SEMIPARAMETRIC ESTIMATION AND PREDICTION FOR TIME SERIES CROSS SECTIONAL DATA *

Olaf Bunke
Department of Mathematics
Humboldt University, Berlin
10099 Berlin

Summary: This paper discusses a methodology which uses time series cross sectional datafor the estimation of a time dependent regression function depending on explanatory variables and for the prediction of values of the dependent variable. The methodology assumes independent observations and is based on an adaptive semiparametric regression estimate depending on the observations from an adaptive running time window. The adaptation consists in the selection of the length (or horizon) of such a window together with one of numerous alternative parametric, nonparametric, additive and semiparametric estimators by minimization of a cross-validation criterion. In the prediction case the window contains only actual and past observations. It is shown, how to asses the influence of explanatory variables by generalized coefficients of determination which are adapted to the special objective of the statistical analysis. This aspect and our regression methodology is illustrated in the case of an analysis of stock market returns. An extended semiparametric methodology is also presented which allows the estimation of additive individual effects and which may essentially improve a traditional panel data analysis.

* The research for this paper was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse" at Humboldt University Berlin. The paper was printed using funds made available by the Deutsche Forschungsgemeinschaft.

1 Introduction

Statistics is applied in numerous fields of application where the dependence between variables and corresponding predictions is of central interest. In such problems a possible dependence on time and also on the special individual, object or sector under observation has to be considered. There is a rich literature on the treatment of such problems under quite restrictive assumptions, eg. on the application of regression methods in cross sectoral studies and on the application of time series analysis. While there are many problems where such methods work successfully, it is well-known, that the influence of the individuals (or sectors) and of time (simultaneously with explanatory variables) is not adequately perceived by these methods. This is the background for the literature on panel data analysis, which assume additive "individual" effects (see e.g. Hsiao (1986) and Baltagi (1996)), and on semiparametric regression models including time as an explanatory variable (see e.g. Engle et al. (1986) and Bunke (1992)). On the other hand it is also clear, that the assumptions of fixed individual effects in panel models leaves few (or sometimes none) degrees of freedom for estimating these effects, while the assumption of random individual effects is not always adequate for the considered applications.

An example is the analysis of the New York and of the German stock market, where the dependence of stock returns on different variables connected with firms (the "individuals") is of interest, e.g. on the firm market capitalization, its "beta-value" and its book-to-market-ratio. The number of firms is large in comparison to the number of years for which a statistical analysis makes sense (see e.g. Fama and French (1995) and Bunke, Sommerfeld and Stehle (1997)). In such a situation there is no hope to get sensible estimates of fixed firm effects as they would be calculated with a panel data analysis.

This paper is oriented towards methodology (in the spirit and as essential extension of the procedures in Bunke (1984), Droge (1991), Bunke (1992) and Bunke, Droge and Polzehl (1995)). It proposes a methodology which leads to sensible estimates of the regression function, leading to a description of the above mentioned dependences on explanatory variables and on time. This methodology is based on an adaptive running time window. The adaptation consists in the selection of the length (or horizon) of such a window together with one of numerous alternative parametric, nonparametric, additive and semiparametric estimators by minimization of a cross-validation estimate of the MSEP (mean square error of prediction). This approach is described in Section 2., while section 3. presents a modification oriented towards prediction, where the time window only contains actual and past observations.

An important problem is often the comparison of the influence of different explanatory variables and the search for the most influencial.

In section 4 we show how to deal with this problem. We allow for some flexibility taking into account, that sometimes it will be more informative to know the ammount of influence (or equivalently the "predictive power") of variables for some transformation or even for certain qualitative properties of the dependent variable. In the stock market case, theses properties could be e.g. a "positive" or "highest" stock return.

A special approach which uses some ideas of the procedures described in section 2. has already been applied by Mai and Polzehl (1991) in the short term prediction of electricity demand, while the procedures in their present complex form are now being applied in the analysis of the German stock market. Some of the results have been already presented in Bunke, Sommerfeld and Stehle (1997).

The last section 5 is devoted to an extension of our methodology by inclusion of an assessment of the dependence on the individuals. We combine our adaptive semiparametric approach with an extension of panel data analysis and for this we adapt a methodology introduced by Bunke and Castell (1998).

2 An adaptive semiparametric regression procedure

We consider observations of real variables $X_1, \ldots, X_k, Y$ with the aim of identifying a dependence of Y on some of the "explanatory" variables $X_1, \ldots, X_k$ which have influence on Y. For each moment $t \in \mathcal{T} = \{1, \ldots, T\}$ we have observations

$$x_{1ti}, \ldots, x_{kti}, y_{ti}$$

of these variables for individuals (or sectors) indexed by $i \in \mathcal{N}_t$. In the most general case are allow different individuals at different moments t (or missing values, if a fixed class of individuals is considered). Let n_t be the number of elements in $\mathcal{N}_t$. We assume the vectors $x_{ti} := (x_{1ti}, \ldots, x_{kti})$ to be fixed and the observations y_{ti} to be realizations of independent random variables $Y_{ti} \quad (t \in \mathcal{T} \ \in \mathcal{N}_t)$.
Leaving out the independence assumption and allowing for correlations between individuals or over time requires additional more complex tools than those proposed in this paper and will be the objective of a forthcoming paper.
We assume the existence of means and variances

$$\text{(2.1)} \qquad \mathrm{E}Y_{ti} = f_t(x_{ti}) =: \mu_{ti}, \quad \mathrm{D}Y_{ti} = \sigma_t^2.$$

The regression functions $f_t : \mathcal{X} \to R^1$, being defined on the range $\mathcal{X}$ of the vector $X := (X_1, \ldots, X_k)$ of explanatory variables and the variances σ_t^2 are unknown. We may also allow the assumption, that the observation vectors (x_{ti}, y_{ti}) are realizations

of independent variables (X_{ti}, Y_{ti}). We use then a "conditional approach", interpreting moments (as in (2.1) and later in (2.3), (2.31)) as conditional moments under the condition $X_{ti} = x_{ti}$ $(t \in \mathcal{T},\ i \in \mathcal{N}_t)$.
Note that in the most general case the regression function f_{ti} and the variances σ_{ti}^2 should also depend on the individual index i (see section 5), but our time dependent model (2.1) is still more general than usual time invariant regression models with homogeneous variance. Although formally there is no difference in the dependence of the moments of Y_{ti} on t and on i, the experiences about dependence on time and on the values of explanatory variables $X_1, \ldots, X_k$ in many real situations suggest, that at least some "smoothness" in such a dependence will be present, that is e.g. for fixed i small changes in time t and in x_{ti} lead to small changes in the moments μ_{ti}, σ_t^2. This allows the application of parametric and nonparametric estimates for μ_{ti} and σ_t^2. A similar assumption about the dependence of a regression function f_{ti} on the index i seems in general not to be sensible. In section 5 we will discuss alternative assumptions leading to a estimation procedure which also considers the direct dependence of f_{ti} on i.

