УДК 519.95

NEW APPROACHES TO REGRESSION IN FINANCIAL MATHEMATICS AND LIFE SCIENCES BY GENERALIZED ADDITIVE MODELS

P. TAYLAN, G.-W. WEBER

This paper introduces into and improves the *theoretical* research done by the authors in the last two years in the applied area of GAMs (generalized additive models) which belong to the modern statistical learning, important in many areas of prediction, e.g., in financial mathematics and life sciences, e.g., computational biology and ecology. These models have the form $\psi(x) = \beta_0 + \sum_{j=1}^{m} f_j(x_j)$, where ψ are functions of the predictors, and they are fitted through local scoring algorithm using a scatterplot smoother as building blocks proposed by *Hastie and Tibshirani* (1987). *Aerts, Claeskens and Wand* (2002) studied penalized *spline* generalized additive models to derive some approximations. We present a mathematical modeling by splines based on a new clustering approach for the input data x, their density, and the variation of the output data y. We bounding (penalizing) second order terms (curvature) of the splines, we include a regularization of the inverse problem, contributing to a more robust approximation. In a first step, we present a refined modification and investigation of the *backfitting algorithm* previously applied to additive models. Then, by using the language of *optimization* theory, we initiate future research on solution methods with mathematical programming.

1. INTRODUCTION

1.1. Learning and Models

In the last decades, learning from data has become very important in every field of science, economy and technology, for problems concerning the public and the private life as well. Modern learning challenges can for example be found in the fields of computational biology and medicine, and in the financial sector. Learning enables for doing estimation and prediction. There are regression, mainly based on the idea of least squares or maximum likelihood estimation, and classification. In statistical learning, we are beginning with deterministic models and, then, we turn to the more general case of stochastic models where uncertainties, noise or measurement errors are taken into account. For a closer information we refer to the book *Hastie, Tibshirani, Friedman* [10]. In classical models, the approach to explain the recorded data y consists of one unknown function only; the introduction of *additive models* (*Buja*, *Hastie, Tibshirani* 1989 [4]) allowed an "ansatz" with a sum of functions which have separated input variables. In our paper, we figure out clusters of input data points x (or entire data points (x, y)), and assign an own function that additively contributes to the understanding and learning from the measured data. These functions over domains (e.g., intervals) depending on the cluster knots are mostly assumed to be splines. We will introduce an *index* useful for deciding about the spline degrees by *density* and *variation* properties of the corresponding data in x and y components, respectively. In a further step of refinement, aspects of stability and complexity of the problem are implied by keeping the curvatures of the model functions under some chosen bounds. The corresponding constrained least squares problem can, e.g., be treated as a *penalized* unconstrained minimization problem. In this paper, for the generalized (penalized) problem, we specify (*modify*) the *backfitting algorithm* which was investigated and applied for additive models. Our new investigation of *generalized additive models* is introduced in the stochastic case and closer presented in the deterministic case.

This paper contributes to both the m-dimensional case of input data separated by the model functions and, as our new alternative, to 1-dimensional input data clustered. Dimensional generalizations of the second interpretation and a combination of both interpretations are possible and indicated. Applicability for data *classification* is noted. We point out advantages and disadvantages of the concept of backfitting algorithm. By all of this, we initiate future research with a strong employing of *optimization* theory.

This paper is related with our research as initiated the papers [16, 17, 19, 20].

1.2. A Motivation of Regression

This paper has been motivated by the approximation of finanical data points (x, y), e.g., coming from the stock market. Here, x represents the input constellation, while y stands for the observed data. The discount function, denoted by $\delta(x)$, is the current price of a risk free, zero coupon bond paying unit of money at time x. We use $y(x)$ to denote the zero-coupon yield curve and to $f(x)$ to denote the instantaneous forward rate curve. These are related to the discount function by

$$\delta(x) = \exp(-xy(x)) = \exp\left(-\int_0^x f(s)ds\right). \qquad (1.1)$$

The term *interest rate curve* can be used to refer to any one of these three related curves.

In a world with complete markets and no taxes or transaction, absence of arbitrage implies that the price of any coupon bond can be computed from an interest rate curve. In particular, if the principal and interest payment of a bond is c_j units of money at time x_j ($j = 1,..., m$), the pricing equation for the bond is

$$\sum_{j=1}^{m} c_j \delta(x_j) = \sum_{j=1}^{m} c_j \exp(-x_j y(x_j)) = \sum_{j=1}^{m} c_j \exp\left(-\int_0^{x_j} f(s)ds\right). \qquad (1.2)$$

The interest rate curve can be estimated if given a set of bond prices. For this reason, let $(B_i)_{i=1,...,N}$ comprise the bonds, $X_1 < X_2 < ... < X_m$ be the set of dates at which principal and interest payments occur, let c_{ij} be the principal and interest payment of the *ith* bond on date X_j, and P_i be the observed price of the *ith* bond. The pricing equation is

$$P_i = \hat{P}_i + \varepsilon_i, \qquad (1.3)$$

where $\hat{P}_i$ is defined by $\hat{P}_i = \sum_{j=1}^{m} c_{ij}\delta\left(X_j\right)$ [18]. The curves of discount $\delta(x)$, yield $y(x)$ and forward rate $f(x)$ can be extracted via linear regression, regression with splines, smoothing splines, etc., using prices of coupon bond. For example, assuming $\mathbf{P} := (P_1,...,P_N)^T$ and $\mathbf{C} := (c_{ij})$, $i = 1,\cdots,N$, $j = 1,\ldots,m$ to be known, denoting the vector of errors or *residuals* (i.e., noise, inaccuracies and data uncertainties) by $\varepsilon := (\varepsilon_1,\ldots,\varepsilon_N)^T$ and writing $\beta := \delta(X) = (\delta(X_1),...,\delta(X_m))^T$, then the pricing equation looks as follows:

$$\mathbf{P} = \mathbf{C}\beta + \varepsilon\,. \tag{1.4}$$

Thus, the equation (1.4) can be seen as linear model with the unknown parameter vector $(\delta(X_1),...,\delta(X_m))^T = \beta$. If we use *linear regression* methods or maximum likelihood estimation and, in many important cases, just least squares estimation, then we can extract $\delta(X)$. For introductory and closer information about these methods from the viewpoints of statistical learning or the theory of inverse problems, we refer to the books of *Hastie, Tibshirani*, *Friedman* [10] and *Aster*, *Borchers*, *Thurber* [2], respectively.

