Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computational Statistics & Data Analysis 50 (2006) 688 – 709

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Bayesian curve estimation by model averaging

Daniel Peña, Dolores Redondas*

Department of Statistics, Universidad Carlos III de Madrid, c/Madrid 126, 28903, Getafe, Madrid, Spain

Received 16 July 2003; received in revised form 21 September 2004; accepted 21 September 2004
Available online 13 October 2004

Abstract

A Bayesian approach is used to estimate a nonparametric regression model. The main features of the procedure are, first, the functional form of the curve is approximated by a mixture of local polynomials by Bayesian model averaging (BMA), second, the model weights are approximated by the BIC criterion and third, a robust estimation procedure is incorporated to improve the smoothness of the estimated curve. The models considered at each sample points are polynomial regression models of order smaller than four, and the parameters are estimated by a local window. The predictive value is computed by BMA, and the posterior probability of each model is approximated by the exponential of the BIC criterion. Robustness is achieved by assuming that the noise follows a scale contaminated normal model, so that the effect of possible outliers is downweighted. The procedure provides a smooth curve and allows a straightforward prediction and quantification of the uncertainty. The method is illustrated with several examples and Monte Carlo experiments.

Keywords: Bayesian model averaging; BIC criterion; Robustness; Nonparametric curve fitting; Local polynomial regression

1. Introduction

A Bayesian approach is used to estimate nonparametrically a regression model

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

given the bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$. We are interested in estimating the functional relationship, m, between the variable y and the explanatory variable x, and to predict the

* Corresponding author. Tel.: 34 916249314; fax: 34 916249849.
E-mail addresses: dpena@est-econ.uc3m.es (D. Peña), redondas@est-econ.uc3m.es (D. Redondas).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.09.008

response for new values of the covariate. The functional form of $m(\cdot)$ is unknown and it is approximated by a mixture of local polynomials estimators.

Both parametric and nonparametric techniques are commonly used to find the regression function $m(\cdot)$. The first parametric approach was to use polynomial regression by selecting the best order of the polynomial to fit the data, see Anderson (1962), Guttman (1967), Hager and Antle(1968), Brooks (1972) and Halpern (1973). The limitations of this approach are due to its global nature, that is, we may need a high order polynomial to approximate the data over the whole range and, even then, the approximation can be poor in wiggly curves. Second, this procedure is very non robust and a simple observation can exert a big influence on the estimated curve.

There is extensive literature for nonparametric techniques, see for example Eubank (1988), Wahba (1990), Hastie and Tibshirani (1990) and Green and Silverman (1994) for a complete survey. Some often used alternatives are piecewise polynomials, splines smoothers and local polynomial regression. The first two methods require selecting the number and positions of the knots. This is not an easy task: a small number of knots reduces the degrees of freedom of the fitted curve and a large number of knots produces overfitting. An excellent review of this topic can be found in Hansen and Kooperberg (2002). Some procedures have been proposed for the automatic selection of the knots, see Wahba (1975), Smith (1982) and Friedman and Silverman (1989). Stone et al. (1997) propose a stepwise approach in which knots can be introduced and deleted and are evaluated by the log-likelihood. From the Bayesian point of view this process can be carried out by using reversible jump Markov chain Monte Carlo (Green, 1995), where the number and the position of the knots are determined by the data, treating both quantities as random variables. See Denison et al. (2002) for a general discussion of this curve-fitting with free-knot procedure. Smith and Kohn (1996) used this Bayesian approach to select the number of knots over a large set for additive regression models. Denison et al. (1998) have applied this method for general univariate and additive models using piecewise polynomials instead of splines, because the first are more flexible for fitting curves that are not smooth. Mallick (1998) proposed estimating the function by taking the order of the polynomial as a random variable and making inference of the joint distribution of both the order of the polynomial and the polynomials coefficients. Liang et al. (2001) introduced an automatic prior setting for the multiple linear regression and they applied the method to Bayesian curve fitting with regression splines. DiMatteo et al. (2001) also applied a free-knot splines approach to data coming from the exponential family by using the BIC criterion as an approximation to the integrated likelihood ratios for the acceptance probabilities. Holmes and Mallick (2003) have applied the free-knot regression to generalized nonlinear modelling. Yau et al. (2003) have proposed a Bayesianvariable selection and model averaging approach to multinomial nonparametric regression which can handle a large number of variables and their interactions. They use a multinomial probit regression model with data augmentation (Albert and Chib, 1993) to turn the multinomial regression problem into a sequence of smoothing problems with Gaussian errors. In general all these procedures require a high computational cost. For instance fitting thin-plate splines to two or more variables requires $O(n^3)$ computations (see Wahba, 1984) and this complexity will increase with the free-knot approach. On the other hand they have a large flexibility, as tightly-spaced knots can produce peaks and widely-spaced knots smooth functions.

Local polynomial regression fits simple parametric models in neighborhoods defined by the regressors. It usually requires a low computational cost and was developed in the works of Stone (1977), Katkovnik (1979), Stone (1980) and Cleveland (1979). See also Loader (1979). Cleveland (1979) and Cleveland and Devlin (1988) proposed a popular procedure, the loess (locally weighted regression), which uses local regression with a kernel around the point of interest and is made robust by using weighted regression. This procedure is fast to compute (of order $O(n)$) but has two main problems. First, it fits the same local model over all the range of the data and thus it has no spatial adaptability and makes the result very dependent on the neighborhood used. Second, it uses M estimators for robustness and it is well know that these estimates are not robust for high leverage observations (see for instance Peña and Yohai, 1999).

In this work we also use local polynomial regression, because of its simplicity and low computational cost, but we introduce two main modifications over previous methods. First, instead of using a fixed degree local polynomial the functional form of the curve is approximated by a mixture of local polynomials by Bayesian model averaging (BMA). Bayesian model averaging leads to forecasts which are a weighted average of the predictive densities obtained by considering all possible polynomial degrees with weights equal to the posterior probabilities of each degree. BMA takes into account the uncertainty about the different models, as was pointed out in the seminal work of Leamer (1978). See George (1999), Raftery et al. (1997), Fernández et al. (2001) and Liang et al. (2001) for interesting applications. In our case, BMA is implemented by fitting local polynomial regression models of degree going from zero to d to the data in a window around each observation, and estimating the unknown regression function by a weighted sum of the values corresponding to the polynomials, with weights equal to the posterior probabilities of each polynomial model. These weights are approximated by the exponential of the BIC criterion (Schwarz, 1978), which approximates the Bayes factor. Second, we made our procedure robust by assuming that the noise may be contaminated. Then the Bayesian estimation provides an automatic downweighting of outliers which takes into account their leverage.

These two modifications keep the main advantages of local polynomial regression methods but provide a more flexible and robust procedure. The use of BMA introduces some spatial adoptability to our procedure, because although we use a fixed window for the local estimation, we allow for a changing polynomial degree. This spatial adoptability is one of the main advantages of the free-knot approach, which is able to change the smoothing applied by taking into account the curvature of the regression function. As shown by Fan and Gijbels (1995), an adaptive bandwidth can be obtained by an adaptive polynomial degree and by using BMA we introduce this adaptive behavior in our procedure. Then, the use of the BIC approximation keeps its simplicity and low computational cost and guarantees that if the true model is a polynomial model of degree smaller than d, then for large sample size the true model will be used. Also, when the true model is not a polynomial, the use of BMA allows us to build credible intervals which take into account the uncertainty about the model and lead to better predictive capability than those built using a single model (Madigan and Raftery, 1994). The use of mixtures of normals for robust estimation has several advantages over classical methods. See for instance Denison et al. (2002). First, we can take into account the leverage of the observations in addition to the residual sizes and avoid the limitation of M estimation in regression. Second, we obtain posterior probabilities for the suspicious observations to be outliers. These mixture estimation methods in general require a high computational cost but we will show in Section 3.1 that, in this problem, we can take into account the information obtained in one local window to simplify the computations for the next window, making the procedure fast and efficient.

The rest of the paper is organized as follows. Section 2 describes the proposed method and presents its main properties. Section 3 develops the modification of the method to make it robust to outliers. Section 4 analyzes some real data sets to illustrate the behavior of the procedure and provides a Monte Carlo comparison with other methods using several simulated benchmark examples proposed in the literature. Finally, Section 5 presents some concluding remarks.

2. The proposed method

Suppose that we have *n* observations (x_i, y_i) which are a sample of independent and identically distributed data from a random variable (X, Y). We assume that these observations are related by

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, and *X* and ε are independent. Further, we suppose that $m(\cdot)$ is a smooth function. It is well known that the family of polynomials of degree smaller than *d*, for *d* large enough, can capture the local structure of any curve. Given a value of *d*, to be discussed below, we consider local polynomial models M_J of the form

$$y_i = \sum_{j=0}^{J} \beta_{Jj}(x_i - \overline{x})^j + \varepsilon_i, \quad J = 0, \ldots, d, \tag{2}$$

for some neighborhood of (x_i, y_i), where $\overline{x}$ is the mean of the *x* variable in the neighborhood. Note that in order to simplify the notation, we write β_{Jj} instead of $\beta_{Jj}^{(i)}$, as the regression parameters are going to depend on the neighborhood. To define this neighborhood, suppose that the *x* observations are all different and ordered, that is, $x_1 < x_2 < \cdots < x_n$, (if they were not different we define the neighborhood over the set of different observations of *x*). Then, for a given observation x_i, the neighborhood around this point is defined by

$$SNN(x_i, w) = \{x_k : x_{i-w} \leqslant x_k \leqslant x_{i+w}\},$$

where *w* is the bandwidth of the window. The number of observations in the window is at least $2w + 1$. We assume that *w* is chosen so that the number of different values of x_k in $SNN(x_i, w)$ is at least $d + 1$, so that the polynomial of degree *d* can be fitted using the data in the window. To take into account the left and right endpoints, where the windows contain fewer observations, we redefined the first and the last windows as $SNN(x_i, w) = \{x_k : x_{\max(1,i-w)} \leqslant x_k \leqslant x_{\min(n,i+w)}\}$.

