Экстраполяция псевдослучайных процессов по максимуму подобия
Extrapolation of pseudorandom number sequence on maximum likeness,
Аспирант И.Чучуева (chuchueva@yandex.ru), профессор, д.т.н. Ю.Н.Павлов (Pavlov@iu3.bmstu.ru)

Развитие информационных систем и, как следствие, накопление огромного количества информации поставило на новый уровень задачи анализа данных. Среди прочих задач одной из актуальных является задача экстраполяции экономических и физических процессов, объединенных во множество псевдослучайных процессов.

В силу того, что задача экстраполяции псевдослучайных процессов не является новой, то до текущего момента была проделана серьезная работа в данной сфере. В соответствии с [1, 2, 3] все алгоритмы можно разделить на несколько групп: регрессионные модели [4,5,6] (линейная регрессия, АРПСС [3]); вероятностные модели [1,2]; метод группового учета аргументов [1]; нейронные сети [1]; классификационно-регрессионные деревья [8]; базовая авторегрессия с условной гетероскедастичностью [7]. Большинство приведенных методов предполагают не только поиск закономерностей внутри процессов, но и учет влияния внешних факторов [3,4,5,6,7,8].

В представленной научной работе в отличие от классических регрессионных методов анализ псевдослучайных процессов основан на предположении того, что существует множество факторов, оказывающих влияние на значения процесса, однако определить степень влияния каждого фактора невозможно по причине объема, конфиденциальности, трудностей в измерении информации. Авторами вводится предположение, что если общее влияние всего множества факторов в какой-то период времени привело к тому, что процесс имел определенный профиль, то существует или когда-то случится такой период времени, когда суперпозиция влияния всего множества факторов приведет к тому, что процесс будет иметь профиль подобный исходному. Данное предположение вводится на базе принципа Дирихле для псевдослучайного процесса с конечным числом внутренних состояний, значение которого рано или поздно повторится.

Пусть существует псевдослучайная последовательность $X(t) = [x_1, x_2, x_3, ...x_T]$ длиной *T*. Тогда введем обозначение

$$X_N^M = [x_N, x_{N+1}, x_{N+2}, x_{N+3}, ...x_{N+M-1}], \qquad (1)$$

вектор длины *M*, лежащий внутри исходного *X(t)* началом которого является момент времени *t=N*. В качестве меры подобия двух векторов внутри одной псевдослучайной последовательности используем линейную корреляцию Пирсона.

$$L_{N,M,J} = corr(X_N^M, X_J^M), \forall N, J \in [1, T-1], \forall M \in [1, T-1] : M + N < T \cup M + J < T \quad (2)$$

$$L_{N,M,J} = \frac{\mathrm{cov}(X_N^M, X_J^M)}{\sqrt{D[X_N^M]} \cdot \sqrt{D[X_J^M]}}, \qquad (3)$$

где $\mathrm{cov}(X_N^M, X_J^M)$ — ковариация исходных векторов, а $D[X_N^M]$ и $D[X_J^M]$ их дисперсии. Причем при *N=J* $L_{N,M,J} \equiv 1, \forall M \in [1, T-N]$.

Тогда **функция подобия**

$$Likeness(X_N^M) = L(i) = \left| corr(X_N^M, X_i^M) \right|, \forall i \in [1, N-1] \qquad (4)$$

возвращает вектор значений модулей коэффициентов линейной корреляции со всеми векторами длины *M*, лежащими левее x_N на оси времени. Результирующий вектор *L(i)* имеет длину *N-1* и назовём **вектор подобия**.

Аспирант И.Чучуева (chuchueva@yandex.ru), профессор, д.т.н. Ю.Н.Павлов (Pavlov@iu3.bmstu.ru)

Тогда вектор $X^M_{i\max}$, соответствующий максимуму вектору подобия

$$L(i\max) = \max(L(i)) = \max(\left|corr(X^M_N, X^M_i)\right|) \forall i \in [1, N-1] \qquad (5)$$

назовем **максимумом подобия** для исходного вектора X^M_N.

Далее введем предположение о том, что если X^M_N и $X^M_{i\max}$ имеют высокое подобие, то есть модуль линейного коэффициента корреляции Пирсона близок к 1, то вектора X^{M+1}_N и $X^{M+1}_{i\max}$ будут также иметь высокое подобие. Данное предположение назовем **предположением о подобии**. На основании предположения о подобии решается задача экстраполяции вектора $X(t) = [x_1, x_2, x_3, ...x_T]$ в точках $X^P_{T+1} = [x_{T+1}, x_{T+2}, x_{T+3}, ...x_{T+P}]$ по максимуму подобия.