We propose two alternative approaches which roughly may be described in the following way:

1. Fitting at each moment t the same semiparametric model M, using the observations at the moment t and possibly at neighbouring moments $t \pm 1, \ldots, t \pm h$ within a "horizon" h. The moments $t - \tau$ are of course excluded, if negative, as well as the moments $t + \tau$, if larger than T. The fitting leads to an estimate $\hat{f}_t^{M,h}$ of f_t. The model M is selected from a class $\mathcal{M}$ of models, simultaneously with the horizon h from the class $\mathcal{H} = \{0, 1, \ldots T-1\}$ of possible horizons. The selection is performed by minimization of the cross-validation estimate $C(M, H)$ of the MSEP $R(M, h)$ (means square error for prediction of a variable Y_{ti}^*, which has the same distribution as Y_{ti} but is independent of the other variables, $Y_{\underline{ti}}$ by $\hat{Y}_{ti} = \hat{f}_t^{M.h}(x_{ti})$):

$$R(M,h) := n^{-1} \sum_{t \in \mathcal{T}} S_t(M,h), \tag{2.2}$$

$$n := \sum_{t \in \mathcal{T}} n_t \quad , \quad S_t(M,h) := \sum_{i \in \mathcal{N}_t} \{\sigma_t^2 + \mathrm{E}|\hat{f}_t^{M,h}(X_{ti}) - f_t(X_{ti})|^2\}. \tag{2.3}$$

Obviously the MSEP is identical with the MSE for the estimation of the values μ_{ti} of the regression function up to a constant which neither depends on the model nor on the horizon.
We call this approach shortly ATFR (adaptive time dependendent fitting of a regression model).

2. Selecting at each moment t a model $M_t \in \mathcal{M}$ together with a time horizon $h_t \in \mathcal{H}_t = \{1, \ldots, \inf[t-1, T-1]\}$ by minimization of the cross-validation estimate $C_t(M,h)$ of the MSEP $R_t(M,h) := n_t^{-1} S_t(M,t)$ for prediction of $Y_t := (Y_{t1}, \ldots, Y_{tn})'$. We call this approach shortly ATMR (adaptive time dependent model selection based regression). In this approach a possible time evolution of the true underlying model for f_t may be followed. But unfortunately on the other side the selection of a model at each moment t will only be done based on the relative few observations within a horizon and therefore possibly lead to higher variability and larger prediction errors. In the approach 1. the model selection is based on a criterion $C(M,h)$ depending in all n observations.

In the following we present the details of the ATFR. The implementation of ATMR follows the same pattern with the obvious modification given by the use of C_t in place of C (see (2.9)) as a selection criterion.

The ATFR consists first in fixing for each horizon $h \in \mathcal{H}$ a class $\mathcal{F}^h = \{\hat{f}^{M,h} \mid M \in \mathcal{M}\}$ of regression estimators

$$\hat{f}^{M,h} : \mathcal{Z}^h \to \mathcal{F} = \{f \mid : \mathcal{X} \to R^1\}, \tag{2.4}$$

where

$$\mathcal{Z}^h = \{z = (t,x,y) \mid t \in \mathcal{T},\ x \in R^{m(t,h),k}, y \in R^{m(t,h)}\}, \tag{2.5}$$

$$m(t,h) = \sum_{\tau=\underline{t}_h}^{\overline{t}_h} n_\tau\ , \qquad \underline{t}_h := \max\{1, t-h\}\ , \qquad \overline{t}_h := \min\{T, t+h\} \tag{2.6}$$

Given the observations

$$x_t := (x_{t1}, \ldots x_{tn_t}) \quad , y_t := (y_{t1}, \ldots, y_{tn_t}) \quad (t \in \mathcal{T}), \tag{2.7}$$

the estimator $\hat{f}^{M,h}$ leads to estimates

$$\hat{f}_t^{M,h} := \hat{f}^{M,h}(t, x_{\underline{t}_h}, \ldots, x_{\overline{t}_h}, y_{\underline{t}_h}, \ldots, y_{\overline{t}_h}) \in \mathcal{F} \tag{2.8}$$

of the regression functions f_t and $\hat{\mu}_{ti}^{M,h} := \hat{f}_t^{M,h}(x_{ti})$, of their values $\mu_{ti} = f_t(x_{ti})$.
The estimates $\hat{\mu}_{ti}^{M,h}$ are based on running time windows $\mathcal{T}_t^h := [\underline{t}_h, \overline{t}_h]$.
For each pair (t,i) we denote in the following by $\tilde{\mu}_{ti}^{M,h}$ the estimate $\hat{f}^{M,h}(x_{ti})$ of μ_{ti} calculated leaving out the observation (x_{ti}, y_{ti}), when estimating the regression function

f_t and of its values μ_{ti}. That is modifying the estimator $\hat{f}^{M,h}$ in the straightforword way to depend on t and on $[m(t,h)-1]k$ and $m(t,h)-1$ dimensional variables x and y resp. (see (2.4)and (2.5)).
The second step in ATFR is the simultaneous selection of the horizon $h \in \mathcal{H}$ and of an estimator $\hat{f}^{M,h}$ from $\mathcal{F}^h$ by minimization of the cross-validation criterion

$$(2.9) \qquad \begin{aligned} C(M,h) &:= n^{-1} \sum_{t \in \mathcal{T}} n_t \, C_t(M,h) \qquad , \\ C_t(M,h) &:= n_t^{-1} \sum_{i \in \mathcal{N}_t} \mid y_{ti} - \tilde{\mu}_{ti}^{M,h} \mid^2 : \end{aligned}$$

$$(2.10) \qquad C(\hat{M},\hat{h}) := \min_{M,h} C(M,h).$$

The "adaptive" estimator $\hat{\hat{f}} := \hat{f}^{\hat{M},\hat{h}}$ is then the final estimator leading to estimates $\hat{\hat{f}}_t$ of f_t (see also (2.8)) and $\hat{\hat{\mu}}_{ti} := \hat{\hat{f}}_t(x_{ti})$ of the values of the regression function.

In a real application it could be even more useful to use an estimate $\hat{f}^{\tilde{M},\tilde{h}}$, which is nearly optimal in the sense of relatively small differences $C(\tilde{M},\tilde{h}) - C(\hat{M},\hat{h})$, but which has an especially simple or appealing structure, possibly allowing an interpretation in the field of application. This modification yields e.g. in the analysis of Bunke, Sommerfeld and Stehle (1997) of the German stock market an estimator $\hat{f}^{\tilde{M},\tilde{h}}$, which is equivalent to the well known model of Fama and French (1995) for the N.Y. stock market.

Now we describe the classes $\mathcal{F}^h$ of estimators $\hat{f}^{M,h}(M \in \mathcal{M})$, in which M characterizes the estimator type, while h determines the number of observations on which it depends. In principle it would be desirable to include many different types of estimators into the class $\mathcal{M}$. The following types of estimators may be seen as proposals already leaing to a very rich and flexible close $\mathcal{M}$, which e.g. has been useful and sufficient in the successful analysis of the German stock market. But also a replacement or an extension by other prefered estimators could as well be allowed, e.g. the inclusion of procedures of CART-type (see Breiman et al. 1994) or of neural-network-type (see White 1989).

