СТАТИСТИЧЕСКИЙ АНАЛИЗ ПАРАМЕТРИЧЕСКИХ ВРЕМЕННЫХ РЯДОВ С ПРОПУСКАМИ НА ОСНОВЕ МОДЕЛЕЙ В ПРОСТРАНСТВЕ СОСТОЯНИЙ

С. В. Лобач

Белорусский государственный университет
Минск, Беларусь
e-mail: lobachS@bsu.by

Рассмотрены временные ряды на основе моделей в пространстве состояний в случае, когда имеют место пропущенные наблюдения. Будущие значения временных рядов также интерпретируются как пропуски, стоящие в конце временного ряда. Проблема прогнозирования временных рядов выступает как проблема статистического анализа временных рядов с пропусками, для ее решения применяется фильтр Калмана.

Ключевые слова: пропущенные наблюдения; модель в пространстве состояний; фильтр Калмана; прогнозирование.

STATISTICAL PARAMETRIC ANALYSIS OF TIME SERIES WITH MISSING OBSERVATIONS BASED ON STATE SPACE MODELS

S. V. Lobach

Belarusian State University
Minsk, Belarus

The article describes the time series based on the state space model when there are missing observations. The future value of a time series is interpreted as well as missing observations stand at the end of the time series. The problem of time series prediction is interpreted as a problem of statistical analysis of time series with missing data, for its decision applied the Kalman filter.

Keywords: the missing observations; state space model; Kalman filter; prediction.

ВВЕДЕНИЕ

В наиболее общей постановке проблема статистического анализа данных с пропусками может быть сформулирована следующим образом. Имеется многомерная выборка наблюдений $(x_t, 0_t), t = 1, ..., T,$ где $x_t = (x_{1t}, ..., x_{nt})' \in R^n$ – вектор наблюдений, $0_t = (0_{1t}, ..., 0_{nt})' \in R^n$ – вектор шаблона пропусков, координаты которого имеют значение «1», если в момент времени t соответствующая координата вектора x_t наблюда-

ется; «0», если соответствующая координата вектора x_t не наблюдается. Проблема состоит в том, чтобы по данным с пропусками построить оптимальные статистические выводы относительно пропущенных значений временного ряда x_t, $t \in 1,2,...,T$, а также относительно будущих значений временного ряда $\{x_\tau\}, \tau > T$. Данной тематике посвящена обширная литература [1–4].

Заполнение пропусков – наиболее общий и гибкий метод решения статистических задач при наличии пропусков. Например, если $n-$ вектор x_t распределен по многомерному нормальному закону распределения вероятностей $N_n(\mu, \Sigma)$ с математическим ожиданием μ и ковариационной матрицей $\sum$, то применяют метод, изложенный в работе [5]. Сначала оценивают μ и $\sum$ выборочным средним и выборочной ковариационной матрицей по имеющимся полным наблюдениям, а затем используют эти оценки для вычисления линейной регрессии пропущенных значений по имеющимся.

EM-АЛГОРИТМ

EM-алгоритм описан в работе [2] и предназначен для вычисления оценок максимального правдоподобия параметров по неполным данным. В дальнейшем он получил широкое распространение и стал применяться для решения других задач статистического анализа данных, например, для разделения смесей, для построения процедур кластерного анализа.

EM-алгоритм состоит из двух шагов.

На шаге «*M*» вычисляют оценку максимального правдоподобия параметра θ модели временного ряда по имеющимся наблюдениям.

На шаге «*E*» находят условное математическое ожидание пропущенных наблюдений $\{x_{tjmis}\}$ по имеющимся наблюдениям и при условии, что параметры модели оценены на шаге «*M*».

Этот процесс продолжается до тех пор, пока не удовлетворяются заданные условия сходимости. До сих пор вопрос о сходимости *EM*-алгоритма остается открытым, получены строгие результаты по сходимости только для некоторых специальных задач статистического анализа данных с пропусками [6].

