Estimation of Nonstationary Linear Regression with Unknown Time-Variability via Continuous Generalization of the Akaike Information Criterion

V. Mottl[1], O. Krasotkina[2], and E. Ezhova[3]

[1] Computing Center of the Russian Academy of Sciences, Moscow, Russia
vmottl@yandex.ru,
[2] Tula State University Tula, Russia
ko180177@yandex.ru,
[3] Moscow Institute of Physics and Technology, Moscow, Russia
lena-ezhova@rambler.ru

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.

Key words: Akaike Information Criterion (AIC), Kullback information maximization principle, nonstationary signal, time-variability of regression coefficients

1 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

$$\hat{\mathbf{c}}(\mathbf{y}) = \arg\max \ln \varPhi(\mathbf{y}\,|\,\mathbf{c}),\ \ln \varPhi(\mathbf{y}\,|\,\mathbf{c}) = \sum\nolimits_{t=1}^{N} \ln \varphi(y_t\,|\,\mathbf{c}). \tag{1}$$

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 $c_i = 0$, $n < i \leq m$:

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

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

The classical AIC is a criterion of choosing the dimension as the most appropriate level of model complexity $\hat{n}(\mathbf{y}) = \arg\max_n \Big[\ln \varPhi\Big(\mathbf{y} \mid \big(\mathbf{c}_n(\mathbf{y}), \mathbf{0}\big)\Big) - n\Big]$ instead of the plain likelihood maximization (1). However, this formula was designed under the assumption that $\bigtriangledown^2_{\mathbf{c}_n\mathbf{c}_n} \ln \varPhi\big(\mathbf{y}\,|\,(\mathbf{c}_n, \mathbf{0})\big)$ 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:

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

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

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

just the desired dimension under the assumption that $\varPhi^*(\mathbf{y}) = \varPhi\big(\mathbf{y} \mid (\mathbf{c}^*_{n^*}, \mathbf{0})\big)$ 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 $\big((y_t,\mathbf{x}_t), t = 1, \ldots, N\big)$

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

is assumed to be changing gradually over the observation interval [7]. 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 $\mathbf{c} = (\mathbf{c}_1 \dots \mathbf{c}_N) \in \mathbb{R}^{kN}$ 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

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

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. The noise variance λ is the structural parameter which determines the time-variability 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 "fuzzy 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 $\big((y_t,\mathbf{x}_t),\ t = 1, \dots, N\big)$ by the nonstationary regression model $\big(y_t = \mathbf{c}_t^T \mathbf{x}_t,\ t = 1, \dots, N\big)$, 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], 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 fuzzy 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 $\Psi(\mathbf{c} \,|\, \lambda)$:

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

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:

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

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.

There exist, at least, two ways of treating the idea (4) under the assumption (7). We study here both of them and show that, in each case, the resulting continuous versions of the criterion boils down to the classical AIC with the respective choice of a priori density $\varPsi(\mathbf{c}\,|\,\lambda)$.

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.

2 Two ways of measuring the Kullback similarity between the model and universe

On the one hand, the mathematical expression of observer's aim would be maximizing the Kullback similarity between $F(\mathbf{y}\,|\,\lambda)$ and $F^*(\mathbf{y})$ like in (4):

$$\lambda^* = \arg\max_{\lambda} \int \big[\ln F(\mathbf{y}\,|\,\lambda)\big] F^*(\mathbf{y}) d\mathbf{y}. \tag{9}$$

This "ideal" criterion suits any actual distribution $F^*(\mathbf{y})$.

On the other hand, the accepted model (7) involves the random parameter $\mathbf{c}$ as a hidden variable. It may be considered as important to fit the joint distribution $H(\mathbf{c},\mathbf{y}\,|\,\lambda) = \varPsi(\mathbf{c}\,|\,\lambda)\varPhi(\mathbf{y}\,|\,\mathbf{c})$ to the hypothetical actual one $H^*(\mathbf{c},\mathbf{y})$. Such an intention makes sense only if the unknown distribution of the universe $F^*(\mathbf{y})$ is assumed to be consistent with the accepted parametric family $\varPhi(\mathbf{y}\,|\,\mathbf{c})$, i.e., if there exists a distribution $\varPsi^*(\mathbf{c})$ such that

$$F^*(\mathbf{y}) = \int \varPhi(\mathbf{y}\,|\,\mathbf{c})\varPsi^*(\mathbf{c}) d\mathbf{c}. \tag{10}$$

