Stat Comput (2007) 17:179–192
DOI 10.1007/s11222-006-9011-x

Feature significance in generalized additive models

B. Ganguli · M. P. Wand

Received: August 2002 / Accepted: November 2006 / Published online: 30 January 2007

Abstract This paper develops inference for the significance of features such as peaks and valleys observed in additive modeling through an extension of the SiZer-type methodology of Chaudhuri and Marron (1999) and Godtliebsen et al. (2002, 2004) to the case where the outcome is discrete. We consider the problem of determining the significance of features such as peaks or valleys in observed covariate effects both for the case of additive modeling where the main predictor of interest is univariate as well as the problem of studying the significance of features such as peaks, inclines, ridges and valleys when the main predictor of interest is geographical location. We work with low rank radial spline smoothers to allow to the handling of sparse designs and large sample sizes. Reducing the problem to a Generalised Linear Mixed Model (GLMM) framework enables derivation of simulation-based critical value approximations and guards against the problem of multiple inferences over a range of predictor values. Such a reduction also allows for easy adjustment for confounders including those which have an unknown or complex effect on the outcome. A simulation study indicates that our method has satisfactory power. Finally, we illustrate our methodology on several data sets.

Keywords Additive models · Best linear unbiased prediction (BLUP) · Bivariate smoothing · Generalised linear mixed models · Geostatistics · Low-rank mixed models · Penalised splines · Penalised quasi-likelihood (PQL)

B. Ganguli (✉)
Department of Statistics, University of Calcutta, 35 Ballygunge Circular Road, Calcutta 700019, India

M. P. Wand
Department of Statistics, School of Mathematics and Statistics, University of New South Wales, Sydney 2052, Australia

1 Introduction

Modeling geographical variation of an outcome can be of interest in many areas of application. For example, increasing use of Geographical Information Systems (GIS) worldwide has lead to a proliferation of spatial data for understanding environmental processes such as global warming. Similarly, in the field of disease mapping, geographical analysis of a health outcome can be useful for prioritising the distribution of preventive services. The methodology developed by Ganguli and Wand (2004) allows for assessment and graphical display of statistically significant features such as peaks, inclines, ridges and valleys on a map of geostatistical data. We extend this method to assessing feature significance in case of non-Gaussian response variables such as binary responses and counts. This includes studying the significance of peaks and valleys in the log-odds or log rate functions when the predictor of interest is univariate as well as the earlier case of determining the significance of peaks, valleys, inclines, ridges etc. when the main predictor of interest is geographical location.

A motivating example is provided by records of mortality rates from 1980–89 from Milan (Zanobetti et al., 2000). Figure 1(a) shows the log-mortality rate plotted as a function of temperature with ±2 pointwise standard errors. The plot was obtained by fitting a Poisson smoothing spline and gives some suggestion of a non-linear effect. Figure 1(b) is the same set of curves, but plotted on the scale of the data. A natural question is whether this non-linearity persists for different amounts of smoothing and after adjustment for other covariates such as pollution and humidity levels.

A second motivating data set is one of lung cancer incidence in Upper Cape Cod, Massachusetts, USA. The data are maintained by the Massachusetts Department of Public Health and summary reports for the period 1982–1992

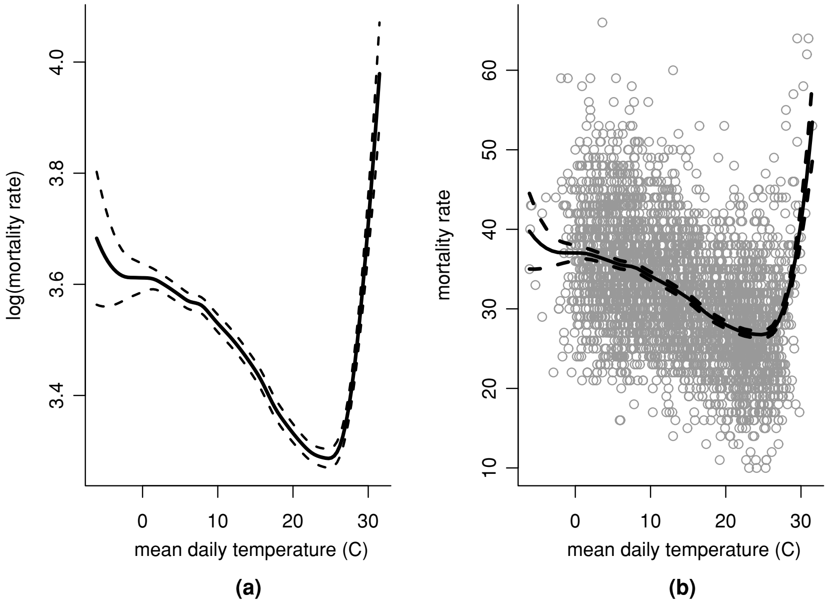


Fig. 1 (a) Estimated log-mortality rate as a function of temperature with ±2 pointwise standard errors for the Milan mortality data. (b) Same curves, but for mortality rate. The temperature/mortality data are also shown

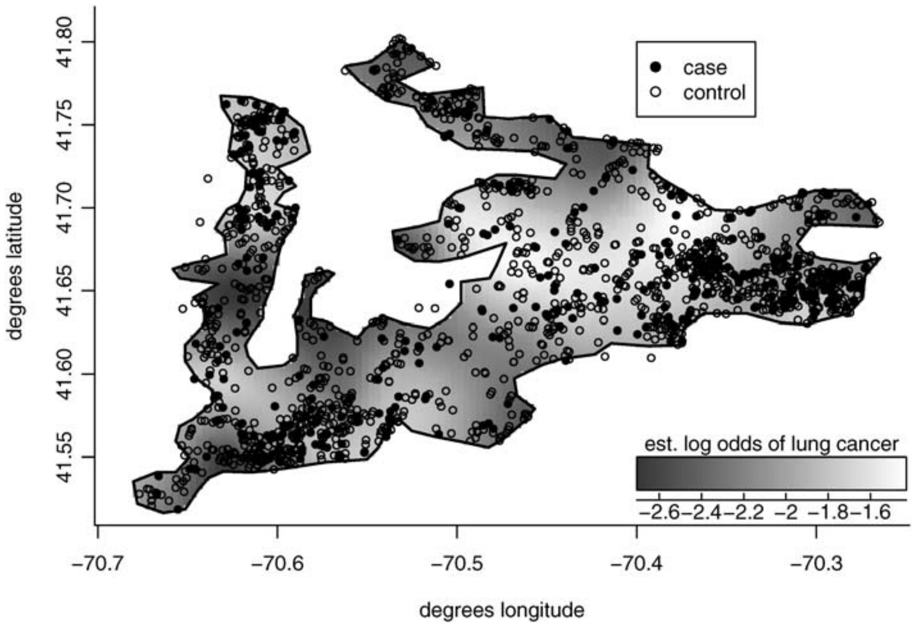


Fig. 2 Map of estimated log-odds of lung cancer in Upper Cape Cod, Massachusetts, USA. The map incorporates adjustment for age and smoking status

showed elevated standardised incidence ratios in several towns of Upper Cape Cod. Figure 2 shows a scatter plot of observed locations superimposed on an map of smoothed incidence rates which suggests possible lung cancer 'hot spots'. The map also accounts for the effects of smoking status and age, considered to be potential confounders.

Ganguli and Wand (2004) use thin plate spline smoothers and a mixed model approximation to make inferences about the statistical significance of observed features such as peaks, valleys, ridges and inclines on a map of geographically referenced measurements of a continuous nature. In this paper, we extend their methodology to assess feature significance in generalised additive modeling. Here, the response is discrete e.g. binary or counts and covariates could include univariate confounders or bivariate predictors such as geographical location. Using a generalised linear mixed model (GLMM) framework leads to a 'seamless' approach which allows direct adjustment for binary predictors but also allows us to avoid making unduly restrictive assumptions about the effect of the continuous univariate or the bivariate predictors. The problem of interest is then to assess the statistical significance of 'features' such as observed peaks and valleys in the adjusted log odds ratio or log rate ratio for a particular covariate of interest before and after adjusting for known confounders which may account for such features.

