UNIVERSIDAD CARLOS III DE MADRID
working papers

Working Paper 03-44
Statistics and Econometrics Series 10
September 2003

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

BAYESIAN CURVE ESTIMATION BY MODEL AVERAGING

Daniel Peña and M. Dolores Redondas*

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 that four, and the parameters of each model are estimated by a local window. The estimated value is computed by BMA, and the posterior probability of each model is approximated by the exponential of the BIC criterion. The 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 some Monte Carlo experiments.

Keywords: Bayesian model averaging, BIC criterion, Robustness, Non parametric curve fitting, Local polynomial regression.

*Peña, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, C/ Madrid, 126, 28903 Getafe. Madrid, e-mail: dpena@est-econ.uc3m.es; Redondas, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, e-mail: redondas@est-econ.uc3m.es.

Bayesian Curve Estimation by Model Averaging

Daniel Peña [a], Dolores Redondas [a,*]

[a] *Department of Statistics, Universidad Carlos III de Madrid, c/Madrid 126, 28903, Getafe, Madrid, Spain.*

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 that four, and the parameters of each model are estimated by a local window. The estimated value is computed by BMA, and the posterior probability of each model is approximated by the exponential of the BIC criterion. The 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 some Monte Carlo experiments.

Key words: Bayesian model averaging, BIC criteron, Robustness, Non parametric curve fitting, Local polynomial regression

1 Introduction

A Bayesian approach is used to estimate non parametrically a regression model

$$y_i = m\left(x_i\right) + \varepsilon_i \qquad i = 1, \ldots, n$$

given the bivariate data $\left(x_1, y_1\right), \ldots, \left(x_n, y_n\right)$. We are interested in estimating the functional relationship m, between the variable y and the explanatory variable x, and to predict the response for new values of the covariate. The functional form of $m\left(\cdot\right)$ 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 $m\left(\cdot\right)$. There is an extensive literature for non parametric techniques, see for example Eubank (1988), Wahba (1990), Hastie and Tibshirani (1990) and Green and Silverman (1994) for a complete survey. 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 for the polynomial regression are due its global nature, that is, we may need a high order polynomial to approximate the data in the whole range of data and even then the approximation can be poor in wiggly curves. Second, the procedure is very non robust and a simple observation can exert a big influence on the estimated curve. Some better alternatives are the piecewise polynomials, the splines smoothers and the local polynomial regression. The first two methods require to select the number and positions of the knots. Smith and Kohn (1996) use a Bayesian approach to splines smoother to select the number of knots over a large number of knots. Liang et al. (2001) propose an automatic prior setting for the multiple linear regression and they applied the method to Bayesian curve fitting with regression splines. With regards to piecewise polynomials, Mallick (1998) makes the polynomial estimation of the curve 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 and Denison et al. (1998) select the knots using reversible jump Markov chain Monte Carlo to obtain the posterior probabilities for the joint distribution of the both the number and the position of the knots. These authors use piecewise polynomials instead of splines because the first are more flexible to modelize curves that are not smooth. Local polynomial regression was used by Cleveland (1979) who proposed a method which uses local regression with a kernel around the point of interest and it is made robust by using a weighted regression. This method has been widely used by its good results and its simplicity.

In this work we also use local polynomial regression but introduce some modifications over previous methods. 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). George (1999) reviews Bayesian model selection and discuses BMA in the context of decision theory. For linear regression models there is an extensive literature, see e.g. Raftery et al. (1997) and Fernández et al. (2002). Liang et al. (2001) propose an automatic prior setting for the parameters of linear regression which allow to implement MCMC methods for a big class of models. 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. The model is made robust by assuming that the noise may be contaminated. Then the Bayesian estimation provides and automatic downweighting of large observations.

The proposed procedure is completely automatic, very simple to apply and to program. Bayesian inference over the model allows us make predictions and optimal credible intervals: on the one hand, the use of the BIC criterion guarantee that if the true model is a polynomial model of degree smaller than d, then the true model will be use to estimate, and on the other hand, if the true model is not a polynomial, the use of BMA allows us to build credible interval that take into account the variability due the uncertainty about the model and gives us better predictive capability than using a single model (Madigan and Raftery (1994)). Another advantage of the proposed method is that we obtain the best model (for the BIC criterion) for each point of the sample, which provides an intuitive notion of the fitted curve.

The rest of the paper is organized as follows. Section 2 describes the proposed method and presents its main properties. Section 3 presents 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 the observations are related by

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

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 know, 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 discuss below, we consider polynomial models M_J of the forms

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

where $\overline{x}$ is the mean of the x variable in the sample considered, and approximate locally the general form $m(x_i)$ by linear combinations of these polynomials. Thus, we compute the posterior probabilities for different polynomial degrees and then estimate $m(x_i)$ at each point by forecasting using Bayesian model averaging (BMA).