Then $H^*(\mathbf{c},\mathbf{y}) = \varPsi^*(\mathbf{c})\varPhi(\mathbf{y}\,|\,\mathbf{c})$, and the "ideal" criterion for choosing λ should be put as

$$\lambda^* = \arg\max_{\lambda} \iint \big[\ln H(\mathbf{c},\mathbf{y}\,|\,\lambda)\big] H^*(\mathbf{c},\mathbf{y}) d\mathbf{c} d\mathbf{y}. \tag{11}$$

We shall see that the concepts (9) and (11) lead to essentially different continuous generalizations of AIC.

3 Basic assumptions and some properties of the parametric density families

Assumptions. We restrict here our consideration only to the case when the parametric density family $\varphi(y\,|\,\mathbf{c})$ yields quadratic logarithmic likelihood functions $\ln \varPhi(\mathbf{y}\,|\,\mathbf{c})$ of parameter $\mathbf{c}$ for samples $\mathbf{y} = (y_j, j = 1, \ldots, N)$ of sufficiently large size N. This may be achieved, for instance, by the Laplace method of Gaussian approximation [8] to $\varPhi(\mathbf{y}\,|\,\mathbf{c})$ in a vicinity of the Bayesian estimate $\hat{\mathbf{c}}_\lambda(\mathbf{y})$ (8). So, the Hessian

$$\mathbf{A} = \bigtriangledown^2_{\mathbf{cc}} \ln \varPhi(\mathbf{y}\,|\,\mathbf{c}) \tag{12}$$

called Fisher information matrix is considered as not depending on the point $\mathbf{c}$ at which it is defined. In particular, for the maximum likelihood estimate $\hat{\mathbf{c}}_\lambda(\mathbf{y})$ (1), even if it is not unique,

$$\begin{aligned} &\ln \Phi(\mathbf{y}\,|\,\mathbf{c}) = \ln \Phi(\mathbf{y}\,|\,\hat{\mathbf{c}}(\mathbf{y})) + (1/2)(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y}))^T\mathbf{A}(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})), \\ &\bigtriangledown_{\mathbf{c}} \log \Phi(\mathbf{y}\,|\,\mathbf{c}) = \mathbf{A}(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})). \end{aligned} \tag{13}$$

Further, practical problems usually suggest constructing $\Psi(\mathbf{c}\,|\,\lambda)$ as a family of degenerate normal densities which are "uniform" in some parallel affine manifolds within $\mathbb{R}^m$. In this case, the notion of mathematical expectation as a point is, generally speaking, inapplicable, it is rather associated with a subspace, and only conventionally may be taken as equal to zero.

We shall assume that the logarithmic a priori densities $\ln \Psi(\mathbf{c} \mid \lambda)$ are quadratic functions which reach their maxima at zero $\bigtriangledown_{\mathbf{c}} \ln \Psi(\mathbf{0} \mid \lambda) = \mathbf{0}$ and are determined by negative semidefinite Hessians

$$\mathbf{B}_\lambda = \bigtriangledown^2_{\mathbf{cc}} \ln \Psi(\mathbf{c}\,|\,\lambda), \tag{14}$$

which are, as a rule, degenerate, so that

$$\begin{aligned} &\ln \Psi(\mathbf{c}\,|\,\lambda) = const_\lambda + (1/2)\mathbf{c}^T\mathbf{B}_\lambda\mathbf{c}, \\ &\bigtriangledown_{\mathbf{c}} \ln \Psi(\mathbf{c}\,|\,\lambda) = \mathbf{B}_\lambda\mathbf{c}. \end{aligned} \tag{15}$$

The dependence of $const_\lambda$ on parameter λ is determined by the specificity of the family of Hessians $\mathbf{B}_\lambda$. In particular, if the Hessians are nondegenerate, the positive definite matrices $-\mathbf{B}_\lambda^{-1}$ are covariance matrices of usual normal distributions with zero mathematical expectations:

$$\begin{aligned} \Psi(\mathbf{c}\,|\,\lambda) &= \frac{1}{\mid -\mathbf{B}_\lambda \mid^{-1/2} (2\pi)^{m/2}} \exp\Big(-\frac{1}{2}\mathbf{c}^T(-\mathbf{B}_\lambda)\mathbf{c}\Big), \\ \ln \Psi(\mathbf{c}\,|\,\lambda) &= \frac{1}{2}\ln|-\mathbf{B}_\lambda| - \ln\Big((2\pi)^{m/2}\Big) + \frac{1}{2}\mathbf{c}^T\mathbf{B}_\lambda\mathbf{c}. \end{aligned} \tag{16}$$

