NEW APPROACHES TO REGRESSION IN FINANCIAL MATHEMATICS BY ADDITIVE MODELS

P. Taylan
Dicle University, Department of Mathematics, Diyarbakır, Turkey
Middle East Technical University, Institute of Applied Mathematics, Ankara, Turkey,
e-mail: ptaylan@dicle.edu.tr

G.-W. Weber
Middle East Technical University, Institute of Applied Mathematics, Ankara, Turkey,
e-mail: gweber@metu.edu.tr

Аддитивные модели принадлежат к технике современного статистического познания, они применяются для прогнозирония во многих областях, таких как финансовая математика, вычислительная биология, медицина, химия и защита окружающей среды. Эти модели используются посредством алгоритма обратного фиттинга, основанного на методе частичных остатков, предложенном Friedman and Stuetzle (1981). В этой статье мы сначала даем введение в проблему и обзор. Затем мы представляем моделирование сплайнами, основанное на новом кластерном подходе для входных данных, плотности этих данных и изменении выходных данных. Наш вклад в метод регрессии с аддитивными моделями состоит в ограничении членов, отвечающих за кривизну сплайна, что приводит к более надежной аппроксимации. Мы предлагаем усовершенствованную модификацию и исследование алгоритма обратного фиттинга применительно к аддитивным моделям. Используя язык теории оптимизации, в частности, метод конического квадратичного программирования, мы инициируем дальнейшие исследования и их практические приложения для программирования.

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* [12]. 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 [5]) 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, we specify (*modify*) the *backfitting algorithm* which contains curvature term and apply for additive models.

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.

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

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

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 $(j = 1, ..., m)$, then the pricing equation for the bond is

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

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

$$P_i = \hat{P}_i + \varepsilon_i, \tag{1.3}$$

where $\hat{P}_i$ is defined by $\hat{P}_i = \sum_{j=1}^{m} c_{ij}\delta(X_j)$ (*Waggoner* 1997 [22]). 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 $\mathbf{C} := (c_{ij})_{i=1,...,N;j=1,...,m}$ to be known, denoting the vector of errors or *residuals* (i. e., noise, inaccuracies and data uncertainties) by $\varepsilon := (\varepsilon_1, ..., \varepsilon_N)^T$ and writing $\boldsymbol{\beta} := \boldsymbol{\delta}(X) = (\delta(X_1), ..., \delta(X_m))^T$, then the pricing equation looks as follows:

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

Thus, the equation (1.4) can be seen as linear model with the unknown parameter vector $(\delta(X_1), ..., \delta(X_m))^T = \boldsymbol{\beta}$. If we use *linear regression* methods or maximum likelihood estimation and, in many important cases, just least squares estimation, then we can extract $\boldsymbol{\delta}(X)$. 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* [12] and *Aster, Borchers, Thurber* [2], respectively.

1.3. Regression

1.3.1. Linear Regression

Linear regression models the relation between two variables by fitting a linear model to observed data. One variable is considered to be an input (explanatory) variable x, and the other is considered to be an output (dependent) variable y. Before attempting to do this fitting, the modeller firstly decides whether at all there is a relationship between x and y. This does not necessarily imply that one variable *causes* the other, but that there is some significant association between the two variables. A *scatterplot* can be a helpful tool in determining the strength of the relationship between two variables [7].

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

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

Here, β_j are unknown parameters or coefficients, the error ε is a Gaussian random variable with expectation 0 and variance σ^2, in short: $\varepsilon \sim N(0, \sigma^2)$, and the variables X_j can be from different sources. We denote the estimation of f by $\hat{f}$. The most popular estimation method is *least squares approximation* which determines the coefficient vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)^T$ to minimize the *residual sum of squares* via observation values (y_i, x_i),

$$RSS(\beta) := \sum_{i=1}^{N} \left(y_i - x_i^T\beta\right)^2 \tag{1.6}$$

or, in matrix notation,

$$RSS\left(\beta\right)=\left(y-X\beta\right)^{T}\left(y-X\beta\right). \tag{1.7}$$

Here, X is the $N\times(m+1)$ matrix with each row being an input vector (with entries 1 in the first column), and y is the N vector of outputs in the training set. Equation (1.7) is a quadratic function in $m+1$ unknown parameters. If $N\geq m+1$ and X has full rank, then our solution vector $\boldsymbol{\beta}$ which minimizes RSS is $\hat{\beta}=\left(X^{T}X\right)^{-1}X^{T}y$. The fitted values at the training inputs are $\hat{y}=X\hat{\beta}=X\left(X^{T}X\right)^{-1}X^{T}y$, where $\hat{y}_{i}=\hat{f}(x_{i})=x_{i}^{T}\hat{\beta}$.

1.3.2. Regression with Splines

In the above regression model, sometimes

$$f(X)=E(Y|\,X_{1},...,X_{m}) \tag{1.8}$$

can be nonlinear and nonadditive. Since, however, a linear model is easy to interpret, we want to represent $f(X)$ by a linear model. Therefore, 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.