The procedure is applied as follows. Suppose that the x observations are ordered, that is, $x_1 < x_2 < \ldots < x_n$, then for a given observation x_i, we define the symmetric nearest neighborhood around this point as

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

where w is the bandwidth of the window. Note that we are supposing that the observations are not repeated, if they are, we define the SNN over the set of different observations of x. As we may have repeated values, 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 $SNN\left(x_i,w\right)$ 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 $SNN\left(x_i,w\right)=\left\{x_k: x_{\max(1,i-w)}\leq x_k\leq x_{\min(n,i+w)}\right\}$.

In this work we make all the inference for the predicted value of a future observation $y_{f_i}=m\left(x_i\right)$ 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, $x_0\in\left(x_1,x_n\right)$, by defining $SNN\left(x_0,w\right)=SNN\left(x_i,w\right)$ where $x_i=\underset{k}{min}\left\|x_k-x_0\right\|$. Then, defining $D_i=\left\{\left(x_k,y_k\right): x_k\in SNN\left(x_i,w\right)\right\}$, the predictive distribution for a new observation at y_{f_i} is given by

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

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

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

where $\widehat{m}\left(x_i\left|D_i,M_J\right.\right)=E(y_{f_i}|D_i,M_J)$ is the expected value for the predictive conditional to the model.

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,

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

Then, the predictive distribution for a new observation, $p\left(y_{f_i} \left| D_i, M_J \right.\right)$, is a t-Student distribution with $v = n_0 - (J+1)$ degrees of freedom, where n_0 is the sample size of $SNN\left(x_i\right)$, mean $E\left(y_{f_i} \left| D_i, M_J \right.\right) = \mathbf{x}_i \widehat{\boldsymbol{\beta}}_J$, where $\widehat{\boldsymbol{\beta}}_J = \left(\widehat{\beta}_{J0}, \ldots, \widehat{\beta}_{JJ}\right)'$ is the vector of usual least-squares estimators for the parameters of the polynomial of degree J, $\mathbf{x}_i = \left(1, \left(x_i - \overline{x}_i\right), \ldots, \left(x_i - \overline{x}_i\right)^J\right)$and $\overline{x}_i = \left\{\sum x_k / n_0 : x_k \in SNN\left(x_i\right)\right\}$, and variance given by $Var\left(y_{f_i} \left| D_i, M_J \right.\right) = \frac{v}{v-2} s_J^2 \left(1 + \left(x_i - \overline{x}_i\right) \left(\mathbf{X}_J' \mathbf{X}_J\right)^{-1} \left(x_i - \overline{x}_i\right)\right)$ where $v s_J^2$ 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.

The posterior probability for a 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

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

where $p\left(\mathbf{y} | \widehat{\boldsymbol{\beta}}_J\right)$ is the likelihood of the model M_J, $\widehat{\boldsymbol{\beta}}_J$ is the MLE of the parameter vector under model M_J, n_0 is the sample size of $SNN\left(x_i\right)$ 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 $BIC\left(M_J\right) = -2S\left(M_J\right)$, and $\exp\left(S\left(M_{J_1}\right) - S\left(M_{J_2}\right)\right)$ approximates the Bayes factor $B_{J_1 J_2}$ with a relative error $O\left(1\right)$. Thus, we can approximate the Bayes factors by

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

and obtain the posterior probability for a model by

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

where $p\left(M_J\right)$ 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

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

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

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

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 very well any curve locally 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 $\widehat{y}_i^w$ the estimated value of $m\left(x_i\right)$ 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

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

We have check by simulation that the results are not very sensible 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 make the procedure fairly robust to the bandwidth parameter chosen.

With regard to the consistency of the proposed method, it can be showed by the consistency of the each polynomial model approach. Also, we can obtain the expressions of the bias and the variance based on the theorem 3.1, pg. 62, in Fan and Gijbels (1996)

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

$$\begin{aligned} Var\left[\widehat{m}\left(x\right)\right] &= Var\left[\sum_{i=0}^{3} p_i \widehat{m}_i\left(x\right)\right] \\ &= \left\{\left(\frac{p_0^2 + p_1^2}{2}\right) + 9\left(\frac{p_2^2 + p_3^2}{2}\right)\right\}\left(\frac{\sigma^2}{c}\frac{1}{nw}\right) + o_p\left(\frac{1}{nw}\right) \end{aligned}$$

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 $m^i\left(x\right)$ indicates the *ith* derivate of the $m\left(x\right)$ function. We are supposing that the marginal density of the observations x, $f\left(x\right)$, is uniform over the range of the data, $f\left(x\right) = c$ and $f'\left(x\right) = 0$.

In order to have a smoother curve the procedure described can be iterated as follows. Let $\widetilde{y}^{(1)}$ by the predicted value obtained by (3), then the observed values (x, y) are replaced by $\left(x, \widetilde{y}^{(1)}\right)$ to obtain $\widetilde{y}^{(2)}$, and in the same way $\widetilde{y}^{(k)}$ can be computed by using $\left(x, \widetilde{y}^{(k-1)}\right)$ as the observed values. In practice, we have found that a small number of iterations, $k = 2$ or $k = 3$ are 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. To reduced this effect, in the next section we propose a procedure for robustifying the method.

