International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:3, No:3, 2009

Estimation of Time-Varying Linear Regression with Unknown Time-Volatility via Continuous Generalization of the Akaike Information Criterion

Elena Ezhova, Vadim Mottl, and Olga Krasotkina

Abstract—The problem of estimating time-varying regression is inevitably concerned with the necessity to choose the appropriate level of model volatility - ranging from the full stationarity of instant regression models to their absolute independence of each other. In the stationary case the number of regression coefficients to be estimated equals that of regressors, whereas the absence of any smoothness assumptions augments the dimension of the unknown vector by the factor of the time-series length. The Akaike Information Criterion is a commonly adopted means of adjusting a model to the given data set within a succession of nested parametric model classes, but its crucial restriction is that the classes are rigidly defined by the growing integer-valued dimension of the unknown vector. To make the Kullback information maximization principle underlying the classical AIC applicable to the problem of time-varying regression estimation, we extend it onto a wider class of data models in which the dimension of the parameter is fixed, but the freedom of its values is softly constrained by a family of continuously nested a priori probability distributions.

Keywords—Time varying regression, time-volatility of regression coefficients, Akaike Information Criterion (AIC), Kullback information maximization principle.

I. Introduction

THE Akaike Information Criterion (AIC) [1] is adopted in data analysis as a simple and effective means of adjusting the most adequate model to the given data set among a discrete succession of nested parametric model classes.

Let the given data set $\mathbf{y} = (y_t, t = 1, \ldots, N)$ be considered as a sample of independent random variables with an unknown density $\varphi^*(y)$, whereas the observer assumes a parametric family $\varphi(y\,|\,\mathbf{c})$, $\mathbf{c} \in \mathbb{R}^m$. It is a typical case that the parameter dimension m is certainly too large for the "actual" density $\varphi^*(y)$ and the size N of the sample, what makes senseless the maximum-likelihood estimate

$$\hat{\mathbf{c}}(\mathbf{y}) = \arg\max \ln \Phi(\mathbf{y}\,|\,\mathbf{c}),\ \ln \Phi(\mathbf{y}\,|\,\mathbf{c}) = \sum_{t=1}^{N} \ln \varphi(y_t\,|\,\mathbf{c}). \quad (1)$$

E. Ezhova is Master Student of the Moscow Institute of Physics and Technology, Department of Intelligent Systems, Moscow, Russia,
e-mail: lena-ezhova@rambler.ru

V. Mottl is Principal Investigator at the Computing Center of the Russian Academy of Sciences, Professor of the Moscow Institute of Physics and Technology, Department of Intelligent Systems, Moscow, and Professor of the Tula State University, Department of Automation and Remote Control, Tula, Russia,
e-mail: vmottl@yandex.ru

O. Krasotkina is Associate Professor of the Tula State University, Department of Automation and Remote Control, Tula, Russia,
e-mail: ko180177@yandex.ru

Manuscript received January 31, 2009; revised March 6, 2009.

The observer's assumption is that the elements of $\mathbf{c}$ are naturally ordered by their "importance". The idea is to truncate the parameter vector $c_i = 0$, $n < i \leq m$:

$$\mathbf{c} = (\mathbf{c}_n, \mathbf{c}_{m-n}), \mathbf{c}_n \in \mathbb{R}^n, \mathbf{c}_{m-n} = \mathbf{0} \in \mathbb{R}^{m-n}. \quad (2)$$

So, the density family $\Phi(\mathbf{y}|\mathbf{c})$ turns into a succession of nested families $\Phi\big(\mathbf{y}\,|\,\mathbf{c} = (\mathbf{c}_n, \mathbf{0})\big)$, $\mathbb{R}^{n_{min}} \subset \cdots \subset \mathbb{R}^{n_{max}}$.

The classical AIC is a criterion of choosing the dimension as the most appropriate level of model complexity $\hat{n}(\mathbf{y}) = \arg\max_n \Big[\ln \Phi\Big(\mathbf{y}\,|\,\big(\hat{\mathbf{c}}_n(\mathbf{y}), \mathbf{0}\big)\Big) - n\Big]$ instead of the plain likelihood maximization (1). However, this formula was designed under the assumption that $\bigtriangledown^2_{\mathbf{c}_n\mathbf{c}_n} \ln \Phi\big(\mathbf{y}\,|\,(\mathbf{c}_n, \mathbf{0})\big)$ is a full-rank matrix at the point of the maximum likelihood, and, so, the estimate $\hat{\mathbf{c}}_n(\mathbf{y})$ is unique. To cover the most general case, the penalty n should be replaced by the rank of this matrix:

$$\hat{n}(\mathbf{y}) = \arg\max_n \Big\{ \ln \Phi\big(\mathbf{y}\,|\,\big(\hat{\mathbf{c}}_n(\mathbf{y}), \mathbf{0}\big)\big) - Rank\big[\bigtriangledown^2_{\mathbf{c}_n\mathbf{c}_n} \ln \Phi\big(\mathbf{y}\,|\,\big(\hat{\mathbf{c}}_n(\mathbf{y}), \mathbf{0}\big)\big)\big] \Big\}. \quad (3)$$

The main idea underlying the AIC is the view of the maximum point of Kulback similarity between the model and universe

$$n^* = \arg\max_n \int \Big[\ln \Phi\big(\mathbf{y}\,|\,(\mathbf{c}^*_n, \mathbf{0})\big)\Big] \Phi^*(\mathbf{y}) d\mathbf{y} \quad (4)$$

just as the desired dimension under the assumption that $\Phi^*(\mathbf{y}) = \Phi\big(\mathbf{y}\,|\,(\mathbf{c}^*_{n^*}, \mathbf{0})\big)$ with some value $(\mathbf{c}^*_{n^*}, \mathbf{0})$, cut out from the unknown $\mathbf{c}^* = (c^*_1, \ldots, c^*_m)$.

