Adaptive Nonstationary Regression Analysis

O. Krasotkina
Tula State University
Tula, Russia
ko180177@yandex.ru

V. Mottl
Computing Center of the Russian Academy of Sciences
Moscow, Russia
vmottl@yandex.ru

Abstract

The problem of finding the most appropriate subset of features or regressors is the generic challenge of Machine Learning problems like regression estimation or pattern recognition. We consider the problem of time-varying regression estimation, which implies also the inevitable necessity to choose the individual appropriate levels of model volatility in each of regressors, ranging from the full stationarity of instant models to their absolute independence in time. The problem is considered from the Bayesian point of view as that of estimating the sequence of regression coefficients associated with the hidden vector state of a stochastic linear dynamic system, whose a priori model includes parameters responsible for both the size of the subset of active regressors and the time-volatility factors of regression coefficients at them. The proposed technique of adaptive time varying regression estimation is built as that of estimating both the state and parameters of the hidden state-space model.

1. Introduction

The most challenging aspect of the problem of regression estimation $y_t : T \to \mathbb{R}$ in some set of observations $t \in T$ is, perhaps, the choice of an appropriate subset $\hat{I} \subseteq I$ of the available set of features $\{x_t^i, i \in I\}$ [1]: $y_t \cong f(x_t^{(i)}, i = 1,...,n)$. The kernel-based approach to estimating dependences, which embeds a set of entities of arbitrary kind into a hypothetical linear space, wipes out the difference not only between numerical features and more complex modalities of object representation, but also between linear and nonlinear models [2], so that it becomes enough to consider only former of them:

$$y_t = \sum_{i=1}^{n} \beta^{(i)} x_t^{(i)} + e_t = \mathbf{x}_t^T \boldsymbol{\beta} + e_t , \qquad (1)$$

However, in many applications, the set of observations $t \in T$ has to be treated rather as a succession $T = \{1,...,N\}$ than a plain set. In most cases, the succession of observations is associated with discrete time, when the stationary regression model (1) turns out to be insufficient. Therefore, the problem of estimating a time-varying regression model

$$y_t = \sum_{i=1}^{n} \beta_t^{(i)} x_t^{(i)} + e_t = \mathbf{x}_t^T \boldsymbol{\beta}_t + e_t , \qquad (2)$$

in which it is required to estimate the succession of regression coefficients $B = (\boldsymbol{\beta}_t, t = 1,...,N)$, has been subject of intensive study in statistical literature during, at least, the recent fifteen years [3,4].

The number of variables in model (2) to be estimated ever exceeds the number of observations. Thus, it is impossible to estimate the time-varying regression model without additional regularization, namely, without taking some a priori assumptions on the hidden sequence of regression coefficients. The way of regularization is suggested by practice.

First, it is typical for the majority of practical situations that the initially assumed set of regressors is much greater than the actual set, so that the majority of regression coefficients are equal to zero. So, the challenge of regressor selection remains actual for the nonstationary regression model, and, for this reason alone, the problem of time-varying regression estimation falls into the competence area of Machine Learning.

Second, it is typical that only a small number of regression coefficients are changing in time, whereas the majority of them remain constant. If we new the names of regressors having constant coefficients, we could drastically reduce the actual number of variables to be estimated. Finally, the hidden regression coefficients are changing sufficiently smoothly in time. If the assumed degree of this smoothness is sufficiently high, the effective freedom of search will be essentially reduced.

In this paper, we propose a procedure that automatically estimates the subset of active regressors, finds, among them, regressors with actually changing coefficients, and estimates the individual time-volatility levels for these coefficients.

2. Flexible Least Squares criterion of time-varying regression estimation

The problem of nonstationary regression estimation (2) has been intensively studied in the literature. The standard means of estimating time-varying models of this

This work is supported by Russian Fund of Basic Research, pr. 06-01-00412

kind is the Flexible Least Squares approach (FLS) first introduced in [3]:

$$J(\beta_t^{(i)}, t=1,...,N, i\in\hat{I}\,|\,\hat{I},\rho)=\sum_{i=1}^{N}\left(y_t-\sum_{i\in\hat{I}}\beta_t^{(i)}x_t^{(i)}\right)^2+\rho\sum_{t=2}^{N}\sum_{i\in\hat{I}}(\beta_t^{(i)}-\beta_{t-1}^{(i)})^2\to\min. \quad (3)$$

