Locally Adaptive Function Estimation for Binary Regression Models

Alexander Jerak[1] and **Stefan Lang**[*1]

[1] Department of Statistics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany.

Summary

In this paper we present a nonparametric Bayesian approach for £tting unsmooth or highly oscillating functions in regression models with binary responses. The approach extends previous work by Lang et al. (2002) for Gaussian responses. Nonlinear functions are modelled by £rst or second order random walk priors with locally varying variances or smoothing parameters. Estimation is fully Bayesian and uses latent utility representations of binary regression models for ef£cient block sampling from the full conditionals of nonlinear functions.

Key words: adaptive smoothing, forest health data, highly oscillating functions, MCMC, random walk priors, unsmooth functions, variable smoothing parameter.

1 Introduction

Nonparametric methods for £tting smooth curves, such as kernel, local or spline regression, are now widely available and accepted. However, these methods can have bad performance when estimating unsmooth functions which have jumps, edges, or which are highly oscillating. Two prominent approaches in nonparametric regression with Gaussian responses that adapt to such spatial heterogeneity are local regression with variable bandwidth (Fan and Gijbels, 1995) or wavelet shrinkage regression (Donoho and Johnstone, 1994). Currently, these methods are restricted to continuous responses and there is a clear lack of methodology and experience for non-Gaussian responses.

In this paper we present a nonparametric fully Bayesian method for £tting unsmooth and highly oscillating functions in regression models with *binary responses*. The approach extends recent work by Lang et al. (2002) for Gaussian responses. Our approach uses a two-stage prior for the unknown regression function. The £rst stage are £rst or second order random walk models as proposed in Fahrmeir and Lang (2001a) and Fahrmeir and Lang (2001b). The second stage consists of analogous smoothness priors for varying variances of the random walk model errors used in the £rst stage leading to locally adaptive dependent variances. The varying variances in our method correspond to variable smoothing parameters and make the prior more ¤exible for modelling functions with differing curvature. We compare our approach with random walk priors with a global variance as well as locally adaptive independent variances. The latter has been already used e.g. by Knorr-Held (1999) in the context of dynamic models.

Bayesian inference is based on latent *utility representations of binary regression models*, see Albert and Chib (1993) for probit models and Holmes and Held (2003) for logit models. The advantage of augmenting the data by latent utilities is that the full conditionals of unknown parameters are Gaussian and ef£cient MCMC sampling schemes developed for Gaussian responses can be exploited.

The rest of this paper is organized as follows: Section 2 describes our Bayesian model for locally adaptive function estimation and gives details about Bayesian inference. Section 3 illustrates the performance of our approach by selected results from an extensive simulation study. In Section 4 the practicability is

* Corresponding author: e-mail: lang@stat.uni-muenchen.de, Phone: +49 89 2180 6404, Fax: +49 89 2180 5040

demonstrated by a complex application on forest health data. The £nal section 5 summarizes the paper and highlights directions for future research.

2 Model speci£cation and Bayesian inference

2.1 Binary response models

Consider regression situations, where observations $(y_t, \boldsymbol{z}_t)$, $t = 1, \ldots, T$, on a binary response Y and covariates $\boldsymbol{Z}$ are given. The most widely used models for binary data are logit or probit models. Given covariates, the responses y_t are binomially distributed, i.e. $y_t \mid \boldsymbol{z}_t \sim B(1, \pi_t)$ with the probability of success $\pi_t = P(y_t = 1 \mid \boldsymbol{z}_t) = E(y_t \mid \boldsymbol{z}_t)$ being modelled as

$$\pi_t = \frac{\exp(\eta_t)}{1 + \exp(\eta_t)}$$

for logit models or

$$\pi_t = \Phi(\eta_t)$$

for probit models. Here, η_t is the predictor that models the in¤uence of the covariates. With a linear predictor

$$\eta_t = \boldsymbol{z}_t' \boldsymbol{\beta} \tag{1}$$

one gets parametric models. In many practical situations, as in our application on forest health data, the assumption of linear effects of the covariates on the predictor is too restrictive. Suppose that $\boldsymbol{z}_t$ is divided into a vector of continuous covariates $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$ whose in¤uence is assumed to be possibly nonlinear, and a vector of categorical covariates $\boldsymbol{w}_t = (w_{t1}, \ldots, w_{tq})'$. Then, we replace the simple linear predictor (1) by the semiparametric additive predictor

$$\eta_t = f_1(x_{t1}) + \cdots + f_p(x_{tp}) + \boldsymbol{w}_t' \boldsymbol{\beta}, \tag{2}$$

where we assume possibly nonlinear effects $f_1, \ldots, f_p$ for the continuous covariates. In this paper, the primary focus is on modelling functions with discontinuities or differing curvature. We will discuss appropriate prior speci£cations for functions of this kind in the next section.
For Bayesian inference, it is quite useful to express binary regression models in terms of latent utilities, see e.g. Fahrmeir and Tutz (2001). Introducing the latent utilities

$$U_t = \eta_t + \epsilon_t, \qquad t = 1, \ldots, T, \tag{3}$$

with i.i.d. errors ϵ_t, we de£ne $y_t = 1$ if $U_t > 0$ and $y_t = 0$ if $U_t \leq 0$.
In general, we assume

$$\epsilon_t \mid \lambda_t \sim N(0, \lambda_t) \tag{4}$$

conditional on some variances λ_t. Depending on the distributional choice for the $\lambda_t' s$ we get different models. The assumption $\lambda_t \equiv 1$, i.e. $\epsilon_t \sim N(0, 1)$, yields a probit model. A logit model is obtained by assuming $\lambda_t = 4\psi_t^2$, where ψ_t follows a Kolmogorov-Smirnov distribution (Devroye, 1986). Hence, ϵ_t is a scale mixture of normal form with a marginal logistic distribution (Andrews and Mallows, 1974). Note, that a logit model could be (well) approximated by assuming a t-distribution for the ϵ_t's with a certain degree ν of freedom. A t-distribution may again be expressed as a scale mixture of normals with $\lambda_t \sim IG(\nu/2, \nu/2)$. An approximative logit model is then obtained with $\nu = 8$ (Albert and Chib, 1993).

2.2 Prior models

For Bayesian inference, the unknown functions f_j of covariates X_j, $j = 1, \ldots, p$, or more exactly the corresponding vectors of function evaluations, are considered as random and must be supplemented by appropriate prior distributions.

We model the unknown function f_j by random walk priors which are commonly used in state space or dynamic models to estimate nonlinear time trends, see e.g. Fahrmeir and Tutz (2001), Ch. 8, and Knorr-Held (1999) for Bayesian inference based on MCMC techniques. Fahrmeir and Lang (2001a) and Fahrmeir and Lang (2001b) use random walk priors for Bayesian inference in generalized additive mixed models. Random walks also serve as a main building block for de£ning Bayesian versions of P(enalized)-splines, see Eilers and Marx (1996) for the classical version and Lang and Brezger (2004) for the Bayesian approach. In this paper we employ random walk priors with *locally adaptive variances* as introduced in Lang et al. (2002) to model functions with discontinuities or differing curvature. We start by reviewing random walks with global variances.

2.2.1 Random walks with global variances

Let $x_j^{(1)} < x_j^{(2)} < \cdots < x_j^{(S_j)}$ denote the ordered sequence of observed values for covariate X_j. De£ne $f_{js} := f_j(x_j^{(s)})$, $s = 1, \ldots, S_j$, and let $\boldsymbol{f}_j = (f_{j1}, \ldots, f_{jS_j})'$ be the vector of function evaluations.

Assuming equidistant covariate values, a common prior for a smooth function f_j is a £rst or second order random walk model

$$f_{js} = f_{j,s-1} + u_{js} \quad (RW1) \quad \text{or} \quad f_{js} = 2f_{j,s-1} - f_{j,s-2} + u_{js} \quad (RW2) \tag{5}$$

with Gaussian errors $u_{js} \sim N(0, \tau_j^2)$ and diffuse priors $f_{j1} \propto constant$, or f_{j1} and $f_{j2} \propto constant$, for initial values, respectively. Both speci£cations act as smoothness priors that penalize too rough functions f_j. A £rst order random walk penalizes abrupt jumps $f_{js} - f_{j,s-1}$ between successive states and a second order random walk penalizes deviations from the linear trend $2f_{j,s-1} - f_{j,s-2}$. Note that a second order random walk is derived by computing second differences, i.e. the differences of neighboring £rst order differences.

Random walk priors can be equivalently de£ned in a more symmetric way via the conditional distributions of f_{js} given the left *and* right neighboring function evaluations, i.e. $f_{j,s-1}$, $f_{j,s+1}$ in case of a £rst order random walk and $f_{j,s-2}$, $f_{j,s-1}$, $f_{j,s+1}$, $f_{j,s+2}$ in case of a second order random walk. For instance, a second order random walk is de£ned in a symmetric way by

$$f_{js} \,|\, \cdot \sim \begin{cases} N(2f_{j2} - f_{j3}, \tau_j^2) & s = 1 \\ N(\frac{2}{5}f_{j1} + \frac{4}{5}f_{j3} - \frac{1}{5}f_{j4}, \tau_j^2/5) & s = 2 \\ N(-\frac{1}{6}f_{j,s-2} + \frac{2}{3}f_{j,s-1} + & s = 3, \ldots, S_j - 2 \\ \frac{2}{3}f_{j,s+1} - \frac{1}{6}f_{j,s+2}, \tau_j^2/6) & \\ N(-\frac{1}{5}f_{j,S_j-3} + \frac{4}{5}f_{j,S_j-2} + \frac{2}{5}f_{jS_j}, \tau_j^2/5) & s = S_j - 1 \\ N(-f_{j,S_j-2} + 2f_{j,S_j-1}, \tau_j^2) & s = S_j \end{cases} \tag{6}$$

where $f_{js} \,|\, \cdot$ denotes the conditional distribution of f_{js} given the left and right neighbors. For $s = 3, \ldots, S_j - 2$ the conditional mean in (6) can be interpreted as a locally quadratic £t to the neighboring parameters, see Figure 1 for an illustration. Similarly, the conditional mean for a £rst order random walk can be interpreted as a locally linear £t.