Regression with splines is a very popular method as for moving beyond linearity [12]. 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 $h_{l}:\mathbb{R}^{m}\rightarrow\mathbb{R}$ be the jth transformation of X or basis function $(l=1,\ldots,M)$. Then, $f(X)$ is modelled by

$$f(X)=\sum_{l=1}^{M}\beta_{l}h_{l}(X), \tag{1.9}$$

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

$$\hat{\beta}=\left(H^{T}(x)H(x)\right)^{-1}H^{T}\left(x\right)Y, \tag{1.10}$$

where $H(x)=(h_{l}(x_{i}))_{\substack{i=1,\ldots,N;\\ l=1,\ldots,M}}$ is the matrix of basis functions evaluated at the input data.

Hence, $\hat{f}\left(X\right)$ becomes estimated by $\hat{f}(X)=h^{T}(X)\hat{\beta}$. For the special case $h_{l}(X)=X_{l}$ $(l=1,...,M)$ the linear model is recovered. Generally, in one dimension $(m=1)$, an *order M spline* with knots $\xi_{\kappa}(\kappa=1,\ldots,K)$ 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 $h_{l}(X)=X^{l-1}$ $(l=1,...,M)$ and $h_{M+\tau}(X)=(X-\xi_{\tau})_{+}^{M-1}$ $(\tau=1,...,K)$, where $(t)_{+}$ stands for the positive part of a value t [12].

Fixed-knot splines are also called *regression splines*. It is necessary to select the order of the spline, the number of knots and their placement. One simple approach is to parameterize a family of splines by the number of basis elements or degrees of freedom and let the observations x_{i} determine the positions of the knots. We shall follow the latter approach

in Section 2; there, we shall define a special *index* for the selection of the spline degrees and, herewith, their orders.

Since the space of spline functions of a particular order and knot sequence is a vector space, there are many equivalent bases for representing them. Truncated power bases, being so conceptually simple, are not attractive numerically, because they can allow big rounding problems. The B-spline basis allows for efficient computations even when the number of knots K is large. For basic information about higher and 1-dimensional splines, we refer to [6].

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* (*Fox* 2002 [8]). But, if the number of independent variables is large in the models, many forms of nonparametric regression do not perform well; in those cases, the number of them must be diminished. This decreasing causes the variances of the estimates to be unacceptably large. It is also difficult to interprete nonparametric regression depending on smoothing spline estimates. To overcome these difficulties, *Stone* (1985) [21] 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) [10] 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, enumerated by the index i, then the ***additive model*** takes the from

$$E\left(Y_i|\, x_{i1}, ..., x_{i\,m}\right) = \beta_0 + \sum_{j=1}^{m} f_j\left(x_{ij}\right). \tag{1.11}$$

The functions f_j are unknown arbitrary and smooth functions and they are mostly considered to be splines, i. e., piecewise polynomial, since, e. g., polynomials themselves have a too strong or early asymptotics to $\pm\infty$ and, by this, they may not be satisfying for data fitting. As we shall explain, these functions are estimated by a smoothing on single ("separated") coordinates or clusters. A standard convention is to assume at x_{ij} : $E\left(f_j\left(x_{ij}\right)\right) = 0$ since, otherwise, there will be a free constant in each of the functions [13]. Here, x_{ij} is the jth coordinate of the ith input data vector; later on, in the backfitting algorithm, these values also serve as the knots of the interpolating (or smoothing) splines which appear there. The estimation of the f_j is done by an algorithm which performs a stepwise smoothing with respect to suitably chosen spline classes, to the points x_{ij}, and to the differences r_{ij} between an average y_i of the observed output data y_{ij} and the sum of our functions evaluated at the interpolation knots x_{ij}.

In our paper, we will introduce a new interpretation: We consider X_j consisting of m variates (coordinates of X) as a value of the jth one of m disjoint clusters of input data which we achieve; the cluster elements are enumerated by x_{ij}. That is, there is the understanding of ith data in the jth component of the input variable (classical separation of variable

approach), and there is the new understanding of the ith points of the jth cluster (I_j) of input data. If these clusters have the same cardinality, which we shall assume without loss of generality, we denote it by N. With this new interpretation we will find and refer to structures, e. g., regularly in time repeating input constellations. We will "separate" them by clusters which we call intervals (maybe in higher-dimensional sense). Here, "regular" means that we can see or without of generality assume correspondences between the sample times in these time intervals, e. g., there are mondays, tuesdays, ... We recall that the output values corresponding to the inputs are denoted by y_{ij}. Averaging over these values with respect to j, delivering $y_i := \sum_{j=1}^{m} y_{ij}$ $(i = 1, ..., N)$, will then represent, e. g., an observation mean over the mondays, tuesdays, etc., respectively. Since our explanations hold for the understanding in the sense of "separation of variables" and the one of "(separated) clusters" as well, we may keep both of them in mind and refer to these two ones in the following.

Additive models have a strong motivation as a useful data analytic tool. Each variable is represented separately in (1.11) 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. Each function f_j is estimated by an algorithm proposed by *Friedman and Stuetzle* (1981) [9] and called ***backfitting algorithm***. As the estimator for $\hat{\beta}_0$, the arithmetic mean (average) of the output data is used: $ave(y_i \mid i = 1, ..., N) := (1/N) \sum_{i=1}^{N} y_i$. This procedure depends on the partial residual against x_{ij}:

$$r_{ij} = y_i - \beta_0 - \sum_{k \neq j} \hat{f}_k (x_{ik}) , \tag{1.12}$$

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 [11]:

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

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

$$r_{ij} = y_i - \hat{\beta}_0 - \sum_{k \neq j}^{m} \hat{f}_k (x_{ik}), \ i = 1, ..., N,$$

$\hat{f}_k$ is updated by smoothing the *partials residuals*,

$r_{ij} = y_i - \hat{\beta}_0 - \sum_{k \neq j}^{m} \hat{f}_k (x_{ik}), \ i = 1, ..., N$, against x_{ij} ***until*** the functions almost do not change.