While the papers [16, 17] refer to the financial sector, the works [19, 20] address the areas of computational biology, environmental protection and the interfaces between both. Actually, finance — the world of prosperity, and development are related with the gene-environment networks.

1.3. Regression

1.3.1. Linear Regression

Provided an input vector $X = (X_1,...,X_m)^T$ of (random) variables and an output variable *Y*, our linear regression model has the form

$$Y = E(Y\,|\,X_1,...,X_m) + \varepsilon = \beta_0 + \sum_{j=1}^{m} X_j\beta_j + \varepsilon\,. \tag{1.5}$$

The linear model either assumes that the regression function $E(Y|X)$ is linear or that linearity means a reasonable approximation. Here, β_j are unknown parameters or coefficients, the error ε is a Gaussian random variable with expectation zero and variance σ^2, written $\varepsilon \sim N(0,\sigma^2)$, and the variables X_j can be from different sources. Typically we have a set of training data $(x_1,y_1),...,(x_N,y_N)$ from which we estimate the parameters β_j. Here, each $x_i = (x_{i1},x_{i2},...,x_{im})^T$ is a vector of feature measurements for the *i*th case. The most popular estimation method is “least squares” which determines the cofficient vector $\beta = (\beta_0,\beta_1,...,\beta_m)^T$ to minimize the *residual sum of squares*

$$\mathrm{RSS}(\beta) := \sum_{i=1}^{N}\left(y_i - x_i^T\beta\right)^2 \quad \text{or} \quad \mathrm{RSS}(\beta) = (Y - X\beta)^{\mathrm{T}}(Y - X\beta)\,. \tag{1.6}$$

Here, X is the $N \times (m+1)$ matrix with each row being an input vector (with a 1 in the first position), and Y is the N vector of outputs in the training set. The second equation in (1.6) is a quadratic function in $m+1$ unknown parameters. If $N \geq m+1$ and X has full rank, then vector β which mimimizes RSS is $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$. The predicted values at an input vector x_0 are given by $\hat{f}(x_0)$; the fitted values at the training inputs are $\hat{y} = X\hat{\beta} = X\left(X^T X\right)^{-1} X^T y$, where $\hat{y} = \hat{f}(x_i)$.

1.3.2. Regression with Splines

In the above regression problems, sometimes $f(X) = E(Y|X_1,...,X_m)$ can be nonlinear and nonadditive. Since, however, a linear model is easy to interpret, we want to represent $f(X)$ by a linear model. Thus, an approximation by a first-order Taylor approximation to $f(X)$ can be used and sometimes even needs to be done. In fact, if N is small or m large, a linear model might be all we are able to use for data fitting without overfitting. As in classification, a linear, Bayes-optimal, decision boundary [10] implies that some monotone transformation of $\Pr(Y=1\,|X)$ is linear in X.

Regression with splines is a very popular method as for moving beyond linearity [10]. Here, we expand or replace the vector of inputs X with additional variables, which are transformations of X and, then, we use linear models in this new space of derived input features. Let X vector of inputs and $h_j : IR^m \to IR$ be the j-th transformation of X or basis function ($j = 1,2,\ldots,M$). Then, $f(X)$ is modelled by $f(X) = \sum_{j=1}^{M} \beta_j h_j(X)$, a linear basis expansion in X. Herewith, the model has become linear in these new variables and the fitting proceeds as in the case of a linear model. In fact, the estimation of β is $\hat{\beta} = = \left(H^T(x)H(x)\right)^{-1} H^T(x)Y$, where $H(x) = (h_j(x_i))_{\substack{i=1,...,m;\\ j=1,...,M}}$ is the matrix of basis functions evaluated at the input data. Hence, f becomes estimated by $\hat{f}(X) = h^T(X)\hat{\beta}$. For the special case $h_j(X) = X_j$ ($j = 1,2,\ldots,M$) the linear model is recovered. Generally, in one dimension ($m=1$), an *order M spline* with knots ξ_κ ($\kappa = 1,2,\ldots,K$) is piecewise polynomial of degree $M-1$, and has continuous derivatives up to order $M-2$. A cubic spline has $M=4$. Any piecewise constant function is an order 1 spline, while the continuous piecewise linear function is an order 2 spline. Likewise the general form for the truncated-power basis set would be $h_j(X) = X^{j-1}$ $(j = 1,2,...,M)$ and $(l = 1,2,...,K)$, where $(\bullet)_+$ stands for the positive part of a value [10].

1.4. Additive Models

1.4.1. Classical Additive Models

We stated that regression models, especially, linear ones, are very important in many applied areas. However, the traditional linear models often fail in real life, since many effects are generally *nonlinear*. Therefore, flexible statistical methods have to be used to characterize nonlinear regression effects; among these methods is *non-parametric regression* [6]. By using the common assumption of linearity, it gives information to explore the data more flexibly, uncovering structure in the data that might otherwise be missed. Many nonparametric methods do not perform well when there is a large number of independent variables in the model. The sparseness of data in this setting inflates the variance of the estimates. The problem of rapidly increasing variance for increasing dimensionality is sometimes referred to as the "curse of dimensionality". Interpretability is another problem with nonparametric regression based on kernel and smoothing spline estimates [11]. To overcome these difficulties [15] proposed *additive models*. These models estimate an additive approximation of the multivariate regression function. Here, the estimation of the individual terms explains how the dependent variable changes with the corresponding independent variables. We refer to *Hastie and Tibshirani* (1986) [8] for basic elements of the theory of additive models.

If we have data consisting of N realizations of random variable Y at m design values, then the additive model takes the from

$$E\left(Y_i \middle| x_{i1},...x_{im}\right) = \beta_0 + \sum_{j=1}^{m} f_j\left(x_{ij}\right). \tag{1.7}$$

Here, the functions f_j are estimated by a smoothing on a single coordinate, and standard convention is to assume at the knots $x_{ij} : E\left(f_j\left(x_{ij}\right)\right) = 0$ [9] (we shall give a justification below). Additive models have a strong motivation as a useful data analytic tool. Each variable is represented separately in (1.7) and the model has an important interpretation feature of some "linear model": Each of the variables separately effects the response surface and that effect does not depend on the other variables. For this reason, if once an additive model can be fit to data, we can plot the *m* coordinate functions separately to examine the roles of the variables in predicting the response. Each function is estimated by an algorithm proposed by *Friedman and Stuetzle* (1981) [7] and called ***backfitting algorithm***. As the estimator for $\hat{\beta}_0$, the arithmetic mean (average) of the output data is used: $\text{ave}(y_i \mid i = 1,...,N) := (1/N)\sum_{i=1}^{N} y_i$.

