Применение Фильтра Калмана в задачах оценивания и прогнозирования.

Часть I

1 Введение

Модели пространства состояний

В процессе изучения современной экономики довольно большое значение имеет динамика ненаблюдаемых переменных,то есть таких, которые не могут быть непосредственно изучены напрямую, а только через какие-либо третьи переменные. Они встречаются при изучении безработицы, динамики инфляции, индексов экономической активности, поведения потребителя, воздействия монетарной политики, финансовых рынков и так далее.

Для моделирования динамики ненаблюдаемых процессов используется так называемое представление в форме пространства состояний *(state-space model)*. Само понятие, первоначально было введено в теории управления, для описания динамической системы, но универсальность применения данного представления привела к тому, что оно было вполне удачно внедрено в процесс изучения динамики ненаблюдаемых переменных в экономике. Основная идея данного представления заключается в том, что выводы о динамике ненаблюдаемой переменной, делаются через динамику косвенных переменных. Раскроем эту идею более подробно далее.

Предположим, что имеется некоторая ненаблюдаемая переменная X_k (здесь и далее нижний индекс k будет обозначать момент времени, в котором рассматривается переменная) и целью ставится изучение её динамики. Данная переменная в модели пространства состояний называется *состоянием.* Также предположим, что существует некоторая наблюдаемая переменная Y_k и, так как она наблюдаемая, по ней имеются все значения в каждый момент времени для k=1,2,.... Более того, предположим, что каким-то образом известна взаимосвязь между наблюдаемой переменной и ненаблюдаемой - уравнение описывающее её называется *уравнением наблюдений (observations)*, или *уравнение измерения (measurement equation)* (в зависимости от источника можно встретить разные названия, в данной работе, будем придерживаться первого варианта). Однако это уравнение заключает в себе и шумы. Также пусть имеется уравнение динамики самого ненаблюдаемого процесса, которое называется *уравнением состояний (state equation) или уравнением перехода (transition equation)*, (в работе будем придерживаться первого варианта). Данное уравнение также содержит в себе некоторый шум. Тогда в общем виде данную модель в форме пространства состояний можно представить следующим образом:

(1) $X_{k+1}=A_kX_k+B_kV_{k+1} \quad \in R^m,$ k=0,1,... (уравнение состояний)
(2) $Y_k=C_kX_k+D_kW_k \quad \in R^d,$ k=0,1,... (наблюдения)

,где

A_k - матрица $m \times m$, B_k - матрица $m \times m$, V_k - модельный шум.

V_k iid $N(0, I_m)$, C_k - матрица $d \times m$, D_k - матрица $d \times d$,

W_k - шум наблюдений/измерений, W_k iid $N(0, I_d)$.

$X_0 \sim N(\mu, \Sigma)$

Предположим ошибки модели и измерений независимыми, в дальнейшем будет показана разница в алгоритме Фильтра Калмана для независимых и зависимых шумов.

Также в разных источниках можно встретить следующую систему уравнений модели пространства состояний:

(1*) $X_{k+1}=A_k X_k+V_{k+1} \quad \in R^m$, k=0,1,... (уравнение состояний)

(2*) $Y_k=C_k X_k+W_k \quad \in R^d$, k=0,1,... (наблюдения)

Со следующими предположениями:

X_0, V_k, W_k и Y_k, как и в предыдушем представлении подчиняются гауссовскому закону распределения и первые три независимы между собой. Отличие состоит в том, что теперь ковариационные матрицы V_k и W_k не единичные, что представляется в следующем виде:

$E[V_k V_l^*] = \delta_{kl} Q_k \quad E[W_k W_l^*] = \delta_{kl} R_k$

$X_0 \sim N(\mu, \Sigma)$

Введем обозначения для двух возможных случаев. При независимых шумах в двух уравнениях, $E[V_k W_l^*] = 0$. В случае зависимости введем обозначения для данной ковариационной матрицы: $E[V_k W_l^*] = \delta_{kl} S_k$ -данное обозначение будет нам полезно далее.

Таким образом, представление модели в одной их данных форм и будет называться формой пространства состояний. Одним из самых распространённых способов для оценки динамики ненаблюдаемой переменной является Фильтр Калмана. Перед переходом непосредственно к самому алгоритму, обратимся к вопросу, почему модели пространства состояний получили такое распространение и почему такое представление может быть во многом полезно.

