Petr Volf

Moving window estimation procedures for additive regression function

Kybernetika, Vol. 29 (1993), No. 4, 391--402

Persistent URL: http://dml.cz/dmlcz/125627

KYBERNETIKA — VOLUME 29 (1993), NUMBER 4, PAGES 389 - 400

MOVING WINDOW ESTIMATION PROCEDURES FOR ADDITIVE REGRESSION FUNCTION[1]

PETR VOLF

The general additive regression function $b(\mathbf{x}) = \sum b_j(x_j)$ is considered and subjected to nonparametric estimation. The method of estimation is inspired by the regressogram approximations to the components of regression function. The procedure using the moving window is then derived, it naturally generalizes to a kernel approach. The method can be applied to the likelihood-based models, in which the value of regression function is a parameter of likelihood of a response variable Y. Suggested moving window algorithm is a variant of Hastie and Tibshirani's [3] local scoring procedure. In order to discuss the quality of obtained results, the method is compared with the approximation by regression splines, treated successfully by Stone [6]. An example illustrates the solution for the logistic regression, the proportional hazard regression model is also examined.

1. INTRODUCTION

The methods for nonparametric estimation and smoothing of curves are in the centre of attention of the data analysts for a long time. The modern equipment enables the statistician to examine the data attentively and to do deep preliminary analysis. Hence the nonparametric estimation of the covariate effect is at least a part of preliminary examination.

Let us consider a pair (X, Y) of real-valued random variables. In a regression model, X is called a covariate, meanwhile Y is a response variable. Let the general regression function be some smooth function $b(x)$, describing the dependence of a response variable Y on a covariate X. Likelihood-based regression model means that the value $b(x)$ is a parameter of likelihood for Y given $X = x$. Examples of this are the normal regression model, in which the regression function stands for $\mathsf{E}(Y|X = x)$, or the logistic regression model. We shall also mention the proportional hazard regression model for survival data.

If the observation is represented by a random sample (X_i, Y_i) of extent n, often the logarithm of likelihood can be expressed as

$$\ell_n = \sum_{i=1}^{n} \ell_1(Y_i,\, b(X_i)), \tag{1}$$

[1]This work was supported by Czech Academy of Sciences grant No. 27 557.

where ℓ_1 is a loglikelihood for one realization of Y, conditioned by a value of X. How to cope with the task of estimation of function b from the log-likelihood? One way may consist in approximation for $b(x)$, by a parametrized function. Every smooth function can be well approached by a linear combination from some basis of functions. For instance, the splines are the popular choice. Sleeper and Harrington [5] illustrate successfully the flexibility of regression splines in the analysis of the form of hazard ratio. Stone [6] used the approximation of regression function by splines in the framework of exponential family of distributions. He proved consistency of this approximation provided the parameters of splines were estimated by (global) maximum likelihood method. Thus, the reparametrization may be considered as an alternative way to solution. From this point of view, the regressogram is a trivial spline, with the order 0.

A widespreaded discussion runs about advantages and capabilities of both approaches – splines and kernel-like smoothing, cf. also discussion to paper of Hastie and Tibshirani [3]. The author does not intend to contribute to arguments of any side, his opinion is that every well-working method is valuable. Although some data-analysts (when joking) claim that one data may be analysed only once and only by one method – in order to avoid contradictions and problems with interpretation of results.

Our approach to estimation of regression function starts from a regressogram approximation. Then it proceeds to the moving window concept, considering simultaneously the additive regression function $b(\mathbf{x}) = \sum_j b_j(x_j)$ in the case of multi-dimensional covariate. It is necessary to stress at once that the additive model can include various transformations of covariates, their interactions (e. g. $x_1 \cdot x_2$), or, say, two-dimensional covariate, so that its idea seems to be sufficiently wide and flexible.

The general features of the method are described in the second part. Part 3 deals with the case of multi-valued logistic regression model. Part 4 considers a rather general case of a counting process with intensities fulfilling the proportional hazard model. The properties of solution are discussed in Part 5. Finally, an example with artificial data is solved numerically and discussed briefly.

Although the moving window procedure is a very flexible and easily computable method, its consistency is not guaranteed by any theoretical result. Only for the case of the normal regression model, in a more general concept of the Alternating Conditional Expectations (ACE) algorithm, Breiman and Friedman [2] show that the solution obtained by the moving window smoothing is the best additive approximation to $\mathsf{E}(Y|\mathbf{X})$. It means also that if $\mathsf{E}(Y|\mathbf{X} = \mathbf{x})$ is an additive function, it is consistently estimable by the moving window approach.