One of the first applications of AIC was modeling of a nonstationary signal on the discrete time axis by dividing the time interval into an unknown number n of blocks and adjusting a locally stationary autoregression model of a fixed order k to each of them [2].

After Akaike's pioneering paper [1], numerous modification of the information-based parsimony principle in model building were proposed [3],[4],[5],[6], among which the Bayesian Information Criterion (BIC) [3] has found the most wide adoption. However, all the known model selection criteria are aimed at the problem of choosing the most appropriate model within a succession of rigidly nested model classes.

The search for ways of generalizing the classical AIC, undertaken in this paper, was prompted by the needs of nonstationary signal analysis when the regression model of the given time series $\big((y_t, \mathbf{x}_t),\ t = 1, \ldots, N\big)$

$$y_t = \mathbf{c}_t^T \mathbf{x}_t + \eta_t,\ \mathbf{c}_t, \mathbf{x}_t \in \mathbb{R}^k,\ \eta_t \sim \mathcal{N}(\eta_t\,|\,0, \delta),\ E(\eta_t, \eta_s) = 0, \quad (5)$$

is assumed to be changing gradually over the observation interval [7], [8]. In this scenario, the dimension of the parameter vector in the family of conditional probability densities $\Phi(\mathbf{y} \mid \mathbf{x}, \mathbf{c})$ is fixed $\mathbf{c} = (\mathbf{c}_1^T \dots \mathbf{c}_N^T)^T \in \mathbb{R}^{kN}$ and k times exceeds the number of observations. Instead, it is assumed that the sequence of regression coefficients to be estimated is a hidden Markov random process

$$\mathbf{c}_t = \mathbf{c}_{t-1} + \xi_t,\ \xi_t \sim \mathcal{N}\left(\xi \,|\, \mathbf{0}, \lambda\delta\mathbf{I}\right),\ E\left(\xi_t \xi_s^T\right) = \mathbf{0}, \quad (6)$$

which starts with an unknown first value $\mathbf{c}_1 \sim \mathcal{N}(\mathbf{c}_1 \,|\, \mathbf{0}, \rho\mathbf{I})$, $\rho \to \infty$, and is excited by zero-mean white noise. We assume the noise variance $\lambda\delta$ in the Markov model (6) to be proportional to that in the observation model (5) because the Bayesian estimate depends only on the ratio between these two variances. The noise-variance coefficient λ is the structural parameter which determines the time-volatility level of the regression coefficients, ranging from full stationarity $\lambda = 0$ to absolute independence of instant regression models $\lambda \to \infty$.

This is a typical example of a softly constrained signal model in which the growing values of λ define a system of continuously nested families of degenerate a priori probability densities $\Psi(\mathbf{c} \mid \lambda)$ starting from the "uniform" distribution in $\mathbb{R}^k$ when $\lambda = 0$ and ending with the "uniform" one in $\mathbb{R}^{kN}$ when $\lambda \to \infty$. This situation suggests the informal notion of some effective dimension of the parameter $\mathbf{c}$ continuously changing from k to kN as λ grows instead of the discrete sequence of integer-valued dimensions. It is required to find the most appropriate value of λ which would provide sufficient approximation of the given time series $\left((y_t,\mathbf{x}_t),\, t = 1, \dots, N\right)$ by the nonstationary regression model $\left(y_t = \mathbf{c}_t^T \mathbf{x}_t,\, t = 1, \dots, N\right)$, on the one hand, and avoid overfitting, on the other.

It is clear that Akaike's criterion is inapplicable to the problem of choosing the real-valued time-volatility parameter $0 < \lambda < \infty$ of the time-varying regression model. In [7], [8], we applied the leave-one-out cross validation embedded into the Kalman-Bucy filter-smoother. However, this principle inevitably leads to the necessity to process the given signal N times in accordance with its length, and destroys, thereby, the originally linear computational complexity of the estimation algorithm with respect to N.

In this paper, with the purpose of extending the computationally perfect Akaike's principle onto the case of data models with continuously changing effective dimension of the unknown parameter, we consider the parametric model of the unknown universe $F^*(\mathbf{y})$ as a continuous mixture of conditional densities from the given family $\Phi(\mathbf{y} \,|\, \mathbf{c})$, $\mathbf{c} \in \mathbb{R}^m$, with some assumed mixing density $\Psi(\mathbf{c} \,|\, \lambda)$:

$$F(\mathbf{y} \,|\, \lambda) = \int \Phi(\mathbf{y} \,|\, \mathbf{c}) \Psi(\mathbf{c} \,|\, \lambda) d\mathbf{c},\ \mathbf{c} \in \mathbb{R}^m. \quad (7)$$

The structural model parameter λ to be adjusted to the observed data set $\mathbf{y}$ is assumed to provide the optimal degree of moderating the too large dimension of $\mathbf{c}$. Once the value of λ is chosen, the Bayesian estimate will be the final result of data analysis:

$$\hat{\mathbf{c}}_\lambda(\mathbf{y}) = \arg\max \left[\ln \Phi(\mathbf{y} \,|\, \mathbf{c}) + \ln \Psi(\mathbf{c} \,|\, \lambda) \right]. \quad (8)$$

We keep to the same idea as (4), namely that of achieving the maximum fit of the model distribution $F(\mathbf{y} \,|\, \lambda)$ (7) to the universe $F^*(\mathbf{y})$ by varying λ. In the particular case, when the structural parameter is a whole positive number $0 \le \lambda \le m$ truncating the ordered elements of the parameter vector $\mathbf{c} = (\mathbf{c}_\lambda,\, \mathbf{c}_{m-\lambda} = \mathbf{0}) \in \mathbb{R}^m$, $\mathbf{c}_\lambda \in \mathbb{R}^\lambda$, the resulting continuous versions of the criterion boils down to the classical AIC with the respective choice of a priori density $\Psi(\mathbf{c} \,|\, \lambda)$.

Finally, we experimentally illustrate the proposed continuous generalization of AIC by its application to the problem of time-varying regression estimation, and compare the results with those obtained by the usual leave-one-out cross validation.