Generalizations to situations with non-equally spaced observations are given in Fahrmeir and Lang (2001a). For instance, a second order random walk could be generalized to

$$f_{js} = \left(1 + \frac{\delta_{js}}{\delta_{j,s-1}}\right) f_{j,s-1} - \frac{\delta_{js}}{\delta_{j,s-1}} f_{j,s-2} + u_{js},$$

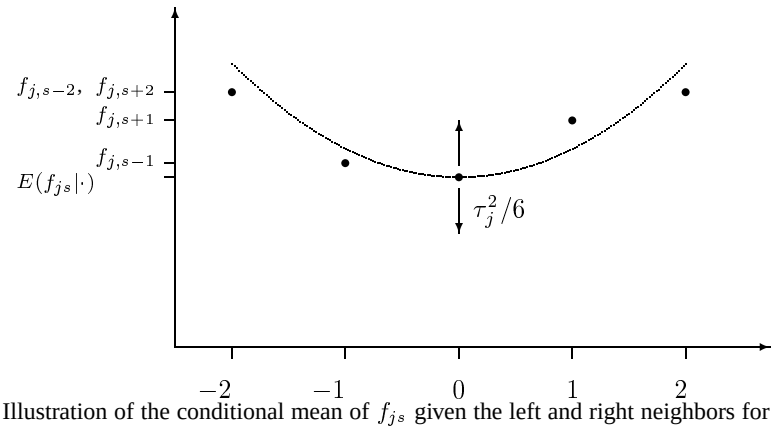

Fig. 1 Illustration of the conditional mean of f_{js} given the left and right neighbors for a second order random walk prior. The conditional mean of f_{js} given the left and right neighbors is obtained by £tting a quadratic polynomial to the four points $(f_{j,s-2}, -2)$, $(f_{j,s-1}, -1)$, $(f_{j,s+1}, 1)$ and $(f_{j,s+2}, 2)$.

where $u_{js} \sim N(0; w_{js}\tau_j^2)$, $\delta_{js} = x_j^{(s)} - x_j^{(s-1)}$ and w_{js} is an appropriate weight, e.g. $w_{js} = \delta_{js}$. To avoid notational confusions, we restrict the presentation for the rest of the paper to the case of equally spaced observations. An extension to the more general case is straightforward and supported by our software.
The amount of smoothness is controlled by the variance parameter τ_j^2 which corresponds to the inverse smoothing parameter in a classical approach. The larger (respectively smaller) the variance, the rougher (respectively smoother) are the estimated functions, see also Figure 1. The amount of smoothness can be estimated simultaneously with the unknown function by de£ning a highly dispersed inverse gamma prior

$$\tau_j^2 \sim IG(a_j, b_j) \tag{7}$$

in a further stage of the hierarchy.

2.2.2 Random walks with locally adaptive variances

For unsmooth or highly oscillating functions, as primarily considered in this paper, the assumption of a global variance or smoothing parameter is not appropriate. We illustrate the dif£culties with a simulated data set taken from our simulation study in Section 3. Figure 6 (g) shows $n = 400$ binary observations simulated from a probit model with predictor $\eta = f(x)$. The underlying "true" function is a variant of the so called Doppler function and is visualized in Figure 6 (h) (dashed line), see Section 3.2. Figure 6 (h) displays function estimates and pointwise credible intervals based on a RW2 with global variance. From left to right the curvature of the function decreases. The global variance is, however, not able to adapt to the changing curvature and yields too wiggled estimates in the less curved parts in the right. If the variance is allowed to adapt appropriately to the decreasing curvature, i.e. it is allowed to decrease as well, the £t improves considerably, see Figure 6 (j).
To overcome the dif£culties we therefore replace the global variance τ_j^2 by *locally adaptive* variances τ_{js}^2. In the following we will discuss approaches with both *stochastic dependent* and *independent* τ_{js}^2. We start with the *stochastic dependent* variant.

Locally adaptive dependent variances

Following Lang et al. (2002) we set $\tau_{js}^2 = \exp(h_{js})$. For the parameters $\boldsymbol{h}_j = (h_{jd}, \ldots, h_{jS_j})'$ we add a second smoothness prior in the form of £rst or second order random walks, i.e.

$$h_{js} = h_{j,s-1} + v_{js} \quad \text{or} \quad h_{js} = 2h_{j,s-1} - h_{j,s-2} + v_{js} \tag{8}$$

with $v_{js} \sim N(0, \sigma_j^2)$. The index d depends on the choice of the prior for f_j, for a RW1 $d = 2$ and for a RW2 $d = 3$. Once again, a highly dispersed inverse gamma prior is assigned to the variance parameter σ_j^2, i.e. $\sigma_j^2 \sim IG(a_j', b_j')$.
The improvement in function estimation that can be achieved through locally adaptive (dependent) variances is demonstrated with the simulated data set already mentioned. Figures 6 (h) and (j) compare the £ts obtained with a global variance and locally adaptive variances.

Locally adaptive independent variances

Alternatively we may assume *independent* local variances τ_{js}^2 instead of dependent variances, which may be particularly useful for functions with discontinuities. Reparameterizing τ_{js}^2 to $\gamma_{js}\tau_j^2$, we assume a continuous mixture of normals

$$u_{js} \,|\, \gamma_{js}, \tau_j^2 \sim N(0, \gamma_{js}\tau_j^2) \tag{9}$$

for the error distributions of the random walk. Assuming i.i.d. inverse gamma priors

$$\gamma_{js} \sim IG(v/2, v/2) \tag{10}$$

for γ_{js} and again prior (7) for τ_j^2, the marginal distribution of the errors is a Student distribution with v degrees of freedom. The case $v = 1$ of a Cauchy distribution is of special interest as a robust prior and is used for the rest of this paper.
We again illustrate the possible improvement in function estimation with a simulated data set. Figure 2 (g) displays $n = 250$ binary observations simulated from a probit model with predictor $\eta = f(x)$. The function f is displayed in Figure 2 (h) (dashed line) and is a variant of the so called Blocks function with three discontinuities. Figures 2 (h) and (i) compare the £ts obtained with £rst order random walks and a global variance and locally adaptive independent variances. Clearly, the jumps are much better detected with locally adaptive variances as they are allowed to increase at the discontinuities and decrease thereafter. Note that the £t based on locally dependent variances in 2 Figure (j) also improves the adaption to the jumps but to a lesser extent as the independent variances. The reason is that the stochastic dependent variances cannot adapt with the same speed as stochastic independent variances. All £ts are comparatively rough, which is, however, not a particular feature of random walk models and in fact caused by the weak information contained in the data. Note that other well known smoothing techniques, e.g. LOESS in Figure 2 (k), yield £ts that are even more jagged. The best result in this respect is again obtained by the random walk with locally adaptive independent variances. Of course, for increasing sample size the resulting £ts are getting smoother, see Section 3 for more details.

2.2.3 Matrix notation

For describing MCMC inference in the next section some matrix notation will be useful.
The predictor (2) can be written in matrix notation as

$$\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{f}_1 + \cdots + \boldsymbol{X}_p\boldsymbol{f}_p + \boldsymbol{W}\boldsymbol{\beta}.$$

Here, $\boldsymbol{X}_j$ are (pseudo) design matrices of size $n \times S_j$. The entry $X_j(i, s)$ in the i-th row and s-th column is one if the s-th ordered covariate value $x_j^{(s)}$ has been observed for observation i and zero otherwise.
The prior for the function evaluations $\boldsymbol{f}_j$ of covariate X_j can be written in terms of a penalty matrix $\boldsymbol{K}_j$. The joint distribution of $\boldsymbol{f}_j$ is easily computed as the product of conditional densities de£ned by (5) and can be written in the general form

$$p(\boldsymbol{f}_j \,|\, \boldsymbol{\tau}_j^2) \propto \exp(-\frac{1}{2}\boldsymbol{f}_j'\boldsymbol{K}_j\boldsymbol{f}_j) \tag{11}$$

where $\boldsymbol{\tau}_j^2 = (\tau_{jd}^2, \ldots, \tau_{jS_j}^2)'$ is the vector of variances. The form of the penalty matrix and the variance vector depends on the *order of random walks* and the *assumptions on the variances*.
The general form of the penalty matrix is

$$\boldsymbol{K}_j = \boldsymbol{D}_k' \boldsymbol{Q}_j \boldsymbol{D}_k \tag{12}$$

where $\boldsymbol{D}_k$ is a difference matrix ($k = 1$ for a RW1 and $k = 2$ for a RW2). The matrix $\boldsymbol{Q}_j = diag(1/\tau_{jd}^2, \ldots, 1/\tau_{jS_j}^2)$ is a diagonal matrix of order $(S_j - k) \times (S_j - k)$ with entries $1/\tau_{js}^2$. For a RW1, $\boldsymbol{D}_1$ is the $(S_j - 1) \times S_j$ upper two-diagonal matrix with entries (-1,1) de£ning the vector of £rst differences $\boldsymbol{D}_1 \boldsymbol{f}_j = (f_{j2} - f_{j1}, \ldots, f_{jS_j} - f_{j,S_j-1})'$. Correspondingly, for a RW2, $\boldsymbol{D}_2$ de£nes the vector of second differences, i.e. $\boldsymbol{D}_2 \boldsymbol{f}_j = \boldsymbol{D}_1(\boldsymbol{D}_1 \boldsymbol{f}_j)$.
If a global variance is assumed, the vector $\boldsymbol{\tau}_j^2$ is given by $\boldsymbol{\tau}_j^2 = (\tau_j^2, \ldots, \tau_j^2)'$ and we obtain $\boldsymbol{Q}_j = 1/\tau_j^2 \boldsymbol{I}$. Assuming locally adaptive variances we get $\boldsymbol{\tau}_j^2 = (\exp(h_{jd}), \ldots, \exp(h_{jS_j}))'$ for the variant with locally dependent variances and $\boldsymbol{\tau}_j^2 = (\gamma_{jd}\tau_j^2, \ldots, \gamma_{jS_j}\tau_j^2)'$ for independent variances. Note that $rank(\boldsymbol{K}_j) = S_j - 1$ for a RW1 and $rank(\boldsymbol{K}_j) = S_j - 2$ for a RW2, i.e. the prior for $\boldsymbol{f}_j$ is improper (the posterior, however, is proper, see Speckman and Sun (2003)).
We £nally note, that the prior for the variance parameters $\boldsymbol{h}_j$ of locally dependent variances can be written as

$$p(\boldsymbol{h}_j \mid \sigma_j^2) \propto \exp(-\frac{1}{2}\boldsymbol{h}_j' \boldsymbol{L}_j \boldsymbol{h}_j),$$

where the penalty matrix $\boldsymbol{L}_j$ is given by $\boldsymbol{L}_j = \frac{1}{\sigma_j^2}\boldsymbol{D}_k'\boldsymbol{D}_k$. This is in complete analogy to the prior (11) for $\boldsymbol{f}_j$.

2.2.4 Additional prior assumptions

We complete our model with a few additional prior assumptions:

1. Priors for the £xed effects parameters $\boldsymbol{\beta}$ are assumed to be independent and diffuse, i.e. $\beta_j \propto const$, $j = 1, \ldots, q$.
2. For given covariates and parameters observations y_t are conditionally independent.
3. Priors for the function evaluations $\boldsymbol{f}_j$, $j = 1, \ldots, p$, and £xed effects are mutually independent.