In this work we make all the inference for the predicted value of a future observation $y_{f_i} = m(x_i)$ corresponding to a given value x_i, although the same analysis can be applied for a new observation x_0 belonging to the range of the data, $x_0 \in (x_1, x_n)$, by defining $SNN(x_0, w) = SNN(x_i, w)$ where $x_i = \min_k \|x_k - x_0\|$.

The procedure is applied as follows. Let $D_i = \{(x_k, y_k) : x_k \in SNN(x_i, w)\}$, for each x_i in the sample, we locally approximate the general form $m(x_i)$ in D_i by a linear combination of the polynomials (2). Thus, using data D_i we compute the posterior probabilities for different polynomial degrees and then estimate $m(x_i)$ at each point by its forecast using BMA. The predictive distribution for a new observation, y_{f_i}, is given by

$$p(y_{f_i} \mid D_i) = \sum_{J=0}^{d} p_J \, p(y_{f_i} \mid D_i, M_J),$$

where $p_J = P(M_J \mid D_i)$ is the posterior probability for the polynomial model of degree J, M_J, given data D_i. The prediction under quadratic loss will be given by $\widehat{m}(x_i \mid D_i) = E(y_{f_i} \mid D_i)$, and we have that

$$\widehat{m}(x_i \mid D_i) = \sum_{J=0}^{d} p_J \widehat{m}(x_i \mid D_i, M_J), \tag{3}$$

where $\widehat{m}(x_i \mid D_i, M_J) = E(y_{f_i} \mid D_i, M_J)$ is the expected value of the predictive distribution conditional to model M_J.

To make the inference about the polynomial models (2), we consider a reference prior distribution by taking a priori the elements of $\boldsymbol{\beta}_J = (\beta_{J0}, \ldots, \beta_{JJ})'$ and σ_J independently and uniformly distributed,

$$p(\boldsymbol{\beta}_J, \sigma_J) \propto \frac{1}{\sigma_J}.$$

Then, the predictive distribution for a new observation, $p(y_{f_i} \mid D_i, M_J)$, is a t-Student distribution with $v = n_0 - (J+1)$ degrees of freedom, where n_0 is the sample size of $SNN(x_i)$, mean $E(y_{f_i} \mid D_i, M_J) = \mathbf{x}_i \widehat{\boldsymbol{\beta}}_J$, where $\widehat{\boldsymbol{\beta}}_J = (\widehat{\beta}_{J0}, \ldots, \widehat{\beta}_{JJ})'$ is the vector of usual least-squares estimators for the parameters of the polynomial of degree J for data in D_i, $\mathbf{x}_i = \left(1, (x_i - \overline{x}_i), \ldots, \ (x_i - \overline{x}_i)^J\right)$ and $\overline{x}_i = \left\{\sum x_k / n_0 : x_k \in SNN(x_i)\right\}$, and variance given by $Var(y_{f_i} \mid D_i, M_J) = \frac{v}{v-2} s_J^2 \left(1 + (x_i - \overline{x}_i)(\mathbf{X}_J' \mathbf{X}_J)^{-1}(x_i - \overline{x}_i)\right)$where vs_J^2 is the standard sum of the squared residuals and $\mathbf{X}_J$ is the design matrix of the polynomial model (2) of degree J fitted to the data in D_i. Note that this estimation is applied for each neighborhood, although, to simplify, we do not include this dependence in the notation.

The posterior probability for model M_J is approximated by the exponential of the BIC criterion, which, as Kass and Raftery (1995) pointed out, approximates the Bayes factor with a relative error $O(1)$. The Schwarz criterion (Schwarz, 1978) for M_J is defined as

$$S(M_J) = \log p\left(\mathbf{y} \mid \widehat{\boldsymbol{\beta}}_J\right) - \frac{1}{2}(J+1) \log n_0,$$

where $p\left(\mathbf{y} \mid \widehat{\boldsymbol{\beta}}_J\right)$ is the likelihood of the model M_J, $\widehat{\boldsymbol{\beta}}_J$ is the MLE of the parameter vector under model M_J for data in D_i, n_0 is the sample size of $SNN(x_i)$ as before and $(J+1)$ is the dimension of the vector $\widehat{\boldsymbol{\beta}}_J$. The Bayesian information criterion (BIC) of a model M_J is $\mathrm{BIC}(M_J) = -2S(M_J)$, and $\exp(S(M_{J_1}) - S(M_{J_2}))$ approximates the Bayes factor $B_{J_1 J_2}$

with a relative error $O(1)$. Thus, we can approximate the Bayes factors by

$$B_{J_1J_2}^{\mathrm{BIC}} = \exp(S(M_{J_1}) - S(M_{J_2})) = \frac{\exp(-0.5\mathrm{BIC}(M_{J_1}))}{\exp(-0.5\mathrm{BIC}(M_{J_2}))}$$

and obtain the posterior probability for a model by

$$p(M_J \mid D_i) \propto p(M_J)\left\{\log\, p(\mathbf{y} \mid \widehat{\boldsymbol{\beta}}_J) - \tfrac{1}{2}(J+1)\log n_0\right\},$$

where $p(M_J)$ is the prior probability for the polynomial model. The likelihood for a normal linear model evaluated at the MLE estimator is easily seen to be

$$p\left(\mathbf{y} \mid \widehat{\boldsymbol{\beta}}_J\right) = (2\pi)^{-n_0/2}\left(\frac{vs_J^2}{n_0}\right)^{-n_0/2} \exp\left\{-\frac{n_0}{2}\right\}$$

and the posterior probability of M_J may be approximated, after absorbing common constants, by $p(M_J \mid D_i) = K_{\mathrm{BIC}}\, p(M_J)(vs_J^2)^{-n_0/2} n_0^{-(J+1)/2}$, where K_{BIC} is obtained by the condition $\sum_{J=0}^{d} p(M_J \mid D_i) = 1$. Then we approximate the posterior probability of the models by

$$p(M_J \mid D_i) \propto s_J^{-n_0} n_0^{-(J+1)/2}. \tag{4}$$

Note that we are applying Bayesian inference locally, so that we do not assume a fixed joint likelihood function for the data, as it is standard in nonparametric statistics. From this point of view our approach can be seen as a Bayesian nonparametric approach in which the prior and the likelihood are specified locally. This provides a flexible approach in situations in which a global model would be very complicated to specify.

In order to apply this method several decisions must be made. First we have to decide about the maximum degree d of the polynomials to be fitted. We propose to take $d = 3$. We have found that this value is large enough to fit any curve locally very well and it avoids the problem of overfitting. Second, we have to decide on the a priori probabilities of the models. Two possible choices are uniform, $p(M_J) = (d+1)^{-1}$ or decreasing with the polynomial degree. We propose the uniform prior for simplicity. The third choice is the bandwidth parameter w. A classical solution is to choose this parameter by cross-validation as follows. Let $\widehat{y}_i^w$ be the estimated value of $m(x_i)$ with bandwidth w, where the observed value y_i is omitted in the estimation of $\widehat{y}_i^w$. Then, the value for w is chosen to minimize the mean squared error

$$\mathrm{MSE}_w = \frac{1}{n}\sum_{i=1}^{n}\left(m(x_i) - \widehat{y}_i^w\right)^2.$$

We have checked by simulation that the results are not very sensitive to the choice of the parameter w. This fact can be explained by the work of Fan and Gijbels (1995). They proposed a method which replaces an adaptive bandwidth by an adaptive order of the polynomial to be fitted, and observed that if the bandwidth parameter is large, then the order chosen for the polynomial order is high, whereas when a small bandwidth is used the order chosen was low. This same effect has been observed in the proposed method, and this compensation effect makes the procedure fairly robust to the bandwidth parameter chosen.

The consistency of the proposed method can be obtained from the consistency of the polynomial model approach. Also, we can obtain the expressions of the bias and the variance based on the Theorem 3.1, page 62, in Fan and Gijbels (1996)

$$\begin{aligned} E[\widehat{m}(x) - m(x) \mid \mathbf{X}] &= E\left[\left\{\sum_{i=0}^{3} p_i \widehat{m}_i(x)\right\} - m(x) \mid \mathbf{X}\right] \\ &= \left(\frac{p_0 + p_1}{2}\right)\frac{w^2}{3}m''(x) + \left(\frac{p_2 + p_3}{2}\right)\frac{w^4}{140}m^{iv}(x) + o_p(w^4), \\ Var[\widehat{m}(x)] &= Var\left[\sum_{i=0}^{3} p_i \widehat{m}_i(x)\right] \\ &= \left\{\left(\frac{p_0^2 + p_1^2}{2}\right) + 9\left(\frac{p_2^2 + p_3^2}{2}\right)\right\}\left(\frac{\sigma^2}{c}\frac{1}{nw}\right) + o_p\left(\frac{1}{nw}\right), \end{aligned}$$

where σ^2 is the residual variance, p_J are the posterior probability of the polynomials models, w is the bandwidth, n is the sample size and $m^i(x)$ indicates the ith derivative of the $m(x)$ function. We are supposing that the marginal density of the observations x, $f(x)$, is uniform over the range of the data, $f(x) = c$ and $f'(x) = 0$.

