Degrees of freedom and model selection in semiparametric additive monotone regression.

Cristina Rueda[1]

University of Valladolid. Spain

Abstract

$\mathbb{R}$ The degrees of freedom of semiparametric additive monotone models are derived using results about projections onto sums of order cones. Two important related questions are also studied, namely, the definition of estimators for the parameter of the error term and the formulation of specific Akaike Information Criteria statistics. Several alternatives are proposed to solve both problems and simulation experiments are conducted to compare the behavior of the different candidates. A new selection criterion is proposed that combines the ability to guess the model but also the efficiency to estimate the variance parameter. Finally, the criterion is used to select the model in a regression problem from a well known data set.

Keywords: additive models, isotonic models, order restricted inference, Akaike information criterion.

1. Introduction

Semiparametric monotone additive models are receiving special attention in the statistical literature because they are pragmatic alternatives to linear and nonparametric models. There are several main advantages. First, the flexibility, including more forms of regression than the rigid linear formulation. Second, the simple additive structure that guarantees, as shown below, the solution of the estimation problem with a relatively simple algorithm. Third, the incorporation of the monotonicity restriction avoids the problem

Email address: crueda@eio.uva.es (Cristina Rueda)

[1]*Address*: Prado de la Magdalena s/n. Facultad de Ciencias. Dpto. Estadistica. University of Valladolid. 47005 Valladolid. Spain.

of defining user-specified choices, such as bandwidth, or smoothing parameters or number and placement of knots, typical drawbacks of nonparametric methods as kernel smoother, smoothing splines and regression splines, being in this sense a robust methodology. And finally, the oracle property, which implies that the rate of convergence of the least-square estimate of each additive component is independent of the number of additive components in the model (see Chen (2009) and references therein).

The usefulness of monotone models is wide, the papers by Morton-Jones et al. (2000), Hussian et al. (2004), and De Boer et al. (2002) give some applications in the biomedical, environmental and toxicology fields respectively. These are only a few among many others in different fields. There are a lot of settings where isotonic models are suitable as monotone relationships frequently appear in real practice. Two illustrative examples are: first, the relation between the risk of getting a disease and exposure, in epidemiological applications, where the risk is often known to decrease with increasing exposure; and second, the prediction of sociological indexes that are known to monotonically change with important predictors, as is the case in the example analyzed in section 4. Besides, the incorporation of linear terms is useful to model dummy explanatory variables and also to decrease the risk of overfitting.

The semiparametric additive monotone regression model is defined by:

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ji} + \sum_{j=1}^{q} h_j(z_{ji}) + \varepsilon_i, i = 1, ..., n, \tag{1}$$

where $y = (y_1, ..., y_n)'$ is the response vector, $x_j = (x_{j1}, ..., x_{jn})', j = 1, ..., p$ and $z_j = (z_{j1}, ..., z_{jn})', j = 1, ...q$, are linearly independent explanatory variables, and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ is a random error term. It is assumed that each $h_j()$ is a monotone function, which is suppose to be monotone increasing without loss of generality, and that $\varepsilon \sim N(0, W^{-1}\sigma^2)$, where $W = diag(w_1, ..., w_n)$ is a matrix of known weights (by instances, in experiments with replicas w_i is the number of replicas under each condition) and σ is unknown. The MLEs $(\widehat{\alpha}, \widehat{\beta}_1, ..., \widehat{\beta}_p, \widehat{h}_1, ..., \widehat{h}_q)$ are the solution of the optimization problem:

$$\min \left(\sum_{i=1}^{n} w_i (y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ji} - \sum_{j=1}^{q} h_j(z_{ji}))^2 \right), \tag{2}$$

subject to the restriction that $\alpha \in \Re$, $\beta \in \Re^p$ and each $h_j()$ is monotone and verifies the standard identifiability condition $\sum_{i=1}^{n} h_j(z_{ji}) = 0$.

In fact, the solution to the optimization problem (2) is not unique, any other set of monotone functions $m_j, j = 1, ..., q$, verifying $m_j(z_{ji}) = \widehat{h}_j(z_{ji}), j = 1, ..., q; i = 1, ..., n$ would have also solved the least square minimization. Define $\theta_0 = \alpha - \sum_{j=1}^{p} \beta_j x_{ji}$; $\theta_j = h_j(z_j), j = 1, ..., q$ adn $\theta = \theta_0 + \theta_1 + ... + \theta_q$. Then, the least-square estimator $\widehat{y} = (\widehat{y}_1, ..., \widehat{y}_n)'$, where $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}_1 x_{1i} + ... + \widehat{\beta}_p x_{pi} + \widehat{h}_1(z_{1i}) + ... + \widehat{h}_q(z_{qi}) = \widehat{\theta}_0 + \widehat{\theta}_1 + ... + \widehat{\theta}_q$, is the L_2-projection with weights W of the observed vector y onto K, $p_w(y/K)$, where $K = L_0 + S_1 + ... + S_q$ is a convex cone in $\Re^n$ defined by the restrictions imposed, L_0 being the linear subspace of dimension $p+1$ spanned by columns in matrix $(1_n, x_1, ..., x_p)$ and each S_j being the order cone associated to z_j, $S_j = \{u \in \Re^n \ / \ u_{1_j} \leq ... \leq u_{n_j}\}$, where $(1_j, ..., n_j)$ is a permutation of $(1, ..., n)$ verifying $z_{j1_j} \leq ... \leq z_{jn_j}$ (if several observations have the same value of the predictor z_j, define S_j using equalities instead of inequalities between the coordinates corresponding to these observations). A simple example that illustrates how the cone K is derived from the initial information on the explanatory variables is given in section 2. Therefore, the model that dealt with in this paper can be expressed in a simplified form as:

$$y = \theta + \varepsilon, \theta \in K; \varepsilon \sim N(0, W^{-1}\sigma^2).$$

Without loss of generality, it can be assumed that $W = I$ since a linear change of variable can be made to the unweighted case by considering $y^* = W^{1/2}y$ and the cone $K^* = W^{1/2}K$. Then, $p_w(y/K) = W^{-1/2}p(y^*/K^*)$, which implies that most properties of the projections in the weighted case are derived from those in the unweighted case, in particular those proved or used in this paper. In addition, the calculation of $p_w(y/K)$ can be accomplished via the PAVA (Pool Adjacent Violators Algorithm) when K is an order cone as S_j, and the algorithm to obtain the projection when K is the sum of several of these cones is a general PAVA, which also works with weights. (see Meyer (1999) and Robertson et al (1988) for the properties of the weighted projections and the computational algorithms). Therefore, in the following the notation for the weighting is dropped to simplify the presentation and the solution to the optimization problem can be expressed as,

$$\widehat{\theta}_K = \arg\min_{\theta \in K} \|y - \theta\|^2, \tag{3}$$

where $\|u\|^2 = \sum_{i=1}^{n} u_i^2$.

The simple additive structure of the model guarantees the solution of the optimization problem with a relatively simple algorithm and good properties of estimators. Mammen and Yu (2007) derive a backfitting algorithm, a cyclic PAVA, to get the estimators, and show the oracle property for the model without linear terms; and Cheng (2009) extends the results to the general case and derives the asymptotic distribution of the regression estimators.

Before the publication of the papers mentioned in the last paragraph, other authors dealt with this type of models. Some important references are Stone (1982) and Stone (1985), who studied the rates of convergence of regression estimators and first showed the oracle property for additive models; Bachetti (1989) who first estimated the additive functions with backfitting; and Huang (2002) who dealt with the case $q = 1$. There have also been many other authors, who have dealt with additive models or isotonic regression models, whose research has been the basis for recent development. We highlight the works by Brunk (1970), Hanson (1973), Dykstra (1983), Hastie and Tibsirani (1986), among many others.