As in Ganguli and Wand (2004), we adopt the 'Scale-Space' approach for assessing significance of such features. 'Scale-Space' is a concept common in computer vision and has been extended to non-parametric curve or surface estimation by Chaudhuri and Marron (1999). For such an approach, the problem of interest is no longer on trying to perform an optimal amount of smoothing so as to obtain a single 'best' estimate of the true underlying relationship. The problem shifts to that of estimating a 'Scale Space' relationship i.e., the different kinds of information that are available by smoothing the data to different extents. Note that the smoothing parameter serves as the natural measure of a 'scale'. The statistical problem then becomes one of performing inference at different levels of smoothing and combining the results thus obtained. Ganguli and Wand (2004) interpret a 'scale space' analysis of data on water pressure measured at several locations in north-western Texas (Cressie, 1989). High levels of smoothing correspond to viewing the surface at a distance and reveal the broad trend in water pressure levels from South West to North East while progressively lower amounts of smoothing correspond to zooming in on the surface and making inferences about finer nuances such as the presence of shelves. Other examples of 'Scale Space' analysis have been discussed by Chaudhuri and Marron (1999) and Godtliebsen et al. (2002, 2004). The advantage of this approach is that it circumvents the need to adjust for estimation bias resulting from an incorrect choice of the optimal amount of smoothing (Chaudhuri and Marron, 2000). Our approach shares the following advantages with that of Kammann and Wand (2003) and Ganguli and Wand (2004):

(1) seamless; due to using a mixed model representation of both kriging and additive models;
(2) model-based and likelihood-driven; our geoadditive model reduces to a generalised linear mixed model and lends itself to estimation of all parameters together with standard errors using maximum likelihood. This includes a likelihood-based choice of the degrees of freedom for each non-linear component. The utility of this for a 'Scale-Space' approach is that this choice serves to provide a benchmark for an informative range of choices of degrees of freedom for the predictor of interest;
(3) low-rank, as defined by Hastie (1996); meaning that the number of basis functions used to construct the function estimates does not grow with the sample size; This leads to a considerable reduction in computational burden since we no longer have to invert a $n \times n$ matrix, n being the sample size which can be prohibitively large especially for disease mapping applications;
(4) permits asymptotically valid inference about feature significance since the number of parameters does not grow with the sample size;
(5) implementable using standard software such as the `glimmix` macro in `SAS` and the `glmmPQL()` function in `R`.

The layout of this paper is as follows: Section 2 outlines the methodology for the general case where it is necessary to model the effect of unknown covariates on a discrete outcome in a flexible but additive manner. Section 3 extends this to incorporate geographical covariates such as latitude and longitude whereupon it becomes necessary to consider the problem of flexible *bivariate* modeling. Section 4 describes how our model formulation makes it is possible to use theory for generalised linear mixed models to test for feature significance. Section 5 investigates the performance of the methodology under various simulation settings. Finally Section 6 presents several applications of the methodology to motivating datasets.

2 Additive model extension

In this section, we shall first demonstrate how to reduce the problem of low rank generalised additive modeling to a generalised linear mixed model formulation. For the sake of simplicity in notation, we consider the case of two predictors, s and t. Our data are then of the form (y_i, s_i, t_1), $1 \le i \le n$, where the $y_i's$ are scalar outcomes. To avoid the restrictiveness of a parametric model for covariate effects, we assume a generalised additive model of the form:

$$\eta_i = g\{E(y_i|s_i, t_i)\} = f_s(s_i) + f_t(t_i), \quad 1 \le i \le n. \tag{1}$$

where $f_s(\cdot)$ and $f_t(\cdot)$ are unspecified but smooth univariate functions and $g(\cdot)$ is a suitably chosen link-function. For example, for binary data $g(\cdot)$ typically corresponds to the logit link and is given by $g(x) = \log(\frac{x}{1-x})$, $0 < x < 1$.

We shall approximate $f_s(s_i)$ and $f_t(t_i)$ by

$$\eta_i = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i$$

where $\mathbf{X} = [\mathbf{1} \; s_i \; t_i]_{1 \le i \le n}$ and $\mathbf{Z} = [\mathbf{Z}_s, \mathbf{Z}_t]$ with $\boldsymbol{\beta}^T = [\beta_0, \boldsymbol{\beta}_s^T \, \boldsymbol{\beta}_t^T]$ and $\mathbf{u}^T = [\mathbf{u}_s^T \mathbf{u}_t^T]$. The columns of $\mathbf{Z}_s$ and $\mathbf{Z}_t$ are basis functions which permit handling of non-linear structure in $f_s(\cdot)$ and $f_t(\cdot)$. The coefficients $\boldsymbol{\beta}$ and $\mathbf{u}$ are chosen by maximising the likelihood corresponding to (2) subject to minimising quadratic penalties of the form $\mathbf{u}_s^T \mathbf{G}_s{}^{-1} \mathbf{u}_s$ and $\mathbf{u}_t^T \mathbf{G}_t{}^{-1} \mathbf{u}_t$. It can be shown that this is equivalent to penalised quasi likelihood (PQL) estimation for the generalised linear *mixed* model

$$f(y|\mathbf{u}) = \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) + \mathbf{1}^T c(\mathbf{y})\}$$

where the coefficients $\mathbf{u}$ are assumed to be random effects generated from

$$f(\mathbf{u}) = (2\pi)^{-\frac{q}{2}} |\mathbf{G}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{G}^{-1} \mathbf{u}\right)$$

with

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_t \end{bmatrix}. \tag{2}$$

PQL has been proposed as a tool for inference and estimation in GLMMs and involves maximisation of the marginal likelihood obtained by integrating out the random effects based on a Laplace approximation of the resulting integral. We elaborate on this in Section 4. Detailed discussions of the algorithm have been presented by Breslow and Clayton (1993) and Wolfinger and O'Connell (1993). Writing $\mathbf{X} = [\mathbf{1}, \mathbf{x}_s, \mathbf{x}_t]$, we shall choose

$$\mathbf{Z}_s = \left[\left|\mathbf{x}_s - {\kappa_k}^s\right|^{(2m-1)}\right]_{1\le i\le n,\, 1\le k\le \kappa_s}$$

and

$$\mathbf{Z}_t = \left[\left|\mathbf{x}_t - {\kappa_k}^t\right|^{(2m-1)}\right]_{1\le i\le n,\, 1\le k\le \kappa_t} \tag{3}$$

where ${\kappa_1}^s, {\kappa_2}^s, \ldots, {\kappa_{K_j}}^s$ and ${\kappa_1}^t, {\kappa_2}^t, \ldots, {\kappa_{K_j}}^t$ are knots in the s and t directions respectively. These are chosen as the observed sample quantiles of the observed s_i's and t_i's with a maximum of 20–40 knots for each. m is a user-specified parameter chosen based on the amount of smoothness assumed in $f_s(\cdot)$ and $f_t(\cdot)$ and in our applications we use $m = 3$.

We take $\mathbf{G}_s$ to be of the form

$$\mathbf{G}_s = {\sigma_s}^2\left(\mathbf{\Omega}_s^{-1/2}\right)\left(\mathbf{\Omega}_s^{-1/2}\right)^T \quad \text{where} \quad \mathbf{\Omega}_s = \left[\left|\kappa_k^s - \kappa_{k'}^s\right|^{(2m-1)}\right]_{1\le k,\, k'\le K_s},$$

with $\mathbf{G}_t$ and $\mathbf{\Omega}_t$ defined analogously. Such a choice corresponds to penalised smoothing with a certain roughness penalty (Green and Silverman, 1994). Further justification for this choice is provided by French et al. (2001).

We reparametrise to

$$\eta_i = (\mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\tilde{\mathbf{u}})_i \tag{4}$$

where

$$\widetilde{\mathbf{Z}} = \mathbf{Z}\begin{bmatrix} \mathbf{\Omega}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_t \end{bmatrix}^{-\frac{1}{2}} = [\widetilde{\mathbf{Z}}_s\widetilde{\mathbf{Z}}_t]$$

$$\text{and} \quad \widetilde{\mathbf{u}} = \begin{bmatrix} \widetilde{\mathbf{u}}^s \\ \widetilde{\mathbf{u}}^t \end{bmatrix} = \begin{bmatrix} \mathbf{\Omega}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_t \end{bmatrix}^{\frac{1}{2}}\mathbf{u}.$$

so that our generalised linear mixed model representation reduces to one with a variance components structure

$$f(y \mid \mathbf{u}) = \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\tilde{\mathbf{u}}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\tilde{\mathbf{u}}) + \mathbf{1}^T c(\mathbf{y})\}. \tag{5}$$

Here $\tilde{\mathbf{u}}_s$ and $\tilde{\mathbf{u}}_t$ are independently distributed with densities

$$f(\tilde{\mathbf{u}}_s) = (2\pi)^{-\frac{q}{2}}{\sigma_s}^{-q}\exp\left(-\frac{1}{2{\sigma_s}^2}\mathbf{u}_s^T\mathbf{u}_s\right)$$

and

$$f(\tilde{\mathbf{u}}_t) = (2\pi)^{-\frac{q}{2}}{\sigma_t}^{-q}\exp\left(-\frac{1}{2{\sigma_t}^2}\mathbf{u}_t^T\mathbf{u}_t\right)$$

This is a form which can be fitted using standard mixed model software such as the `glimmix` macro in `SAS` and the `glmmPQL()` function in `R`.

Note that our model corresponds to 'low-rank' smoothing since the number of basis functions stays fixed at $K_s + K_t + 2$ regardless of the sample size. For large values of n such as commonly arises in epidemiological applications, this reduces the computational burden of implementing the PQL algorithm.

