delta_{\min}$ and $\delta_{\max}$. Moreover, the posterior mean may be smooth, as the model permits variability in the number, locations and heights of the steps. Consequently, our approach can recover both smooth and discontinuous functional shapes. This result is well-known for the estimation of a single probability density function using a piecewise approximation (Heikkinen & Arjas, 1998). The proof of Theorem 1 is in the Supplementary Material.

In a fully Bayesian framework, we would need priors for η and ω. However, the normalizing constant of $\pi(\Delta_1, \ldots, \Delta_K \mid \omega, \eta)$ in (8) is intractable, unless $\omega = 0$. This leads to our novel inferential approach for ω in §2·6. In terms of setting η, Proposition 1 implies that higher values of η will generally lead to smoother surfaces. Alternatively, one may learn about η by considering the case $\omega = 0$. We can then specify a conjugate Beta prior for $1/\eta$ and sample from the full conditional Beta posterior; the performance of this approach is explored in §3.

2·5. *Inference and analysis of the marked point processes*

Our scheme to sample from the posterior density in (9) is based on Saarela & Arjas (2011). Initially, $\Delta_{k,i}$ ($k = 1, \ldots, K; \; i = 1, \ldots, I$) is empty and so $\lambda_k = \delta_{\min}$. The K sets are then

updated sequentially. We first select one of the processes $\Delta_{k,1}, \ldots, \Delta_{k,I}$ $(k = 1, \ldots, K)$ with equal probability. For the sampled process Δ_{k,i^*}, we randomly propose one of three moves, implying local changes of λ_k. A birth move adds a point (ξ^*, δ^*) to Δ_{k,i^*}, where ξ^* is sampled uniformly on X_{i^*}. Given ξ^*, the associated mark δ^* is sampled uniformly, subject to monotonicity being preserved. A death move removes a point from Δ_{k,i^*}, maintaining reversibility. A shift move changes the location and mark of a point in Δ_{k,i^*}, subject to the partial order imposed by the monotonicity constraints being maintained. See the Appendix for details and the acceptance probabilities. We implemented this scheme in C++ and a simulation study to verify correctness is provided in the Supplementary Material.

Realizations sampled from the posterior distribution are rich and facilitate detailed analysis of $\lambda_1, \ldots, \lambda_K$. Thinning of the Markov chains is needed to reduce autocorrelation. Posterior mean estimates for λ_k are obtained by averaging over the stored realizations. The mean and quantiles of the posterior distribution are accessible for any $x \in X$ by deriving $\lambda_k(x)$ for each sample. Further, the samples facilitate the detection of discontinuities; see the Supplementary Material.

2·6. *Estimation of* ω

The performance of our approach relies on a suitable ω in (8). If ω is too high, spatial variation is oversmoothed, while overfitting may occur if ω is too small. Since the normalizing constant of (8) is intractable, we cannot sample from the full conditional distribution of ω via an additional Gibbs step within the scheme in §2·5. Further, while there exists a rich literature on handling intractable normalizing constants (Beaumont et al., 2002; Møller et al., 2006; Andrieu & Roberts, 2009), these approaches cannot be adapted since efficient sampling from the prior distribution in (8) is infeasible. Hence, we estimate ω prior to inference on $\Delta_1, \ldots, \Delta_K$.

One approach is s-fold cross-validation: the data for each of the K regions are split into s subsets of equal size. The sampling scheme in §2·5 is then performed s times with varying training and test data. Parameter values are compared by the posterior mean squared error for the test data points. In order to keep the number of evaluated values for ω small, we combine cross-validation with the global optimization algorithm of Jones et al. (1998).

Efficient global optimization postulates a sequential design strategy to detect global extrema of a black-box function r. The algorithm is widely applied in simulations if r is costly to evaluate and the parameter space Z is small (Roustant et al., 2012). The rationale is to model r by a Gaussian process R which is updated sequentially. Specifically, the proposal $z^* \in Z$ is selected to maximize the expected improvement

$$E\left[\max\left\{r_{\text{opt}} - R(z), 0\right\}\right], \qquad z \in Z, \tag{11}$$

where r_{opt} denotes the current optimum. Hence, (11) represents the potential of $r(z)$ to be smaller than r_{opt}. The proposal is evaluated until its expected improvement falls below a critical value, corresponding to r_{opt} being sufficiently close to the unknown minimum of r. As this approach balances local exploration of the areas likely to provide good model fit, and a global search, a suitable solution is generally found after a reasonable number of evaluations.

When estimating ω, interest lies in the minimum of the cross-validation function CV(ω). Algorithm 1 sketches our approach. Since efficient global optimization can only be applied to a closed set, we first derive an upper bound. An initial bound ω_u is increased until its mean squared error is greater than that for $\omega = 0$ by a sufficient amount; $\beta = 2$ in Algorithm 1 proved reasonable in our simulations. Once ω_u is fixed, an initial proposal $\omega^* \in [0, \omega_u]$ is made, guaranteeing that R in (11) is fitted with at least three data points. We use the DiceOptim R package (Roustant et al., 2012) to derive the expected improvement and run multiple s-fold cross-validations with the same ω to reduce the dependence on the split of the data. The mean and variance of the

mean squared error across the repetitions are used to fit G. We then repeatedly perform cross-validation and update ω^* until the maximum expected improvement falls below the critical value α. To conclude, we set ω to the value ω_{opt} that provided the lowest mean squared error.