However, there is still an unsolved and important question around these models that has to do with the determination of the degrees of freedom or the dimensionality of the model. Several authors have dealt with the question of dimensionality. Meyer and Woodroofe (2000) dealt with the univariate monotone and shape regression models, while Kato (2009) dealt with the question in shrinkage regression with application to the Lasso. For a general convex cone C, these authors introduced the concept of degrees of freedom of the associated model, also called the divergence, which is defined by $D_C(y)$ as follows:

$$D_C(y) = div_C(\widehat{\theta}_C) = \sum_{i=1}^{n} \frac{\partial}{\partial y_i} \widehat{\theta}_i(y), \tag{4}$$

where, $\widehat{\theta}_C(y) = p(y/C) = (\widehat{\theta}_1(y), ..., \widehat{\theta}_n(y))'$.

When $q = 0$ ($p = 0$ and $q = 1$), the cone K is defined using inequality restrictions and $D_K(y)$ is easily derived as $dim(L_K^y)$, where L_K^y is the subspace defined by the inequalities that the projection verifies as equalities and verifies $p(y/K) = p(y/L_K^y)$. It is also given by counting among these inequalities, the maximum number being linearly independent. Something similar happens in the applications given in Kato (2009), including the Lasso among others.

In these cases, the convex cone is expressed explicitly by a set of linear inequalities and equalities.

However, the derivation of $D_K(y)$ is not straightforward in the general case $q > 0$ and $p > 0$ from the results in previous papers. The aim of this paper is to achieve it in a simple way.

The derivation of $D_K(y)$ is also relevant for solving two important aspects in model fitting the estimation of σ and the model selection. In the former, the degrees of freedom is used to correct the bias of the MLE estimator and in the latter to derive the AIC measure penalty term. Several authors have considered the first problem in univariate regression, Meyer (2000) and Rueda et al. (2010) among others. But again, as far as we know, the question is not solved in the general case. On the other hand, the problem of variable selection is a very important problem in statistical modeling that has recently gained a lot of attention in semiparametric models (see Li and Liang (2008) and Xiao et al. (2010) among others). The AIC approach is considered in this paper for this question. In this paper, several alternative proposal are given to solve both questions which are validated with numerical simulations and in a example.

The outline for the rest of the paper is as follows: in section 2, a practical formula for $D_K(y)$ for semiparametric additive monotone models is derived using new algebraic results in relation with the projection onto the sum of order cones. In section 3, the question of the estimation of σ is discussed and a new AIC criterion for model selection that uses a corrected bias estimator for σ is proposed. The estimator and the criterion are validated using simulation experiments. Finally, the results are applied in section 4 to the well known Prestige data set and some general conclusions are given in section 5.

2. Degrees of freedom

In this section, the subscript K is eliminated from $\widehat{\theta}$ and $D(y)$ to make the presentation easier. The solution to the optimization problem (3) is then given by $\widehat{\theta} = p(y/K)$, which is the maximum likelihood estimator of θ.

As Meyer and Woodroofe (2000) have shown, the divergence $D(y)$ defined in (4), gives the degrees of freedom of the model. When K is a convex polyhedral, $K = \{u \in \Re / a_i' u \leq 0, i = 1, ..., m\}$, $D(y)$ is the dimension of the linear subspace where the projection is obtained, $D(y) = \dim(L_K^y)$,

$L_K^y = \left\{ u \in \Re / a_i' u = 0, \text{ for all } i \text{ for which } a_i' \widehat{\theta}(y) = 0 \right\}$.

Let $F_K^y = L_K^y \cap K$, the set of polyhedral cones $F_{K_y}^y$ are called faces of K and each L_K^y is the linear subspace associated with the face F_K^y . The faces of a polyhedral cone are involved in the algebraic result derived in this section.

Lemma 2.1(i), below, shows that $K = L_0 + S_1 + ... + S_q$ is a polyhedral cone (this is a known result given here for completeness). Lemma 2.1 (ii) and (iii) give other results on projections on convex cones that will be used later on.

However, $D(y)$ cannot be derived from Lemma 2.1(i) in a straightforward manner because it does not provide an explicit version of K as a set of linear inequality restrictions.

In the simple case of L_0 and each S_j being orthogonal linear subspaces, we have, from the properties of projections, that $p(y/K) = p(y/L_0) + p(y/S_1) + ... + p(y/S_q)$ and also that $D(y) = \dim(L_0) + \dim(S_1) + ... + \dim(S_q)$. Even in the case where orthogonality is not verified but each S_j is a linear subspace, the dimension can be derived from the individual dimensions and dimensions on the intersections. In the case $q = 1$, the equation is very simple: $D(y) = dim(L_0 + S_1) = \dim(L_0) + \dim(S_1) - \dim(L_0 \cap S_1)$. This is not true when S_j are cones, as the example shows. The example also illustrates the derivation of $D(y)$ in a simple case.

Example Let $n = 3$, $z = (3, 2, 1)$, $x = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}$.

$L_0 = \{u \in \Re^3 / u_1 = u_2\}$, $S_1 = \{u \in \Re^3 / u_1 \geq u_2 \geq u_3\}$, $L_0 \cap S_1 = \{u \in \Re^3 / u_1 = u_2 \geq u_3\}$. $\dim(L_0) + \dim(S_1) - \dim(L_0 \cap S_1) = 2 + 3 - 2 = 3$.

Let $y = (y_1, y_2, y_3) = (1, 2, 2)$, the backfitting algorithm (Cheng (2009))is applied to find $\widehat{\theta}(y)$. The runs of the algorithm are as follows:

$\widehat{\theta}_0^{(0)}(y) = (\overline{y}, \overline{y}, \overline{y})$; $\widehat{\theta}_0^{(1)}(y) = (0, 0, 0)$

$\widehat{\theta}_0^{(1)}(y) = p(y - \widehat{\theta}_1^{(0)}(y)/L_0) = (\frac{y_1+y_2}{2}, \frac{y_1+y_2}{2}, y_3)$;

$\widehat{\theta}_1^{(1)}(y) = p(y - \widehat{\theta}_0^{(1)}(y)/S_1) = (0, 0, 0)$.

Then, $\widehat{\theta}(y) = \widehat{\theta}_0(y) + \widehat{\theta}_1(y) = (\frac{y_1+y_2}{2}, \frac{y_1+y_2}{2}, y_3) + (0, 0, 0)$.

Now, $D(y) = \sum_{i=1}^{3} \frac{\partial}{\partial y_i} \widehat{\theta}_i(y) = \sum_{i=1}^{3} \frac{\partial}{\partial y_i} \sum_{j=0}^{1} \widehat{\theta}_{ji}(y) = 2$.

In the general case, it is much more complicated to find $D(y)$ because the backfitting algorithm does not give an explicit version of $\widehat{\theta}$ in terms of y. To

solve the problem, as in the example above, any representation of $\widehat{\theta}$ will be used from the backfitting algorithm that is given by: $\widehat{\theta} = \widehat{\theta}_0 + \widehat{\theta}_1 + ... + \widehat{\theta}_q$, $\widehat{\theta}_0 \in L_0$, $\widehat{\theta}_j \in S_j$, $j = 1, ..., q$.

To make this representation useful to derive $D(y)$, lemma 2.2 (i) proves that $D(y) = \dim(L_0 + L_1^y + ... + L_q^y)$, where each L_j^y is the linear subspace associated with a face of the cone S_j such as $\widehat{\theta}_j \in L_j^y$. Lemma 2.2(ii) also provides an easy way to calculate $D(y)$ in practical applications.

The proofs of the two lemmas are deferred to the appendix.

Lemma 2.1. *(i) Let L be a linear subspace and let C_j, $j = 1, ..., q$ be polyhedral cones. Then $K = L + C_1 + ... + C_q$ is a polyhedral cone.*

(ii) Let L and C be a linear subspace and a convex cone respectively and let $L \subset C$. Then, $C = L + C \cap L^{\perp}$ and $p(y/C) = p(y/L) + p(y/L^{\perp} \cap C)$.

