ПРИМЕНЕНИЕ ФИЛЬТРА КАЛМАНА ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА ВРЕМЕННЫХ РЯДОВ С ПРОПУСКАМИ

В.И. Лобач

Белорусский Государственный университет, пр. Независимости, 4, 220030 г. Минск, Беларусь, lobach@bsu.by

Рассматривается модификация фильтра Калмана для анализа временных рядов с пропусками. Рассматривается два варианта пропусков-детерминированный и случайный.

***Ключевые слова*:** фильтр Калмана, временные ряды, пропуски в наблюдениях, модели в пространстве состояний.

APPLICATION OF THE KALMAN FILTER FOR STATISTICAL ANALYSIS OF TIME SERIES WITH MISSING DATA

V.I. Lobach

Belarusian State University, Independence Ave., 4, 220030 Minsk, Belarus, lobach@bsu.by

A modification of the Kalman filter for analyzing time series with gaps is considered. Two options for omissions are considered: deterministic and random.

Keywords: Kalman filter, time series, gaps in observations, state space models.

Введение

Самые ранние исследования, использующие методологию пространства состояний, относятся, скорее, к техническим приложениям, чем к статистическим. Пионерской работой в этом направлении исследований является работа Калмана [1]. В этой работе Калман высказал две важные идеи. Во-первых, он показал, что очень большой класс проблем может быть сформулирован в терминах моделей в пространстве состояний. Во-вторых, в силу марковского свойства модели в пространстве состояний вычислительные алгоритмы могут быть представлены в рекуррентной форме, что является очень удобным для компьютерной реализации алгоритмов.

Когда используется подход Бокса-Дженкинса [2] для прогнозирования временных рядов, исключается тренд, сезонная составляющая путем взятия конечных разностей, поэтому невозможно восстановить тренд, сезонную компоненту, хотя именно эти компоненты временного ряда представляют иногда особый интерес для эконометрических приложений.

При использовании подхода, основанного на концепции пространства состояний, таких проблем не возникает. Практические исследования показывают, что модели типа $\mathrm{ARIMA(p,d,q)}$достаточно хорошо описывают широкий класс данных, однако эти модели могут быть представлены в форме моделей в пространстве состояний. Подход Бокса-Дженкинса возможен, если после взятия конечных разностей получается стационарный временной ряд, что является слабостью этого метода. При использовании моделей в пространстве состояний такой проблемы не возникает.

В данной работе на основе моделей в пространстве состояний предлагается алгоритм прогнозирования $\mathrm{AR(p)}$моделей временных рядов с пропусками.

Модели в пространстве состояний

Рассмотрим модель в пространстве состояний [3]

$$\beta_{t+1} = F_t\beta_t + \epsilon_t, \tag{1}$$

$$z_t = H_t^{'}\beta_t + \eta_t, t \geq 0, \tag{2}$$

где β_t – случайный вектор размерности k, называемый вектором состояний в момент времени t. Предполагается, что β_0 имеют нормальное распределение $N(m,P)$, а также, что k-мерные вектора ϵ_t независимые и одинаково распределённые по закону $N(0,Q)$, $\eta_t \sim N(0,R)$. Матрица F_t размерности $k \times k,$, называемая матрицей переходов, является неслучайной. Первое из приведённых уравнений называется уравнением состояния. Оно определяет распределение β_t для любых $t \geq 0$. Вектор z_t, именуемый вектором измерений, имеет размерность n. Вектора η_t взаимно независимы с векторами ϵ_t и одинаково распределены по закону $N(0,R)$. Второе из приведённых уравнений называется уравнением измерения. Оно позволяет определить распределение z_t. Оба эти уравнения опреде-

ляют распределение гауссовского процесса $(\beta_t, z_t), t \geq 0$. Матрица H_t размерности $n \times k$ является неслучайной. Она именуется матрицей измерений. Важное различие между z_t и β_t заключается в том, что z_t наблюдаются, в то время как β_t, вообще говоря, частично наблюдаются либо не наблюдаются вовсе. В любом случае информацией, доступной на момент времени t, является $z_1, \ldots, z_t$.

Проблемами, с которыми сталкивается, используя модели в пространстве состояний, являются проблемы, связанные с численным определением условных математических ожиданий. Могут быть выделены три вида этой проблемы. Проблема фильтрации основывается на вычислении $E\{\beta_t \mid z_1, \ldots, z_t\}$, которое является оптимальным приближением β_t по информации, доступной на момент времени t. Проблемы сглаживания основываются на вычислении $E\{\beta_s \mid z_1, \ldots, z_t\}$, в случае когда $s < t$. И проблемы прогнозирования основываются на численном определении $E\{\beta_s \mid z_1, \ldots, z_t\}$, $E\{z_s \mid z_1, \ldots, z_t\}$, когда $s > t$.. Практически важной является решение проблемы наличия пропусков в наблюдениях.