В связи с тем, что в качестве меры подобия был использован коэффициент линейной корреляции, то экстраполированные значения будут определяться следующим образом:

$$X^P_{T+1} = A \cdot X' \ , \qquad (6)$$

где A — матрица линейных коэффициентов размерностью 2x1, а X' часть исходной последовательности X^T_1, определяемая исходя из равенств, представленных ниже. В частном случае при $P = 1$, то есть, в случае, когда необходимо экстраполировать процесс лишь в одной точке, равенство (6) принимает вид простой линейной зависимости:

$$x_{T+1} = a_1 \cdot x' + a_0 \,, \qquad (7)$$

В общем случае при необходимости экстраполяции *P* точек для определения матрицы *A*, возьмем вектор X^M_{T-M+1} и найдем его максимум подобия.

$$X^M_{i\max} : L(i\max) = \max(Likeness(X^M_{T-M+1})), \forall i \in [1, T-P] \qquad (8)$$

Считаем, что для векторов X^M_{T-M+1} и $X^M_{i\max}$ верно равенство (9), которое расшифровывается в выражении (10)

$$X^M_{T-M+1} = A \cdot X^M_{i\max} + \varepsilon^M \,, \qquad (9)$$

$$\begin{bmatrix} x_{T-M} \\ x_{T-M+1} \\ \dots \\ x_T \end{bmatrix} = [a_1, a_0] \cdot \begin{bmatrix} x_{i\max} & 1 \\ x_{i\max+1} & 1 \\ \dots & \dots \\ x_{i\max+M} & 1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_M \end{bmatrix}, \qquad (10)$$

где вектор ε^M - вектор ошибок аппроксимации. Аппроксимация

$$X^M_{T-M+1} = A \cdot X^M_{i\max} \,, \qquad (11)$$

позволяет определить матрицу *A*, решая уравнение (11):

$$A = ((X^M_{i\max})^T \cdot X^M_{i\max})^{-1} \cdot (X^M_{i\max})^T \cdot X^M_{T-M+1} \qquad (12)$$

В соответствии с предположением о подобии в качестве X' берем вектор $X' = X^P_{i\max+M+1}$, то есть вектор, лежащий на оси времени сразу за вектором максимума подобия. Положения векторов наглядно представлены на рис 1.

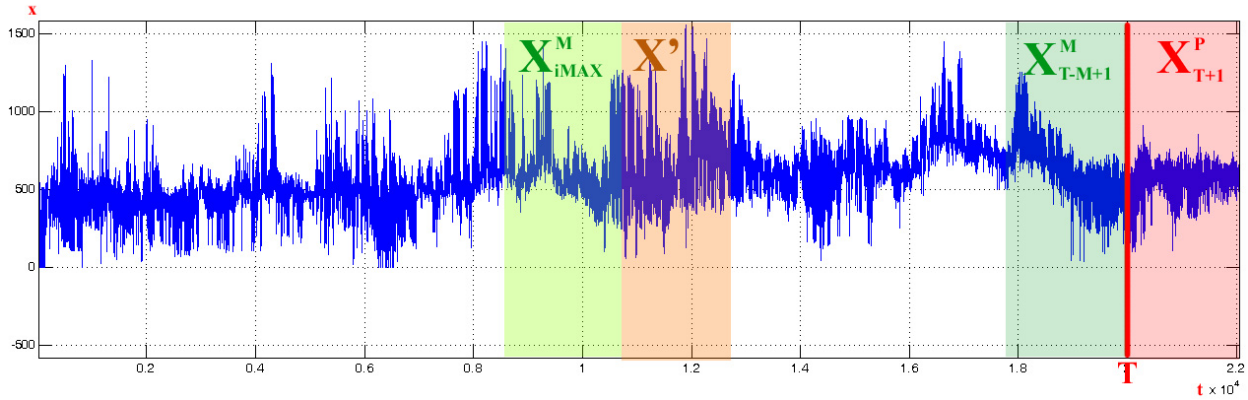

Рис 1. Положения векторов X^P_{T+1}, X^M_{T-M+1}, $X^M_{i\max}$ и X' на оси времени

По предложенному алгоритму решается задача экстраполяции псевдослучайной последовательности $X(t)=[x_1,x_2,x_3,...x_T]$ в точках $X^P_{T+1}=[x_{T+1},x_{T+2},x_{T+3},...x_{T+P}]$ по максимуму подобия.