Algorithm 1. Combination of efficient global optimization and cross-validation.

Set initial upper bound ω_u, critical value α and factor β
Perform cross-validation for $\omega = 0$ and store CV(0)
While CV(ω_u) $<$ β CV(0)
  Increase ω_u
  Perform cross-validation for ω_u and store CV(ω_u)
Set initial proposal ω^*, e.g. $\omega^* = \omega_u/2$
Initialize maximum expected improvement $M > \alpha$
While $M > \alpha$
  Perform cross-validation for ω^* and store CV(ω^*)
  Fit Gaussian process R
  Update ω^* and M
Return value ω_{opt} which provided smallest error

3. SIMULATION STUDY

3·1. *Introduction*

We aim to demonstrate that our methodology improves estimates if similarities between functions exist, and is robust otherwise. Furthermore, we examine sensitivity to the prior parameters p and η in expression (8).

Responses for region k ($k = 1, \ldots, K$) are simulated independently from

$$y_k \mid x_k \sim \text{Normal}\left\{\lambda_k(x_k), \theta_k\right\},$$

where $x_k \in [0,1]^2$. As described in §2·3, we define $\lambda_k(x) = \mu_k + \varphi_k(x)$, and perform inference on $\mu_k \in \mathbb{R}$ and $\varphi_k : [0,1]^2 \to [\delta_{\min}, \delta_{\max}]$. The likelihood (1) is then

$$f\left\{y_k \mid \varphi_k\left(x_k\right), \mu_k, \theta_k\right\} = \left(\frac{1}{2\pi\theta_k}\right)^{1/2} \exp\left[-\frac{1}{2\theta_k}\left\{y_k - \mu_k - \varphi_k(x_k)\right\}^2\right].$$

An intrinsic conditional autoregressive prior (Besag et al., 1991; Rue & Held, 2005) is defined for $(\mu_1, \ldots, \mu_K)$ and imposes a spatial structure. Here, $\mu_1, \ldots, \mu_K$ are updated separately via a random walk Metropolis step and the hyperparameter in $\pi\left(\mu_1, \ldots, \mu_K\right)$ is updated via Gibbs sampling (Knorr-Held, 2003). Furthermore, we assign the prior distribution $1/\theta_k \sim \text{Gamma}\left(1, 0{\cdot}001\right)$ $(k = 1, \ldots, K)$ and update $\theta_1, \ldots, \theta_K$ via Gibbs sampling.

Here, X is the square spanned by the minimum and maximum observed value in each explanatory variable across the K regions. The boundaries are set to $\delta_{\min} = -1$ and $\delta_{\max} = 4$. We assess performance via the absolute difference of the posterior mean estimate $\hat{\lambda}_k$ and the true function λ_k, over a regular 100×100 grid on X. Only grid points contained in the convex hull of the observed values of x_k $(k = 1, \ldots, K)$ are considered. Improvements are discussed with respect to the setting $\omega = 0$, which imposes no dependence.

Algorithm 1 is applied with $\beta = 2,\ \alpha = \text{CV}(0)/1000$ and $\omega_u = 50$. We increase ω_u by factor 10 until CV$(\omega_u) < 2$ CV(0). For each proposed ω, five repetitions of 10-fold cross validation are performed. A fold consists of 50,000 iterations and every 100th sample is stored after a burn-in period of 25,000 iterations. In addition to the expected improvement criterion, we stop if 30 proposals have been considered; this occurred once in all our simulations. Birth, death and shift

Table 1. *Mean (*$\times 100$*) and standard deviation (*$\times 100$*) of the absolute difference between truth and posterior mean estimate for* (λ_1, λ_2) *in Studies 1 to 5 in Fig. 3 for* $\eta = (10, 1000, \hat{\eta})$ *and* $\omega = 0$*. The final column refers to an estimated monotonized generalized additive model*

Study	Function	$\eta = 10$	$\eta = 1000$	$\eta = \hat{\eta}$	$\omega = 0$	GAM
1	λ_1	1·8 (2·1)	1·8 (2·0)	1·8 (2·1)	1·8 (2·1)	3·2 (3·4)
	λ_2	2·5 (2·7)	3·0 (3·3)	2·8 (3·1)	4·6 (5·6)	4·6 (4·9)
2	λ_1	1·6 (3·2)	1·6 (3·2)	1·6 (3·2)	1·t6 (3·3)	7·0 (7·0)
	λ_2	2·8 (3·3)	3·1 (3·7)	3·1 (3·9)	4·6 (5·6)	9·0 (9·0)
3	λ_1	1·3 (1·2)	1·1 (1·1)	1·1 (1·1)	1·1 (1·0)	0·5 (0·5)
	λ_2	2·0 (1·6)	1·9 (1·5)	1·9 (1·4)	2·3 (2·2)	1·2 (0·9)
4	λ_1	3·1 (9·2)	2·7 (7·4)	2·8 (7·8)	2·8 (7·5)	6·7 (8·6)
	λ_2	4·2 (9·2)	4·0 (7·8)	4·1 (8·0)	5·9 (10·9)	7·3 (9·9)
5	λ_1	1·4 (1·3)	1·3 (1·2)	1·3 (1·2)	1·3 (1·2)	0·8 (0·7)
	λ_2	2·3 (2·7)	2·2 (2·7)	2·3 (2·7)	2·4 (2·8)	3·4 (3·1)

moves are proposed with probabilities 0·3, 0·3 and 0·4. Estimates for $\Delta_1, \ldots, \Delta_K$ are based on 3,000,000 iterations, with the first 1,000,000 discarded, and then every 1000th sample stored.