3 Robustifying the method

The method can be made robust to reduce the influence of the outliers in the local estimation by modeling the residuals by a mixture of normals. This model was introduced by Tukey (1960) and studied by Box and Tiao (1968). Suppose that observations $\mathbf{y}$ are generated by the model (1), where now the errors ε_i are a random variable with a normal 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 a $N\left(0, k^2\sigma^2\right)$ distribution. To make inference about this model, we introduce a dummy variable δ, $\delta_i = 1$ if $Var\left(\varepsilon_i\right) = k^2\sigma^2$ 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, making the variable change

$\mathbf{Y}_h = \mathbf{V}^{1/2}\mathbf{Y}$, $\mathbf{X}_h = \mathbf{V}^{1/2}\mathbf{X}$, we can apply standard inference results for linear models. The BMA predictive distribution for the future observation y_{f_i} given the data D_i, will be given by

$$p\left(y_{f_i}\left|D_i\right.\right) = \sum_{h=0}^{2^n} \sum_{J=0}^{d} p\left(y_{f_i}\left|D_i, M_J, \boldsymbol{\Delta}_h\right.\right) p_{Jh} \tag{5}$$

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

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

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

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

and the variance of the predictive distribution is

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

where

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

is the standard sum of the squared residuals.

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

The second term for the weights, is computed by

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

where $p\left(\mathbf{y} \left|\boldsymbol{\Delta}_h, M_J\right.\right)$ is the marginal distribution of the data, given a model M_J and a configuration $\boldsymbol{\Delta}_h$, $p\left(\boldsymbol{\Delta}_h \left|M_J\right.\right)$ is the prior probability of a configuration, which not depends of the model M_J, $p\left(\boldsymbol{\Delta}_h \left|M_J\right.\right) = p\left(\boldsymbol{\Delta}_h\right) = \alpha^{n_h}\left(1-\alpha\right)^{n-n_h}$ and n_h is the number of elements with high variance in the configuration $\boldsymbol{\Delta}_h$, $n_h = \sum \delta_i$. Finally, $p\left(M_J\right)$, 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, $p\left(\mathbf{y} \left|\boldsymbol{\Delta}_h, M_J\right.\right)$ the likelihood of the model for the parameters $\theta_J = (\boldsymbol{\beta}_J, \sigma_J)$ can be written as

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

where $\mathbf{X}_{Jn_h}$ indicates the rows of $\mathbf{X}_J$ corresponding to the observations with variance $k^2\sigma_{Jh}^2$, and $\mathbf{X}_{J(n-n_h)}$ corresponding to the observation with variance

σ^2_{Jh}. The marginal density is obtained by integrating θ_J, $p(\mathbf{y}\,|\boldsymbol{\Delta}_h, M_J) \propto (\widehat{s}^2_{Jh})^{-(n-J+1)/2} \left|\mathbf{X}'_J\mathbf{V}\mathbf{X}_J\right|^{-1/2}$ and finally we can obtain the expression for the marginal of the configuration,

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

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

3.1 Implementation

The scale contaminated normal model has the problem that the inference is over the 2^n possible configurations of the data and it requires intensive computation. Although we have many local problems with small sample size, the number of computations grows in exponential form, for example, for windows size $n_0 = 20$, it 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 Vernedelli and Wasserman, 1991 and Justel and Peña, 1996) but the local character of the estimation implies to solve approximately $n - n_0$ local problems which requires intensive computation. Note that in this problem we may take advantage from 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 a set of observations belonging to a 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}. We propose a procedure to obtain a good approximation to the posterior inference, that takes into account these characteristic of the problem. First, we obtain the configurations with highest probability in the first windows and second, using the information, we 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

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

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 $j = 1, 2, ..h - 1$ 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 this two groups. Peña and Tiao (1992) proposed to study the differences between the probabilities $P(A_i A_j)$, and $P(A_i) P(A_j)$, where A_i is the event that x_i is an outlier, and consider as potential outliers to 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 a 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 $A = \left\{ x_i : \frac{P(\mathbf{\Delta}_i|D)}{P(\mathbf{\Delta}_0|D)} \geq 3 \right\}$.

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

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

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 windows, D_1, we use this information to select the configurations in the next windows, 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 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 $p\left(\mathbf{\Delta}_h^i \left| \mathbf{y}\right.\right)$ for all the configurations in the windows D_i which have not negligible probability. We select the set of M configurations $\nabla_{D_i} = \{\mathbf{\Delta}_1^i, \ldots, \mathbf{\Delta}_M^i\}$ with highest posterior probability. Now, we move to the next windows, D_{i+1}, and let $\nabla_{RD_i} = \left\{ \mathbf{\Delta}_1^R, \ldots, \mathbf{\Delta}_{2^{m_i^R}}^R \right\}$ 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 deleted from ∇_{D_i} the terms corresponding to the observations which are not in D_{i+1}. Let $\mathbf{\Delta}_k^{*i}$ be the configuration obtained form $\mathbf{\Delta}_k^i \in \nabla_{D_i}$ 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 $\left\{\left[\mathbf{\Delta}_1^{*i} \cup \mathbf{\Delta}_1^R\right], \ldots, \left[\mathbf{\Delta}_1^{*i} \cup \mathbf{\Delta}_{2^{m_i^R}}^R\right], \ldots, \left[\mathbf{\Delta}_M^{*i} \cup \mathbf{\Delta}_1^R\right], \ldots, \left[\mathbf{\Delta}_M^{*i} \cup \mathbf{\Delta}_{2^{m_i^R}}^R\right]\right\}$ where $\left[\mathbf{\Delta}_k^{*i} \cup \mathbf{\Delta}_l^R\right]$ represents the $\mathbf{\Delta}_j^{*i}$ configuration for the observations which belongs to D_i and the configuration $\mathbf{\Delta}_l^R$ 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 choose 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 a experiment on the efficacy of crash helmets in simulated motor-cycle crash, and it is described in detail by Schmidt et al. (1981). The second one is the Ethanol data. The data includes 88 measurement of two variables from a 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 on 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.