As to the unknown distribution of the universe $F^*(\mathbf{y})$, we shall assume that the family $\Phi(\mathbf{y}\,|\,\mathbf{c})$ is consistent with it in the sense that there exists an unknown density $\Psi^*(\mathbf{c})$ which allows for the representation

$$F^*(\mathbf{y}) = \int \Phi(\mathbf{y}\,|\,\mathbf{c})\Psi^*(\mathbf{c})d\mathbf{c} \tag{17}$$

Properties. For a fixed λ, the Bayesian estimate $\hat{\mathbf{c}}_\lambda(\mathbf{y})$ (8) is unique if the Hessian $\bigtriangledown_{\mathbf{cc}}[\ln \Phi(\mathbf{y}\,|\,\mathbf{c}) + \ln \Psi(\mathbf{c}\,|\,\lambda)] = \mathbf{A} + \mathbf{B}_\lambda$ is negative definite. This is the case in most practical situations even if $\mathbf{A}$ (12) is degenerate and, so, the maximum likelihood estimate $\hat{\mathbf{c}}(\mathbf{y})$ (1) is not uniquely defined. More over, $\mathbf{A}+\mathbf{B}_\lambda$ is usually nondegenerate even if both $\mathbf{A}$ and $\mathbf{B}_\lambda$ (14) 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 (13), that, if $\mathbf{A}$ is full-rank matrix $Rank(\mathbf{A}) = n$, 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 $Rank(\mathbf{A}) < n$, (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 (13) and (15) are met, the random Bayesian estimate (8) is a linear function of the likelihood estimate

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

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

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

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

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

4 The principle of maximum fit to the actual distribution of the observed variable

Criterion. An immediate realization of criterion (9) 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 $\ln F(\mathbf{y}\,|\,\lambda)$ (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 (9) by the mathematical expectation of $\ln \Phi(\mathbf{y} \,|\, \hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))$:

$$\hat{\lambda} = \arg\max_\lambda \int \Big\{ \int \Big\{ \int \big[\ln \Phi(\mathbf{y} \,|\, \hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))\big] \Phi(\tilde{\mathbf{y}} \,|\, \mathbf{c}) d\tilde{\mathbf{y}} \Big\} \Phi(\mathbf{y} \,|\, \mathbf{c}) d\mathbf{y} \Big\} \Psi^*(\mathbf{c}) d\mathbf{c}. \tag{25}$$

Proposition 1. *Under the assumptions (13) (15),*

$$\begin{aligned} &\int \Big\{ \int \Big\{ \int \big[\ln \Phi(\mathbf{y} \,|\, \hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))\big] \Phi(\tilde{\mathbf{y}} \,|\, \mathbf{c}) d\tilde{\mathbf{y}} \Big\} \Phi(\mathbf{y} \,|\, \mathbf{c}) d\mathbf{y} \Big\} \Psi^*(\mathbf{c}) d\mathbf{c} = \\ &\int J_1(\lambda \,|\, \mathbf{y}) F^*(\mathbf{y}) d\mathbf{y}, \qquad\qquad (26) \\ &J_1(\lambda \,|\, \mathbf{y}) = \ln \Phi\big(\mathbf{y} \,|\, \hat{\mathbf{c}}_\lambda(\mathbf{y})\big) - Tr\Big[\mathbf{A}(\mathbf{A} + \mathbf{B}_\lambda)^{-1}\Big]. \end{aligned}$$

Proof is based on the quadratic representation of $\ln \Phi(\mathbf{y} \,|\, \mathbf{c})$ (13) at $\mathbf{c} = \hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})$ and equalities (19)-(23).