(iii) Let L and C be a linear subspace and a convex cone respectively. Then, $p(y/L + C) = p(y/L) + p(y/L^{\perp} \cap (L + C))$.

Lemma 2.2. *Let $K = L_0 + S_1 + ... + S_q$ be a convex cone where L_0 is a linear subspace and each S_j is an order cone. For a given $y \in \Re^n$, the backfitting algorithm gives: $p(y/K) = \widehat{\theta} = \widehat{\theta}_0 + \widehat{\theta}_1 + ... + \widehat{\theta}_q$, $\widehat{\theta}_0 \in L_0$, $\widehat{\theta}_j \in S_j$, $j = 1, ..., q$. Then,*

(i) $D(y) = \dim(L_0 + L_1^y + ... + L_q^y)$, where each L_j^y is a linear subspace associated with a face of S_j that has $\widehat{\theta}_j$ as an interior point.

(ii) Let us denote $\{a_{ji}\}_{i=1}^{N_j}$ as the N_j different components of $\widehat{\theta}_j$. For each $j = 1, ..., q$ and $i = 1, ..., N_j$, let v_j^i be the n-dimensional vector defined by

$$v_{jk}^i = 1 \Longleftrightarrow \widehat{\theta}_{jk} = a_{ji}, v_{jk}^i = 0 \Longleftrightarrow \widehat{\theta}_{jk} \neq a_{ji}, k = 1, ..., n.$$

Then, for each j, $\dim(L_j^y) = N_j$, and the set $\left\{v_j^i\right\}_{i=1}^{N_j}$ is a minimal set of generators for L_j^y.

Remark From the generators for the subspaces, L_j^y $j = 1, .., q$, given in Lemma 2.2(ii), select a maximal set of linearly independent vectors and define a matrix, V, with these vectors as columns. Then, $rank(V, 1_n, x_1, ...x_p) = D(y)$.

Moreover, from the properties of projection onto the sum of subspaces, we have that,

$$D(y) = \dim(L_0 + L_1^y + ... + L_q^y) = p + 1 + \dim(L_1^y + ... + L_q^y) - \dim(L_0 \cap (L_1^y + ... + L_q^y))$$

and,

$$\dim(L_1^y+...+L_q^y) = \sum_j \dim(L_j^y) - \sum_{l,j} \dim(L_l^y \cap L_j^y) + \sum_{l,m,j} \dim(L_l^y \cap L_m^y \cap L_j^y) - ....$$

$$= \sum_j N_j - \sum_{l,j} N_{l,j} + \sum_{l,m,j} N_{lmj} - ...,$$

where N_j is the number of different values in the set $\left\{\widehat{\theta}_{ji}\right\}_{i=1}^n$, N_{lj} is the number of different values that occur simultaneously (in the same coordinates) in the sets $\left\{\widehat{\theta}_{ji}\right\}_{i=1}^n$ and $\left\{\widehat{\theta}_{li}\right\}_{i=1}^n$, etc... We then have a constructive equation that shows the contribution of the linear component, the monotone components and the intersections to $D(y)$.

3. Variance estimation and AIC statistics

3.1. Discussion and definitions

Let us consider first the estimation of σ^2. The problem of the bias of $\widehat{\sigma}^2_{MLE} = \frac{\left\|y-\widehat{\theta_K}\right\|^2}{n}$ is a well known problem in linear modeling, where $K = L$ is a linear subspace and an unbiased estimator is usually defined as $\widehat{\sigma}^2_{UB} = \frac{\left\|y-\widehat{\theta_K}\right\|^2}{n-D_K(y)}$, in which $D_K(y) = p + 1$ is the total number of coefficients in the linear model.

Sampson et al. (2003) deal with the problem in monotone regression when $n = 2$ and propose the use of $\widehat{\sigma}^2_{UB}$. Also, Rueda et al. (2010) use the latter estimator to estimate the variance in a univariate monotone mixed model.

On the other hand, in monotone univariate regression problems, Meyer and Woodroofe (2000) propose an estimator for σ^2 that corrects the bias, as follows:

$$\widehat{\sigma}^2_{MW} = \frac{\left\|y - \widehat{\theta}_K\right\|^2}{n - Min(1.5D_K(y), n/2)}.$$

The estimator above uses the penalty $1.5D_K(y)$ instead of $D_K(y)$ in the denominator. An insight behind this correction is given below in Lemma 3.1. Moreover, the final version of the penalty as $Min(1.5D_K(y), n/2)$ assure that $\widehat{\sigma}^2_{MW}$ has good asymptotic properties (Meyer and Woodroofe (2000)).

In this paper, two estimators are defined for the semiparametric model following the research of these authors. Lemma 3.1, below, shows an inequality that will be useful to derive the estimators. The proof of the lemma is deferred to the appendix.

Lemma 3.1. *Let $y = \theta + \varepsilon$, $\theta \in C; \varepsilon \sim N_n(0, \sigma^2 I)$ where, $C = L + C_0$, L is a linear subspace of dimension r and C_0 a convex cone. Then, $\exists c \in \Re, 2 \geq c \geq 1$ such as:*

$$E\frac{\left\|y - \widehat{\theta}_C\right\|^2}{\sigma^2} = n - c(ED_C(y) - r) - r.$$

From lemma 3.1, applied to $K = L_0 + S_1 + ... + S_q$, we have the following equation,

$$E\frac{\left\|y - \widehat{\theta}_K\right\|^2}{\sigma^2} = n - c(ED_K(y) - p - 1) - (p + 1),$$

for a given unknown c, $2 \geq c \geq 1$.

The quantity c depends on θ but it is not strainforward to estimate. We consider two values that have also been considered before in the literature. Firstly, the smallest value, $c = 1$, is considered as it is the natural extension to the linear case and it has been successfully used by Rueda et al (2010), even though it provides a positive biased estimator. Secondly, in order to correct the positive bias, we consider $c = 1.5$ as a compromise intermediate value that has been also considered by Meyer and Woodroofe (2000). These authors have studied the behavior of other choices and their recommendation is to use $c = 1.5$ to correct the bias of the estimator with the inclusion of an upper limit. The corresponding estimators for σ^2 are given by:

$$\widehat{\sigma}^2_{K,1.5} = \frac{\left\|y - \widehat{\theta}_K\right\|^2}{n - Min(1.5D_K(y) - 0.5(p+1), n/2)}; \quad \widehat{\sigma}^2_{K,1} = \frac{\left\|y - \widehat{\theta}_K\right\|^2}{n - D_K(y)}.$$

In addition, the performance of both estimators is compared with the simulation experiments below, showing that the MSE are very close to each other relative to that of the MLE.

The AIC approach is considered for the second question. The AIC is a very popular criterion to model selection in a broad class of statistical problems. From the definition of the first AIC statistics, Akaike (1973), there has been a very large number of papers dedicated to the definition of alternative AIC for specific applications. Isotonic models are no exception: Kato (2009) proposes to use a standard AIC statistic with penalty term equal to $2D_K(y)$, while Zhao and Peng (2002) and Liu et al (2009) propose AIC measures with smaller penalty terms. However, these measures have not been validated in regression scenarios. Different proposals to solve both questions are provided and validated in this paper.

Now, in order to define the AIC measures, let us consider for the moment that σ is a known parameter. Usually, the AIC is defined as a penalized loglikelihood. $AIC(\widehat{\theta}) = -2l(\widehat{\theta}) + 2k$, where k is the number of parameters in the model and accounts for the bias when estimating the expected log-likelihood ($l(\widehat{\theta})$. In the context of the model subject to restrictions on the parameters given by a cone K, Anraku (1999) shows that the bias is the following quantity:

$b(\theta) = \frac{n}{2} + \frac{1}{2\sigma^2} E_\theta(\left\|\theta - \widehat{\theta}_K\right\|^2 - \left\|y - \widehat{\theta}_K\right\|^2)$.