Each class $\mathcal{F}^h$ may consist of several subclasses

$$\mathcal{F}_r^h = \{\hat{f}^{Mh} | M \in \mathcal{M}_r\}, \quad \mathcal{M} = \bigcup_r \mathcal{M}_r.$$

1. $\mathcal{M}_1$: Parametric estimators

We consider the class $\mathcal{M}_1$ of parametric models M of the form

$$(2.11) \qquad g_M(x|b) = T_0^{-1}[p_q(T_1[x_1], \ldots, T_k[x_k]; \tilde{T}|b)],$$

where $x = (x_1, \ldots, x_k) \in R^k$. The function p_q in (2.11) is a nonlinear extension of a polynomial of order q in the possibly nonlinearly transformed variables $\tilde{x}_\kappa = T_\kappa[x_\kappa]$ $(\kappa = 1, \ldots, k)$:

$$\begin{aligned} & p_q(\tilde{x}_1, \ldots, \tilde{x}_k; \tilde{T}|b) = \\ & = b_0 + \sum_{j_1=1}^{k} b_{j_1} \tilde{x}_{j_1} + \sum_{j_1, j_2=1}^{k} b_{j_1 j_2} T_{j_1 j_2}[\tilde{x}_{j_1} \tilde{x}_{j_2}] \\ (2.12) \qquad & + \ldots + \sum_{j_1, \ldots, j_q=1}^{k} b_{j_1, \ldots, j_q} T_{j_1, \ldots, j_q}[\tilde{x}_{j_1}, \ldots, \tilde{x}_{j_q}], \end{aligned}$$

where $\tilde{T} = (T_{11}, T_{12}, \ldots, T_{kk}, \ldots, T_{k \ldots k})$ and where each T_j $(j = (j_1, \ldots, j_r))$ is a transformation from a Box-Cox-type class $\mathcal{T}_j$ of four standard transformations:

$$(2.13) \qquad T_j[z] = \begin{cases} z & \text{(identical)} \\ (z + a_j)^{-1} & \\ \ln[(z + d_j)/s_j] & \\ \exp[c_j z] & \end{cases}$$

In (2.13) we denote by s_j^2 the empirical variance of the argument $z = \tilde{x}_{j_1} \ldots \tilde{x}_{j_r}$ in the transformation T_j:

$$(2.14) \qquad s_j^2 := n^{-1} \sum_{t,i} |z_{ti} - n^{-1} \sum_{\underline{t}, \underline{i}} z_{\underline{t}\underline{i}}|^2.$$

The constants a_j, d_j, c_j in (2.13) may be choosen in such a way, that the nonidentical transformations are as nonlinear as possible over the range of the corresponding argument, e.g. as proposed in Droge (1991) or Bunke, Droge and Polzehl (1994). The variables $\tilde{X}_\kappa = T_\kappa[X_\kappa]$ are transformations of the variables $X_\kappa (\kappa = 1, \ldots, k)$ by transformations T_κ from $\mathcal{T}_\kappa$. The transformation T_0 is an element of the class $\mathcal{T}_0$. Leaving out some (or none) of the terms in the polynomial (2.12), say the terms with indices in a set $\mathcal{J}$, leads to different models (2.11). These models g_M are obviously determined by the vector $\xi := (T_0, T_1, \ldots, T_k, \tilde{T})$ of transformations, the polynomial order q and the index set $\mathcal{J}$, so that we write $M = (\xi, q, \mathcal{J})$. The set $\mathcal{M}_1$ would be the set of all such M, subject to some convenient restrictions in order to limit the computational effort and to allow for easier interpretation. This could be a restriction $q \leq \overline{q}$ on the order q and (or) on the number of terms in the polynomial. E.g. in the stock market analysis of Bunke, Sommerfeld and Stehle

(1997), the restriction $\overline{q} = 3$ and of a maximal number of 12 terms proved to work. The heuristical background of a model (2.11) is that the regression of the possibly nonlinearly transformed dependent variable $\tilde{Y} = T_0[Y]$ on transformed explanatory variables $\widetilde{X}_1, \ldots, \widetilde{X}_k$ is approximated by a (possibly nonlinearly extended) polynomial of order q. The choice $T_0 = T_1 = \ldots = T_k =$ identical transformation means an approximation of the original dependent variable by a polynomial in the original explanatory variables, but the possibility of nonlinear transformations may lead to a better approximation and consequently to more accurate estimates of μ_{ti}. The estimate

$$\hat{f}_t^{M,h} := g_M(\bullet|\hat{b}_t^{M,h}) \tag{2.15}$$

of the regression function f_t is defined by ordinary least squares fitting within the time window $\mathcal{T}_t^h := [\underline{t}_h, \overline{t}_h]$

$$S_{t,M,h}(\hat{b}_t^{M,h}) = \min_b S_{t,M,h}(b), \tag{2.16}$$

using the sum of squared deviations corresponding to the window:

$$S_{t,M,h}(b) = \sum_{\tau=\underline{t}_h}^{\overline{t}_h} \sum_{i \in \mathcal{N}_\tau} |\, y_{\tau i} - g_M(x_{\tau i}|b)|^2. \tag{2.17}$$

The heuristical background for such an estimate based on the time window is the following:
If the model $g_M(x|b)$ is used for approximating the regression function f_t, there will be for each t some ("pseudo-true" or "projection") parameter value b_t with

$$b_t = \arg\min_b \sum_{i \in \mathcal{N}_t} E\,|\, f_t(X_{ti}) - g(X_{ti}\,|\,b)\,|^2. \tag{2.18}$$

If the regression function f_t and therefore b_t are believed to depend "smoothly" on t, then for each fixed t the values b_t will not differ much at moments τ from a sufficiently narrow window $\mathcal{T}_t^h$. The artificial assumption that these values b_τ are exactly identical will lead to the LSE $\hat{b}_b^{M,h}$ of b based on $m(t,h)$ observations. This estimator obviously would be better than a LSE based solely on the n_τ observations $x_{ti}, y_{ti} \quad (i \in \mathcal{N}_t)$, if the differences between the pseudo-true values b_τ are sufficiently small.

2. $\mathcal{M}_2$: Nonparametric kernel estimators

We consider possibly transformed kernel estimators $\hat{f}^{M,h}$ based on a multiplicative kernel smoothing of the possibly transformed observations $\tilde{x}_{\tau i}, \tilde{y}_{\tau i}$ for $\tau \in \mathcal{T}_t^h$ defined by different kernel functions K and bandwidths λ (see e.g. Härdle (1990)). Possibly $\lambda = (\lambda_1, \ldots, \lambda_k)$ is used as a vector of bandwidths assigned to the variables $\widetilde{X}_1, \ldots, \widetilde{X}_\kappa$. We may write $M = (K, \lambda, \xi)$ with $\xi = (T_0, T_1, \ldots, T_k)$ and consider the estimator $\hat{f}_t^{M,h} = \hat{f}_t^{M,h}[\bullet]$ of f_t defined by

$$\hat{f}_t^{M,h}[x] = T_0^{-1}\Big[\sum_{\tau=\underline{t}_h}^{\overline{t}_h} \sum_{i \in \mathcal{N}_\tau} \widetilde{K}_x^{M,h}(\tau, i) T_0[y_{\tau i}]\Big], \quad (x \in \mathcal{X}), \tag{2.19}$$

$$\widetilde{K}_x^{M,h}(\tau, i) = K_x^{M,h}(\tau, i)[\sum_{\underline{\tau}=\underline{t}_h}^{\overline{t}_h} \sum_{\underline{i} \in \mathcal{N}_{\underline{\tau}}} K_x^{M,h}(\underline{\tau}, \underline{i})]^{-1}, \tag{2.20}$$

$$K_x^{M,h}(\tau, i) = \prod_{\kappa=1}^{k} K(\lambda_\kappa^{-1} \,|\, T_\kappa[x_\kappa] - T_\kappa[x_{\kappa\tau i}] \,|). \tag{2.21}$$

If we restrict the transformations T_κ to the classes $\mathcal{T}_\kappa$ (see 1.), the kernels K to a class of few standard kernels (e.g. Epanechikov, triangular and normal) and λ to a convenient grid Λ in a finite interval $[0, \lambda_{\max})^k$, then we have fixed the set $\mathcal{M}_2$ for the admitted estimators $\hat{f}^{M,h}$ $(M \in \mathcal{M}_2)$.

It is interesting, that in the stock market analysis of Bunke, Sommerfeld and Stehle (1997) the estimators of this subclass and also of all following subclasses turn out to be worse (in the sense of cross-validation) than the best in the parametric class given by $\mathcal{M}_1$, which is flexible and better interpretable.