3 Geostatistical extension

Suppose that the data are of the form $(\mathbf{x}_i, y_i)$, $1 \le i \le n$, where as before the $y_i's$ are scalar outcomes and $\mathbf{x}_i = (x_{i1}, x_{i2})^T$, $1 \le i \le n$ now represent n geographical locations. A general non-parametric formulation should then consider a *bivariate* smoothing model of the form

$$\eta_i = g\{E(y_i|\mathbf{x}_i)\} = f(x_{i1}, x_{i2}), \quad 1 \le i \le n \tag{6}$$

where $f(\cdot)$ is an unspecified but sufficiently smooth bivariate function. A natural extension of the methods proposed by Kammann and Wand (2003) is to consider a mixed model representation of the form

$$f(y|\mathbf{u}) = \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{x}}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{x}}\mathbf{u}) + \mathbf{1}^T c(\mathbf{y})\}$$

$$f(\mathbf{u}) = (2\pi)^{-\frac{q}{2}}|\mathbf{G}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{G}^{-1}\mathbf{u}\right) \tag{7}$$

where

$$\mathbf{X} = \left[\mathbf{1}_{n\times 1}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, {\mathbf{x}_{i1}}^2, \mathbf{x}_{i1} \odot \mathbf{x}_{i2}, {\mathbf{x}_{i2}}^2, \ldots, {\mathbf{x}_{i1}}^{m-1}, \ldots, {\mathbf{x}_{i2}}^{m-1}\right]_{1\le i\le n}$$

and the basis functions comprising the columns of $\mathbf{Z}_{\mathbf{x}}$ contains *bivariate* radial basis functions, analogous to $\mathbf{Z}_s$ and $\mathbf{Z}_t$. Here $\odot$ denotes componentwise multiplication. A reasonable choice for $\mathbf{Z}_{\mathbf{x}}$ is

$$\mathbf{Z}_{\mathbf{x}} = \left[r\left(\left\|\mathbf{x}_i - \boldsymbol{\kappa}_k^{\mathbf{x}}\right\|\right)\right]_{\substack{1\le i\le n \\ 1\le k\le K^{\mathbf{x}}}} \tag{8}$$

where $r(x) = x^{(2m-2)}\log(x)$. As in the one-dimensional case (Section 2) thin plate spline theory motivates

$$\mathbf{G}_{\mathbf{x}} = \sigma_{\mathbf{x}}^2\mathbf{\Omega}_{\mathbf{x}} \quad \text{where} \quad \mathbf{\Omega}_{\mathbf{x}} = \left[r\left(\left\|\mathbf{x}_i - \boldsymbol{\kappa}_k^{\mathbf{x}}\right\|\right)\right]_{1\le k,k'\le K^{\mathbf{x}}}.$$

Reparametrisation to $\mathbf{Z}_{\mathbf{x}}\boldsymbol{\Omega}_{\mathbf{x}}^{-1/2}$ again facilitates implementation via standard GLMM software. m is a user-specified parameter chosen from considerations of the smoothness desired in $f(\cdot)$. In our application, we use $m = 3$.

The remaining choice to be made is that of the bivariate knots $\boldsymbol{\kappa}_k^{\mathbf{x}}, 1 \le k \le K^{\mathbf{x}}$. An effective strategy is to use a *space filling algorithm* (e.g. Nychka and Saltzman, 1998). The function `cover.design()` in the `R` package `fields` facilitates this algorithm. An alternative option is to use a clustering algorithm such as CLARA (Kaufman and Rousseuw, 1990) which is available in the `R` package `cluster`.

We reduce the model to one with an approximate variance component structure by reparametrising $\mathbf{Z}_{\mathbf{x}}$ to $\tilde{\mathbf{Z}}_{\mathbf{x}}$ as before We can easily combine the models in (5) and (7) to arrive at the single generalised linear mixed model

$$f(y|\mathbf{u}) = \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) + \mathbf{1}^T c(\mathbf{y})\}$$

$$f(\mathbf{u}) = (2\pi)^{-\frac{q}{2}}|\mathbf{G}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{G}^{-1}\mathbf{u}\right).$$

Here

$$\mathbf{X} = \big[\mathbf{1}\; s_i\; t_i\; \mathbf{x}_{i1}, \mathbf{x}_{i2}, {\mathbf{x}_{i1}}^2, \mathbf{x}_{i1} \odot \mathbf{x}_{i2}, {\mathbf{x}_{i2}}^2, \ldots, {\mathbf{x}_{i1}}^{m-1}, \ldots, {\mathbf{x}_{i2}}^{m-1}\big]_{1\le i\le n} \quad \text{and} \quad \mathbf{Z} = [\tilde{\mathbf{Z}}_s \tilde{\mathbf{Z}}_t \tilde{\mathbf{Z}}_{\mathbf{x}}]$$

with $\tilde{\mathbf{Z}}_s$ and $\tilde{\mathbf{Z}}_t$ are defined as in (3) and $\tilde{\mathbf{Z}}_{\mathbf{x}}$ is as defined in (8). Extension to the case where there are more than two covariates is immediate.

4 Inference

As mentioned in Section 1, one advantage of using the mixed model-based approach is that we are able to get likelihood-driven estimates of all parameters.

4.1 Asymptotic distribution

The PQL approach involves maximisation of the marginal likelihood obtained by integrating out the random effects

$$\mathcal{L}(\boldsymbol{\beta}, {\sigma_s}^2, {\sigma_t}^2, {\sigma_{\mathbf{x}}}^2) = \int f(y|\mathbf{u}_t, \mathbf{u}_s) f(\mathbf{u}_s) f(\mathbf{u}_t) d\mathbf{u}_s d\mathbf{u}_t \quad (9)$$

based on a Laplace approximation of the above integral. This can be shown to reduce to an iteratively reweighted least squares scheme with Fisher scoring updates given by

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{bmatrix} \longleftarrow (\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \mathbf{B})^{-1}\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{y}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} \quad (10)$$

where

$$\mathbf{C} = [\mathbf{X}, \mathbf{Z}]; \mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} = \operatorname{diag}\{\exp(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}})\};$$

$$\mathbf{y}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} = {\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}}^{-1}\{\mathbf{y} - \exp(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}})\}$$

Details are provided by Breslow and Clayton (1993) and Wolfinger and O'Connell (1993). Defining $\mathbf{B} = [\begin{smallmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{smallmatrix}]$, approximate standard errors can be based on the above estimating equations leading to

$$\widehat{\text{cov}}\left(\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{bmatrix}\bigg|\mathbf{u}\right) \simeq (\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \mathbf{B})^{-1} \times \mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}(\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \mathbf{B})^{-1}$$

Inference is based on the asymptotic result

$$\begin{bmatrix}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\\ \hat{\mathbf{u}} - \mathbf{u}\end{bmatrix} \sim N\big(\mathbf{0}, (\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \mathbf{B})^{-1} \times \mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}(\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \mathbf{B})^{-1}\big) \quad (11)$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are the PQL estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$ respectively. Note that the low-rank approach is crucial to the validity of this asymptotic statement since otherwise the number of parameters would grow with the sample size leading to more complicated asymptotics.

4.2 Degrees of freedom

The degrees of freedom of a fit is defined as the trace of the hat matrix defined as

$$\mathbf{H} = \mathbf{C}(\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C} + \mathbf{B})^{-1}\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}} \quad (12)$$

The motivation for this definition is provided by the expression for the Fisher update in Eq. (10) and its analogy to linear additive modeling. The degree of freedom for a predictor can be obtained by considering only the corresponding columns of the $\mathbf{C}$ matrix in (12) and provides a measure of the amount of smoothing which is comparable across various smoothing techniques.

Using a generalised linear mixed model-based approach leads to automatic likelihood-based selection of the amount of smoothing for each additive component. This is obtained by maximising an approximate 'REML -type' criterion obtained from the iteratively reweighted least squares scheme as shown by Breslow and Clayton (1993). The utility of this for a 'Scale-Space' approach is that the REML choice serves as a benchmark for an informative range of choices of degrees of freedom for the main predictor of interest while the degrees of freedom can be fixed at the REML value for all other predictors. Estimation can be performed for specified choices of degrees of freedom by first solving (12) for restrictions on the variance components.

4.3 Derivative estimation and global confidence bands

First, we consider the case where the additive model does not have a geostatistical component. In this case, our features of interest are peaks and valleys and we can base our conclusions about feature significance on the statistical significance of the first and second derivatives of the linear predictor. Assessing feature significance on the scale of the linear predictor is equivalent to assessing features on the scale of the link function $g(\cdot)$ which is typically monotonic.

We consider a grid of locations over the observed sample values of the predictor of interest. We wish to test our null hypothesis of a zero first or second derivative at each of these locations. For our model the vector of first derivatives of over n_g grid locations is a linear function of $(\boldsymbol{\beta}, \mathbf{u})^T$ of the form

$$f^{(1)} = \mathbf{X}^{(1)}\boldsymbol{\beta} + \mathbf{Z}^{(1)}\mathbf{u} = \mathbf{C}^{(1)}\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{u}\end{bmatrix} \tag{13}$$

where $\mathbf{X}^{(1)}$ and $\mathbf{Z}^{(1)}$ are appropriately defined design matrices and

$$\mathbf{C}^{(1)} = [\mathbf{X}^{(1)}\ \mathbf{Z}^{(1)}]$$

so that we wish to test 'point-wise' whether each component of $f^{(1)}$ is significantly non zero.