This procedure depends on the partial residual against x_{ij}:

$$r_{ij} = y_i - \beta_0 - \sum_{k \neq j} \hat{f}_k\left(x_{ik}\right) \tag{1.8}$$

and consists of estimating each smooth function by holding all the other ones fixed. In a framework of *cycling* from one to the next iteration, this means the following [9]:

initialization $\hat{\beta}_0 = ave(y_i \mid i = 1,...,N)$, $\hat{f}_j(x_{ij}) \equiv 0,\ \forall i, j$;

```
**cycle** $j = 1,...,m,1,...,m,1,...,$
    $r_{ij} = y_i - \hat{\beta}_0 - \sum_{k \neq j}^{m} \hat{f}_k(x_{ik}), \quad i = 1,...,N,$
    $\hat{f}_j$ is updated by smoothing the *partials residuals*,
    $r_{ij} = y_i - \hat{\beta}_0 - \sum_{k \neq j}^{m} \hat{f}_k(x_{ik}), \quad i = 1,...,N$, against $x_{ij}$;
**until** the functions almost do not change.
```

The backfitting procedure is also called *Gauss-Seidel* algorithm. To prove its **convergence** [4] reduced the problem to the solution of a corresponding homogeneous system, analyzed by a linear fixed point equation of the form $\hat{\mathbf{T}}\mathbf{f} = \mathbf{f}$. In fact, to represent the effect on the homogeneous equations of updating the *j*th component under Gauss-Seidel algorithm, the authors introduced the linear transformation

$$\hat{T}_j : IR^{Nm} \to IR^{Nm}, \quad f \mapsto \left(f_1^T \dots S_j^T\left(-\sum\nolimits_{k \neq j} f_k\right) \dots f_m^T\right)^T.$$

A full cycle of this algorithm is determined by $\hat{T} = \hat{T}_m \hat{T}_{m-1} \dots \hat{T}_1$; then, $\hat{T}^l$ correspond l full cycles. If all smoothing splines S_j are symmetric and have eigenvalues in $[0,1]$, then the backfitting algorithm always *converges*. In Subsection 2.6, we will come back closer to the algorithm and the denotation used here.

1.4.2. Additive Models Revisited

In our paper, we allow a different and new motivation: In addition to the approach given by a *separation* of the variables x_j done by the functions f_j, now we perform a *clustering* of the input data of the variable *x* by a partitioning of the domain into cubes Q_j or, in the 1-dimensional case: intervals I_j, and a determination of f_j with reference to the knots lying in Q_j (or I_j), respectively. In any such a case, a cube or interval is taking the place of a dimension or coordinate axis. We will mostly refer to the case of one dimension; the higher dimensional case can then be treated by a combination of separation and clustering. That clustering can incorporate any kind of periods of seasons assumed, any comparability or correspondence of successive time intervals, etc. Herewith, the functions f_j are more considered as allocated to sets I_j (or Q_j) rather than depending on some special, sometimes arbitrary elements of those sets (input data) or output values associated. This new interpretation and usuage of additive models (or generalized ones, introduced next) is a key step of this paper.

2. GENERALIZED ADDITIVE MODELS

To extend the additive model to a wide range of distribution families, *Hastie and Tibshirani* (1990) [11] proposed *generalized additive models* (*GAM*) which are among the most practically used modern statistical techniques. These models enable the mean of the dependent variable to be an additive predictor through a link function. Many widely used statistical models belong to this general class; they include additive models for Gaussian data, nonparametric logistic models for binary data, and nonparametric log-linear models for Poisson data.

2.1. Definition of a Generalized Additive Model

If we have $X_1,...,X_m$, being m covariates comprised by the m-tuple $X=(X_1,...,X_m)^T$, then, in our regression setting, a *generalized additive model* has the form

$$G(\mu(X))=\psi(X)=\beta_0+\sum_{j=1}^{m} f_j(X_j). \tag{2.1}$$

Here, the function f_j are unspecified ("nonparametric") and $\theta=(\beta_0, f_1,\ldots$ $\ldots,f_m)^T$ is the unknown parameter to be estimated; G is the link function. The incorporation β_0 as some average outcome allows us to assume $E(f_j(x_{ij}))=0$ ($j=1,\ldots,m$). Often, the unknown functions f_j are elements of a finite dimensional space consisting, e.g., of splines and these functions depending on the cluster knots are mostly assumed to be splines; the spline orders (or degrees) are suitably choosen depending on the density and variation properties of the corresponding data in *x* and *y* components, respectively. Then, our problem of specifying θ becomes a finite-dimensional parameter estimation problem.

2.2. Clustering of Input Data

2.2.1. Introduction

Clustering is the process of organizing objects into $I_1,...,I_m$ groups or, higher dimensionally: $Q_1,...,Q_m$, whose elements are similar in some way. A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. For example, we can easily identify some cluster out of a finite number of clusters into which the data can be divided with respect to the similarity criterion *distance*. We put two or more objects belonging to the same cluster if they are "close" according to a given distance (in this case, geometrical distance).

Differently from usual clustering, in this paper, we understand clustering always as being accompanied by a ***partitioning*** of the (input) space, including space coverage. In other words, it will mean a classification in the absense of different labels or categories. Especially, the clusters shall not be overlapping, and the partitions containing the clusters shall also be pairwise disjoint, except at the boundaries. Instead of a general introduction into cluster and classification methods, we give the following information only.

2.2.2. Clustering for Generalized Additive Models

Financial markets have different kinds of trading activities. These activities work with considerably long horizons, ranging from days and weeks to months and years. For this reason, we may have any kind of data. These data can sometimes be problematic for being used at the models, for example, given a longer horizon with sometimes less frequent data recorded, but to other times highly frequent measurements. In those cases, by the differences in data density and, possibly, data variation, the underlying reality and the following model will be too unstable or inhomogeneous. The process may be depending on unpredictable market behaviour or external events like naturally calamity. Sometimes, the structure of data is has particular properties. These may be a larger variability or a handful of outliers. Sometimes we do not have any meaningful data. For instance, share price changes will not be available when stock markets are closed at weekends or holidays.