3. $\mathcal{M}_3$: Semiparametric model.

We consider estimates $\hat{f}^{M,h}$ in the semiparametric model

$$f_t^M(x) = h[g_M(x|b)], \tag{2.22}$$

where h is assumed to be an unknown smooth ("link") function, g_M is defined by (2.11), and $M = (K, \lambda, \xi, q, \mathcal{J})$. The estimator

$$\hat{f}_t^{M,h}[x] = \hat{h}^{M,h}[g_M(x|\hat{b}^{M,h})] \tag{2.23}$$

may be determined adapting the approach of Ichimura (1993), using a "preliminary kernel estimate"

$$\hat{h}_b^{M,h}[\gamma] := \sum_{\tau=\underline{t}_h}^{\overline{t}_h} \sum_{i\in\mathcal{N}_t} \tilde{K}_b^{M,h}(\gamma,\tau,i) y_{\tau i} \tag{2.24}$$

where

$$\tilde{K}_b^{M,h}(\gamma,\tau,i) = K_b^{M,h}(\gamma,\tau,i)[\sum_{\underline{\tau}=\underline{t}_h}^{\overline{t}_h} \sum_{\underline{i}\in\mathcal{N}_{\underline{\tau}}} K_b^{M,h}(\gamma,\underline{\tau},\underline{i})]^{-1}, \tag{2.25}$$

$$K_b^{M,h}(\gamma,\tau,i) = K(\lambda^{-1} \mid \gamma - g(x_{\tau i} \mid b) \mid). \tag{2.26}$$

and by least squares

$$\hat{b}^{M,h} = \arg\min_b \sum_{\tau=\underline{t}_h}^{\overline{t}_h} \sum_{i\in\mathcal{N}_t} \mid y_{\tau i} - \hat{h}_b^{M,h}[g_M(x_{\tau i} \mid b)] \mid^2. \tag{2.27}$$

Varying the kernel K the bandwidth λ over a grid in $[0, \lambda\max)$ and ξ as in 1. and 2. determines the class of semiparametric estimators

$$\{\hat{f}^{M,h} \mid M \in \mathcal{M}_3\}.$$

4. $\mathcal{M}_4$: Additive models

We consider estimators based on models

$$f(x|g) = T_0^{-1}\Big[\sum_{\kappa=1}^{k} g_\kappa(\tilde{x}_\kappa) + \sum_{\kappa=2}^{k}\sum_{\underline{\kappa}=1}^{k-1} g_{\kappa\underline{\kappa}}(\tilde{x}_\kappa, \tilde{x}_{\underline{\kappa}})\Big], \tag{2.28}$$

which are "additive (with second order interaction terms)" for the dependence of the transformed dependent variable $\widetilde{Y}$ on the transformed variables $\widetilde{X}_\kappa$. The functions appearing as components in $g := (g_1, g_2, \ldots, g_{\kappa\kappa-1})$ are assumed to be smooth. Leaving out some (or none) of the terms in (2.28) and varying $\xi = (T_0, \ldots, T_k)$ leads to the different transformed additive models $f^M(\tilde{x} \mid g)(M \in \mathcal{M}_4)$.
The vector g of functions in a model $f^M(x|g)$ may be estimated for observations $\tilde{X}_{\tau i}, \tilde{Y}_{\tau i}$ with $\tau \in \mathcal{T}_t^h$ by backfitting (see Hastie and Tibshirani (1990)). We arrive at estimates $\hat{g}^{M,h}$ and

$$\hat{f}^{M,h}[x] := f(T_1[x_1], \ldots, T_k[x_k] \,|\, \hat{g}^{M,h}). \tag{2.29}$$

5. $\mathcal{M}_5$: Partially parametric additive models

By $\mathcal{M}_5$ we consider the following combinations of the models from $\mathcal{M}_1$ and $\mathcal{M}_4$:

$$f_M(x|b,g) = T_0^{-1}[p_q^{\mathcal{J}}(T_1[x_1], \ldots, T_k[x_k]; \tilde{T}|b)] + \sum_{\kappa \in \mathcal{K}} g_\kappa(T_\kappa[x_\kappa]), \tag{2.30}$$

where $\mathcal{J}$ is the set of indices of the terms excluded in the polynomial p_q of order q and $\mathcal{K}$ characteristizes the variables X_κ included in the additive part. The estimation of b, g is done iteratively by backfitting and least squares (see Hastie and Tibshirani (1990)).

Remark 1.: From the above description of the subclasses $\mathcal{M}_j$ it is apparent, that our class $\{\hat{f}^{M,h} | M \in \mathcal{M}\}$ of estimators contains numerous alternatives of different forms. Provided that the true regression function f_t is not very irregular, it will be likely, that for each fixed $t \in \mathcal{T}$ there will be a function of one of the forms introduced in 1- to 5. which is near to the regression function f_t (that is, there is a small bias or "model error"), so that our adaptive estimates $\hat{\hat{\mu}}_{ti}$ of μ_{ti} should be relatively accurate, provided the number n_t of individuals considered at each moment t is sufficiently large. The large number of estimators and the calculation of their values and of corresponding cross-validation criteria for all moments t demand a considerable computational effort, but the gain in estimation accuracy in comparison to the application of a standard regression program (say for linear regression with model selection or for additive regression) would be the reward. As an example for the application of the procedure in the analysis of the German stock market, where the observation were of 3 variables for $n_t = 150$ firms during $T = 22$ years, the computation using an IBM risk 6000 workstation demanded two to three hours.

Remark 2.: The minimal cross-validation value $\hat{C} = C(\hat{M}, \hat{h})$ will be a rough estimate of the MSEP

$$\mathrm{MSEP}(\hat{\hat{f}}) = \frac{1}{n} \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{N}_t} \{\sigma_t^2 + \mathrm{E} \,|\, \hat{\hat{\mu}}_{ti} - \mu_{ti} \,|^2\} \tag{2.31}$$

for the adaptive estimator $\hat{\hat{f}}$, although obviously it is underestimating it. In our experience (by simulations in simple situations) often for moderate sample

sizes the MSEP($\hat{\hat{f}}$) seems to be only somewaht larger (10 to 20 %) than the minimal value $\hat{C}$, which does not take into account the variability of the "optimal" $\hat{M}, \hat{h}$ caused by their dependence on the observations. But at least the order of magnitude of the squared prediction errors may be assessed by $\hat{C}$.