It is straight forward to show that:

$b(\theta) = \frac{1}{2\sigma^2} E_\theta(\|y-\theta\|^2 + \left\|\theta - \widehat{\theta}_K\right\|^2 - \left\|y - \widehat{\theta}_K\right\|^2) =$

$= \frac{1}{\sigma^2} E_\theta(< y-\theta, \widehat{\theta}_K - \theta >) = E_\theta(D_K(y))$,

where the last equality follows from Stein's (1981) lemma.

Thus, we can define an AIC criterion for restricted regression when σ is known as follows:

$$AIC(\widehat{\theta}_K) = -2l(\widehat{\theta}_K) + 2E_\theta D_K(y).$$

In the framework of simple order restricted mean problems, several authors have dealt with the question, giving different proposals. Anraku(1999) proposes using $ORIC_A(\widehat{\theta}) = -2l(\widehat{\theta}) + 2B$, where $B = \inf_\theta E_\theta(D_K(y)) = E_{\theta_0}(D_K(y))$, where $\theta_0 \in L_K$, the largest subspace verifying $L_K \subset K$. Then, B is usually too small, except when $\theta \in L_K$, which is a very strong assumption.

Also, Zhao and Peng (2002) and Liu et al. (2009) propose the use of a penalty term defined by $2\lambda D_K(y)$, where $\lambda < 1$ and is chosen from different values, depending on the number of replications. These authors focus on specific questions, like the detection of the multiplicity of the largest parameter.

They consider simulations with very small values of n ($n \leq 5$) and compare the performance of the new criterion only against the $ORIC_A(\widehat{\theta}_K)$ and the standard AIC with a penalty term equal to $D_K(y)$. These results are not very relevant for our problem, as n is much higher in regression contexts, there are no replications and the focus is on the model selection within a wider family of models. However, for comparative purposes, an AIC measure defined using a penalty $2_K(y)$ with $\lambda < 1$ has also been considered, to show the effect of reducing the penalty. From the several choices in the literature, which usually depends on n, (see Zhao and Peng (2002)), we have selected, for our simulations, the value $\lambda = 0.75$. It is not worth to test other alternatives because numerical result clearly point out that $\lambda = 1$ is the best choice in this context.

Moreover, there is the problem of σ being unknown. It is hopeless to derive the corresponding AIC criterion in this case if the properties of the candidate estimators for σ cannot be obtained. A simple approach is therefore followed and the following general criterion is proposed:

$$AIC_{\lambda,\mu}(\widehat{\theta}_K) = -2l(\widehat{\theta}_K, \widehat{\sigma}_{K,\mu}) + 2\lambda D_K(y),$$

where $\lambda \in \{1, 0.75\}$ and $\mu \in \{1, 1.5\}$. In preliminary studies, smaller values for λ have also been considered but, as their performance is bad, they are not include here in order to simplify the output.

For a given pair (λ, μ), let $K^* = \arg_K \min AIC_{\lambda,\mu}(\widehat{\theta}_K)$ be the cone associated to the selected model and $\widehat{\sigma}^* = \widehat{\sigma}_{K^*,\mu}$ the corresponding estimator for σ. The MSE of $\widehat{\sigma}^*$ will be estimated in the simulations, besides the probability of correct detection, to determin the goodness of the $AIC_{\lambda,\mu}(\widehat{\theta}_K)$ criterion.

On the other hand, when the MLE for σ is used instead of $\widehat{\sigma}_{K,\mu}$ the AIC statistic reduces to:

$$AIC_{\lambda,MLE}(\widehat{\theta}_K) = \log(\widehat{\sigma}_{MLE}) + 2\lambda D_K(y),\ \lambda \in \{1, 0.75\}\,.$$

$AIC_{1,MLE}(\widehat{\theta}_K)$ is the most widely used AIC statistic in the literature when σ is unknown. As shown below, the performance of $AIC_{1,MLE}(\widehat{\theta})$ is worse than the new proposals in most scenarios.

Finally, for the case when σ is known, we consider:

$$AIC_{\lambda}(\widehat{\theta}_K) = -2l(\widehat{\theta}_K) + 2\lambda D_K(y),\ \lambda \in \{1, 0.75\}\,.$$

The combination of the different options gives eight criteria (six for the case when σ is unknown and two for when σ is known) that are compared with simulation experiments in the next subsection.

3.1.1. Monte Carlo studies

Two simulation experiments, A and B, have been conducted. A uses simulated independent explanatory variables and B uses the explanatory variables from the example in section 4. In both cases, the sample size equals 102, which is the sample size in the example.

The data generating model in experiment A is an additive regression model of the form:

$y = \sum_{j=1}^{3} m_j(u_j) + \varepsilon,$

with predictor vector u_1, u_2, u_3 i.i.d $U[0,1]^{102}$, $\varepsilon \sim N_{102}(0, \sigma^2 I)$ and $\sigma \in \{1, 5\}$ .The functions $\sum m_j()$ have been defined, in four different forms, following the suggestions of other authors (Borra and Ciaccio(2002), Curtis and Ghosal(2010) and Yang(2008)) as follows:

$$\begin{aligned} M1 &: exp(1.1u_1^3) + log((e^2-1)u_2 + 1) - \sin(2\pi u_3); \\ M2 &: exp(1.1u_1^3) - \sin(2\pi u_2) + u_3; \\ M3 &: -\sin(2\pi u_1) + u_2 + 5u_3; \\ M4 &: u_1 + 2u_2 + 3u_3. \end{aligned}$$

$M1$ is a nonparametric model in the three components ($p = 0, q = 3$), $M2$ is nonparametric in the first two components and linear in the third ($p = 1, q = 2$), $M3$ is nonparametric in the first component and linear in the rest ($p = 2, q = 1$) and $M4$ is a linear model ($p = 3, q = 0$). In each scenario, the same four models have been fitted with intercept.

On the other hand, the data generating model in experiment B, is given by:

$y = \sum_{j=1}^{3} h_j(x_j) + \varepsilon,$

the predictor vectors x_1, x_2, x_3 being the explanatory variables in the Prestige data set (section 4), $\varepsilon \rightsquigarrow N_{102}(0, \sigma^2 I)$ and $\sigma \in \{2, 10\}$. The selection of other values for σ to those in experiment 1 allows to show the performance of the approaches under different uncertainty levels. The functions $h_j()$ have been defined to imitate the relationship between the explanatory variables and the Prestige Score, the response in the real problem, in four different

$\widehat{\sigma}$	σ	M1	M2	M3	M4	H1	H2	H3	H4
$\widehat{\sigma}_{1.5}$,	low	0.0086	0.0076	0.0066	0.0056	0.3527	0.1795	0.0306	0.0245
$\widehat{\sigma}_1$	low	0.0108	0.0056	0.0060	0.0056	0.3249	0.2131	0.0435	0.0245
$\widehat{\sigma}_{MLE}$	low	0.0330	0.0131	0.0069	0.0401	0.9850	0.7440	0.2013	0.0269
$\widehat{\sigma}_{1.5}$	high	0.1535	0.1604	0.1459	0.1406	1.0487	0.9070	0.6305	0.6124
$\widehat{\sigma}_1$	high	0.1951	0.1648	0.1483	0.1406	2.1531	1.6040	0.7622	0.6124
$\widehat{\sigma}_{MLE}$	high	0.4897	0.3194	0.2158	0.1609	6.8484	4.6655	1.7758	0.6731

Table 1: MSE for $\widehat{\sigma}_{1.5}$, $\widehat{\sigma}_1$, and $\widehat{\sigma}_{MLE}$ under different scenarios.

forms, as follows:

$$\begin{aligned} H1 &: 15 log x_1 + x_2^4/2000 - log(x_3+1); \\ H2 &: 15 log x_1 + x_2^4/2000; \\ H3 &: 8(x_1 - \overline{x}_1)/s_{x_1} + x_2^4/2000; \\ H4 &: 8(x_1 - \overline{x}_1)/s_{x_1} + (x_2 - \overline{x}_2)/s_{x_2}. \end{aligned}$$

$H1$ is a monotone model in the three components, $H2$ is monotone in the first two components, $H3$ is linear in the first component and monotone in the second and $H4$ is a linear model in the first two components. In each scenario the same four models have been fitted.