As mentioned earlier, we wish to perform a 'Scale -Space' analysis for one predictor of interest. Accordingly, we use the REML degrees of freedom for this predictor to determine a reasonable range of choices of degrees of freedom for this predictor while we fix the amount of smoothing for all other predictors at the corresponding REML default. For each specified choice of degrees of freedom for the main predictor, we can estimate the vector of partial derivatives by $\mathbf{C}^{(1)}[\begin{smallmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{smallmatrix}]$. Using the asymptotic theory for PQL estimation as outlined in (11), we can standardise the estimates of $f^{(1)}(\mathbf{x})$ over a grid of values $\mathbf{x}$ by its corresponding standard error to construct Wald-type statistics $W(\mathbf{x})$ of the form $(\hat{f}^{(1)}(\mathbf{x}))^2/\widehat{\text{Var}}(\hat{f}^{(1)}(\mathbf{x}))$ to test the null hypothesis of a zero partial derivative.

However, we are interested in testing the hypothesis of a null first derivative pointwise over all n_g locations so that it is necessary to guard against the problem of multiple inferences. Using our generalised linear mixed model formulation, we can find a $100(1-\alpha)\%$ *simultaneous* confidence region for the derivatives $f^{(1)}(\mathbf{x})$ at all n_g locations. We do this by simulating the asymptotic distribution of the maximum of the Wald statistics, $W(\mathbf{x})$ over the grid i.e. given a desired overall level of significance α we find, by simulation, a cutoff value q_α satisfying

$$P\left(\max_{\mathbf{x}} W(\mathbf{x}) \leq q_\alpha\right) = 1 - \alpha$$

At each simulation we perform the following steps:

1. Generate a realisation from the distribution of $[\begin{smallmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\mathbf{u}}-\mathbf{u}\end{smallmatrix}]$ specified in (11).
2. Use the formula for the partial derivatives in (13) to get realisations of $\hat{f}^{(1)} - f^{(1)}$.
3. Using these, calculate the Wald statistics $W(\mathbf{x})$. Estimated variances required for standardising the estimates obtained above are given by the diagonal entries of

$$\mathbf{C}^{(1)}\left(\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}+\mathbf{B}\right)^{-1}\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}\left(\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}+\mathbf{B}\right)^{-1}\mathbf{C}^{(1)T}$$

4. Hence calculate $\max_{\mathbf{x}} W(\mathbf{x})$.

We recommend performing about 1000 simulations. The empirical $100(1-\alpha)$th percentile is then our required estimate, say $\widehat{q_\alpha}$. We reject the null hypothesis of a zero gradient at a particular grid location $\mathbf{x}$ if $w(\mathbf{x}) > \widehat{q_\alpha}$ where $w(\mathbf{x})$ denotes the value of the observed Wald statistic at grid-point $\mathbf{x}$. This ensures an overall nominal level of significance α over the grid.

We can easily extend the above methodology to assess significant curvature over the grid by noting as before that the vector of second derivatives of the linear predictor over the $n_g$grid locations are also linear function of $(\boldsymbol{\beta}, \mathbf{u})^T$ of the form. Information about the statistical significance of slope and curvature can be displayed graphically using SiZer plots. (Chaudhuri and Marron, 1999).

In case, the main predictor of interest is geographical location, we need to find rotation-invariant measures of slope and curvature. Slope can be assessed based on the length of gradient at a grid location $\mathbf{x}$ given by

$$G(\mathbf{x}) = \sqrt{f^{(0,1)}(\mathbf{x})^2 + f^{(1,0)}(\mathbf{x})^2}$$

where $f^{(i,j)}(\mathbf{x})$ denotes the partial derivatives of order (i, j) at $\mathbf{x}$. Our null hypothesis of interest is then

$$H_0 : G(\mathbf{x}) = 0,$$

and we wish to test this point-wise for all grid points $\mathbf{x}$. For our model the vectors of first partial derivatives of orders $(0, 1)$ and $(1, 0)$ over n_g grid locations are linear function of $(\boldsymbol{\beta}, \mathbf{u})^T$ of the form

$$f^{(1,0)} = \mathbf{X}^{(1,0)}\boldsymbol{\beta} + \mathbf{Z}^{(1,0)}\mathbf{u} = \mathbf{C}^{(1,0)}\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{u}\end{bmatrix}$$
$$f^{(0,1)} = \mathbf{X}^{(0,1)}\boldsymbol{\beta} + \mathbf{Z}^{(0,1)}\mathbf{u} = \mathbf{C}^{(0,1)}\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{u}\end{bmatrix} \tag{14}$$

where $\mathbf{X}^{(0,1)}, \mathbf{Z}^{(0,1)}$ and $\mathbf{X}^{(1,0)}, \mathbf{Z}^{(1,0)}$ are appropriately defined design matrices, $\mathbf{C}^{(1,0)} = [\mathbf{X}^{(1,0)}\ \mathbf{Z}^{(1,0)}]$ and $\mathbf{C}^{(0,1)} =$

$[\mathbf{X}^{(0,1)}\ \mathbf{Z}^{(0,1)}]$. Hence for each specified choice of the smoothing parameter, we can estimate the partial derivatives by $\mathbf{C}^{(1,0)}[\begin{smallmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{smallmatrix}]$ and $\mathbf{C}^{(0,1)}[\begin{smallmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{smallmatrix}]$ where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are the PQL estimates. The explicit calculations for $(\mathbf{X}^{(1,0)}, \mathbf{Z}^{(1,0)}, \mathbf{X}^{(0,1)}, \mathbf{Z}^{(0,1)})$ have been given in Ganguli and Wand (2004) and are available from the authors. Using the asymptotic theory for PQL estimation as outlined in (11), we can find the joint asymptotic distribution of $(f^{(0,1)}(\mathbf{x}), f^{(1,0)}(\mathbf{x}))^T$ at each grid-location $\mathbf{x}$ and hence construct Wald-type statistics $W(\mathbf{x})$.

As before, we correct for multiple comparisons by finding a $100(1-\alpha)\%$ *simultaneous* confidence region for the gradients $G(\mathbf{x})$ at all n_g locations. We do this by simulating the asymptotic distribution of the maximum of the Wald statistic over the grid i.e. given a desired overall level of significance α we find, by simulation, a cutoff value q_α satisfying

$$P\Big[\max_{\mathbf{x}} W(\mathbf{x}) \le q_\alpha\Big] = \alpha$$

As before, at each simulation we perform the following steps:

1. Generate a realisation from the distribution of $[\begin{smallmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\mathbf{u}}-\mathbf{u}\end{smallmatrix}]$ specified in (11)
2. Use the formula for the partial derivatives in (14) to get realisations of $\hat{f}^{(1,0)} - f^{(1,0)}$ and $\hat{f}^{(0,1)} - f^{(0,1)}$.
3. Using these, calculate the Wald statistics $W(\mathbf{x})$. For $\hat{f}^{(1,0)} - f^{(1,0)}$ estimated variances and covariances for the Wald statistics are given by the diagonal entries of

$$\mathbf{C}^{(1,0)}\big(\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}+\mathbf{B}\big)^{-1}\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}\big(\mathbf{C}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{C}+\mathbf{B}\big)^{-1} \times \mathbf{C}^{(1,0)^T}.$$

   An analogous matrix is required for $\hat{f}^{(0,1)} - f^{(0,1)}$.
4. Hence calculate $\max_{\mathbf{x}} W(\mathbf{x})$.

The empirical $100(1-\alpha)$ th percentile is our required estimate, $\widehat{q_\alpha}$ and we reject the null hypothesis of a zero gradient at a particular grid location $\mathbf{x}$ if $w(\mathbf{x}) > \widehat{q_\alpha}$ where $w(\mathbf{x})$ denotes the observed value of the Wald statistic at location $\mathbf{x}$.

For assessing local curvature of a two-dimensional surface, we base our inference on the eigenvalues of the Hessian matrices over the grid. These provide a rotation-invariant means of assessing local curvature and have geometrical significance of being the greatest and least amounts of curvature at a point with the directions of curvature being given by the corresponding eigenvectors.