Figure 1 shows the estimated curve for the Helmets data, where the parameter estimated by cross validation is $w = 12$. 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. In 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.

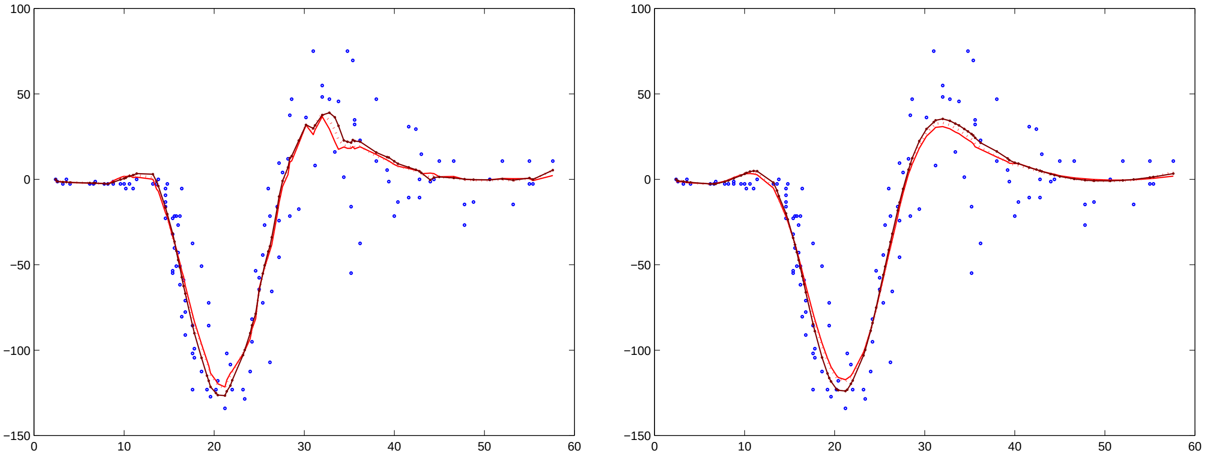

Fig. 1. Curves fit for Helmets data. The left figure shows the curve for the standard method (solid line), the robust method with parameters $(\alpha = 0.05, k^2 = 3)$ (dotted line) and the robust method with $(\alpha = 0.1, k^2 = 5)$ (solid line with marks *). The right figure shows the second iteration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

Figure 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 one 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 observed that there are small differences among the three curves and none of them is completely smooth. Note that as the data is homogeneous in this case 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. It can be seen that the three curves obtained are smooth and very similar.

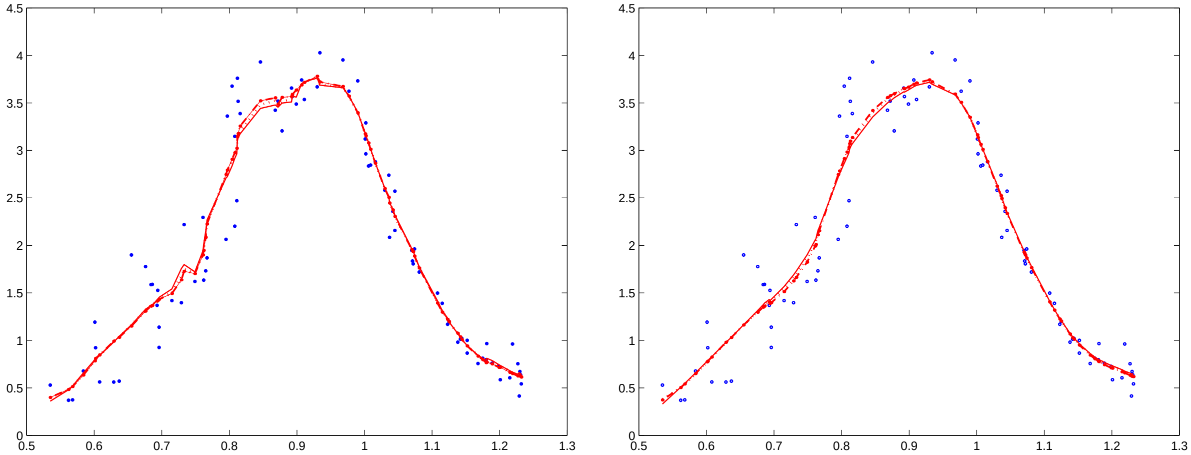

Fig. 2. Curves fit for Ethanol data. The left figure shows the curve fitted by the standard method (solid line), the robust method with parameters ($\alpha = 0.05, k^2 = 3$) (dotted line) and the robust method with ($\alpha = 0.1, k^2 = 5$) (solid line with marks *). The right figure shows the second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

Figure 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.