The subset of regressors $\hat{I}\subseteq I$ and coefficient ρ are the parameters of this criterion. The first term $\sum_{t=1}^{N}\left(y_t-\sum_{i=1}^{n}\beta_t^{(i)}x_t^{(i)}\right)^2$ of the FLS criterion (3) stands for the approximation of the observations y_t. As to the second term $\rho\sum_{i=1}^{n}\sum_{t=1}^{N}(\beta_t^{(i)}-\beta_{t-1}^{(i)})^2$, it is responsible for the oveall time-volatility of regression coefficients. The greater value of $\rho>0$ the "smoother" are the estimated asset weights $\boldsymbol{\beta}_t$ in time and the smaller is the actual "dimension" of the problem. If $\rho\to\infty$, the criterion (3) turns into the plain least squares method $\hat{\beta}_1^i=...=\hat{\beta}_N^i$.

The key element of the FLS criterion is treating the regression coefficients as independent hidden processes assumed *a priori* to possess the Markov property $\beta_t^{(i)}=\beta_{t-1}^{(i)}+\xi_t^{(i)}$, where $\xi_t^{(i)}$ are independent normal white noises with zero mathematical expectations $E(\xi_t^{(i)})=0$ and variances $E\left((\xi_t^{(i)})^2\right)=\rho E\left(e_t^2\right)$.

Minimization of criterion (3) is equivalent to solving the system of linear equation of very high dimension $N\,|\,\hat{I}\,|$. The FLS criterion (3) is a quadratic function of the variables $(\boldsymbol{\beta}_t, t=1,...,N)$ with block-tridiagional matrix, so, finding its minimum point is provided by the double-sweep method applied to the respective system of linear equations. The resulting algorithm is completely equivalent, on the one hand, to a continuous version of the dynamic programming procedure, and, on the other hand, to the Kalman-Bucy filter-smoother [3,5]. These three equivalent procedures have the linear computational complexity with respect to the number of vector variables, what is of great importance for time series analysis, because the length of the observable time series $(Y,X)=((y_t,\mathbf{x}_t), t=1,...,N)$ is often not fixed.

As a rule, it is impossible to choose the appropriate subset $\hat{I}\subset I$ and the appropriate value of the smoothness parameter λ a priori.

3. Adaptive Flexible Least Squares criterion

In this paper, we propose a modification of the FLS criterion (3) which we call the adaptive FLS. This adaptive criterion allow first, to automatically choose the subset of best regressors, second, select among them the subset of regressors with really changing coefficients, and, third, determine the volatility parameter ρ for the changing regression coefficients.

Let $(\mathbf{x}_t, t=1,...,N)$, $\mathbf{x}_t=(x_t^{(i)}, i=1,...,n)$, be a given sequence of regressors not subject to any probabilistic modeling. We consider the time series to be processed $(y_t, t=1,...,N)$ (2) as the observable part of a two-component random process, whose hidden part is the unknown sequence of time-varying regression coefficients $\left(\boldsymbol{\beta}_t=(\beta_t^{(i)}, i=1,...,n), t=0,1,...,N\right)$. The main point of the regressors selection technique we propose here is the a priori probabilistic model of the hidden random sequence of regression coefficients $\boldsymbol{\beta}_t=(\beta_t^{(i)}, i=1,...,n)$. First, its components are considered as a priori independent. Second, the values of the regression coefficients are formed by the identical autoregression models

$$\beta_t^{(i)}=\frac{\delta^{(i)}}{\delta^{(i)}+\lambda^{(i)}}\beta_{t-1}^{(i)}+\xi_t^{(i)},$$

where $\xi_t^{(i)}$ are independent normal white noises with zero mathematical expectations $E(\xi_t^{(i)})=0$ and variances $E\left((\xi_t^{(i)})^2\right)=\frac{\delta^{(i)}\lambda^{(i)}}{\delta^{(i)}+\lambda^{(i)}}$.