Proposition 1 suggests a way of forming a continuous analog of the classical AIC. Despite the fact that the density $\Psi^*(\mathbf{c})$ in (17) remains unknown and, so, the original criterion (25) is computationally intractable, the equality (26) shows that the easily computable function $J_1(\lambda \mid \mathbf{y})$ 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:

$$\hat{\lambda}(\mathbf{y}) = \arg\max_\lambda \Big\{ \ln \Phi\big(\mathbf{y} \,|\, \hat{\mathbf{c}}_\lambda(\mathbf{y})\big) - Tr\Big[\mathbf{A}(\mathbf{A} + \mathbf{B}_\lambda)^{-1}\Big] \Big\}. \tag{27}$$

This is just a continuous generalization of AIC (3). Comparison of (27) and (3) suggests interpretation of the penalty term $Tr\Big[\mathbf{A}(\mathbf{A} + \mathbf{B}_\lambda)^{-1}\Big]$ as a conventional "fuzzy dimension" of the parameter $\mathbf{c}$ whose choice is constrained by distribution $\ln \Psi(\mathbf{c} \mid \lambda)$.

Algorithm of discrete search. However, to find the most appropriate $\hat{\lambda}$, it is required to compute the criterion (27) for a succession of tentative values $\lambda^{(1)} < \ldots < \lambda^{(M)}$ with a sufficiently small step, just as when the usual AIC is applicable.

We shall see in the next Section that the alternative criterion (11) allows, at least in principle, to find $\hat{\lambda}$ along with $\hat{\mathbf{c}}_{\hat{\lambda}}(\mathbf{y})$ as result of a joint iterative procedure, and to avoid thereby discrete enumeration of tentative values of λ.

5 The principle of maximum fit to the actual joint distribution of the observed variable and hidden parameter

Criterion. The principle (11) does not lend itself to numerical realization, either, not only for the reason that the joint distribution $H^*(\mathbf{c}, \mathbf{y} \,|\, \lambda)$ is unknown, but also because the random parameter $\mathbf{c}$ is hidden from observation. We resort to the same trick as in the previous Section - imagine the existence of an independent sample $\tilde{\mathbf{y}}$ and replace $\ln H(\mathbf{c},\mathbf{y}\,|\,\lambda)$ by the mathematical expectation of $\ln H\big(\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}),\mathbf{y}\,|\,\lambda\big)$:

$$\hat{\lambda} = \arg\max \int\int\bigg\{\int\Big[\ln H\big(\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}),\mathbf{y}\,|\,\lambda\big)\Big]\varPhi(\tilde{\mathbf{y}}\,|\,\mathbf{c})d\tilde{\mathbf{y}}\bigg\}H^*(\mathbf{c},\mathbf{y})d\mathbf{c}d\mathbf{y}.$$

Here $\ln H\big(\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}),\mathbf{y}\mid\lambda\big) = \ln\varPhi\big(\mathbf{y},\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})\big) + \ln\varPsi\big(\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})\mid\lambda\big)$ and $H^*(\mathbf{c}\mid\mathbf{y}) = \varPhi(\mathbf{y}\,|\,\mathbf{c})\varPsi^*(\mathbf{c})$. We obtain the criterion

$$\begin{gathered}\hat{\lambda} = \arg\max_{\lambda}\int\Bigg\{\int\bigg\{\int\Big[\ln\varPhi\big(\mathbf{y},\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})\big)+ \\ \ln\varPsi\big(\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})\,|\,\lambda\big)\Big]\varPhi(\tilde{\mathbf{y}}\,|\,\mathbf{c})\bigg\}\varPhi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}\Bigg\}\varPsi^*(\mathbf{c})d\mathbf{c}.\end{gathered} \tag{28}$$

which differs from (25) only by the presence of the additional summand $\ln\varPsi\big(\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})\,|\,\lambda\big)$.

Proposition 2. *Under the assumptions (13) and (15),*

$$\begin{gathered}\int\Bigg\{\int\bigg\{\int\Big[\ln\varPhi\big(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})\big) + \ln\varPsi\big(\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})\,|\,\lambda\big)\Big]\varPhi(\tilde{\mathbf{y}}\,|\,\mathbf{c})\bigg\}\varPhi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y}\Bigg\}\varPsi^*(\mathbf{c})d\mathbf{c} = \\ \int J_2(\lambda\,|\,\mathbf{y})F^*(\mathbf{y})d\mathbf{y}, \\ J_2(\lambda\,|\,\mathbf{y}) = \ln\varPhi\big(\mathbf{y}\,|\,\hat{\mathbf{c}}_\lambda(\mathbf{y})\big) + \ln\varPsi\big(\hat{\mathbf{c}}_\lambda(\mathbf{y})\,|\,\lambda\big) - Tr\Big[\mathbf{A}(\mathbf{A}+\mathbf{B}_\lambda)^{-1}\Big].\end{gathered} \tag{29}$$