For a more general family of models, the results of Stone [6], mentioned above, can be used in order to support our conviction about the quality of the moving window smoothing. We discuss the connection between the moving window concept and the approximation by regression splines–polynomials on fixed windows.

Volf in [8] deals exclusively with the proportional hazard regression models and solves several simulated examples, in order to show a good performance of the method. In the example illustrating the paper of Sleeper and Harrington [5], the result of smoothing by the splines is compared graphically with the result obtained

by the local scoring.

2. LIKELIHOOD-BASED ESTIMATION PROCEDURE

Let us first consider the one-dimensional covariate X, with values in some finite interval $\mathcal{X} \subset R$. The construction of a regressogram means that the domain $\mathcal{X}$ is divided into M disjoint intervals I_m (their choice depends on the analyst), the function $b(x)$ is approximated as $\sum_{m=1}^{M} \beta_m \cdot \mathbf{1}[x \in I_m]$. Now, after inserting into the loglikelihood, the parameters β_m are estimated in ordinary way, which searches for solution of the equations $\partial \ell_n / \partial \beta_m = 0$, $m = 1, \dots M$. If the loglikelihood is of the form (1), then its first and second derivatives are

$$\frac{\partial \ell_n}{\partial \beta_m} = \sum_{i=1}^{n} \mathbf{1}[X_i \in I_m] \cdot \ell_1'(Y_i, \beta_m) \tag{2}$$

$$\frac{\partial^2 \ell_n}{\partial \beta_m \, \partial \beta_k} = \sum_{i=1}^{n} \mathbf{1}[X_i \in I_m] \cdot \ell_1''(Y_i, \beta_m) \quad \text{for } m = k, \quad = 0 \text{ otherwise.}$$

The step from estimation of the regressogram to the moving window estimation is quite straightforward. If we wish to estimate the value of $b(x)$ at a point $x = z$, we take b as a constant b_z in some chosen neighborhood (window) around z, say, in $\mathcal{O}(z)$. Then b_z is treated as a parameter, we have to solve the equation $\partial \ell_n / \partial b_z = 0$.

If the loglikelihood has the form (1), then

$$\frac{\partial \ell_n}{\partial b_z} = \sum_{i} \mathbf{1}\left[X_i \in \mathcal{O}_z\right] \cdot \ell_1'(Y_i, b_z). \tag{3}$$

Now, (3) contains only the derivatives of a "local" loglikelihood. It is the basis for the idea of the local scoring (or local likelihood) algorithm. However, when the form (1) does not hold, the derivatives do not contain the local results only. It is clearly visible in Example 2 which deals with the proportional hazard regression model.

Example 1. Logistic regression with two-valued response.
Let $\mathsf{P}(Y = 0 \,|\, x) = 1/(1 + \exp b(x))$, $\mathsf{P}(Y = 1 \,|x) = 1 - \mathsf{P}(Y = 0 \,|\, x)$. Then

$$\ell_n = \sum_{i=1}^{n} \{b(X_i) \cdot \mathbf{1}[Y_i = 1] - \ln(1 + \exp b(X_i))\}.$$

Example 2. Proportional hazard model for survival times and for i.i.d. sample $\{Y_i, \delta_i, X_i,\ i = 1, \dots, n\}$, where Y_i is an observed value and δ_i is the indicator of censoring. It means that $\delta_i = 1$ when Y_i is observed survival time, $\delta_i = 0$ if Y_i is less than survival time, the ith observation is censored at time moment Y_i. The inference for the hazard proportion $b(x)$ is based on the logarithm of Cox's partial

likelihood (cf. Andersen and Gill [1]), namely on

$$\ell_n = \sum_{i=1}^{n} \delta_i \ln \left\{ \frac{\exp b(X_i)}{\sum_{j=1}^{n} \exp b(X_j) \cdot I_j(i)} \right\},$$

where $I_j(i) = 1$ if $Y_j \geq Y_i$, $\quad I_j(i) = 0$ otherwise.