Table 1 Characterisation of curvature

Feature	Characterisation
Hole	$\hat{\lambda}_+(\mathbf{x}), \hat{\lambda}_-(\mathbf{x}) > q_T$
Long valley	$\hat{\lambda}_+(\mathbf{x}) > q_T,\ \lvert\hat{\lambda}_-(\mathbf{x})\rvert < q_T$
Saddle point	$\hat{\lambda}_+(\mathbf{x}) > q_T,\ \hat{\lambda}_-(\mathbf{x}) < -q_T$
Long ridge	$\lvert\hat{\lambda}_+(\mathbf{x})\rvert < q_T,\ \hat{\lambda}_-(\mathbf{x}) < -q_T$
Peak	$\hat{\lambda}_+(\mathbf{x}), \hat{\lambda}_-(\mathbf{x}) < -q_T$

For two-dimensional data, the eigenvalues $\lambda_+(\mathbf{x})$ and $\lambda_-(\mathbf{x})$ can be obtained explicitly as

$$\lambda_\pm(\mathbf{x}) = \Big[f^{(2,0)}(\mathbf{x}) + f^{(0,2)}(\mathbf{x}) \pm\sqrt{\{f^{(2,0)}(\mathbf{x}) - f^{(0,2)}(\mathbf{x})\}^2 + 4f^{(1,1)}(\mathbf{x})^2}\Big]\Big/2 \qquad (15)$$

where $f^{(i,j)}(\mathbf{x})$ denotes the partial derivative of order (i, j) at $\mathbf{x}$. Our hypothesis of interest is

$$H_0 : T(\mathbf{x}) = 0$$

at all grid locations $\mathbf{x}$ where $T(\mathbf{x}) = \max\{|\lambda_+(\mathbf{x})|, |\lambda_-(\mathbf{x})|\}$. Noting as before that all partial derivatives $f^{(i,j)}$ are linear functions of $(\boldsymbol{\beta}, \mathbf{u})$, we can estimate them as before using the PQL estimates to construct our test-statistic $\hat{T}(\mathbf{x}) = \max\{|\hat{\lambda}_+(\mathbf{x})|, |\hat{\lambda}_-(\mathbf{x})|\}$. using the formula in (15).

Due to the non-linear nature of $T(\mathbf{x})$, we cannot construct a confidence region for $(\lambda_+(\mathbf{x}), \lambda_-(\mathbf{x}))$ over the grid as before. However we can simulate the *null* distribution of $\max_{\mathbf{x}} \hat{T}(\mathbf{x})$ to arrive at a cutoff value q_T which ensures a pre-specified overall level of significance α. Since curvature estimates can be noisy, we also estimate the variance of $\max_{\mathbf{x}} T(\mathbf{x})$ from the simulation and find the cutoff for the standardised statistic.

If H_0 is rejected at a grid location $\mathbf{x}$, we use the characterisation developed by Godtliebsen et al. (2002, 2004) to determine the probable nature of curvature at $\mathbf{x}$, listed in Table 1.

As with the gradient, estimation and inference about curvature is conducted across a range of choices of degrees of freedom.

4.4 Effective sample size

We define the effective sample size (ESS) at a grid-point $\mathbf{x}$ as

$$ESS_{\mathbf{x}} = \frac{\widehat{\mathrm{Var}}(y)}{\widehat{\mathrm{Var}}(\widehat{E}(y|\mathbf{x}))}.$$

In performing inference, we ignore regions in which the effective sample size is less than 5 since the gradient estimate in such regions is likely to be unstable. Justification for using this definition is provided by Ruppert and Wand (1994) who show that this ratio equals the sample size used for local fits in case of non-parametric local regression using a uniform kernel. When the kernel used is non-uniform, the authors comment that this ratio provides a reasonable definition of 'effective sample size' since it is penalised by larger conditional variance of y given $\mathbf{x}$ and sparse data near $\mathbf{x}$. Additional justification is provided by Fan et al. (1995).

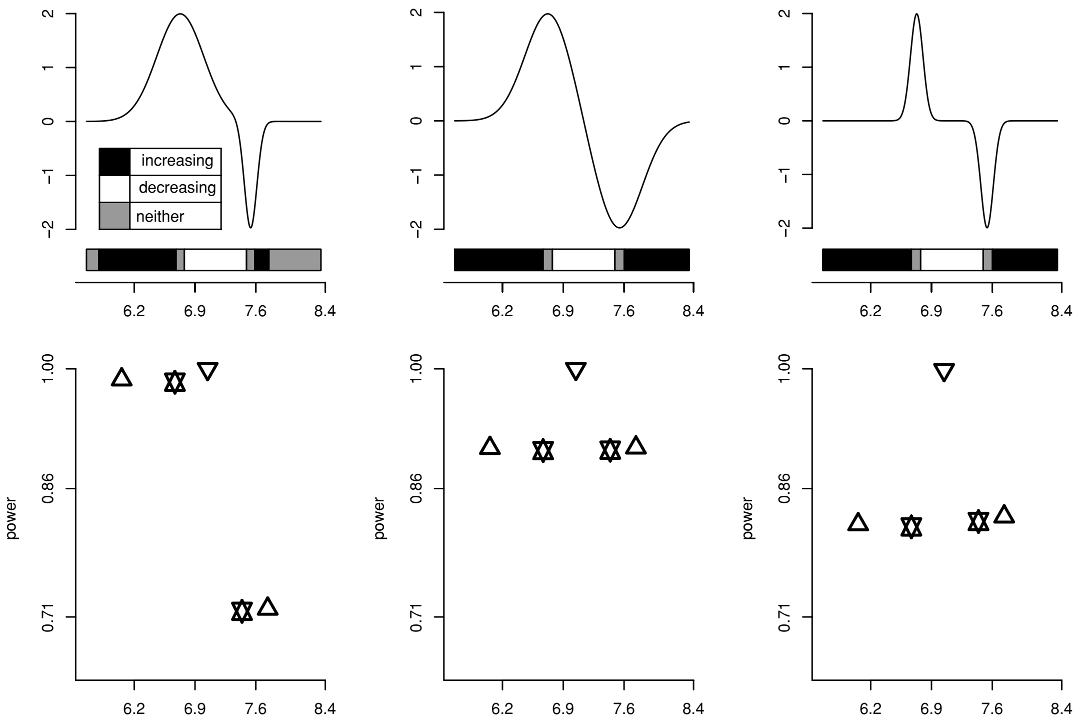


Fig. 3 Summary of simulation results for three different versions of the log-odds ratio; $n = 300$. For each version, the upper panels show the true log odds function. The bar at the base corresponds to the qualitative properties of the function as indicated by the legend. The lower panels plot the empirical power to detect qualitative features of the log odd functions. The key for the symbols is: (△) = increase, (▽) = decrease, (✡) = turning point

An alternative definition which can be used only at the observed data-points and which can be justified using the concept of 'equivalent kernels' (Green and Silverman, 1994) is

$$ESS_{\mathbf{x}_i} = \sum_{j=1}^{n} h_{ij}/h_{ii}$$

where $\mathbf{H} = [h_{i,j}]$ is the hat matrix as defined in (12). The two definitions are identical in the case of Gaussian data with an identity link and for the applications considered, led to identical results at observed sample locations for non-Gaussian data.

5 Simulation study

We conducted several simulation studies with binary outcomes and a single predictor for varying choices of sample sizes, sample design, levels of significance and the shape of the log-odds ratio as a function of the predictor. Unless otherwise indicated a sample size of 300 was used.

Figure 3 shows a graphical summary of the simulation results. The upper panel shows the true log-odds as a function of the predictor with varying levels of curvature. These were obtained as a linear combination of two normal curves. The bar below each plot represents the truth for the analytic derivative. Note that the grey bars correspond to regions in which the derivative is of very small magnitude (< 0.0001) since it is never exactly zero. The lower panel shows the simulated probabilities of variously identifying (a) only the initial rise (b) both the initial rise and the initial dip i.e. the peak (c) only the subsequent dip (d) both the subsequent dip and the subsequent rise i.e. the trough. Empirical power, say for the first increase was measured by first finding the proportion of times the first derivative was found to be significantly positive at at least one location over the first black bar separately for each choice of degrees of freedom. This was then averaged across degrees of freedom to yield the empirical power. The plots generally indicate that features with less curvature are more likely to be identified with a power of around 0.9–1 while the more abrupt features are identified with a power of around 0.7–0.85.

Figure 4 shows simulated probabilities for $n = 100, 200$ and 300. Figure 5 shows simulated probabilities for the case of a $N(7, 1)$ distributed predictor and an equispaced design respectively while Fig. 6 shows simulated probabilities for $\alpha = 0.01$ and 0.05. These simulations indicate that the probabilities of detecting features are not particularly sensitive to sample size although the probability of detecting the the peak and valley are somewhat higher in the region of 0.95 and 0.7 respectively in case of a sample size of 300 as compared to