The following three parts of fig. 1 are showing some important cases of input data distribution and clustering: the *equidistant case* (fig. 1,*a*)) where all points can be put into one cluster (or interval) I_1, the *equidistant case with regular breaks* (weekends, holidays, etc.; (fig. 1,*b*) where the regularly neighbouring points and the free days could be put in separate cluster intervals I_j, and the *general case* (cf. (fig. 1,*c*)) where there are many interval I_j of different interval lengths and densities. We remark that we could also include properties of the output data *y* into this clustering; for the ease of exposition, however, we disregard this aspect.

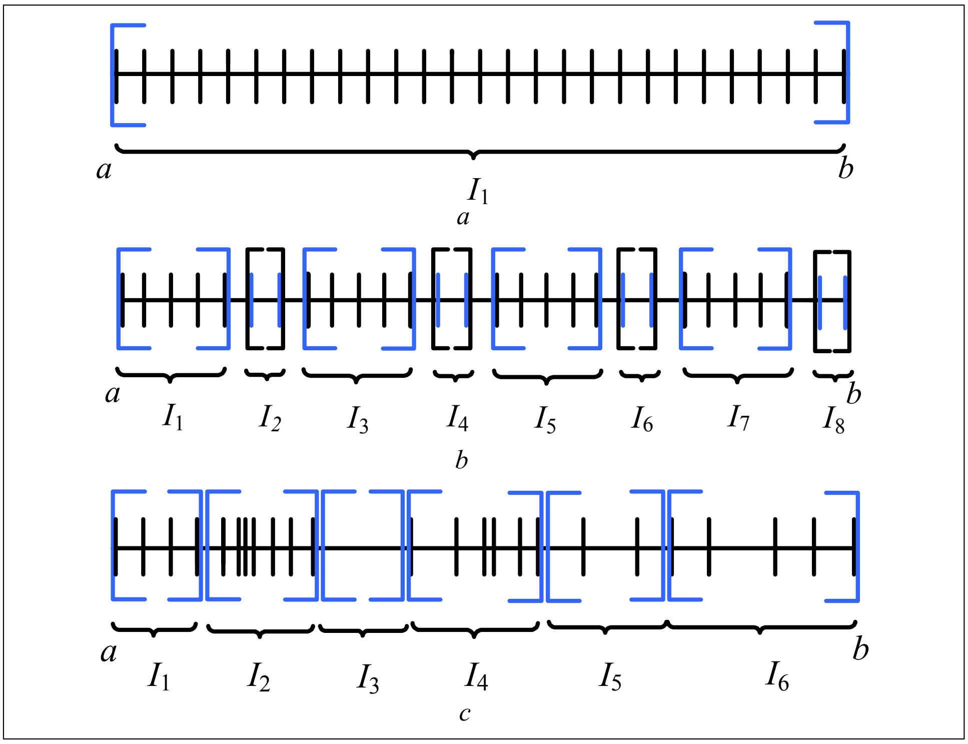

Fig. 1. Three important cases of input data distribution and its clustering: *a* — equidistance, *b* — equidistance with breaks, and *c* — general case

In the following, we will take into account the data variation; to get and impression of this, please have a look at fig. 2.

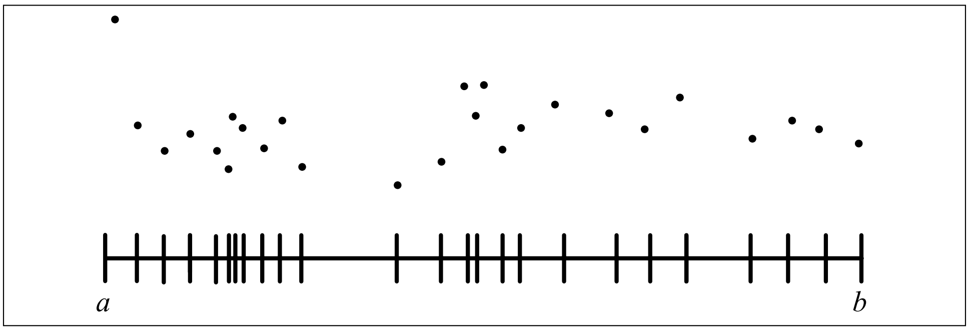

Fig. 2. Example of a data (scatterplot); here, we refer to case (fig. 1,*c*)

For the sake of simplicity, we assume from now on that the number N_j of input data points x_{ij} in each cluster I_j is the same, say, $N_j \equiv N \; (j = 1,...,m)$. Otherwise there will be no approximation need at data points missing and the residuals of our approximation were 0 there. Furthermore, given the output data y_{ij} we denote the aggregated value over the all the *i*th output values of the clusters by

$$y_i := \sum_{j=1}^{m} y_{ij} \quad (i = 1,..., N).$$

In the example of case (fig. 1,*b*), this data summation refers to all the days *i* from monday to friday. Herewith, the cluster can also have a chronolocial meaning. By definition, up to the division by m, the values y_i are averages of the output values y_{ij}.

Before we come to a closer understanding of data density and variation, we proceed with our introduction of splines. In fact, the selection of the splines orders, degrees and classes will essentially be influenced by indices based on densities and variations (Subsection 2.5).

2.3. Splines

Let $x_{1j}, x_{2j},..., x_{Nj}$ be N distinct knots of $[a,b]$, where $a \le x_{1j} < x_{2j} < \ldots$ $\ldots < x_{Nj} \le b$. The function $f_k(x)$ on the interval $[a,b]$ (or in $\mathbf{R}$) is a spline of some degree *k* relative to the knots x_{ij} if

$$f_k \big|_{[x_{ij}, x_{i+1j}]} \in IP_k \text{ (polynomial of degree } \le k;\ i = 1,..., N-1), \qquad (2.2)$$

$$f_k \in C^{k-1}[a,b]. \qquad (2.3)$$

To characterize a spline of degree *k*, $f_{k,i} := f_k \big|_{[x_{ij}, x_{i+1j}]}$ can be represented by

$$f_{k,i}(x) = \sum_{l=0}^{k} g_{li} (x - x_{ij})^l \quad \left(x \in \left[x_{ij}, x_{i+1j}\right]\right).$$

For a closer information about spline we refer to [5, 12].