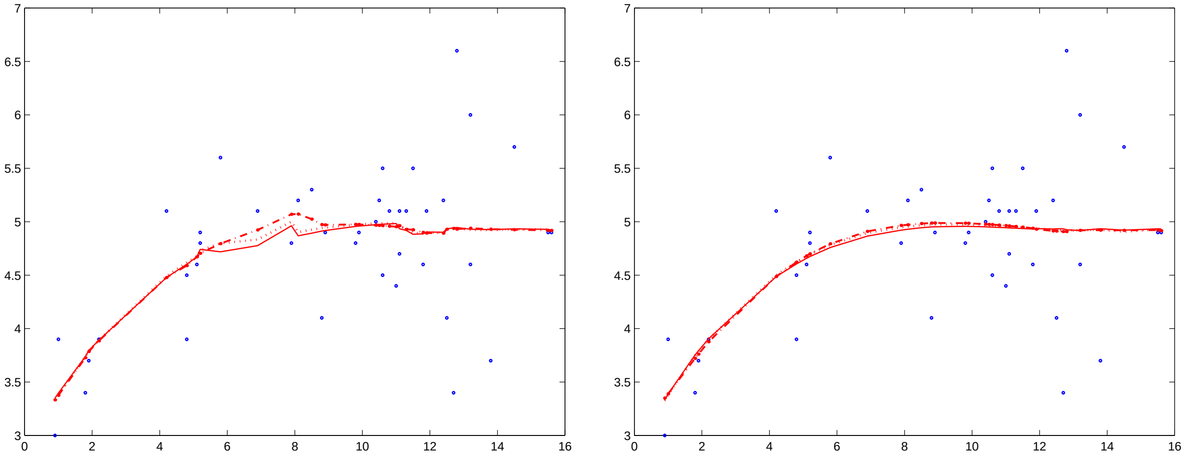

Fig. 3. Curves fit for Diabetes data. The left figure shows the curve fitted by the standard method (solid line), the robust method with parameters ($\alpha = 0.01, k^2 = 3$) (dotted line) and robust method with ($\alpha = 0.05, k^2 = 7$) (solid line with marks *). The right figure shows the second iteration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

4.1 Monte Carlo experiment

We compare the behavior of the proposed method to the popular loess method due to Cleveland (1979) which is implemented in many computer programs. The comparison is made by using four simulated function proposed by Donoho and Johnstone (1994) which have been used often in the literature for comparison purposes (see Denison el al., 1998). The four simulated functions are:

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

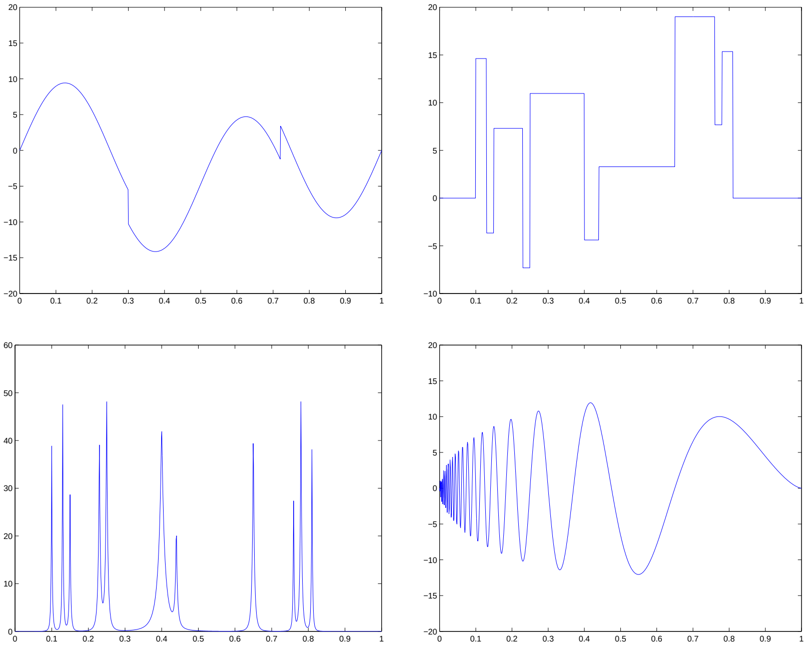

Fig. 4. The simulated functions used to compare the proposed method: Heavisine, Blocks, Bumps and Doppler.