This partial likelihood has not the form (1). Nevertheless, let us compute its first derivatives with respect to value b_z in a neighbourhood $\mathcal{O}(z)$ of a point $z \in \mathcal{X}$:

$$\frac{\partial \ell_n}{\partial b_z} = \sum_i \delta_i \left\{ \mathbf{1}[X_i \in \mathcal{O}_z] - \frac{\exp b_z \cdot \sum_j I_j(i) \cdot \mathbf{1}[X_j \in \mathcal{O}_z]}{\sum_j \exp b(X_j) \cdot I_j(i)} \right\}. \tag{4}$$

The numerical iteration is the most frequently used way how to solve the likelihood equations. As a rule, the procedures need the second derivative of the loglikelihood, which in the case (1) yields

$$\frac{\partial^2 \ell_n}{\partial b_z^2} = \sum_i \mathbf{1}\{x_i \in \mathcal{O}_z\} \cdot \ell_1''(Y_i, b_z).$$

When the Newton–Raphson procedure is applied, the step from sth to $(s+1)$-st iteration is given by the following expression:

$$b_z^{(s+1)} = b_z^{(s)} - \frac{\partial \ell}{\partial b_z} \Big/ \frac{\partial^2 \ell}{\partial b_z^2}, \tag{5}$$

where the derivatives are evaluated at $b^{(s)}(x)$.

Hastie and Tibshirani [3] recommend to incorporate a smoothing directly into every step (5), they suggest the modification

$$b_z^{(s+1)} = \text{ smooth } \left[b_z^{(s)} - \frac{\partial \ell}{\partial b_z} \Big/ \text{ smooth } \left(\frac{\partial^2 \ell}{\partial b_z^2} \right) \right].$$

The notion of smoothing can have a very wide meaning, from weighted mean to, say, local parametrized regression.

Both examples mentioned above allow also another iteration procedure, which differs from (5) and which does not use the second derivatives. Moreover, after smoothing the results at each point, we shall "secondarily" smooth the final result. The procedure will be described in the following parts of the paper.

Let us now consider the K-dimensional covariate $\mathbf{X}$, with values in some bounded interval $\mathcal{X} \subset R_K$. When the dimension of $\mathbf{X}$ increases, the data are sparse and the method using the K-dimensional windows becomes ineffective. Then the additive "hypothesis" is available. The general additive regression model means that the regression function is

$$b(\mathbf{x}) = \sum_{k=1}^{K} b_k(x_k).$$

The component functions b_k should be nonparametrically estimated. The technique is essentially the same as for the one-dimensional case, but the inner loops has

to be incorporated to the procedure. This loop computes (at each sth step of the "outer" loop) successively all $b_k^{(s)}$, $k = 1, \ldots, K$, at all chosen points z_k. At least the values at all realized points x_{ki} are needed for further computation. Here k denotes the component, i denotes the case, $i = 1, \ldots, n$.

Let z be a point from the domain of X_1, say. The derivation of loglikelihood (1) with respect to $b_1(z)$ now yields

$$\frac{\partial \ell_n}{\partial b_1(z)} = \sum_{i=1}^{n} \mathbf{1}\,[X_{1i} \in \mathcal{O}_z] \cdot \ell_1' \left(Y_i,\, b_1(z) + \sum_{k=2}^{K} b_k(X_{ki}) \right).$$

It is seen that the actual estimates (i.e. estimates obtained from the last preceding step) of other component functions b_k, $k = 2, 3, \ldots, K$, have to be available.

3. LOGISTIC REGRESSION MODEL

Let Y be a random variable with $M + 1$ possible values from $\{0, 1 \ldots, M\}$. The logistic model describes the dependence of probability distribution of Y on a (K-dimensional) covariate $\mathbf{X}$. The model assumes that

$$\mathsf{P}(Y = 0 \,|\, \mathbf{X}) = 1/S(\mathbf{X}), \qquad \mathsf{P}(Y = y \,|\, \mathbf{X}) = \exp(C(y, \mathbf{X}))/S(\mathbf{X}),$$

$$\text{with} \quad S(\mathbf{X}) = 1 + \sum_{m=1}^{M} \exp C(m, \mathbf{X}),$$

when $y = 1, 2, \ldots, M$. Moreover, the additive version of the model considers additive functions $C(m, \mathbf{x}) = \sum_{k=1}^{K} C(m, k, x_k)$. The form of the log-likelihood has been sketched in Example 1, now it enlarges to

$$\ell_n = \sum_{i=1}^{n} \left\{ \sum_{m=1}^{M} \mathbf{1}[Y_i = m] \cdot \sum_{k=1}^{K} C(m, k, X_{ki}) - \ln S(\mathbf{X}_i) \right\}. \tag{6}$$