2.4. Variation and Density

Density is a measure of mass **per unit of volume.** The higher an object's density, the higher its mass per volume. Let us assume that we have $I_1,...,I_m$ intervals; then, the density of the input data x_{ij} in the *j*-th interval I_j is defined by $D_j :=$ $(\text{number of point } x_{ij} \text{ in } I_j)\big/\text{length of } I_j$. This definition can be directly generalized to the higher dimensional case of cubes Q_j rather than intervals I_j, by referring to the cubes' volumes. *Variation* is a quantifiable difference between individual measurements. Every repeatable process exhibits variation. If over the interval I_j we have the data $(x_{1j}, y_{1j}),\ldots,(x_{Nj}, y_{Nj})$, then the variation of these data refers to the output dimension *y* and it is defined as $V_j := \sum_{i=1}^{N-1} \left| y_{i+1j} - y_{ij} \right|$. If this value is big, at many data points the rate of change of the angle between any approximating curve and its tangent would be big, i.e., its curvature could be big. Otherwise, the curvature could be expected to be small. In this sense, high curvature over an interval can mean a highly oscillating behaviour. The occurrence of outliers y_{ij} may contribute to this very much and mean instability of the model.

2.5. Index of Data Variation

Still we assume that $I_1,...,I_p$ (or $Q_1,...,Q_m$) are the intervals (or cubes) according to the data grouped. For each interval I_j (cube Q_j), we define the associated index of data variation by $Ind_j := DV$ or, more generally, $Ind_j :=$ $:= d_j(D_j)v_j(V_j)$, where d_j, v_j are some positive, strongly monotonically increasing functions selected by the modeller. In fact, from both the viewpoints of data fitting and complexity (or stability), cases with a high variation distributed over a very long intervall are very much less problematic than cases with a high variation over a short intervall. The multiplication of variation terms with density terms due to each interval found by clustering is representing this difference.

We determine the degree of the splines f_j with the help of the numbers Ind_j. If such an index is low, then we can choose the spline degree (or order) to be small. In this case, the spline may have a few coefficients to be determined and we can find these coefficients easily using any appropriate solution method for the corresponding spline equations. If the number Ind_j is big, then we must choose a high degree of the spline. In this case, the spline may have a more complex structure and many coefficients have to be determined; i.e., we may have many system equations or a high dimensional vector of unknows; to solve this could become very difficult. Also, a high degree of splines $f_1, f_2,..., f_m$, respectively, causes high curvatures or oscillations, i.e., there is a high "energy" implied; this means a higher (co)variance or instability under data perturbations. As the extremal case of high curvature we consider *non*smoothness meaning an instantaneous movement at a point which does not obey to any tangent.

The previous words introduced a model-free element into our explanations. Indeed, as indicated in Subsection 2.4, the concrete determining of the spline degree can be done adaptively by the implementer who writes the code. From a close mathematical perspective we propose to introduce discrete *thresholds* γ_ν and to assign to all the intervals of indices $Ind \in [\gamma_\nu, \gamma_{\nu+1})$ the same specific spline degrees. This determination and allocation has to base on the above reflections and data (or residuals) given.

For the above reasons, we want to impose some control on the oscillation. To make the oscillation smaller, the curvature of each spline must be bounded by the penalty parameter. We introduce a *penalty parameter* into the criterion of minimizing RSS, called *penalized sum or squares* PRSS now:

$$\mathrm{PRSS}(\beta_0, f_1, ..., f_m) := \sum_{i=1}^{N} \left\{ y_i - \beta_0 - \sum_{j=1}^{m} f_j(x_{ij}) \right\}^2 + \sum_{j=1}^{m} \mu_j \int_a^b \left[f_j''(t_j) \right]^2 dt_j \ . \ (2.4)$$

The first term measures the goodness of data fitting, while the second term is a penalty term and defined by means of the functions' curvatures. Here, the interval $[a,b]$ is the union of all the intervals I_j. In the case of separation of variables, the interval bounds may also depend on j, i.e., they are $[a_j, b_j]$. For the sake of simplicity, we sometimes just write $\int$ and refer to the interval limits given by the context. There are also further refined curvature measures, especially, one with the input knot distribution implied by Gaussian bell-shaped density functions; these appear as additional factors in the integrals and have a cutting-off effect. For the sake of simplicity, we shall focus on the given standard one now and turn to the sophisticated model in a later study.

In (2.4), $\mu_j \geq 0$ are tuning or *smoothing* parameters and they represent a tradeoff between first and second term. Large values of μ_j yield smoother curves, smaller values result in more fluctuation. It can be shown that the minimizer of PRSS is an additive spline model: Each of the functions f_j is a spline in the component X_j, with knots at x_{ij} $(i = 1, \ldots, N)$. However, without further restrictions on the model, the solution is not unique. The constant β_0 is not identifiable since we can add or substract any constants to each of the functions f_j, and adjust β_0 accordingly. For example, one standard convention is to assume that $\sum_{j=1}^{m} f_j(x_{ij}) = 0 \ \ \forall i$, the function average being zero over the corresponding data (e.g., of mondays, tuesdays, ... , fridays, respectively). This can be achieved by means of a *bias* (*intercept*) β_0 in front of the sum $\sum_{j=1}^{m} f_j(x_{ij})$ in the additive approach. In this case, $\hat{\beta}_0 = \mathrm{ave}\,(y_i \,|\, i = 1, 2, \ldots, N)$, as can be seen easily.

We firstly want to have $\sum_{i=1}^{N}\left\{y_i-\beta_0-\sum_{j=1}^{m}f_j(x_{ij})\right\}^2\approx 0$ and, secondly, $\sum_{j=1}^{m}\int\left[f_j''(t_j)\right]^2 dt_j\approx 0$ or being sufficiently small, at least bounded. In the backfitting algorithm, these approximations, considered as equations, will give rise to expected or *update* formulas. For these requests, let us introduce

$$F(\beta_0,f):=\sum_{i=1}^{N}\left\{y_i-\beta_0-\sum_{j=1}^{m}f_j(x_{ij})\right\}^2 \text{ and } g_j(f):=\int\left[f_j''(t_j)\right]^2 dt_j-M_j,$$

where $f=(f_1,f_2,...,f_m)^T$. The terms $g_j(f)$ can be interpreted as curvature integral values minus some prescribed upper bounds $M_j>0$. Now, the combined standard form of our regression problem subject to the constrained curvature condition takes the following form:

$$\text{Minimize}\quad F(\beta_0,f)\ \text{ subject to }\ g_j(f)\le 0\quad (j=1,...,m). \tag{2.5}$$