ВРЕМЕННЫЕ РЯДЫ НА ОСНОВЕ МОДЕЛЕЙ В ПРОСТРАНСТВЕ СОСТОЯНИЙ И ФИЛЬТР КАЛМАНА

Предположим, что имеется временной ряд $\{x_t\}$, $t \in 1,2,...,T$. Можно считать, что случайный n-вектор x_t описывает текущее состояние некоторой системы в момент времени t. Обычно переменные состояния не наблюдаются точно, а регистрируются другие переменные y_t, связанные функциональной или статистической зависимостью с переменными состояния x_t.

Моделью в пространстве состояний называется система, состоящая из двух векторных линейных уравнений:

$$x_t = Fx_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N_n(0, Q); \tag{1}$$

$$y_t = Hx_t + \eta_t, \quad \eta_t \sim N_m(0, R); \tag{2}$$

$$x_0 \sim N_n(\alpha_0, P_0), \quad t = 1, 2, ..., T,$$

где F, H, Q, R — матрицы размеров $(n \times n)$, $(m \times n)$, $(n \times n)$, $(m \times m)$ соответственно, в общем случае они неизвестны либо являются известными функциями от параметров. Начальные условия α_0, P_0 предполагаются известными. Случайные векторы $(\varepsilon_t, \eta_{te}), t = 1, ..., T,$ являются последовательностью независимых гауссовских случайных векторов с нулевыми математическими ожиданиями и ковариационными матрицами Q и R.

Оказывается, что большинство известных параметрических моделей временных рядов могут быть представлены в форме моделей в пространстве состояний (1), (2). Главным ограничением модели (1), (2) является линейность модели. Элементы матрицы F и H могут быть нестационарными, уравнения (1) и (2) могут быть нелинейными.

Для оценивания неизвестных параметров временного ряда (3) может быть использован фильтр Калмана [7], примененный к модели (5). Фильтр Калмана [7] позволяет рекуррентно вычислить оценки $x_{t|t} = E\{x_t \mid y_0^t\}$ и матрицу ковариаций $P_{t|t} = Var\{x_t \mid y_0^t\}$, где $y_0^t = \{y_0, ..., y_t\}$ – наблюдения до момента времени t включительно.

Введем обозначения:

$$x_{t|t-1} = E\{x_t \mid y_o^{t-1}\}; \qquad y_{t|t-1} = E\{y_t \mid y_o^{t-1}\};$$
$$V_{t|t} = E\{(x_t - x_{t|t})(x_t - x_{t\}t})'\}, \quad V_{t|t-1} = E\{(x_t - x_{t|t-1})(x_t - x_{t|t-1})'\};$$
$$M_{t|t-1} = E\{(y_t - y_{t|t-1})(y_t - y_{y|t-1})'\}; \quad \upsilon_t = y_t - y_{t|t-1} = y_t - Hx_{t|t-1},$$

где $x_{t|t-1}, y_{t|t-1}$ являются прогнозными значениями x_t и y_t, полученными по наблюдениям $\{y_\tau\}$ до момента $t-1$ включительно. Матрицы $V_{t|t}, V_{t|t-1}, M_{t|t-1}$ представляют собой матрицы ковариаций ошибок оценивания; ν_t – остатки оценивания.

Для применения фильтра Калмана требуется задание начальных условий. Более того, чтобы применить формулы (6), (7) в момент времени $t = 1$, необходимо знать $x_{1|0}, v_{1|0}$. Оценка $v_{1|0}$ означает оптимальную оценку x_1 при условии, что нет никакой информации, т. е. наблюдения отсутствуют. Эта проблема инициализации решается просто, если даны параметры a_0 и P_0, в этом случае $x_{1|0} = a_0, \quad v_{1|0} = P_0$. В противном случае необходимо задать априорные оценки этих параметров.

В [8] показано, что проблема пропущенных наблюдений легко решается для временных рядов, представленных в форме моделей в пространстве состояний. Если наблюдение y_t полностью пропущено в момент времени t, то в формулах (6), (7), (12) нужно положить $\nu_t = 0, \quad K(t) = 0$. В статье данный подход предлагается применить для будущих значений временного ряда, которые предлагается рассматривать как пропущенные.