II. Time-varying regression model

Observation model. In the problem of time-varying regression estimation (5)-(6), the Bayesian estimate of the hidden sequence of regression coefficients $\mathbf{c} = (\mathbf{c}_1^T \cdots \mathbf{c}_N^T)^T \in \mathbb{R}^m$, $m = kN$, depends only on the ratio λ of assumed variances in observation δ and state $\delta\lambda$, but its statistical properties essentially depend on the observation-noise variance.

To put the model into an explicit form, we consider the column vectors $\mathbf{y} = (y_1 \cdots y_N)^T \in \mathbb{R}^N$ and $\mathbf{c} = (\mathbf{c}_1^T \cdots \mathbf{c}_N^T)^T \in \mathbb{R}^{kN}$, as well as the block-diagonal matrix $\mathbf{X} = (\mathbf{X}_{ts},\, t, s = 1, \dots, N)$ of total dimension $(kN \times N)$ with diagonal column-blocks $\left(\mathbf{X}_{tt} = \mathbf{x}_t(k \times 1),\, t = 1, ..., N\right)$ and nondiagonal blocks $\mathbf{X}_{ts} = \mathbf{0}$ $(k \times 1)$, $t \ne s$.

We shall consider here as random only the observations $\mathbf{y} = (y_1 \cdots y_N)$ and treat the sequence of regressors $(\mathbf{x}_1 \cdots \mathbf{x}_N)$ as fixed. Then, for the observation noise variance conventionally taken as equal to unity $\delta = 1$, the observation model (5) will produce the parametric family of probability densities

$$\Phi(\mathbf{y} \,|\, \mathbf{c}) = \mathcal{N}(\mathbf{y} \,|\, \mathbf{X}^T\mathbf{c}, \mathbf{I}) = \frac{1}{(2\pi)^{N/2}} \exp\Big[-\frac{1}{2}\Big((\mathbf{y} - \mathbf{X}^T\mathbf{c})^T(\mathbf{y} - \mathbf{X}^T\mathbf{c})\Big)\Big], \quad (9)$$

and the logarithmic likelihood function

$$\ln \Phi(\mathbf{y} \,|\, \mathbf{c}) = const - \frac{1}{2}\sum_{t=1}^N \left(y_t - \mathbf{x}_t^T\mathbf{c}_t\right)^2 = const - \frac{1}{2}\Big((\mathbf{y} - \mathbf{X}^T\mathbf{c})^T(\mathbf{y} - \mathbf{X}^T\mathbf{c})\Big), \quad (10)$$

whose negative semidefinite Hessian is usually called Fisher information matrix:

$$\mathbf{A} = \bigtriangledown^2_{\mathbf{cc}} \ln \Phi(\mathbf{y} \,|\, \mathbf{c}). \quad (11)$$

For time-varying regression, the Hessian $\mathbf{A} = -\mathbf{X}\mathbf{X}^T$ $(kN \times kN)$ is always degenerate as being block-diagonal matrix with diagonal degenerate blocks $\mathbf{A}_{tt} = \mathbf{x}_t\mathbf{x}_t^T (k \times k)$, $t = 1, \dots, N$. If the regressors $\left[(x_{it}, t = 1, \dots, N),\ i = 1, \dots, k\right]$ are linearly independent, the rank of $\mathbf{A}$ reaches its maximum value $Rank(\mathbf{A}) = N$.

State - space model. With $\delta = 1$, the hidden Markov model of regression coefficients (6) is expressed by the family of normal a priori distributions with zero mathematical expectations:

$$\Psi(\mathbf{c} \,|\, \rho, \lambda) = \mathcal{N}(\mathbf{c}_1 \,|\, \mathbf{0}, \rho\mathbf{I}) \prod_{t=2}^N \mathcal{N}(\mathbf{c}_t \,|\, \mathbf{c}_{t-1}, \lambda\mathbf{I}) = \frac{1}{\rho^{k/2}\lambda^{k(N-1)/2}(2\pi)^{kN/2}} \exp\Big[-\frac{1}{2}\Big(\frac{1}{\rho}\mathbf{c}_1^T\mathbf{c}_1 + \frac{1}{\lambda}\sum_{t=2}^N (\mathbf{c}_t - \mathbf{c}_{t-1})^T(\mathbf{c}_t - \mathbf{c}_{t-1})\Big)\Big],$$

$$\ln\Psi(\mathbf{c}\,|\,\rho,\lambda)=const-\frac{1}{2}\Big[k\ln\rho+k(N-1)\ln\lambda+\frac{1}{\rho}\mathbf{c}_1^T\mathbf{c}_1+\frac{1}{\lambda}\sum\nolimits_{t=2}^{N}(\mathbf{c}_t-\mathbf{c}_{t-1})^T(\mathbf{c}_t-\mathbf{c}_{t-1})\Big)\Big].$$

In practice, the a priori information on the first vector of regression coefficients $\mathbf{c}_1$ in the Markov state-space model (6) is hardly available, therefore, we put the variance ρ equal to a sufficiently large number $\rho\to\infty$. Under this assumption, we come the logarithmic a priory density

$$\ln\Psi(\mathbf{c}\,|\,\lambda)=const-\frac{1}{2}\Big[k(N-1)\ln\lambda+\frac{1}{\lambda}\sum\nolimits_{t=2}^{N}(\mathbf{c}_t-\mathbf{c}_{t-1})^T(\mathbf{c}_t-\mathbf{c}_{t-1})\Big]. \tag{12}$$

The Bayesian estimation of time-varying regression coefficients $\hat{\mathbf{c}}_\lambda(\mathbf{y})=\hat{\mathbf{c}}_\lambda(\mathbf{y},\mathbf{X})=\big(\mathbf{c}_{1,\lambda}(\mathbf{y},\mathbf{X}),\ldots,\mathbf{c}_{N,\lambda}(\mathbf{y},\mathbf{X})\big)$ (8) is provided via minimization of the Flexible Least Squares criterion

$$\hat{\mathbf{c}}_\lambda(\mathbf{y})=\arg\min\Big[\sum\nolimits_{t=1}^{N}\big(y_t-\mathbf{x}_t^T\mathbf{c}_t\big)^2+\frac{1}{\lambda}\sum\nolimits_{t=2}^{N}(\mathbf{c}_t-\mathbf{c}_{t-1})^T(\mathbf{c}_t-\mathbf{c}_{t-1})\Big] \tag{13}$$

by the Kalman-Bucy filter-smoother [7], [8] for the time proportional to N.

