УДК 004.9311

# ОБОБЩЕНИЕ ИНФОРМАЦИОННОГО КРИТЕРИЯ АКАИКЕ ДЛЯ ВЫБОРА ЗНАЧЕНИЙ НЕПРЕРЫВНЫХ ПАРАМЕТРОВ В МОДЕЛЯХ ДАННЫХ

**© Ежова Е.О., Моттль В.В., Красоткина О.В.**

Московский Физико-Технический институт
Институтский пер, 9, г. Долгопрудный, 141700, Россия
e-mail: *lena-ezhova@rambler.ru*

ВЦ РАН
ул Вавилова, 40, г. Москва, 117967, Россия
e-mail: *vmottl@yandex.ru*

ТулГУ
пр-т Ленина, 92, г. Тула, 300600, Россия
e-mail: *krasotkina@uic.tula.ru*

**Abstract**. The crucial restriction of the Akaike Information Criterion (AIC) as means of adjusting a model to the given data set within a succession of nested parametric model classes is the assumption that the classes are rigidly defined by the growing dimension of an unknown vector parameter. We extend the Kullback information maximization principle underlying the classical AIC 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 class of continuously nested a priori probability distributions We illustrate the proposed continuous generalization of AIC by its application to the problem of time-varying regression estimation which implies the inevitable necessity to choose the time-variability of regression coefficients treated a nonstationary model of the given signal.

## Введение

*Широко используемый в современном анализе данных информационный критерий Акаике (AIC)* [1] *является простым и эффективным способом выбора наиболее адекватного класса модели из упорядоченного дискретного множества вложенных классов моделей.*

В классической постановке критерия обычно рассматривается выборка $\mathbf{y} = (y_j, j = 1, \dots, N)$ независимых случайных величин с неизвестной плотностью распределения $\varphi^*(y)$, принадлежащей некоторому параметрическому семейству $\varphi(y \mid \mathbf{c})$, $\mathbf{c} \in \mathbb{R}^m$. Часто размерность вектора параметров $m$ оказывается очень большой и существенно превосходит размер обучающей выборки $N$, что делает бессмысленным применение для оценивания вектора параметров $\mathbf{c}$ принципа максимального правдоподобия.

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

Если же предположить, что элементы вектора $\mathbf{c}$ обладают естественной упорядоченностью по степени значимости и при этом $c_i = 0$, $n < i \le m$:

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

то это позволит нам рассмотреть параметрическое семейство $\Phi(\mathbf{y}\,|\,\mathbf{c})$ как последовательность вложенных классов моделей $\Phi\big(\mathbf{y}\mid\mathbf{c}=(\mathbf{c}_n\,|\,\mathbf{0})\big)$, $\mathbb{R}^{n_{min}}\subset\cdots\subset\mathbb{R}^{n_{max}}$.

*Критерий AIC в классической постановке является способом оценивания подходящей размерности вектора параметров, как меры сложности модели* $\hat{n}=\arg\max_n\Big[\ln\Phi\big(\mathbf{y}\,|\,\mathbf{c}_n(\mathbf{y})\big)-n\Big]$. Однако это формула получена в предположении, что гессиан $\nabla^2_{\mathbf{c}_n\mathbf{c}_n}\ln\Phi(\mathbf{y}\,|\,\mathbf{c}_n,\mathbf{0})$ в точке максимального правдоподобия имеет полный ранг, а значит и оценка $\hat{\mathbf{c}}_n(\mathbf{y})$ — единственная. В более общем случае заменим штраф $n$ на ранг матрицы

$$\hat{n}=\arg\max_n\Big\{\ln\Phi\big(\mathbf{y}\,|\,\mathbf{c}_n(\mathbf{y}),\mathbf{0}\big)-\mathrm{rank}\big[\nabla^2_{\mathbf{c}_n\mathbf{c}_n}\ln\Phi(\mathbf{y}\,|\,\mathbf{c}_n,\mathbf{0})\big]\Big\} \qquad (3)$$

В основе классического AIC лежит принцип максимизации информации по Кульбаку между моделью плотности распределения и настоящей гипотетической плотностью распределения.