АНАЛИЗ НЕПОЛНЫХ ДАННЫХ

Пусть только некоторые компоненты вектора y_t наблюдаются в момент времени t: $i_1(t) < i_2(t) < ... < i_{l_t}(t)$, тогда наблюдаемым в момент времени t является вектор $y_t^* = S_t y_t$, где S_t является матрицей размера $l_t \times m$, которая принимает значение «1» в позициях $\left(1, i_1(t), ..., \left(l_t, i_{l_t}(t)\right)\right)$ и «0» – в других. В этом случае модель (1), (2) сводится к модели:

$$x_t = Fx_{t-1} + \varepsilon_t; \tag{13}$$

$$y_t^* = H^* x_t + \eta_t^*, \tag{14}$$

где $y_t^* = S_t y_t$, $H^* = S_t H$, $\eta_t^* = S_t \eta_t$.

Заметим, что модель (13), (14) является моделью с полными данными, к ней может быть применен фильтр Калмана.

Прогнозирование на шаг вперед осуществляется по формулам:

$$x_{t|t-1} = Fx_{t-1|t-1}, \quad y_{t|t-1} = Hx_{t|t-1}.$$

Риск прогнозирования определяется матрицами ковариации ошибок оценивания (9), (10).

АНАЛИЗ ДАННЫХ С ПРОПУСКАМИ МОДЕЛИ ARIMA(2,1,2)

Рассмотрим следующие модели пропусков:

– пропуски в фиксированные моменты времени;

– пропуски в случайные моменты времени.

i_t – СВ ~ $Bi(1, p)$,

$P\{i_t = 1\} = p$; $P\{i_t = 0\} = 1 - p$, $p = 0{,}2$.

Таблица

Оценка параметров при анализе данных с пропусками

	α_1	α_2	β_1	β_2
Реальные значения	0,927	–0,079	–0,15	1,05
Значения без пропусков	1,08117	–0,058	–0,017	0,5799
Детерминированный шаблон				
0–20	1,2306	–0,2092	–0,1371	0,2554
20–40	1,2168	–0,1768	–0,12702	0,2701
40–60	1,223	–0,192	–0,1307	0,239
10–20, 40–50	1,2166	–0,186	–0,1317	0,2334
40–50, 90–100	1,2493	–0,2619	–0,1365	0,2063

Случайный шаблон				
Вероятность 0,2	1,181	–0,081	–0,8508	0,911
Вероятность 0,3	1,278	–0,089	–0,162	1,213
Вероятность 0,4	1,38	–0,163	–0,1158	1,067
Вероятность 0,5	1,269	–0,105	–0,1325	1.007
Вероятность 0,6	1,247	–0,0989	–0,889	0,8499
Вероятность 0,7	1,328	–0,149	–0,841	0,903
Вероятность 0,8	1,327	–0,164	–0,645	0,834

Как видно из таблицы, в детерминированном шаблоне пропусков место, в котором допущены пропуски, влияет на оценки слабо. Оценки детерминированного и случайного шаблона существенно отличаются. Это обусловлено сильным разбросом случайных пропусков. Оценки случайного шаблона более приближены к реальным значениям. Пропуски ухудшают точность оценивания. При увеличении вероятности пропусков точность оценивания уменьшается.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Green W. Econometric Analysis // New York, 2000.
2. Литтл Р., Рубин Д. Статистический анализ данных с пропусками //М. : Наука, 1991.
3. Shafer J. Analysis of incomplete data // London, 1997.
4. Харин Ю. С. Оптимальность и робастность в статистическом прогнозировании // Минск : БГУ, 2008.
5. Hurtly H., Hocking R. The analysis of incomplete data // Biometrics. 1971. Vol. 27. P. 783–808.
6. Jeff Wu C.F. On the Convergence Properties of the EM Algorithm // The Annals of statistics. Vol. 11. № 1, 1983. P. 95–103.
7. Kalman R., Busy R. New results in linear filtering and prediction theory // Transaction ASME Journal of Basic Engineering. Series D. Vol.83. P. 95–108.
8. Harvey A., Koopman S., Shepard N. State space and unobserved component models // Theory and applications. London. Cambridge University Press, 2004.