In order to have a smoother curve the procedure can be iterated. As iterating the procedure does not have a clear Bayesian justification, this stage can be skipped, although we have found in practice (see Section 4) that it often leads to better results. The iteration can be carried out as follows. Let $\widetilde{y}^{(1)}$ by the predicted value obtained by (3) in the first application of the procedure or the first iteration. Then the observed values (x, y) are replaced by the output of the first iteration, $(x, \widetilde{y}^{(1)})$, and the procedure is applied to this modified data set to obtain $\widetilde{y}^{(2)}$, which is the output in the second iteration. In the same way the values in the kth iteration, $\widetilde{y}^{(k)}$, can be computed by using the output of the $(k-1)$th iteration $(x, \widetilde{y}^{(k-1)})$ as input data. In practice, we have found that a single iteration is enough to produce a good result.

A possible problem when applying this procedure is that a single outlier observation can have a large effect on the estimated models. In the next section we propose a robustified version of the method.

3. Robustifying the method

The method can be made robust to reduce the influence of the outliers in the local estimation by modelling the residuals by a mixture of normals. This model was introduced by Tukey (1960) and studied by Box and Tiao (1968). Suppose that observations **y** are generated by the model (1), where now the errors ε_i are random variables with the mixture distribution

$$\varepsilon_i \sim (1-\alpha)N\left(0, \sigma^2\right) + \alpha N\left(0, k^2\sigma^2\right),$$

where α is the prior probability that one observation comes from the $N(0, k^2\sigma^2)$ distribution and $k > 1$. To make inference about this model, a set of dummy variables δ_i are defined by $\delta_i = 1$ if $Var(\varepsilon_i) = k^2\sigma^2$ and $\delta_i = 0$ otherwise. Let $\mathbf{\Delta}_k = (\delta_1 = l_1, \ldots, \delta_n = l_n)$ be a possible configuration of the data, where $l_i = 0, 1$. Then there are 2^n possible classifications of the observations into the two components of the mixture. Let $\mathbf{V}$ be a diagonal matrix with elements (i, i), v_{ii} equal to 1 if $\delta_i = 0$ and $v_{ii} = 1/k^2$ if $\delta_i = 1$. Then, by making the transformation $\mathbf{Y}_h = \mathbf{V}^{1/2}\mathbf{Y}$, $\mathbf{X}_h = \mathbf{V}^{1/2}\mathbf{X}$, standard inference results for linear models can be applied. The BMA predictive distribution for the future observation y_{f_i} given the data D_i, will be given by

$$p(y_{f_i} \mid D_i) = \sum_{h=0}^{2^n} \sum_{J=0}^{d} p(y_{f_i} \mid D_i, M_J, \mathbf{\Delta}_h) p_{Jh}, \tag{5}$$

which is a mixture of $(d + 1) \times 2^n$ distributions, $p(y_{f_i} \mid D_i, M_J, \mathbf{\Delta}_h)$, where the weights, for each model and each configuration of the data, are given by $p_{Jh} = p(M_J \mid \mathbf{\Delta}_h, D_i) p(\mathbf{\Delta}_h \mid D_i)$. We compute the predicted value $\widehat{m}(x_i \mid D_i)$ as the expected value of the predictive distribution $p(y_{f_i} \mid D_i)$,

$$\widehat{m}(x_i \mid D_i) = \sum_{J=0}^{d} \sum_{h=0}^{2^n} p_{Jh} \widehat{m}(x_i \mid D_i, M_J, \mathbf{\Delta}_h).$$

Given the model and the configuration, the predictive distribution $p(y_f \mid D_f, M_J, \mathbf{\Delta}_h)$ for a new observation x_f, is a *t*-student distribution $t(v, \mathbf{x}_f \widehat{\boldsymbol{\beta}}_{Jh}, h)$ with $v = n - (J + 1)$ degrees of freedom. The expected values $\widehat{m}(\mathbf{x}_f \mid D_f, M_J, \mathbf{\Delta}_h) = \mathbf{x}_f \widehat{\boldsymbol{\beta}}_{Jh}$ is the mean of the distribution, $\mathbf{x}_f = \left(1, (x_f - \overline{x}_f), \ldots, (x_f - \overline{x}_f)^J\right)$, $\overline{x}_f = \{\sum x_k/n_0 : x_k \in SNN(x_f)\}$, and $\widehat{\boldsymbol{\beta}}_{Jh}$ are the estimated parameters given the $\mathbf{\Delta}_h$ configuration and the model M_J,

$$\widehat{\boldsymbol{\beta}}_{Jh} = (\mathbf{X}'_{Jh}\mathbf{X}_{Jh})^{-1}\mathbf{X}'_{Jh}\mathbf{Y}_h = (\mathbf{X}'_J\mathbf{V}\mathbf{X}_J)^{-1}\mathbf{X}'_J\mathbf{V}\mathbf{Y}$$

and the variance of the predictive distribution is

$$\frac{v}{v-2}\widehat{s}^2_{Jh}\left(1 + (x_f - \overline{x}_f)(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}(x_f - \overline{x}_f)\right),$$

where

$$v\widehat{s}^2_{Jh} = \left(\mathbf{Y}_h - \mathbf{X}_{Jh}\widehat{\boldsymbol{\beta}}_h\right)'\left(\mathbf{Y}_h - \mathbf{X}_{Jh}\widehat{\boldsymbol{\beta}}_h\right) = \mathbf{Y}'[\mathbf{V} - \mathbf{V}\mathbf{X}_J\left(\mathbf{X}'_J\mathbf{V}\mathbf{X}_J\right)^{-1}\mathbf{X}'_J\mathbf{V}]\mathbf{Y}$$

is the standard sum of the squared residuals.

The weights of the mixture are given by $p_{Jh} = p(M_J \mid \mathbf{\Delta}_h, D_f) p(\mathbf{\Delta}_h \mid D_f)$, where the first term, $p(M_J \mid \mathbf{\Delta}_h, D_f)$, is the posterior probability of the models given a configuration $\mathbf{\Delta}_h$. We approximate this term by the exponential of the BIC, given by (4), where $\widehat{s}^2_J$ is replaced by $\widehat{s}^2_{Jh}$ which depends on the model and on the configuration. The integration constant is computed by using the restriction that the sum of the posterior probabilities of the four polynomials models, given each configuration, $\mathbf{\Delta}_h$, is one.

The second term for the weights, is computed by

$$p(\boldsymbol{\Delta}_h \mid \mathbf{y}) = K_2 p(\mathbf{y} \mid \boldsymbol{\Delta}_h) p(\boldsymbol{\Delta}_h) = K_2 \sum_{J=0}^{d} p(\mathbf{y} \mid \boldsymbol{\Delta}_h, M_J) p(\boldsymbol{\Delta}_h \mid M_J) p(M_J),$$

where $p(\mathbf{y} \mid \boldsymbol{\Delta}_h, M_J)$ is the marginal distribution of the data, given a model M_J and a configuration $\boldsymbol{\Delta}_h$, $p(\boldsymbol{\Delta}_h \mid M_J)$ is the prior probability of a configuration, which does not depend on the model M_J, $p(\boldsymbol{\Delta}_h \mid M_J) = p(\boldsymbol{\Delta}_h) = \alpha^{n_h}(1-\alpha)^{n-n_h}$ and n_h is the number of elements with high variance in the configuration $\boldsymbol{\Delta}_h$, $n_h = \sum \delta_i$. Finally, $p(M_J)$, the prior probabilities, are equal for all the models and this term is absorbed by the integration constant.

In order to compute the marginal density, $p(\mathbf{y} \mid \boldsymbol{\Delta}_h, M_J)$ the likelihood of the model for the parameters $\boldsymbol{\theta}_J = (\boldsymbol{\beta}_J, \sigma_J)$ can be written as

$$\begin{aligned}
& f(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}_J, M_J, \boldsymbol{\Delta}_h) \\
&\quad = (2\pi)^{-n_h/2} \sigma_{Jh}^{-n_h} k^{-n_h} \exp\left\{ -\frac{1}{2\sigma_{Jh}^2 k^2} \left(\mathbf{Y}_{n_h} - \mathbf{X}_{Jn_h}\boldsymbol{\beta}_J\right)' \left(\mathbf{Y}_{n_h} - \mathbf{X}_{Jn_h}\boldsymbol{\beta}_J\right) \right\} \\
&\qquad \times (2\pi)^{-(n-n_h)/2} \sigma_{Jh}^{-(n-n_h)} \\
&\qquad \times \exp\left\{ -\frac{1}{2\sigma_{Jh}^2} \left(\mathbf{Y}_{(n-n_h)} - \mathbf{X}_{J(n-n_h)}\boldsymbol{\beta}_J\right)' \left(\mathbf{Y}_{(n-n_h)} - \mathbf{X}_{J(n-n_h)}\boldsymbol{\beta}_J\right) \right\} \\
&\quad = (2\pi)^{-n/2} \sigma_{Jh}^{-n} k^{-n_h} \\
&\qquad \times \exp\left\{ -\frac{1}{2\sigma_{Jh}^2 k^2} \left(\mathbf{V}^{1/2}\mathbf{Y} - \mathbf{V}^{1/2}\mathbf{X}_J\boldsymbol{\beta}_J\right)' \left(\mathbf{V}^{1/2}\mathbf{Y} - \mathbf{V}^{1/2}\mathbf{X}_J\boldsymbol{\beta}_J\right) \right\},
\end{aligned}$$

where $\mathbf{X}_{Jn_h}$ indicates the rows of $\mathbf{X}_J$ corresponding to the observations with variance $k^2\sigma_{Jh}^2$, and $\mathbf{X}_{J(n-n_h)}$ those corresponding to observations with variance σ_{Jh}^2. The marginal density is obtained by integrating $\boldsymbol{\theta}_J$, $p(\mathbf{y} \mid \boldsymbol{\Delta}_h, M_J) \propto \left(\widehat{s}_{Jh}^2\right)^{-(n-J+1)/2} \mid \mathbf{X}_J'\mathbf{V}\mathbf{X}_J \mid^{-1/2}$ and we can obtain the expression for the marginal probability of the configuration,

$$\begin{aligned}
p(\boldsymbol{\Delta}_h \mid \mathbf{y}) &= K_2 \sum_{J=0}^{d} p(\mathbf{y} \mid \boldsymbol{\Delta}_h, M_J) p(\boldsymbol{\Delta}_h) \\
&= K_2 \sum_{J=0}^{d} \left(\widehat{s}_{Jh}^2\right)^{-(n-J+1)/2} \mid \mathbf{X}_J'\mathbf{V}\mathbf{X}_J \mid^{-1/2} \alpha^{n_h}(1-\alpha)^{n-n_h},
\end{aligned}$$

where the constant K_2 is computed by using the condition $\sum_{h=0}^{2^n} p(\boldsymbol{\Delta}_h \mid \mathbf{y}) = 1$.