Математические модели временных рядов могут быть сведены к моделям в пространстве состояний различными способами, то есть неоднозначно, что обусловлено возможностью различным образом задавать вектор состояний, а также матрицы переходов и измерений.

Сведение моделей авторегрессии к модели в пространстве состояний

Рассмотрим модель авторегрессии AR(p)

$$z_t = \phi_1 z_{t\text{-}1} + \ldots + \phi_p z_{t\text{-}p} + \epsilon_t,$$

где $\epsilon_t \sim N\left(0, \sigma^2\right)$.

Приведем один из вариантов привидение этой модели к форме моделей пространстве состояний. Пусть вектор состояния задан следующим образом:

$$\beta_t = \left(z_t, z_{t-1}, \ldots, z_{t-p+1}\right)^T,$$

а вектора w_t зададим следующим образом:

$$w_t = \left(\epsilon_t, 0, \ldots, 0\right)^T,$$

тогда модель в пространстве состояний для модели авторегрессии порядка p имеет вид:

$$\beta_t = \begin{pmatrix} \phi_1 & \cdots & \phi_p \\ 1 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \\ & & \end{pmatrix} \beta_{t-1} + w_t,$$

$$z_t = \begin{pmatrix} 1 & 0 & \ldots & 0 \end{pmatrix} \beta_t.$$

Аналогичный результат можно получить для модели авторегресии – скользящего среднего ARMA(p, q) и ARIMA(p,d,q).

Фильтр Калмана для статистического анализа ARMA-модели временного ряда

Ковариационный фильтр Калмана – это алгоритм, используемый для рекуррентного вычисления «отфильтрованных» векторов состояния. Ниже приведены формулы, задающие фильтр в классическом случае. Этот фильтр используется для сглаживания зашумлённых данных.

Введём следующие обозначения:

$$\hat{\beta}_{(t|t)} = E\{\beta_t \mid z_1, \ldots, z_t\},$$

$$\beta_{(t|t-1)} = E\{\beta_t \mid z_1, \ldots, z_{t-1}\},$$

$$\hat{z}_{(t|t-1)} = E\{z_t \mid z_1, \ldots, z_{t-1}\},$$

$$\Sigma_{(t|t)} = E\left\{\left(\beta_t - \beta_{(t|t)}\right)\left(\beta_t - \beta_{(t|t)}\right)'\right\},$$

$$\Sigma_{(t|t-1)} = E\left\{\left(\beta_t - \beta_{(t|t-1)}\right)\left(\beta_t - \beta_{(t|t-1)}\right)'\right\},$$

$$M_{(t|t-1)} = E\left\{\left(z_t - \hat{z}_{(t|t-1)}\right)\left(z_t - \hat{z}_{(t|t-1)}\right)'\right\},$$

$$\tilde{z}_t = z_t - \hat{z}_{(t|t-1)} = z_t - H_t'\beta_{(t|t-1)}.$$

Таким образом, $\beta_{(t|t-1)}$ и $\hat{z}_{(t|t-1)}$ – прогнозные значения β_t, z_t, составленные в момент времени $t-1$. Матрицы $\Sigma_{(t|t)}, \Sigma_{(t|t-1)}, M_{(t|t-1)}$ – средне-

квадратичные ошибки прогнозирования, а $\tilde{z}_t$ – остатки МНК в регрессии z_t на предыдущие значения.

Фильтр Калмана для модели (1), (2) определяется соотношениями [4,5]:

$$\beta_{(t|t)} = \beta_{(t|t-1)} + K_t \tilde{z}_t, \tag{3}$$

где коэффициенты фильтра вычисляется следующим образом:

$$K_t = \Sigma_{t|t-1} H_t \left(H_t' \, \Sigma_{t|t-1} H_t + R \right)^{-1}, \tag{4}$$

$$\Sigma_{t|t} = \left(I - K_t H_2' \right) \Sigma_{t|t-1}, \tag{5}$$

$$\beta_{(t+1|t)} = F_t \beta_{(t|t)}, \tag{6}$$

$$\Sigma_{t+1|t} = F_t \Sigma_{t|t} F_t' + Q, \tag{7}$$

$$\hat{z}_{(t+1|t)} = \mathrm{H}_{t+1}' \beta_{(t+1|t)}, \tag{8}$$

$$M_{t+1|t} = H_{t+1}' \Sigma_{t+1|t} H_{t+1} + R. \tag{9}$$

Рассмотрим не классическую реализацию фильтра Калмана для зашумлённых данных, а его робастную модификацию, которую можно было бы применить к временному ряду с пропущенными значениями в определённые периоды наблюдений.