The backfitting procedure is also called *Gauss — Seidel* algorithm. To prove its **convergence**, *Buja and Hastie* [5] 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 jth component under Gauss — Seidel algorithm, the authors introduced the linear transformation

$$\hat{T}_j : IR^{Nm} \to IR^{Nm} \quad \mathbf{f} \mapsto \begin{pmatrix} \mathbf{f}_1 \\ \vdots \\ S_j \left(- \sum_{k \neq j} \mathbf{f}_k \right) \\ \vdots \\ \mathbf{f}_m \end{pmatrix} . \tag{1.13}$$

A full cycle of this algorithm is determined by $\hat{\mathbf{T}} = \hat{\mathbf{T}}_m \hat{\mathbf{T}}_{m-1} ... \hat{\mathbf{T}}_1$; then, $\hat{\mathbf{T}}^l$ correspond l full cycles. Here, S_j is a smoothing spline operator which performs the interpolation at the knots x_{ij}. In the case where all smoothing S_j are symmetric and have eigenvalues in [0, 1], then the backfitting algorithm always *converges*. In Subsection 2.5, 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 higher dimensional 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 usage of additive models (or generalized ones, introduced next) is a key step of this paper.

1.5. A Note on Generalized Additive Models

1.5.1. Introduction

To extend the additive model to a wide range of distribution families, *Hastie and Tibshirani* (1990) [13] proposed *generalized additive models* (*GAM*) which are among the most practically used modern statistical techniques. Many often used statistical models belong to this general class, e. g., additive models for Gaussian data, nonparametric logistic models for binary data, and nonparametric log-linear models for Poisson data.

1.5.2. Definition of a Generalized Additive Model

If we have m covariates (or values of clusters) $X_1, X_2, ..., X_m$, comprised by the m-tuple $X = (X_1, ..., X_m)^T$, and a response Y to the input X assumed to the have exponential family density $h_Y(y, \alpha, \varpi)$ with the mean $\mu = E(Y | X_1, ..., X_m)$ linked to the predictors through a link function G. Here, α is called the natural parameter and ϖ is the dispersion parameter. Then, in our regression setting, a *generalized additive model* takes the form

$$G(\mu(X)) = \psi(X) = \beta_0 + \sum_{j=1}^{m} f_j(X_j). \tag{1.14}$$

Here, the function f_j are unspecified ("nonparametric") and $\theta = (\beta_0, f_1, ..., f_m)^T$ is the unknown parameter to be estimated; G is the link function. The incorporation β_0 as some average outcome allows us to assume $E(f_j(X_j)) = 0$ $(j = 1, ..., m)$. 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. In future research, we will extend the new methods of this paper to *generalized* additive models.

2. Investigation of the Additive Model

In this section, we analytically and numerically investigate the additive model. Before we introduce and study our modified backfitting algorithm, we approach the input and output data and address the aspect of stability.

2.1. Clustering of Input Data

2.1.1. Introduction

Clustering is the process of organizing objects into groups $I_1, I_2, ..., I_m$ or, higher dimensionally: $Q_1, Q_2, ..., Q_m$, 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. 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) [2, 14]. In this paper, differently from the classical understanding, we always interpret clustering 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. The aim of clustering is to determine the intrinsic grouping in a set of unlabeled data. Therefore, we decide about clustering methods which depend on a criterion. This criterion must be supplied by the user, in such a way that the result of the clustering will suit his needs [16]. Clustering algorithms can be applied in many fields like marketing, biology, libraries, book ordering, insurance, city-planning or earthquake studies. For further information we refer to [3].

For each clusters, namely, I_j (or Q_j), we denote the elements by x_{ij}. This interpretation is new, since according to the classical understanding of additive models, there is a separation of variables made by the model functions added and x_{ij} is just the jth coordinate of the ith input data vector.

2.1.2. Clustering for 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 models used need to be able to cope with such values, inventing values to fill such gaps is not a good way to proceed. Government and private research groups moved from sampling and analyzing financial data annually to monthly, to weekly, to daily, and, now, intradaily. One could choose to aggregate the financial data to fixed intervals of time, justifying this action by the large number of recording errors and the transient reactions to news that occur during intense trading periods. Naturally, some information would be lost due to aggregation. Too large or small an aggregating interval, or the changing of the interval when it should remain constant, could cause information to be lost or distorted. Methods which address the irregularity of ultra-high frequency financial data are needed.