The auxiliary variables $\delta^{(i)}\ge 0$ and $\lambda^{(i)}\ge 0$ perform the function of adaptation. If $\lambda^{(i)}\to 0$ then $E\left((\xi_t^{(i)})^2\right)\to 0$, and the sequence of regression coefficients at the i-th regressor will remain constant $\beta_t^i=\beta_{t-1}^i$ with some a-priori unknown value. If $\delta^{(i)}\to 0$ then $E\left((\xi_t^{(i)})^2\right)\to 0$ together with $\delta^{(i)}/(\delta^{(i)}+\lambda^{(i)})\to 0$, and the i-th sequence of regression coefficients turns into the zero constant. The nonzero variables $\delta^{(i)}$ form the set of relevant regressors $\hat{I}=\{i:\delta_i>0\}$, and the nonzero variables $\lambda^{(i)}$ extract from them the subset of nonstationary regression coefficients $\hat{I}_{\text{var}}=\{i,\lambda^{(i)}\ge 0\}\subseteq I$. The product of noise variances $E(\xi_t^{(i)})^2$ at all the regressors defines the volume of the concentration ellipsoid for the random vector $(\beta_t^{(i)}, i\in I)$. If this volume tends to zero, the random deviations of all the regression coefficients from each other and from zero are decreasing.

The variables $\lambda^{(i)}$ и $\delta^{(i)}$ control only the ratio between time volatility levels of different regression coefficients but do not affect the general time-volatility level of the model. This fact leads to the necessity to fix the volume of the concentration ellipsoid $\prod_{i\in I}\frac{\delta^{(i)}\lambda^{(i)}}{\delta^{(i)}+\lambda^{(i)}}=1$. The general time volatility is specified by the observation noise variance in the model of nonstationary regression (2) $E(e_t)=0$, $E\left((e_t)^2\right)=\rho$.

So, we have defined, first, the conditional a priori distribution of the hidden sequence of regression coefficients $\Psi(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_N\,|\,\delta^{(i)},\lambda^{(i)}, i\in I)$ and, second, the conditional distribution of the observed time series

$\Phi(y_1,...,y_N \mid \boldsymbol{\beta}_1,...,\boldsymbol{\beta}_N,\rho)$. It is clear that the a posteriori joint distribution density of the hidden vector sequence of regression coefficients and variances of their components related to single regressors will be proportional to the product

$$P(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_N,\delta^{(i)},\lambda^{(i)},\ i\in I \mid y_1,...,y_N,\rho) \propto$$
$$\Phi(y_1,...,y_N \mid \boldsymbol{\beta}_1,...,\boldsymbol{\beta}_N,\rho)\Psi(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_N \mid \delta^{(i)},\lambda^{(i)},\ i\in I).$$

It appears natural to take the maximum point of this a posteriori density as the estimate of the sequence of time-varying regression coefficients along with the variances indicating participation of each of regressors in the model:

$$(\hat{\boldsymbol{\beta}}_0,\hat{\boldsymbol{\beta}}_1,...,\hat{\boldsymbol{\beta}}_N,\hat{\delta}^{(i)},\hat{\lambda}^{(i)},\ i\in I \mid \rho) = \arg\max P(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_N,\delta^{(i)},\lambda^{(i)},\ i\in I \mid y_1,...,y_N,\rho). \quad (4)$$

Theorem. The maximum point of the a posteriori density (4) by $(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_N,\delta^{(i)},\lambda^{(i)},\ i\in I)$ is the minimum point of the criterion

$$J(\beta_t^{(i)},t=1,...,N,\delta_i,\lambda_i,i\in I \mid \rho) = \sum_{i=1}^{N}\left(y_t - \sum_{i\in I}\beta_t^{(i)}x_t^{(i)}\right)^2 + \rho\sum_{t=2}^{N}\sum_{i\in I}\frac{\delta^{(i)}+\lambda^{(i)}}{\delta^{(i)}\lambda^{(i)}}\left(\beta_t^{(i)} - \frac{\delta^{(i)}}{\delta^{(i)}+\lambda^{(i)}}\beta_{t-1}^{(i)}\right)^2 \to \min, \quad (5)$$
$$\prod_{i\in I}\frac{\delta^{(i)}\lambda^{(i)}}{\delta^{(i)}+\lambda^{(i)}} = 1.$$

In contrast to (3), the adaptive FLS is applied to the whole set of regressors I.

As we see, if the parameters $(\delta^{(i)},\lambda^{(i)},\ i\in I)$ are fixed, the resulting criterion practically coincides with the FLS criterion (3).

However, the presence of the additional variables $(\delta^{(i)},\lambda^{(i)},\ i\in I)$ is extremely important. If some $\delta^{(i)}\to 0$, the criterion drastically penalizes the deflection of the entire sequence of the respective time-varying regression coefficient $(\beta_0^{(i)},\beta_1^{(i)},...,\beta_N^{(i)})$ from zero, and practically excludes the i th regressor from the model. If some $\lambda^{(i)}\to 0$, the neighboring values of the hidden process are practically equal to each other, and the i th regression coefficient is almost constant in time.