A total of 16 scenarios, eight models with low and high values for σ, have been simulated, 100 replications were generated in each scenario and for each data set four models were fitted being among them the correct one. The simulation addresses two questions. First, which estimator of σ is preferred in terms of MSE?. Second, which of the AIC statistics defined in section 3.1 performs better?. For the first question, assuming the correct model, the empirical estimators of the MSE are derived for the new proposals. For the second question, the four models fitted are considered and for each of the 16 scenarios and each criterion, $AIC_{\lambda,\mu}$, the frequency of correct detection (FCD) is derived and the empirical estimator of the MSE of $\widehat{\sigma}^*$ are obtained, where $\widehat{\sigma}^*$ has been defined in section 3.1. A good criterion should be one with a high FCD that gives an accurate estimator for σ under the selected model. The FCD of AIC_λ and $AIC_{\lambda,MLE}$ and the MSE of σ_{MLE} have also been obtained for comparative purposes.

The results are given in tables 1, 2, 3 and 4.

AIC	σ	M1	M2	M3	M4	H1	H2	H3	H4	Mean
$AIC_{1,1.5}$	low	0.78	**0.41**	**0.53**	**0.61**	**0.99**	0.53	0.15	**0.70**	**0.59**
$AIC_{1,1}$	low	0.90	0.40	0.44	0.36	0.93	**0.59**	0.16	0.46	0.53
$AIC_{1,MLE}$	low	0.99	0.36	0.23	0.25	**0.99**	0.53	0.02	0.32	0.46
$AIC_{.75,1.5}$	low	0.98	0.30	0.21	0.13	**0.99**	0.47	0.01	0.26	0.42
$AIC_{.75,1}$	low	**1.00**	0.20	0.10	0.05	0.98	0.54	0.02	0.10	0.37
$AIC_{.75,MLE}$	low	**1.00**	0.17	0.04	0.01	**0.99**	0.47	0.01	0.04	0.34
AIC_{1}	low	0.79	0.42	0.52	0.73	0.58	0.72	0.70	0.78	0.66
$AIC_{.75}$	low	1.00	0.26	0.16	0.16	0.75	0.63	0.28	0.39	0.45
$AIC_{1,1.5}$	high	0.13	0.09	0.15	**0.59**	0.32	0.59	**0.25**	**0.58**	**0.34**
$AIC_{1,1}$	high	0.18	0.12	0.14	0.55	0.43	0.58	0.19	0.53	**0.34**
$AIC_{1,MLE}$	high	0.24	0.11	0.15	0.48	0.51	0.58	0.14	0.49	**0.34**
$AIC_{.75,1.5}$	high	0.46	**0.17**	**0.17**	0.24	0.45	**0.61**	0.21	0.31	0.33
$AIC_{.75,1}$	high	0.50	0.16	0.12	0.18	**0.60**	0.48	0.12	0.26	0.30
$AIC_{.75,MLE}$	high	**0.58**	0.10	0.10	0.16	0.74	0.42	0.11	0.16	0.30
AIC_{1}	high	0.13	0.09	0.12	0.59	0.23	0.64	0.23	0.72	0.34
$AIC_{.75}$	high	0.48	0.16	0.16	0.26	0.45	0.62	0.20	0.33	0.33

Table 2: FCD for different AIC statistics and simulated sccenarios.

AIC	σ	M1	M2	M3	M4	H1	H2	H3	H4
$AIC_{1,1.5}$	low	0.0087	0.0069	0.0063	0.0089	0.3531	0.2233	0.1454	0.0315
$AIC_{1,1}$	low	0.0084	0.0070	0.0097	0.0191	0.3249	0.2536	0.2050	0.1305
$AIC_{.75,1.5}$	low	0.0114	0.0068	0.0063	0.0095	0.3531	0.2252	0.1501	0.0402
$AIC_{.75,1}$	low	0.0106	0.0069	0.0108	0.0228	0.3269	0.2545	0.2130	0.1125
$AIC_{1,1.5}$	high	0.1535	0.1464	0.1516	0.1563	1.0950	1.0444	0.8907	0.7613
$AIC_{1,1}$	high	0.1812	0.1763	0.1828	0.1928	2.1289	2.0359	1.6390	1.0471
$AIC_{.75,1.5}$	high	0.1586	0.1476	0.1547	0.1590	1.0256	0.9962	0.9109	0.7924
$AIC_{.75,1}$	high	0.1995	0.1853	0.2064	0.2187	2.1790	2.0205	1.7979	1.1877

Table 3: MSE of the corresponding $\widehat{\sigma}$ for each criterion and simulated model.

3.2. Conclusions

As explained below, there are clear, winning candidates among the estimators for σ and the AIC criteria, namely, the bias corrected estimator, $\widehat{\sigma}_{1.5}$, and the AIC defined using this latter estimator and a penalty term equal to $2D_K(y)$, which corresponds with $AIC_{1,1.5}$.

Table 1 gives the MSE of the three estimators for σ for the 16 different model choices. Assuming that the true model is known, $\widehat{\sigma}_{1.5}$ outperforms $\widehat{\sigma}_1$ in most scenarios. Only for low σ and selected scenarios have we found that $\widehat{\sigma}_1$ has a smaller MSE. Moreover, compared with the $\widehat{\sigma}_{MLE}$, both estimators have a smaller MSE in the 16 scenarios. Note that in the particular case of linear models $M4$ and $H4$ the MSE of the $\widehat{\sigma}_{1.5}$ equals the MSE of $\widehat{\sigma}_1$ which is the unbiased estimator in the linear model framework.

In table 2, the FCD for the AIC measures considered is shown. Attending to the figures in the last column of the table, where the mean value of the FCD across scenarios is obtained, it can be concluded that when σ is assumed unknown, $AIC_{1,\mu}$ outperforms $AIC_{.75,\mu}$ and that with the new proposals, $AIC_{1,\mu}$ outperforms the classical $AIC_{1,MLE}$.

On the other hand, when σ is assumed to be known, AIC_1 is also preferred to $AIC_{.75.}$ Note that these two measures are only comparable among themselves as sigma is assumed known.

Looking at each scenario (the best performer is given in bold), the best behavior is again exhibited by $AIC_{1,1.5}$, which outperforms the rest in terms of FCD, except when the true model is complex, which is favored by $AIC_{.75,\mu}$ or $AIC_{.75,MLE}$. However, $AIC_{.75,\mu}$ performs badly in the rest of the scenarios and $AIC_{.75,MLE}$ even worse.

From the results in table 2, it could be concluded that the AIC statistic defined using $\widehat{\sigma}_{1.5}$, and with a penalty term equal to $2D_K(y)$, is the best performer.

In table 3, the $\widehat{MSE}$ of $\widehat{\sigma}^*$ are given for each scenario and the new criteria. These figures are useful to evaluate each criterion as an estimation method, as this also gives insights about the behavior when the selected model is not the correct one. In terms of $\widehat{MSE}(\widehat{\sigma}^*)$, $AIC_{1,1.5}$ also gives the best results, except when the true model is complex. However, even in these settings, this criterion is the second best performer (except for $H1$ and σ low), and the $\widehat{MSE}(\ \widehat{\sigma}^*)$ are close to those in table 1 obtained under the

AIC	f_c	f_s
$AIC_{1,1.5}$	0.55	0.45
$AIC_{1,MLE}$	0.71	0.29

Table 4: Frequency of complex f_c and simple f_s model selection.

true scenario. In order to show if $AIC_{1,1.5}$ favors complex or simple models compared with the standard $AIC_{1,MLE}$, the frequency of selection of complex models $(M1, M2, H1, H2)$ and simple models $(M3, M4, H3, H4)$ has been included in the 1600 samples generated using both criteria, and the results are given in table 4. From the figures in table 4, it can be concluded that the criterion $AIC_{1,1.5}$ favors complex models, but in a less extended form than $AIC_{1,MLE}$ does.