Cross-validation principle of choosing the time-volatility level. Up to now, the leave-one-out cross validation has been the only known method of estimating the state-space variance λ responsible for the time-volatility of regression coefficients [7], [8].

In the full time series $\big((y_1,\mathbf{x}_1),\ldots,(y_N,\mathbf{x}_N)\big)$, single elements $t=1,\ldots,N$ are skipped one-by-one $\big((y_1,\mathbf{x}_1),\ldots,(y_{t-1},\mathbf{x}_{t-1}),(y_{t+1},\mathbf{x}_{t+1}),\ldots,y_N,\mathbf{x}_N)\big)$, each time via replacing the sum $\sum_{t=1}^{N}\big(y_t-\mathbf{x}_t^T\mathbf{c}_t\big)^2$ in (13) with a tentative value of λ by the truncated sum $\sum_{s=1,s\neq t}^{N}\big(y_s-\mathbf{x}_s^T\mathbf{c}_s\big)^2$, and the optimal vector sequences $\big(\mathbf{c}_{s,\lambda}^{(t)}(\mathbf{y}),s=1,\ldots,N\big)$ are found, where the upper index (t) means that the observation $(y_t,\mathbf{x}_t)$ was omitted when computing the respective estimate. For each t, the instantaneous squared prediction error is calculated using the respective single estimate $\big(y_t-(\mathbf{c}_{t,\lambda}^{(t)})^T\mathbf{x}_t\big)^2$. The cross-validation criterion for the given λ is computed as the average over all of the local squared prediction errors

$$CV(\lambda\mid\mathbf{y})=\frac{1}{N}\sum\nolimits_{t=1}^{N}\big(y_t-(\mathbf{c}_{t,\lambda}^{(t)})^T\mathbf{x}_t\big)^2. \tag{14}$$

The value $\hat{\lambda}(\mathbf{y})=\arg\min_\lambda CV(\lambda\mid\mathbf{y})$ is recommended as the optimal time-volatility parameter for the given time series.

This method is quite time consuming, because, to compute one value of the criterion (14), it is required to run the dynamic programming procedure N times for each skipped element of the signal.

III. Summary of basic properties of the assumed parametric density families

Below, in Section IV, when studying the way of extending the classical AIC onto a wider class of probabilistic data models, we are not going to restrict our consideration to only the problem of time-varying regression estimation.

We consider the given data set $\mathbf{y}$, in general, as a realization of a random variable defined by a parametric family of probability densities $\Phi(\mathbf{y}\mid\mathbf{c})$ in the respective space of observations. The only assumptions are, first, that the logarithmic likelihood $\ln\Phi(\mathbf{y}\mid\mathbf{c})$, $\mathbf{c}\in\mathbb{R}^m$, is concave quadratic function around the maximum-likelihood estimate $\hat{\mathbf{c}}(\mathbf{y})$ (1), even if it is not unique, as it is in the case of time-varying regression model $m=kN>N$, and, second, that the Hessian $\mathbf{A}$ $(m\times m)$ (11) does not depend on the random data $\mathbf{y}$:

$$\ln\Phi(\mathbf{y}\,|\,\mathbf{c})=\ln\Phi\big(\mathbf{y}\,|\,\hat{\mathbf{c}}(\mathbf{y})\big)+\frac{1}{2}\big(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})\big)^T\mathbf{A}\big(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})\big),\quad \bigtriangledown_{\mathbf{c}}\log\Phi(\mathbf{y}\,|\,\mathbf{c})=\mathbf{A}\big(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})\big). \tag{15}$$

In the particular case of time-varying regression estimation, we have $\mathbf{A}=-\mathbf{X}\mathbf{X}^T$.

Further, we suppose that

$$\ln\Psi(\mathbf{c}\,|\,\lambda)=const_\lambda+\frac{1}{2}\mathbf{c}^T\mathbf{D}_\lambda\mathbf{c},\quad \bigtriangledown_{\mathbf{c}}\ln\Psi(\mathbf{c}\,|\,\lambda)=\mathbf{D}_\lambda\mathbf{c}, \tag{16}$$

which assumption covers, in particular, the time-varying regression model (12) with

$$const_\lambda=const-\frac{1}{2}k(N-1)\ln\lambda,\quad \mathbf{D}_\lambda=-(1/\lambda)\mathbf{B}, \tag{17}$$

where $\mathbf{B}$ is block-three-diagonal matrix $(kN\times kN)$ with the diagonal $\big(\mathbf{I},2\mathbf{I},\ldots,2\mathbf{I},\mathbf{I}\big)$ and two off-diagonals $\big(-\mathbf{I},\ldots,-\mathbf{I}\big)$ formed by identity matrices $\mathbf{I}(k\times k)$.

For a fixed λ, the Bayesian estimate $\hat{\mathbf{c}}_\lambda(\mathbf{y})$ (8) is unique if the Hessian $\bigtriangledown_{\mathbf{cc}}\big[\ln\Phi(\mathbf{y}\,|\,\mathbf{c})+\ln\Psi(\mathbf{c}\,|\,\lambda)\big]=\mathbf{A}+\mathbf{D}_\lambda$ is negative definite. This is the case in most practical situations even if $\mathbf{A}$ is degenerate and, so, the maximum likelihood estimate $\hat{\mathbf{c}}(\mathbf{y})$ is not uniquely defined. More over, $\mathbf{A}+\mathbf{D}_\lambda$ is usually nondegenerate even if both $\mathbf{A}$ and $\mathbf{D}_\lambda$ are degenerate.