2.3 Bayesian inference via MCMC

For binary regression models useful and ef£cient sampling schemes can be developed on the basis of the latent variables representation with Gaussian errors de£ned in (3) and (4).
Bayesian inference is based on the posterior augmented by the latent variables $\boldsymbol{U} = (U_1, \ldots, U_T)'$ with variances $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T)'$ introduced in (3) and (4). The general form of the posterior is given by

$$\begin{aligned} p(\ldots, \boldsymbol{f}_j, \boldsymbol{\tau}_j^2, \ldots, \boldsymbol{\beta}, \boldsymbol{U}, \boldsymbol{\lambda} \mid \boldsymbol{y}) \quad \propto \quad & p(\boldsymbol{y} \mid \boldsymbol{U}) p(\boldsymbol{U} \mid \boldsymbol{\lambda}, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta}) \\ & p(\boldsymbol{\lambda}) \prod_{j=1}^{p} \left\{ p(\boldsymbol{f}_j \mid \boldsymbol{\tau}_j^2) p(\boldsymbol{\tau}_j^2) \right\} \end{aligned}$$

with $p(\boldsymbol{y} \mid \boldsymbol{U}) = \prod_t p(y_t|U_t)$. The conditional likelihood $p(y_t|U_t)$ is given by

$$p(y_t|U_t) = I(U_t > 0)I(y_t = 1) + I(U_t \leq 0)I(y_t = 0),$$

due to the fact that $p(y_t|U_t)$ is one if U_t obeys the constraint imposed by the observed value of y_t. We use $p(\boldsymbol{\tau}_j^2)$ as a generic symbol for the prior of the variances of the random walks $\boldsymbol{f}_j$. For a random walk with

a global variance $p(\boldsymbol{\tau}_j^2) = p(\tau_j^2)$ is an inverse gamma density de£ned by (7). In case of locally adaptive variances we have $p(\boldsymbol{\tau}_j^2) = p(\boldsymbol{h}_j \mid \sigma_j^2)p(\sigma_j^2)$ for dependent variances and $p(\boldsymbol{\tau}_j^2) = \prod_s p(\gamma_{js})p(\tau_j^2)$ for independent variances where $p(\gamma_{js})$ and $p(\tau_j^2)$ are de£ned in (10) and (7).
MCMC sampling is based on successive updating of the parameter blocks $U_t, \lambda_t, t = 1, \ldots, T, \boldsymbol{f}_j, \boldsymbol{\tau}_j^2, j = 1, \ldots, p$ and $\boldsymbol{\beta}$. In the following we derive the full conditionals of the parameter blocks and discuss how the parameters are updated in an MCMC sampler by drawing random numbers from the full conditionals:

Updating the latent utilities U_t's and their variances λ_t

The algorithm for updating U_t and λ_t depends on the assumptions about the categorical regression model. We may distinguish between probit models, logit models and models with t-distributed error in (4):

- *Probit models*
  Assuming a probit model only the latent utilities U_t must be updated because the variances $\lambda_t \equiv 1$ are non-stochastic. The full conditionals for the U_t's are truncated normals. For $y_t = 1$ we obtain

  $$U_t|\cdot \sim N(\eta_t, 1)I(U_t > 0) \tag{13}$$

  and for $y_t = 0$ we get

  $$U_t|\cdot \sim N(\eta_t, 1)I(U_t \leq 0). \tag{14}$$

  Drawing random numbers from a truncated normal distribution poses no further problems, see e.g. Robert (1995) for an algorithm.

- *Logit models*
  For binary logit models, sampling becomes more complicated and less ef£cient. The main difference to the probit case is that sampling the additional variance parameters λ_t is computationally intensive. Holmes and Held (2003) propose to update U_t and λ_t *jointly*. More speci£cally, the full conditional of U_t, λ_t is factorized into

  $$\begin{aligned} p(U_t, \lambda_t \mid \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta}, y_t) &= p(U_t \mid \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta}, y_t) \\ &\quad\; p(\lambda_t \mid U_t, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta}) \end{aligned}$$

  and updated parameters are then obtained by £rst drawing from the marginal distribution $p(U_t \mid \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta}, y_t)$ of the $U_t's$ followed by drawing from $p(\lambda_t \mid U_t, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta})$. The marginal densities of the $U_t's$ are truncated logistic distributions while $p(\lambda_t \mid U_t, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta})$ is not of standard form. Detailed algorithms for sampling from both distributions can be found in Holmes and Held (2003), Appendix A3 and A4. This updating scheme is considerably slower than the scheme for probit models. The main reason is that drawing random numbers from $p(\lambda_t \mid U_t \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta})$ is based on rejection sampling and therefore time consuming. From a computational point of view we therefore prefer probit models because updates of the full conditionals of the latent utilities are faster.

- *t-distributed errors*
  If a t-distribution is assumed for the errors, the full conditionals for λ_t are inverse gamma distributions, i.e.

  $$\lambda_t \mid \cdot \sim IG(\nu/2 + 1/2, \nu/2 + (U_t - \eta_t)^2/2). \tag{15}$$

  Hence, the latent utilities U_t and variances λ_t can be updated by £rst drawing from the full conditional of the $U_t's$ given in (13) and (14) followed by updating λ_t by drawing from (15).

Updating the vectors of function evaluations $\boldsymbol{f}_j$

The advantage of augmenting the posterior by the latent variables is that the full conditionals for the vector of function evaluations $\boldsymbol{f}_j$ become (multivariate) Gaussian, allowing the usage of the sampling schemes developed for Gaussian responses in Lang et al. (2002) with only minor changes. The precision matrix $\boldsymbol{P}_j$ and the mean $\boldsymbol{m}_j$ of the Gaussian full conditional are given by

$$\boldsymbol{P}_j = \boldsymbol{X}_j'\boldsymbol{\Lambda}\boldsymbol{X}_j + \boldsymbol{K}_j, \quad \boldsymbol{m}_j = \boldsymbol{P}_j^{-1}\boldsymbol{X}_j'\boldsymbol{\Lambda}(\boldsymbol{U} - \tilde{\boldsymbol{\eta}}), \tag{16}$$

where $\boldsymbol{\Lambda} = diag(1/\lambda_1, \ldots, 1/\lambda_T)$ and $\tilde{\boldsymbol{\eta}}$ is the part of the predictor associated with all remaining effects in the model.

Updating the variance parameters $\boldsymbol{\tau}_j^2$

The updating scheme for $\boldsymbol{\tau}_j^2$ depends on the assumption about the variances of the random walk priors $\boldsymbol{f}_j$.

- *Random walk with a global variance*
  The full conditional for the global variance τ_j^2 is an inverse gamma distribution, i.e.

$$\tau_j^2 \mid \cdot \sim IG(a_j + \frac{1}{2}rank(\boldsymbol{K}_j), b_j + \frac{1}{2}\sum_{s=d}^{S_j} u_{js}^2) \tag{17}$$

  and u_{js} determined by (5). Hence, updating of τ_j^2 can be done by a simple Gibbs step. Having sampled τ_j^2 we set $\boldsymbol{\tau}_j^2 = (\tau_j^2, \ldots, \tau_j^2)'$ and recompute the penalty matrix $\boldsymbol{K}_j$ according to (12).

- *Random walk with locally adaptive dependent variances*
  It turns out, that updating of the variance parameters vector $\boldsymbol{h}_j$ in one step is not feasible because of too small acceptance rates. Therefore, the parameter vector $\boldsymbol{h}_j$ must be further divided into smaller blocks $\boldsymbol{h}_{j,[lr]} = (h_{jl}, \ldots, h_{jr})'$, usually of size 10-20. The full conditionals for the variance parameters $\boldsymbol{h}_{j[lr]}$ are not in closed form. We use an MH-algorithm with conditional prior proposals of Knorr-Held (1999) for drawing from the full conditionals $p(\boldsymbol{h}_{j[lr]} \mid \cdot)$. MH steps consist of drawing a proposal $\boldsymbol{h}_{j,[lr]}^*$ from the *conditional prior*

$$p(\boldsymbol{h}_{j,[lr]} \mid h_{jd}, \ldots, h_{j,l-1}, h_{j,r+1}, \ldots, h_{jS_j}, \sigma_j^2) \tag{18}$$

  and accepting it with probability

$$\min\left(1, \frac{\prod_{s=l}^{r} p(f_{js} \mid f_{j,s-1}, \ldots, f_{j1}, h_{js}^\star)}{\prod_{s=l}^{r} p(f_{js} \mid f_{j,s-1}, \ldots, f_{j1}, h_{js})}\right). \tag{19}$$

  The conditional distributions $p(f_{js} \mid f_{j,s-1}, \ldots, f_{j1}, h_{js})$ are Gaussian, de£ned by the random walk priors (5). The conditional prior distribution of $\boldsymbol{h}_{j,[lr]}$ given the rest is a multivariate Gaussian distribution. Its mean and covariance matrix can be written in terms of the precision matrix $\boldsymbol{L}_j$ of $\boldsymbol{h}_j$. Let $\boldsymbol{L}_{j[lr]}$ denote the sub-matrix of $\boldsymbol{L}_j$, given by the rows and columns numbered l to r and let $\boldsymbol{L}_{j,[1,l-1]}$ and $\boldsymbol{L}_{j,[r+1,S_j]}$ denote the matrices left and right of $\boldsymbol{L}_{j,[lr]}$. Then the (conditional) mean $\boldsymbol{\mu}_{lr}$ and the covariance matrix $\boldsymbol{\Sigma}_{lr}$ are given by

$$\boldsymbol{\mu}_{lr} = \begin{cases} -\boldsymbol{L}_{j,[lr]}{}^{-1}\boldsymbol{L}_{j,[r+1,S_j]}\boldsymbol{h}_{j,[r+1,S_j]} & l = d \\ -\boldsymbol{L}_{j,[lr]}{}^{-1}\boldsymbol{L}_{j,[d,l-1]}\boldsymbol{h}_{j,[d,l-1]} & r = S_j \\ -\boldsymbol{L}_{j,[lr]}{}^{-1}(\boldsymbol{L}_{j,[d,l-1]}\boldsymbol{h}_{j,[d,l-1]} & \text{else} \\ +\boldsymbol{L}_{j,[r+1,S_j]}\boldsymbol{h}_{j,[r+1,S_j]}) & \end{cases} \tag{20}$$

and

$$\boldsymbol{\Sigma}_{lr} = \boldsymbol{L}_{j,[lr]}{}^{-1}, \tag{21}$$

respectively. Details about ef£cient computation of the mean $\boldsymbol{\mu}_{lr}$ and the choice of the block size $r - l + 1$ can be found in Fahrmeir and Lang (2001a).

The full conditionals for the variance parameters σ_j^2 are inverse gamma distributions given by

$$\sigma_j^2 \,|\, \boldsymbol{h}_j \sim \text{IG}\left(a_j' + \frac{rank(\boldsymbol{L}_j)}{2},\, b_j' + \frac{1}{2}\sum_{s=d+1}^{S_j} v_{js}^2\right)$$

and v_{js} determined by (8). Thus, updating of σ_j^2 can be done by simple Gibbs steps.

- *Random walk with locally adaptive independent variances*
  If we assume locally independent variances $\boldsymbol{\tau}_j^2 = (\tau_{j1}^2, \ldots, \tau_{jS_j}^2)'$ with $\tau_{js}^2 = \gamma_{js}\tau_j^2$, the sampling schemes facilitate considerably. Updating of the variance parameters is straightforward because the full conditionals for γ_{js}, $s = 1, \ldots, S_j$ are inverse Gamma distributions with

$$\gamma_{js} \,|\, \cdot \sim IG\left(\nu/2 + 1/2, \nu/2 + \frac{u_{js}^2}{2\tau_j^2}\right). \tag{22}$$

  For the full conditional of τ_j^2 we obtain

$$\tau_j^2 \sim IG\left(a_j + \frac{1}{2}rank(\boldsymbol{K}_j), b_j + \frac{1}{2}\sum_{s=d}^{S_j}\frac{u_{js}^2}{\gamma_{js}}\right). \tag{23}$$

Updating linear effects parameters $\boldsymbol{\beta}$

The full conditionals of linear effects parameters $\boldsymbol{\beta}$ are multivariate Gaussian with precision matrix $\boldsymbol{P}_{\boldsymbol{\beta}}$ and mean $\boldsymbol{m}_{\boldsymbol{\beta}}$ given by

$$\boldsymbol{P}_{\boldsymbol{\beta}} = \boldsymbol{W}'\boldsymbol{\Lambda}\boldsymbol{W}, \quad \boldsymbol{m}_{\boldsymbol{\beta}} = \boldsymbol{P}_{\boldsymbol{\beta}}^{-1}\boldsymbol{W}'\boldsymbol{\Lambda}(\boldsymbol{U} - \tilde{\boldsymbol{\eta}}). \tag{24}$$

Thus, updating of $\boldsymbol{\beta}$ is done by Gibbs steps.