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 $Var\left(f\right) = 7^2$. The errors are generated by $\varepsilon_i \sim N\left(0, \sigma^2\right)$, where σ^2 is chosen so that the root of the signal noise ratio $\left(RSNR = \sqrt{\frac{var(f)}{\sigma^2}}\right)$ are $3, 5, 7$ and 10. The simulation are based in 1000 points. The four simulated functions are showed in the Figure 4.

The proposed method is based on the use of a uniform kernel $W_1\left(x, x_i\right) = 1$ if $x \in SNN\ \left(x_i\right)$. In these simulations we compare the use of the uniform kernel with the mixture of the four polynomials models with the kernels used by Cleveland (1979) in his loess method with a fixed polynomial degree $d = 1$ or $d = 3$. The kernels are bisquare weight function, $B\left(x\right) = \left(1 - x^2\right)^2$ for $|x| < 1$, and the 'tricube' function $T\left(x\right) = \left(1 - |x|^3\right)^3$. In both kernels x is reescaled

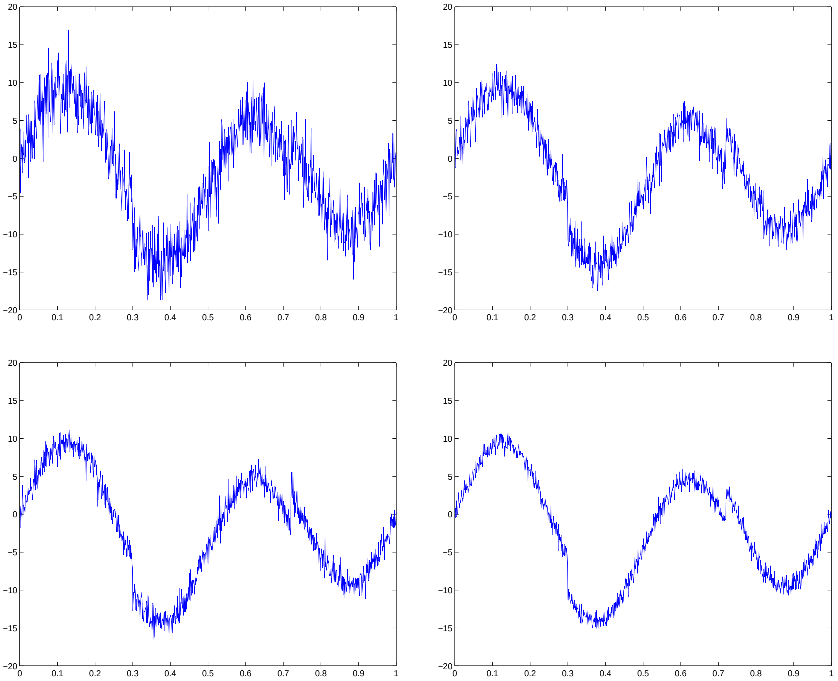

Fig. 5. Heavisine function with different root of signal-noise ratio: 3, 5, 7, 10.

by $\frac{(x-x_i)}{h_i}$ where h_i is the distance $|x - x_i|$ from x to the rth nearest neighbor.

The mean of the squared error, $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}\left(x_i\right) - m\left(x_i\right)\right)^2$, are showed in the next tables, where $\widehat{m}\left(x_i\right)$ is computed by six different procedures. The first two called BMA1 and BMA2 in the tables, are the proposed method with 1 and 2 iterations. The third and fourth are computed using the loess with polynomial of degree 1 and 3, and a bisquare kernel and are called B1 and B3. The last two methods correspond to degree 1 and 3 using the tricube kernel and are called T1, T3. The results are based in 1000 replications of the simulated curve. The simulated curves with the four levels of RSNR=$\{3, 5, 7, 10\}$ are showed in the figures 5, 6, 7, and 8.

Table 1 shows the mean and the standard deviation, in small letter size, of the MSE of the 1000 replications of the function Hevisine. We can observed that the smallest MSE is obtained by BMA2, the proposed method with two iterations of the algorithm. Also, we can observed that the bisquare kernel is

Table 1

ECM obtaned for the Heavisine data with four different signal to noise ratio.

$RSNR$	BMA1	BMA2	B1	B3	T1	T3
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
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
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
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