3 A modified adaptive procedure for prediction

In some applications the objective of the analysis is not the estimation of the regression function but the prediction of the value y_{ti} on the basis of the knowledge of the value x_{ti} of the explanatory variables and of the past observations $x_{\tau\underline{i}}, y_{\tau\underline{i}}$ $(\tau < t, \underline{i} \in \mathcal{N}_\tau)$. Then it is possible to use the adaptive procedure of section 2., provided it is properly modified in view of the prediction objective. We have again two alternative approaches:

1. Prediction by ATFR (PATFR)

The procedure is the same as ATFR, but using only the observations from the time window $t \in \mathcal{H}_{t,h} = [\max\{1, t-h\}, t-1]$ when defining the estimators $\hat{f}^{M,h}$, that is leaving out variables in (2.5) and (2.8) corresponding to the moments $t, t+1, \ldots$.. The predictor of y_{ti} will then be $\hat{y}_{ti} := \hat{f}_t^{M,h}(x_{ti})$. The horizon h will be restricted to $\mathcal{H} = \{1, 2, \ldots, T-1\}$. The cross-validation criterion (2.4) for estimator or model selection is not adequate in the prediction situation and has to be replaced by a criterion $C_{T_\mathcal{M}}$, which depends only on observations for moments t up to the moment $T_\mathcal{M}$ (the "training time interval"), at which the selection of the horizon h and of the estimator (or equivalently of the predictor) $\hat{f}^{M,h}$ has to be done:

$$C_{T_\mathcal{M}}(M,h) = \Big[\sum_{t=1}^{T_\mathcal{M}} n_t\Big]^{-1} \sum_{t=1}^{T_\mathcal{M}} \sum_{i \in \mathcal{N}_t} |y_{ti} - \hat{y}_{ti}^{M.h}|^2, \tag{3.1}$$

After selection of $(\hat{M}, \hat{h})$ minimizing $C_{T_\mathcal{M}}$ the same estimator $\hat{f}_{ti}^{\hat{M},\hat{h}}$ will be used for prediction by $\hat{y}_{ti} = \hat{f}_t^{\hat{M},\hat{h}}(x_{ti})$ at all further moments $t > T_\mathcal{M}$. Sometimes it may be interesting at the final moment $t = T$ to deal with the complete set of data, that is to take $T_\mathcal{M} = T$, and to know which predictor would have been the best and which prediction accuracy would have been obtained when the same predictor is used for all moments $t \in \mathcal{T}$. In such a case it would be convenient to use C_T for the choice of M and h.

2. Prediction by ATMR (PATFR)

This is the modification of ATMR analogous to PATFR. The horizon h_t and the estimator $\hat{f}^{M_t,h_t}$ will be chosen at each moment $t \in \mathcal{T}$ and the adequate cross-validation criterion would be the corresponding squared prediction error at the previous moment $t-1$:

$$C^t(M,h) := (n_{t-1})^{-1} \sum_{i \in \mathcal{N}_{t-1}} |y_{t-1\,i} - \hat{y}^{M,h}_{t-1i}\,|^2 \tag{3.2}$$

4 Assessing the influence of variables

Usual objectives in the analysis of observations $x_{ti\kappa}, y_{ti}$ of the variables $X_1, \ldots X_k, Y$ at moments t and for individuals or sectors i are the assessment of the influence of the different explanatory variables X_κ on Y and the search for the most influencial variable or for a group of most influencial variables. Correlations or their squares as coefficients of determination are traditional tools for such an analysis.
In the situation given by the assumptions of section 2. with random variables X_{ti} the traditional definition has to be adequately modified. The ("ratio-type-")multiple coefficient of determination or squared multiple correlation ratio between Y and a subgroup $X^s := (X_1, \ldots, X_s)$ of s explanatory variables ($s \leq k$) may be defined in extension of the usual definition (where $T = n_t = 1$, see Rao (1973)) by

$$R^2(Y; (X_1, \ldots, X_s)) = B_s(X^s) := 1 - \frac{R_s(f^s)}{R_s(f^0)}, \tag{4.1}$$

where

$$R_s(f) = \sum_{t,i} \mathrm{E}\,|Y_{ti} - f_{ti}(X^s_{ti})|^2, \tag{4.2}$$

and

$$f^s_{ti}(x^s_{ti}) = \mu^s_{ti} := \mathrm{E}(Y_{ti}|X^s_{ti} = x^s_{ti}) \tag{4.3}$$

$$f^0_{ti} = f^0 := n^{-1} \sum_{t,i} \mathrm{E}Y_{ti}. \tag{4.4}$$

The partial coefficient of determination for variables $X_1, \ldots, X_s$ with $s < k$ is then defined by

$$R^2(Y, (X_1, \ldots, X_s) \,|\, (X_{s+1}, \ldots, X_k)) := \tag{4.5}$$
$$B_k(X^s) = \frac{R^2(Y, (X_1, \ldots, X_k)) - R^2(Y, (X_{s+1}, \ldots, X_k))}{1 - R^2(Y, (X_{s+1}, \ldots, X_k))}.$$

These coefficients of determination $B_s (s \leq h)$ are essentially normalized MSEP's (with values in $[0, 1]$) and therefore measure the "predictive power" of the variables $X_1, \ldots, X_s$ for predicting values of Y at the different moments t and individuals i. The partial coefficient $B_k(X^s)$ $(s < k)$ is usually interpreted as describing the linear or nonlinear dependence of Y on $X_1, \ldots, X_s$ under elimination of the influence of the other variables $X_{s+1}, \ldots, X_k$.

As the coefficients depend on the unknown distributions of the independent random vectors (X_{ti}, Y_{ti}) they have to be estimated. For this we may use our adaptive estimates, which we will denote by $\hat{\hat{f}}^s$, if applied to a situation with the observations $x_{ti1}, \ldots, x_{tis}, y_{ti}$.
The conditional means f^s given by (4.3) are estimated by $\hat{\hat{f}}^s$, while for each $t \in \mathcal{T}$ the observation mean

$$\hat{f}^0 = \overline{y} := n^{-1} \sum_{ti} Y_{ti} \tag{4.6}$$

will be an estimate of f_t^0. This yields the estimate

$$\hat{B}_s(X^s) := 1 - \frac{\sum\limits_{t,i} | \, y_{ti} - \hat{\hat{f}}_{ti} \, |^2}{\sum\limits_{t,i} | \, y_{ti} - \overline{y} \, |^2} \tag{4.7}$$

of the coefficient of determination (4.1).
While the coefficients (4.7) and the corresponding estimates of partial coefficients estimate the predictive paper which would have the explanantory variables, if the distributions P_{ti} of the variables (X_{ti}, Y_{ti}) would be known, it is even more interesting to know the predictive power attained without this knowledge, given only the n observations X_{ti}, Y_{ti}. A corresponding "cross-validated" coefficient of determination would be smaller, because the errors in estimating the means f^s and f^0 have to be taken into consideration: under unknown P_{ti} the predictors for Y_{ti} used in (4.1) will be $\hat{\hat{f}}_{ti}^s$ in place of the unknown optimal predictors f^s given by (4.3) $(s = 0, 1, \ldots, k)$. The "cross-validated coefficients of determination" $\widetilde{B}_s(X_1, \ldots, X_s)$ is obtained by

(4.7) replacing $\hat{\hat{f}}^s{}_{ti}$ by the analogous estimate $\tilde{f}^s_{ti}$ calculated leaving out the observation (x_{ti}, y_{ti}).

The Stock Market example

An example for an application of these coefficients is the analysis of the German stock market in Bunke, Sommerfeld and Stehle (1997), were it was shown, that the book-to-market ratio has the highest partial coefficient of determination among the considered variables and thus may be considered as most influential for the stock return. This fact has been already noticed (with less statistical justification) in empirical stock market research. In this application even the highest partial coefficient of determination is relatively small. This illuminates the well known fact, that stock returns may hardly be predicted with sensible accuracy.
On the other side, it may as well be possible that some more rough or summary properties z_{ti} of stock returns y_{ti} (in place of their exact ammount) may be predicted with a higher accuracy. Examples for such derived qualitative properties:

A. Positive or nonpositive stock returns:

$$z_{ti} = \begin{cases} 1 & y_{ti} > 0 \\ 0 & y_{ti} \leq 0 \end{cases} \tag{4.8}$$

If a function $g : \{(t,x)|t \in \mathcal{T}, x \in R^{n_t}\} \to \{0,1\}^{n_t}$ is used for the prediction $\hat{z}_t = g(t, x_{t1}, \ldots, x_{tn_t})$ of the vector $z_t := (z_{t1}, \ldots, z_{tn_t})$, then the prediction error has a natural description by the loss function

$$L(y_t, \hat{z}_t) = \sum_i |\hat{z}_{ti} - z_{ti}| = \sharp\{i|\hat{z}_{ti} \neq z_{ti}\}. \tag{4.9}$$

An alternative loss function may be even more appearing:

$$L(y_t, \hat{z}_t) = \sum_i (y^+_{ti} - \hat{z}_{ti} y_{ti}), \quad (y^+ := \max\{0, y\}). \tag{4.10}$$

This is just the loss in return at the moment t relative to the "optimal strategy" (investing a unit amount of capital amount in the asset i, if (and only if) the return y_{ti} is positive), if the capital is just assigned to the asset i, if the prediction $\hat{z}_{ti}$ is positive, that is, if the asset i is predicted to have positive return at the moment t.