Convergence of the sampled Markov chains for Δ_k $(k = 1, \ldots, K)$ is checked via the trace plots of $\lambda_k(x)$ for ten random points in X. Posterior mean plots and trace plot examples are provided in the Supplementary Material. We also applied our methodology to non-Gaussian settings; an example with binomial response data is presented in the Supplementary Material. The C++ and R code for all simulations is provided in the Supplementary Material.

3·2. *Sensitivity analysis on* η

We explore general performance and sensitivity to η based on five simulations with $K = 2$ regions. Columns 1 and 2 in Fig. 3 illustrate the five pairs of (λ_1, λ_2). Across all studies, $\lambda_k(x_k) \in [0, 2]$ $(k = 1, 2)$. For each study, 1,000 and 100 data points are sampled for regions 1 and 2, respectively, with $\theta_k = 0{\cdot}5^2$ and $x_k \sim \text{Unif}([0, 1]^2)$ $(k = 1, 2)$. This setting explores the potential benefits of borrowing statistical information from region 1 when estimating λ_2. We fix the prior parameter $p = 1$ and consider three settings for η: (i) $\eta = 10$, (ii) $\eta = 1000$ and (iii) $\eta = \hat{\eta}$. Here, $\hat{\eta}$ is the posterior mean estimate for η in the case $\omega = 0$ as described in §2·4.

We also estimate a monotonized generalized additive model for each region separately and derive the same summary statistics as for our approach. We first fit a generalized additive model (Hastie & Tibshirani, 1990) and then apply the projection by Lin & Dunson (2014); plots of the estimated surfaces are provided in the Supplementary Material.

Study 1 and 2 consider the case $\lambda_1 = \lambda_2$ and Table 1 shows reduced error measures, particularly for region 2, compared to the setting when $\omega = 0$. Figure 3 illustrates that both smooth surfaces and discontinuities are recovered well. In Study 3 and Study 4, λ_1 and λ_2 are similar and the conclusions are consistent with those for Study 1 and Study 2. Study 5 considers the case of λ_1 being smooth while λ_2 is piecewise linear. Table 1 shows no worsening in the error measures, demonstrating robustness of our methodology. The prospect of variable selection described in §2·3 has been examined for λ_2 in Study 5, where $\lambda_2(x)$ depends only on $x_{2,1}$. Almost all sampled points are in $\Delta_{2,1}$, hence the results indicate $x_{2,2}$ to be redundant.

Table 1 shows that our approach performs better than the fitted monotonized generalized additive model, unless the true function is smooth. The results also indicate a low sensitivity to η. In particular, higher values of η yield improved results if the true function is smooth, as in Study