We £nally summarize the resulting sampling scheme:

Summary of sampling scheme

1. *For $t = 1, \ldots, T$ update U_t and λ_t:*
   - *Probit models:* Update U_t by drawing from the truncated normal distribution given in (13) and (14). The variances $\lambda_t \equiv 1$ are not stochastic.
   - *Logit models:* Joint update U_t and λ_t by £rst drawing U_t from $p(U_t \,|\, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta}, y_t)$ followed by drawing λ_t from $p(\lambda_t \,|\, U_t, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p, \boldsymbol{\beta})$, see Holmes and Held (2003) for details.
   - *t-distributed errors:* Update U_t by drawing from the truncated normal distribution given in (13) and (14). Update λ_t by drawing from (15).

2. *For $j = 1, \ldots, p$ update $\boldsymbol{f}_j$:*
   Update $\boldsymbol{f}_j$ by drawing from the multivariate Gaussian distribution with mean and precision matrix given in (16).

3. *For $j = 1, \ldots, p$ update variance parameters $\boldsymbol{\tau}_j^2$:*

   - *Global variances:* Update τ_j^2 by drawing from (17) and compute $\boldsymbol{\tau}_j^2 = (\tau_j^2, \ldots, \tau_j^2)'$.
   - *Locally adaptive dependent variances:* Update the blocks $\boldsymbol{h}_{j,[lr]}$ of $\boldsymbol{h}_j$ by MH steps with conditional prior proposals. Draw proposals $\boldsymbol{h}_{j,[lr]}^*$ from the conditional prior (18) with mean and covariance matrix given by (20) and (21) and accept them with probability (19). Compute $\boldsymbol{\tau}_j^2 = (\exp(h_{jd}), \ldots, \exp(h_{jS_j}))'$.
   - *Locally adaptive independent variances:* Update γ_{js} for $s = d, \ldots, S_j$ by drawing from (22). Update τ_j^2 by sampling from (23). Compute $\boldsymbol{\tau}_j^2 = (\gamma_{jd}\tau_j^2, \ldots, \gamma_{jS_j}\tau_j^2)'$.

4. *For j=1,...,p recompute the penalty matrices $\boldsymbol{K}_j$.*

5. *Update $\boldsymbol{\beta}$:*
   Updating is done by sampling from the Gaussian distribution with parameters given in (24).

3 Simulation studies

To illustrate the performance of our locally adaptive approaches we carried out two simulation studies for binomial probit models with different settings for the true regression function. Binomial response vectors $y = (y_1, \ldots, y_T)$ were generated by assuming $\eta_t = f(x_t)$ and drawing $B(n, \Phi(\eta_t))$ distributed random variables y_t, $t = 1, \ldots, T$. To assess the dependence of results on the number of observations we used $n = 1$, $n = 3$ and $n = 6$.
For the true regression function f we considered the two cases depicted in Figures 2 and 6 (dashed lines). In the £rst case ($T = 250$) a discontinuous step function for f and in the second case ($T = 400$) the regression function

$$f(x) = \sqrt{x(1-x)}\sin\left\{\frac{2\pi(1+2^{(9-4j)/5})}{x+2^{(9-4j)/5}}\right\}, \qquad x \in [0,1],\ j = 4,$$

characterized by differing curvature and medium spatial variability taken from Ruppert and Carroll (2000) was used. For each of the two situations we generated 250 replications and applied the three approaches with global variances, locally dependent and independent variances described in Section 2 to each replication. For comparison with standard software, we additionally applied the penalized spline approach by Wood (2000) implemented in the R package MGCV (Wood, 2001) and LOESS (Loader, 1999) implemented in S-plus.
A similar simulation study based on logit models rather than probit models shows virtually identical results. Therefore, and to keep the paper in reasonable length, results for logit models are not presented.
In the following, we present in Subsection 3.1 results for the step function and in Subsection 3.2 results for the function with differing curvature. In some cases, particulary for the second function, the difference between $n = 3$ and $n = 6$ is quite small. If this is the case, the presentation of results is restricted to $n = 1$ and $n = 3$.

3.1 Regression function with discontinuities

Facing a regression function with discontinuities, the best results for all approaches were usually obtained by using a RW1 prior for the regression function f and, in case of the approach with locally dependent variances, a RW1 prior for the variance function h. We therefore restrict the presentation to these cases

and denote them in the following with RW1 (global approach), TRW1 (locally independent variances) and RW1VRW1 (locally dependent variances).
Figures 2 - 4 display for $n = 1$, $n = 3$ and $n = 6$ boxplots of $\ln(MSE) = \ln(1/T \sum_{t=1}^{T}(f(x_t) - \hat{f}_t(x_t))^2)$ and regression function estimates averaged over the 250 replications for the various estimators, see panels (a)-(f). In panels (h)-(l) estimates corresponding to the median £t of TRW1 including pointwise 95% credible (or con£dence) intervals are given. In a Bayesian approach based on MCMC simulations, pointwise credible intervals are simply obtained by computing the respective quantiles of the sampled function evaluations. Panels (a) and (b) of Figure 5 show for $n = 1$ and $n = 3$ variance function estimates, again averaged over the 250 replications. Panels (c)-(f) show the coverage probabilities of pointwise credible (or con£dence) intervals for a nominal level of 95%. From Figures 2 - 5 we can draw the following conclusions:

- The by far best results in terms of bias and MSE are obtained with locally adaptive independent variances (TRW1). Satisfactory results are also obtained with the approach RW1VRW1. As could have been expected, LOESS and MGCV perform worst because these approaches are not designed for estimating functions with discontinuities.

- For $n = 1$, inspection of individual estimates reveals that only one observation per covariate value is in many cases not enough to recover the underlying step function satisfactorily. With $n = 3$ and $n = 6$ observations, the true curve can be recovered satisfactorily. For $n = 6$ and TRW1 the bias almost completely disappears.

- The jumps in the regression function are best re¤ected in the variance functions for TRW1. For the approach with locally dependent variances RW1VRW1 the variance functions show only for $n = 3$ (and $n = 6$) observations a signi£cant increase of variances at all jumps. The increase of variances is, of course, less pronounced for RW1VRW1 as for TRW1.

- Even for $n = 1$, for both approaches with adaptive variances the coverage rates are closer to the nominal level than for the approach with a global variance. The approach TRW1 reveals a dramatic improvement of coverage rates at the jumps compared to RW1 and RW1VRW1. Around the jumps, the coverage rates of LOESS and MGCV are far below the nominal level.

3.2 Regression function with differing curvature

In contrast to a regression function with discontinuities, the best results for all approaches were usually obtained by using a RW2 prior for the regression function f and, in case of the approach with locally dependent variances, a RW1 prior for the variance function h. We therefore restrict the presentation to these cases and denote them in the following with RW2 (global approach), TRW2 (locally independent variances) and RW2VRW1 (locally dependent variances).
Figures 6 and 7 display in panels (a) for $n = 1$ and $n = 3$ boxplots of $\ln(MSE)$, in panels (b)-(f) regression function estimates averaged over the 250 replications, and in panels (h)-(l) estimates corresponding to the median £t of RW2VRW1 including pointwise 95% credible (or con£dence) intervals. Average variance function estimates are depicted in panels (a) and (b) of Figure 8. The coverage rates of pointwise credible intervals are shown in panels (c)-(f). The results displayed in Figures 6 - 8 can be interpreted as follows:

- Not surprisingly, the approach RW2VRW1 with locally dependent variances clearly outperforms the global approach RW2 and the locally independent approach TRW2. Most striking is the severe bias for $n = 1$ obtained with TRW2 at the local minima and maxima of the curve. A possible explanation might be that the approach with local independent variances is too ¤exible in situations where the true probabilities of success are close to one or zero (as is the case at the minima and maxima of the curve).

- As could have been expected, LOESS and MGCV are more competitive for smooth functions with differing curvature than for functions with discontinuities. However, our approach RW2VRW1 still outperforms both standard methods.
- The decrease of spatial variability in the regression function is accurately re¤ected by the variance functions obtained from RW2VRW1 while for TRW2 the variances stay more or less constant at the level of the approach with global variance RW2.
- The coverage rates for RW2, TRW2 and RW2VRW1 are generally close to the nominal level. For TRW2 and $n = 1$, however, the coverage is below the nominal level close to the local minima and maxima of the true curve. The main reason is the strong bias in this area. The coverage rates of the standard methods LOESS and MGCV are sometimes considerably below the nominal level in the moderately oscillating areas of the function.

4 Application to forest health data

In this section we demonstrate the practicability of our methods by an application to forest health data. We analyze the in¤uence of calendar time, age, canopy density CP and location L on the health state of trees ($y = 1$ for a damaged tree and $y = 0$ otherwise). Data have been collected in yearly forest damage inventories carried out in the forest district of Rothenbuch in northern Bavaria from 1983 to 2001. There are 80 observation points with occurrence of beeches spread over an area extending about 15 km from east to west and 10 km from north to south, see Figure 11. A detailed data description can be found in Göttlein and Pruscha (1996).
We used a binary probit model with predictor

$$y_{it} = f_1(t) + f_2(age_{it}) + f_{spat}(L_i) + \beta \cdot CP_i$$

for tree i, $i = 1, \ldots, 80$ and year t, $t = 1983, \ldots, 2001$. Here, age_{it} is the age of the tree in years at the beginning of the observation period, L_i is the location of tree i, and CP_i is the canopy density at the stand in percent (0%,10%,20%,...,100%). Preliminary examination of the data reveal that the effect of canopy density is linear. Therefore CP is included as a usual linear effect with a diffuse prior for the regression coef£cient.
Although this is only a demonstrating example, it is important to consider possible spatial heterogeneity of the data for a realistic modelling approach which captures the most important features of the data. For that reason we included a spatial effect f_{spat}. We assigned a Markov random £eld prior (Besag et al., 1991), with the neighborhood ∂_L of trees including all trees L' with euclidian distance $d(L, L') \leq 1.2$ km, see also Fahrmeir and Lang (2001b). Thus, our model is an example for a regression model with *geoadditive predictor* (Kammann and Wand, 2003) and demonstrates one of the main advantages of Bayesian inference for semiparametric regression based on MCMC simulation: models can be easily extended to more complex formulations.
As a starting point we used random walk priors with global variances. We tested all four combinations of £rst and second order random walks for f_1 and f_2. In terms of the DIC (Spiegelhalter et al., 2002), the combinations RW1,RW1 and RW1,RW2 performed best, see Table 1. Figure 9 (a) and (b) depicts estimated functions f_1 and f_2 for the combination RW1,RW1 and Figure 10 (a) and (b) for the combination RW1,RW2. For illustration, the estimated spatial effect for the RW1,RW1 combination is shown in Figure 11. We see that trees recover after the bad years around 1986, but after 1992 health status declines to a lower level again. As we might have expected, younger trees are in healthier state than the older ones. Note also, that the incorporation of the spatial effect into the model is quite important since the estimated effect suggests considerable spatial heterogeneity.
Starting from the four models with global variances, experiments with our spatially adaptive random walk priors gave evidence for a jump of the age effect around age 20 and hints for a smoothly varying variance of

Year	Age	$D(\theta)$	pD	DIC
RW1	RW1	845.45	79.20	1003.85
RW1	RW2	864.76	69.78	1004.32
RW2	RW1	854.65	77.70	1010.05
RW2	RW2	873.74	67.23	1008.20
RW1VRW1	TRW1	847.64	77.64	1002.92
RW1VRW1	TRW2	845.43	75.68	996.79
RW2VRW1	TRW1	855.89	76.43	1008.75
RW2VRW1	TRW2	851.69	75.30	1002.29

Table 1 Forest health example. Comparison of deviance evaluated at posterior mean estimate $D(\bar{\theta})$, effective number of model parameters pD and deviance information criterion DIC.

the time trend. We therefore replaced the global variance of the age effect by locally independent variances (9) and the global variance of the time trend by locally dependent variances (8) with a £rst order random walk for h_s. In terms of the DIC, all models with locally adaptive variances (sometimes clearly) outperform the results obtained with global variances, see Table 1. The best result is obtained with the combination RW1VRW1,TRW2. For comparison with the global variances Figure 9 (c) and (d) shows results for the combination RW1VRW1,TRW1 and Figure 10 (c) and (d) for the combination RW1VRW1,TRW2. The respective panels (e) and (f) display the estimated locally varying variance functions. Results for the spatial effect remain almost unchanged compared to our basis models and are therefore not replicated. As could have been expected, the estimated jump for the age effect is steeper with a £rst order random walk rather than a second order random walk for f_2.