3.1. Implementation

The scale contaminated normal model has the problem that the inference is made over the 2^n possible configurations of the data and it requires intensive computation. Although in our case we have many local estimation problems with small sample size, the number of computations grows in exponential form. For example, for window size $n_0 = 20$, the procedure requires computing approximately 10^6 posterior probabilities for the models, for each one of the $n - n_0$ windows.

The problem has been solved in the literature using the Gibbs sampler, (see Verdinelli and Wasserman, 1991 and Justel and Peña, 1996) but the local character of the estimation implies solving approximately $n - n_0$ local problems which requires intensive computation. Note that in this problem we may take advantage of the fact that the inference in a given window gives us information about the inference in the next window, because they will only differ in a few observations. Suppose we have computed the posterior probabilities for all the configurations of the data corresponding to the set of observations belonging to window D_i. The next window, D_{i+1}, is obtained from the previous one by deleting some observations in the left extreme of D_i and adding some new observations in the right hand of the D_{i+1}. Thus, we can obtain the configurations with highest probability in the first windows and use this information to obtain the configurations with highest probabilities in the next window, D_{i+1}.

For this first window, if the sample size is small enough, the simplest solution is to carry out an exhaustive study of the configurations. Otherwise, an alternative fast method which allows an automatic implementation was proposed by Peña and Tiao (1992). Suppose that we have a sample of size n and that we can classify the observations in two groups. The first includes n_1 observations of potential outliers and the second the remaining $n_2 = n - n_1$ observations which we believe have a high probability of not being an outlier. Then, as

$$\binom{n}{h} = \sum_{j=0}^{h} \binom{n_1}{j} \binom{n_2}{h-j} = \binom{n_1}{h} + \sum_{j=0}^{h-1} \binom{n_1}{j} \binom{n_2}{h-j}$$

instead of studying all the combinations of h outliers out of n we can compute all the combinations of h outliers out of the n_1 potential set of outliers and a sample of the combinations which include $j = 1, 2, \ldots, h-1$ outliers and a small sample of all the combinations of h points out of n_2. In order to do so we need to divide the observations in these two groups. Peña and Tiao (1992) proposed studying the differences between the probabilities $P\left(A_i A_j\right)$, and $P\left(A_i\right) P\left(A_j\right)$, where A_i is the event that x_i is an outlier, and consider as potential outliers those observations in which both probabilities were different.

To apply this idea to the problem, the set of potential outliers is identified as follows:

(1) Compute the posterior probabilities for all the configurations which have the number of outliers less or equal to 2. Let $\mathbf{\Delta}_0$ be the configuration without outliers, $\mathbf{\Delta}_i$ the configuration with only one outlier, the observation x_i and $\mathbf{\Delta}_{ij}$ the configuration in which only the elements (x_i, x_j) are outliers.

(2) Include in the set of potential outliers the isolated outliers defined by the set $A = \left\{ x_i : \frac{P(\mathbf{\Delta}_i \mid D)}{P(\mathbf{\Delta}_0 \mid D)} \geqslant 3 \right\}$.

(3) Include also the partially masked outliers as those belonging to the set $B = \left\{ x_j : \frac{P(\mathbf{\Delta}_{i,j} \mid D)}{P(\mathbf{\Delta}_i \mid D)} \geqslant 3, \ x_i \in A \right\}$.

(4) Include also the completely masked outliers, defined by the elements of the set $C = \left\{(x_i, x_j) : \frac{P(\boldsymbol{\Delta}_{i,j} \mid D)}{P(\boldsymbol{\Delta}_0 \mid D)} \geqslant 3,\ (x_i, x_j) \notin (A \cup B)\right\}$.

The set of potential outliers is formed by elements belonging to $(A \cup B \cup C)$.

Once the configurations of outliers and good points with highest probability are detected for the first window, D_1, we use this information to select the configurations in the next window, D_2. In the same way we use the information of D_i to select the configurations of D_{i+1} in a recursive form. In order to do so we introduce some notation: let $LD_i = D_i \backslash D_{i+1}$, the left part of D_i, the set of observations belonging to D_i which do not belong to D_{i+1}, m_i^L the cardinal of LD_i, similarly let RD_i the right part of D_i and m_i^R the cardinal of RD_i.

Suppose that we have the posterior probabilities $p\left(\boldsymbol{\Delta}_h^i \mid \mathbf{y}\right)$ for all the configurations in the window D_i which do not have negligible probability. We select the set of M configurations $\nabla_{D_i} = \left\{\boldsymbol{\Delta}_1^i, \ldots, \boldsymbol{\Delta}_M^i\right\}$ with highest posterior probability. Now, we move to the next window, D_{i+1}, and let $\nabla_{RD_i} = \left\{\boldsymbol{\Delta}_1^R, \ldots, \boldsymbol{\Delta}_{2^{m_i^R}}^R\right\}$ be the $2^{m_i^R}$ possible configurations for the m_i^R new observations with are incorporated in D_{i+1}. In addition we have to delete from ∇_{D_i} the terms corresponding to the observations which are not in D_{i+1}. Let $\boldsymbol{\Delta}_k^{*i}$ be the configuration obtained from $\boldsymbol{\Delta}_k^i \in \nabla_{D_i}$ by deleting the first m_i^L terms. Then, the configurations with highest probabilities in the next window D_{i+1} will belong to the set $\left\{\left[\boldsymbol{\Delta}_1^{*i} \cup \boldsymbol{\Delta}_1^R\right], \ldots, \left[\boldsymbol{\Delta}_1^{*i} \cup \boldsymbol{\Delta}_{2^{m_i^R}}^R\right], \ldots, \left[\boldsymbol{\Delta}_M^{*i} \cup \boldsymbol{\Delta}_1^R\right], \ldots, \left[\boldsymbol{\Delta}_M^{*i} \cup \boldsymbol{\Delta}_{2^{m_i^R}}^R\right]\right\}$ where $\left[\boldsymbol{\Delta}_k^{*i} \cup \boldsymbol{\Delta}_l^R\right]$ represents the $\boldsymbol{\Delta}_j^{*i}$ configuration for the observations which belong to D_i and the configuration $\boldsymbol{\Delta}_l^R$ for the new observation incorporated. If there are not repeated observations in the data set and $m_i^R = 1$, for all the windows D_i, then we can choose M big enough to guarantee that the best configurations are selected. In data sets with repeated observations, M should be chosen moderate to avoid expensive computations.

4. Examples

To illustrate the methods developed, we consider three data set frequently analyzed in the nonparametric curve fitting. The first one is the Helmets data. The data consists of accelerometer readings taken through time in an experiment on the efficacy of crash helmets in simulated motor-cycle crashes, described in detail by Schmidt et al. (1981). The second one is the Ethanol data. The data includes 88 measurements of two variables from an experiment in which ethanol was burned in a single cylinder automobile test engine (Brinkman, 1981). The two variables measured are the concentration of nitric oxide (NO) and nitrogen dioxide (NO_2) in engine exhaust and the equivalence ratio at which the engine was run (a measure of the richness of the air-ethanol mix). The third example is the Diabetes data. It includes two variables measured in children with insulin-dependent diabetes. The variables are the age of the children and the level of serum C-peptide, and were obtained from Sockett et al. (1987). We have analyzed the same subset of 43 observations that appear in Hastie and Tibshirani (1990) which use this data to show the effect of several smoothers in Chapter 2 of their book.

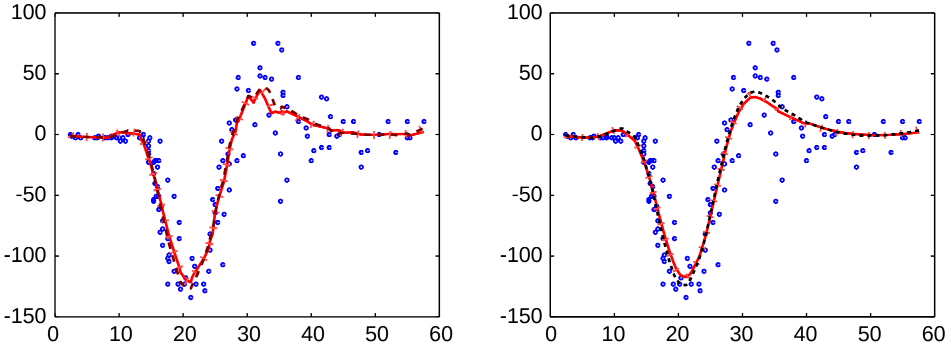

Fig. 1. Curves fit for Helmets data. The left figure shows the curve for the standard method (solid line), the robust method with parameters $(\alpha = 0.05, k^2 = 3)$ (dotted line) and the robust method with $(\alpha = 0.1, k^2 = 5)$ (dashed line). The right figure shows the second iteration of the procedure for these three cases.