2 Модели представимые в форме пространства состояний

Главное достоинство моделей пространства состояний, состоит в том, что многие модели, не включающие в себя ненаблюдаемые переменные, например, такие как, многие модели временных рядов, могут быть представлены в данной форме. Линейные модели с ненаблюдаемыми компонентами, также можно свести к форме моделей пространства состояний. Более того, большое количество нелинейных моделей также может быть представлено в данной форме. Рассмотрим несколько примеров таких представлений ниже. В процессе представления моделей в пространстве состояний будем пользоваться обозначениями уравнений (1) и (2).

№1. Регресионная модель с меняющимися коэффициентами (TVP)

Рассмотрим классическую запись модели:

$y_k = \sum_{i=1}^{m} \theta_{i,k} z_{ik} + \eta_k$

$\theta_{i,k+1} = \alpha_i \theta_{i,k} + \varepsilon_{i,k}$; *i=1,2,...m*

$E[\eta_k] = E[\varepsilon_{i,k}] = 0$; *i=1,2,...,m*

$var[\eta_k] = \sigma^2_{\eta_k} var[\varepsilon_{i,k}] = \sigma^2_{\varepsilon_{i,k}}$; *i=1,2,...,m,*

причем y_k и $z_{i,k}$ считаем наблюдаемыми, и необходимо оценить $\theta_{i,k}$

Тогда пусть вектор X_k имеет вид:

$$X_k = \begin{bmatrix} \theta_{1,k} \\ \theta_{2,k} \\ \vdots \\ \theta_{m,k} \end{bmatrix}, \text{ тогда } \; X_{k+1} = \begin{bmatrix} \theta_{1,k+1} \\ \theta_{2,k+1} \\ \vdots \\ \theta_{m,k+1} \end{bmatrix}$$

Уравнение (2) примет следующий вид: $X_{k+1} = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_m \end{bmatrix} \begin{bmatrix} \theta_{1,k} \\ \theta_{2,k} \\ \vdots \\ \theta_{m,k} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,k+1} \\ \varepsilon_{2,k+1} \\ \vdots \\ \varepsilon_{m,k+1} \end{bmatrix}$

То есть $A_k = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_m \end{bmatrix}$ и $V_{k+1} = \begin{bmatrix} \nu_{1,k+1} \\ \vdots \\ \\ \nu_{m,k+1} \end{bmatrix}$

Уравнение (1): $y_k = [z_{1,k}, \dots, z_{m,k}] \begin{bmatrix} \theta_{1,k} \\ \theta_{2,k} \\ \vdots \\ \theta_{m,k} \end{bmatrix} + \eta_k$

Основным минусом такого представления является то, что шумы присутствуют как в уравнении наблюдений, так и в уравнении состояний. Ниже будут показаны примеры, когда шум присутствует только в одном из уравнений.

№2. Процесс скользящего среднего - MA(q).

В общем виде процесс скользящего среднего записывается следующим образом:

$y_k = \mu + \varepsilon_k + \sum_{i=1}^{q} \theta_i \varepsilon_{k-i}$

$\varepsilon_k \;\; iid \;\; N(0, \sigma^2)$

$E[\varepsilon_k \varepsilon_t] = 0, \;\; k \neq t$

Пусть $X_k = \begin{bmatrix} \varepsilon_k \\ \varepsilon_{k-1} \\ \vdots \\ \varepsilon_{k-q} \end{bmatrix}$

Уравнение (1): $X_{k+1} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0_{q+1,q+1} \end{bmatrix} X_k + \begin{bmatrix} \varepsilon_{k+1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

Уравнение (2): $y_k = \mu + [1, \theta_1, \dots, \theta_q] \begin{bmatrix} \varepsilon_k \\ \varepsilon_{k-1} \\ \vdots \\ \varepsilon_{k-q} \end{bmatrix} = \mu + [1, \theta_1, \dots, \theta_q] X_k$

В отличии от прошлого представления, можно заметить, что в уравнении (2) нет модельного шума.