Our task consists in successive estimation of all functions $C(m, k, x)$ as a functions of $x = x_k$. Let us imagine that we have already got some estimates of the regression functions from the sth step of the outer loop. In order to proceed with $(s + 1)$-st step of estimation, we need to know the estimates of $C(m, k, x)$ at all realized points x_{ki} $i = 1, \ldots, n$, $k = 1, \ldots, K$. Let z be a point in the domain of X_j, $\mathcal{O}_z$ be its neighborhood (an interval around z). The actualized $(s + 1)$-st estimation of $f_m = C(m, j, z)$ is obtained from the solution of (local) likelihood equation

$$\frac{\partial \ell_n}{\partial f_m} = \sum_{i=1}^{n} \mathbf{1}\,[X_{ji} \in \mathcal{O}_z] \left\{ \mathbf{1}[Y_i = m] - \frac{\exp(f_m) \cdot \exp C_j(m, \mathbf{X}_i)}{S(\mathbf{X}_i)} \right\} = 0, \tag{7}$$

where $C_j(m, \mathbf{x}_i) = \sum_{k=1}^{K} \mathbf{1}[k \neq j]\, C(m, k, x_{ki})$. We can estimate the value of f_m simultaneously for all $m = 1, \ldots, M$. The equations (7), in which j and z are fixed, can be solved separately for each m, or it can be solved as an M-dimensional equation. When computing the example described in Part 6, we used separate evaluation

for one value of m after another. The procedure then proceeds to another point z in the domain of X_j. When the values of $C(m, j, x_{ji})$ in all realized points x_{ji} and for each m are estimated, $i = 1, \ldots, n$, $m = 1, \ldots, M$, the algorithm starts to compute the estimates of $C(m, j+1, x_{j+1,i})$. All these computations are a part of the inner loop. It iterates through all $j = 1, \ldots, K$. Only then the algorithm may proceed to a further $(s+2)$-nd step of the outer loop, which again runs for $j = 1$ to K. The iterations are repeated until the convergence of all estimated functions $C(m, k, \cdot)$. How can be the convergence of functions checked and recognized? After every step, for every component $C(m, j, \cdot)$ we can construct the optimal least squares line through the points $C(m, j, x_{ji})$, $i = 1, \ldots, n$. The changes of the parameters of the line can serve as a criterion of iteration progress and as an indicator of convergence.

The usual way how to solve (7) consists in an iteration with the help of the second derivative of loglikelihood, for instance it may follow the scheme (5). In our example with the logistic model,

$$\frac{\partial^2 \ell_n}{\partial f_m^2} = \sum_{i=1}^{n} \frac{\exp(f_m) \cdot \exp C_j(m, \mathbf{X}_i)}{S(\mathbf{X}_i)} \left\{ \frac{\exp(f_m) \cdot \exp C_j(m, \mathbf{X}_i)}{S(\mathbf{X}_i)} - 1 \right\} \cdot \mathbf{1}[X_{ji} \in \mathcal{O}_z],$$

if again $f_m = C(m, j, z)$, j and z are fixed. But the form of equation (7) suggests also another procedure of iterative estimation. If (7) is solved directly for f_m, it yields

$$f_m = -\ln \left\{ \sum_{i,z} \frac{\exp C_j(m, \mathbf{X}_i)}{S(\mathbf{X}_i)} \Big/ \sum_{i,z} \mathbf{1}[Y_i = m] \right\}, \tag{8}$$

where the sums are through $\{i = 1, \ldots, n : X_{ji} \in \mathcal{O}_z\}$. The "inner" iteration again proceeds through all $m = 1, \ldots, M$, then through all $z = x_{ji}$ (realized points), and it renovates successively the estimates of component functions for $j = 1, \ldots, K$.

4. PROPORTIONAL HAZARD REGRESSION MODEL

The model is a popular choice for the description of covariate effect in life events history analysis. Especially, the Cox model is an often used representative of the model. It is able to analyse the censored data, its semiparametric form can be identified easily. However, the Cox model restricts the log hazard ratio to be linear in the covariates. A proportional hazard model considering a more general hazard function has an intensity $\lambda(t|x) = a(t) \cdot \exp(b(\mathbf{x}))$, where $b(\mathbf{x})$ is an unspecified smooth function. The estimation of proper function b can be based on K-dimensional kernel procedure (Volf [7]).