5 Conclusions

This paper presents a practical approach for £tting highly oscillating or unsmooth functions in binary regression models. The simulation study in Section 3 suggests that for highly oscillating functions the approach with locally dependent variances performs superior to locally independent variances and simple random walk models with a global variance. For jump functions results are superior with locally independent variances. It has also been shown that standard methods like LOESS or MGCV are not capable of handling features like jumps or differing curvature appropriately.
We see the following directions for future research:

- *Other response distributions*
  Our approach can be extended to models with multicategorical responses by using similar latent utility representations as for binary responses, see Fahrmeir and Lang (2001b) and Holmes and Held (2003). For general responses from an exponential family sampling schemes based on latent utilities are no longer available. A possible approach could be based on variants of iteratively weighted least squares proposals for the nonlinear functions f_j as proposed for generalized linear models by Gamerman (1997) and for semiparametric regression by Brezger and Lang (2003).

- *Model choice*
  Another aspect for future research concerns model choice. The introduction of locally adaptive function estimates considerably complicates model choice, because one has to decide not only whether a covariate should be included into the model or not, but also how the covariate effect should be modelled. In our application we used the DIC as a goodness of £t measure. The drawback of model choice via the DIC is that only a limited number of models can be tested. For the future, we plan to develop Bayesian inference techniques that allow estimation and model choice (to some extent) simultaneously.

Acknowledgements This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 “Statistische Analyse diskreter Strukturen”. We are grateful to two anonymous referees for their many valuable suggestions to improve the £rst version of the paper.

References

Albert, J. and Chib, S., 1993: Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669-679.

Andrews, D.F. and Mallows, C.L., 1974: Scale mixtures of normal distributions. *Journal of the Royal Statistical Society B*, 36, 99-102.

Besag, J., York, J. and Mollie, A., 1991: Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Brezger, A. and Lang, S., 2003: Generalized additive regression based on Bayesian P-splines. SFB 386 Discussion paper 321, Department of Statistics, University of Munich.

Devroye, L., 1986: *Non-uniform random variate generation.* Springer-Verlag, New York.

Donoho, D. L. and Johnstone, I. M., 1994: Ideal spatial adaption by wavelet shrinkage. *Biometrika*, 81, 425-455.

Eilers, P.H.C. and Marx, B.D., 1996: Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statistical Science*, 11, 89-121.

Fahrmeir, L. and Lang, S., 2001a: Bayesian inference for generalized additive mixed models based on Markov random £eld priors. *Journal of the Royal Statistical Society C*, 50, 201-220.

Fahrmeir, L. and Lang, S., 2001b: Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the Institute of Statistical Mathematics*, 53, 10-30

Fahrmeir, L. and Tutz, G., 2001: *Multivariate statistical modelling based on generalized linear models*, Springer–Verlag, New York.

Fan, J. and Gijbels, I., 1995: Data-driven bandwidth selection in local polynomial £tting: variable bandwidth and spatial adaption. *Journal of the Royal Statistical Society B*, 57, 371-394.

Gamerman, D., 1997: Ef£cient sampling from the posterior distribution in generalized linear models. *Statistics and Computing*, 7, 57–68.

Göttlein, A. and Pruscha, H., 1996: Der Ein¤uss von Bestandskenngrössen, Topographie, Standort und Witterung auf die Entwicklung des Kronenzustandes im Bereich des Forstamtes Rothenbuch, *Forstwissenschaftliches Centralblatt*, 114, 146-162.

Holmes, C., and Held, L., 2003: On the simulation of Bayesian binary and polychotomous regression models using auxiliary variables. Technical report. Available at: `http://www.stat.uni-muenchen.de/~leo`

Kamman, E. E. and Wand, M. P., 2003: Geoadditive models. *Journal of the Royal Statistical Society C*, 52, 1-18.

Knorr-Held, L., 1999: Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics*, 26, 129-144.

Lang, S. and Brezger, A., 2004: Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13, 183-212.

Lang, S., Fronk, E.-M. and Fahrmeir, L., 2002: Function estimation with locally adaptive dynamic models. *Computational Statistics*, 17, 479-500.

Loader, C., 1999: *Local Regression and Likelihood.* Springer Verlag, New York.

Ruppert, D. and Carroll, R. J., 2000: Spatially adaptive penalties for spline £tting. *Australian and New Zealand Journal of Statistics*, 42, 205-223.

Robert, C.P., 1995: Simulation of truncated normal variables. *Statistics and Computing*, 5, 121-125.

Speckmann, P.L. and Sun, D.C., 2003: Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90, 289-302.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A., 2002: Bayesian measures of model complexity and £t. *Journal of the Royal Statistical Society B*, 65, 583 - 639.

Wood, S.N., 2000: Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. *Journal of the Royal Statistical Society B*, 62, 413-428.

Wood, S.N., 2001: mgcv: GAMs and Generalized Ridge Regression for R. *R News*, 1, 20-25.

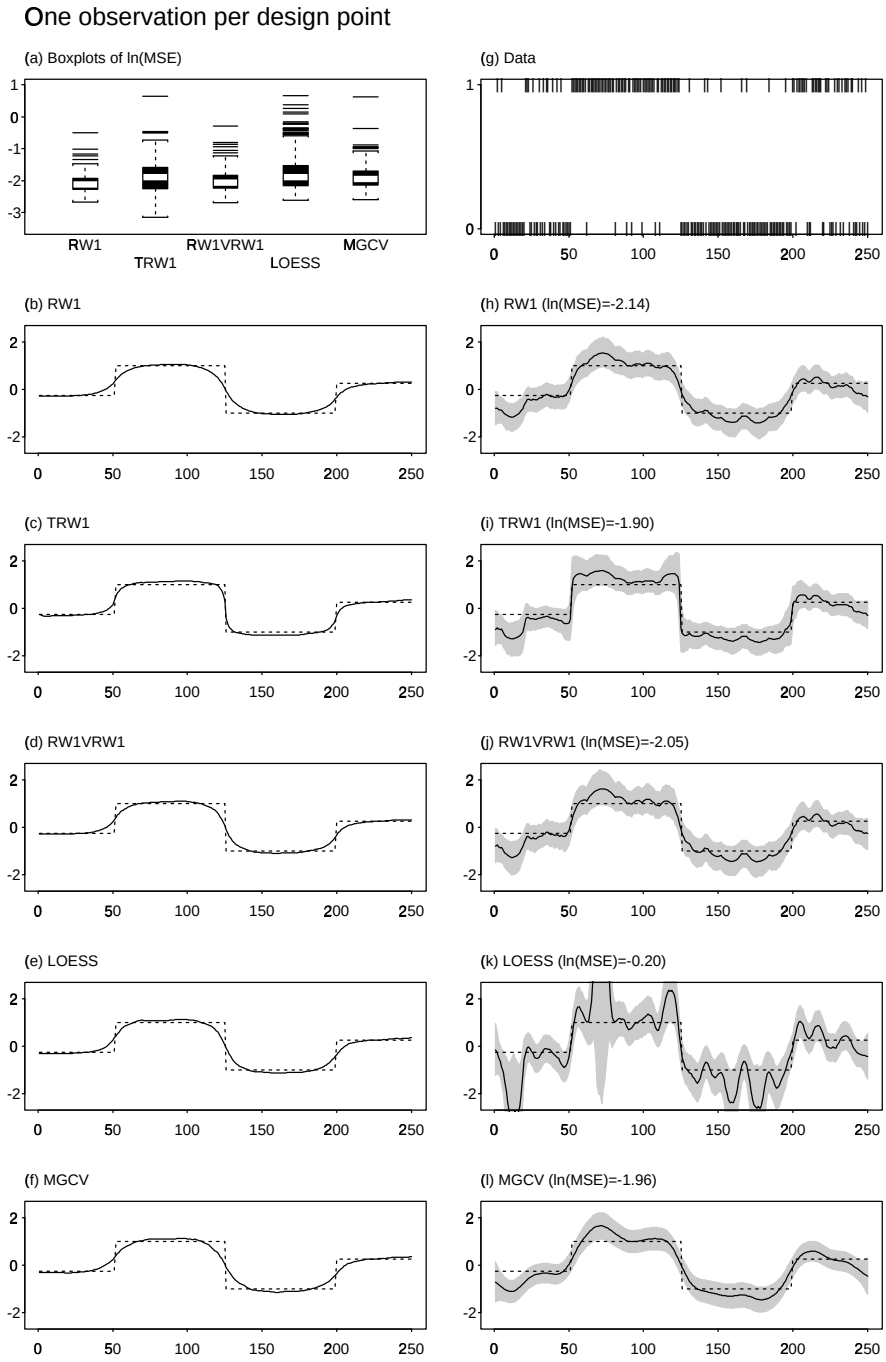

Fig. 2 Simulation results for regression function with discontinuities and one observation per design point. (a) Boxplots of ln(MSE). (b)-(f) Averaged posterior mean/ML estimates (—) and true function (· · ·). (g) Response corresponding to median £t of TRW1. (h)-(l) Posterior mean/ML estimates (—), true function (· · ·) and pointwise 80 % credible/con£dence intervals.