4. The iterative minimization procedure

For finding the minimum point of the modified FLS criterion (5) with the fixed structural parameter ρ, we apply the Gauss-Seidel iteration to both groups of variables $(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_N)$ and $(\delta^{(i)},\lambda^{(i)},\ i\in I)$ starting with the initial values $(\delta^{(i),0},\lambda^{(i),0},i=1,...,n)$.

At each iteration, the current approximations $(\delta^{(i),k},\lambda^{(i),k},i\in I)$ turn (5) into the usual FLS criterion (3) with respect to the regression coefficients $(\boldsymbol{\beta}_0^k,\boldsymbol{\beta}_1^k,...,\boldsymbol{\beta}_N^k)$, which can be easily minimized by the standard Kalman-Bucy filter and smoother [6]. Once the regression coefficients are found, the next values of the variances $(\delta^{(i),k},\lambda^{(i),k},i\in I)$ are defined by the following expression which is easy to prove:

$$\delta^{(i).k+1} = \frac{a^{(i).k+1}}{1-a^{(i).k+1}}\lambda^{(i).k+1},$$
$$\lambda^{(i).k+1} = \frac{1}{a^{(i).k+1}}\frac{\sum_{t=2}^{N}(\beta_t^{(i),k} - a^{(i).k+1}\beta_{t-1}^{(i),k})^2}{\left[\prod_{l\in I}\sum_{t=2}^{N}(\beta_t^{(l),k} - a^{(l).k+1}\beta_{t-1}^{(l),k})^2\right]},$$

where

$$0 < a^{(i).k+1} = (a_0, \frac{\sum_{t=2}^{N}\beta_{t-1}^{(i),k}\beta_t^{(i),k}}{\sum_{t=2}^{N}(\beta_{t-1}^{(i),k})^2}, a_1) < 1,$$
$$0 \approx a_0 < a_1 \approx 1.$$

These two steps are repeated until the iterative process converges. It takes usually not more than ten iterations.

The important property of proposed procedure is quickly tending to zero values $\delta^{(i),k}\to\hat{\delta}^{(i)}\approx 0$ the majority of parameters $\delta^{(i),k}$. As a result, the estimation algorithm suppresses redundant regressors. In addition, the majority of coefficients $\lambda^{(i),k}$ tend to zero values, too, $\lambda^{(i),k}\to\hat{\lambda}^{(i)}\approx 0$. This means that the estimation algorithm tries to interpret as many regression coefficients as constant in time, and to explain the output variable by a few number of remaining regression coefficients which are actually time-varying.

For finding the appropriate values of the structural parameter ρ, the leave-one-out cross-validation technique described in [6] is to be applied.

5. Case study: Reverse-engineering of an investment portfolio

In this section, we discuss a practical problem which is concerned with the necessity of intelligent time-varying regression estimation which includes, first, regressor selection and, then, further selection of regressor having nonstationary coefficients. This is the problem of recovering, from publicly available data, the time-varying structure of an investment portfolio, which is usually strongly hidden from public. A typical example of an investment company is a hedge fund, which accumulates money of common people with the purpose of saving it from devaluation.

Let the capital of an investment company be fully invested in n financial instruments or securities (stocks, bonds, currencies, etc.) in proportions denoted by coefficients, which are changing in time and are unknown. These coefficients are just the subject of public interest. What is observed is only the periodic information (daily, weekly, quarterly, monthly) on the so-called return of the investment company, namely, the sequence of relative changes in the portfolio's monetary volume. Each company is obliged to report this information to governmental institutions, whereas the volume itself remains secret.

It can be proved that, under some additional financial assumptions [7], the sequence of periodic returns of an investment company is linear combination of daily returns

of assets in which the capital is invested. The time-varying coefficients of this linear combination have the sense of proportions of capital sharing between the kinds of assets. So, we come to the problem of time-varying regression estimation.

As a rule, the set of assets in which the capital is actually invested is essentially smaller than the set of all potential assets. So, it is strongly required to find the subset of regressors with non-zero regression coefficients. In addition, managing an investment portfolio is usually performed through trading a few number of assets, whereas the other kinds of assets remain untraded. Thus, the problem of finding the time-varying regression coefficients in contrast to time-constant ones is of extreme importance here.

In this Section, we present an example of application of the proposed methodology to a real-life portfolio, namely, Long Term Capital Management (LTCM).