The following three parts of Fig. 1 are showing some important cases of input data distribution and clustering: the *equidistant case* (cf. (a)) where all points can be put into one cluster (or interval) I_1, the *equidistant case with regular breaks* (weekends, holidays, etc.; cf. (b) where the regularly neighbouring points and the free days could be put in separate cluster intervals I_j, and the *general case* (cf. (c)) where there are many interval I_j of different

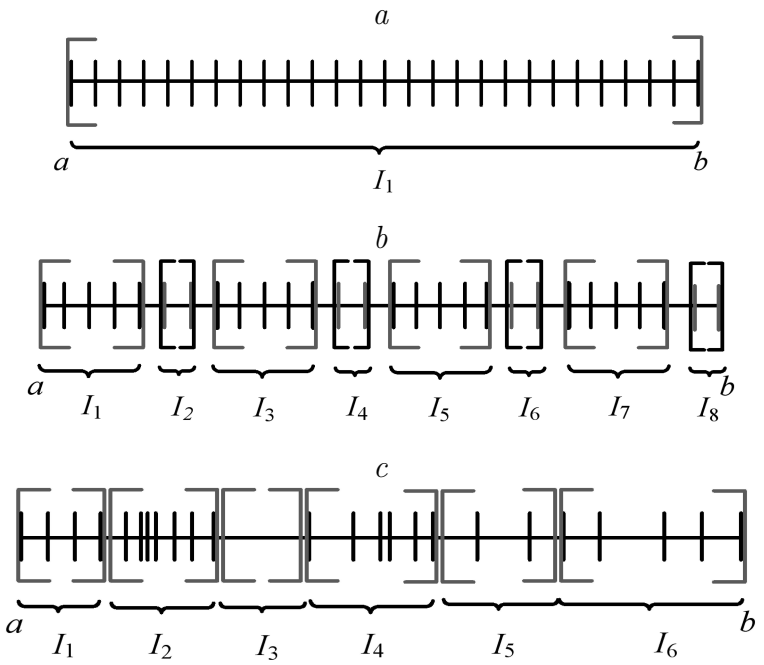

Fig. 1. Three important cases of *input data distribution* and its *clustering*: equidistance (a), equidistance with breaks (b), and general case (c).

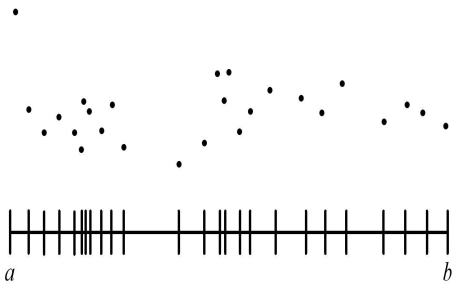

Fig. 2. Example of a *data* (scatterplot); here, we refer to case (c) of Fig. 1.

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.

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

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 ith output values of the clusters by

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

In the example of case $(1, b)$, 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 2.3).

2.2. Interpolating Splines

Let $x_{1j}, x_{2j}, ..., x_{Nj}$ be N distinct knots of $[a, b]$, where $a < x_{1j} < x_{2j} < ... < x_{Nj} < b$. The function $f_k(x)$ on the interval $[a, b]$ (or in $\mathbb{R}$) is a spline of some degree k relative to the knotsx_{ij} if [18]

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

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

Here, $C^{k-1}[a, b]$ is the set of functions defined on the interval $[a, b]$ which can be extended on an open neighbourhood of the interval such that it is $(k-1)$-times continuously differentiable at each element of the neighbourhood. To characterize a spline of degree k, $f_{k,i} := f_k|[x_{ij}, x_{i+1\,j}]$ can be represented by

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

There are $(k + 1)(N - 1)$ coefficients g_{li} to be determined. Furthermore, it has to hold $f^{(l)}_{k,i-1,}(x_{ij}) = f^{(l)}_{k,i}(x_{ij})$ $(i = 1, ..., N-2;\ l = 0, ..., k-1)$. Then, there are $k(N-2)$ conditions, and the remaining degrees of freedom are $(k + 1)(N - 1) - k(N - 2) = k + N - 1$ [18].

2.3. 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 jth interval I_j is defined by

$$D_j := \frac{\text{number of point } x_{ij} \text{ in } I_j}{\text{length of } I_j}.$$

This definitions can be directly generalized to the higher dimensional interval rather than intervals I_j, by referring to the higher dimensional volumes.

Variation is a quantifiable difference between individual measurements. Every repeatable process exhibits variation. If over the interval I_j we have the data $(x_{1j}, y_{1j}), ..., (x_{Nj}, y_{Nj})$, then the variation of these data refers to the output dimension y and it is defined as

$$V_j := \sum_{i=1}^{N-1} |y_{i+1j} - y_{ij}|.$$

Since we do the spline interpolation in the course of the algorithm with respect to the residuals r_{ij}, we can also in every iteration separately refer to a variation, defined by $V_j := \sum_{i=1}^{N-1} |r_{i+1j} - r_{ij}|$; for simplicity, we suppressed the iteration index. The change of the reference outputs could be made after some iterations. We point out that this policy and the determination of the spline degrees discussed in Subsection 2.4 are left to the practitioner who is looking at the given data and follows the algorithm or, if a closed model is preferred rather than adaptive elements, they can be decided based on thresholds.

If the variation 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} (of r_{ij}) may contribute to this very much and mean instability of the model.