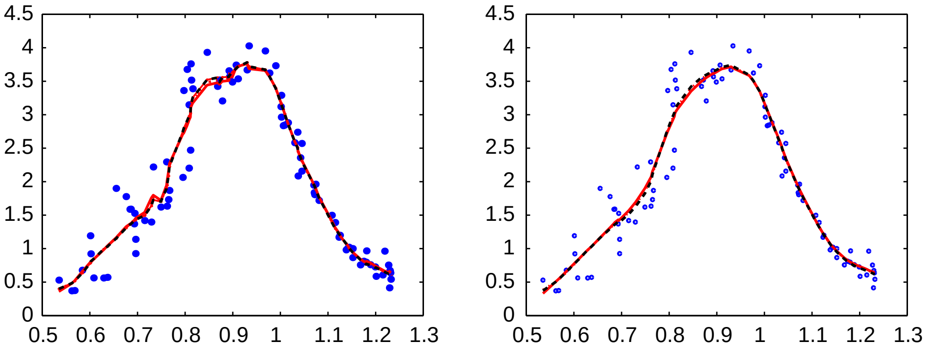

Fig. 2. Curves fit for Ethanol data. The left figure shows the curve fitted by the standard method (solid line), the robust method with parameters $(\alpha = 0.05, k^2 = 3)$ (dotted line) and the robust method with $(\alpha = 0.1, k^2 = 5)$ (dashed line). The right figure shows the second iteration of the procedure for these three cases.

Fig. 1 shows the estimated curve for the Helmets data, with $w = 12$ estimated by cross validation. The figure in the left-hand side shows the estimated curve with the procedure presented in Section 2 and two robust curve estimates with parameters $(\alpha = 0.01, k^2 = 3)$ and $(\alpha = 0.1, k^2 = 5)$. It can be seen that the smoothness of the curve increases with the prior proportion of outliers. On the right hand a second iteration for each of these three cases are shown and it can be seen that these curves are very smooth and the differences among them are very small.

Fig. 2 shows the estimated curve for the Ethanol data. In this data set the value of the parameter w obtained by minimizing the MSE for cross validation is $w = 10$. The three curves shown are the ones obtained by the standard estimation and two obtained by a robust approach with the same values of the parameters as in the previous example $(\alpha = 0.01, k^2 = 3)$ and $(\alpha = 0.1, k^2 = 5)$. We can observe that there are small differences among the three

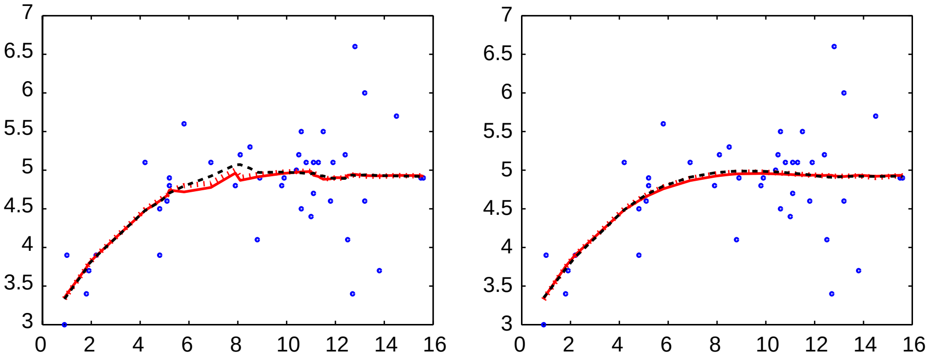

Fig. 3. Curves fit for diabetes data. The left figure shows the curve fitted by the standard method (solid line), the robust method with parameters ($\alpha = 0.01$, $k^2 = 3$) (dotted line) and robust method with ($\alpha = 0.05$, $k^2 = 7$) (dashed line). The right figure shows the second iteration of the procedure for these three cases.

curves and none of them is completely smooth. Note that as the data is homogeneous the robustification does not modify the standard estimation. In the right hand figure we show the second iteration of the procedure in the three cases, the three curves obtained are smooth and very similar.

Fig. 3 shows the fitting curve for the diabetes data in the first two iterations of the algorithm. The window which minimizes the MSE for cross validation is now $w = 22$, and the sample size is 43. The lack of smoothness observed in the curve fitted by the standard procedure corresponds to the incorporation of the extreme observations around $x_i = 13$. The robust estimate of the curve reduces this effect. Apart from the variability at this point there are small differences among the fitted curves due to the large window used. The second iteration of the procedures leads to similar fitted curves.

4.1. Monte Carlo experiment

We compare the behavior of the proposed method to the popular loess method of Cleveland (1979) which is implemented in many computer programs, and to the Bayesian free-knot splines approach by DiMatteo et al. (2001) as implemented in the BARS code which can be downloaded from http:// www.stat.cmu.edu/~jliebner/. The comparison is made by using four simulated functions proposed by Donoho and Johnstone (1994) which have been used often in the literature for comparison purposes (see Denison et al., 1998). The four simulated functions are:

$$\begin{aligned}
&\text{Heavisine} \quad f(x) = [4\sin(4\pi x) - sgn(x - 0.3) + \varepsilon_3 - sgn(0.72 - x)],\\
&\text{Blocks} \quad f(x) = \sum h_j^{(2)} K(x - x_j) + \varepsilon_4 \quad K(x) = (1 + sgn(x))/2,\\
&\text{Bumps} \quad f(x) = \sum h_j^{(1)} K((x - x_j)/w_j) + \varepsilon_5 \quad K(x) = (1 + |x|)^{-4},\\
&\text{Doppler} \quad f(x) = \sqrt{x(1 - x)}\sin(2.1\pi/(x + 0.05)) + \varepsilon_6,
\end{aligned}$$

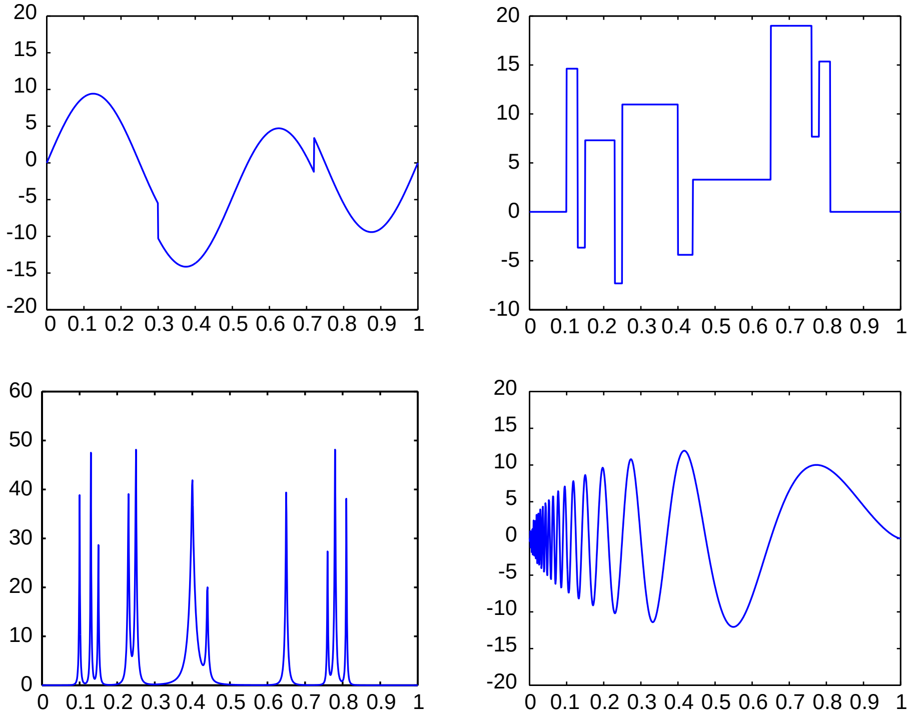

Fig. 4. The four simulated functions used to compare the proposed method: Heavisine, Blocks, Bumps and Doppler.

where $x_j = \{0.1, 0.13, 0.15, 0.23, 0.25, 0.4, 0.44, 0.65, 0.76, 0.78, 0.81\}$, $h_j^{(1)} = \{4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4.2\}$, $h_j^{(2)} = \{4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 5.1, -4.2\}$ and $w_j = \{0.005, 0.005, 0.006, 0.01, 0.01, 0.03, 0.01, 0.01, 0.005, 0.008, 0.005\}$. These functions are standardized to $Var(f) = 7^2$. The errors are generated by $\varepsilon_i \sim N(0, \sigma^2)$, where σ^2 is chosen so that the root of the signal-noise ratio $\left(\mathrm{RSNR} = \sqrt{\frac{var(f)}{\sigma^2}}\right)$ are 3, 5, 7 and 10. The simulation are based on 1000 points. The four simulated functions are showed in Fig. 4.