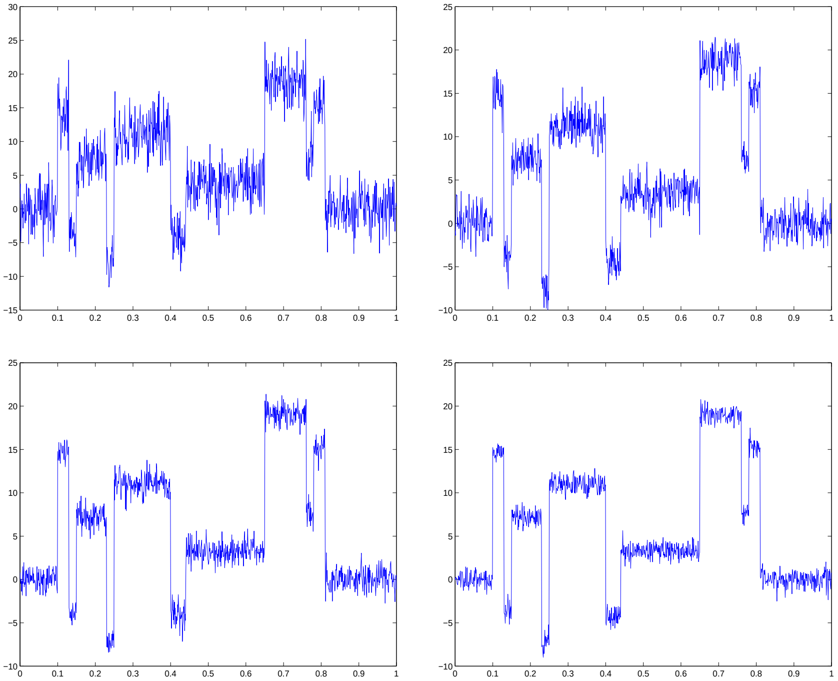

Fig. 6. Blocks function with different root of signal-noise ratio: 3, 5, 7, 10.

slightly better than the tricube, and that the linear fit in loess is better than the cubic fit. The biggest differences among the procedures are observed with the signal to noise ratio is the largest, RSNR=10.

Table 2 shows the result obtained for the function Blocks. The best results are again obtained for BMA2, the second iteration of the algorithm, and again

Table 2

ECM obtained for the Blocks data with four different signal to noise ratio.

$RSNR$	BMA1	BMA2	B1	B3	T1	T3
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
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
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
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

Table 3

ECM obtained for the Bumps data with four different signal to noise ratio.

$RSNR$	BMA1	BMA2	B1	B3	T1	T3
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
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
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
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

the linear fit is better than the cubic and the bisquare kernel slightly better than the tricube. However, for the functions Bumps and Doppler (see tables 3 and 4) the best results are obtained with BMA1, the first iteration of the algorithm. This is not surprising as these functions are not very smooth and a second iteration smooths the picks in the case of the bumps data and the maximum and minimum in the case of the Doppler data. With regards to loess the ressults are the same as before: the linear fit with the bisquare kernel has the best performance.

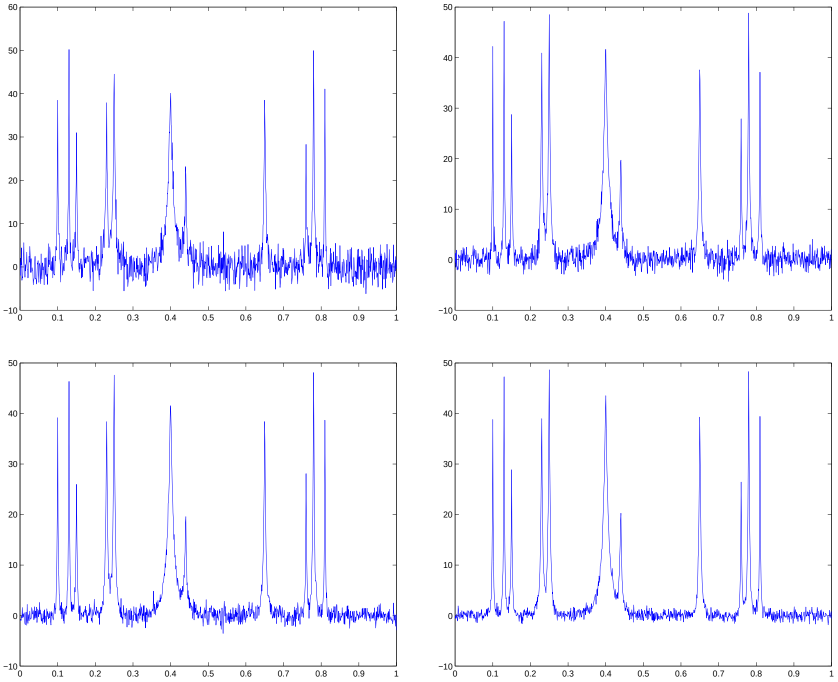

Fig. 7. Bumps function with different root of signal-noise ratio: 3, 5, 7, 10.