4. Prestige data

The Canadian occupational prestige data from the census 1971 (Fox(1997)) is a popular data set that has been analyzed by several authors. A recent reference where this data set has been analyzed is Griffin and Steel (2010) where a fully bayesian nonparametric approach is adopted. Prestige score(y) on 102 occupations, is linked to three explanatory variables, average income (in \$1000s)(x_1), education (in years) (x_2) and the percentage of incumber that are women (x_3). It is assumed that the score increases with the values of x_1 and x_2 and decreases with x_3.

Eight candidate models has been considered according to how the auxiliary information is used to obtain the estimators. Other models have been discarded as they give clearly worse fits. The description of the models is given in table 1, within the values of $AIC_{1,1.5}$, $\widehat{\sigma}^2_{1.5}$, $D(y)$, and $AIC_{1,MLE}$.

From the $AIC_{1,1.5}$ values in table 5, the model m_5, which includes x_1 in a linear form and x_2 in a nonparametric form with 15 degrees of freedom and $\widehat{\sigma} = 7.37$, is selected. It is interesting to note that using $AIC_{1,MLE}$ the most complex model which includes the three explanatory variables in a nonparametric form, would have been selected. This fact agrees with the conclusions following the simulation results, that $AIC_{1,MLE}$ favors the more complex models.

Other authors have adopted a fully nonparametric approach to analyze these data, using x_1 and x_2, but this is the first time a semiparametric model has been proposed, and also the first time a model selection criterion that

model	linear	monotone	$AIC_{1,1.5}$	$\widehat{\sigma}^2_{1.5}$	$D(y)$	$AIC_{1,MLE}$
m_1	x_2	-	554.562	82.869	2	554.543
m_2	-	x_2	547.911	73.203	16	544.548
m_3	x_1, x_2	-	524.307	60.998	3	524.261
m_4	x_2	x_1	525.937	59.028	19	523.912
m_5	x_1	x_2	**518.014**	54.338	15	515.369
m_6	-	x_1, x_2	524.394	54.541	29	511.188
m_7	x_1, x_2, x_3	-	526.253	61.567	4	526.172
m_8	-	x_1, x_2, x_3	525.501	54.867	30	**511.145**

Table 5: Model description and fitting results with the Prestige data.

tries to choose between parametric and nonparametric alternatives has been used.

5. Conclusions and future research

The problem of the derivation of the degrees of freedom, $D_K(y)$, for semiparametric monotone models has been solved. This quantity is incorporate in new AIC measures and in the estimators for the variance parameter, which are shown to be useful in model selection. Besides, $D_K(y)$ is also useful to derive inferential tools as hypothesis tests in nested models which is a question to be dealt with in our future research.

It has been shown, by simulation experiments and in the example, that semiparametric models compares favorably with linear alternatives, being the fitting and model selection steps easily achieved. For researchers who dislike the non continuity of the monotone fitted curve with the cyclic PAVA, the consideration of a two-step fitting process is proposed. In a first step, the linear and nonparametric terms defining the model are determined with the $AIC_{1,1.5}$ and, in the second step, the nonparametric modeling is performed using alternative approaches, following the researcher's preferences. Within these alternatives is a hybrid approach that produces monotone estimators, with properties similar to those of nonparametric regression estimators applying a smoother to each monotone component obtained from the backfitting algorithm (for details see Mukerjee(1988)).

Finally, there are several interesting extensions to model (1) that will also benefit from the results derived in the present paper.

The first extension is the semiparametric additive mixed model:

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ji} + \sum_{j=1}^{q} h_j(z_{ji}) + u + \varepsilon_i,$$

where $u \sim N(0, \sigma_u^2)$ is a random effect. Rueda et al. (2010), deal with this type of model, when $p = 0$ and $q = 1$, to solve small area estimation problems.

The second extension is the general isotonic model where the monotone restrictions are replaced by more general shape restrictions, including concave relationships among others. An interesting reference dealing with isotonic regression models is Meyer (2008), where univariate shape models ($p = 0$ and $q = 1$) are studied.

6. APPENDIX

6.1. Proof of lemma 2.1

(i) A polyhedral cone, K, is one that can be defined by a set of linear inequalities: $K = \{u \in \Re^n / a_i' u \leq 0, i = 1, ..., m\}$. Let us denote by $span^+ \{b_1, ..., b_s\}$ the subset of $\Re^n$ defined by the non negative linear combinations of $\{b_1, ..., b_s\}$.

It is a known result (Goldman and Tucker (1956)) that a convex cone C is a polyhedral cone, if and only if, it is finitely generated (with positive coefficients) by a finite number of vectors, that is:

$$\exists a_1, ..., a_m : C = \{u \in \Re^n / a_i' u \leq 0, i = 1, ..., m\}$$
$$\Leftrightarrow \exists b_1, ..., b_s : C = \left\{ u \in \Re^n / u = \sum_{i=1}^{s} \alpha_i b_i, \alpha_i > 0 \right\} = span^+ \{b_1, ..., b_s\}.$$

Therefore, as C_j are polyhedral cones, we have that, $C_j = span^+ \left\{b_{j1}, ..., b_{js_j}\right\}, j = 1, ..., q$, also, as L is a subspace, $\exists \{d_1, ..., d_l\}$ such as $L = span \{d_1, ..., d_l\} = span^+ \{d_1, ..., d_l\} + span^+ \{-d_1, ..., -d_l\}$. Then, $K = span^+ \{b_{11}, ..., b_{1s_1}\} + ... + span^+ \left\{b_{q1}, ..., b_{qs_q}\right\} + span^+ \{d_1, ..., d_l\} + span^+ \{-d_1, ..., -d_l\}$ is also a polyhedral cone.

(ii) let $y \in C$, $p(y/L^{\perp}) = y - p(y/L) \in C$ because $L \subset C$. Then, $p(y/L^{\perp}) = p(y/L^{\perp} \cap C)$ and $y = p(y/L) + p(y/L^{\perp} \cap C) \in L + L^{\perp} \cap C$. We have proven: $C \subset L + L^{\perp} \cap C$.

The opposite is also true as $L \subset C$.

Now, let $y \in \Re^n$, $p(y/C) = p(p(y/C)/L) + p(p(y/C)/L^{\perp})$ where, $p(p(y/C)/L^{\perp}) = p(y/C) - p(p(y/C)/L) \in C$, because $L \subset C$. Then,

$$p(p(y/C)/L^{\perp}) = p(p(y/C)/L^{\perp} \cap C) \text{ and } p(y/C) = p(p(y/C)/L) + p(p(y/C)/L^{\perp} \cap C)$$

Moreover, from lemma 2.2 in Raubertas (1986),

$$p(p(y/C)/L) = p(y/L) \text{ and } p(p(y/C)/L^{\perp} \cap C) = p(y/C \cap L^{\perp})$$

and (ii) follows.

(iii) From (ii), as $L \subset L + C$, we have that: $L + C = L + (L + C) \cap L^{\perp}$ and

$p(y/L + C) = p(y/L) + p(y/(L + C) \cap L^{\perp})$ and the result follows.

6.2. Proof of lemma 2.2

(i) To prove the result several properties of projections onto polyhedral cones will be used (see Meyer(1999) and references therein). From lemma 2.1 (ii), $K = L_0 + S_1 + ... + S_q = L_0 + ((L_0 + S_1 + ... + S_q) \cap L_0^{\perp}) = L_0 + K_0$, where $K_0 = (L_0 + S_1 + ... + S_q) \cap L_0^{\perp}$ and $\dim(L_0) = p + 1$. Also, from Lemma 2.1(iii),

$$p(y/K) = p(y/L_0) + p(y/K_0) \tag{5}$$

and $D_K(y) = p + 1 + D_{K_0}(y)$.