B. High, moderate, low or negative stock return

A more refined view at the stock return would classify them, e.g. as high, moderate, low or negative.
Assigning formal number to these properties we have e.g. for some positive thresholds $\eta_1 < \eta_2$:

$$z_{ti} = \begin{cases} 1 & \text{if} \quad y_{ti} \in \mathcal{Y}_1 = (\eta_2, \infty) \\ 2 & \text{if} \quad y_{ti} \in \mathcal{Y}_2 = (\eta_1, \eta_2] \\ 3 & \text{if} \quad y_{ti} \in \mathcal{Y}_3 = [0, \eta_1] \\ 4 & \text{if} \quad y_{ti} \in \mathcal{Y}_4 = (-\infty, 0) \end{cases} \tag{4.11}$$

If for each i a wrong prediction of z_{ti} by $\hat{z}_{ti}$ is measured by a loss $c(z_{ti}, \hat{z}_{ti}) > 0$ and a correct prediction by zero loss $c(z_{ti}, z_{ti}) = 0$, we would have the loss function

$$L(y_t, \hat{z}_t) = \sum_i c(z_{ti}, \hat{z}_{ti}). \tag{4.12}$$

C. Highest stock return

In the context of stock markets an asset $z_t = z_t(y_t) \in \mathcal{N}_t$ with highest return

$$\max_{i \in \mathcal{N}_t} y_{ti} = y_{tz_t} \tag{4.13}$$

should be of special interest.
We see that the property depends at each fixed moment t on the whole vector $y_t = (y_{t1}, \ldots, y_{tn_t})$ of firm returns, and its prediction $\hat{z}_t = g(t, x_t)$ has to be done using the vector $x_t = (x_{t1}, \ldots, x_{tm_t})$ of values of explanatory variables for all firms i. A sensible loss function would be

$$L_t(y_t, \hat{z}_t) = y_{tz_t} - y_{t\hat{z}_t}, \tag{4.14}$$

which is the loss in return compared with the highest return, if a unit amount of capital is invested in the asset $\hat{z}_t$.

These examples suggest a generalized definition of the coefficient of determination, which is more flexible and may be adapted to applied problems like the above problems A., B., C. We consider for each $t \in \mathcal{T}$ a function $z_t : R^{n_t} \to \mathcal{Z}$ with values in a set $\mathcal{Z}$ and a loss function $L_t : R^{n_t} \times \mathcal{Z} \to R^1$. A predictor is a function $g^s : \mathcal{T} \times R^{s \cdot n_t} \to Z$. The prediction of $z_t = z_t(y_t)$ using the vector x_t^s of values of

the explanatory variables $X_1, \ldots, X_s$ for all individuals $i \in \mathcal{N}_t$ at the moment t is then $\hat{z}_t^s = g^s(t, x_t^s)$. The average prediction power over the time interval $\mathcal{T}$ will be described by the risk

$$R_s(g) = T^{-1} \sum_{t \in \mathcal{T}} \mathrm{E} L_t(Y_t, g(t, X_t^s)) \tag{4.15}$$

and the corresponding generalized coefficients of determination would be defined by (4.1) and (4.4). Here g^s is a function minimizing $R_s(g)$ over all predictors g and $g^0 : \mathcal{T} \to R^1$ a constant function minimizing $R_s(g)$ over all constants $g \in R^1$.
In the special above mentioned cases we obtain the following formulae for the coefficients of determination:

Case A (with loss function (4.10):

It is easy to see, that the optimal predictor g^s is given by

$$g^s(t, x_t^s) = (g_{t1}^s, \ldots, g_{tn_t}^s) \tag{4.16}$$

and

$$g_{ti}^s = \text{ sign } \{\mathrm{E}(Y_{ti} \mid X_{ti}^s = x_{ti}^s)\} \tag{4.17}$$

while g^0 is given by

$$g^0 := \text{ sign } \{n^{-1} \sum_{t,i} \mathrm{E} Y_{ti}\}. \tag{4.18}$$

Thus the estimated coefficients of determination will be

$$\hat{B}_s(X^s) := \frac{\sum_{t,i}(\ (\hat{\hat{f}}_{ti}^s)^+ - \overline{y}^+)}{\sum_{t,i}(y_{ti}^+ - \overline{y}^+)}. \tag{4.19}$$

Case B.

It is easy to see and well known from discriminant analysis, that the optimal predictor is given by (4.16) and

$$g_{ti}^s = \min\{w_0 \in W = \{1, 2, 3, 4\} \mid \ell_{ti}^s(w_0 \mid x_{ti}^s) = \min_{w \in W} \ell_{ti}^s(w \mid x_{ti}^s)\}, \tag{4.20}$$

where

$$\ell^s_{ti}(w \mid x^s_{ti}) = \sum_{v \neq w} c(v,w) P(Y_{ti} \in \mathcal{Y}_v \mid X^s_{ti} = x^s_{ti}). \tag{4.21}$$

The trivial predictor g^0 is determined by

$$g^0 := \min\{w_0 \in \mathcal{W} \mid \ell(w_0) = \min_{w \in \mathcal{W}} \ell(w)\}, \tag{4.22}$$

$$\ell(w) := \sum_{v \in \mathcal{W}} c(v,w) n^{-1} \sum_{t,i} P_{ti}(Y_{ti} \in \mathcal{Y}_v). \tag{4.23}$$

To obtain estimated coefficients of determination it is necessary to estimate the probabilities in (4.21) and (4.23). This is an estimation problem, which is parallel to our problem of estimating the regression function and deserves a separate treatment, to which we will devote a forthcoming paper. With such estimates we would obtain estimates $\widehat{\ell^s}(w|x)$ of $\ell^s(w|x), \hat{\ell}(\omega)$ of $\ell(\omega)$ and of the coefficient of determination

$$\widehat{B_s}(X^s) := \frac{\sum_{ti}(\min_w \hat{\ell}(w) - \min_w \hat{\ell}^s_{ti}(w))}{\sum_{ti}(\min_w \hat{\ell}(w))}. \tag{4.24}$$

It is obvious, that more generally $\mathcal{W}$ may be any finite set corresponding to a partition $R^1 = \sum_{w \in \mathcal{W}} y_w$. In the special case of a simple 0-1-loss $c(v,w) = \delta_{vw}$ (δ_{vw} : Kronecker symbol; $v, w \in \mathcal{W}$) the estimated coefficient of determination would be

$$\hat{B}_s(X^s) = \frac{\sum_{t,i}[\max_w \hat{P}^s_{ti}(w) - \max_w \hat{P}(w)]}{\sum_{t,i}[1 - \max_w \hat{P}(w)]} \tag{4.25}$$

where $\hat{P}^s_{ti}(w)$ and $\hat{P}(w)$ are the above mentioned estimates of the probabilities $P(Y_{ti} \in \mathcal{Y}_w \mid X^s_{ti} = x^s_{ti})$ and $n^{-1} \sum_{t,i} P(Y_{ti} \in \mathcal{Y}_w)$ resp.