In the tables the mean of the squared errors, $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{m}(x_i) - m(x_i))^2$, is presented with $\widehat{m}(x_i)$ computed by eight different procedures. BMA1 and BMA2 both use the method proposed in Sections 2 and 3 of this paper with 1 and 2 iterations, respectively. This iterations are made as explained in Section 2, that is, the predicted value obtained in the first application of the procedure is used as data in the second application of the procedure. LB1, LB3, LT1 and LT3 use the loess method as proposed by Cleveland (1979). LB1 and LB3 with a bisquare weight function, $B(x) = \left(1 - x^2\right)^2$ for $|x| < 1$, and polynomial of degrees $d = 1$ or $d = 3$, respectively and LT1 and LT3 with the tricube kernel, $T(x) = \left(1 - |x|^3\right)^3$, and again degrees 1 and 3, respectively. In both kernels x is rescaled by $(x - x_i)/h_i$ where h_i is the distance $|x - x_i|$ from x to the rth nearest neighbor. Finally, BARS is the approach of DiMatteo et al. (2001). We also include a column with the number of knots (mean and

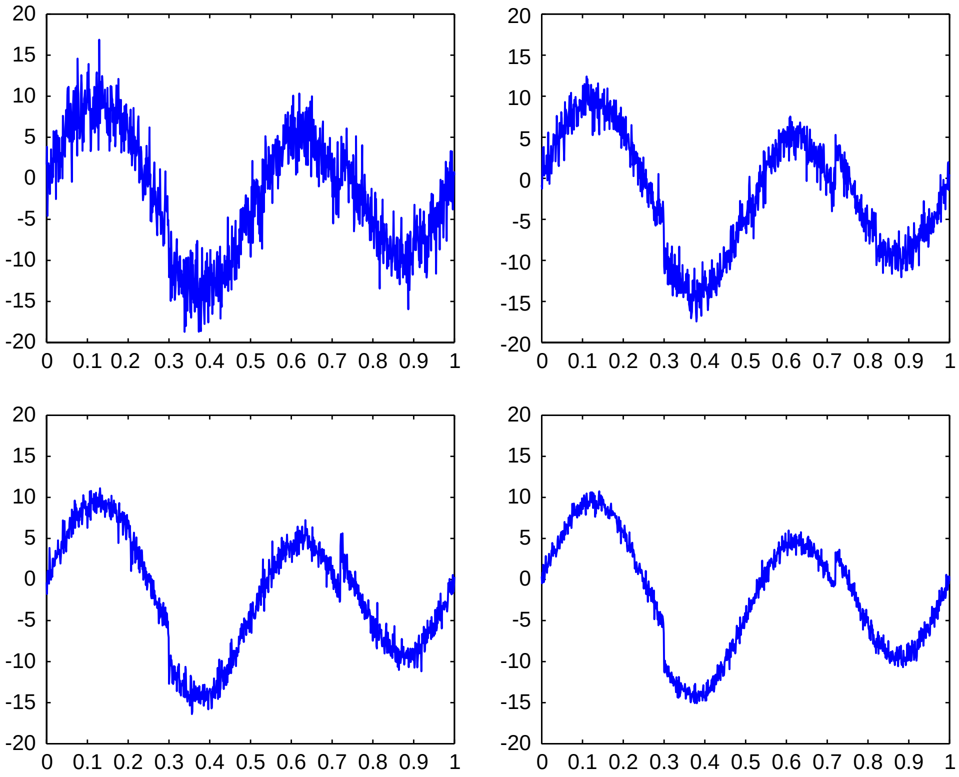

Fig. 5. Heavisine function with different root of signal-noise ratio: 3, 5, 7, 10.

standard error) used to fit the curve by this last procedure. The results are the mean of 1000 replications. The simulated curves with the four levels of root signal-to-noise ratio, RSNR = {3, 5, 7, 10}, are shown in Figs. 5–8.

Table 1 shows the mean and the standard deviation, in small letter size, of the MSE of the 1000 replications of the function Heavisine. We can observe that when the ratio signal-to-noise is small, RSNR = 3, the smallest MSE is obtained by BMA2, the proposed method with two iterations of the algorithm, but when this ratio increases the best performance is obtained by BARS, which uses between 9 and 13.5 knots to fit the curve. The number of knots grows when the RSNR grows. With regards to the loess method, the bisquare kernel is slightly better than the tricube, and the linear fit works better than the cubic fit.

Table 2 shows the result obtained for the function Blocks. The results are similar to the previous ones. When the signal-to-noise ratio is small, RSNR = 3, BMA2, the second iteration of the proposed algorithm, has the best performance. However, when this ratio increases the best results are obtained by BARS. Again for the loess procedure the linear fit is better than the cubic and the bisquare kernel slightly better than the tricube. We also observe that although the Blocks function presents eleven discontinuity points, and it is constant between them, the BARS procedure uses 50 knots for the highest level of RSNR,

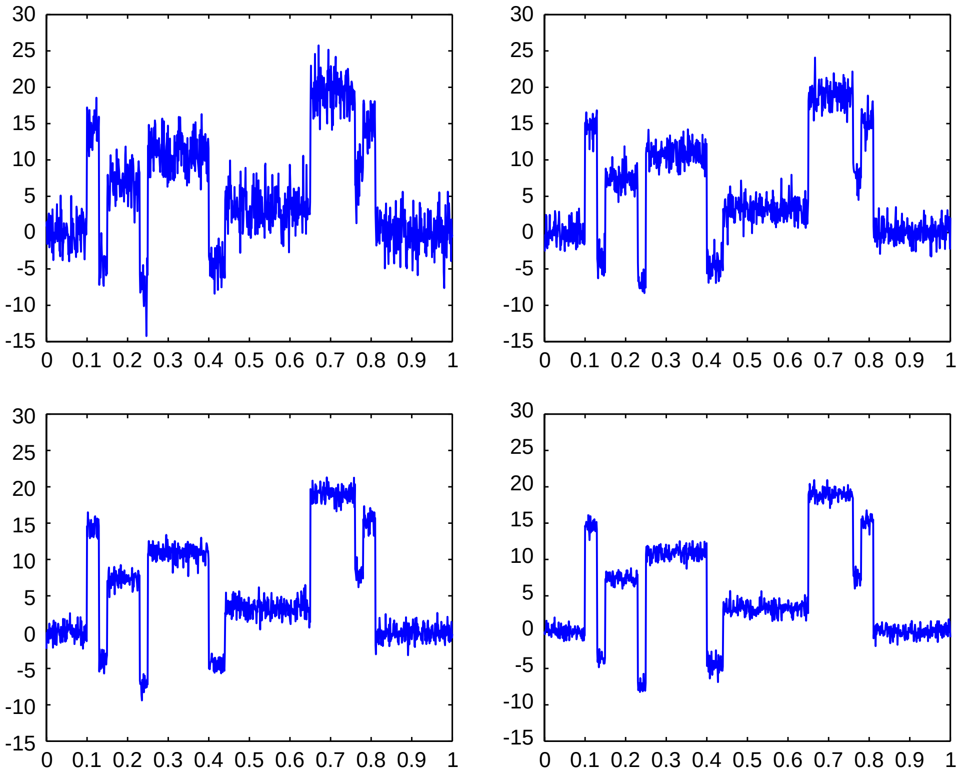

Fig. 6. Blocks function with different root of signal-noise ratio: 3, 5, 7, 10.

Table 1
MSE obtained for the Heavisine data with four different signal-to-noise ratio

RSNR	BMA1	BMA2	LB1	LB3	LT1	LT3	BARS	#Knots
3	0.2869 0.0445	0.2634 0.0443	0.2709 0.0437	0.3690 0.0544	0.2745 0.0436	0.3739 0.0548	0.3210 0.0821	9.09 1.06
5	0.1566 0.0183	0.1458 0.0173	0.1629 0.0173	0.2264 0.0409	0.1653 0.0172	0.2249 0.0371	0.0992 0.0267	12.05 0.78
7	0.1075 0.0108	0.1016 0.0095	0.1137 0.0092	0.1411 0.0160	0.1160 0.0093	0.1411 0.0143	0.0449 0.0153	13.06 0.45
10	0.0748 0.0060	0.0707 0.0051	0.0791 0.0050	0.1279 0.0093	0.0809 0.0050	0.1464 0.0169	0.0224 0.0064	13.49 0.58

The procedures compared are the proposed BMA with 1 and 2 iterations (BMA1 and BMA2), four implementation of loess (LB1, LB2, LT1, LT3) and the free-knots splines BARS.

RSNR = 10. For the functions Bumps and Doppler (see Tables 3 and 4) the results are different as now in all cases the best results are obtained by the BARS procedure. In these cases, the BMA1, the first iteration of the algorithm presents better results than BMA2. This is not surprising as these functions are not smooth and a second iteration smooths the picks,

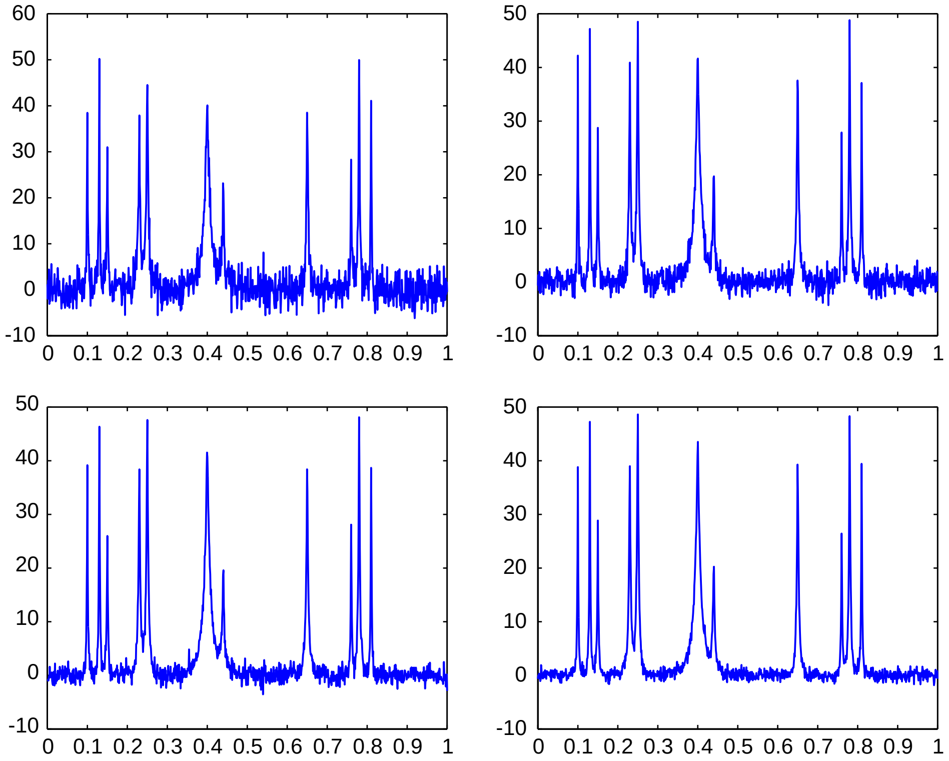

Fig. 7. Bumps function with different root of signal-noise ratio: 3, 5, 7, 10.