Proof of this Proposition is based on the same reasons as that of Proposition 1. The equality (29) shows that function $J_2(\lambda\,|\,\mathbf{y})$ is an unbiased estimate of the criterion (28). Its immediate maximization yields another version of the continuous AIC:

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

Joint criterion and iterative optimization algorithm. It is clear that $\hat{\lambda}$ found by criterion (30) and the respective Bayesian estimate $\hat{\mathbf{c}}_{\hat{\lambda}}(\mathbf{y})$ satisfy the joint optimization condition

$$\begin{gathered}(\hat{\lambda},\hat{\mathbf{c}}_{\hat{\lambda}}) = \arg\max Q(\lambda,\mathbf{c}), \\ Q(\lambda,\mathbf{c}) = \ln\varPhi(\mathbf{y}\,|\,\mathbf{c}) + \ln\varPsi(\mathbf{c}\,|\,\lambda) - Tr\Big[\mathbf{A}(\mathbf{A}+\mathbf{B}_\lambda)^{-1}\Big].\end{gathered} \tag{31}$$

The sum $\Big\{\ln\varPhi(\mathbf{y}\mid\mathbf{c}) + \ln\varPsi(\mathbf{c}\mid\lambda)\Big\}$ is concave function of $\mathbf{c}$ by assumption. If, in addition, the difference $\Big\{\ln\varPsi(\mathbf{c}\mid\lambda) - Tr\Big[\mathbf{A}(\mathbf{A}+\mathbf{B}_\lambda)^{-1}\Big]\Big\}$ is unimodal function of λ, having the only maximum point $(d/d\lambda)\Big\{\ln\varPsi(\mathbf{c}\,|\,\lambda) - Tr\Big[\mathbf{A}(\mathbf{A}+\mathbf{B}_\lambda)^{-1}\Big]\Big\} = 0$, which fact depends on the structure of matrix function $\mathbf{B}_\lambda$, the criterion $Q(\lambda,\mathbf{c})$ will have the only maximum point with respect to the joint variable $(\lambda,\mathbf{c})\in\mathbb{R}^{m+1}$.

6 A particular case: The classical AIC

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

$$\Psi(\mathbf{c}\,|\,\lambda) = \textstyle\prod_{i=1}^{\lambda} \psi_i(c_i\,|\,\lambda),\ \psi_i(c_i\,|\,\lambda) = \mathcal{N}(c_i\,|\,0, \sigma^2),\ \sigma \to \infty,$$
$$\Psi(\mathbf{c}\,|\,\lambda) \cong const = 0, \quad \ln \Psi(\mathbf{c}\,|\,\lambda) \cong const < 0.$$

Since only the first part of the vector parameter is free in the conditional density $\Phi(\mathbf{y} \mid \mathbf{c}_\lambda,\, \mathbf{c}_{m-\lambda})$, the Hessian $\mathbf{A}_\lambda = \bigtriangledown^2_{\mathbf{c}_\lambda \mathbf{c}_\lambda} \ln \Phi(\mathbf{y} \mid \mathbf{c}_\lambda, \mathbf{0})$ is a matrix $(\lambda \times \lambda)$. Under these assumptions, both versions of the continuous AIC (27) and (30) reduce to the criterion (3):

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

7 Time-varying regression estimation

In the problem of time-varying regression estimation (5)-(6), the Bayesian estimate of the hidden sequence of regression coefficients $\mathbf{c} = (\mathbf{c}_1^T \cdots \mathbf{c}_N^T)^T \in \mathbb{R}^{kN}$ 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 $\mathbf{X} = (\mathbf{X}_{ts},\, t, s = 1, \ldots, N)$ of total dimension $(kN \times N)$ with diagonal column-blocks $\mathbf{X}_{tt} = (\mathbf{x}_t, t = 1, ..., N)$ $(k \times 1)$ and nondiagonal blocks $\mathbf{X}_{ts} = \mathbf{0}$ $(k \times 1)$, $t \neq s$.