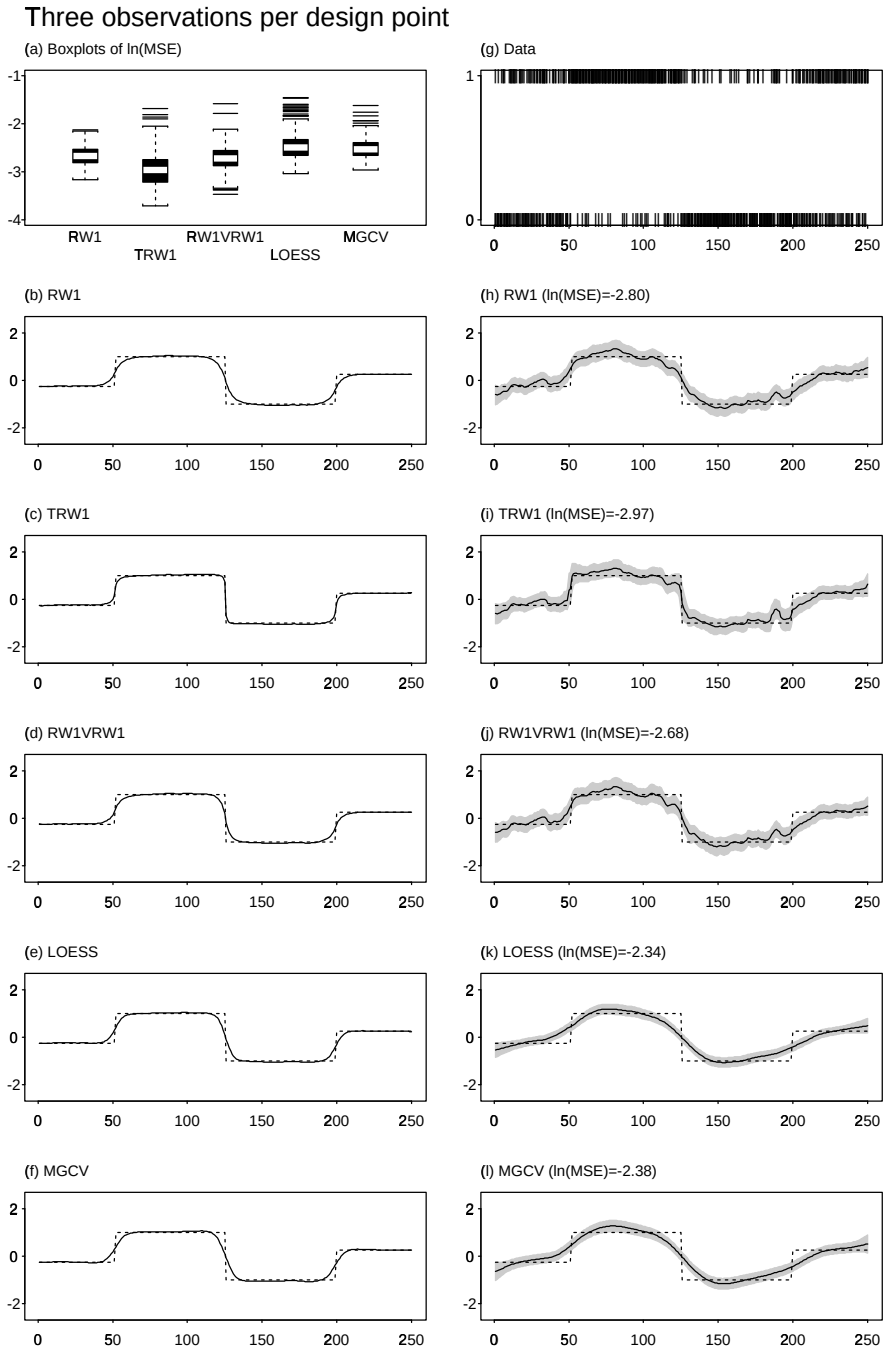

Fig. 3 Simulation results for regression function with discontinuities and three observations per design point. (a) Boxplots of ln(MSE). (b)-(f) Averaged posterior mean/ML estimates (—) and true function (· · ·). (g) Response corresponding to median £t of TRW1. (h)-(l) Posterior mean/ML estimates (—), true function (· · ·) and pointwise 80 % credible/con£dence intervals.

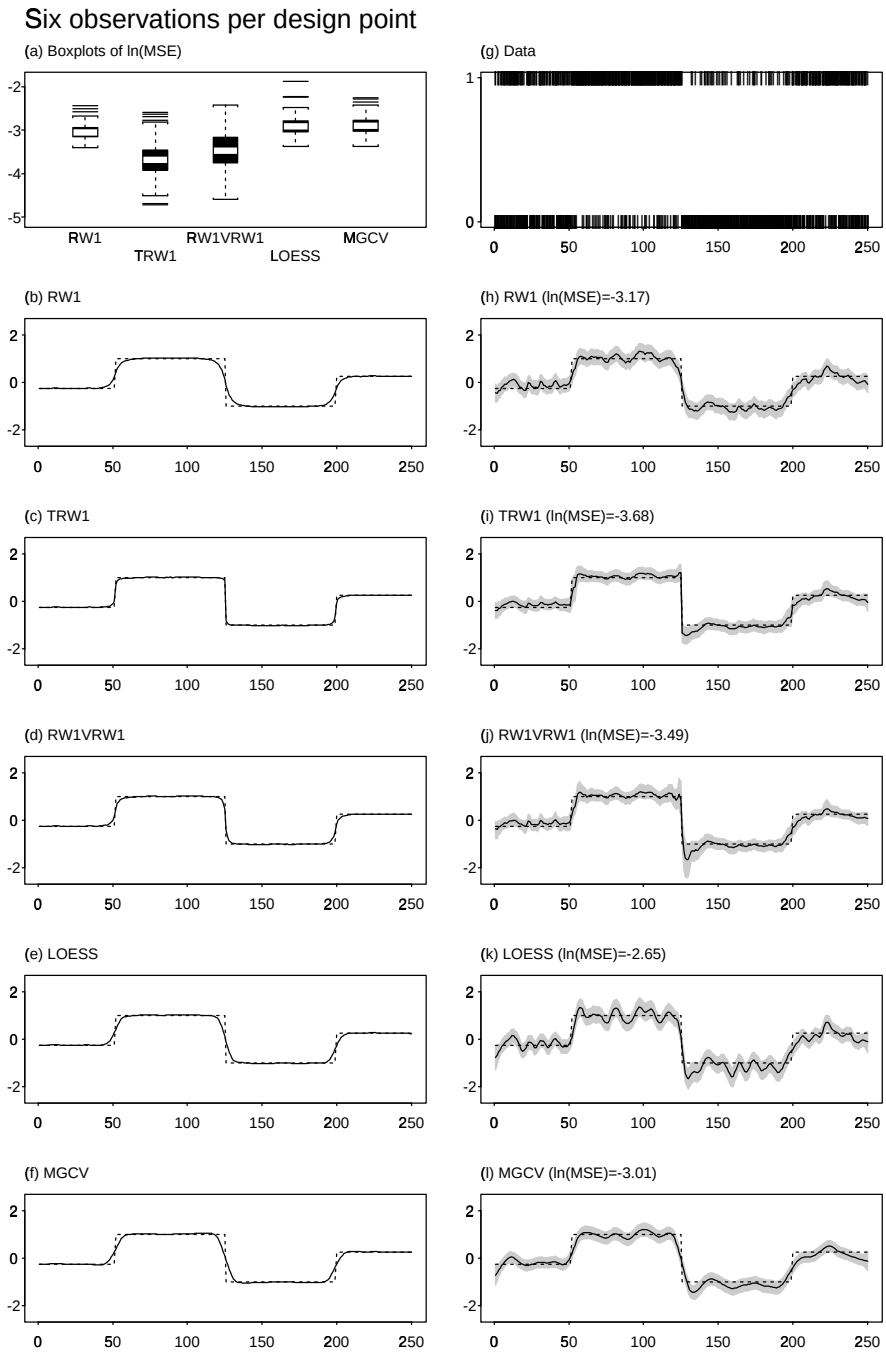

Fig. 4 Simulation results for regression function with discontinuities and six observations per design point. (a) Boxplots of ln(MSE). (b)-(f) Averaged posterior mean/ML estimates (—) and true function (· · ·). (g) Response corresponding to median £t of TRW1. (h)-(l) Posterior mean/ML estimates (—), true function (· · ·) and pointwise 80 % credible/con£dence intervals.

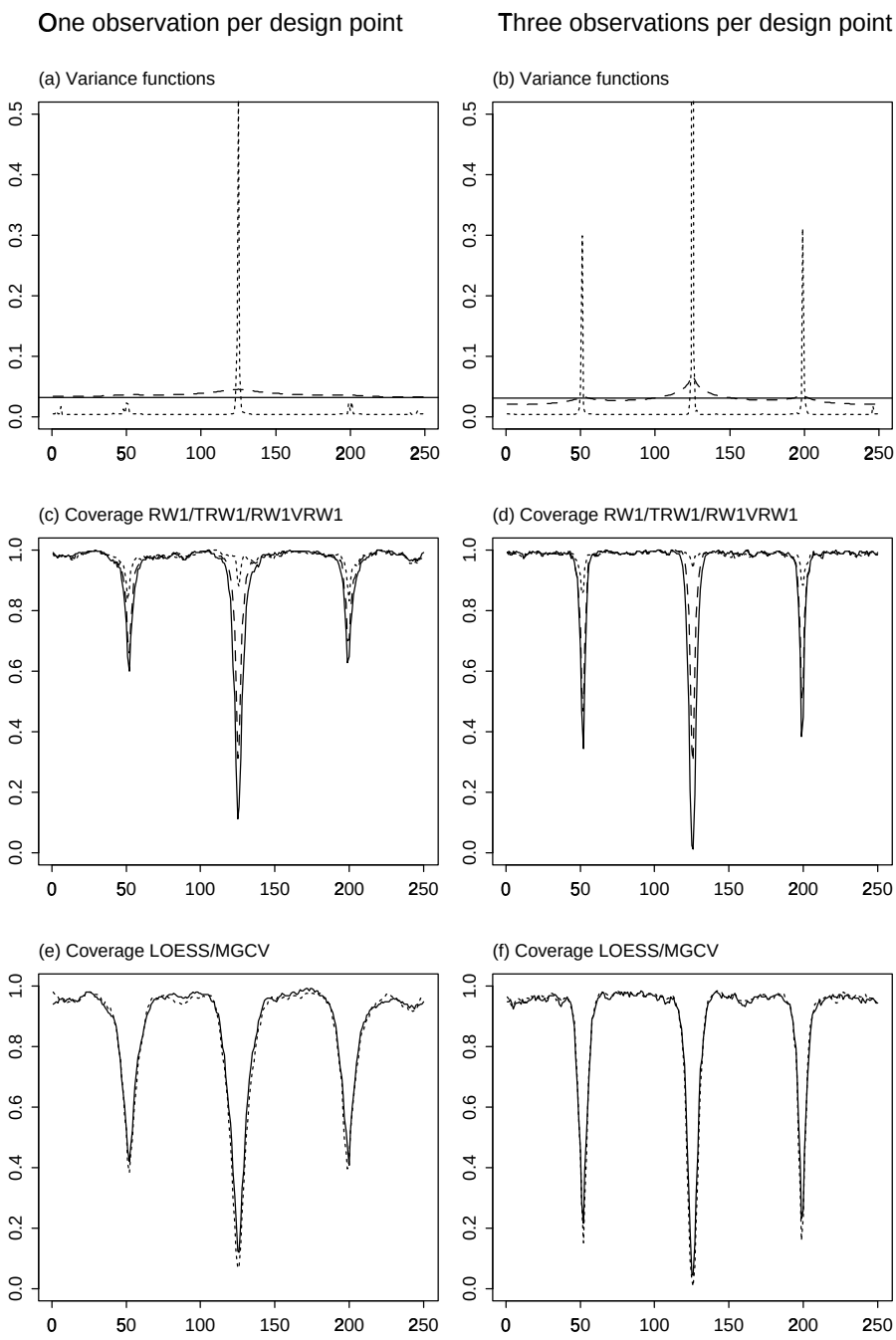

Fig. 5 Simulation results for regression function with discontinuities and one/three observations per design point. (a)-(b) Averaged posterior median estimates for variance functions given global variance (—), locally dependent variances (- -) and locally independent variances (· · ·). (c)-(d) Coverage of pointwise 95 % credible intervals given global variance (—), locally dependent variances (- -) and locally independent variances (· · ·). (e)-(f) Coverage of pointwise 95 % con£dence intervals given LOESS (—) and MGCV (· · ·).