Отметим также, что в основе экстраполяции лежит линейная регрессия (6), а, следовательно, возможен учет влияния внешних факторов на исследуемый процесс в случаях, когда данный учет необходим. Тогда

$X^P_{T+1}=A\cdot X'$, где

$$X'=\begin{bmatrix} x_{i\max+M+1} & y_{i\max+M+1} & \dots & 1 \\ \dots & \dots & \dots & \dots \\ x_{i\max+M+P} & y_{i\max+M+P} & \dots & 1 \end{bmatrix}, \qquad (13)$$

где Y – вектор значений независимой переменной. Однако в данном случае в качестве меры подобия необходимо использовать квадрат множественного коэффициента корреляции [9] вместо линейной корреляции Пирсона. Алгоритм экстраполяции с учетом внешних факторов, а также поведение ошибки будет опубликован в одной из следующих статей.

В заключение статьи приводим примеры реализации экстраполяции псевдослучайных кривых.

1) Экстраполяция кривой фьючерсных цен на природный газ на Нью-Йоркской товарной бирже (NYMEX, www.nymex.com) за период с 01.10.2008 по 01.05.2009 (7 месяцев) на 24 значение вперед (почасовые значения на следующий день) – средняя ошибка экстраполяции составила 2,36%.

2) Экстраполяция Торгового Графика (потребления) по Сибирской ценовой зоне ОРЭМ (Оптовый рынок электроэнергии и мощности, www.atsenergo.ru) за период с 01.03.2008 по 01.03.2009 (12 месяцев) на 24 значения вперед (почасовые значения на следующий день) – средняя ошибка экстраполяции составила 1.39%.

3) Экстраполяция цен РСВ (рынок на сутки вперед) по Европейской ценовой зоне ОРЭМ за период с 01.01.2009 по 28.02.2009 (2 месяца) на 24 значения вперед (почасовые значения на следующий день) – средняя ошибка составила 7.94%.

Оценка точности экстраполяции производилось при помощи MAPE (mean absolute percentage error) – средняя абсолютная ошибка в процентах, определяемая по формуле:

$$MAPE=\frac{1}{n}\cdot\sum_{i=1}^{n}\frac{|P_i^{Forecast}-P_i^{Fact}|}{P_i^{Fact}}\cdot 100\% \qquad (14)$$

Аспирант И.Чучуева (chuchueva@yandex.ru), профессор, д.т.н. Ю.Н.Павлов (Pavlov@iu3.bmstu.ru)

В статье рассмотрен метод экстраполяции псевдослучайных процессов на основании максимума подобия, а также продемонстрированы некоторые результаты, которые позволяют говорить о состоятельности данного подхода. В дальнейших статьях планируется представить подробный анализ результатов экстраполяции различных псевдослучайных процессов.

1. Э.Е. Тихонов, «Прогнозирование в условиях рынка», Невинномысск, 2006 г.
2. В. И. Суслов, Н. М. Ибрагимов, Л. П. Талышева, А. А. Цыплаков, «Эконометрия», И: Новосибирский государственный университет, 2005 г.
3. А.А. Грешилов, В.А. Стакун, А.А. Стакун, «Математические методы построения прогнозов», И: Москва, Радио и связь, 1997 г.
4. Дж. Бокс, Г. Дженкинс, «Анализ временных рядов», 1967 г.
5. Prajakta S. Kalekar, «Time series Forecasting using Holt-Winters and Exponential Smoothing», Kanwal Rekhi School of Information Technology, 2004
6. Uwe Hassler and JÄurgen Wolters, «Autoregressive Distributed Lag Models and Cointegration», 2005
7. Reinaldo C. Garcia, «A GARCH Forecasting Model to Predict Day-Ahead Electricity Prices», German Institute of Economic Research, DIW (Berlin), Germany, 2003
8. M. Sc. Jingfei Yang, «Power System Short-term Load Forecasting», Elektrotechnik und Informationstechnik der Technischen Universität Darmstadt, 2006
9. Herve Abdi, Multiple Correlation Coefficient, The University of Texas at Dallas, 2007