However, a more-dimensional covariate causes the data sparse and the global kernel approach loses its effectivity. Therefore, let us return to the model of the additive influence of covariates to the log hazard. Now the log hazard ratio has K components, $b(\mathbf{x}) = \sum_{j=1}^{K} b_j(x_j)$. The analyst has to identify suitable functions $b_1, \ldots, b_K$ and also the underlying common hazard function $a(t)$, or better, its cumulative version $A(t) = \int_0^t a(s)\, ds$. Evident ambiguity (with respect to additional constant in b_j's) can be overcome by proper normalization of the functions. Volf [8]

describes a method of estimation for a particular but frequent case, when each object has constant values of covariates. The method consists in alternating sequential computing of functions b_j and A. The procedure has been tested successfully, by simulated examples as well as by real data.

In the sequel, we shall consider a more general design, based on the counting processes. We have simultaneously to enlarge the model and to allow the time-dependent (random) processes of covariates $\mathbf{X}_i(t)$, $i = 1, \ldots, n$. In fact, such a system ceased to have the proportional hazards, although, for fixed $\mathbf{X} = \mathbf{x}$, the proportional hazard model holds. The counting process $N(t) = N_1(t), \ldots, N_n(t)$ is a set of right-continuous random step functions on $[0, \mathcal{T}]$, with steps $+1$. It is assumed that no two components step simultaneously. In this model, the components need not to be i.i.d., the recurrent jumps are allowed. $N_i(t)$ simply counts the events of ith kind or of ith object in the life history.

The model is fully described by the (random) hazard rates for counting processes $N_i(t)$, namely $\lambda_i(t) = a(t) \cdot \exp b\,(X_i(t)) \cdot I_i(t)$, $i = 1, \ldots, n$, $t \in [0, \mathcal{T}]$, where $I_i(t)$ is an indicator of risk set. It means that $I_i(t) = 1$ if the ith object is in the risk set at moment t, $I_i(t) = 0$ otherwise. The inference is based on Cox's partial likelihood. Its logarithm is

$$\ell_n = \sum_{i=1}^{n} \int_0^{\mathcal{T}} \ln \frac{e^{b(X_i(t))}}{\sum_{j=1}^{n} e^{b(X_j(t))} I_j(t)} \, \mathrm{d}\, N_i(t).$$

By the way, if we define again the underlying baseline cumulative hazard function $A(t) = \int_0^t a(s)\, \mathrm{d}s$, there exists its generalized maximum likelihood estimator $\hat{A}(t) = \int_0^t \frac{\mathrm{d}\bar{N}(s)}{\sum_j \exp b(X_j(s)) I_j(s)}$, where $\bar{N} = \sum_{i=1}^{n} N_i$. In the frame of Cox's model, this estimator is strongly consistent and asymptotically normal. However, here the analogy with the survival time model ends.

Let us now return to the idea of the kernel (moving window, or m-nearest neighbor) estimate for function $b(x)$. Inspired by (4) of Example 2, dealing with the one-dimensional case, we may suggest the iteration scheme $b^{(s+1)}(z) = h\,(b^{(s)}, z)$, where

$$h(b, z) = \\ = -\ln\left[\sum_i \int_0^{\mathcal{T}} \frac{\sum_j \mathbf{1}\,[X_j(t) \in \mathcal{O}(z)]\, I_j(t)}{\sum_j \exp b(X_j(t)) I_j(t)} \mathrm{d}N_i(t) \Big/ \sum_i \int_0^{\mathcal{T}} \mathbf{1}\,[X_i(t) \in \mathcal{O}(z)]\, \mathrm{d}N_i(t)\right]. \quad (9)$$

Here $\mathcal{O}(z)$ denotes our moving window-neighborhood of point $z \in \mathcal{X}$.

Do not forget that $N_i(t) - s$ are the step-wise functions, with steps $+1$ at the moments of "counts". In a survival time model it corresponds to moments $(Y_i, \delta_i = 1)$. It is seen, that we need not register all trajectories of $X_j(t)$, but only their values $X_j(S_i)$, where S_i are the moments of counts of $N_i(t)$, and we are able to register them only if $I_j(S_i) = 1$.