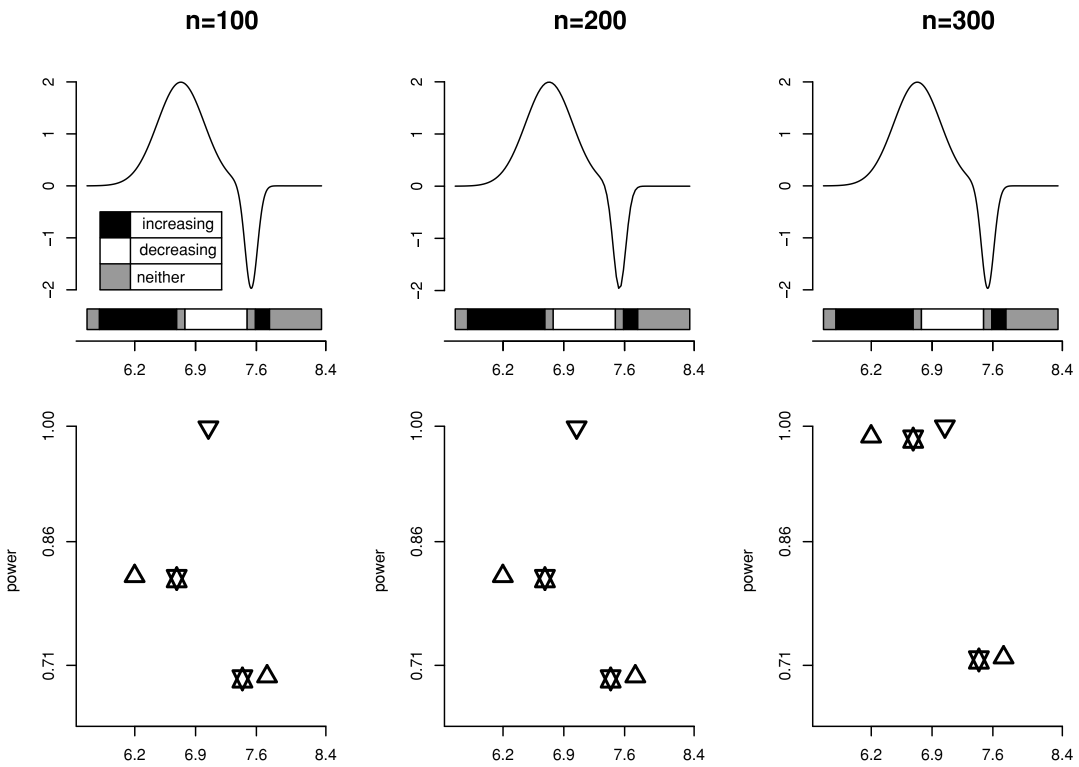


Fig. 4 Summary of simulation results for $n = 100, 200, 300$. For each version, the upper panels show the true log odds function. The bar at the base corresponds to the qualitative properties of the function as indicated by the legend. The lower panels plot the empirical power to detect qualitative features of the log odd functions. The key for the symbols is: (△) = increase, (▽) = decrease, (✡) = turning point

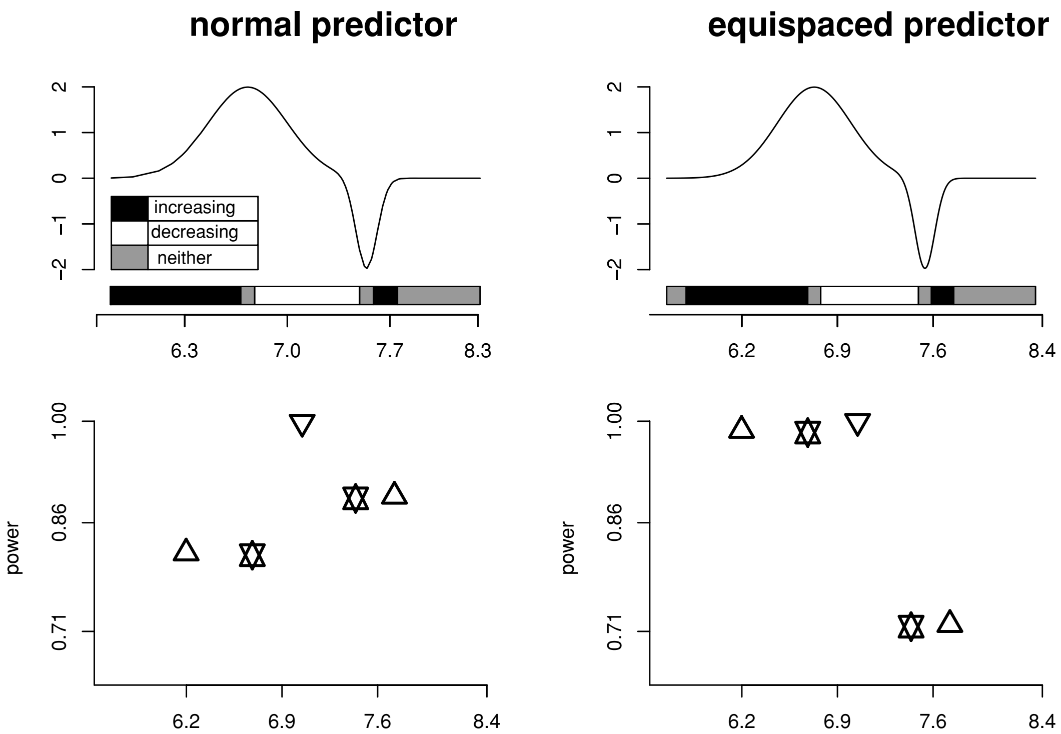


Fig. 5 Summary of simulation results for a normally distributed predictor and an equispaced design; $n = 300$. For each version, the upper panels show the true log odds function. The bar at the base corresponds to the qualitative properties of the function as indicated by the legend. The lower panels plot the empirical power to detect qualitative features of the log odd functions. The key for the symbols is: (△) = increase, (▽) = decrease, (✡) = turning point

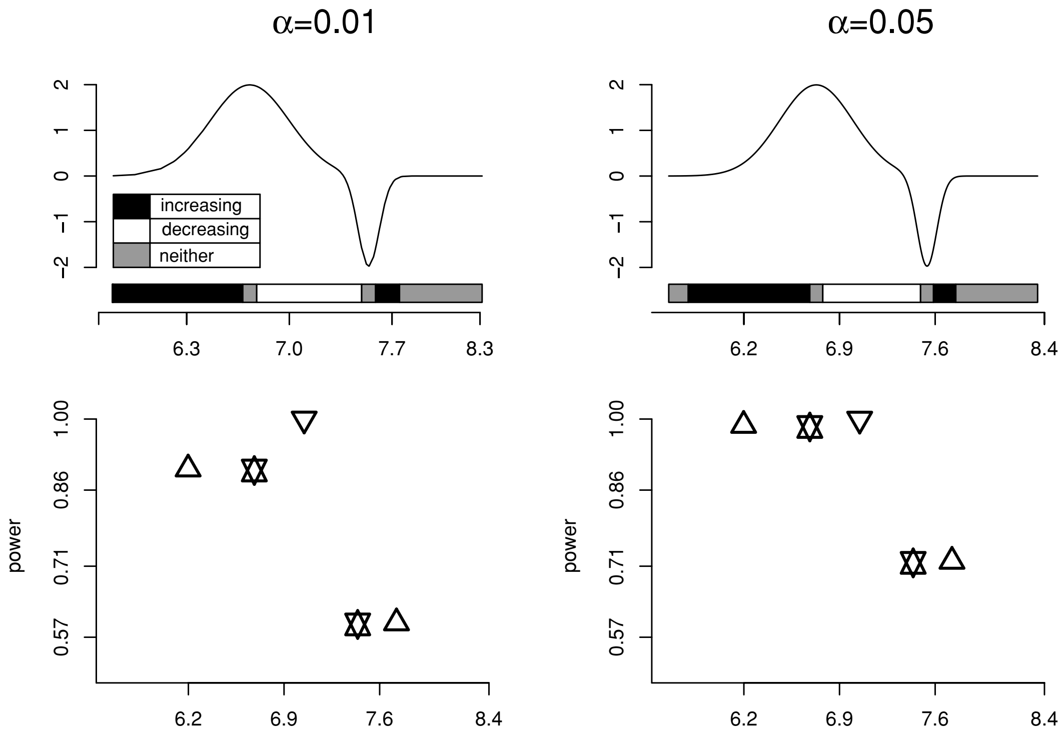


Fig. 6 Summary of simulation results for $\alpha = 0.01, 0.05$ and $n = 300$. For each version, the upper panels show the true log odds function. The bar at the base corresponds to the qualitative properties of the function as indicated by the legend. The lower panels plot the empirical power to detect qualitative features of the log odd functions. The key for the symbols is: (△) = increase, (▽) = decrease, (✡) = turning point

figures of 0.9 and 0.6–0.65 respectively for the cases $n = 100$ and 200. The dependence of the power on the overall level of significance required seems to be of a similar nature with higher probabilities of detecting the peak and trough in the case $\alpha = 0.05$ as compared to the case $\alpha = 0.01$. In case of a varying design, it appears that the power to detect the initial rise, peak and dip are higher for an equispaced design while the power to detect the subsequent valley and rise are higher in case of the normally distributed predictor. This could be explained by the higher clustering of sample values around these two latter and sharper features in case of a normal design.

6 Examples

In this section, we first present two applications of our methodology for additive modeling. The first is to the mortality data from Milan (Zanobetti et al., 2000) and the second is to data on union membership as a function of wage (Berndt, 1991) which show similar suggestions of a non-linear effect in univariate analysis with a default amount of smoothing.