In what follows, we shall need some more detailed properties of the relationship between $\hat{\mathbf{c}}(\mathbf{y})$ and $\hat{\mathbf{c}}_\lambda(\mathbf{y})$.

Let the random sample $\mathbf{y}$ be produced by a probability distribution $\Phi(\mathbf{y}\,|\,\mathbf{c})$ with some fixed parameter value $\mathbf{c}$. It is well known for a much wider class of conditional densities than the above-specified class (15), that, if $\mathbf{A}$ is full-rank matrix $Rank(\mathbf{A})=m$, the random maximum likelihood estimate $\hat{\mathbf{c}}(\mathbf{y})$ is unbiased

$$\int\hat{\mathbf{c}}(\mathbf{y})\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}=\mathbf{c}, \tag{18}$$

and its conditional covariance matrix is completely determined by the Fisher information matrix:

$$\int\big(\hat{\mathbf{c}}(\mathbf{y})-\mathbf{c}\big)\big(\hat{\mathbf{c}}(\mathbf{y})-\mathbf{c}\big)^T\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}=-\mathbf{A}^{-1}. \tag{19}$$

In the more general case, if $Rank(\mathbf{A})<m$, (18) and (19) should be treated as

$$\int\mathbf{A}\big(\hat{\mathbf{c}}(\mathbf{y})-\mathbf{c}\big)\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}=\mathbf{0}, \tag{20}$$

$$\int\Big[\mathbf{A}\big(\hat{\mathbf{c}}(\mathbf{y})-\mathbf{c}\big)\Big]\Big[\mathbf{A}\big(\hat{\mathbf{c}}(\mathbf{y})-\mathbf{c}\big)\Big]^T\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}=-\mathbf{A}. \tag{21}$$

If (15) and (16) are met, the random Bayesian estimate (8) is a linear function of the likelihood estimate

$$\hat{\mathbf{c}}_\lambda(\mathbf{y})=(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\mathbf{A}\hat{\mathbf{c}}(\mathbf{y}) \tag{22}$$

with conditional covariance matrix relative to the fixed value of parameter $\mathbf{c}$

$$\int\big(\hat{\mathbf{c}}_\lambda(\mathbf{y})-\hat{\mathbf{c}}_\lambda(\mathbf{c})\big)\big(\hat{\mathbf{c}}_\lambda(\mathbf{y})-\hat{\mathbf{c}}_\lambda(\mathbf{c})\big)^T\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}= -(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}, \tag{23}$$

where $\hat{\mathbf{c}}_\lambda(\mathbf{c})$ is the conditional mathematical expectation

$$\hat{\mathbf{c}}_\lambda(\mathbf{c})=\int\hat{\mathbf{c}}_\lambda(\mathbf{y})\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}=(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\mathbf{A}\hat{\mathbf{c}}. \tag{24}$$

IV. Measuring the Kullback similarity between the model and universe: The principle of maximum fit to the actual distribution of the observed variable

It appears natural to mathematically express the observer's aim as maximizing the Kullback similarity between $F(\mathbf{y}\,|\,\lambda)$ and $F^*(\mathbf{y})$ like in (4):

$$\lambda^*=\arg\max_\lambda\int\big[\ln F(\mathbf{y}\,|\,\lambda)\big]F^*(\mathbf{y})d\mathbf{y}. \tag{25}$$

This "ideal" criterion suits for any actual distribution $F^*(\mathbf{y})$, but we assume here that it is consistent with the accepted parametric family $\Phi(\mathbf{y}\mid\mathbf{c})$, i.e., there exists a distribution $\Psi^*(\mathbf{c})$ such that

$$F^*(\mathbf{y})=\int\Phi(\mathbf{y}\,|\,\mathbf{c})\Psi^*(\mathbf{c})d\mathbf{c}. \tag{26}$$

However, an immediate realization of criterion (25) is impossible even for the reason alone that the actual distribution $F^*(\mathbf{y})$ is unknown. The maximization of the likelihood function for the only available sample $\ln F(\mathbf{y}\mid\lambda)$ (7) as an unbiased estimate of the criterion is also senseless, because it will prefer the values of the structural parameter suppressing moderation of the too large dimension of $\mathbf{c}\in\mathbb{R}^m$.

To overcome "the curse of the only sample", we apply the respective generalization of Akaike's reasoning underlying the classical AIC [1], namely, imagine the existence of another independent sample $\tilde{\mathbf{y}}$ yielding the random Bayesian estimate $\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})$ (8), and replace $\ln F(\mathbf{y}\mid\lambda)$ in (25) by the mathematical expectation of $\ln\Phi(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))$:

$$\hat{\lambda}=\arg\max_\lambda\int\Big\{\int\Big\{\int\big[\ln\Phi(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))\big]\Phi(\tilde{\mathbf{y}}\,|\,\mathbf{c})d\tilde{\mathbf{y}}\Big\}\times\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}\Big\}\Psi^*(\mathbf{c})d\mathbf{c}. \tag{27}$$

Theorem 1. Under the assumptions (15) and (16),

$$\int\Big\{\int\Big\{\int\big[\ln\Phi(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))\big]\Phi(\tilde{\mathbf{y}}\,|\,\mathbf{c})d\tilde{\mathbf{y}}\Big\}\times\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}\Big\}\Psi^*(\mathbf{c})d\mathbf{c}=\int J(\lambda\,|\,\mathbf{y})F^*(\mathbf{y})d\mathbf{y}, \tag{28}$$

$$J(\lambda\,|\,\mathbf{y})=\ln\Phi\big(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\mathbf{y})\big)-Tr\Big[\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\Big]. \tag{29}$$

Proof is based on the quadratic representation of $\ln\Phi(\mathbf{y}\,|\,\mathbf{c})$ (15) at $\mathbf{c}=\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})$ and equalities (19)-(23).