Для того, чтобы модифицировать фильтр Калмана, применим следующую процедуру. Если в момент времени t наблюдение отсутствует, то в формулах (3)-(9) вместо истинного значения наблюдения z_t будем использовать его оценку по предыдущим $t-1$ наблюдением, таким образом, остаток регрессии $\tilde{z}_t$ на предыдущие значения становится тождественно равным нулю. Это влечёт за собой то, что в формулах (3)-(9) пропадает второе слагаемое, то есть для вектора состояния этап коррекции, по сути, пропускается, так как нет значения, по которому было бы возможно провести коррекцию.

Таким образом, если в момент времени t наблюдаемое значение отсутствует, то формулы (3), (8) обращаются в следующие формулы:

$$\hat{z}_{t+1} = 0, \qquad \hat{\beta}_{(t|t)} = \beta_{(t|t-1)}.$$

На основе вышеописанных формул может быть запущен процесс построения оценок параметров модели, заложенных в вектор состояния динамической системы, а также может быть построен прогноз будущих значений временного ряда.

Оценивание параметров AR(2)-модели временного ряда

Рассмотрим задачу оценивания параметров (2) модели временного ряда и прогнозирования его значений с учётом имеющихся пропусков, описываемых детерминированным или случайным шаблоном.

В качестве неизвестных истинных значений параметров для модели AR(2) возьмём значения $\phi_1 = 1.5,\ \phi_2 = -1$ для генерации нестационарного временного ряда, а также значения $\phi_1 = 0.5,\ \phi_2 = 0.3$ для стационарного временного ряда.

Результаты оценивания параметров AR(2) модели временного ряда с пропусками на основе модифицированного фильтра Калмана приведены в табл. 1, 2.

Таблица 1

Результаты компьютерных экспериментов на модельных данных на первом наборе параметров

$\phi_1 = 1.5,\ \phi_2 = -1.0$	Детерминированный шаблон. Пропуски 46-50, 96-100	Случайный шаблон P(пропуск)=0.1
ϕ_1	1,474	1,446
ϕ_2	-0,973	-0,949

Таблица 2

Результаты компьютерных экспериментов на модельных данных на втором наборе параметров

$\phi_1 = 0.5,\ \phi_2 = -0.3$	Детерминированный шаблон Пропуски 46-50, 96-100	Случайный шаблон P(пропуск)=0.1
ϕ_1	0,472	0,484
ϕ_2	-0,283	-0,295

Из вышеприведенных таблиц можно видеть, что при малой доле пропусков (например, 10%) на случайном шаблоне оценки значений временного ряда получаются более точными, чем на детерминированном

шаблоне. Однако при увеличении доли пропусков оценки как параметров, так и значений на случайном шаблоне становятся неустойчивыми. Это связано с тем, что присутствие пропусков в начале наблюдений сильно ухудшает качество оценок, а в детерминированном шаблоне мы намеренно вставляли пропуски не в начало наблюдений, а ближе к середине. Впрочем, если повторить эксперимент для детерминированного шаблона, но добавить пропуски в начальные моменты наблюдений (например, с 3-го по 53-ый момент), то имеет место та же картина неустойчивости, что мы видим в таблице для случайного шаблона.

Изучим свойства полученных оценок для значений временного ряда. Для этого построим одношаговый прогноз на этой же реализации временного ряда и проанализируем ковариационную матрицу, полученную на этом шаге. Ковариационная матрица выглядит следующим образом:

$$\Sigma_{(100|100)} = \begin{pmatrix} 0.001 & -0.001 \\ -0.001 & 0.001 \end{pmatrix}.$$

Отметим, что в качестве начального приближения для ковариационной матрицы использовалась единичная матрица с диагональными элементами, равными 10. Таким образом, видим, что диагональные элементы сходятся к нулевым значениям.

Библиографические ссылки

1. *Kalman R.E., Busy R.C.* New results in liner filtring and prediction theory. Trans. ASME, J. of Basic Engineering, 1961. V.83, №1.

2. *Бокс Дж., Дженкинс Г.* Анализ временных рядов. М. : Мир, 1974.

3. *Durbin J., Harvey A., Koopman S. J., Shephard N.* State Space and Unobserved Component Models: Theory and Applications. Shephard Cambridge University Press, 2004.

4. *Charles K. Chui, Guanrong Chen* Kalman Filtering with Real-Time Applications. Springer, 2009.

5. *Липцер Р. Ш., Ширяев А.Н.* Статистика случайных процессов. М : Наука, 1974.