Figure 7 is a SiZer plot for a univariate Poisson mixed model fit to the Milan data. The vertical axis indexes degrees of freedom with the ‘REML’ choice indicated by the horizontal black line. We interpret a SiZer plot by considering each value of the predictor plotted on the horizontal axis and looking across the vertical axis to check if there is at least one value at which the corresponding colour indicates the presence of a significant feature at that value. Features are indicated using the colour key. A detailed discussion of SiZer plots is provided by Chaudhuri and Marron (1999). Thus, the plot suggests that mortality is relatively constant as a function of average daily temperature except for a few dips towards the middle of the curve but increases beyond temperatures of about 30 degrees Celsius.

Figure 8 shows the SiZer plot from an additive modeling of the same data. Let temp_i, TSP_i, humid_i and mort_i be the average daily temperature, measured total suspended particles, average daily humidity and total all-cause mortality on day i. holiday_i is an indicator of whether or not day i corresponded to a holiday. Our additive model regression equation is

$$\log \mathrm{E}(\text{mort}_i) = \beta_0 + \beta_1(\text{holiday}_i) + f_1(\text{temp}_i) + f_2(\text{humid}_i) + f_3(\text{TSP}_i)$$

The SiZer plot from the adjusted analysis also indicates that other covariates remaining the same, mortality increases as a function of temperature at higher ends of the range but also suggests a range towards the middle where mortality decreases as a function of temperature.

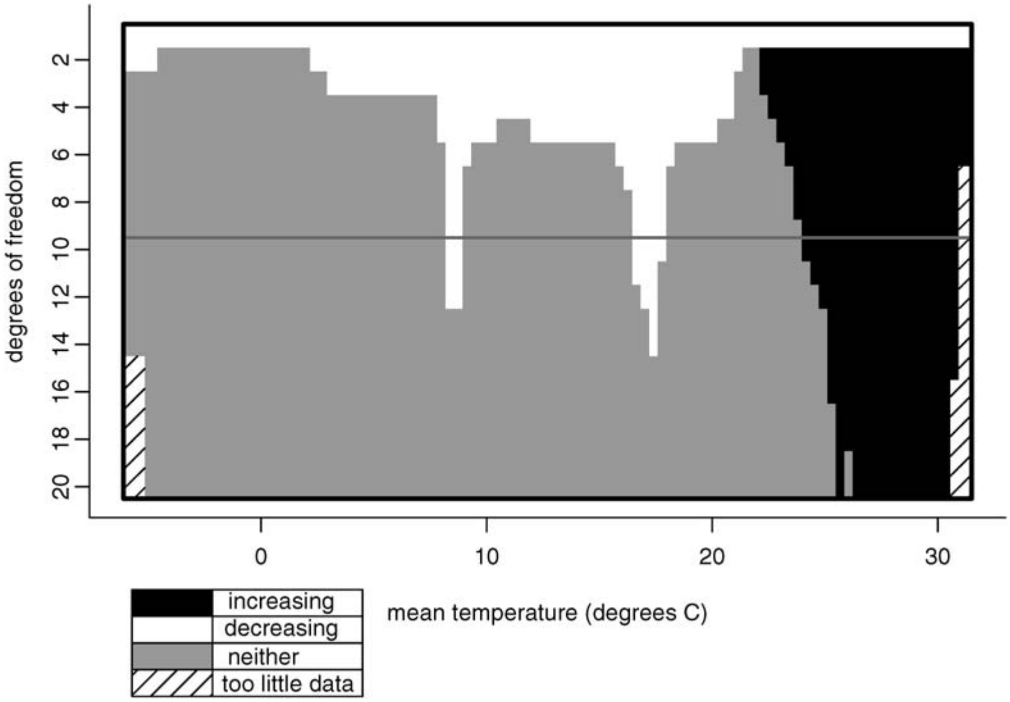


Fig. 7 SiZer map for unadjusted log mortality ratio. The horizontal line corresponds to degrees of freedom estimated via REML

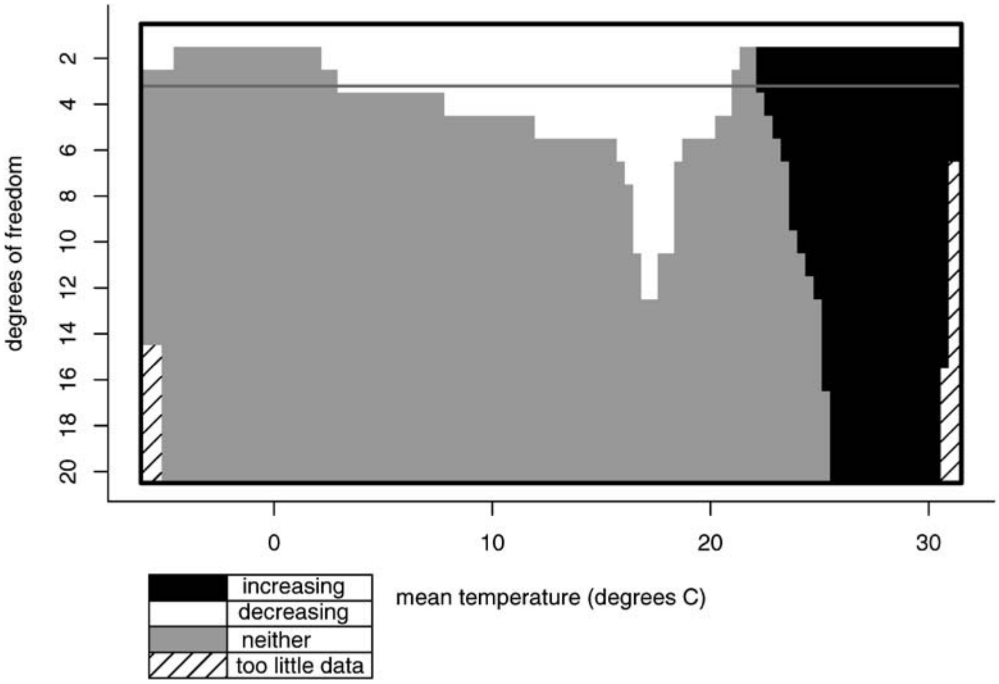


Fig. 8 SiZer map for log mortality ratio adjusted for other confounders. The horizontal line corresponds to degrees of freedom estimated via REML

Figure 9 is a SiZer plot showing significant curvature in the adjusted log mortality ratio for temperature for the Milan data. The plot is interpreted similar to a SiZer plot for significant gradient except that now significance relates to convexity or concavity of the estimated ratio. Thus, the plot suggests significant convexity towards both ends of the range but no statistically significant departure from piecewise linearity over the rest of the range.

Figures 10 and 11 show SiZer plots for the odds of union membership as a function of wages from univariate and

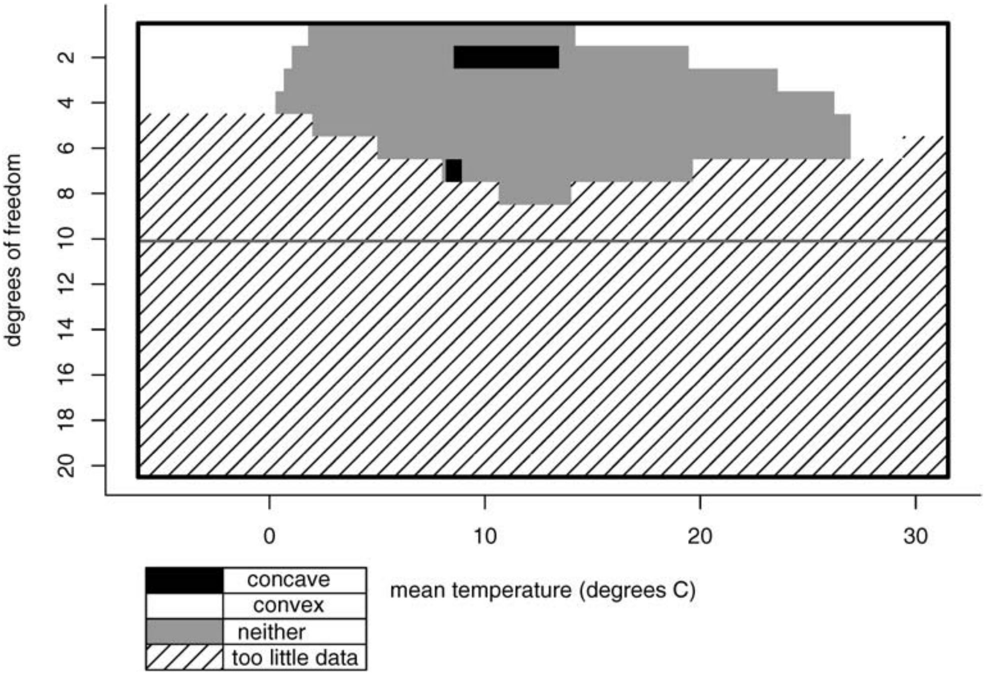
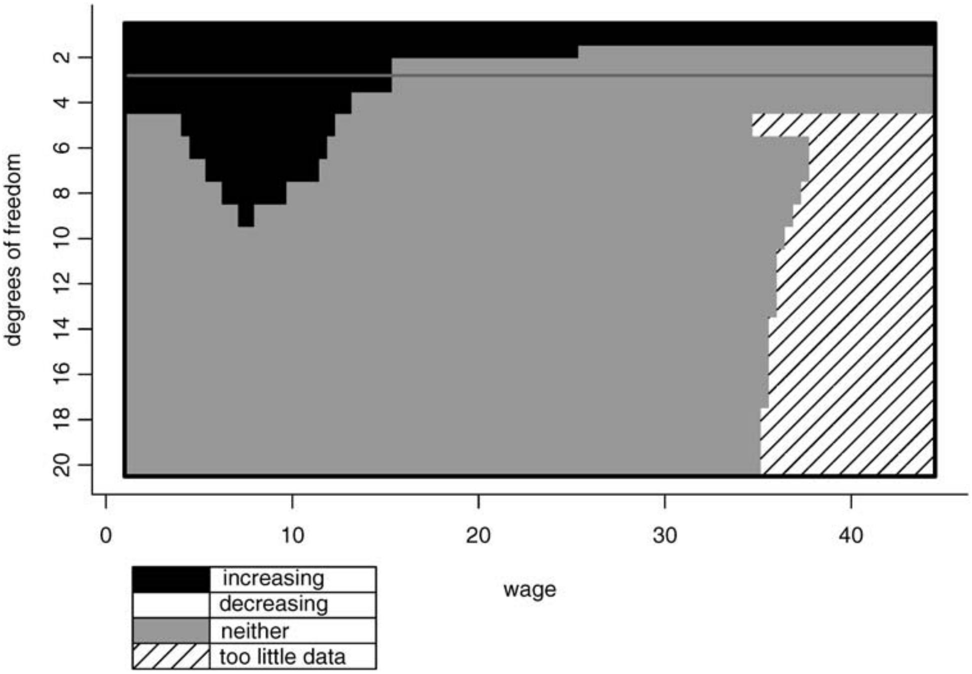