Theorem 1 suggests a way of forming a continuous analog of the classical AIC. Despite the fact that the density $\Psi^*(\mathbf{c})$ in (26) remains unknown and, so, the original criterion (27) is computationally intractable, the equality (28) shows that the easily computable function $J(\lambda\mid\mathbf{y})$ is an unbiased estimate of the full criterion. As a reasonable compromise, which is analogous to Akaike's reasoning, this function may be immediately maximized with respect to the sought-for value of the structural parameter:

$$\hat{\lambda}(\mathbf{y})=\arg\max_\lambda J(\lambda\,|\,\mathbf{y})= \arg\max_\lambda\Big\{\ln\Phi\big(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\mathbf{y})\big)-Tr\Big[\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\Big]\Big\}. \tag{30}$$

This is just a continuous generalization of AIC (3). Comparison of (30) and (3) suggests interpretation of the penalty term $Tr\Big[\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\Big]$ as a conventional effective dimension of the parameter $\mathbf{c}$ whose choice is constrained by the a priori distribution $\ln\Psi(\mathbf{c}\mid\lambda)$.

V. A particular case: The classical AIC

Let the structural parameter be a whole positive number $0\le\lambda\le m$ truncating the ordered elements of the parameter vector $\mathbf{c}=(\mathbf{c}_\lambda,\mathbf{c}_{m-\lambda})\in\mathbb{R}^m$ as in (2) with $n=\lambda$, i.e. $\mathbf{c}_\lambda\in\mathbb{R}^\lambda$, $\mathbf{c}_{m-\lambda}=\mathbf{0}\in\mathbb{R}^{m-\lambda}$. The absence of any a priori information on vector $\mathbf{c}$ may be expressed in terms of an "almost uniform" normal distribution:

$$\Psi(\mathbf{c}_\lambda\,|\,\lambda)=\textstyle\prod_{i=1}^{\lambda}\psi_i(c_i),\ \psi_i(c_i)=\mathcal{N}(c_i\,|\,0,\sigma^2),\ \sigma^2\to\infty,$$

$$\Psi(\mathbf{c}_\lambda\,|\,\lambda)\cong const=0,\quad \ln\Psi(\mathbf{c}_\lambda\,|\,\lambda)\cong const\ll 0.$$

Since only the first part of the vector parameter is free in the conditional density $\Phi(\mathbf{y}|\mathbf{c}_\lambda,\,\mathbf{c}_{m-\lambda})$, the Hessian $\mathbf{A}_\lambda=\bigtriangledown^2_{\mathbf{c}_\lambda\mathbf{c}_\lambda}\ln\Phi(\mathbf{y}|\mathbf{c}_\lambda,\mathbf{0})$ is a matrix $(\lambda\times\lambda)$. Under these assumptions, the continuous AIC (30) reduces to the criterion (3):

$$\max_{\mathbf{c}_\lambda}\ln\Phi(\mathbf{y}\,|\,\mathbf{c}_\lambda,\mathbf{0})-Rank\,(\mathbf{A}_\lambda)\to\max_\lambda.$$

VI. The continuous AIC for the time-volatility of time-varying regression coefficients

The form of the continuous effective dimension $Tr\left[\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\right]$ in (30) as function of the structural parameter λ depends on how it occurs in the Hessian of the logarithmic a priori density $\mathbf{D}_\lambda$ (16). If it is a strictly increasing function of λ, the logarithmic likelihood at the a posteriori optimal point will be increasing function, too, tending to a constant $\ln\Phi\big(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\mathbf{y})\big)\to const$ as $\lambda\to\infty$. It should be expected, in this case, that $J(\lambda|\,\mathbf{y})$ has a maximum.

In the general case, the most appropriate value $\hat{\lambda}(\mathbf{y})$ can be found by computing the criterion for a succession of tentative values $\lambda^{(1)}<\ldots<\lambda^{(M)}$ with a sufficiently small step, just as when the classical AIC (3) or the cross-validation criterion (14) is applied.

A more detailed study of $J(\lambda\mid\mathbf{y})$ requires making more specific assumptions on the Hessian $\mathbf{D}_\lambda$ as matrix function of λ. In this Section, we consider the specificity of the time-varying regression model which allows, first, for easy computation of the penalty term in the continuous AIC (30) and, second, for a numerical iterative maximization of the criterion instead of the plain search in a sufficiently dense set of fixed values $\lambda^{(1)}<\ldots<\lambda^{(M)}$.

In accordance with notations accepted in (10) and (16), the penalty term in the criterion (30) will have the form

$$Tr\Big[\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\Big] = Tr\Big[\mathbf{X}\mathbf{X}^T\big(\mathbf{X}\mathbf{X}^T+\frac{1}{\lambda}\mathbf{B}\big)^{-1}\Big].$$

Let the symmetric inverse sum of matrices $\big(\mathbf{X}\mathbf{X}^T+\frac{1}{\lambda}\mathbf{B}\big)^{-1}$ be represented in block-wise form as $\big(\mathbf{X}\mathbf{X}^T+\frac{1}{\lambda}\mathbf{B}\big)^{-1} = \mathbf{G}_\lambda = (\mathbf{G}_{\lambda,\,t},\ s,t=1,\ldots,N)$ with square blocks $\mathbf{G}_{\lambda,ts} = \mathbf{G}_{\lambda,st}^T$. Then, since matrix $\mathbf{X}\mathbf{X}^T$ is block-diagonal, the penalty term will depend only on the diagonal blocks of $\mathbf{G}_\lambda$:

$$Tr\Big[\mathbf{A}(\mathbf{A}+\mathbf{D}_\lambda)^{-1}\Big] = Tr\Big[\mathbf{X}\mathbf{X}^T\big(\mathbf{X}\mathbf{X}^T+\tfrac{1}{\lambda}\mathbf{B}\big)^{-1}\Big] = \sum_{t=1}^{N} Tr\big(\mathbf{x}_t\mathbf{x}_t^T\mathbf{G}_{\lambda,\,tt}\big).$$