Table 2
MSE obtained for the Blocks data with four different signal-to-noise ratio

RSNR	BMA1	BMA2	LB1	LB3	LT1	LT3	BARS	#Knots
3	2.0494 0.0907	1.9042 0.0808	2.0050 0.0811	2.2307 0.0767	2.0674 0.0811	2.2730 0.0769	2.3972 0.3352	26.41 2.35
5	1.6817 0.0509	1.5643 0.0379	1.6366 0.0376	1.8625 0.0324	1.7019 0.0377	1.9003 0.0325	1.3003 0.4955	34.06 5.88
7	1.5821 0.0356	1.4763 0.0232	1.5405 0.0237	1.7673 0.0196	1.6068 0.0238	1.8043 0.0199	0.6827 0.6269	43.31 9.43
10	1.5271 0.0251	1.4278 0.0151	1.4879 0.0155	1.7148 0.0117	1.5548 0.0155	1.7512 0.0116	0.3791 0.5820	51.07 10.51

The procedures compared are the proposed BMA with 1 and 2 iterations (BMA1 and BMA2), four implementation of loess (LB1, LB2, LT1, LT3) and the free-knots splines BARS.

for the bumps data, or the extremes, for the Doppler data. With regards to loess the results are the same as before: the linear fit with the bisquare kernel has the best performance. Also, we can observe that for the wiggly curve (see Tables 2 and 3), the BARS method uses a large quantity of knots, and this may be the reason for the huge standard deviation of the estimated MSE for the blocks and bumps functions.

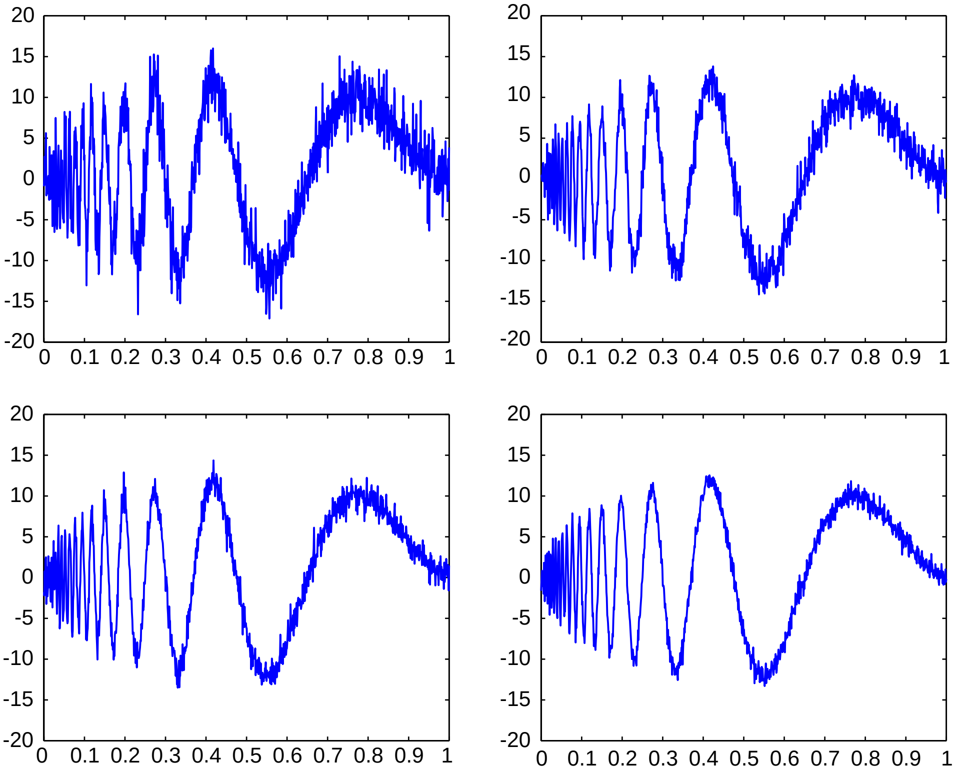

Fig. 8. Doppler function with different root of signal-noise ratio: 3, 5, 7, 10.

Table 3
MSE obtained for the Bumps data with four different signal-to-noise ratio

RSNR	BMA1	BMA2	LB1	LB3	LT1	LT3	BARS	#Knots
3	6.6577 0.2094	6.7748 0.1491	6.8940 0.1297	8.3255 0.1136	7.2484 0.1288	8.5622 0.1144	4.2794 4.5466	49.93 14.20
5	6.2017 0.1294	6.3904 0.0862	6.5223 0.0718	7.9670 0.0608	6.8808 0.0717	8.2016 0.0615	3.4574 4.7053	56.10 17.53
7	6.0877 0.0913	6.3014 0.0611	6.4385 0.0511	7.8787 0.0438	6.7981 0.0510	8.1110 0.0446	2.4413 4.2878	63.35 18.16
10	6.0097 0.0630	6.2435 0.0383	6.3788 0.0335	7.8223 0.0295	6.7384 0.0332	8.0543 0.0300	2.4897 4.7674	66.95 20.19

The procedures compared are the proposed BMA with 1 and 2 iterations (BMA1 and BMA2), four implementation of loess (LB1, LB2, LT1, LT3) and the free-knots splines BARS.

4.2. Simulation with outliers

To show the behavior of the method when there are outliers in the sample, we repeat the simulation for the first two functions of the previous section, Heavisine and Blocks, but now

Table 4
MSE obtained for the Doppler data with four different signal-to-noise ratio

RSNR	BMA1	BMA2	LB1	LB3	LT1	LT3	BARS	#Knots
3	1.0856 0.0809	1.1069 0.0765	1.2312 0.0815	1.3782 0.0832	1.2655 0.0814	1.4068 0.0838	0.7952 0.1533	23.95 1.84
5	0.7284 0.0349	0.8025 0.0308	0.8713 0.0355	1.0230 0.0332	0.9085 0.0354	1.0476 0.0336	0.3514 0.0643	30.28 1.94
7	0.6284 0.0215	0.7155 0.0193	0.7701 0.0250	0.9213 0.0203	0.8078 0.0248	0.9446 0.0200	0.2053 0.0395	35.16 2.36
10	0.5717 0.0137	0.6693 0.0116	0.7187 0.0157	0.8695 0.0125	0.7569 0.0156	0.8921 0.0125	0.0993 0.0160	41.19 2.51

The procedures compared are the proposed BMA with 1 and 2 iterations (BMA1 and BMA2), four implementation of loess (LB1, LB2, LT1, LT3) and the free-knots splines BARS.

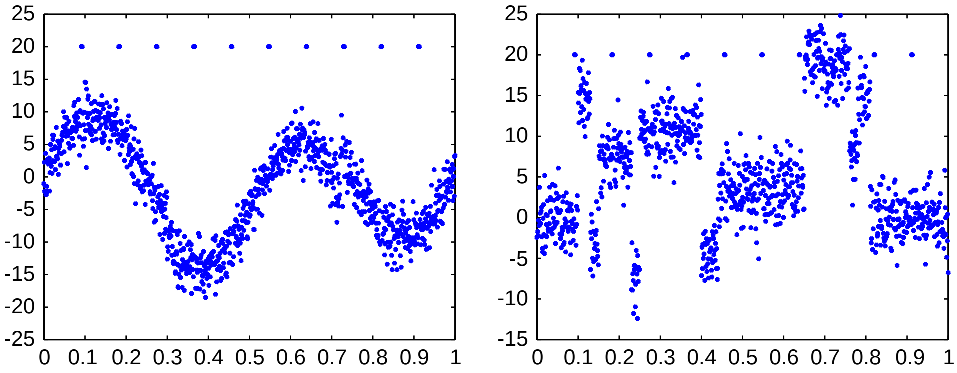

Fig. 9. A replication of the simulated outliers with RSNR $= 3$.

Table 5
MSE obtained for the Heavisine data with outliers

RSNR	BMA1	BMA2	LB1	LB3	BARS	#Knots
3	3.0956 0.2893	1.1771 0.1431	24.3396 4.4235	7.6680 2.6390	5.5503 3.3667	15.80 1.17
5	2.6917 0.2412	1.1110 0.1061	46.2537 3.5707	20.3535 3.2370	8.8345 3.5595	25.89 1.59
7	2.4896 0.1753	1.0880 0.0768	53.5682 2.6243	27.9767 2.6776	10.7109 3.1900	31.97 1.25
10	2.2456 0.1420	1.0371 0.0605	57.0434 1.7318	31.6018 2.2795	11.3895 3.0064	33.85 0.98

adding 3% of outliers. They have been added in groups of three consecutive outliers equal-spaced in the interval [0,1] and always with $y = 20$. One sample of the data configuration obtained with this distribution of the outliers and with a RSRN $= 3$ is presented in Fig. 9.