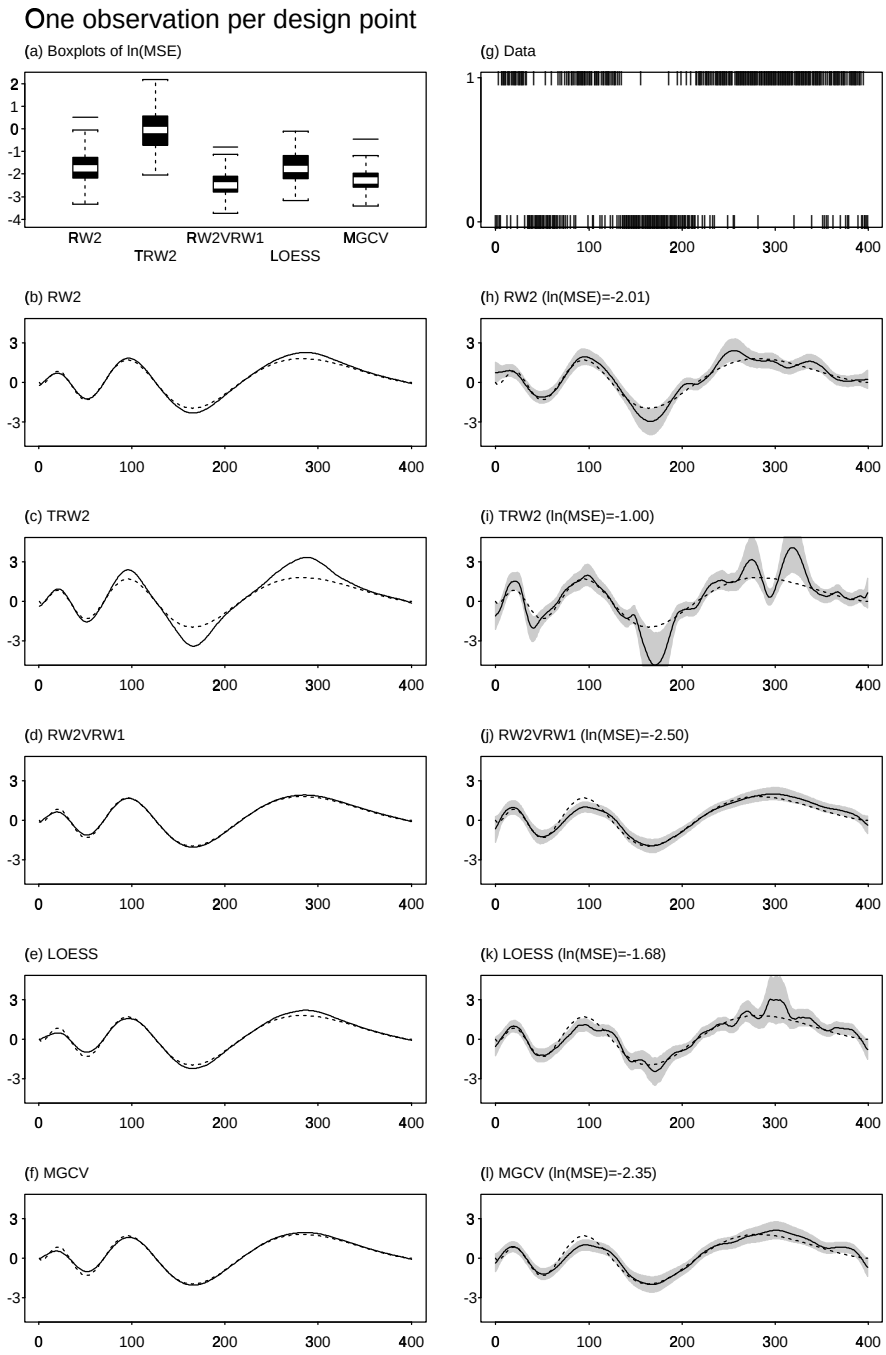

Fig. 6 Simulation results for regression function with differing curvature and one observation per design point. (a) Boxplots of ln(MSE). (b)-(f) Averaged posterior mean/ML estimates (—) and true function (· · ·). (g) Response corresponding to median £t of RW2VRW1. (h)-(l) Posterior mean/ML estimates (—), true function (···) and pointwise 80 % credible/con£dence intervals.

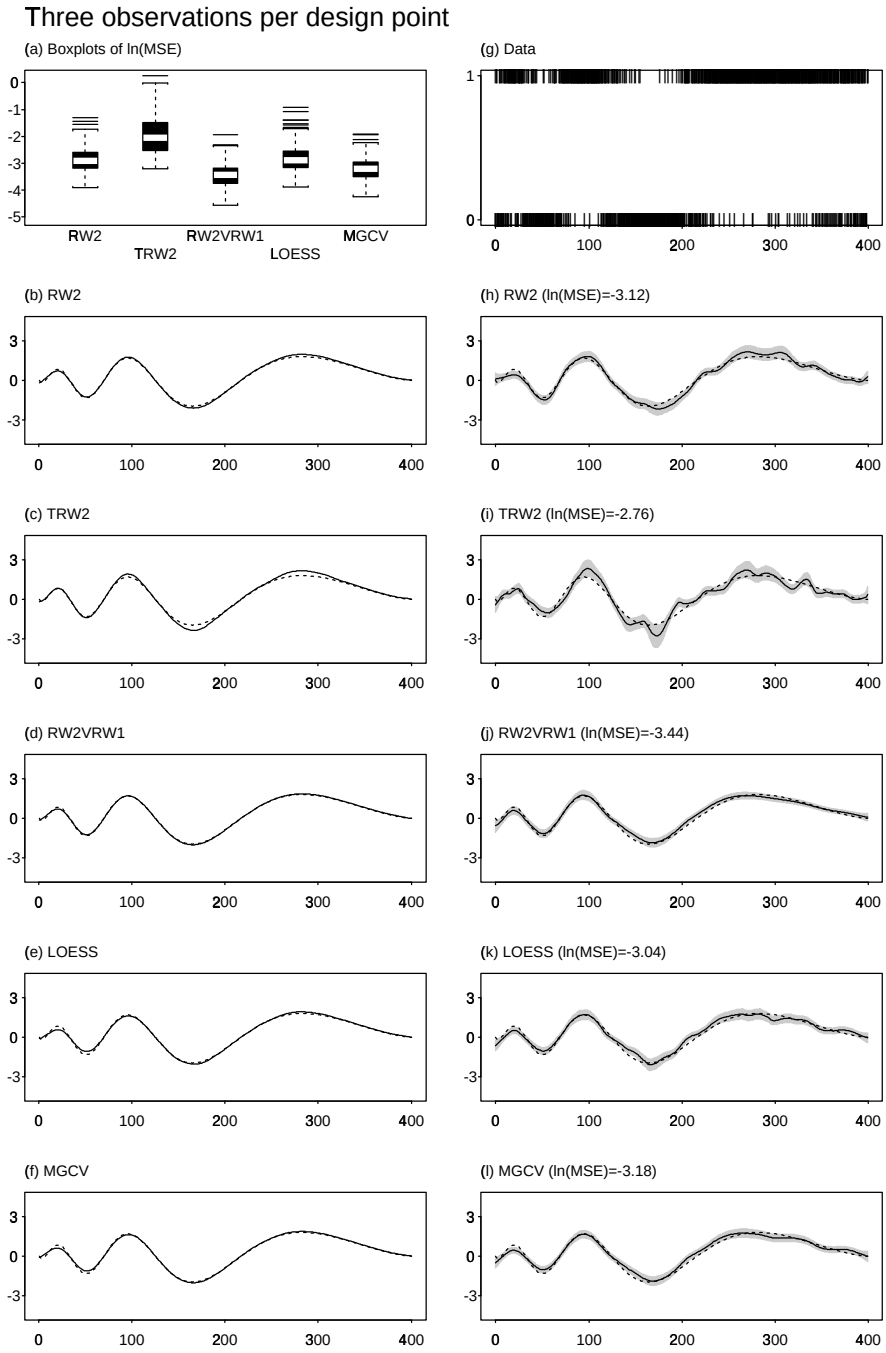

Fig. 7 Simulation results for regression function with differing curvature and three observations per design point. (a) Boxplots of ln(MSE). (b)-(f) Averaged posterior mean/ML estimates (—) and true function (· · ·). (g) Response corresponding to median £t of RW2VRW1. (h)-(l) Posterior mean/ML estimates (—), true function (···) and pointwise 80 % credible/con£dence intervals.

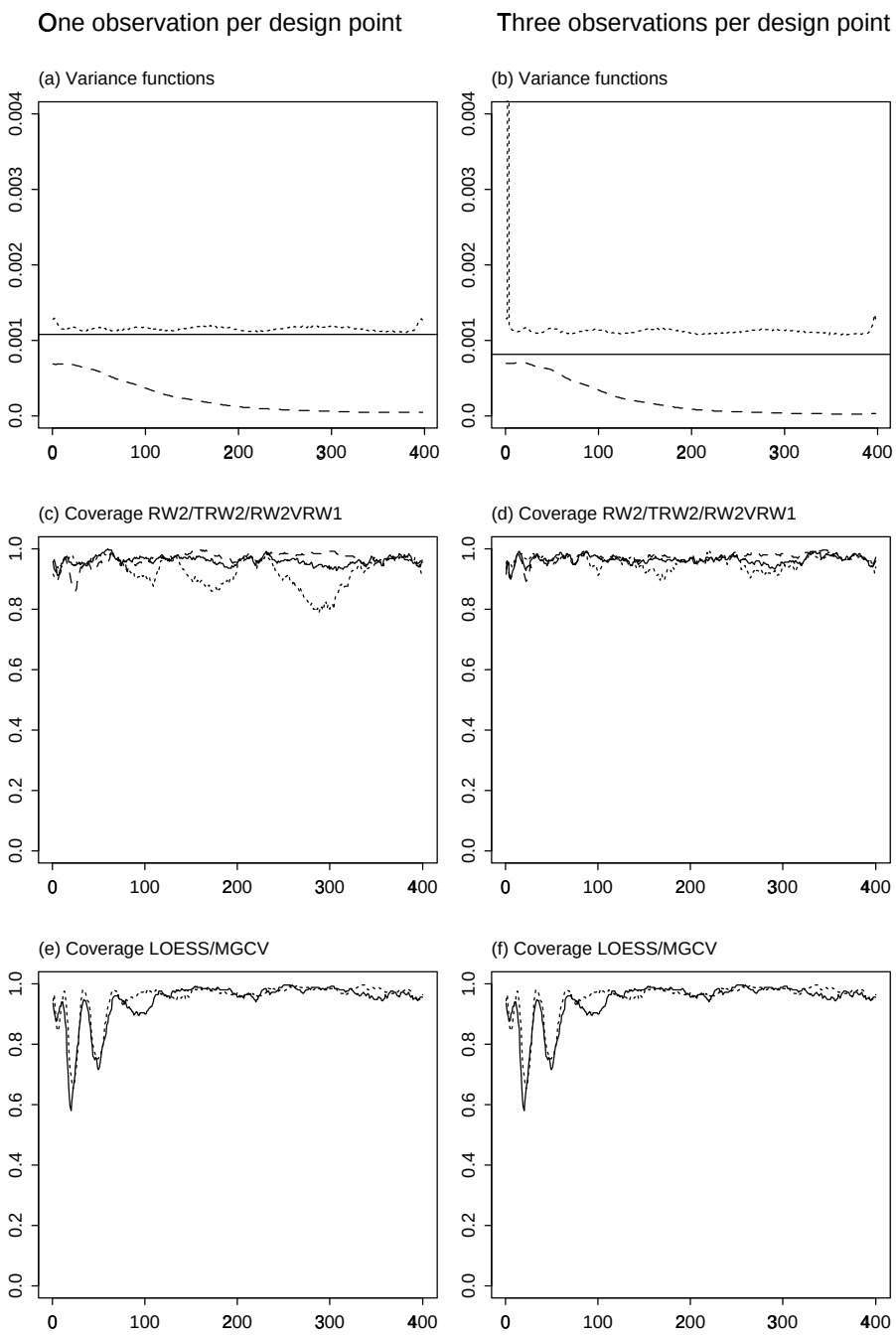

Fig. 8 Simulation results for regression function with differing curvature and one/three observations per design point. (a)-(b) Averaged posterior median estimates for variance functions given global variance (—), locally dependent variances (- -) and locally independent variances (· · ·). (c)-(d) Coverage of pointwise 95 % credible intervals given global variance (—), locally dependent variances (- -) and locally independent variances (· · ·). (e)-(f) Coverage of pointwise 95 % con£dence intervals given LOESS (—) and MGCV (· · ·).