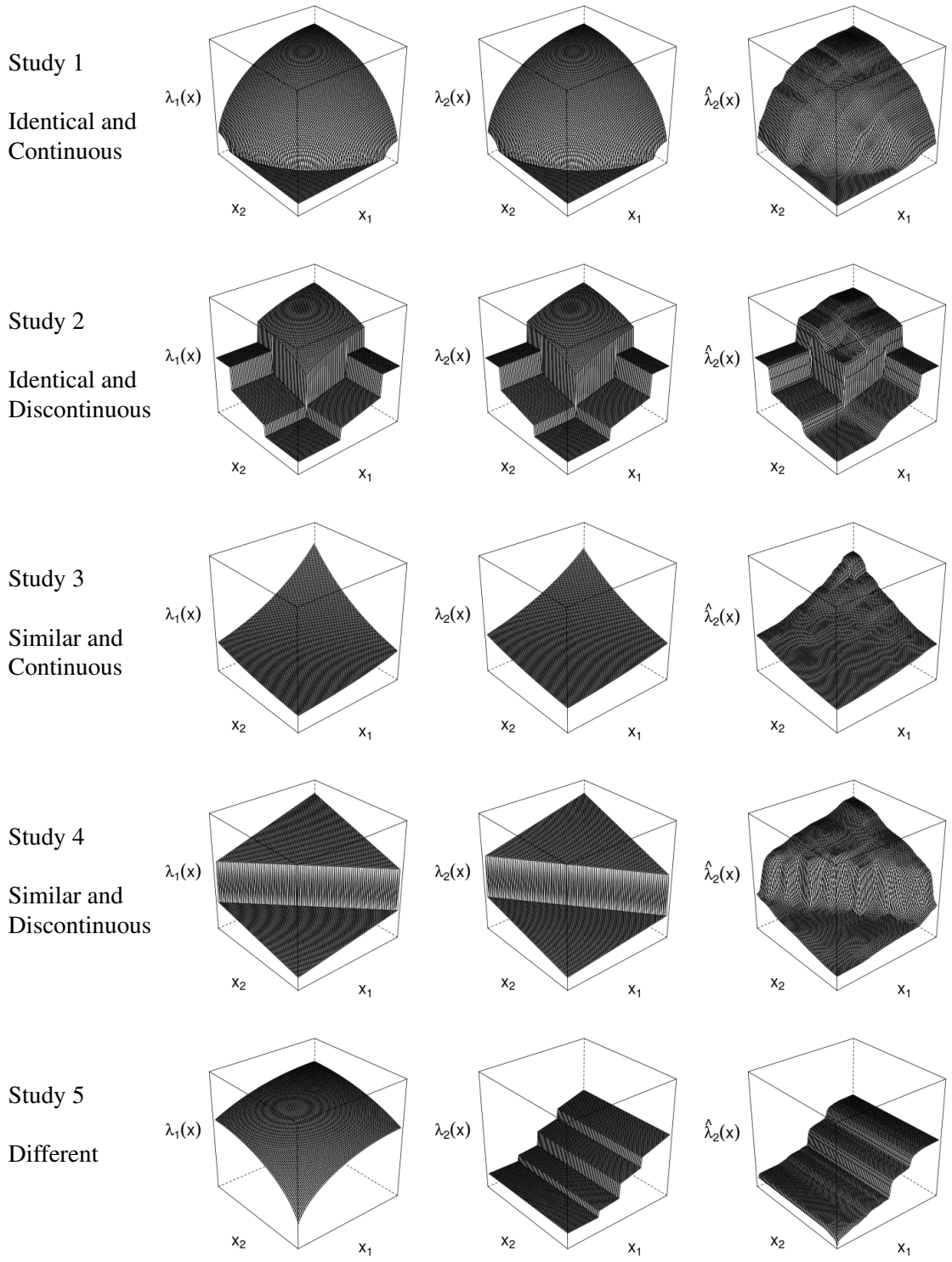

Fig. 3. True functions λ_1 (Column 1) and λ_2 (Column 2), and the posterior mean estimate $\hat{\lambda}_2$ obtained for $\eta = 1000$ (Column 3) for the simulations in §3·2.

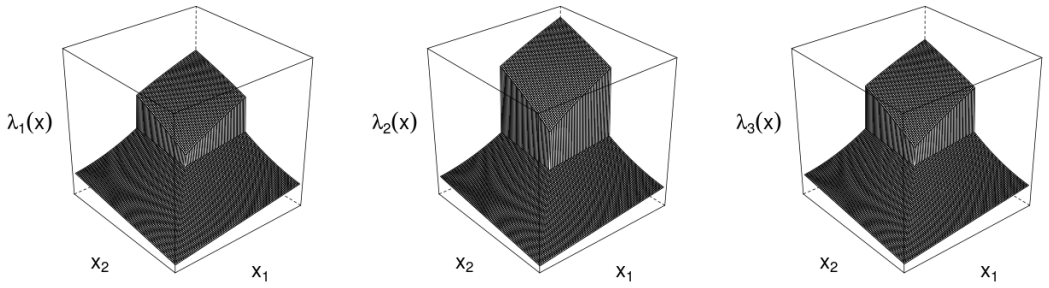

Fig. 4. True functions $(\lambda_1, \lambda_2, \lambda_3)$ in §3·3. The $\lambda_k(x)$-axis $(k = 1, 2, 3)$ is from 0 to 3.

Table 2. *Mean* ($\times 100$) *and standard deviation* ($\times 100$) *of the absolute difference between the truth and posterior mean estimate for* $(\lambda_1, \lambda_2, \lambda_3)$ *in Fig. 4 for the settings* $p = (1{\cdot}0, 0{\cdot}2, 0{\cdot}6, 2{\cdot}0)$ *and* $\omega = 0$

Study	$p = 1{\cdot}0$	$p = 0{\cdot}2$	$p = 0{\cdot}6$	$p = 2{\cdot}0$	$\omega = 0$
1	5·9 (8·3)	5·3 (7·9)	5·2 (7·4)	5·8 (8·5)	6·3 (9·3)
2	6·3 (9·5)	5·9 (10·0)	6·1 (10·0)	6·1 (10·1)	7·0 (11·9)
3	6·4 (8·8)	5·9 (9·0)	6·0 (9·1)	6·9 (9·7)	6·8 (9·6)

3, or requires a large number of points to be approximated, as in Study 4. Conversely, $\eta = 10$ performs better in Study 1 and Study 2, as it does not tend to interpolate linearly when functions switch between a zero and non-zero slope. These findings are consistent with §2: a higher value for η tends to produce smoother estimates, as the sampled functions have more but smaller steps.

3·3. *Sensitivity analysis on* p

We consider $K = 3$ regions with region 2 adjacent to regions 1 and 3 while region 1 and 3 are non-adjacent. Figure 4 shows the true functions $(\lambda_1, \lambda_2, \lambda_3)$, which all exhibit a discontinuity at $(0{\cdot}5, 0{\cdot}5)$, and are more similar for $x_k \in [0, 1]^2 \setminus [0{\cdot}5, 1{\cdot}0]^2$ than for $x_k \in [0{\cdot}5, 1{\cdot}0]^2$ $(k = 1, 2, 3)$. The distribution of x_k $(k = 1, 2, 3)$ varies across three studies while λ_1, λ_2 and λ_3 remain unchanged. Specifically, the studies explore the performance of our approach, subject to the relative intensity of points in subsets of X for which the functions are similar.