Now, from lemma 2.1(i), K_0 is polyhedral and then, K_0 is defined by a subset of generators, as follows,

$$\exists \{\delta_i, i = 1, ..., M\} \subset L_0^{\perp} : K_0 = span^{+} \{\delta_i, i = 1, ..., M\} = \left\{ u \in L_0^{\perp} / u = \sum_{i=1}^{M} b_i \delta_i, b_i \geq 0 \right\}$$

and also by a subset of inequality restrictions:

$$\exists \{\gamma_j, j = 1, ..., m\} \subset L_0^{\perp} : K_0 = \left\{ u \in L_0^{\perp} / u' \gamma_j \leq 0, j = 1, .., m \right\},$$

where $span^{+}\{\gamma_j, j=1,...,m\} = K_0^p \cap L_0^{\perp}$.

Moreover, from proposition 3 in Meyer(1999), for each $y \in L_0^{\perp}$ we can define the sets $I^y \subset \{1,...,M\}$, $J^y \subset \{1,..,m\}$ and the corresponding subspaces : $L(I^y) = span\{\delta_i, i \in I^y\}$, $L(J^y) = span\{\gamma_j, j \in J^y\}$ such as:

$$p(y/K_0) = p(y/L(I^y)), p(y/K_0^p) = p(y/L(J^y)) \tag{6}$$

and

$$L(I^y)^{\perp} \cap L_0^{\perp} = L(J^y), L(I^y) = (L(J^y))^{\perp} \cap L_0^{\perp}. \tag{7}$$

On the other hand, let be $\widehat{\theta}_j, j = 0, 1, ..., q$ a sequence given by the backfitting algorithm, we have that:

$$p(y/K) = \widehat{\theta}_0 + \widehat{\theta}_1 + ... + \widehat{\theta}_q, \widehat{\theta}_0 \in L_0, \widehat{\theta}_i \in S_j.$$

Let $v_j^i, i = 1, ..., n-1; j = 1, ...q$ be the generators of the order cones, then, $S_j = span^{+}\left\{v_j^i, i = 1, ..., n-1\right\}$. Now, for each $y \in \Re^n$ and each $\widehat{\theta}_j$, there exists a set of indexes I_j^y such as $L_j^y = span\left\{v_j^i, i \in I_j^y\right\}$ determine $F_j^y = L_j^y \cap S_j$, which is the face of the cone S_j that has $\widehat{\theta}_j$ as an interior point. Then, necessarily,

$$\widehat{\theta}_j = \sum_{i \in I_j^y} \lambda_{ij} v_j^i, \lambda_{ij} > 0, \forall i, j. \tag{8}$$

Moreover, from (5) and lemma 2.2 in Raubertas (1986), we have that:

$$p(y/K) = p(y/L_0) + p(y/K_0) = \widehat{\theta} = p(\widehat{\theta}/L_0) + p(\widehat{\theta}/K_0) = p(y/L_0) + p(\widehat{\theta}/K_0) \Longrightarrow$$
$$p(y/K_0) = p(\widehat{\theta}/K_0) = \widehat{\theta} - p(\widehat{\theta}/L_0) = \widehat{\theta}_1 + ... + \widehat{\theta}_q - p(\widehat{\theta}_1 + ... + \widehat{\theta}_q/L_0).$$

Now, from the last equality, (5), and (8) we have that,

$$p(y/K_0) = \sum_{j=1}^{q} \sum_{i \in I_j^y} \lambda_{ij} v_j^i - \sum_{j=1}^{q} \sum_{i \in I_j^y} \lambda_{ij} v_j^{iL} = \sum_{j=1}^{q} \sum_{i \in I_j^y} \lambda_{ij} v_j^{iK},$$

where, for each i and j, $v_j^i = v_j^{iL} + v_j^{iK}, v_j^i \in K, v_j^{iL} \in L_0, v_j^{iK} \in K_0$.

Now, from (6) and (7) it follows,

$\forall l \in L(J^y)$, $0 = \gamma_l' p(y/K_0) = \sum_{j=1}^{q} \sum_{i \in I_j^y} \lambda_{ij} \gamma_l' v_j^{iK}$.

However, $\lambda_{ji} > 0$ from (8), and $\gamma_l' v_j^{iK} \leq 0, \forall i \in I_j^y, j = 1, ..., q, l \in J^y$ (from the definition of K_0,) which implies, $\gamma_j' v_j^{iK} = 0, \forall i \in I_j^y, j \in J^y, l \in J^y$. This last property and (7) imply that,

$$\begin{aligned} v_j^{iK} &\in (L(J^y))^{\perp} \cap L_0^{\perp}, \forall i \in I_j^y, j \in J^y \Rightarrow v_j^{iK} \in L(I^y), \forall i \in I_j^y, j \in J^y \\ &\Rightarrow v_j^i = v_j^{iL} + v_j^{iK} \in L_0 + L(I^y), \forall i \in I_j^y, j \in J^y. \end{aligned}$$

Thus, we have proven that,

$$L_0 + L_1^y + ... + L_q^y \subset L_0 + L(I^y). \tag{9}$$

Now, from the equality $p(y/K) = p(p(y/K)/L_0 + L_1^y + ... + L_q^y)$, which is a consequence of $p(y/K) \in L_0 + L_1^y + ... + L_q^y$, obtained from the backfitting algorithm, and the statements (5) and (6) we have that,

$$\begin{aligned} p(y/K) = p(p(y/K)/L_0 + L_1^y + ... + L_q^y) &= p(p(y/L_0) + p(y/K_0)/L_0 + L_1^y + ... + L_q^y) \\ &= p(p(y/L_0) + p(y/L(I^y))/L_0 + L_1^y + ... + L_q^y). \end{aligned}$$

Finally, this last statement, the fact that L_0 and $L(I^y)$ are orthogonal subspaces and (9) imply:

$$p(y/K) = p(p(y/L_0 + L(I^y))/L_0 + L_1^y + ... + L_q^y) = p(y/L_0 + L_1^y + ... + L_q^y)$$

and the result follows.

(ii) The subspaces associated with faces of order cones have the following general expression:

$$L = \{u \in \Re^n / u_k = u_l, (k,l) \in P\}, \text{where } P \subset \{(k,l), k = 1, ..., n; l = 1, ..n\}.$$

Let u_0 be any vector belonging to L with a maximal number of different components. Let us denote this number by N and let us denote the different values by $\{u_{0i}\}_{i=1}^N$. Then, $N = dim(L)$ and a minimal set of generators for L is given by the set $\{v^i\}_{i=1}^N$, where:

$$v_k^i = 1 \Longleftrightarrow u_{0k} = u_{0i}; v_k^i = 0 \Longleftrightarrow u_{0k} \neq u_{0i}, k = 1, ..., n.$$

Now, the subspace associated with the face containing $\widehat{\theta}_j$ as an interior point is defined as follows: $L_j^y = \left\{u \in \Re^n / u_k = u_l, (k,l) \in P_j^y\right\}$, where

$P_j^y = \left\{(k,l)/\widehat{\theta}_{jk} = \widehat{\theta}_{jl}, k, l \in \{1, ..., n\}\right\}$. $\widehat{\theta}_j$ is a vector belonging to L_j^y with a maximal number of different components (otherwise other case $\widehat{\theta}_j$ would not be an interior point of the corresponding face). Moreover, if $\{a_{jk}\}_{k=1}^{N_j}$ are the values of the N_j different components of $\widehat{\theta}_j$, then $dim(L_j^y) = N_j$ and a minimal set of generators for L_j^y is given by the set $\left\{v_j^i\right\}_{i=1}^{N_j,}$ where,

$$v_{jk}^i = 1 \Longleftrightarrow \widehat{\theta}_{jk} = a_{ji}; v_{jk}^i = 0 \Longleftrightarrow \widehat{\theta}_{jk} \neq a_{ji}, k = 1, ..., n$$

and the result follows.