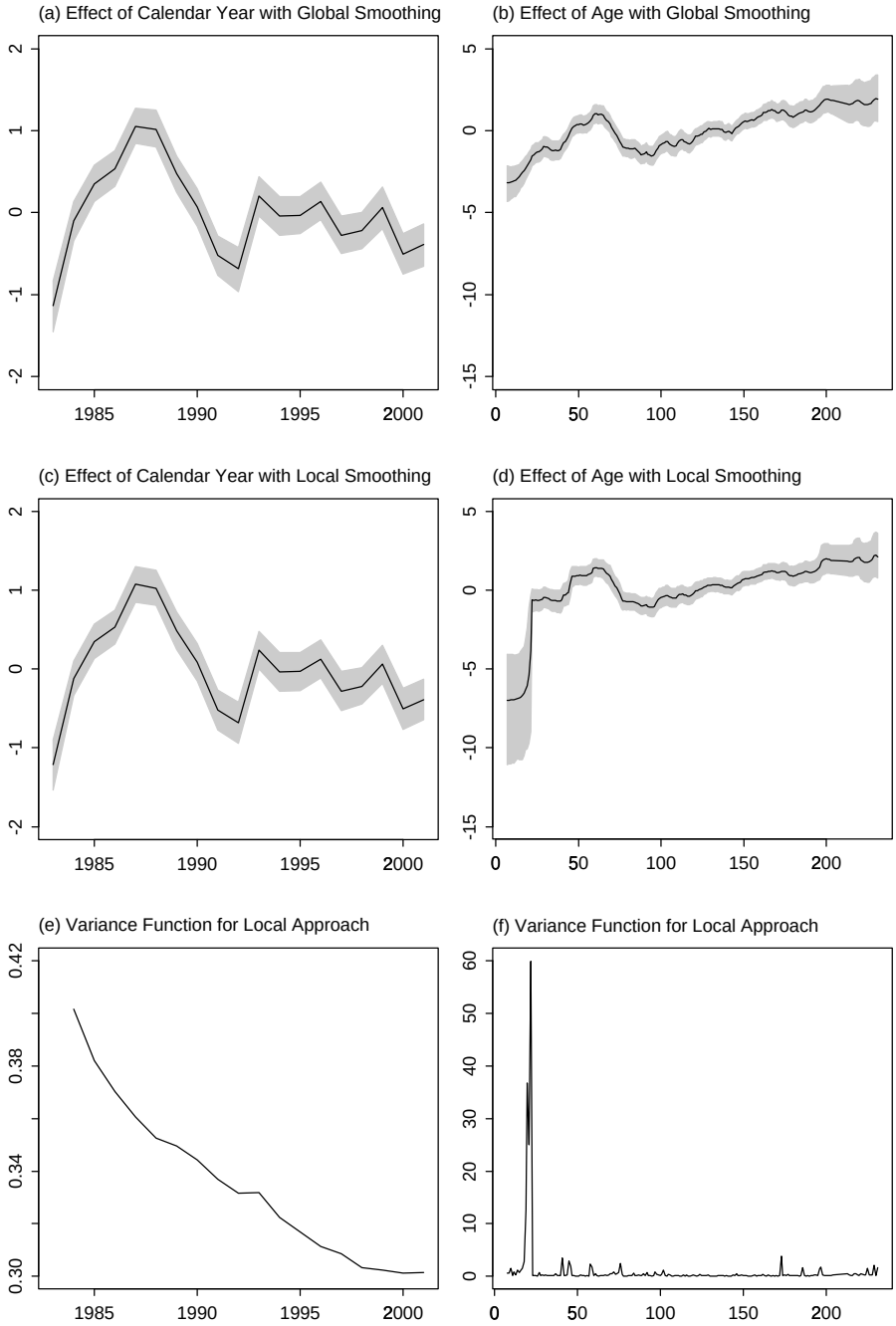

Fig. 9 Forest health example. (a)-(b) Posterior mean estimates (—) and 80 % credible intervals (· · ·) in global approach with RW1 for both calendar year and age effect. (c)-(d) Posterior mean estimates (—) and 80 % credible intervals (· · ·) in local approach with RW1VRW1 for calendar year and TRW1 for age effect. (e)-(f) Posterior median estimates for variance functions in local approach.

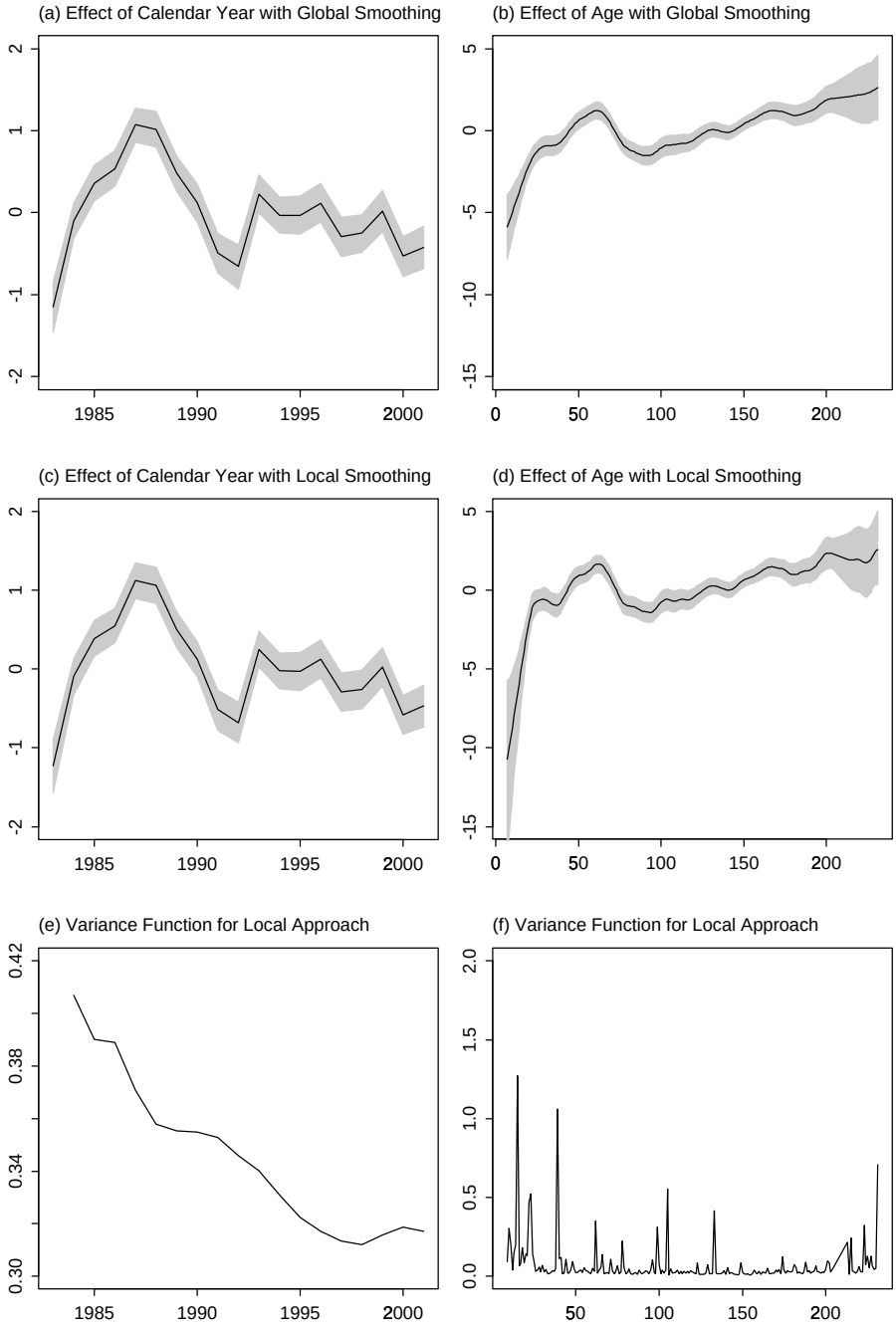

Fig. 10 Forest health example. (a)-(b) Posterior mean estimates (—) and 80 % credible intervals (· · ·) in global approach with RW1 for calendar year and RW2 for age effect. (c)-(d) Posterior mean estimates (—) and 80 % credible intervals (· · ·) in local approach with RW1VRW1 for calendar year and TRW2 for age effect. (e)-(f) Posterior median estimates for variance functions in local approach.

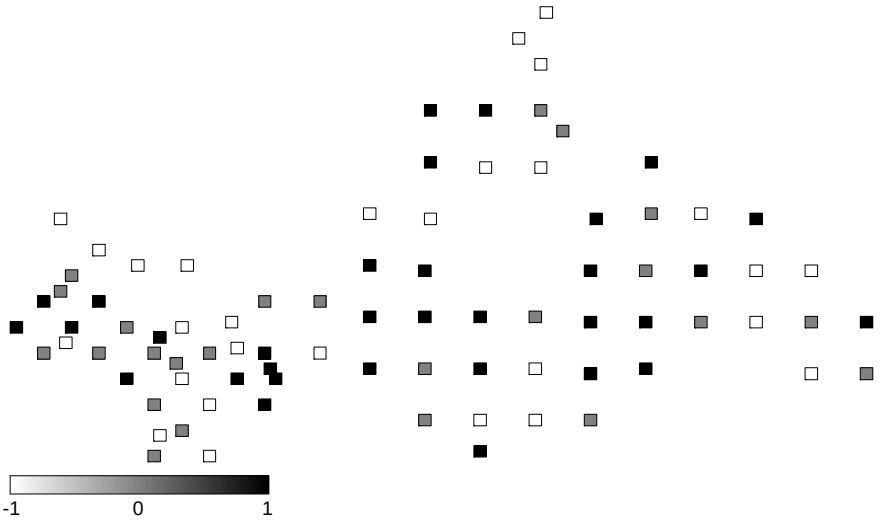

Fig. 11 Forest health example. Posterior probabilities (based on a nominal level of 80%) for the spatial effect of the model with £rst order random walks and global variances for f_1 and f_2. Black spots indicate a positive, white spots a negative and grey spots a non-signi£cant effect.