We generate 200 data points for each region with variance $\theta_k = 0{\cdot}2^2$ $(k = 1, 2, 3)$. The three studies vary with respect to the number of observations sampled on $[0{\cdot}5, 1{\cdot}0]^2$ for regions 1 and 3 while $x_2 \sim \text{Unif}([0, 1]^2)$ in all of them. Study 1 considers the case $x_k \sim \text{Unif}([0, 1]^2)$ $(k = 1, 3)$. In Study 2, 150 data points are sampled uniformly from $[0{\cdot}5, 1{\cdot}0]^2$ for regions 1 and 3, while only 25 observations are sampled from this subset in Study 3. The remaining 175 and 50 data points in Study 2 and Study 3, respectively, are sampled uniformly from $[0{\cdot}0, 1{\cdot}0]^2 \setminus [0{\cdot}5, 1{\cdot}0]^2$.

We compare four settings for p. The first, $p = 1$, yields the integrated squared difference in expression (2). Settings $p = 0{\cdot}2$ and $p = 0{\cdot}6$ allow for stronger dependence in the lower function values, while $p = 2$ imposes increased dependence for higher function values. The other parameters are fixed to $\eta = 1000$, $d_{1,2} = d_{2,3} = 1$ and $d_{1,3} = 0$.

Table 2 shows that we improve upon $\omega = 0$, except for $p = 2$ in Study 3, and indicates sensitivity to the prior parameter p, as the settings with $p < 1$ perform best. Since a setting $p < 1$ imposes higher dependence on the lower function values, we effectively borrow information across the functions in Fig. 4 to improve estimates for the lower function values without induc-

Fig. 5. Map of the Norwegian municipalities in §4.

ing a large bias on the upper function values. As such, our extended discrepancy measure based on (3) has benefits when compared to the integrated squared distance. Table 2 further indicates that the sensitivity to p depends on where most of the data are observed: if data fall in areas where the functions differ, the sensitivity is lower. The individual summary statistics for each function are provided in the Supplementary Material.

4. CASE STUDY

We consider the Norwegian insurance and weather data used by Haug et al. (2011) and Scheel et al. (2013). The data provide the daily number of insurance claims due to precipitation, surface water, snow melt, undermined drainage, sewage back-flow or blocked pipes at municipality level from 1997 to 2006. Further, the average number of policies held per month and multiple daily weather metrics, such as the amount of precipitation, are recorded.

Table 2 in Scheel et al. (2013) indicates that a generalized linear model underpredicts high numbers of claims, perhaps, due to threshold effects, as the risk of localized flooding only exists for high daily precipitation levels. While linearity may be too strong an assumption, the risk per property increases with precipitation levels, motivating the application of our methodology. We consider the $K = 11$ municipalities in Fig. 5 and explore the effect of precipitation $R_{k,t}$ and $R_{k,t-1}$ $(k = 1, \ldots, 11)$ on day t and $t-1$, as Haug et al. (2011) and Scheel et al. (2013) find these to be the most informative explanatory variables.

Let $N_{k,t}$ and $A_{k,t}$ denote the number of claims and policies, respectively, on day t for municipality k. We model $N_{k,t}$ as binomial with the logit of the daily claim probability, $p_{k,t}$, given by $\lambda_k(R_{k,t}, R_{k,t-1})$. As in §3, we define $\lambda_k(R_{k,t}, R_{k,t-1}) = \mu_k + \varphi_k(R_{k,t}, R_{k,t-1})$ and estimate μ_k and φ_k $(k = 1, \ldots, 11)$. Formally,

$$N_{k,t} \sim \text{Binomial}(A_{k,t}, p_{k,t}), \qquad \text{logit}\, p_{k,t} = \mu_k + \varphi_k(R_{k,t}, R_{k,t-1}).$$

An intrinsic conditional autoregressive prior is defined for $\mu_1, \ldots, \mu_{11}$, and the boundaries of φ_k $(k = 1, \ldots, 11)$ are set to $\delta_{\min} = 0$ and $\delta_{\max} = 10$. The set X is derived as the square spanned by the observed minima and maxima of $R_{k,t}$ across all municipalities and years.

We set $d_{k,k'} = 1$ in (4) if municipalities k and k' are adjacent and $d_{k,k'} = 0$ otherwise. To avoid oversmoothing threshold effects, we select $\eta = 10$, based on our results in §3·2. The sensitivity analysis in §3·3 motivates setting $p < 1$ since we expect the municipalities to exhibit similar vul-

Table 3. *Sum of squared errors of the daily number of claims for 2001 and 2003 for four models with estimates being based on the remaining 8 years between 1997 and 2006*

Municipality	$\omega = \omega_{\text{opt}}$	$\omega = 0$	Constant	Linear model
Ås	14·0	14·1	13·9	14·3
Asker	360·0	357·7	372·5	331·0
Bærum	215·2	234·3	915·1	679·3
Frogn	8·2	8·2	8·5	12·3
Hurum	17·4	17·3	17·7	17·1
Nesodden	20·7	20·5	20·5	20·2
Oppegård	36·1	36·8	26·2	27·6
Oslo	440·4	438·3	412·2	452·3
Røyken	55·9	56·5	63·5	53·3
Ski	39·1	39·2	38·2	38·8
Vestby	18·7	18·8	18·5	18·9
Overall	1225·7	1241·7	1906·8	1665·1

nerability to small amounts of precipitation, while differences in infrastructure, for example, may lead to different effects for higher precipitation levels. We set $p = 0{\cdot}5$. The functions $\lambda_1, \ldots, \lambda_{11}$ are estimated based on 1,000,000 iteration steps,with every 500th sample stored after a burn-in period of 200,000 iterations.