6.3. Proof of lemma 3.1

The proof follows the same steps to close result close to that found in the paper by Meyer and Woodofre(2000). From Stein's (1981) identity:

$$E\left\|y - \widehat{\theta}_C\right\|^2 = E\left\|y - \theta\right\|^2 - 2E < y - \theta, \widehat{\theta}_C - \theta > + E\left\|\widehat{\theta}_C - \theta\right\|^2 =$$

$$= n\sigma^2 - 2\sigma^2 E D_C(y) + E\left\|\widehat{\theta}_C - \theta\right\|^2. \tag{10}$$

Now, from lemma 2.1(iii), and the orthogonality between L and $L^{\perp}$, we have that:

$$E\left\|\theta - \widehat{\theta}_C\right\|^2 = E\left\|p(\theta/L) - p(y/L)\right\|^2 +$$

$$+E\left\|p(\theta/(L + C_0) \cap L^{\perp}) - p(y/(L + C_0) \cap L^{\perp})\right\|^2 \geq r\sigma^2. \tag{11}$$

Moreover, from properties of projections and the Stein identity again:

$$0 \leq E < y - \widehat{\theta}_C, \widehat{\theta}_C - \theta >= E < y - \theta, \widehat{\theta}_C - \theta > -E\left\|\widehat{\theta}_C - \theta\right\|^2 =$$

$$= \sigma^2 E D(y) - E\left\|\widehat{\theta}_C - \theta\right\|^2. \tag{12}$$

Now, from (10)and (11) we have that:

$$\frac{E\left\|y - \widehat{\theta}_C\right\|^2}{\sigma^2} \geq n - 2ED(y) + r = n - 2(ED(y) - r) - r, \tag{13}$$

and from (10)and (12) we have that:

$$\frac{E\left\|y-\widehat{\theta}_C\right\|^2}{\sigma^2} \leq n - ED(y) = n - (ED(y) - r) - r. \tag{14}$$

Then, the result follows from (13) and (14).

Acknowledgements: This research was partially supported by Spanish DGES (grant MTM 2009-11161).

[1] Akaike, H. (1973). Information Theory and the extension of the maximum likelihood principle. In: *Petrov, B. N., Csaki, F. (Eds), Proceedings of the Second International Symposium on Information Theory.* Akademiai Kiado, Budapest, pp. 267-281.

[2] Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika.*, 86, pp. 141-152.

[3] Bachettti, P. (1989). Additive isotonic models. *J. Amer. Statist. Assoc.*, 84, pp.289-294.

[4] Borra,S. Di Ciaccio, A. (2002). Improving nonparametric regression methods by bagging and boosting. *Comp. Stat. Data. An.*, 38 pp. 407-420

[5] Brunk, H.D. (1970). Estimation of isotonic regression (with discussion). In: *Puri, M.L. (Ed). Nonparametric Techniques in Statistical Inference.* Cambridge University Press, Cambridge.

[6] Cheng, G. (2009). Semiparametric additive isotonic regression. *J. Stat.Plan.Infer.*, 130, pp. 1980-1991.

[7] Curtis, M. S. and Ghosal, S. (2010). Fast Bayesian Mod
el Assessment for Nonparametric Additive Regression. www4.stat.ncsu.edu/ ghoshal/papers/BayesVarSel.pdf

[8] De Boer, W.J., Besten, P.J. and Ter Braak, C. F. (2002). Statistical analysis of sediment toxicity by additive monotone regession splines. *Ecotoxicology.* 11, pp. 435-50.

[9] Dykstra, R. (1983). An algorithm for restricted least squares regression. *Ann.Statist.*, 3, pp. 401-421.

[10] Fox, J. (1997). *Applied regression analysis, linear models and related methods.* Sage Publications, McMaster. University, Hamilton, Ontario, Canada.

[11] Goldman, A.J. and Tucker, A. W. (1956). Polyhedral convex cones. *Ann. of Math. Studies.*, 38, pp. 19-40.

[12] Griffin, J.E. and Steel, M.F.J. (2010). Bayesian nonparametric modelling with the Dirichlet process regression smoother. *Stat. Sinica.* (forthcoming).

[13] Hanson, D.L., Pledger, G. and Wright, F.T. (1973). On consistency in monotonic regression. *Ann.Statist.*, 3, pp. 401-421.

[14] Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statist.Sci.*, 3, pp. 297-310.

[15] Huang, J. (2002). A note on estimating a partly linear model under monotonicity constraints. *J. Stat.Plan.Infer.*, 107, pp. 345-351.

[16] Hussian, M., Grimvall, A., Burdakov, O. and Sysoev, O. (2004). Monotonic regression for assessment of trends in environmental quality data. *ECCOMAS 2004.* P. Neittaanmäki, T. Rossi, K. Majava, and O. Pironneau (eds.). V. Capasso and W. Jäger (assoc. eds.). Jyväskylä.

[17] Kato, K. (2009). On the degrees of freedom in shrinkage estimation. *J. Multivariate Anal.*, 100, pp. 1338-1352.

[18] Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.*, 36, pp. 261-286.

[19] Liu, T., Lin, N., Shi, N. and Zhang, R.(2009). Information criterion-based clustering with order-restricted candidate profiles in short time-course microarray experiments. *Bioinformatics.*, 10:146.

[20] Mammen, E. and Yu, K. (2007). Additive isotone regression. *Lecture. Notes-Monograph Series.*, 55, pp. 179-195.

[21] Meyer, M. (1999). An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. *J. Stat.Plan.Infer.*, 81, pp. 13-31.

[22] Meyer, M. and Woodrofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann.Statist.*, 28, pp. 1083-1104.

[23] Meyer, M. (2008). Inference using Shape-restricted regression Splines. *Ann.Statist.*, 2, pp. 1013-1033.

[24] Morton-Jones, T., Diggle, P., Parker, L., Dickinson, H.O. and Binks, K.(2000). Additive isotonic regression models in epidemiology. *Stat.Med.*,9, pp. 849-59.

[25] Mukerjee, H. (1988). Monotone nonparametric regression. *Ann.Statist.*,16, pp. 741-750.

[26] Raubertas, R.F., Lee C.C. and Nordheim, E.V. (1986). Hypothesis Test for normal means constrained by linear inequalities. *Commun.Statisti.-Theor.Meth.*, 15, pp. 2809-2833.

[27] Robertson, T., Wright, F.T., Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. Wiley.

[28] Rueda, C., Menéndez, J.A. and Gomez, F. (2010). Small Area Estimators based on restricted Mixed models. *TEST.*, 19, pp. 558-579.

[29] Sampson, A. R., Singh, H. and Whitaker, L.R. (2003). Order restricted estimators: some bias results. *Stat.Probabil.Lett.*, 61, pp. 299-308.

[30] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann.Statist.*, 9, 1135-1151.

[31] Stone,C.J. (1982). Optimal Global rates of convergence for nonparametric regression.*Ann.Statist.*, 10, 1040-1053.

[32] Stone,C.J. (1982). Additive Regression and other nonparametric models.*Ann.Statist.*, 13, 689-705.

[33] Yang, L. (2008). Confidence Band for Additive Regression Model. *Journal of Data Science.*, 6, pp. 207-217

[34] Zhao, L. and Peng, L. (2002). Model selection under order restrictions. *Stat.Probabil.Lett.*, 57, pp. 301-306.

[35] Xiao, N, Zhang,D. and Zhang,H.H.(2010) Variable selection for semi-parametric mixed models in longitudinal studies. *Biometrics.*, 66, pp. 79-88.