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

есть желаемая размерность в предположении, что $\Phi^*(\mathbf{y})=\Phi(\mathbf{y}\mid\mathbf{c}^*_{n^*})$ с некоторым значением $(\mathbf{c}^*_{n^*}\mathbf{0})$, вырезанного из неизвестного $\mathbf{c}^*=(c^*_1,\ldots,c^*_m)$

*Одним из первых применений AIC было моделирование нестационарного сигнала на дискретной временной оси, разделенной на на неизвестное количество n интервальных блоков, и проверка локальной стационарности модели авторегрессии с фиксированным порядком k на каждом их них* [2].

*Со времен публикации первой статьи Акаике было предложено много модификаций этого критерия* [3, 4, 5, 6]. *Среди них Байесовский информационный критерий (BIC)* [3] *нашел более широкое применение. Однако все они были нацелены на выбор размерности вектора параметров для случая известной упорядоченности его элементов по степени значимости.*

*В данной работе предлагается совершенно новое обобщение критерия Акаике, которое было вызвано необходимостью анализа нестационарного сигнала* $(\mathbf{y},\mathbf{x})=\big((y_t,\mathbf{x}_t),t=1,\ldots,N\big)$,*регрессионная модель которого*

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

*меняется на интервале наблюдения.* Очевидно, что при этом размерность вектора параметров в семействе условных плотностей распределения $\Phi(\mathbf{y}\mid\mathbf{x},\mathbf{c})$ оказывается фиксированной $\mathbf{c}=(\mathbf{c}_1\ldots\mathbf{c}_N)\in\mathbb{R}^{kN}$ и в $k$ раз превосходит количество наблюдений. Вместо этого *предполагается, что искомая последовательность коэффициентов представляет собой случайный марковский процесс*

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

*который начинается с неизвестного первого значения* $\mathbf{c}_1\sim N(\mathbf{c}_1\,|\,\mathbf{0},\rho\mathbf{I})$, $\rho\to\infty$. *Параметр дисперсии шума* $\lambda$ *является структурным параметром априорной модели и отвечает за степень временной нестационарности коэффициентов регрессии.*

Это типичный пример задачи, в которой плавное изменение параметра $\lambda$ определяет систему непрерывно вложенных априорных плотностей распределения $\Psi(\mathbf{c}\,|\,\lambda)$ вектора параметров модели, начиная от «однородного» распределения в $\mathbb{R}^k$ при $\lambda = 0$ до «однородного» распределения в $\mathbb{R}^{kN}$ при $\lambda \to \infty$. Такая ситуация фактически представляет собой введение вместо дискретной последовательности целочисленных размерностей понятия «размытой размерности» вектора параметров $\mathbf{c}$, непрерывно меняющейся от $k$ до $kN$ при увеличении параметра $\lambda$.*Естественно, что классический критерий AIC оказывается неприменимым для выбора наиболее подходящего для данного сигнала* $(\mathbf{y},\mathbf{x})$ *значения параметра* $0 < \lambda < \infty$.

В этой статье мы рассматриваем параметрическую модель плотности распределения неизвестной генеральной совокупности $F^*(\mathbf{y})$ как смесь условной плотности из заданного семейства $\Phi(\mathbf{y}\,|\,\mathbf{c})$, $\mathbf{c} \in \mathbb{R}^m$ и априорной плотности распределения вектора параметров $\Psi(\mathbf{c}\,|\,\lambda)$:

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

*Значение структурного параметра модели* $\lambda$, *оцененное по наблюдаемой выборке* $\mathbf{y}$, *обеспечивает оптимальную степень сокращения слишком большой размерности вектора параметров* $\mathbf{c}$. Как только значение $\lambda$ выбрано, результат анализа представляет собой байесовскую оценку вектора параметров $\mathbf{c}$

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

Мы будем эксплуатировать ту же идею, что и в (4), т.е. будем с помощью варьирования параметра $\lambda$ пытаться обеспечить наилучшее приближение модельного распределения $F(\mathbf{y}\mid\lambda)$ (15) и неизвестного распределения генеральной совокупности $F^*(\mathbf{y})$

*Мы предлагаем в этой статье два способа обобщения критерия AIC, а также покажем, что классический AIC может быть получен как частный случай обоих способов при принятии специальных предположений об априорной плотности* $\Psi(\mathbf{c}\,|\,\lambda)$.

*Наконец, мы опишем результаты модельного эксперимента, сравнивающего предложенный обобщенный информационный критерий оценки непрерывного структурного параметра модели с критерием скользящего контроля на задаче анализа нестационарного сигнала.*

## 1. Два способа вычисления расстояния по Кульбаку между неизвестной истинной плотностью распределения и ее моделью

С одной стороны, идея максимизации информации по Кульбаку о распределении выборки значений наблюдаемой переменной $F^*(\mathbf{y})$, содержащейся в модельной плотности $F(\mathbf{y}\,|\,\lambda)$, есть математическое ожидание: $\int\left[\ln F(\mathbf{y}\,|\,\lambda)\right]F^*(\mathbf{y})d\mathbf{y}$. Идея максимизации этой информации путем выбора подходящего значения $\lambda$ приводит к критерию:

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

Этот способ выбора параметра $\lambda$ подходит для любого вида истинной плотности распределения $F^*(\mathbf{y})$.

С другой стороны, рассматриваемая нами модель (15) включает в себя произвольный параметр $\mathbf{c}$, как скрытую переменную. Таким образом, мы можем выбирать модель совместной плотности распределения скрытой и наблюдаемой переменной $H(\mathbf{c},\mathbf{y} \mid \lambda) = \Psi(\mathbf{c} \mid \lambda)\Phi(\mathbf{y} \mid \mathbf{c})$ наиболее близкую к истинному распределению $H^*(\mathbf{c},\mathbf{y})$.Этот способ имеет смысл, только если неизвестная истинная плотность распределения $F^*(\mathbf{y})$ согласуется с принятым параметрическим семейством распределений $\Phi(\mathbf{y}\,|\,\mathbf{c})$, то есть существует такое распределение $\Psi^*(\mathbf{c})$, что выполняется

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

Тогда $H^*(\mathbf{c},\mathbf{y}) = \Psi^*(\mathbf{c})\Phi(\mathbf{y}\,|\,\mathbf{c})$, и «идеальный» критерий выбора $\lambda$ принимает вид

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

Мы увидим, что формализации (9) и (11) приведут к существенно различным обобщениям классического информационного критерия Акаике для выбора значения непрерывного структурного параметра.

## 2. Основные предположения и некоторые свойства семейств параметрических плотностей

**Предположения.** Мы ограничимся здесь рассмотрением случая параметрических семейств плотностей распределения $\varphi(y \mid \mathbf{c})$, для которых логарифмическая функция правдоподобия $\ln\Phi(\mathbf{y} \mid \mathbf{c})$ ассимптотически квадратична в окрестности оценки максимальногоправдоподобия $\mathbf{c}$ , т.е. для достаточно большого размера $N$ выборки $\mathbf{y} = (y_j, j = 1, \ldots, N)$ можно считать, что

$$\ln\Phi(\mathbf{y}\,|\,\mathbf{c}) = \ln\Phi(\mathbf{y}\,|\,\hat{\mathbf{c}}(\mathbf{y})) + (1/2)(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y}))^T\mathbf{A}_N(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})),\ \bigtriangledown_{\mathbf{c}}\log\Phi(\mathbf{y}\,|\,\mathbf{c}) = A_N(\mathbf{c}-\hat{\mathbf{c}}(\mathbf{y})). \tag{12}$$

Причем гессиан $\mathbf{A}_N = \bigtriangledown^2_{\mathbf{cc}}\ln\Phi(\mathbf{y}\,|\,\mathbf{c})$, называемый информационной матрицей Фишера, не зависит от точки $\mathbf{c}$, в которой определен.

Рассмотрим теперь семейство плотностей априорного распределения скрытой переменной $\Psi(\mathbf{c}\,|\,\lambda)$. Будем полагать, что каждая из этих плотностей является нормальной, возможно вырожденной, с нулевым математическим ожиданием и ковариационной матрицей, определяемой значением структурного параметра $\lambda$. Это приводит к тому, что логарифмическая функция правдоподобия $\ln\Psi(\mathbf{c}\,|\,\lambda)$ есть квадратичная функция, достигающая своего максимального значения в нуле $\bigtriangledown_{\mathbf{c}}\log\Psi(\mathbf{0}\,|\,\lambda) = \mathbf{0}$ и определяемая своим Гессианом $\mathbf{B}_\lambda = \bigtriangledown^2_{\mathbf{cc}}\ln\Psi(\mathbf{c}\,|\,\lambda)$, так что

$$\ln\Psi(\mathbf{c}\,|\,\lambda) = const + (1/2)\mathbf{c}^T\mathbf{B}_\lambda\mathbf{c}. \tag{13}$$

Что касается неизвестной плотности распределения выходной переменной $F^*(\mathbf{y})$, то мы будем предполагать, что оно согласуется с семейством плотностей $\Phi(\mathbf{y} \mid \mathbf{c})$ в том смысле, что существует неизвестная плотность $\Psi^*(\mathbf{c})$, которая допускает представление

$$F^*(\mathbf{y}) = \int \Phi(\mathbf{y}\,|\,\mathbf{c})\Psi^*(\mathbf{c})d\mathbf{c} \tag{14}$$

**Свойства.** Рассмотрим произвольную выборку $\mathbf{y}$, порожденную вероятностным распределением $\Phi(\mathbf{y} \mid \mathbf{c})$ с некоторым фиксированным значением параметра $\mathbf{c}$. Хорошо известно, что для гораздо более широкого класса условных распределений, чем те, что описаны выше (12), таких, что если $\mathbf{A}_N$ есть матрица с полным рангом $\mathrm{rank}(\mathbf{A}_N) = n$, произвольная оценка максимума правдободия $\hat{\mathbf{c}}(\mathbf{y})$оказывается несмещенной

$$\int \hat{\mathbf{c}}(\mathbf{y})\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y} = \mathbf{c}, \tag{15}$$

а ее условная ковариационная матрица полностью определяется информационной матрицей Фишера:

$$\int \big(\hat{\mathbf{c}}(\mathbf{y}) - \mathbf{c}\big)\big(\hat{\mathbf{c}}(\mathbf{y}) - \mathbf{c}\big)^T \Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y} = -\mathbf{A}_N^{-1}. \tag{16}$$

В более общем случае, когда $\mathrm{rank}(\mathbf{A}_N) < n$, (22) и (16) имеют вид:

$$\int \mathbf{A}_N\big(\hat{\mathbf{c}}(\mathbf{y}) - \mathbf{c}\big)\big(\hat{\mathbf{c}}(\mathbf{y}) - \mathbf{c}\big)^T \Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y} = \mathbf{0}, \tag{17}$$

$$\int \Big[\mathbf{A}_N\big(\hat{\mathbf{c}}(\mathbf{y}) - \mathbf{c}\big)\Big]\Big[\mathbf{A}_N\big(\hat{\mathbf{c}}(\mathbf{y}) - \mathbf{c}\big)\Big]^T \Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y} = -\mathbf{A}_N \tag{18}$$

Если условия (12) и (20) выполнены, то произвольная Байессовская оценка (8) есть линейная функция от оценки максимального правдоподобия

$$\hat{\mathbf{c}}_\lambda(\mathbf{y}) = (\mathbf{A}_N + \mathbf{B}_\lambda)^{-1}\mathbf{A}_N\hat{\mathbf{c}}(\mathbf{y}) \tag{19}$$

с условной ковариационной матрицей относительно фиксированного значения параметра $\mathbf{c}$

$$\int \big(\hat{\mathbf{c}}_\lambda(\mathbf{y}) - \hat{\mathbf{c}}_\lambda(\mathbf{c})\big)\big(\hat{\mathbf{c}}_\lambda(\mathbf{y}) - \hat{\mathbf{c}}_\lambda(\mathbf{c})\big)^T \Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y} = -(\mathbf{A}_N + \mathbf{B}_\lambda)^{-1}\mathbf{A}_N(\mathbf{A}_N + \mathbf{B}_\lambda)^{-1}, \tag{20}$$

где $\hat{\mathbf{c}}_\lambda(\mathbf{c})$ есть условное матеметическое ожидание

$$\hat{\mathbf{c}}_\lambda(\mathbf{c}) = \int \hat{\mathbf{c}}_\lambda(\mathbf{y})\Phi(\mathbf{y}\,|\,\mathbf{c})d\mathbf{y} = (\mathbf{A}_N + \mathbf{B}_\lambda)^{-1}\mathbf{A}_N\hat{\mathbf{c}}. \tag{21}$$

## 3. Критерий максимума информации о распределении наблюдаемой переменной

Непосредственная реализация критерия (9) невозможна хотя бы потому, что истинное распределение $F^*(\mathbf{y})$ неизвестно. Максимизация функции правдободобия по одной доступной реализации $\ln F(\mathbf{y} \mid \lambda)$, как несмещенной оценки критерия, также бессмысленно, так как при этом будут предпочтительны значения структурного параметра, приводящие к слишком большим размерностям $\mathbf{c} \in \mathbb{R}^m$.

Для того чтобы преодолеть «проклятие единственной выборки», мы применим идею компромисса, обосновывающего классический информационный критерий Акаике ([1]), а именно вообразим существование другой независимой выборки $\tilde{\mathbf{y}}$. Пусть по ней получена произвольная байессовская оценка $\hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})$ (8). Заменим $\ln F(\mathbf{y} \mid \lambda)$ в (9) на математическое ожидание $\ln \Phi(\mathbf{y} \,|\, \hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))$:

$$\hat{\lambda} = \arg\max_\lambda \int \Big\{ \int \Big\{ \int \big[\ln \Phi(\mathbf{y} \,|\, \hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}}))\big] \Phi(\tilde{\mathbf{y}} \mid \mathbf{c}) d\tilde{\mathbf{y}} \Big\} \Phi(\mathbf{y} \,|\, \mathbf{c}) d\mathbf{y} \Big\} \Psi^*(\mathbf{c}) d\mathbf{c}. \quad (22)$$

**Предложение 1.** При предположениях (12) и (20),

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

$$\int J_1(\lambda \,|\, \mathbf{y}) F^*(\mathbf{y}) d\mathbf{y}, \quad (23)$$

$$J_1(\lambda \,|\, \mathbf{y}) F^*(\mathbf{y}) = \ln \Phi\big(\mathbf{y} \,|\, \hat{\mathbf{c}}_\lambda(\mathbf{y})\big) - Tr\Big[\mathbf{A}_N (\mathbf{A}_N + \mathbf{B}_\lambda)^{-1}\Big].$$

*Доказательство.* основано на квадратичном представлении $\ln \Phi(\mathbf{y} \mid \mathbf{c})$ (12) в $\mathbf{c} = \hat{\mathbf{c}}_\lambda(\tilde{\mathbf{y}})$ в выражениях (16)— (20).

Эта теорема указывает на построение непрерывного аналога классического AIC. Хотя распределение $\Psi^*(\mathbf{c})$ в (12) по-прежнему неизвестно, а значит непосредственно применить критерий (22) невозможно, но выражение (23) дает легко вычисляемую функцию $J_1(\lambda \mid \mathbf{y})$, которая является несмещенной оценкой полного критерия. Аналогично рассуждениям Акаике, эту функцию можно также максимизировать по искомому значению структурного параметра:

$$\hat{\lambda} = \arg\max_\lambda \Big\{ \ln \Phi\big(\mathbf{y} \,|\, \hat{\mathbf{c}}_\lambda(\mathbf{y})\big) - Tr\Big[\mathbf{A}_N (\mathbf{A}_N + \mathbf{B}_\lambda)^{-1}\Big] \Big\}. \quad (24)$$

Это и есть обобщенный информационный критерий Акаике (6). Сравнение критериев (24) и (6) позволяет интерпретировать штрафной член $Tr\Big[\mathbf{A}_N (\mathbf{A}_N + \mathbf{B}_\lambda)^{-1}\Big]$, как условную «размытую размерность» параметра $\mathbf{c}$, выбор которого ограничено распределением $\ln \Psi(\mathbf{c} \mid \lambda)$. □

## 4. Критерий максимума информации о совместном распределениии наблюдаемой переменной и скрытого параметра

Критерий (11) также невозможно вычислительно реализовать, не только по тому, что совместное распределение $H^*(\mathbf{c},\mathbf{y} \mid \lambda)$ неизвестно, но также из-за того, что призвольный параметр $\mathbf{c}$ скрыт от наблюдателя. Как и в предыдущем разделе мы применим компромисс, заключающийся в использовании независимой воображаемой выборки $\tilde{\mathbf{y}}$ и замене $\ln H(\mathbf{c},\mathbf{y} \mid \lambda)$ на математическое ожидание $\ln H\big(\hat{\mathbf{c}}(\tilde{\mathbf{y}}),\mathbf{y} \,|\, \lambda\big)$:

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

Здесь $\ln H\big(\hat{\mathbf{c}}(\tilde{\mathbf{y}}),\mathbf{y}\,|\,\lambda\big) = \ln\Phi\big(\mathbf{y},\hat{\mathbf{c}}(\tilde{\mathbf{y}})\big) + \ln\Psi\big(\hat{\mathbf{c}}(\tilde{\mathbf{y}})\,|\,\lambda\big)$ и $H^*(\mathbf{c}\,|\,\mathbf{y}) = \Phi(\mathbf{y}\,|\,\mathbf{c})\Psi^*(\mathbf{c})$. Мы получаем критерий

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

который отличается от (22) только наличием дополнительного слагаемого $\Psi\big(\hat{\mathbf{c}}(\tilde{\mathbf{y}})\,|\,\lambda\big)$.

**Предложение 2.** При принятых предположениях (12) и (20),

$$\begin{gathered}\int\left\{\int\left\{\int\Big[\ln\Phi\big(\mathbf{y}\,|\,\hat{\mathbf{c}}(\tilde{\mathbf{y}})\big) + \ln\Psi\big(\hat{\mathbf{c}}(\tilde{\mathbf{y}})\,|\,\lambda\big)\Big]\Phi(\tilde{\mathbf{y}}\,|\,\mathbf{c})\right\}\Phi(\mathbf{y}\mid\mathbf{c})d\mathbf{y}\right\}\Psi^*(\mathbf{c})d\mathbf{c} = \\ \int J_2(\lambda\,|\,\mathbf{y})F^*(\mathbf{y})d\mathbf{y}, \\ J_2(\lambda\,|\,\mathbf{y}) = \ln\Phi\big(\mathbf{y},\hat{\mathbf{c}}(\mathbf{y})\big) + \ln\Psi\big(\hat{\mathbf{c}}(\mathbf{y})\,|\,\lambda\big) - Tr\Big[\mathbf{A}_N(\mathbf{A}_N+\mathbf{B}_\lambda)^{-1}\Big].\end{gathered} \quad (26)$$

*Доказательство.* этого утверждения основано на таких же рассуждениях, которые были сделаны для Предложения 1.

Выражение (26) показывает, что функция $J_2(\lambda\mid\mathbf{y})$ есть несмещенная оценка критерия (25). Его непосредственная максимизация и есть другая версия обобщения классического AIC

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

□

## 5. Частный случай: классический информационный критерий Акаике

Пусть структурный параметр принимает целые положительные числа $0 \le \lambda \le m$ и урезает вектор параметров с упорядоченными элементами $\mathbf{c} = (\mathbf{c}_\lambda,\mathbf{c}_{m-\lambda}) \in \mathbb{R}$, так же как и в (5) с $n = \lambda$, то есть $\mathbf{c}_\lambda \in \mathbb{R}^\lambda$, $\mathbf{c}_{m-\lambda} \in \mathbb{R}^{m-\lambda}$. Никакой априорной информации о векторе $\mathbf{c}$ нет, то есть

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

Так как только первая часть вектора параметров входит в условную плотность $\Phi(\mathbf{y}\,|\,\mathbf{c}_\lambda,\mathbf{c}_{m-\lambda})$, то Гессиан $\mathbf{A}_{N,\lambda} = \bigtriangledown^2_{\mathbf{c}_\lambda\mathbf{c}_\lambda}\ln\Phi(\mathbf{y}\mid\mathbf{c}_\lambda,\mathbf{0})$ есть матрица $(\lambda\times\lambda)$.

При принятых предположениях, обе версии обобщения AIC (24) и (27) приводят к критерию (6):

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

## 6. Применение критерия Акаике в задаче оценивания нестационарной регрессии: модельные эксперименты

В задаче оценки нестационарной регрессии (5)–(6) сама байессовская оценка скрытой последовательности коэффициентов регрессии $\mathbf{c} = (\mathbf{c}_1^T \cdots \mathbf{c}_N^T)^T \in \mathbb{R}^{kN}$ зависит только от отношения $\lambda$ предполагаемых дисперсий шума в уравнениях наблюдения (5) и состояния (6), в том время ее статистические свойства существенно определяется дисперсией шума в модели наблюдения. Байесовская оценка вектора параметров $\mathbf{c}$ может быть получена минимизацией критерия Flexible Least Squares

$$\sum\nolimits_{t=1}^{N} \left(y_t - \mathbf{x}_t^T \mathbf{c}_t\right)^2 + (1/\lambda) \sum\nolimits_{t=2}^{N} (\mathbf{c}_t - \mathbf{c}_{t-1})^T (\mathbf{c}_t - \mathbf{c}_{t-1}) \to \min(\mathbf{c})$$

с помощью фильтра-интерполятора Калмана-Бьюси [7].

Представим модель в явной форме. Мы будем полагать, что $\mathbf{y} = (y_1 \cdots y_N)^T \in \mathbb{R}^N$ и $\mathbf{c} = (\mathbf{c}_1^T \cdots \mathbf{c}_N^T)^T \in \mathbb{R}^{kN}$ есть вектор-столбцы, $\mathbf{X} = (\mathbf{X}_{ts}, t, = 1, \ldots, N)$ есть блочная матрица размера $(kN \times N)$ с блоками $\mathbf{X}_{ts} = (\mathbf{x}_t, \text{если} t \neq s)$ $(k \times 1)$, $\mathbf{B}_{\lambda,\rho}(kN \times kN)$ есть квадратная блочно-трехдиагональная матрица с диагональю $\big((1/\rho + 1/\lambda)\mathbf{I}, (2/\lambda)\mathbf{I}, \ldots, (2/\lambda)\mathbf{I}, (1/\lambda)\mathbf{I}\big)$ и не диагоналями $\big(-(1/\lambda)\mathbf{I}, \ldots, -(1/\lambda)\mathbf{I}\big)$, где $\mathbf{I}$ есть единичная матрица размера $(k \times k)$. Положим также дисперсию наблюдаемого шума равной единице $\delta = 1$, тогда модель (5) будет тогда давать функцию максимального правдоподобия $\ln \Phi(\mathbf{y} \,|\, \mathbf{c}, \mathbf{X}) = \ln N(\mathbf{y} \mid \mathbf{X}^T \mathbf{c}, \mathbf{I}) = const + (1/2)\mathbf{c}^T \mathbf{A}_N \mathbf{c}$, гессиан которой $\mathbf{A}_N = -\mathbf{X}\mathbf{X}^T$ $(kN \times kN)$ всегда вырожден и, если регрессоры $(x_{it}, t = 1, \ldots, N)$ линейно независимы, имеет максимальный ранг $\mathrm{rank}(\mathbf{A}_N) = N$. Скрытый марковская модель коэффициентов регрессии (6) выражается семейством априорных плотностей распределения $\ln \Psi(\mathbf{y} \,|\, \lambda, \rho) = \ln N(\mathbf{c} \,|\, \mathbf{0}, \mathbf{B}_{\lambda,\rho}^{-1}) = const + (1/2)|\mathbf{B}_{\lambda,\rho}| - (1/2)\mathbf{c}^T \mathbf{B}_{\lambda,\rho} \mathbf{c}$.

*Мы проанализировали* 200 *независимых реализаций случайного процесса* (5) *длинной* $N = 50$ *, полученного как линейная комбинация трех регрессоров* $(x_{it}, t = 1, \ldots, N)$, $i = 1, \ldots, k$, $k = 3$, *представляющих собой случайный белый шум с нулевым средним, с коэффициентами регрессии, взятыми как синусоидальные последовательности* $c_{it}^* = 4 \sin\big((2\pi/N)t + (2\pi/3)(i-1)\big)$ *смещенные друг относительно друга по фазе* 10%. Дисперсией шума в модели наблюдения составляла 10 процентов $\delta = 0.1\left((1/N) \sum_{i=1}^{N} (\mathbf{x}_t^T \mathbf{c}_t)^2\right)$.

Предполагается, что нет никакой априорной информации о векторе коэффициентов в первый момент времени, то есть $\rho \to \infty$. Зависимость «эффективной размерности» последовательности коэффициентов регрессии $(\mathbf{c}_1 \cdots \mathbf{c}_N)$ от предполагаемой дисперсии $\lambda$, вычисленная по единственной реализации произвольной последовательности регрессоров, изображена на Рис. 1. Эта размерность равна числу регрессоров в случае нулевой дисперсии $\lambda \to 0$ и достигает длины временных серий при $\lambda \to \infty$.

*Для каждой из* 200 *смоделированных временных серий, были вычислены* 3 *значения параметра дисперсии* $\hat{\lambda}$*, во-первых, по принципу двух версий обобщенного критерия Акаике* (24) и (27)*, во-вторых, традиционным методом скользящего контроля* [7]. *Затем мы применили каждое из полученных значений к оставшимся* 199 *временным сериям, как к контрольному множеству, и сравнили истинную последовательность коэффициентов регрессии* $(\mathbf{c}_1^* \cdots \mathbf{c}_N^*)$ *с полученной оценкой* $(\hat{\mathbf{c}}_{1,\hat{\lambda}} \cdots \hat{\mathbf{c}}_{N,\hat{\lambda}})$ *по критерию*

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

*Мы получили следующие результаты:*

| Критерий | $\hat{\lambda}$ | $\varepsilon_{\hat{\lambda}}$ |
|---|---|---|
| Максимум близости распределения наблюдаемой переменной | 0.010 | 0.012 |
| Скользящий контроль | 0.033 | 0.034 |
| Максимум близости совместного распределения наблюдаемой переменной и скрытого параметра | 0.072 | 0.055 |

*Формально, первая версия критерия Акаике показала наилучшие результаты, хотя результаты всех трех критериев очень близки друг к другу. Эксперименты выявили важный факт, что два фундаментально различных подхода: новый класс непрерывного обобщения информационного критерия Акаике и традиционный принцип скользящего контроля не превосходят друг друга в вопросе выбора наиболее подходящего значения параметра дисперсии, отвечающего за нестационарность регрессионой модели. В то же время непрерывный AIC несравненно лучше с вычислительной точки зрения.*

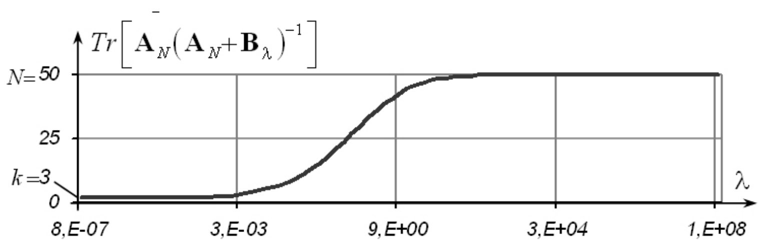


*Рис.* 1. Эффективная размерность последовательности коэффициентов регрессии как функция от $\lambda$ (логарифмический масштаб)

## Список литературы

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 fnd Akaik's Information Criterian (AIC): The general theory ands its analitical 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., Krasotcina 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.

***Статья поступила в редакцию 08.05.2008***