To assess predictive performance, observations for 2001 and 2003 are stored as test data and $\lambda_1, \ldots, \lambda_{11}$ are estimated from the remaining eight years. We consider two competing models: (i) the average daily number of claims in the municipality over the training period and (ii) a linear model with spatially varying parameters (Assunção, 2003). The latter is estimated via 10,000 iterations of a random walk Metropolis scheme, with the first 1,000 samples discarded.

Table 3 shows that our approach is the best in terms of overall predictive performance. The small scale of improvement from $\omega = 0$ to $\omega = \omega_{\text{opt}}$ is due to the large number of training data points; important structures in $\lambda_1, \ldots, \lambda_{11}$ are likely to be captured without borrowing statistical information from adjacent municipalities. Posterior mean plots for Oslo and Hurum are provided in the Supplementary Material, but the function values are omitted for confidentiality reasons.

The largest improvement is achieved for Bærum, which has the highest $N_{k,t}$ over the test period. Hence, the increased flexibility of our approach captures the dynamics leading to large numbers of claims better than competing models. For the other municipalities, the models perform similarly, due to there being zero high-claim days over the test period. This is also indicated by the predictive error of the constant mean model being low for most municipalities. Our approach performs slightly worse than the linear model for Asker, which is due to a single observation $N_{k,t}$. In particular, high precipitation levels caused a count $N_{k,t}$ which was the highest over the full 10-year period.

5. Discussion

Our modelling framework can be extended to a spatio-temporal context. Assume that the intercept changes between observations but the effect of the explanatory variables is temporally stationary. We can then define $\lambda_{k,t}(x) = \mu_{k,t} + \varphi_k(x)$ $(k = 1, \ldots, K)$, similar to §3 and §4. Temporal structure on $\mu_{k,1}, \ldots, \mu_{k,T_k}$ is, for instance, imposed via an autoregressive model. Our approach can also be extended to a setting for which $\lambda_1, \ldots, \lambda_K$ change at specified time points, with temporal structure being imposed analogously to the spatial structure using time-adjacency.

An aspect not discussed is the selection of the number I of marked point processes representing λ_k ($k = 1, \ldots, K$). Since we considered examples with $m = 2$ explanatory variables, $I = 3$. In higher dimensions, however, one may want to restrict I. Assume there exists prior knowledge that continuous variable $x_{k,h}$ ($h = 1, \ldots, m$) is informative and let $X = [0, 1]^m$. The set of processes could then be defined based on the non-empty subsets of $\{1, \ldots, m\}$ which contain h. Consequently, we would represent λ_k ($k = 1 \ldots, K$) via 2^{m-1}, instead of $2^m - 1$, processes.

Our methodology performs well for regression problems with $m = 2$ to $m = 5$ explanatory variables. However, as for other flexible approaches, such as generalized additive models, some issues arise for higher dimensions. Firstly, the computational cost for calculating the prior ratio grows exponentially with m. We reduce this cost by deriving the subset of X affected by the proposal before evaluating the integral in expression (4). Secondly, the monotonicity constraint becomes less restrictive with increasing dimension, leading to potential overfitting. Larger sets of explanatory variables can be accommodated by imposing an additive or semi-parametric structure on λ_k ($k = 1, \ldots, K$), where the lower-dimensional monotonic functions are then estimated jointly. Consequently, our methodology can be applied to higher-dimensional regression problems, but we would recommend a pre-analysis.

Our work can be extended in several ways, such as the construction of other discrepancy measures based, for instance, on the Kullback–Leibler divergence. When estimating ω, parallelized computing techniques, allocating the folds to multiple processors, can reduce the computational time. Further, we arbitrarily fixed the number of folds to $s = 10$ but the value for ω also depends on the number of data points. A larger number of folds may return a more robust estimate.

ACKNOWLEDGEMENT

Rohrbeck gratefully acknowledges funding by the EPSRC via the STOR-i Centre for Doctoral Training. This paper was also financially supported by the Norwegian Research Council. Our work greatly benefited from discussions with Jonathan Tawn, Paul Fearnhead, Elija Arjas, Christopher Nemeth, Sylvia Richardson, Lawrence Bardwell, Jamie Fairbrother, David Hofmeyr, Rob Shooter, Jennifer Wadsworth and Ida Scheel. We also thank Ida Scheel for providing access to the insurance and weather data. Finally, we would like to thank the editors and two referees for suggestions that substantially improved the presentation of the work.

SUPPLEMENTARY MATERIAL

Supplementary material available on *Biometrika* online contains a dependence model for functions with varying support, details on the prior and the sampling scheme, the proofs of Proposition 1 and Theorem 1, an algorithm to detect discontinuities, a simulation study to verify correctness of our implementation, posterior mean plots, trace plots to illustrate mixing and convergence, an example with binomial response data, data plots and posterior mean estimates for two municipalities of the case study, and the C++ and R code for §3.