The collapse of this hedge fund in 1998 is by far the most dramatic hedge fund story to date. LTCM's troubles began in May-June 1998 [8]. By the end of September 1998, a month after the Russian crisis, the fund had lost 92% of its December 1997 assets. Applying Adaptive Nonstationary Analysis to monthly returns of the fund, we attempted to determine major factors explaining the fall of LTCM. For our analysis, we used the returns of assets classes in which the LTCM's capital can be invested. These returns values are provided by Lehman Brothers and Merrill Lynch.

We varied the smoothness parameter ρ in the interval from the minimum value $\rho = 10^{-8} \cong 0$, which is equivalent to absolute independence of instant models, to the maximum value $\rho = 10^{8}$ providing full stationarity. For each value of ρ, we applied the iterative procedure proposed in section 4 to the given time series. Application of the leave-one-out cross-validation technique of finding the optimal smoothness parameter ρ gave the value $\rho = 10$.

The respective result is shown in Figure 1. Asset exposures $(\beta_t^{(1)}, ..., \beta_t^{(n)})$ for each time period are "stacked" along the vertical axis with respect to the sign. The negative positions correspond to hedge or, in other words, debt capital. What is especially important here is the fact that the proposed algorithm suppressed the redundant assets and, as it is well seen, only 8 of 18 initial regressors occur in the final model.

Another important aspect is the fact that 7 of 8 remaining assets weights are estimated as almost completely stationary. The only weight recognized as time-varying corresponds to the fund's capital share invested in ML Emerging Bonds, the government bonds from developing financial markets, including Russian short-term treasury government bonds (the so-called ГКО in the Russian abbreviation). So, our analysis suggests the hypothesis that the crash of LTCM could be due just to the price drop of the Russian government bonds in July-August 1998.

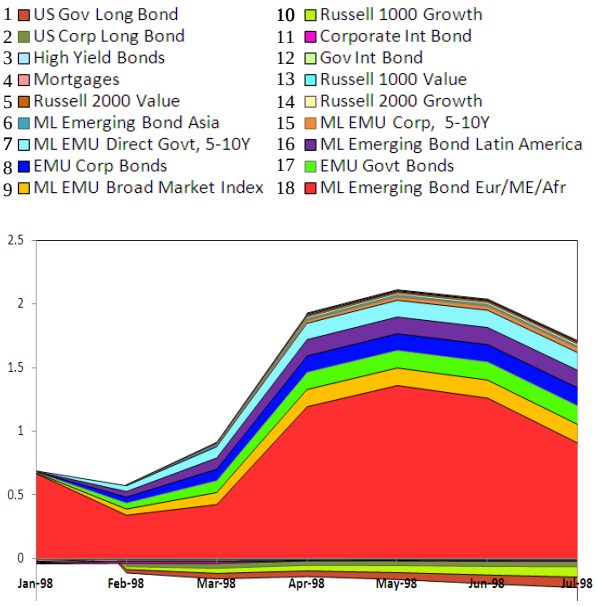

Figure 1. Adaptive nonstationary analysis of the LTCM fund.

6. References

[1]. Jain A., Zongker D. Feature selection: Evaluation, application, and small sample performance. IEEE Trans. on Pattern Analysis and Machine Intelligence, February 1997, Vol. 19, no. 2, pp. 153-158.

[2]. V. Mottl, O. Krasotkina, O. Seredin, I Muchnik. Kernel fusion and feature selection in machine learning. Proceedings of the 8th IASTED International Conference on Intelligent Systems and Control. Cambridge, USA, October 31 – November 2, 2005.

[3]. R. Kalaba, L. Tesfatsion. Time-varying linear regression via flexible least squares. International Journal on Computers and Mathematics with Applications, 1989, Vol. 17, pp. 1215-1245.

[4]. Wells C. The Kalman Filter in Finance. Kluwer Academic Publishers, 1996.

[5]. W. Schneider. Analytical uses of Kalman filtering in econometrics – a survey. Statistical Papers, 1988, Vol. 29, pp. 3-33.

[6]. Markov M., Krasotkina O., Mottl V., Muchnik I. Time-varying regression model with unknown time-volatility for nonstationary signal analysis. Proceedings of the 8th IASTED International Conference on Signal and Image Processing. Honolulu, Hawaii, USA, August 14-16, 2006.

[7]Sharpe W.F. Asset allocation: Management style and performance measurement. *The Journal of Portfolio Management*, Winter 1992, pp. 7-19.

[8] Ph. Jorion. Risk management lessons from long-term capital management. European Financial Management, September, 2000.