Let us now consider the situation with multidimensional covariate processes $\mathbf{X}(t) = (X_1(t), \ldots X_K(t))$ and suppose the additive form of function b, $b(\mathbf{x}) = \sum_{j=1}^{K} b_j(x_j)$. Then the inner loop has to be incorporated to our iteration scheme. It

computes successively all components $b_1^{(s)}$ to $b_K^{(s)}$, then we proceed to the $(s+1)$-st step of "outer" iteration.

In order to obtain a generalization for iteration (9) with an additive regression function, let us imagine that when estimating, say, function $b_\ell(x_\ell)$, we have already estimated all functions $b_m(x_m)$, $m = 1, 2, \ldots, K$, during the preceding step of the outer loop.

Quite analogically to the one-dimensional case, from the equation $\partial \ell_n / \partial b_\ell(z) = 0$ we can suggest the following scheme for the moving window estimation procedure:

$$b_\ell^{(s+1)}(z) = -\ln\left[\sum_i \int_0^{\mathcal{T}} \frac{R_\ell(z, b^{(s)}, t)}{S_0(b^{(s)}, t)}\, \mathrm{d}N_i(t) \Big/ \sum_i \int_0^{\mathcal{T}} \mathbf{1}\left[X_{\ell i} \in \mathcal{O}_\ell(z)\right] \mathrm{d}N_i(t)\right],$$

where now

$$R_\ell(z, b, t) = \sum_{j=1}^{n} \mathbf{1}\left[X_{\ell j}(t) \in \mathcal{O}_\ell(z)\right] \cdot \exp\left\{\sum_{k=1}^{K} \mathbf{1}[k \neq \ell] \cdot b_k(X_{kj}(t))\right\} \cdot I_j(t),$$

$S_0(b, t) = \sum_{j=1}^{n} \exp\{b(\mathbf{X}_j(t))\} \cdot I_j(t)$ and $\mathcal{O}_\ell(z)$ is a chosen window around z in the domain of ℓth covariate.

The inner loop now iterates through $\ell = 1, \ldots, K$ and gives the values of $b_\ell^{(s+1)}(z)$ at every chosen z (we need at least the values at all observed $x_{\ell j}(T_i)$ provided $I_j(T_i) = 1$, where T_i are the moments of counts, $i, j = 1, \ldots, n$). The first inner loop may start from $b_1 = \cdots = b_K \equiv 0$ or from another convenient initial guess.

5. REMARKS ON CONSISTENCY

Stone [6] has examined the family of exponential-type regression models. Their loglikelihood has the form (1) with

$$\ell_1(Y, \mathbf{X}) = c(\theta(\mathbf{X})) \cdot Y + d(\theta(\mathbf{X})), \tag{10}$$

where c, d, e are known functions, θ is a regression function of our interest. The functions c, d are required to be twice continuously differentiable, with $c' > 0$. Stone has proved that under mild conditions a unique (as to the additive shift) additive function $b(\mathbf{x}) = \sum_{j=1}^{K} b_j(x_j)$ exists, closests to $\theta(\mathbf{x})$ in the sense of the Kullback–Leibler distance. Leaving this aspect of the problem apart, we assume that the regression function has already the additive form. From this point of view, the second result of Stone [6] is important. Stone has considered the polynomial splines (of chosen order) approximating each component b_j. Thus, the model is reparametrized by a finite number of parameters, they are then estimated by means of the standard (global) maximum likelihood method. It suffices to assume that:

1. The distribution of $\mathbf{X}$ is absolutely continuous on $\mathcal{X}$, with its density bounded away from zero and infinity.

2. Function b is Lipschitz continuous on $\mathcal{X}$.

3. The knots of the splines are chosen equidistantly and their number is proportional to n^γ, where γ is chosen properly from (0,1).

4. There are positive constants r and R such that

$$\mathsf{E}(\exp(s\,Y)|\mathbf{X}=\mathbf{x}) \leq R \quad \text{for} \quad |s| \leq r \quad \text{and} \quad \mathbf{x} \in \mathcal{X}.$$

Then, when n increases to infinity, this approximation by splines yields the consistent estimates of functions b_j. Stone gives also the order of convergence.

By the way, even the $M+1$ valued logistic regression model can be regarded as an M-dimensional representant of the exponential family. Let us recall its loglikelihood (6). It has the form (1), with

$$\ell_1(Y,\mathbf{X}) = \sum_{m=1}^{M} Y_m^* b^m(\mathbf{X}) - \ln\left\{1 + \sum_{m=1}^{M} \exp b^m(\mathbf{X})\right\},$$

where $Y_m^* = \mathbf{1}[Y=m]$.