APPENDIX

Details of the sampling scheme for the marked point processes

We present the acceptance probabilities for the three moves in §2·5; more details are provided in the Supplementary Material. For notational simplicity, let birth and death moves be proposed with equal probability and let Δ_{k,i^*} ($k - 1, \ldots, K;\ i^* = 1, \ldots, I$) denote the marked point process to be updated.

A birth move proposes the addition of a point (ξ^*, δ^*) to Δ_{k,i^*}. Since this increases the dimension of the parameter space, the acceptance probability has to be derived as described by Green (1995). The mapping for adding a point is equal to the identity function and, hence, the determinant of the Jacobian in the acceptance probability is equal to 1. Further, the proposal densities $q(\xi^*)$ and $q\left(\delta^* \mid \xi^*, \Delta_k\right)$ cancel with parts of the prior $\phi\left(\Delta_k \mid \eta\right)$. Formally, the acceptance probability is

$$\min\left\{1, \prod_{t=1}^{T_k} \frac{f\left\{y_{k,t} \mid \lambda_k^*\left(x_{k,t}\right), \theta_k\right\}}{f\left\{y_{k,t} \mid \lambda_k\left(x_{k,t}\right), \theta_k\right\}} \times \prod_{k' \neq k} \frac{\exp\left[-\omega\, d_{k,k'} \int_X \gamma_p\left\{\tilde{\lambda}_k^*(x), \tilde{\lambda}_{k'}(x)\right\}\, \mathrm{d}x\right]}{\exp\left[-\omega\, d_{k,k'} \int_X \gamma_p\left\{\tilde{\lambda}_k(x), \tilde{\lambda}_{k'}(x)\right\}\, \mathrm{d}x\right]} \times \left(1 - \frac{1}{\eta}\right) \frac{N_k + 1}{N_k + I}\right\}.$$

A death or shift move is rejected if Δ_{k,i^*} contains no points. Otherwise, a death move selects one of the n_{k,i^*} existing points with equal probability and proposes to remove it. The acceptance probability for a death move is then

$$\min\left\{1, \prod_{t=1}^{T_k} \frac{f\left\{y_{k,t} \mid \tilde{\lambda}_k^*\left(x_{k,t}\right), \theta_k\right\}}{f\left\{y_{k,t} \mid \tilde{\lambda}_k\left(x_{k,t}\right), \theta_k\right\}} \times \prod_{k' \neq k} \frac{\exp\left[-\omega\, d_{k,k'} \int_X \gamma_p\left\{\tilde{\lambda}_k^*(x), \tilde{\lambda}_{k'}(x)\right\}\, \mathrm{d}x\right]}{\exp\left[-\omega\, d_{k,k'} \int_X \gamma_p\left\{\tilde{\lambda}_k(x), \tilde{\lambda}_{k'}(x)\right\}\, \mathrm{d}x\right]} \times \frac{1}{1 - \frac{1}{\eta}} \frac{N_k + I - 1}{N_k}\right\}.$$

Finally, a shift move changes both the location and mark of an existing point, subject to the partial ordering of the locations in $\Delta_{k,1} \ldots, \Delta_{k,I}$, induced by the monotonicity constraint, being maintained. First, one of the n_{k,i^*} points in Δ_{k,i^*} is selected with equal probability. The proposed location ξ^* is then sampled uniformly on the subset of X_i which maintains the total order in each component of the locations; see Saarela & Arjas (2011) for details. The proposed mark δ^* is then sampled uniformly, subject to the monotonicity constraints. Formally, the acceptance probability is

$$\min\left\{1, \prod_{t=1}^{T_k} \frac{f\left\{y_{k,t} \mid \tilde{\lambda}_k\left(x_{k,t}\right), \theta_k\right\}}{f\left\{y_{k,t} \mid \tilde{\lambda}_k\left(x_{k,t}\right), \theta_k\right\}} \times \prod_{k' \neq k} \frac{\exp\left[-\omega\, d_{k,k'} \int_X \gamma_p\left\{\tilde{\lambda}_k(x), \tilde{\lambda}_{k'}(x)\right\}\, \mathrm{d}x\right]}{\exp\left[-\omega\, d_{k,k'} \int_X \gamma_p\left\{\tilde{\lambda}_k(x), \tilde{\lambda}_{k'}(x)\right\}\, \mathrm{d}x\right]}\right\}.$$

REFERENCES

ANDRIEU, C. & ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**, 697–725.

ASSUNÇÃO, R. M. (2003). Space varying coefficient models for small area data. *Environmetrics* **14**, 453–473.

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.

BACCHETTI, P. (1989). Additive isotonic model. *J. Am. Statist. Assoc.* **84**, 289–294.

BARLOW, R. & BRUNK, H. (1972). The isotonic regression problem and its dual. *J. Am. Statist. Assoc.* **67**, 140–147.

BARRON, A., SCHERVISH, M. J. & WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27**, 536–561.

BEAUMONT, M. A., ZHANG, W. & BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

BELL, M. L., MCDERMOTT, A., ZEGER, S. L., SAMET, J. M. & DOMINICI, F. (2004). Ozone and short-term mortality in 95 US urban communities, 1987–2000. *J. Am. Med. Assoc.* **292**, 2372–2378.