2.4. Index of Data Variation

Still we assume that $I_1, ..., I_m$ (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

$$Ind_j := D_j V_j$$

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 2.3, 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 PRRS* now:

$$PRSS(\beta_0, f_1, ..., f_m) := \sum_{i=1}^{N} \left\{ y_i - \beta_0 - \sum_{j=1}^{m} f_j(x_{ij}) \right\}^2 + \sum_{j=1}^{m} \varphi_j \int_a^b \left[f_j''(t_j) \right]^2 dt_j. \qquad (2.1)$$

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.

We also note that the definition of *PRSS* in (2.1) can be extended to the higher dimensional case by using the corresponding higher dimensional integrals. However, one basic idea of the (generalized) additive models just consists in the separation of the variables.

In (2.1), $\varphi_j \geq 0$ are tuning or *smoothing* parameters and they represent a tradeoff between first and second term. Large values of φ_j yield smoother curves, smaller values result in more fluctuation. It can be shown that the minimizer of *PRRS* 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,...,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). In this case, $\hat{\beta}_0 = ave(y_i | i$=1,...,$N)$, 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

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

where $f = (f_1, ..., f_m)^T$. 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

$$\begin{array}{l} \text{Minimize} \quad F(\beta_0, f) \\ \text{subject to} \quad g_j(f) \leq 0 \quad (j = 1, ..., m). \end{array} \tag{2.2}$$

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

$$L\left((\beta_0, f), \varphi\right) = \sum_{i=1}^{N} \left\{ y_i - \beta_0 - \sum_{j=1}^{m} f_j(x_{ij}) \right\}^2 + \sum_{j=1}^{m} \varphi_j \left(\int \left[f_j''(t_j) \right]^2 dt_j - M_j \right), \tag{2.3}$$

where $\varphi := (\varphi_1, ..., \varphi_m)^T$. Here, $\varphi_j \geq 0$ are auxilary *penalty parameters* introduced in [4]. In the light of our optimization problem, they can now be seen as *Lagrange multipliers* associated with the constraints $g_j \leq 0$. The *Lagrangian dual problem* takes the form

$$\max_{\varphi \geq 0} \min_{(\beta_0, f)} L\left((\beta, f), \varphi\right). \tag{2.4}$$

The solution of this optimization problem (2.4) will help us for determining the smoothing parameters φ_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 additive model.

2.5. Modified Backfitting Algorithm for Additive Model

2.5.1. Additive Model Revisited

For the additive model (cf. Subsection 1.4), we will modify the backfitting algorithm used before for fitting additive model. For this reason, we will use the following theoretical setting in term of conditional expectation (*Buja, Hastie and Tibshirani* (1989) [5]), where $j = 1, 2, ..., m$:

$$f_j(X_j) = P_j \left(Y - \beta_0 - \sum_{k \neq j} f_k(X_k) \right) := E \left(Y - \beta_0 - \sum_{k \neq j} f_k(X_k) | \, X_j \right). \tag{2.5}$$

Now, to find more robust estimation for $f_j(X_j)$ in our additive model, let us add the term $-\sum_{k=1}^{m} \varphi_k \int [f_k''(t_k)]^2 dt_k$ to equation (2.5). In this case, (2.5) will become the update formula

$$f_j(X_j) \leftarrow P_j \left(Y - \beta_0 - \sum_{k \neq j} f_k(X_k) \right) - \left(\sum_{k=1}^{m} \varphi_k \int [f_k''(t_k)]^2 dt_k \right) =$$

$$= E\left(Y - \beta_0 - \sum_{k\neq j} f_k(X_k)|\, X_j\right) - \left(\sum_{k=1}^{m} \varphi_k \int [f_k''(t_k)]^2 dt_k\right) \tag{2.6}$$

or

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

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

where $\sum\limits_{k\neq j}^{m} \varphi_k \int \left[\hat{f}_k''(t_k)\right]^2 dt_k =: c_j$ (constant, i. e., not depending on the knots); the functions $\hat{f}_j$ are unknown and to be determined in the considered iteration. By using the notation c_j, we underline that the weighted sum of integrals which c_j denotes is just a scalar value and, herewith, introducing a constant shift in the following. Therefore, we can write equation (2.6) as

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

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

$$Z_j(X_j) \leftarrow E\left(Y - \beta_0 - \sum_{k\neq j} Z_k\,(X_k)|X_j\right). \tag{2.7}$$

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 [5]. If this idea is applied to additive model $\eta(\mathbf{X})$, then the minimizer of $E\,(Y - \eta(X))^2$ will give the closest additive approximation to $E\,(Y\,|\mathbf{X})$. Equivalently, the following system of *normal equations* is necessary and sufficient for $\mathbf{Z} = (Z_1, ..., Z_m)^T$ to minimize $E\,(Y - \eta(\mathbf{X}))^2$ (for the formula without intercept β_0, we refer to [5]):

$$\begin{pmatrix} I & P_1 & . & . & P_1 \\ P_2 & I & . & . & P_2 \\ . & . & . & . & . \\ . & . & . & . & . \\ P_m & P_m & . & . & I \end{pmatrix} \begin{pmatrix} Z_1\,(X_1) \\ Z_2\,(X_2) \\ . \\ . \\ Z_m\,(X_m) \end{pmatrix} = \begin{pmatrix} P_1\,(Y - \beta_0\mathbf{e}) \\ P_2\,(Y - \beta_0\mathbf{e}) \\ . \\ . \\ P_m\,(Y - \beta_0\mathbf{e}) \end{pmatrix}, \tag{2.8}$$