Then, for the observation noise variance conventionally taken as equal to unity $\delta = 1$, the observation model (5) will produce the likelihood function $\ln \Phi(\mathbf{y}\,|\,\mathbf{c}) = \ln \mathcal{N}(\mathbf{y}\,|\,\mathbf{X}^T\mathbf{c}, \mathbf{I})$

$$\ln \Phi(\mathbf{y}\,|\,\mathbf{c}) = const - (1/2)\Big((\mathbf{y} - \mathbf{X}^T\mathbf{c})^T \mathbf{A} (\mathbf{y} - \mathbf{X}^T\mathbf{c})\Big). \tag{32}$$

The negative semidefinite Hessian $\mathbf{A} = -\mathbf{X}\mathbf{X}^T$ $(kN \times kN)$ is block-diagonal matrix with diagonal blocks $\mathbf{A}_{tt} = \mathbf{x}_t\mathbf{x}_t^T (k \times k)$, $t = 1, \ldots, N$. It is always degenerate and, if the regressors $\big[(x_{it}, t = 1, \ldots, N),\, i = 1, \ldots, k\big]$ are linearly independent, has the maximum rank $Rank(\mathbf{A}) = N$. With $\delta = 1$, the hidden Markov model of regression coefficients (6) is expressed by the family of a priori densities

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

It is assumed that no a priori information on the first vector of regression coefficients is available, that is $\rho \to \infty$ and $\mathcal{N}(\mathbf{c}_1 \,|\, \mathbf{0}, \rho \mathbf{I}) \to const_{\mathbf{c}_1} \equiv 0$. In logarithmic form, for sufficiently large but finite ρ, we have

$$\begin{aligned}
&\ln \Psi(\mathbf{c} \,|\, \lambda) \cong const + (1/2)k(N-1)\ln(1/\lambda) - (1/2)(1/\lambda)\mathbf{c}^T\mathbf{B}\mathbf{c} = \\
&const + (1/2)k(N-1)\ln(1/\lambda) + (1/2)\mathbf{c}^T\mathbf{B}_\lambda\mathbf{c}, \qquad (33)\\
&\mathbf{B}_\lambda = -(1/\lambda)\mathbf{B}.
\end{aligned}$$

The structure of the negative semidefinite Hessian $\mathbf{B}_\lambda$ is defined by the degenerate square block-tridiagonal matrix $\mathbf{B}$ $(kN \times kN)$ with the diagonal $(\mathbf{I}, 2\mathbf{I}, \dots, 2\mathbf{I}, \mathbf{I})$ and two off-diagonals $(-\mathbf{I}, \dots, -\mathbf{I})$ formed by identity matrices $\mathbf{I}$ $(k \times k)$.

The Bayesian estimate of regression coefficients $\hat{\mathbf{c}}_\lambda(\mathbf{y}, \mathbf{X}) = \big(\mathbf{c}_{1,\lambda}(\mathbf{y}, \mathbf{X}), \dots, \mathbf{c}_{k,\lambda}(\mathbf{y}, \mathbf{X})\big)$ is provided via minimization of the Flexible Least Squares criterion

$$\hat{\mathbf{c}}_\lambda(\mathbf{y}, \mathbf{X}) = \arg\min\Big\{\sum\nolimits_{t=1}^{N} \big(y_t - \mathbf{x}_t^T\mathbf{c}_t\big)^2 + (1/\lambda)\sum\nolimits_{t=2}^{N}(\mathbf{c}_t - \mathbf{c}_{t-1})^T(\mathbf{c}_t - \mathbf{c}_{t-1})\Big\}$$

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

In accordance with notations accepted in (32) and (33), the penalty term in both criteria (27) and (30) will have the form

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

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

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

So, to compute the penalty term in the criteria (27) and (30), it is enough, instead of full inverting the sum of matrices for each tentative value of λ, to compute the diagonal blocks of inversion. This can be easily done by a slight modification of the double-sweep method.

8 Ground-truth experiments

We analyzed 200 independent realizations of the random process (5) of length $N = 50$ with three regressors $(x_{it}, t = 1, \dots, N)$, $i = 1, \dots, k$, $k = 3$, generated as Markov model $c_{it}^* = c_{it-1}^* + \xi_{it}, i = 1, \dots, k,\ t = 1, \dots, N$, and the sinusoidal "actual" sequences of coefficients $c_{it}^* = 4\sin\big((2\pi/N)t + (2\pi/3)(i-1)\big)$ mutually shifted by phase, and 10% noise variance $\delta = 0.1\Big((1/N)\sum_{i=1}^{N}(\mathbf{x}_t^T\mathbf{c}_t)^2\Big)$.