Sometimes the likelihood is more complicated, e. g. in the case of the proportional hazard regression model. New results of Kooperberg et al. [4] prove consistency even for the spline approximation of hazard regression.

Let us now try to transfer the results of Stone to the moving window estimation. Let us recall again the regressogram approximation, in the case of loglikelihood (1) and one-dimensional covariate X. Its (global) maximum likelihood solution (2) leads in fact to the local likelihood iterations, because the matrix of the second derivatives is diagonal. Thus, the only difference between (3) and (2) consists in the use of the moving window instead of the fixed one. That is why the consistency property of Stone (which applies also to the regressogram – a "trivial" spline of order 0) holds also for the moving window solution. The sufficient conditions are the same as above, instead of increasing number of knots the decreasing width of window has to be considered, proportional to $n^{\gamma-1}$. Again, the result is well known in the case of the kernel estimation of regression function $\mathsf{E}(Y|X=x)$ in the normal regression model.

Unfortunately, the same statement does not hold when the additive regression function of multi-dimensional covariate is considered. Even in the framework of the exponential family of models, Hastie and Tibshirani [3] express a mere "conjecture" that their result of local scoring does not differ significantly from the approximation by splines.

However, the number of the splines-generated parameters is high, the direct computational task of global maximum likelihood would be too large. Therefore one should search for some sequential procedure, computing iteratively one subset of parameters after another. Such a procedure is again comparable with the local likelihood approach. But the optimality of such a procedure is not guaranteed.

6. NUMERICAL EXAMPLE WITH LOGISTIC MODEL

This part describes the data and the solution of an artificial example. Nevertheless, the case may represent a real situation. Let us imagine a company (say a kind of an

academic Institute), budget of which was affected by the economic problems in the country (probably the country from Central Europe). Therefore it was necessary to reduce the staff of the Institute. Simultaneously, some people are leaving the Institute voluntarily, they are searching for better paid jobs in the slowly developing private sector.

The sample has been collected during the critical period of the last two years. That is why the values of all covariates may be considered as constant in the time. The response variable characterizes the kind of leaving (or not leaving) the job during followed period. The data have the following structure:

$$\{\delta_i, X_{1i}, X_{2i}, X_{3i}, X_{4i},\ i = 1, \ldots, n = 185\}.$$

Here n is the number of employeed, the response variable $\delta = 1$ when the employee was fired (42 cases), $\delta = 2$ when the individual left his job voluntarily (20 cases, retired employees are included in this group), $\delta = 0$ for remaining employees. The covariables have the following meaning: X_2 is the length of the previous employment in the Institute, up to the moment of event (in case of $\delta > 0$) or to the moment when the data have been collected ($\delta = 0$). It is measured in years. Its values are from 0 to 14. X_3 characterizes the category of the job: 1 – scientist (40 cases), 2 – specialist (98), 3 – administration (21), 4 – technical staff (16), 5 – unqualified assisting employees (10). $X_4 = 1$ for men (107), $= 2$ for women (78). X_1 is the age of the individual, again in years, at the moment of leaving the job or of collecting the data. Its range is from 20 to 60 years.

We wish to reveal and describe the influence of covariates on the probability of individuals to remain in the Institute, to be fired or to leave, respectively. The appropriate mathematical model is the model for the response variable (δ) and for its regression on the covariables $X_1, \ldots, X_4$.

For this example, the analysis of the dependence of the response on the covariates will be accomplished in the frame of the logistic model, by the iteration procedure (8). As the fourth covariate acquires two values only, its influence can be described fully by a linear function $b_4(x_4) = \alpha + \beta \cdot x_4$. This assumption can be incorporated into the computing procedure.

The results of estimation are summarized in Table 1 and Figure 1. After 9 iterations the convergence has been achieved. Table 1 displays the parameters of optimal lines (and correlation and variance analysis) led through estimated points of functions $C(m, j, x)$. This linear analysis has been done before a final (secondary) smoothing. Thus, the least squares procedure has been weighted by the number of tied values of a covariate. It concerned especially the third covariate. The weighting with respect to the variance of results in a window has not been considered. Figure 1 then displays smoothed estimates of functions $C(m, j, x)$ for first three covariates, i.e. $j = 1 \sim$ age, $j = 2 \sim$ duration of employment in the Institute, $j = 3 \sim$ category of employee. Two distinct events were considered, for $m = \delta = \{1, 2\}$ for two distinct reasons of departure.