where $\mathbf{e}$ is the N-vector or entries 1; or, in short,

$$\mathbf{PZ} = \mathbf{Q}\,(\mathbf{Y} - \boldsymbol{\beta}_0)\,.$$

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 $(Y, \mathbf{X})$ in the (2.8) by their realizations $(y_i, \mathbf{x}_i)$, $\mathbf{x}_i = (x_{i1}, ..., x_{im})$, and the conditional

expectations $P_j = E(\cdot \mid X_j)$ by smoothers S_j on x_j,

$$\begin{pmatrix} I & S_1 & . & . & S_1 \\ S_2 & I & . & . & S_2 \\ . & . & . & . & . \\ . & . & . & . & . \\ S_m & S_m & . & . & I \end{pmatrix} \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ . \\ . \\ \mathbf{z}_m \end{pmatrix} = \begin{pmatrix} S_1(\mathbf{y} - \hat{\beta}_0\mathbf{e}) \\ S_2(\mathbf{y} - \hat{\beta}_0\mathbf{e}) \\ . \\ . \\ S_m(\mathbf{y} - \hat{\beta}_0\mathbf{e}) \end{pmatrix}. \tag{2.9}$$

For (2.9) we briefly write (in our estimation notation used from now on):

$$\hat{\mathbf{P}}\mathbf{z} = \hat{\mathbf{Q}}\left(\mathbf{y} - \hat{\boldsymbol{\beta}}_0\right) =: \hat{\mathbf{Q}}\mathbf{y}_1.$$

Here, referring to the base spline representation (cf. (1.9)) over I_j, $S_j = (h_{jl}(x_i))_{\substack{i=1,\ldots,N \\ l=1,\ldots,N}}$ are smoothing matrices of type $N \times N$ (i is the row index and l the column index), $\mathbf{z}_j$ are N-vectors representing the spline function $\hat{f}_j + \varphi_j \int \left[\hat{f}''_j(t_j)\right]^2 dt_j$ in a canonical form (1.12), i. e., $\sum\limits_{l=1}^{N} \theta_{jl} h_{jl}(X)$ (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, (2.9) is an $(Nm \times Nm)$-system of *normal equations*. The solutions to (2.9) 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 $\mathbf{z}_j \leftarrow S_j\left(\mathbf{y} - \hat{\beta}_0\mathbf{e} - \sum\limits_{k \neq j} \mathbf{z}_k\right)$.

In case, we want to emphasize $\hat{\boldsymbol{\beta}}_0$ among the unknowns, i. e., $\left(\hat{\boldsymbol{\beta}}_0^T, \mathbf{z}_1^T, ..., \mathbf{z}_m^T\right)^T$, then equation (2.9) can equivalently be represented in the following quadratic form:

$$\begin{pmatrix} O & O & O & . & . & O \\ I & I & S_1 & . & . & S_1 \\ I & S_2 & I & S_2 & . & S_2 \\ . & . & . & I & . & . \\ . & . & . & . & . & . \\ I & S_m & . & . & . & I \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}}_0 \\ \mathbf{z}_1 \\ \mathbf{z}_2 \\ . \\ . \\ \mathbf{z}_m \end{pmatrix} = \begin{pmatrix} Oy \\ S_1 y \\ S_2 y \\ . \\ . \\ S_m y \end{pmatrix}, \tag{2.10}$$

where O is the $N \times N$ matrix with zero entries. There is a variety of efficient methods for solving the system (2.9), which depend on both the number and types 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 $\hat{\mathbf{P}}\mathbf{z} = \hat{\mathbf{Q}}\mathbf{y}_1$ 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 [17].

2.5.2. 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 (2.9). It coincides with the backfitting algorithm. If in the

algorithm we write $\hat{z}_j = \hat{f}_j + \varphi_j \int \left[\hat{f''}_j\,(t_j)\right]^2 dt_j$ (in fact, the functions $\hat{f}_j$ are unknowns), then, the pth iteration in the backfitting or Gauss — Seidel includes the additional penalized curvature term. Not forgetting the step-wise update of the penalty parameterφ_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*** $j = 1, 2, ..., m,\ 1, 2, ..., m,\ 1, 2, ..., m, ...$

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

This iteration is done until the individual functions do not change: Here, in each iterate, $\hat{z}_j$ is by the spline with reference to the knots x_{ij} found by the values $y_i - \hat{\beta}_0 - \sum\limits_{k \neq j} \hat{z}_k\,(x_{ik})$ $(i = 1, ..., N)$, i. e., by the other functions $\hat{z}_k$ and, finally, by the functions $\hat{f}_k$ and the penalty (smoothing) parameter φ_k. Actually, since by definition it holds $\hat{z}_j = \hat{f}_j + \varphi_j \int \left[\hat{f''}_j(t_j)\right]^2 dt_j$, throughout the algorithm we must have a *book keeping* about both $\hat{f}_j$ and the curvature effect $\varphi_j \int \left[\hat{f''}_j(t_j)\right]^2 dt_j$ controlled by the penalty parameter φ_j which we can update from step to step. This book keeping is guaranteed since $\hat{f}_j$ 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 $\varphi_j \int \left[\hat{f''}_j(t_j)\right]^2 dt_j$ is constant, the second order derivative of $\hat{z}_j$ is

$$\hat{z}''_j(t_j) = \hat{f''}_j(t_j);$$

this yields

$$\varphi_j \int \left[\hat{f''}_j(t_j)\right]^2 dt_j := \varphi_j \int \left[\hat{z}''_j(t_j)\right]^2 dt_j$$

and, herewith,

$$\hat{f}_j := \hat{z}_j - \varphi_j \int \left[\hat{f''}_j(t_j)\right]^2 dt_j.$$