Case C.

Here it the optimal predictor g^s is given by (4.16) and (see(4.3))

$$g^s_{ti} = \min\{i_0 \in \mathcal{N}_t \mid \mu^s_{ti_0} = \max_{i \in \mathcal{N}_t} \mu^s_{ti}\}, \tag{4.26}$$

while the trivial predictor g^0 is given by $\overline{y}_{.i} = T^{-1} \sum_t y_{ti}$ and

$$g^0 := \min\{i_0 \mid \mathrm{E}\overline{Y}_{.i_0} = \max_i \mathrm{E}\overline{Y}_{.i}\}. \tag{4.27}$$

The estimated coefficient of determination will then be

$$\hat{B}_s(X^s) := \frac{\sum_t (\max_i \hat{\hat{f}}^s_{ti} - \max_i \overline{y}_{.i})}{\sum_t (\max_i y_{ti} - \max_i \overline{y}_{.i})}. \tag{4.28}$$

5 Estimating the individual effects: an adaptive semiparametric alternative to parametric panel data analysis

The estimation procedures of the sections 2. and 3. were based on the at least approximatively valid assumption of a regression function $f_t(x_{ti})$ depending on the individual i only trough the values of the explanatory variables. Without this restriction are would have a regression function $f(t, i, x_{ti})$ depending on time, the individual i and the values x_{ti} of the explanatory variables. The literature on the statistical analysis of panel data offers some procedures for regression functions of this type, although under other relatively restrictive assumptions. An example would be a time independent pseudo-linear fixed effect panel model

$$f(i, x) = \sum_{j=1}^{P} \beta_j g_j(x) + \nu_i \quad (t \in \mathcal{T}, i \in \mathcal{N}) \tag{5.1}$$

with the restriction $\sum_i \nu_i = 0$ (see Hsiao 1986). Here and in the following we will assume $\mathcal{N}_t \equiv \mathcal{N}$ and $n_t = n_1$, that is the same individuals at every moment $t \in \mathcal{T}$. The estimation of the individual effects ν_i makes sense only if the time interval $\mathcal{T}$ is sufficiently large compared with the number n_1 of individuals. This is often not the case and even more, if in reality the parameters β_j, ν_i change sometimes or even continuously with the time t. The same comments apply to more general semiparametric models of the type

$$f(t, i, x) = f_t[x] + \nu_{ti}, \tag{5.2}$$

where f_t itself may possibly follow some parametric or semiparametric model. The second term ν_{ti} describes the individual effect which are not produced solely by the

explanatory variables and which may possibly vary with the time t.

In this section we propose a procedure for the estimation of regression functions (5.2) under the assumptions

$$\mathrm{E}y_{ti} = f(t, i, x_{ti}) := \mu_{ti} \quad , \quad DY_{ti} = \sigma^2$$

, which is based on

(i) the approach of Bunke and Castell (1998) developed for regression on qualitative variables (for estimating ν_{ti} in (5.2)) combined with

(ii) our approach of section 2. (for estimating the function f_t in (5.2)).

The approach has an analoguous heuristic background as that of section 2. We use a partition

$$\pi : \mathcal{N} = \sum_{j=1}^{q_\pi} \mathcal{N}_j^\pi \tag{5.3}$$

of the set $\mathcal{N}$ of individuals, which for the sake of single presentation is assumed to be independent of the time t. The use of time dependent partitions π_t presents no additional difficulties and allows possibly a higher accuracy. For each $i \in \mathcal{N}$ there is a $j = j(i, \pi)$ with $i \in \mathcal{N}_j^\pi$. The individuals $\underline{i} \in \mathcal{N}_j^\pi$ may sometimes be characterized as "neigbours" of i, but in general this will be only a formal characterization, especially when i is a purely nominal index and therefore there is no sensible definition of a "distance" between individuals.

The formal assumptions of identical individual effects $\nu_{\underline{t},\underline{i}}$ for fixed t, i (in analogy to the approach of Bunke and Castell 1998) and

$$\underline{t} \in \mathcal{T}_t^h := [\underline{t}_h, \overline{t}^h] \quad , \quad \underline{i} \in \mathcal{N}_{j(i,\pi)}^\pi \tag{5.4}$$

and of the identifiability condition

$$\sum_{i \in \mathcal{N}} \nu_{ti} = 0 \qquad (t \in \mathcal{T}) \tag{5.5}$$

lead for each fixed $t \in \mathcal{T}$ to a linear model $\mathcal{L}_t^\pi$ for the vector ν_t of individual effects ν_{ti} $(i \in \mathcal{N})$ obeying $\nu_{t\underline{i}} = \nu_{ti}$ for $j(\underline{i}, \pi) = j(i, \pi)$. The vector ν_t is dermined by the vector $\eta_t^\pi = (\eta_{t1}^\pi, \ldots, \eta_{tq_\pi}^\pi)$ given by

$$\eta_{tj}^\pi = \nu_{ti} \quad \text{for} \quad i \in \mathcal{N}_j^\pi. \tag{5.6}$$

The linear model $\mathcal{L}_t$ may be written in the form

$$(5.7) \qquad \mathcal{L}_t^\pi = \{\nu_t \in R^{n_1} \mid \nu_{ti} = (a_{ti}^\pi)'\eta_t^\pi \ , \ i \in \mathcal{N}\}$$

with some fixed vectors $a_{t,i}^\pi$.

For fixed t a "local" estimator of the terms $\nu_{\underline{t}i}(\underline{t} \in \mathcal{T}_t^h \ , \ i \in \mathcal{N})$ in (5.2) based on the assumption of identical individual effects $\nu_{\underline{t},\underline{i}}$ under (5.4) will intuitively be a good estimate, even if assumption of "identical effects" is replaced by "at most moderately differing effects". The idea of our procedure is to find a partition and a horizon h hopefully leading to such a situation for all $t \in \mathcal{T}$. The tool for this is to minimize a cross-validation criterion over a set $\prod_0$ of partitions. $\prod_0$ may be the set of all partitions of $\mathcal{N}$ or some subset of especially interesting partitions. For instance some of the subsets $\mathcal{N}_j^\pi$ may be fixed in advance due to preliminary knowledge in the field of application. In the example of the stock market, there will be groups of firms i with similar production profile and economic background, so that their individual effects may assumed to be similar.

Our approach consists in choosing simultaneously the horizon $h \in \mathcal{H}$, a partition $\pi \in \prod_0$ and an estimator from a class $\{\hat{f}^{M,h,\pi} \mid M \in \mathcal{M}\}$ of estimators by minimization of the cross-validation criterion

$$(5.8) \qquad C(M,h,\pi) := \frac{1}{n} \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{N}} \mid y_{ti} - \tilde{\mu}_{ti}^{M,h,\pi} \mid^2.$$

$\tilde{\mu}_{ti}^{M,h,\pi}$ denotes the "local" estimate of μ_{ti} calculated under (5.2) and (5.4) with the estimator $\hat{f}^{M,h,\pi}$, but leaving out the observation (x_{ti}, y_{ti}).