It was assumed that no a priori information on the first vector of regression coefficients is available $\rho \to \infty$. The dependence of the "efficient dimension" of the regression coefficient sequence $(\mathbf{c}_1 \cdots \mathbf{c}_, i = N)$ on the assumed time-variability λ of regression coefficients computed for one realization of the random regressor sequence is shown in Fig. 1. This dimension equals the number of regressors in the case of zero variability $\lambda \to 0$ and approaches the length of the time series if $\lambda \to \infty$.

For each of 200 simulated time series, three values of the time-variability parameter were computed using, first, two versions of the continuous Akaike criterion (27) and (30), and, second, the traditional leave-one-out cross validation technique [7]. Then, we applied each of the chosen values to the remaining 199 time series as the control set, and compared the estimated and the ground-truth sequence of the regression coefficient $(\mathbf{c}_1^* \cdots \mathbf{c}_N^*)$ and $(\hat{\mathbf{c}}_{1,\hat{\lambda}} \cdots \hat{\mathbf{c}}_{N,\hat{\lambda}})$ by the criterion

$$\varepsilon_{\hat{\lambda}} = \sum_{t=1}^{N} (\hat{\mathbf{c}}_{t,\hat{\lambda}} - \mathbf{c}_t^*)^T (\hat{\mathbf{c}}_{t,\hat{\lambda}} - \mathbf{c}_t^*) \Big/ \sum\nolimits_{t=1}^{N} (\mathbf{c}_t^*)^T \mathbf{c}_t^*$$

.

We obtained the following results:

Criteria	Mean value $\varepsilon_{\hat{\lambda}}$	
	Sinusoidal sequence of coefficients $c_{it}^* = 4\sin\big((2\pi/N)t + \varphi_i\big)$	Coefficients generated as Markov model $c_{it}^* = c_{it-1}^* + \xi_{it}$
Continious Generalization of AIC (1)	0.02	0.016
Continious Generalization of AIC (2)	0.004	0.013
Leave-one-out cross validation	0.003	0.046

As is seen, in the case of sinusoidal sequence of coefficients leave-one-out cross validation principle has demonstrated the best results. It may be explain that formally sinusoidal sequence of coefficients don't satisfied Markov model, which was used as the main assumption by forming AIC. If "actual" coefficients are satisfied this model, then a continious generalization of AIC recommended more appropriate value of λ. The main advantages of AIC comparing with leave-one-out cross validation principle are firstly, having strict one extremum, secondly, the capability to recieve more adequate estimation of parameter, thirdly, at the same time, the continuous AIC is incomparably more preferable from the computational viewpoint.

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.

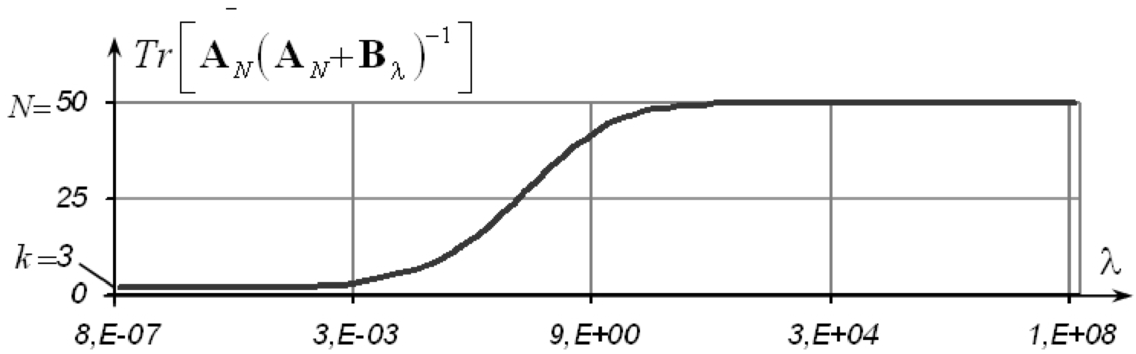

Fig. 1. The efficient dimension of regression coefficient sequence as function in the logarithmic scale)

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. Bishop C.M. *Pattern Recognition and Machine Learning.* Springer, 2006.