Now, *PRSS* can be interpreted in *Lagrangian* form as follows:

$$L\left((\beta_0,f),\mu\right)=\sum_{i=1}^{N}\left\{y_i-\beta_0-\sum_{j=1}^{m}f_j(x_{ij})\right\}^2+\sum_{j=1}^{m}\mu_j\left(\int\left[f_j''(t_j)\right]^2 dt_j-M_j\right), \tag{2.6}$$

where $\mu:=(\mu_1,...,\mu_m)^T$. Here, $\mu_j\ge 0$ are auxilary *penalty parameters* introduced in [3]. In the light of our optimization problem, they can now be seen as *Lagrange multipliers* associated with the constraints $g_j\le 0$. The *Lagrangian dual problem* takes the form

$$\max_{\mu\ge 0}\ \min_{(\beta_0,f)} L\left((\beta,f),\mu\right). \tag{2.7}$$

The solution of this optimization problem (2.7) will help us for determining the smoothing parameters μ_j and, in particular, the functions f_j will be found, likewise their bounded curvatures $\int\left[f_j''(t_j)\right]^2 dt_j$. Herewith, a lot of future research is initialized which can become an alternative to the backfitting algorithm concept. In this paper, we go on with refining and discussing the backfitting concept for the generalized additive model.

2.6. Modified Backfitting Algorithm for Generalized Additive Model

2.6.1. Generalized Additive Model Revisited

For the generalized additive model (cf. Section 2.1), we will modify the backfitting algorithm used before for fitting additive model (cf. Subsection 1.3). For this reason, we will use the following theoretical setting in term of conditional expectation (*Buja, Hastie and Tibshirani* (1989) [4]), where $j=1,\dots,m$:

$$f_j(X_j) = P_j\left(Y - \beta_0 - \sum_{k \neq j} f_k(X_k)\right) := E\left(Y - \beta_0 - \sum_{k \neq j} f_k(X_k) \middle| X_j\right). \quad (2.8)$$

Here, $P_j(\cdot)$ denote the conditional expectation value $E\left(\cdot \middle| X_j\right)$. Now, to find $f_j(X_j)$ in our generalized addive model, let us add the term $-\sum_{k=1}^{m} \mu_k \int \left[f_k''(t_k)\right]^2 dt_k$ to equation (2.8). In this case, (2.8) will become the update formula

$$f_j(X_j) + \mu_j \int \left[f_j''(t_j)\right]^2 dt_j \leftarrow$$

$$\leftarrow P_j\left(Y - \beta_0 - \sum_{k \neq j} f_k(X_k)\right) - \left(\sum_{k \neq j}^{m} \mu_k \int \left[f_k''(t_k)\right]^2 dt_k\right) =$$

$$= E\left(Y - \beta_0 - \sum_{k \neq j} f_k(X_k) \,|X_j\right) - \left(\sum_{k \neq j}^{m} \mu_k \int \left[f_k''(t_k)\right]^2 dt_k\right), \quad (2.9)$$

where $\sum_{k \neq j}^{m} \mu_k \int \left[f_k''(t_k)\right]^2 dt_k = c_j$ (constant, i.e., not depending on the knots); the functions $\hat{f}_j$ are unknown and to be determined in the considered iteration. Therefore, we can write equation (2.9) as

$$f_j(X_j) + \mu_j \int \left[f_j''(t_j)\right]^2 dt_j \leftarrow$$

$$\leftarrow E\left(Y - \beta_0 - \sum_{k \neq j}\left(f_k(X_k) + \mu_k \int \left[f_k''(t_k)\right]^2 dt_k\right) \middle| X_j\right).$$

If we denote $Z_k(X_k) = f_k(X_k) + \mu_k \int \left[f_k''(t_k)\right]^2 dt_k$ (the same for j), then we get the update formula

$$Z_j(X_j) \leftarrow E\left(Y - \beta_0 - \sum_{k \neq j} Z_k(X_k) \,|\, X_j\right). \quad (2.10)$$

For random variables (Y, X), the conditional expectation $f(x) = E(Y|X = x)$ minimizes $E(Y - f(X))^2$ over all L_2 functions f [4]. If this idea is applied to our generalized additive model, then the minimizer of $E(Y - \psi(\boldsymbol{X}))^2$ will give the closest additive approximation to $E(Y\,|\,\boldsymbol{X})$. Equivalently, the following system of *normal equations* is necessary and sufficient for $\mathbf{Z} = (Z_1, Z_2, ..., Z_m)^T$ to minimize $E(Y - \psi(\boldsymbol{X}))^2$ (for the formula without intercept β_0, we refer to [4]):

$$\begin{pmatrix} I & P_1 & . & . & P_1 \\ P_2 & I & . & . & P_2 \\ . & . & . & . & . \\ . & . & . & . & . \\ P_m & P_m & . & . & I \end{pmatrix} \begin{pmatrix} Z_1(X_1) \\ Z_2(X_2) \\ . \\ . \\ Z_m(X_m) \end{pmatrix} = \begin{pmatrix} P_1(Y-\beta_0\mathbf{e}) \\ P_2(Y-\beta_0\mathbf{e}) \\ . \\ . \\ P_m(Y-\beta_0\mathbf{e}) \end{pmatrix}, \qquad (2.11)$$

where $\mathbf{e}$ is the N-vector or entries 1; or, in short, $\mathbf{PZ} = \mathbf{Q}(Y - \beta_0)$. Here, $\mathbf{P}$ and $\mathbf{Q}$ represent the matrix and vector of operators, respectively. If we want to apply normal equation to any given discrete experimental data, we must change the variables (Y, X) in the (2.11) by their realizations $(y_i, \mathbf{x}_i)$, $\mathbf{x}_i = = (x_{i1}, x_{i2}, ..., x_{im})^T$, and the conditional expectations $P_j = E(\cdot | X_j)$ by smoothers S_j on x_j,

$$\begin{pmatrix} I & S_1 & . & . & S_1 \\ S_2 & I & . & . & S_2 \\ . & . & . & . & . \\ . & . & . & . & . \\ S_m & S_m & . & . & I \end{pmatrix} \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ . \\ . \\ \mathbf{z}_m \end{pmatrix} = \begin{pmatrix} S_1(\mathbf{y}-\hat\beta_0\mathbf{e}) \\ S_2(\mathbf{y}-\hat\beta_0\mathbf{e}) \\ . \\ . \\ S_m(\mathbf{y}-\hat\beta_0\mathbf{e}) \end{pmatrix}. \qquad (2.12)$$