The example has been constructed in order to demonstrate the usefulness of additive regression models and in order to check the procedure of solution. No test of significance of regression has been applied in order to support the conclusions.

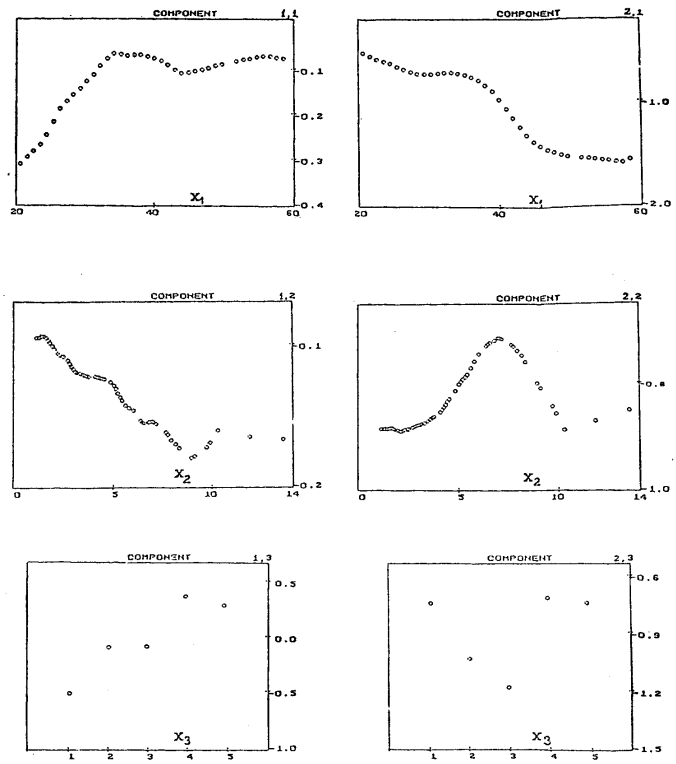

Figure 1.

Such a test might be accomplished in a traditional way, for a parametric logistic model, i.e. for a linear approximation to functions $C(m, j, x)$. The values of intercepts and slopes from Table 1 can be considered as preliminary estimates of the parameters of this parametric model.

Table 1.

m	j	intercept	slope	correl	var
1	1	−0.3434	0.0058	0.2805	0.0381
2	1	0.4507	−0.0386	−0.7858	0.0885
1	2	−0.0753	−0.0119	−0.1600	0.0383
2	2	−0.8912	0.0136	0.1286	0.0776
1	3	−0.7376	0.2369	0.8075	0.0394
2	3	−0.6800	−0.1006	−0.7246	0.0480
1	4	−0.5424	1.1668	1	0
2	4	1.3862	1.3485	1	0

(Received January 28, 1993.)

REFERENCES

[1] P. K. Andersen and R. D. Gill: Cox's regression model for counting processes: a large sample study. Ann. Statist. *10* (1982), 1100–1120.
[2] L. Breiman and J. H. Friedman: Estimating optimal transformations for multiple regression and correlation. J. Amer. Statist. Assoc. *80* (1985), 580–597.
[3] T. Hastie and R. Tibshirani: Generalized additive models (with discussion). Statist. Science *1* (1986), 297–318.
[4] C. Kooperberg, C. J. Stone and Y. K. Truong: The L_2 rate of convergence for hazard regression. Techn. Report of California Univ. No. 390 (1993).
[5] L. A. Sleeper and D. P. Harrington: Regression splines in the Cox model with application to covariate effects in liver disease. J. Amer. Statist. Assoc. *85* (1990), 941–949.
[6] C. J. Stone: The dimensionability reduction principle for generalized additive models. Ann. Statist. *14* (1986), 590–606.
[7] P. Volf: A large sample study of nonparametric proportion hazard regression model. Kybernetika *26* (1990), 404–415.
[8] P. Volf: Estimation procedures for nonparametric regression models of lifetime. In: Trans. of 11-th Prague Conference, Academia, Prague 1992.

Petr Volf, CSc., Ústav teorie informace a automatizace AV ČR (Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic), Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic.