Научный вестник НГТУ. – 2013. – № 3(52)

УДК 519.237.5

Статистические критерии обнаружения разладки регрессии с циклическим трендом*

А.П. КОВАЛЕВСКИЙ

В работе построены статистические критерии, предназначенные для выявления изменений математического ожидания модели с циклическим трендом. Класс таких критериев основан на декорреляции и нормировке значений эмпирического моста. Наряду с этим классом рассматривается критерий, основанный на максимальном отклонении эмпирического моста от оси абсцисс. Сравнение критериев на численном примере показывает преимущество критерия, основанного на максимальном отклонении эмпирического моста.

Ключевые слова: статистический критерий, циклический тренд, разладка, гауссовский процесс, эмпирический мост.

ВВЕДЕНИЕ

Известно, что любое периодическое колебание с траекториями из может быть единственным образом представлено разложением в ряд Фурье по синусоидам. В настоящей работе рассматривается модель с конечным числом взаимно ортогональных гармоник и аддитивным случайным шумом в дискретном времени. Эта модель известна в литературе [1–3] как модель линейной регрессии с циклическим трендом. Изучение статистических критериев обнаружения разладки в этой модели предпринято в связи с исследованием колебаний строительных конструкций [4]: ставится задача определения изменений прочностных характеристик конструкции по исследованию ее колебаний.

Через n обозначим количество наблюдений. Будем предполагать, что моменты наблюдения равноотстоят друг от друга и что период наблюдений состоит из целого числа периодов колебаний. Таким образом, приходим к модели

$$Y_i^{(n)} = a_0 + \sum_{k\in M} \left(a_k \cos(2\pi ki/n) + b_k \sin(2\pi ki/n)\right) + \varepsilon_i,\ i = 1,\ldots,n. \qquad (1)$$

Через M обозначено некоторое конечное подмножество множества натуральных чисел. Относительно ошибок наблюдений делаются следующие стохастические предположения: случайные величины независимы и одинаково распределены с нулевым математическим ожиданием и ненулевой дисперсией σ^2. Задача состоит в том, чтобы построить класс статистических критериев, различающих основную гипотезу, определенную моделью (1), и альтернативную гипотезу, состоящую в том, что в некоторый момент времени значение a_0 в модели (1) заменяется на некоторое отличное от него значение b_0. При альтернативной гипотезе предполагается, что это изменение происходит только один раз за весь период наблюдений. После построения статистических критериев необходимо их сравнить и выбрать более мощный.

Естественным подходом к обнаружению разладки, т. е. изменения параметров модели в процессе наблюдения, является анализ регрессионных остатков $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$, где $\hat{Y}_i$ получены

* *Статья получена 17 января 2013 г.*

по формуле (1) с заменой истинных значений параметров a_0, a_k, b_k, $k \in M$, на их оценки по методу наименьших квадратов $\hat{a}_0, \hat{a}_k, \hat{b}_k$.

Простейшим критерием обнаружения разладки является критерий, основанный на сравнении средних значений остатков, подсчитанных по первой и второй половинам наблюдений. Ясно, что разность средних должна нормироваться среднеквадратическим отклонением или его выборочным аналогом. В результате получаем статистический критерий, близкий к критерию Стьюдента: большие по модулю значения нормированной разности средних свидетельствуют против основной гипотезы.

В параграфе 2 изложена асимптотическая теория и рассмотрен класс критериев, включающий приведенный выше. В параграфе 3 проводится моделирование процесса и исследуется вопрос об относительной эффективности предложенных критериев.

1. АСИМПТОТИЧЕСКАЯ ТЕОРИЯ

Обозначим через $\hat{\Delta}_j = \hat{\varepsilon}_1 + \ldots + \hat{\varepsilon}_j$ суммы остатков регрессии, $\hat{\Delta}_0 = 0$, а через $Z_n = \{Z_n(t), t \in [0,1]\}$ случайную ломаную, построенную по точкам $(k/n, \hat{\Delta}_k/(\sigma\sqrt{n}))$, $k = 0, \ldots, n$.

В статье МакНила [4] доказана слабая сходимость этой случайной ломаной в равномерной метрике в пространстве $C(0,1)$ к центрированному гауссовскому процессу B_f с ковариационной функцией

$$K_f(s,t) = \min(s,t) - st - \frac{1}{2\pi^2} \sum_{k \in M} \frac{1}{k^2} \big((1 - \cos 2\pi ks)(1 - \cos 2\pi kt) + \sin 2\pi ks \sin 2\pi kt\big), \quad (1)$$

$s,t \in [0,1]$. Пользуясь известными формулами тригонометрии, преобразуем (2) к более компактному виду:

$$K_f(s,t) = \min(s,t) - st - \frac{2}{\pi^2} \sum_{k \in M} \frac{1}{k^2} \sin \pi ks \sin \pi kt \cos \pi k(s-t). \quad (2)$$

Наряду с процессом Z_n рассмотрим *эмпирический мост* $\tilde{Z}_n$ – случайную ломаную, построенную по точкам $(k/n, \hat{\Delta}_k/(\tilde{\sigma}\sqrt{n}))$, $k = 0, \ldots, n$, где $\tilde{\sigma} = \left(\sum_{i=1}^n \hat{\varepsilon}_i^2 / n\right)^{1/2}$.

В определении эмпирического моста среднеквадратическое отклонение случайных ошибок σ заменено на его выборочный аналог. Таким образом, вычисление эмпирического моста не требует знания каких-либо характеристик модели (1).

Вместе с тем $\tilde{\sigma}$ является состоятельной оценкой параметра σ в силу состоятельности оценок коэффициентов регрессии, ограниченности косинуса и синуса, сходимостей $\sum_{i=1}^n \varepsilon_i / n \to 0$, $\sum_{i=1}^n \varepsilon_i^2 / n \to \sigma^2$ с вероятностью единица. Поэтому предельная теорема для Z_n верна также для эмпирического моста.

Рассмотренный во введении критерий, основанный на разности средних по первой и второй половине наблюдений, в терминах эмпирического моста основан на статистике $\tilde{Z}_n(1/2)$. Действительно, если число наблюдений n четно, то, так как сумма остатков регрессии в модели (1) равна нулю, разность между суммой первых $n/2$ и последних $n/2$ остатков равна $2\hat{\Delta}_{n/2}$. При делении на $\tilde{\sigma}\sqrt{n}$ получаем $2\tilde{Z}_n(1/2)$.

При нечетном n логично использовать ту же статистику как обеспечивающую симметрию между первой и второй половинами наблюдений.

Обозначим

$$J_1 = \tilde{Z}_n^2(1/2)/DB_f(1/2) = \tilde{Z}_n^2(1/2)/\left(\frac{1}{4} - \frac{2}{\pi^2}\sum_{k\in M}\frac{1}{k^2}\sin^2\frac{\pi k}{2}\right). \qquad (4)$$

При верной основной гипотезе статистика J_1 сходится слабо к распределению χ_1^2. Отметим, что для конечного множества M всегда $DB_f(1/2) > 0$, а если бы множество M включало в себя все положительные нечетные числа, то предельный процесс B_f в точке $1/2$ имел бы вырожденное в нуле распределение.

Предложенный критерий можно обобщить следующим образом. Вместо одной точки $1/2$ взять d точек: $\frac{1}{d+1},\dots,\frac{d}{d+1}$, и рассмотреть декоррелированные и нормализованные (в соответствии с корреляционной функцией (3)) значения эмпирического моста $\tilde{Z}_n$ в этих точках. Тогда сумма квадратов этих значений будет иметь в пределе хи-квадрат распределение с d степенями свободы в силу того, что сумма квадратов линейных функций от значений эмпирического процесса в фиксированных точках является непрерывным (в равномерной метрике) функционалом от эмпирического моста. И полученный таким образом критерий, и статистику, на которой он основан, будем обозначать J_d. Обозначим

$$\mathbf{z}_d = (z_{1,d},\dots,z_{d,d})^T = \left(\tilde{Z}_n\left(\frac{1}{d+1}\right),\dots,\tilde{Z}_n\left(\frac{d}{d+1}\right)\right)^T.$$

Через C_d обозначим ковариационную матрицу вектора

$\mathbf{b}_d = \left(B_f\left(\frac{1}{d+1}\right),\dots,B_f\left(\frac{d}{d+1}\right)\right)^T$:

$$C_d = \mathrm{E}\,\mathbf{b}_d{}^T\mathbf{b}_d.$$

Статистика J_d вычисляется по формуле

$$J_d = \mathbf{z}_d{}^T C_d{}^{-1}\mathbf{z}_d. \qquad (5)$$

Опишем подробнее критерии J_2 и J_3. Для построения критерия J_2 используем значения $\tilde{Z}_n(1/3)$ и $\tilde{Z}_n(2/3)$. Согласно (3) предельное значение $B_f(1/3)$ имеет дисперсию

$$b_{11} = K_f(1/3,1/3) = \frac{2}{9} - \frac{2}{\pi^2}\sum_{k\in M}\frac{1}{k^2}\sin^2\frac{\pi k}{3}.$$

В силу формул приведения, дисперсия предельного значения $B_f(2/3)$ такая же: $b_{22} = b_{11}$. Ковариация случайных величин $B_f(1/3)$ и $B_f(2/3)$ равна $b_{12} = \frac{1}{9} - \frac{1}{\pi^2}\sum_{k\in M}\frac{1}{k^2}\sin^2\frac{\pi k}{3} = b_{11}/2$.

Согласно (5) статистика

$$J_2 = \frac{4}{3b_{11}}\left(\widetilde{Z}_n^2(1/3) - \widetilde{Z}_n(1/3)\widetilde{Z}_n(2/3) + \widetilde{Z}_n^2(2/3)\right)$$

сходится слабо к распределению χ_2^2.

Для построения критерия J_3 выполняем те же действия со значениями $\widetilde{Z}_n(1/4)$, $\widetilde{Z}_n(1/2)$ и $\widetilde{Z}_n(3/4)$. Особенно простой вид имеет статистика J_3 в случае, когда M состоит из k, кратных 4. В этом случае матрица ковариаций равна

$$C_3 = \frac{1}{16}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix},$$

и статистика

$$J_3 = 8\left(\widetilde{Z}_n^2(1/4) - \widetilde{Z}_n(1/4)\widetilde{Z}_n(1/2) + \widetilde{Z}_n^2(1/2) - \widetilde{Z}_n(1/2)\widetilde{Z}_n(3/4) + \widetilde{Z}_n^2(3/4)\right)$$

сходится слабо к распределению χ_3^2.

Наряду с критериями J_d будем использовать критерий, основанный на статистике $J = \sup_{t\in[0,1]} |\widetilde{Z}_n(t)|$. Как показано в работе [5], этот критерий является асимптотически наиболее мощным в широком классе критериев в предположении, что $M = \varnothing$, т. е. в модели случайной выборки.

В общем случае (при $M \neq \varnothing$) для предельного закона статистики J нет аналитического описания, из общей теории [6] известна лишь грубая асимптотика больших уклонений

$$\ln P\{\sup_{t\in[0,1]} |B_f(t)| \geq x\} \sim -\frac{x^2}{2\sup_{t\in[0,1]} K_f(t,t)}$$

при $x \to \infty$. Если M состоит только из четных чисел, то $\ln P\{\sup_{t\in[0,1]} |B_f(t)| \geq x\} \sim -2x^2$.

2. СРАВНЕНИЕ КРИТЕРИЕВ

Смоделируем процесс (1) и сравним критерии, основанные на статистиках J_1, J_2, J_3 и J. Положим $n = 1024$, $M = \{4;16\}$, $b_1 = b_2 = \sigma = 1$, $a_0 = a_1 = a_2 = 0$. В случае разладки будем полагать $b_0 = 0,2$, а момент разладки τ равномерно распределенным на целых числах от 1 до n. Графики траектории процесса изображены на рис. 1.

Отметим, что периоды синусоид можно определить на основе быстрого преобразования Фурье (рис. 2). Для процесса с разладкой он имеет очень похожий вид. После оценивания параметров и вычитания прогнозных значений получаем остатки регрессии (рис. 3). Непосредственно на основании изучения рисунка затруднительно сделать вывод о том, произошла ли разладка в каждом конкретном случае. Поэтому вычислим значения статистик J_1, J_2, J_3 и J и достигнутые ими уровни значимости. Результаты приведены в табл. 1.

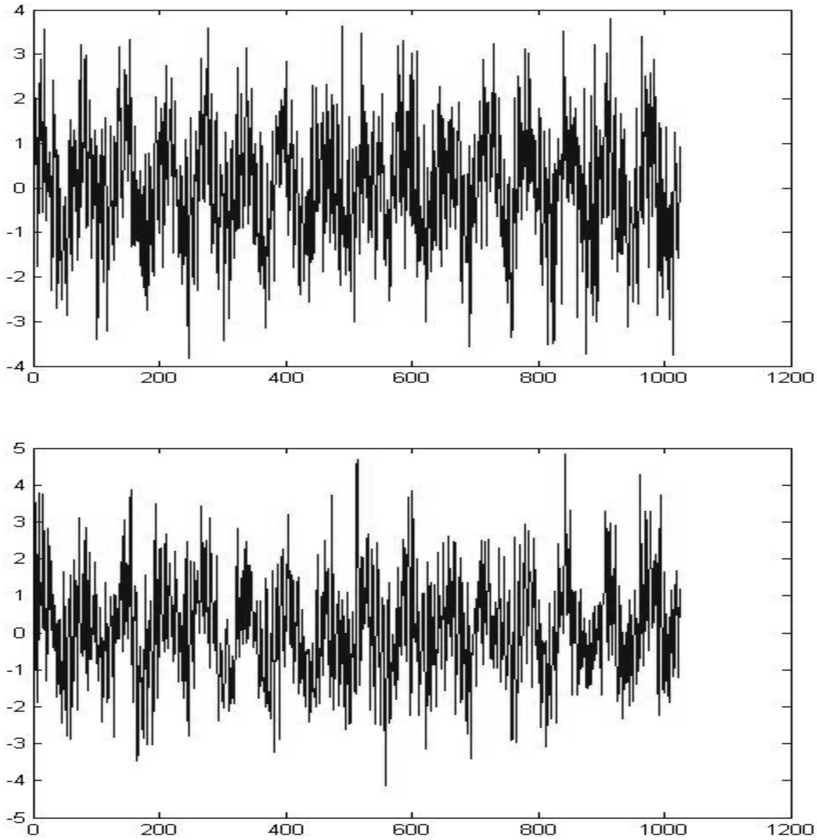

Рис. 1. График траектории процесса
(вверху – без разладки; внизу – с разладкой)

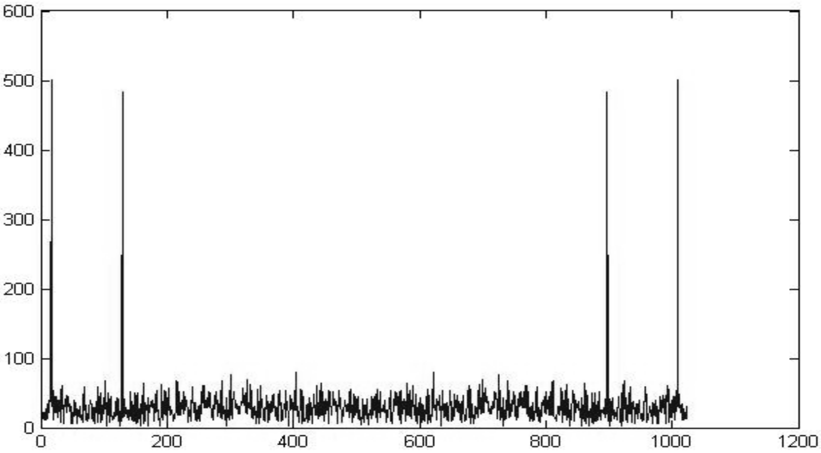

Рис. 2. График быстрого преобразования Фурье для процесса без разладки

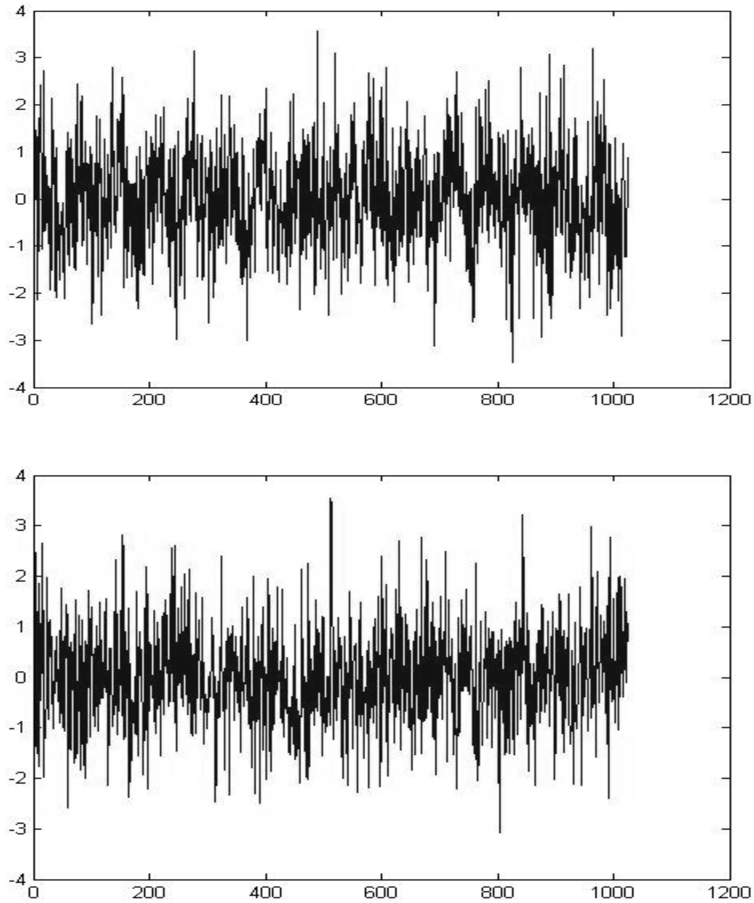

Рис. 3. Остатки регрессии (вверху – без разладки, внизу – с разладкой)

Таблица 1

Значения статистик критериев и достигнутые уровни значимости для примера моделирования процесса

Используемый критерий	Без разладки		С разладкой	
	Значение статистики	Достигнутый уровень значимости	Значение статистики	Достигнутый уровень значимости
J_1	0.65	0.42	2.42	0.12
J_2	1.75	0.42	6.34	0.04
J_3	2.91	0.41	5.91	0.12
J	0.82	0.51	1.32	0.06

В приведенном примере каждый из рассмотренных критериев не дает оснований предполагать разладку в случае, когда разладки на самом деле нет. В случае, когда разладка есть, критерии J_1 и J_3 отвергают основную гипотезу на уровне 0,12, критерий J на уровне 0,06, а критерий J_2 на уровне 0,04. Отметим, что уровень для критерия J требует коррекции, так как точное предельное распределение его статистики неизвестно.

Для того чтобы скорректировать критерий J, а также проверить соответствие критических уровней для остальных критериев, анализируем поведение критериев при отсутствии разладки. Эмпирические уровни значимости вычисляем по результатам 20000 моделирований процесса. Для критериев J_1, J_2, J_3 эмпирические уровни значимости отличаются от выбранного теоретического уровня 0,05 не более чем на 0,002, а для критерия J эмпирический уровнень значимости принимает значительно более низкое значение 0,0389. Это происходит вследствие того, что для этого критерия неизвестно точное предельное распределение, а лишь его грубая (логарифмическая) асимптотика. Поэтому скорректируем критерий J, выбрав в качестве нового уровня $0,05^2 / 0,0389 \approx 0,0643$. Повторный численный эксперимент по 20000 результатам моделирования дает эмпирический уровень значимости 0,0510. Всюду в дальнейшем используется этот исправленный критерий J для уровня значимости 0,05.

В случае разладки будем полагать момент разладки τ равномерно распределенным на целых числах от 1 до n, а значению b_0 математического ожидания после разладки будем последовательно придавать значения от 0,2 до 1 с шагом 0,2. Проведем вычисления с исправленным критерием J. Результаты вычислений приведены в табл. 2. Погрешность каждого вычисления не превосходит 0,01 с надежностью не менее 0,95.

Таблица 2

Эмпирические мощности критериев

Используемый критерий	Величина изменения математического ожидания					
	0.0	0.2	0.4	0.6	0.8	1.0
J_1	0.05	0.40	0.69	0.80	0.85	0.87
J_2	0.05	0.43	0.74	0.82	0.86	0.88
J_3	0.05	0.43	0.79	0.86	0.90	0.91
J	0.05	0.50	0.80	0.88	0.91	0.93

Из таблицы видно, что исправленный критерий J, основанный на максимальном отклонении эмпирического моста, при рассматриваемой альтернативной гипотезе является более мощным, чем критерии J_1, J_2, J_3, основанные на декорреляции и нормировке значений эмпирического моста в фиксированных точках. В то же время с ростом d мощность критерия J_d все более приближается к мощности критерия J.

ЗАКЛЮЧЕНИЕ

В работе построены статистические критерии обнаружения разладки процесса гармонических колебаний со случайным шумом.

Класс таких критериев основан на декорреляции и нормировке значений эмпирического моста. Для построения критерия J_d вычисляется корреляционная матрица значений предельного гауссовского процесса в точках $1/(d+1)$, …, $d/(d+1)$. Затем теоретические значения в этих точках ортогонализуются, нормируются, возводятся в квадрат и складываются. В результате получаем распределение хи-квадрат с d степенями свободы. Это распределение является предельным для статистики от эмпирического моста, построенной с использованием тех же преобразований.

Более детально описаны критерии J_1, J_2, J_3. Для них статистики критериев приведены в явном виде (для критерия J_3 только в случае, когда число наблюдений n кратно четырем). Наряду с ними рассматривается критерий, основанный на максимальном отклонении эмпирического моста от оси абсцисс. Этот критерий требует корректировки по результатам моделирования, так как для него известна только грубая асимптотика предельного распределения.

Сравнение критериев проводится на численном примере и показывает преимущество критерия, основанного на максимальном отклонении эмпирического моста в случае, когда альтернатива состоит в однократном изменении математического ожидания в случайный момент времени, равномерно распределенный на интервале наблюдения. Однако критерии, основанные на измерениях в нескольких точках, приближаются к нему по мощности с ростом числа точек.

СПИСОК ЛИТЕРАТУРЫ

[1] **Андерсон Т.** Статистический анализ временных рядов / Т. Андерсон. – М.: Мир, 1976.

[2] **Lu Q.Q.** Linear regression under multiple changepoints / Q.Q. Lu. – Athens, Georgia, 2004.

[3] **Дрейпер Н.Р.** Прикладной регрессионный анализ / Н.Р. Дрейпер, Г. Смит. – М.: Диалектика, 2007.

[4] **Шахраманьян А.М.** Системы мониторинга и прогноза технического состояния зданий и сооружений. Теория и практика / А.М. Шахраманьян // Русский инженер. – 2011. – № 1 (28). – С. 54–64.

[5] **MacNeill I.B.** Limit processes for sequences of partial sums of regression residuals / I.B. MacNeill // Annals of probability. – 1978. – Vol. 6. – № 4. – P. 695–698.

[6] **Гусарова Г.В.** Критерии наличия разладки / Г.В. Гусарова, А.П. Ковалевский, А.Г. Макаренко // Сибирский журнал индустриальной математики. – 2005. – Т. VIII. – №. 4 (24). – С. 18–33.

[7] **Kallianpur J.** Freidlin-Wentzell type estimates for abstract Wiener spaces / J. Kallianpur, H. Odaira // Sankhya. – 1978. – Ser. A. – Vol. 40. – P. 116–137.

Ковалевский Артем Павлович, кандидат физико-математических наук, доцент кафедры высшей математики Новосибирского государственного технического университета, доцент кафедры высшей математики физического факультета Новосибирского государственного университета. Основное направление научных исследований – теория вероятностей и математическая статистика. Имеет более 50 публикаций, в том числе 1 монографию. E-mail: pandorra@ngs.ru.

A.P. Kovalevskii

Statistical tests of a change-point in a regression with a cyclic trend

We propose statistical tests of change point in a model with a cyclic trend. A class of tests use decorrelation and renormalization of values of an empirical bridge. Another test is based on a maximal deviation of the empirical bridge. We compare these tests by calculations on an example and show an advantage of test based on the maximal deviation.

Key words: statistical test, cyclic trend, change-point, Gaussian process, empirical bridge.