Thus, to compute the penalty term in the criterion (30), it is enough, instead of full inverting the sum of matrices for each value of λ, to compute the diagonal blocks of inversion $\mathbf{G}_{\lambda,\,tt}$. Generally speaking, the problem of finding each column of blocks $(\mathbf{G}_{\lambda,\,1t},\ldots,\mathbf{G}_{\lambda,\,Nt})$ is that of solving a system of linear equation, which is block-three-diagonal because so is the initial matrix $\big(\mathbf{X}\mathbf{X}^T+\frac{1}{\lambda}\mathbf{B}\big)$ to be inverted. Hence, in each column, the symmetrically indexed block $\mathbf{G}_{\lambda,\,tt}$ can be easily computed by a slight modification of the sweep method without computing other blocks.

Theorem 2. For any sequence of vector regressors $(\mathbf{x}_t,\ t=1,\ldots,N)$ such that their elements $\big[(x_{it}, t=1,\ldots,N),\ i=1,\ldots,k\big]$ are linearly independent, the following assertions are met:

$$\begin{cases} \lim\limits_{\lambda\to 0} Tr\left[\mathbf{X}\mathbf{X}^T\big(\mathbf{X}\mathbf{X}^T+\dfrac{1}{\lambda}\mathbf{B}\big)^{-1}\right] = k, \\ \lim\limits_{\lambda\to\infty} Tr\left[\mathbf{X}\mathbf{X}^T\big(\mathbf{X}\mathbf{X}^T+\dfrac{1}{\lambda}\mathbf{B}\big)^{-1}\right] = N. \end{cases} \tag{31}$$

Proof is omitted here. However, this mathematical result appears to be expectable. Indeed, if $\lambda\to 0$, the model loses its time-volatility, and k instantaneous values of the regression coefficients completely determine the entire sequence. On the contrary, if $\lambda\to\infty$, there is no a priori information on the kN coefficients, but the maximum number of unknown values which can be inferred from N observations if noise is completely absent equals just the rank of the Hessian of the likelihood function $\mathbf{X}\mathbf{X}^T$, namely, the number of observations.

The first item of the AIC criterion (30) is the logarithmic likelihood function at the point of the Bayesian estimate of the sequence of regression coefficients $\ln\Phi\big(\mathbf{y}\mid\hat{\mathbf{c}}_\lambda(\mathbf{y})\big)$. This is a strictly increasing random function of λ, which tends to its constant maximum-likelihood value $\ln\Phi\big(\mathbf{y}\mid\hat{\mathbf{c}}(\mathbf{y})\big)$ (15) as $\lambda\to\infty$. The difference $J(\lambda\mid\mathbf{y})$ (29) is also a random function, but it remains practically always unimodal. A typical plot of the continuous AIC and its constituents as functions of λ is shown below in Figure 1.

The strong unimodality of the continuous AIC allows for computing its maximum point by an appropriate method of one-dimensional optimization. We use the method of golden section, which consists in finding a sequence of shortening intervals (λ'_s,λ''_s), $s=0,1,2,...$, starting with a sufficiently large initial interval $(\lambda'_0=0,\ \lambda''_0=\lambda_{max})$.

VII. Comparative Specification of Regression Time-Volatility via Continuous Akaike Criterion and Cross-Validation

The experiments are built by the following scheme:

– simulating two sufficiently large sets of 200 random time series each with two known relatively smooth sequences of three-dimensional regression coefficients $(\mathbf{c}_1^{*(1)},\ldots,\mathbf{c}_N^{*(1)})$ and $(\mathbf{c}_1^{*(2)},\ldots,\mathbf{c}_N^{*(2)})$ having essentially different styles of oscillation;

– inferring the Bayesian estimates $(\hat{\mathbf{c}}_{\lambda,1},\ldots,\hat{\mathbf{c}}_{\lambda,N})$ from each realization by the Flexible Least Squares criterion (13) with choosing the time-volatility parameter λ via the continuous AIC and leave-one-out cross-validation;

– comparing the four averaged estimation errors.

One of the sequences of ground-truth regression coefficients $(\mathbf{c}_1^{*(1)},\ldots,\mathbf{c}_N^{*(1)})$ was built in full accordance with the theoretically assumed normal Markov model (12). On the contrary, the second sequence $(\mathbf{c}_1^{*(2)},\ldots,\mathbf{c}_N^{*(2)})$ was formed by three sinusoidal functions of time $c_{it}^{*(2)} = 4\sin\big((2\pi/N)t+(2\pi/3)(i-1)\big)$ mutually shifted by phase.

All the 400 time series $\big((y_1,\mathbf{x}_1),\ldots,(y_N,\mathbf{x}_N)\big)$ had the length $N=50$ and were simulated with the same sequence of regressors $(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ obtained as the set of independent random normal three-dimensional vectors with zero mean and the same variance of independent components. The random output values y_t (5) were generated with 10% noise variance $\delta = 0.1\big((1/N)\sum_{i=1}^{N}(\mathbf{x}_t^T\mathbf{c}_t^*)^2\big)$ depending on the respective ground-truth sequence of regression coefficients.

Finally, for each of time series, the two estimates of regression coefficients $(\hat{\mathbf{c}}_{\lambda,1},\ldots,\hat{\mathbf{c}}_{\lambda,N})$ were compared with the respective ground-truth model by the criterion of relative mean deflection of the estimate from the model $\varepsilon = \sum_{t=1}^{N}(\hat{\mathbf{c}}_{t,\hat{\lambda}}-\mathbf{c}_t^*)^T(\hat{\mathbf{c}}_{t,\hat{\lambda}}-\mathbf{c}_t^*)/\sum_{t=1}^{N}(\mathbf{c}_t^*)^T\mathbf{c}_t^*$.