2.5.3. Discussion about Modified Backfitting Algorithm

If we consider our optimization problem on (2.1) (cf. also (2.4)) as fixed with respect to φ_j, then we can carry over the *convergence theory* about backfitting (see Section 1.3) to the present modified backfitting, replacing the functions $\hat{f}_j$ by $\hat{z}_j$.

However, at least approximatively, we have to guarantee feasibility also, i. e.,

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

If $\int \left[\hat{f''}_j(t_j)\right]^2 dt_j \leq M_j$, then we preserve the value of φ_j for $l \leftarrow l+1$; otherwise, we increase φ_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 φ_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,

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

2.5.4. Alternative Approaches

If we have $I_1,...,I_m$ interval and we chosen penalty terms with the distribution of the knots taken into account by bell-shaped density functions multiplied at the squared second-order derivatives $\left[\hat{f}''_j(t_j)\right]^2$, then our problem *PRSS* takes the form

$$PRSS(\beta_0,f_1,...,f_m):=\sum_{i=1}^{N}\left\{y_i-\beta_0-\sum_{j=1}^{m}f_j(x_{ij})\right\}^2+$$

$$+\sum_{j=1}^{m}\varphi_j\int\exp\left[-\frac{(x_{ij}-\bar{x}_j)^2}{\sigma_j^2}\right]\left[\hat{f}''_j(t_j)\right]^2 dt_j$$

(cf. (2.1) and (2.3)), where i and j have the same meaning as before, namely, for enumerating the data and spline functions, respectively. Herewith, there would be a "cut-off effect" outside of the intervals I_j and the modified penalty terms even more forced to be closer to zero (Fig. 3). Here $\bar{x}_j$ and σ_j^2 are respectively average value and variance for x_{ij} knots which is in the I_j interval and they are defined respectively,

$$\bar{x}_j=\frac{1}{N_j}\sum_{i=1}^{Nj}x_{ij}\,,\quad \sigma_j^2=\frac{1}{N_j-1}\sum_{i=1}^{Nj}(x_{ij}-\bar{x}_j)^2.$$

Besides of this and further improvements of the additive model and the corresponding modified backfitting algorithm being possible, the previous discussions teach us that the developed methods of *continuous optimization theory* will become an important complementary

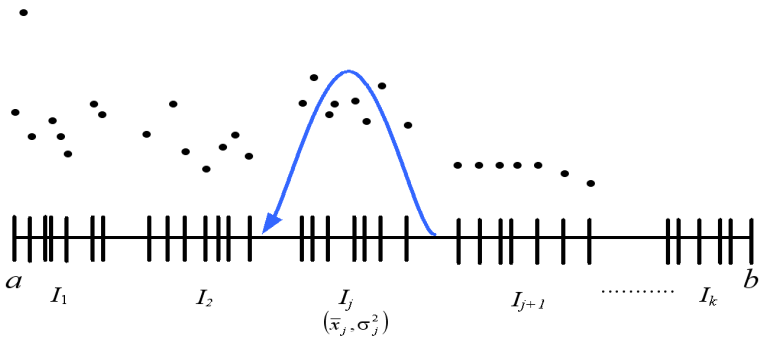

Fig. 3. Cutting-off in data approximation.

concept and alternative to the concept of backfitting algorithm. This will be subject of future research.

2.6. 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 as well as an interior point algorithm related with a closed approach to our problems and real-world applications by conic quadratic programming. In the following last section, will will sketch this second approach. A comparison and possible combination of these two algorithmic strategies is what we recommend in this pioneering paper.

3. Concluding Remarks and Outlook

This paper has given a contribution to the discrete approximation or regression of data in 1- and multivariate cases. The additive models have been investigated, input data grouped by clustering, its density measured, data variation quantified, spline classes selected by indices and thresholds, and their curvatures bounded with the help of penalization. Then, the backfitting algorithm which is also applicable for data classification has become modified. This *modified backfitting algorithm* which we present in our pioneering paper gives a new tool in the arsenal of approximating the unknown functions $\hat{f}_j$ while governing the instability caused. What is more, our paper offers to the colleagues from the practical areas of real-world examples a number of refined versions and improvements. But, this algorithm has some disadvantage in the iterative update of the penalty parameter. Indeed, this update changes the convergence behaviour of the algorithm. For this reason, a further utilization of modern optimization has been recommended [20], diminishing the adaptive and model-free elements of the algorithm. For this, if we turn to optimization problem (2.2), we can

equivalently write this optimization problem in the following form:

$$\begin{aligned} &\min_{t,\beta_0,f} t, \\ &\text{where} \quad \sum_{i=1}^{N} \left\{ y_i - \beta_0 - \sum_{j=1}^{m} f_j(x_{ij}) \right\}^2 \leq t^2, \quad t \geq 0, \\ &\qquad \int \left[f_j''(t_j) \right]^2 dt_j \leq M_j \quad (j = 1, 2, ..., m). \end{aligned}$$