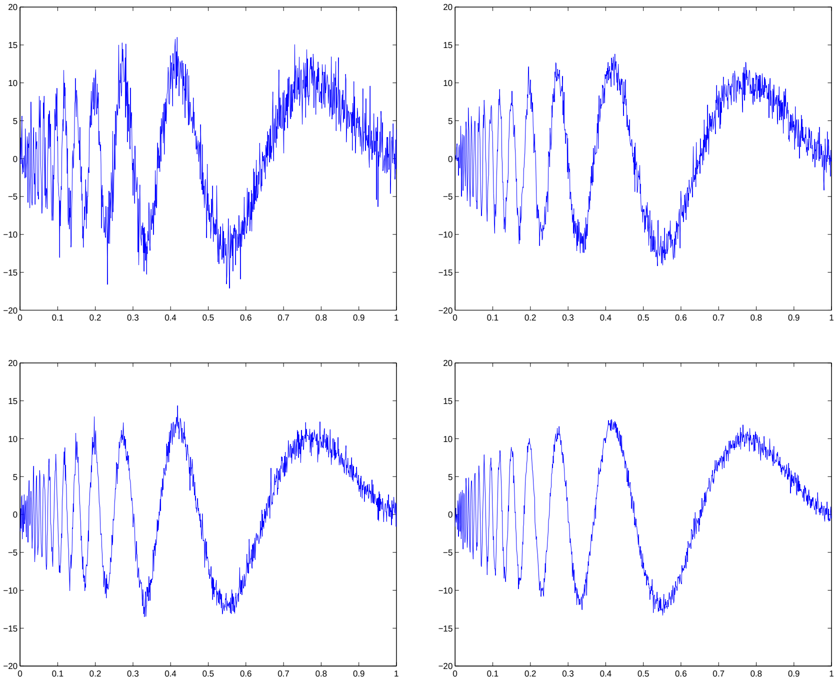

Fig. 8. Doppler function with different root of signal-noise ratio: 3, 5, 7, 10.

5 Concluding Remarks

In this article a new method for fitting a curve is proposed. The proposed method is very simple to apply and to programme. Furthermore, it is com-

Table 4

ECM obtained for the Doppler data with four different signal to noise ratio.

RSRN	BMA1	BMA2	B1	B3	T1	T3
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
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
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
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

pletely automatic and the Bayesian inference provides the predictive distribution and credible intervals. The method takes into account the possible polynomial models fitted locally to the data, and the consistency of the BIC criterion , which provides the weights, guarantees that if the true model is a polynomial of degree less than four, then asymptotically the true model will be used for estimation.

In the Monte Carlo results we show that the method works better than methods of similar complexity, with kernels which take into account the distance to the point of interest.

References

Anderson, T. W., 1962. The choice of the degree of a polynomial regression as a multiple decision problem. The Annals of Mathematical Statistics 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 cylindrer engine study of efficiency and exhaust emissions. SAE Transactions 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 smothing scatterplots. Journal of the American Statistical Association 74 (368), 829–836.

Denison, D. G. T., Mallick, B. K., Smith, A. F. M., 1998. Automatic Bayesian curve fitting. Journal of the Royal Statistical Society, Ser B. 60 (2), 333–350.

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

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

Fan, J., Gijbels, I., 1995. Adaptive order polynomial fitting: Bandwidth robustification and bias reduction. Journal of Computational and Graphical Statistics 4 (3), 213–227.

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

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

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

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

Guttman, I., 1967. The use of the concept of a future observation in goodness-of-fit problems. Journal of the Royal Statistical Society, Ser B. 29 (1), 83–100.

Guttman, I., Peña, D., Redondas, D., 2003. A Bayesian approach for predicting polynomial regression of unknown degree. Working Papers 03-21. Universidad Carlos III de Madrid. .

Hager, H., Antle, C., 1968. The choice of the degree of a polynomial model. Journal of the Royal Statistical Society, Ser B. 30, 469–471.

Halpern, E. F., 1973. Polynomial regression from a Bayesian approach. Journal of the American Statistical Association 68, 137–143.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Capman and Hall, London.

Hoeting, J., Madigan, D., Raftery, A., Volinsky, C., 1999. Bayesian model averaging: A tutorial. Statistical Science 14 (4), 382–417.

Justel, A., Peña, D., 1996. Gibbs sampling will fail in outlier problem with strong masking. Journal of Computational and Graphical Statistics. 5 (2), 176–189.

Kass, R., Raftery, A., 1995. Bayes factor. Journal of the American Statistical Association 90 (430), 773–795.

Leamer, E., 1978. Bayesian Statistics: An Introduction. John Wiley and Sons.

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

Madigan, D., Raftery, A., 1994. Model selection and accounting for model uncertainty in graphical models using occam's window. Journal of the American Statistical Association 89, 1535–1546.

Mallick, B. K., 1998. Bayesian curve estimation by polynomial of random order. Journal of Statistical Planning and Inference 70, 91–109.

Peña, D., Tiao, G., 1992. Bayesian Robustness Functions for Linear Models. J.M. Bernardo, J.O. Berger, A. P. Dawid and A.F.M. Smith. Oxford

University Press.

Raftery, A., Madigan, D., Hoeting, J., 1997. Bayesian model averaging for linear regression model. Journal of the American Statistical Association 92 (437), 179–191.

Rupert, D., Wand, M., 1994. Multivariate locally weighted least squares regression. The Annals of Statistics 22, 1346–1370.

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 efects of impact. Tech. rep., EEC Research Program 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. The Annals of Statistics 6, 461–464.

Smith, M., Kohn, R., 1996. Nonparametric regression using Bayesian variable selection. Journal of 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. Diabet 30, 453–459.

Tukey, J., 1960. A Survey of Sampling From Contaminated Distributions. Contributions to Probability and Statistics: Volume Dedicated to Harold Hetelling. Stanford, CA: Stanford University Press.

Verdinelli, I., Wasserman, L., 1991. Bayesian analysis of outlier problems using the gibbs samper. Statistics and Computing 1, 105–117.

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