The estimators $\hat{f}^{M.h,\pi}$ are constructed in the following way :

Case 1: $M \in \mathcal{M}_1$

In this case we assume for fixed $\pi \in \prod_0, t \in \mathcal{T}$ and for moments $\underline{t} \in \mathcal{T}_t^h$ and individuals $i \in \mathcal{N}$ a parametric model

$$(5.9) \qquad f^{M,h,\pi,t}(\underline{t}, i, x) = g_M(x \mid b_t) + (a_{\underline{t}i}^\pi)'\eta_t^\pi.$$

with parameters b_t, η_t^π. Fitting by ordinary least squares to the observations $x_{\underline{t}}, y_{\underline{t}} \quad \underline{t} \in \mathcal{T}_t^h$ leads to the estimates $\hat{b}_t^{M,\pi,h}, \hat{\eta}_t^{M,\pi,h}$ and to the estimates

$$(5.10) \qquad \hat{f}^{M,h,\pi,t}(t,i,x) := g_M(x \mid \hat{b}_t^{M,\pi,h}) + (a_{ti}^\pi)'\hat{\eta}_t^{M,\pi,h}$$

of $f(t,i,x)$. The estimates (5.10) may be interpreted as "running semiparametric estimates" w.r.t. the variable t and as "regressograms" w.r.t. the variable i. The models $g_M \quad (M \in \mathcal{M}_1)$ are described in section 2.
The estimates of the individual effects ν_{ti} are given by $\hat{\nu}_{ti}^{M,h,\pi} := (a_{ti}^{\pi})' \hat{\eta}_t^{\pi,h}$.

A triple $(\hat{M}, \hat{h}, \hat{\pi}) \in \mathcal{M}_1 \times \mathcal{H} \times \prod_0$ minimizing (5.6) finally leads to the "adaptive estimate" $\hat{\hat{f}}$ of f given by

$$\hat{\hat{f}}(t,i,x) := \hat{f}^{\hat{M},\hat{h},\hat{\pi},t}(t,i,x).$$

Case 2: $M \in \mathcal{M}_s \quad (s = 2,3,4,5)$.

In these cases we use adaptive partially parametric ($s = 2,4,5$) or semi-parametric ($s = 3$) estimates of $f(t,i,x)$ of the form

$$\hat{\hat{f}}(t,i.x) = \hat{\hat{f}}_t[x] + \hat{\nu}_{ti}, \quad \hat{\nu}_{ti} = (a_{ti}^{\pi})' \hat{\eta}_t. \tag{5.11}$$

The estimates $\hat{\hat{f}}_t$ will be determined in an analoguous way as described in section 2 for $\hat{\hat{f}}$, but taking into consideration the parametric forms given by the individual effects and adapting simultaneously M, h and the partition π as in the case 1.
For instance, when $s = 2$ the estimate may be constructed in the following way:
For fixed $h \in \mathcal{H}, M \in \mathcal{M}_2, \pi \in \prod_0, t \in \mathcal{T}, \eta_t \in R^{q_\pi}$ we take the "1st stage transformed kernel estimate"

$$f_t^{M,h,\pi,\eta_t}[x] := T_0^{-1}\Big\{ \sum_{\tau=\underline{t}_h}^{\overline{t}_h} \sum_{i \in \mathcal{N}_\tau} \widetilde{K}_x^M(\tau,i)\, T_0[y_{\tau i} - (a_{\tau i}^{\pi})' \eta_t] \Big\} \tag{5.12}$$

of the first term in (5.2) corresponding to the time window $\underline{t} \in \mathcal{T}_t^h$. In (5.12) we use the same notation as in (2.19). Least squares fitting leads to the estimates

$$\hat{f}^{M,h,\pi}(t,i,x) := f^{M,h,\pi,\hat{\eta}_t^{M,h,\pi}}[x] + (a_{ti}^{\pi})' \hat{\eta}_t^{M,h,\pi} \tag{5.13}$$

where $\eta_t = \hat{\eta}_t^{M,h,\pi}$ minimizes the sum of squares

$$S_{t,M,h,\pi}(\eta_t) := \sum_{\tau=\underline{t}_h}^{\overline{t}_h} \sum_{i \in \mathcal{N}_\tau} | y_{\tau i} - \hat{f}_t^{M,h,\pi,\eta_t}[x_{\tau i}] - (a_{\tau i}^{\pi})' \eta_t |^2. \tag{5.14}$$

The adaptive estimates (5.11) are given by

$$(5.15) \qquad \hat{\hat{f}}_t[x] := f_t^{\hat{M},\hat{h},\hat{\pi},\hat{\hat{\eta}}_t}[x], \quad \hat{\hat{\eta}}_t := \hat{\eta}_t^{\hat{M},\hat{h},\hat{\pi}},$$

where $(\hat{M}, \hat{h}, \hat{\pi})$ minimizes the corresponding cross-validation (5.8).

Remark: The selection of an optimal partition $\hat{\pi}$ may require excessive computational effort, if the number n of individuals is large, because then the number $B(n)$ of all partitions is very large

$$(5.16) \qquad B(n) = \sum_{q=1}^{n} \sum_{j=1}^{q} (-1)^{q-1} j^n \,/\, j!\,(q-j)!$$

(see e.g. Stanley (1997)). Practical devices in this situation are to choose a sensible but only moderately large class $\prod_0$ of partitions or alternatively to select only a "suboptimal" partition $\hat{\pi}$ in $\prod_0$ by a stepwise procedure, as for instance described in Bunke and Castell (1998).

References:

Baltagi, B.H. (1996). *Econometric Analysis of Panel Data*, Wiley, Chichester

Bunke, O. (1984). Selecting variables and models in regression analysis. Some tools and suggestions for a strategy. Preprint Nr. 68. Humboldt University, Department of Mathematics, Berlin

Bunke, O. (1992). Semiparametric modelling for a variable depending on time and explanatory variables. In P.G.M. van der Heyden et al. (editors), *Statistical Modelling,* Verlag J. Eul, Köln, 115-126

Bunke, O., Droge, B. and Polzehl, J. (1995). Model selection, transformations and variance estimation in nonlinear regression. *Discussion Paper* **52**, Sonderforschungsbereich 373, Humboldt University, Berlin

Bunke, O. and Castell, E. (1998). Regression and contrast estimates based on adaptive regressograms depending on qualitative explanatory variables, *Discussion Paper* **??**, Sonderforschungsbereich 373, Humboldt University, Berlin

Bunke, O., Sommerfeld, V. and Stehle, R. (1997). Semiparametric modelling of the cross-section of expected stock returns. *Discussion Paper* **95**, Sonderforschungsbereich 373, Humboldt University, Berlin

Droge, B. (1992). On a computer programme for the selection of variables and models in regression analysis. In: Model Oriented Data-Analysis (Eds. V. Fedorov, W. G. Müller and I. N. Vuchkov), Physica, Heidelberg, 181-192.

Engle, R.F., Granger, C.W.J., Rice, T. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81**, 310-320

Fama, E.F. and French, R.F. (1995). Size and book-to-market factors in earnings and returns. *Journal of Finance* **50**, 131-155

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press. Cambridge

Hastie, T.T. and Tibshirani, R.J. (1990). *Generalized additive models.* Chapman & Hall, London

Hsiao, C. (1986). *Analysis of Panel Data*, Cambridge University Press, Cambridge

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Econometrica, 61, 387-421.

Mai, K. and Polzehl, J. (1991). Regression based short term prediction of electrical load for a power system. Technical report, Humboldt University, Department of Mathematics

Rao, R.C. (1965). *Linear Statistical Inference*, Wiley, New York

Stanley, R.D. (1997). *Enumerative Combinatorics*, Cambridge University Press, Cambridge

White, H. (1989). Learning in artificial neural networks: a statistical perspective. Neural Computation **1**, 425-464.

Aknowledgement:

The author is grateful to V. Sommerfeld for comments which lead to valuable improvements of this paper.