Fig. 9 SiZer map for curvature in log mortality ratio for Milan data adjusted for other confounders. The horizontal line corresponds to degrees of freedom estimated via REML

Fig. 10 SiZer plot for log odds of trade-union membership unadjusted for other covariates. The horizontal line corresponds to degrees of freedom estimated via REML



additive modeling. Both plots indicate that the odds initially increase as a function of wage and subsequently remain unchanged. Figure 10 has a white bar suggesting a decline for very high wages at a level of smoothing which is close to the default specified in S-PLUS. The grey regions towards the higher ranges suggests that this could be due to the influence of a few outlying observations. The white bar does not show up on the SiZer plot for the adjusted analysis.

We also performed a formal additive analysis of the female lung cancer rates presented in Section 1 adjusting for smoking status and age of participants. The model fitted was

$$\log\left(\frac{P[\text{case}_i]}{1-P[\text{case}_i]}\right) = \beta_0 + \beta_1(\text{current-smoker}_i) \\ + \beta_2(\text{past-smoker}_i) + f(\text{age}_i) + g(\mathbf{x}_i)$$

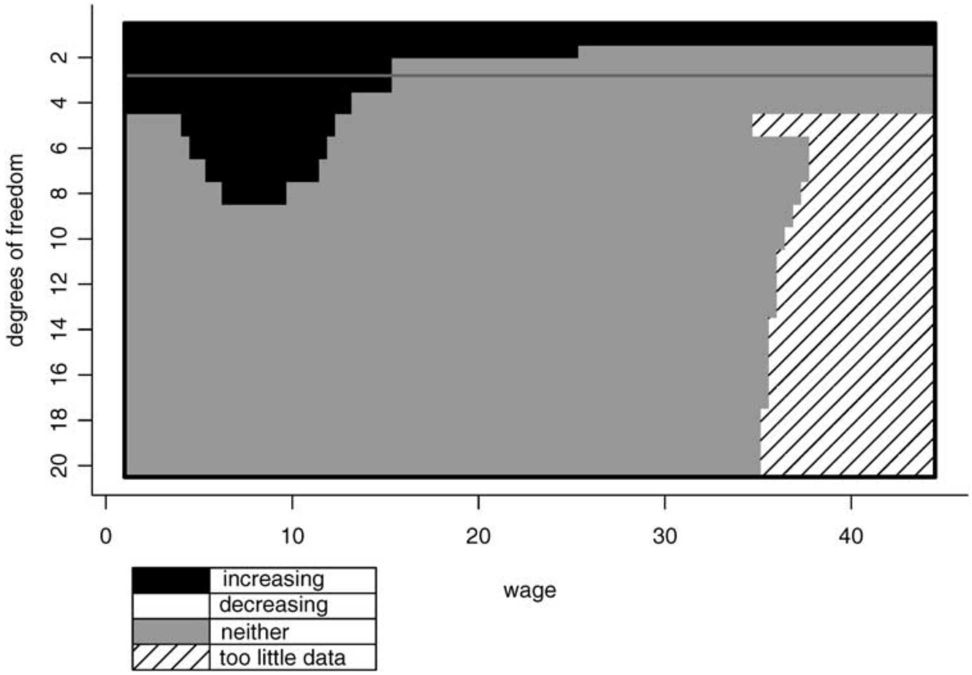


Fig. 11 SiZer plot for log odds of trade-union membership adjusted for other covariates. The horizontal line corresponds to degrees of freedom estimated via REML

where case$_i$ is an indicator of whether or not the ith respondent had lung cancer, current-smoker$_i$ and past-smoker$_i$ are indicators of whether or not the ith respondent was respectively a current or a past smoker, age$_i$ is the age of the ith respondent and $\mathbf{x}_i$ denotes the location of the ith respondent measured in terms of latitude and longitude. While, the analysis indicated a broad trend from North West to South East. None of the plots find any statistically significant gradients or curvatures.

7 Closing remarks

In this paper, we extend the methodology presented by Ganguli and Wand (2004) to assess feature significance in generalised additive models. Using a generalised linear mixed model formulation enables us to derive likelihood-based estimates of model parameters. Further extensions worth exploring include the possibility of handling additional complications which can be addressed using a mixed model likelihood-based approach. These include modeling data missingness, allowing for interactions between predictors or accounting for outliers.

The Scale-Space approach of Godtliebsen et al. (2004) leads to inferences which are simultaneous over surface locations but not over levels of smoothing. It is conceptually simple to modify our approach to also account for simultaneous inference over levels of smoothing since we can instead simulate the asymptotic distribution of the maximum Wald statistic with the maximum being taken over the grid as well as over levels of smoothing. However, the question that arises here is whether such an approach would lead to confidence regions which are too large to be informative.

Acknowledgments We are grateful to Dr. J. S. Marron for comments and suggestions. The research of B. Ganguli was funded by U. S. National Institute of Environmental Health Sciences grant R01-ES10844-01.

References

Berndt E.R. 1991. The Practice of Econometrics: Classical and Contemporary. Addison-Wesley: Reading, Massachusetts.

Breslow N.E. and Clayton D.G. 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88: 9–25.

Chaudhuri P. and Marron J.S. 1999. SiZer for exploration of structures in curves. Journal of the American Statistical Association 94: 807–823.

Chaudhuri P. and Marron J.S. 2000. Scale space view of curve estimation. The Annals of Statistics 28: 408–428.

Cressie N. 1989. Geostatistics. The American Statistician 43: 197–202.

Fan J., Heckman N.E., and Wand M.P. 1995. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. Journal of the American Statistical Association 90: 141–150.

French J.L., Kammann E.E., and Wand M.P. 2001. Comment on paper by Ke and Wang. Journal of the American Statistical Association 96: 1285–1288.

Ganguli B. and Wand M.P. 2004. Feature significance in geostatistics. Journal of Computational and Graphical Statistics 13: 954–973.

Godtliebsen F., Marron J.S., and Chaudhuri P. 2002. Significance in scale space for bivariate density estimation. Journal of Computational and Graphical Statistics 11: 1–22.

Godtliebsen F., Marron J.S., and Chaudhuri P. 2004. Statistical significance of features in digital images. Image and Vision Computing 13: 1093–1104.

Green P.J. and Silverman B.W. 1994. Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London.

Hastie T. 1996. Pseudosplines. Journal of the Royal Statistical Society, Series B 58: 379–396.

Kammann E.E. and Wand M.P. 2003. Geoadditive models. Applied Statistics 52: 1–18.

Kaufman L. and Rousseeuw P.J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Nychka D. and Saltzman N. 1998. Design of Air Quality Monitoring Networks. In: D. Nychka, L. Cox, and W. Piegorsch (Eds.), Case Studies in Environmental Statistics, Lecture Notes in Statistics, Springer-Verlag, pp. 51–76.

Ruppert D. and Wand M.P. 1994. Multivariate locally weighted least squares regression. The Annals of Statistics 22: 1346–1370.

Wolfinger R. and O'Connell M. 1993. Generalized linear mixed models: A pseudo-likelihood approach. Journal Statistical Computation and Simulation 48: 233–243.

Zanobetti A., Wand M.P., Schwartz J., and Ryan L.M. 2000. Generalized additive distributed lag models. Biostatistics 1: 279–292.