In estimation notation (2.12) equation can be written $\hat{\mathbf{P}}\mathbf{z} = \hat{\mathbf{Q}}(\mathbf{y} - \hat\beta_0) =: \hat{\mathbf{Q}}y_1$. Here, $S_j = (h_{jl}(x_i))_{\substack{i=1,...,N \\ j=1,...,N}}$ are smoothing matrices of type $N \times N$, $\mathbf{z}_j$ are N-vectors representing the spline function $\hat f_j + \mu_j \int \left[\hat f_j''(t_j)\right]^2 dt_j$ in a canonical form, i.e., $\sum_{l=1}^{N} \theta_{jl} h_{jl}(X)$ (with the number of unknown equal to the number of conditions). In this notation, without loss of generality we already changed from lower spline degrees d_j to a maximal one d, and to the order N. Furthermore, (2.12) is an ($Nm \times Nm$)-system of *normal equations*. The solutions to (2.12) satisfy $\mathbf{z_j} \in \Re(S_j)$, where $\Re(S_j)$ is the range of the linear mapping S_j, since we update by $\mathbf{z_j} \leftarrow S_j\left(\mathbf{y} - \hat\beta_0\mathbf{e} - \sum_{k \neq j} \mathbf{z_k}\right)$ In case we want to emphasize $\hat\beta_0$ among the unknowns, i.e., $\left(\hat\beta_0^T, \mathbf{z}_1^T, ..., \mathbf{z}_m^T\right)^T$, again equation (2.12) can equivalently be written for this situation.

There is a variety of efficient methods for solving the system (2.12), which depend on both the number and typs of smoother used. If the smoother matrix S_j is a $N \times N$ nonsingular matrix, then the matrix $\hat{\mathbf{P}}$ will be a nonsingular ($Nm \times Nm$)-matrix; in this case, the system $\hat{\mathbf{P}}\mathbf{z} = \hat{\mathbf{Q}}y_1$ has a unique solution. If the smoother matrices S_j are not guaranteed to be invertible (nonsingular) symmetric, but just arbitrary ($N \times N$)-matrices, we can use a generalized inverses S_j^-

(i.e., $S_j S_j^- S_j = S_j$) and $\hat{\mathbf{P}}^-$. For closer information about generalized solution and matrix calculus we refer to [12].

2.6.2. Modified Backfitting Algorithm

Gauss-Seidel method, applied to blocks consisting of vectorial component $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m$, exploits the special structure of (2.12). It coincides with the backfitting algorithm. If in the algorithm we write $\hat{z}_j = \hat{f}_j + \mu_j \int \left[\hat{f}_j''(t_j)\right]^2 dt_j$ (in fact, the functions $\hat{f}_j$ are unknowns), then, the *l*th iteration in the backfitting or Gauss-Seidel includes the additional penalized curvature term. Not forgetting the stepwise update of the penalty parameter μ_j but not mentioning it explicitely, then the framework of the procedure looks as follows:

1. **initialize** $\hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\hat{f}_j \equiv 0 \;\Rightarrow\; \hat{z}_j \equiv 0$, $\forall j$;

2. **cycle** $j = 1, 2, \ldots m$,

$$\hat{z}_j \leftarrow S_j \left[\left\{ y_i - \hat{\beta}_0 - \sum_{k \neq j} \hat{z}_k(x_{ik}) \right\}_{i=1}^{N} \right].$$

This iteration is done until the individual functions do not change: Here, in each iterate, $\hat{z}_j$ is by the spline function related with the knots x_{ij} and found by the values $y_i - \hat{\beta}_0 - \sum_{k \neq j} \hat{z}_k(x_{ik})$ $(i = 1, 2, ..., N)$. In the other words, by the other functions $\hat{z}_k$ and, finally, by the functions $\hat{f}_k$ and the penalty (smoothing) parameter μ_k. Actually, since by definition it holds $\hat{z}_j = \hat{f}_j + \mu_j \int \left[\hat{f}_j''(t_j)\right]^2 dt_j$, throughout the algorithm we must have a *book keeping* about both $\hat{f}_j$ and the curvature effect $\mu_j \int \left[\hat{f}_j''(t_\mathrm{j})\right]^2 dt_j$ controlled by the penalty parameter μ_j which we can update from step to step. This book keeping is guaranteed since $\hat{f}_\mathrm{j}$ and the curvature $\int \left[\hat{f}_j''(t_j)\right]^2 dt_j$ can be determined via $\hat{z}_j$. Since the value of $\mu_j \int \left[\hat{f}_j''(t_j)\right]^2 dt_j$ is constant, the second order derivative of $\hat{z}_j$ is $\hat{z}_j''(t_j) = \hat{f}_j''(t_j)$; this yields $\mu_j \int \left[\hat{f}_j''(t_j)\right]^2 dt_j := \mu_j \int \left[\hat{z}_j''(t_j)\right]^2 dt_j$ and, herewith, $\hat{f}_j := \hat{z}_j - \mu_j \int \left[\hat{f}_j''(t_j)\right]^2 dt_j$.

2.6.3. Discussion about Modified Backfitting Algorithm

If we consider our optimization problem on (2.4) (see also (2.7)) as fixed with respect to μ_j, then we can carry over the *convergence theory* about additive models (see Section 1.3) to the present generalized additive model, replacing the functions $\hat{f}_{\mathrm{j}}$ by $\hat{z}_j$. However, at least approximatively, we have to guarantee feasibility also, i.e.,

$$\int \left[\hat{f}_j''(t_j)\right]^2 dt_j \le M_j \quad (j = 1,...,m).$$

If $\int \left[\hat{f}_j''(t_j)\right]^2 dt_j \le M_j$, then we preserve the value of μ_j for $l \leftarrow l+1$; otherwise, we increase μ_j. But this update changes the values of $\hat{z}_j$ and, herewith, the convergence behaviour of the algorithm. What is more, the modified backfitting algorithm bases on both terms in the objective function to be approximated by 0; too large an increase of μ_j can shift too far away from 0 the corresponding penalized curvature value in the second term. The iteration stops if the functions f_j become stationary, i.e., not changing very much and, if we request it, if $\sum_{i=1}^{N} \left\{ y_i - \beta_0 - \sum_{j=1}^{m} f_j(x_{ij}) \right\}^2$ becomes sufficiently small, i.e., lying under some error threshold ε, *and*, in particular, $\int \left[\hat{f}_j''(t_j)\right]^2 dt_j \le M_j$ ($j = 1,2,\ldots,m$).