BERGERSEN, L. C., THARMARATNAM, K. & GLAD, I. K. (2014). Monotone splines lasso. *Comp. Statist. Data Anal.* **77**, 336–351.

BESAG, J., YORK, J. & MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43**, 1–20.

BOWMAN, A. W. & AZZALINI, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Clarendon Press.

BOWMAN, A. W., JONES, M. C. & GIJBELS, I. (1998). Testing monotonicity of regression. *J. Comp. Graph. Statist.* **7**, 489–500.

BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26**, 607–616.

BRUNK, H. D., EWING, G. M. & UTZ, W. R. (1957). Minimizing integrals in certain classes of monotone functions. *Pacific J. Math.* **7**, 833–847.

CAHILL, M. & MULLIGAN, G. (2007). Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review* **25**, 174–193.

CONGDON, P. (2006). A model for non-parametric spatially varying regression effects. *Comp. Statist. Data Anal.* **50**, 422–445.

FANG, Z. & MEINSHAUSEN, N. (2012). LASSO isotone for high-dimensional additive isotonic regression. *J. Comp. Graph. Statist.* **21**, 72–91.

FARAH, M., KOTTAS, A. & MORRIS, R. D. (2013). An application of semiparametric Bayesian isotonic regression to the study of radiation effects in spaceborne microelectronics. *Appl. Statist.* **62**, 3–24.

FOTHERINGHAM, A. S., BRUNSDON, C. & CHARLTON, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. Hoboken, NJ: Wiley.

GELFAND, A. E. & KUO, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika* **78**, 657–666.

GEORGII, H.-O. (2011). *Gibbs Measures and Phase Transitions*. Berlin: de Gruyter, 2nd ed.

GHOSAL, S., SEN, A. & VAN DER VAART, A. W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28**, 1054–1082.

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

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Boca Raton, Fla.: Chapman & Hall.

HAUG, O., DIMAKOS, X. K., VÅRDAL, F., J., ALDRIN, M. & MEZE-HAUSKEN, E. (2011). Future building water loss projections posed by climate change. *Scandinavian Actuarial Journal* **2011**, 1–20.

HEIKKINEN, J. & ARJAS, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scand. J. Statist.* **25**, 435–450.

JONES, D. R., SCHONLAU, M. & WELCH, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* **13**, 455–492.

KNORR-HELD, L. (2003). Some remarks on Gaussian Markov random field models for disease mapping. In *Highly Structured Stochastic Systems*, P. J. Green, N. L. Hjort & S. Richardson, eds. Oxford: Oxford University Press, pp. 203–207.

LIN, L. & DUNSON, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* , 303–317.

LUSS, R., ROSSET, S. & SHAHAR, M. (2012). Efficient regularized isotonic regression with application to gene–gene interaction search. *Ann. Appl. Statist.* **6**, 253–283.

MØLLER, J., PETTITT, A. N., REEVES, R. & BERTHELSEN, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**, 451–458.

PENTTINEN, A., STOYAN, D. & HENTTONEN, H. M. (1992). Marked point processes in forest statistics. *Forest Science* **38**, 806–824.

RAMSAY, J. O. (1998). Estimating smooth monotone functions. *J. R. Statist. Soc. B* **60**, 365–375.

RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis*. New York: Springer, 2nd ed.

ROUSTANT, O., GINSBOURGER, D. & DEVILLE, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *J. Statist. Softw.* **51**, 1–55.

ROYSTON, P. (2000). A useful monotonic non-linear model with applications in medicine and epidemiology. *Statist. Med.* **19**, 2053–2066.

RUE, H. & HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, Fla.: Chapman & Hall.

SAARELA, O. & ARJAS, E. (2011). A method for Bayesian monotonic multiple regression. *Scand. J. Statist.* **38**, 499–513.

SCHEEL, I., FERKINGSTAD, E., FRIGESSI, A., HAUG, O., HINNERICHSEN, M. & MEZE-HAUSKEN, E. (2013). A Bayesian hierarchical model with spatial variable selection: The effect of weather on insurance claims. *Appl. Statist.* **62**, 85–100.

SCOTT, J. G., SHIVELY, T. S. & WALKER, S. G. (2015). Nonparametric Bayesian testing for monotonicity. *Biometrika* **102**, 617–630.

SHIVELY, T. S., SAGER, T. W. & WALKER, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. R. Statist. Soc. B* **71**, 159–175.

TUTZ, G. & LEITENSTORFER, F. (2007). Generalized smooth monotonic regression in additive modeling. *J. Comp. Graph. Statist.* **16**, 165–188.

WAKEFIELD, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics* **8**, 158–183.

WALKER, S. G. & HJORT, N. L. (2001). On Bayesian consistency. *J. R. Statist. Soc. B* **63**, 811–821.

WALLER, L. A. & GOTWAY, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: Wiley.

WILSON, A., REIF, D. M. & REICH, B. J. (2014). Hierarchical dose–response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics* **70**, 237–246.

ZHANG, L. & SHI, H. (2004). Local modeling of tree growth by geographically weighted regression. *Forest Science* **50**, 225–244.

[*Received April* 2012. *Revised October* 2015]