As mentioned previously, the functions f_j are elements in a corresponding spline spaces. Instead of the integrals $\int \left[f_j''(t_j) \right]^2 dt_j$ we may use the approximative discretized form of Riemann sums, e. g., by evaluating the base splines $f_j''(\cdot)$ at the knots x_{ij}. Then, we can write our optimization problem equivalently as

$$\begin{aligned} &\min_{t,\beta_0,\theta} t, \\ &\text{where} \quad \|W(\beta_0, \theta)\|_2^2 \leq t^2, \\ &\qquad \|V_j(\beta_0, \theta)\|_2^2 \leq M_j \quad (j = 1, 2, ..., m), \\ &\qquad 0 \leq t, \end{aligned}$$

where $\omega_i := \sqrt{x_{i+1j} - x_{ij}}$ $(i = 1, 2, ..., N-1)$, $V_j^T = \left(f_j''(x_{1j})\omega_1, ..., (x_{N-1j})\omega_{N-1} \right)$ and

$$W(\beta_0, \theta) := (y_1 - \beta_0 - \sum_{j=1}^{m} f_j(x_{1j}), \ \ldots, \ y_N - \beta_0 - \sum_{j=1}^{m} f_j(x_{Nj}))^T.$$

Herewith, our optimization program has turned to a *conic quadratic programming* problem [14] which can be solved by *interior point* algorithms [15, 19]. In future, this approach and the main one presented in our paper will be further compared and combined by us, and applied on real-word data. By all of this we give a contribution to a better understanding of data from the financial world and many other practical areas, and to a more refined instrument of prediction.

The authors express their gratitude to Prof. Dr. Bülent Karsasözen (IAM, METU) for hospitality given to Pakize Taylan in DOSAP program, to him and to Prof. Dr. Hayri Körezlioğlu (IAM, METU), Prof. Dr. Klaus Schittkowski (University of Bayreuth, Germany), Prof. Dr. Peter Spellucci (Darmstadt University of Technology, Germany), Dr. Ömür Uğur and Dr. Çoşkun Küçüközmen (both from IAM, METU) for valuable discussions, and to the unknown referee for his valuable criticisim.

References

[1] Akume D., Weber G.-W. Cluster algorithms: theory and methods // J. of Comp. Technol. 2002. Vol. 7, N 1. P. 15–27.

[2] Aster A., Borchers B., Thurber C. Parameter Estimation and Inverse Problems. N.Y.: Acad. Press, 2004.

[3] Bock H.H., Sokoowski A., Jajuga K. Classification, Clustering, and Data Analysis: Recent Advances and Applications. B.: Springer-Verl., 2002.

[4] Boyd S., Vandenberghe L. Convex Optimization. L.: Cambridge Univ. Press, 2004.

[5] Buja A., Hastie T., Tibshirani R. Linear smoothers and additive models // The Ann. Stat. 1989. Vol. 17, N 2. P. 453–510.

[6] De Boor C. Practical Guide to Splines. B.: Springer-Verl., 2001.

[7] Easton V.J., McColl J.H. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[8] Fox J. Nonparametric Regression, Appendix to an R and S-Plus Companion to Applied Regression. Sage Publ., 2002.

[9] Friedman J.H., Stuetzle W. Projection pursuit regression // J. Amer. Statist. Assoc. 1981. Vol. 76. P. 817–823.

[10] Hastie T., Tibshirani R. Generalized additive models // Statist. Sci. 1986. Vol. 1, N 3. P. 297–310.

[11] Hastie T., Tibshirani R. Generalized additive models: some applications // J. Amer. Statist. Assoc. 1987. Vol. 82, N 398.

[12] Hastie T., Tibshirani R., Friedman J.H. The Element of Statistical Learning. N.Y.: Springer-Verl., 2001.

[13] Hastie T.J., Tibshirani R.J. Generalized Additive Models. N.Y.: Chapman and Hall, 1990.

[14] Nemirovskii A. Modern Convex Optimization: Lecture Notes. Israel Institute of Technology, 2005.

[15] Nesterov Y.E., Nemirovskii A.S. Interior Point Methods in Convex Programming. SIAM, 1993.

[16] Politicnico di Milano, Dipartimento de Elettronica e Informazione, A Tutorial on Clustering Algorithms. http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/

[17] Pringle R.M., Rayner A.A. Generalized Inverse Matrices With Applications to Statistics. Hafner Publ., 1971.

[18] Quarteroni A., Sacco R., Saleri F. Numerical Mathematics. Texts in Applied Mathematics 37. Springer, 1991.

[19] Renegar J. Mathematical View of Interior Point Methods in Convex Programming. SIAM, 2000.

[20] Spellucci P. Personal Communication. Germany: Darmstadt Univ. of Technology, 2006.

[21] Stone C.J. Additive regression and other nonparametric models // The Ann. Stat. 1985. Vol. 13, N 2. P. 689–705.

[22] Waggoner D.F. Spline methods for extractions interest rate curves coupon bound prices. Federal Reserve Bank of Atlanta Working Paper 97-10. 1997.

Received for publication 17 July 2006