We obtained the following results:

Time-variability estimation criterion	Mean deflection of the estimate from the model	
	Markov-model coefficients	Sinusoidal coefficients
Continuous Akaike	0.016	0.021
Leave-one-out cross validation	0.046	0.015

The plots of function $J(\lambda|\mathbf{y})$ (29) along with its constituents $\ln\Phi\big(\mathbf{y}\mid\hat{\mathbf{c}}_\lambda(\mathbf{y})\big)+\ln\Psi\big(\hat{\mathbf{c}}_\lambda(\mathbf{y})\mid\lambda\big)$ and $-Tr\Big[\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T+\frac{1}{\lambda}\mathbf{B})^{-1}\Big]$ experimentally computed from one of the random time series representing the sinusoidal sequence of regression coefficients are shown in the Figure 1 below. All other realizations of $J(\lambda\mid\mathbf{y})$ display the same strong unimodality.

VIII. Discussing the Experimental Results and Conclusion

The continuous Akaike criterion, just as the classical AIC, on one hand, and the cross validation as a representative

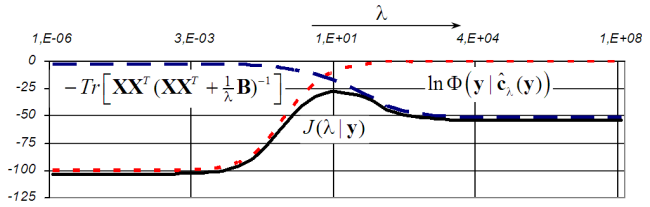

Fig. 1. An example of function $J(\lambda\,|\,\mathbf{y})$ and its constituents.

of the large group of resampling methods, on the other, are essentially different by the principle of adjusting a model to the given data set within a succession of nested model classes.

The idea of resampling is quite a straightforward imitation of the dream of an immediate access to the universe. The leave-one-out procedure is one of possible ways of drawing several subsets from, actually, the only available sample and treating them as though these were the samples taken from the universe. This is the only self-deception of the resampling, instead, it is free of any assumptions on some "expected" properties of the real world.

On the contrary, the very idea of maximizing the Kulback similarity between the class of nested probabilistic data models and the random universe in the form of (25) is based on the hope that the assumed class of a priori densities $\Psi(\mathbf{c}\mid\lambda)$ "almost contains" the hypothetical "actual" distribution $\Psi^*(\mathbf{c})$. If this is a justified hope, the informational approach must essentially outperform cross-validation, but if not, it will rather lose the competition.

The results of our experiments look as confirming these philosophical considerations. Indeed, the assumed first-order Markov model of the hidden sequence of regression coefficients coincides with the actual mechanism of forming the first ground-truth sequence $(\mathbf{c}_1^{*(1)},\ldots,\mathbf{c}_N^{*(1)})$ but is in contradiction with the second one $(\mathbf{c}_1^{*(1)},\ldots,\mathbf{c}_N^{*(1)})$, because a sinusoidal signal can be generated only by a second-order Markov model. The fact that the continuous AIC has won in the case of Markov-model coefficients and lost the competition at the sinusoidal sequence may be referred just to this difference.

Acknowledgments

This work is supported by the Russian Foundation of Basic Research, Grants 08-01-12023 and 08-01-00695-a.

References

[1] Akaike H. A new look at the statistical model idendification. *IEEE Trans. on Automatic Control*, Vol. IC-19, No.6, December 1974, pp. 716-723.
[2] Kitagawa G., Akaike H. A procedure for the modeling of no-stationary time series. *Ann. Inst. Statist. Math.*, Vol. 30, Part B, 1987, pp. 351-363.
[3] Scharz G. Estimating the dimtnsion of the model. *The Annals of Statistics*, Vol. 6,No.2, 1978, pp. 461-464.
[4] Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrica*, Vol. 52, No.3, September 1987.
[5] Spiegelhalter D., Best N., Carlin B. Van der Linde A. Bayesian mesures of model complexity and fit. *Journal of the Royal Statistical Society*. Series B (Statistical Methodology), Vol. 64, No.4, 2002, pp. 583-639.
[6] Rodrigues C.C. The ABC of model selection: AIC, BIC and new CIC. *AIP Conference Proceedings*, Vol. 803, November 23, 2005, pp. 80-87.
[7] Markov M., Krasotkina O., Mottl V., Muchnik I. Time-varying regression model with unknown time-volatility for nonstationary signal analyses. *Proceedings of the 8th IASTED Internation Conference on Signal and Image Processing*. Honolulu, Hawaii, USA, August 14-16, 2006.
[8] Markov M., Muchnik I., Mottl V., Krasotkina O. Dynamic analysis of hedge funds. *Proceedings of the 3rd IASTED Internation Conference on Financal Engineering and Applications*. Cambridge, Massachusetts, USA, October 9-11, 2006.
[9] Bishop C.M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Elena Ezhova received the B.S. Degree in Computer Science from the Moscow Institute of Physics and Technology, Russia, in 2008. She is working now towards the M.S. Degree in Computer Science at the Department of Intelligent Systems. Her scientific interests include computational statistics, machine learning, methodology of improving generalization ability in data analysis.

Vadim Mottl received the Ph.D. Degree in 1979 and the D.Sci. Degree in 1994, both in Computer Science, from the Institute of Control Sciences of the Russian Academy of Sciences in Moscow. He is now Principal Investigator at the Computing Center of the Russian Academy of Sciences, and Professor at the Moscow Institute of Physics and Technology and Tula State University. His scientific interests embrace theoretical aspects of pattern recognition and machine learning with applications to signal and image analysis.

Olga Krasotkina received the Ph.D. Degree in Computer Science from the Computing Center of the Russian Academy of Sciences in 2003. She is now Associate Professor of the Tula State University, Department of Automation and Remote Control. Her scientific interests are concentrated around adaptation of the machine learning ideology to the needs of signal analysis.