The results obtained in the comparison of the methods are shown in Tables 5 and 6. The methods included are the proposed procedure, BMA1 and BMA2, the loess method with bisquare kernel and polynomial degrees 1 and 3, LB1 and LB3, and the BARS method. We

Table 6
MSE obtained for the Blocks data with outliers

RSRN	BMA1	BMA2	LB1	LB3	BARS	#Knots
3	4.0725 0.1504	4.1641 0.1391	6.6320 0.2058	6.2982 0.4627	5.1663 0.7927	36.51 1.55
5	3.7817 0.1200	3.9714 0.1043	9.3803 0.5900	5.4019 0.1845	5.7939 0.9939	51.66 1.95
7	3.6300 0.0843	3.8811 0.0688	13.5658 0.7973	5.6632 0.7042	6.1844 0.9925	63.59 1.95
10	3.5143 0.0491	3.8182 0.0445	17.9794 0.6079	5.8036 0.2424	6.3139 0.8695	67.62 1.67

have not included in this table the loess method with kernel tricube, LT1 and LT3, because of their bad results. These results illustrate the danger of using a model based approach, such as BARS, blindly when there are outliers in the sample. The first conclusion from Tables 5 and 6 is that the performance of the BARS method, as it can be expected, is now much worse than before, and this procedure seems to be very sensitive to outliers. Both the proposed procedure and the loess method include some robust estimation and thus although their MSE increases for the larger variability due to the outliers, they are much less affected by it than the BARS method. For both functions, Heavisine and Blocks, the best result are obtained by BMA2, the second iteration of the proposed algorithm. We can observed that, when the RSNR grows, the MSE increases for the BARS method. This is due to the fact that when the residual noise decreases, the outliers have a larger relative size and their effect in introducing biased in the estimation of the curve increases. This effect also appears in general in loess, although more in the Heavisine function than in the Blocks function. On the other hand, with the proposed procedure the MSE decreases with the variance of the noise, as it should be for a robust procedure which is able to identify and downweight the outliers.

5. Concluding remarks

In this article a new method for fitting a curve by local polynomials is proposed. It introduces more flexibility in the local fitting by using BMA, and robustness by using mixtures of normals for the noise. The proposed method is simple to apply and to programme and completely automatic. We have shown in a Monte Carlo study that this method works better than others of similar computational complexity.

The main ideas of the method can be generalized to vector valued regressors $\mathbf{x} \in R^p$ in a straightforward way. First, we need to define a neighborhood in the space of independent variables and a distance function. Second, to avoid the curse of dimensionality problem, we have to control the number of parameters of the surface when p increases so that the number of parameters are a small fraction of the sample size. This can be done in a number of ways. The first is to assume an additive mode (see Hastie and Tibshirani, 1990), in order to reduce the p dimensional function $m(\mathbf{x})$ to a sum of p univariate functions in which the linear effects enter independently into the model. See Gustafson (2000) for some Bayesian generalizations of this approach. A second alternative is to use regression trees, as in the Treed procedure (see Alexander and Grimshaw, 1996) in which the sample is divided using a variable at each step. A third alternative is to use index models, in which $m(\mathbf{x})$ is written as a sum of functions of some index or components variables $z_i = \mathbf{g}_i'\mathbf{x}$. Finally, a fourth possibility is to use projection pursuit regression (Friedman and Stuetzle, 1981). The extension of this approach comparing these alternative ways to avoid the curse of dimensionality will be the subject of further research.

Acknowledgements

We acknowledge support from Grant BEC2000-0167, MCYT, Spain.

References

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. J. Am. Statist. Assoc. 88, 669–679.

Alexander, W.T., Grimshaw, S.D., 1996. Treed regression. J. Comput. Graphical Statist. 5, 156–175.

Anderson, T.W., 1962. The choice of the degree of a polynomial regression as a multiple decision problem. Ann. Math. Statist. 33, 255–265.

Box, G.E.P., Tiao, G.C., 1968. A Bayesian approach to some outlier problems. Biometrika 55, 119–129.

Brinkman, N.D., 1981. Ethanol fuel—a single cylinder engine study of efficiency and exhaust emissions. SAE Trans. 90, 1410–1427.

Brooks, R.J., 1972. A decision theory approach to optimal regression designs. Biometrika 59, 563–571.

Cleveland, W., 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Statist. Assoc. 74, 368, 829–836.

Cleveland, W., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Statist. Assoc. 83, 403, 596–610.

Denison, D.G.T., Mallick, B.K., Smith, A.F.M., 1998. Automatic Bayesian curve fitting. J. Roy. Statist. Soc. Ser. B 60, 2, 333–350.

Denison, D.G.T., Holmes, C.C., Mallick, B.K., Smith, A.F.M., 2002. Bayesian Methods for Nonlinear Classification and Regression, Wiley, New York.

DiMatteo, I., Genovese, C.R., Kass, R.E., 2001. Bayesian curve-fitting with free-knot splines. Biometrika 88, 1055–1071.

Donoho, D., Johnstone, I., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, 3, 425–455.

Eubank, R., 1988. Spline Smoothing and Nonparametric Regression, Marcel Dekker, New York.

Fan, J., Gijbels, I., 1995. Adaptive order polynomial fitting: bandwidth robustification and bias reduction. J. Comput. Graphical Statist. 4, 3, 213–227.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and Its Applications, Chapman & Hall, London.

Fernández, C., Ley, E., Steel, M., 2001. Benchmark priors for Bayesian model averaging. J. Econometrics 100, 381–427.

Friedman, J.H., Silverman, B.W., 1989. Flexible parsimonious smoothing and additive modeling (with discussion). Technometrics 31, 3–39.

Friedman, J.H., Stuetzle, W., 1981. Projection pursuit regression. J. Am. Statist. Assoc. 76, 817–823.

George, E., 1999. Bayesian model selection. In: Kotz, S., Read, C., Banks, D. (Eds.), Encyclopedia of Statistical Science Update 3. Wiley, New York, pp. 39–46.

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

Green, P.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, Chapman & Hall, London.

Gustafson, P., 2000. Bayesian regression modeling with interactions and smooth effects. J. Am. Statist. Assoc. 95, 795–806.

Guttman, I., 1967. The use of the concept of a future observation in goodness-of-fit problems. J. Roy. Statist. Soc. Ser. B 29, 1, 83–100.

Hager, H., Antle, C., 1968. The choice of the degree of a polynomial model. J. Roy. Statist. Soc. Ser. B 30, 469–471.

Halpern, E.F., 1973. Polynomial regression from a Bayesian approach. J. Am. Statist. Assoc. 68, 137–143.

Hansen, M.H., Kooperberg, C., 2002. Spline adaption in extended linear models. Statist. Sci. 17, 1, 2–51.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models, Chapman & Hall, London.

Holmes, C.C., Mallick, B.K., 2003. Generalized nonlinear modeling with multivariate free-knot regression splines. J. Am. Statist. Assoc. 98, 462, 352–368.

Justel, A., Peña, D., 1996. Gibbs sampling will fail in outlier problem with strong masking. J. Comput. Graphical Statist. 5, 2, 176–189.

Kass, R., Raftery, A., 1995. Bayes factor. J. Am. Statist. Assoc. 90, 430, 773–795.

Katkovnik, V.Y., 1979. Linear and nonlinear methods of nonparametric regression analysis. Soviet Autom. Control 5, 35–46.

Leamer, E., 1978. Bayesian Statistics: An Introduction, Wiley, New York.

Liang, F., Truong, Y., Wong, W., 2001. Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting. Statist. Sinica 11, 1005–1029.

Loader, C., 1979. Local Regression and Likelihood, Springer, Berlin.

Madigan, D., Raftery, A., 1994. Model selection and accounting for model uncertainty in graphical models using occam's window. J. Am. Statist. Assoc. 89, 1535–1546.

Mallick, B.K., 1998. Bayesian curve estimation by polynomial of random order. J. Statist. Plann. Inference 70, 91–109.

Peña, D., Tiao, G., 1992. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Robustness Functions for Linear Models. Bayesian Statistics, vol. 4. Oxford University Press, Oxford, pp. 365–388.

Peña, D., Yohai, V., 1999. A fast procedure for outlier diagnostics in large regression problems. J. Am. Statist. Assoc. 94, 434–445.

Raftery, A., Madigan, D., Hoeting, J., 1997. Bayesian model averaging for linear regression model. J. Am. Statist. Assoc. 92, 179–191.

Schmidt, G., Mettern, R., Schueler, F., 1981. Biomechanical investigation to determine physical and traumatogical differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under the effects of impact. Technical Report, EEC ResearchProgram on Biomechanics of Impacts. Final Report. Phase III. Project G5. Institut für Rechtsmedizin, University of Heidelberg, Germany.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Smith, P.L., 1982. Curve fitting and modelling with splines using statistical variable selection techniques. NASA report 166034, Langely Research Centre, Hampton, VA.

Smith, M., Kohn, R., 1996. Nonparametric regression using Bayesian variable selection. J. Econometrics 75, 317–343.

Sockett, E., Daneman, D., Clarson, C., Ehrich, R., 1987. Factors affecting and patterns of residual insulin secretion during the first year of type i (insulin dependent) diabetes mellitus in children. Diabetes 30, 453–459.

Stone, C.J., 1977. Consistent nonparametric regression. Ann. Statist. 5, 595–645.

Stone, C.J., 1980. Optimal rates of convergence for nonparametric estimators. Ann. Statist. 8, 1348–1360.

Stone, C.J., Hansen, M., Kooperberg, C., Truong, Y.K., 1997. Polynomial splines and their tensor products in extended linear modeling. Ann. Statist. 25, 1371–1470.

Tukey, J., 1960. A survey of sampling from contaminated distributions. Contributions to Probability and Statistics: Volume Dedicated to Harold Hetelling, Stanford University Press, Stanford, CA.

Verdinelli, I., Wasserman, L., 1991. Bayesian analysis of outlier problems using the gibbs sampler. Statist. Comput. 1, 105–117.

Wahba, G., 1975. Smoothing noisy data with spline functions. Numer. Math. 24, 383–393.

Wahba, G., 1984. Cross validated spline methods for direct and indirect sensing experiments. In: David, H.A., David, H.T. (Eds.), Statistics: An Appraisal. Iowa State University Press, Ames, pp. 205–235.

Wahba, G., 1990. Spline Models for Observational Data, Society for Industrial and Applied Mathematics, Philadelphia.

Yau, P., Kohn, R., Wood, S., 2003. Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. J. Comput. Graphical Statist. 12, 1, 23–54.