2.7. On a Numerical Example

Numerical applications arise in many areas of science, technology, social life and economy with, in general, very huge and firstly unstructured data sets; in particular, they may base on data from *financial mathematics*. These data can be got, e.g., from *Bank of Canada* (http://www.bankofcanada.ca/en/rates/interest-look.html) as daily, weekly and monthly; they can be regularly partioned, which leads to a *partitioning* (clustering) of the (input) space, and indices of data variation can be assigned accordingly. Then, we decide about the degrees of the spline depending of the location of the indices between thresholds γ_ν. In this entire process, the practitioner has to study the structure of the data. In particular, the choice on the cluster approach at all, or of the approach on separation of variables, or of a combination of both, has to be made at an early stage and in close collaboration between the financial analyst, the optimizer and the computer engineer. At Institute of Applied Mathematics of METU, we are in exchange with the experts of its Department of Financial Mathematics, and this application is initiated. Using the splines which we determine by the modified backfitting algorithm, an approximation for the unknown functions of the additive model can be iteratively found. There is one adaptive element remaining in this iterative process: the update the penalty parameter, in connection with the observation of the convergence behaviour. Here, we propose the use and implementation of our algorithm and, to overcome its structural frontiers given by the choice of the penalty parameter in the course of the program, a use of *conic quadratic programming* with *interior point algorithm* applied. A comparison and possible combination of these two algorithmic strategies is what we recommend in this pioneering paper.

3. CONCLUDING

This basic and more theoretical paper has given a contribution to the discrete approximation or regression of data in 1- and multivariate cases. Generalized additive models have been investigated, input data grouped by clustering, its density measured, data variation quantified, spline classes selected by indices and their curvatures bounded with the help of penalization. Then, the backfitting algorithm which is also applicable for data classification has become modified and the further utilization of modern optimization recommended [14]. By this we have contributed to a better understanding of data from the financial world and life sciences, to a more refined instrument of prediction. In the paper [23], we extended our approach by spline from discrete or Gaussian approximation to the continuous type of *Chebychev* approximation, by this representing the occurence of errors and uncertainy in modern technology, decision making and negotiations. In the work, we made a connection to CO_2 *emission control*, visualizing dynamics and simulations also. There is a lot of work waiting in future research and application, and we cordially invite to this.

Acknowledgement: The authors express their gratitude to Prof. Dr. Alexander Makarenko and Prof. Dr. Peter I. Bidyuk for encouragement and help, and to the referee for valuable notes and proposals. This study carried out my postdoct duration at the Middle East Technical University.

REFERENCES

1. *Aerts M., Claeskens G. and Wand M.P.* Some theory for penalized spline generalized additive models, J. Statist. Planning and Inference 103 (2002).
2. *Aster A., Borchers B. and Thurber C.* Parameter Estimation and Inverse Problems, Academic Press, 2004.
3. *Boyd S. and Vandenberghe L.* Convex Optimization. — Cambridge University Press, 2004.
4. *Buja A., Hastie T. and Tibshirani R.* Linear smoothers and additive models —The Ann. Stat. 17, 2 (1989).
5. *De Boor C.* Practical Guide to Splines. — Springer Verlag, 2001.
6. *Fox J.* Nonparametric regression, Appendix to an R and S-Plus Companion to Applied Regression. — Sage Publications, 2002.
7. *Friedman J.H. and Stuetzle W.* Projection pursuit regression. — J. Amer. Statist Assoc. — 76 (1981).
8. *Hastie T. and Tibshirani R.* Generalized additive models. — Statist. Science. — 1, 3 (1986).
9. *Hastie T. and Tibshirani R.* Generalized additive models: some applications. — J. Amer. Statist. Assoc. — 82, 398 (1987).
10. *Hastie T., Tibshirani R. and Friedman J.H.* The Element of Statistical Learning. — Springer Verlag, New York, 2001.
11. *Hastie T.J. and Tibshirani R.J.* Generalized Additive Models, New York, Chapman and Hall, 1990.
12. *Pringle R.M. and Rayner A.A.* Generalized Inverse Matrices With Applications to Statistics. — Hafner Publishing, 1971.

13. *Quarteroni A., Sacco R. and Saleri F.* Numerical Mathematics. — Texts in Applied Mathematics 37. — Springer, 1991.

14. *Spellucci P.* Personal communication, Darmstadt University of Technology. — Germany (2006).

15. *Stone C.J.* Additive regression and other nonparametric models. — The Annals of Statistics. — 13, 2 (1985).

16. *Taylan P. and G.-W. Weber.* New approaches to regression in financial mathematics by additive models. — To appear in Journal of Computational Technologies (2007).

17. *Taylan P., Weber G.-W. and Beck A.* New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology. To appear in the special issue of the journal Optimization at the occasion of the 5-th Ballarat Workshop on Global and Non-Smooth Optimization: Theory, Methods and Applications, November 28–30, 2006.

18. *Waggoner D.F.* Spline methods for extractions interest rate curves coupon bound prices. — Federal Reserve Bank of Atlanta Working Paper. — 97–10 (1997).

19. *Weber G.-W., Tezel A., Taylan P., Soyler A. and Cetin M.* On dynamics and optimization of gene-environment networks. To appear in the special issue of Optimization in honour of the 60th birthday of Prof. Dr. H.Th. Jongen.

20. *Weber G.-W., Uğur Ö., Taylan P. and Tezel A.* On optimization, dynamics and uncertainty: a tutorial for gene-environment networks. To appear in the special issue of Discrete Applied Mathematics "Networks in Computational Biology".

21. *Weber G.-W., Alparslan-Gök S.Z. and Söyler B.* A new mathematical approach in environmental and life sciences: gene-environment networks and their dynamics, invited paper submitted to Environmental Modeling & Assessment.

Received 26.03.2007

From the Editorial Board: the article corresponds completely to supmitted manuscript.