№3. Авторегрессионный процесс порядка p - AR(p)

В общем виде процесс AR(p) можно записать в виде:

$(y_{k+1} - \mu) = \phi_1(y_k - \mu) + \phi_2(y_{k-1} - \mu) + ... + \phi_p(y_{k-p+1} - \mu) + \varepsilon_{k+1}$

$V(\varepsilon_k) = \sigma^2 \qquad E[\varepsilon_k \varepsilon_t] = 0, \;\; k \neq t$

Пусть

$X_k = \begin{bmatrix} y_k - \mu \\ y_{k-1} - \mu \\ \vdots \\ y_{k-p+1} - \mu \end{bmatrix}$

Тогда уравнение (1) примет вид: $X_{k+1} = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0_{pp} \end{bmatrix} X_k + \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \varepsilon_{k+1}$

Уравнение (2): $y_k = \mu + [1, 0, \dots, 0] X_k$

Аналогично с предыдущей моделью шум присутствует только в одном из уравнений, а именно в уравнении состояний.

№4. Модель авторегрессии — скользящего среднего - ARMA (p,q)

Как правило, всегда есть несколько способов представления модели (если она представима) в виде модели пространства состояний. И так как процесс ARMA весьма популярен и используется во многих работах, нами будут рассмотрены два возможных способа представления данной модели.

В общем виде процесс ARMA (p,q) может быть записан в виде:
$y_{k+1} = \phi_1 y_k + \phi_2 y_{k-1} + ... + \phi_p y_{k-p+1} + \varepsilon_{k+1} + \sum_{i=1}^{q} \theta_i \varepsilon_{k-i+1}$
$V(\varepsilon_k) = \sigma^2 \qquad E[\varepsilon_k \varepsilon_t] = 0$

Способ 1.(Aoki) Выше были рассмотрены модели AR(p) и MA(q), так как процесс ARMA в свою очередь "состоит"из них, то можно определить вектор состояний следующим образом:

$$X_k = \begin{bmatrix} y_k \\ y_{k-1} \\ \vdots \\ y_{k-p+1} \\ \varepsilon_k \\ \varepsilon_{k-1} \\ \vdots \\ \varepsilon_{k-q+1} \end{bmatrix}$$

Тогда уравнение (1):

$$X_{k+1} = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p & \theta_1 & \theta_2 & \dots & \theta_{q-1} & \theta_q \\ 1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0_{pp} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0_{q+p,q+p} \end{bmatrix} X_k + \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1_{p+1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \varepsilon_{k+1}$$

Уравнение (2): $y_k = [1, 0, \dots, 0] X_k$

Однако, нетрудно, заметить, что размерность матрицы $A_k : (q+p) \times (q+p)$, а самого вектора состояний (q+p). Поэтому ниже будет рассмотрено альтернативное представление ARMA(p,q).

Способ 2.(Hamilton) В данном подходе вводится переменная состояния

$\xi_k \in R^{max[p,q+1]}$ В общем виде и в матричной форме подход Гамильтона описывается следующим уравнением состояния:

$\xi_{k+1} = A\xi_k + \nu_{k+1}$

Пусть r=max[p,q+1].

Примем при j>p $\phi_j = 0$ и $\theta_j = 0$ при j>q. Тогда первоначальная модель примет вид:

$y_k = \phi_1 y_{k-1} + \phi_2 y_{k-2} + ... + \phi_r y_{k-r} + \varepsilon_k + \sum_{i=1}^{r-1} \theta_i \varepsilon_{k-i+1}$

Тогда в явном виде уравнение (1) выглядит следующим образом:

$$\xi_{k+1} = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_r \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0_{rr} \end{bmatrix} \xi_k + \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \varepsilon_{k+1}$$

Уравнение (2): $y_k = [1, \theta_1, \theta_2 \dots, \theta_{r-1}]\xi_k$

Найдем какими должны быть элементы данного вектора ξ_k и убедимся в том, что данная модель является представлением исходной модели.

Пусть $\xi_{j,k}$ - jый элемент вектора ξ_k, тогда зная, что второй элемент вектора ξ_{k+1} равен первого элементу вектора ξ_k, то есть в введенных обозначениях

$\xi_{2,k+1}$=$\xi_{1,k}$, тогда

$\xi_{3,k+1} = \xi_{2,k} = \xi_{1,k-1}$, и так далее. В общем виде:

$\xi_{j,k+1}$=$L^{j-1}\xi_{1,k+1}$, где L-оператор сдвига

Из уравнения (1) для первого элемента вектора ξ_{k+1} получим следующее равенство:

$\xi_{1,k+1}$=$(\phi_1 + \phi_2 L + \phi_3 L^2 + ... + \phi_r L^r)\xi_{1,k} + \varepsilon_{k+1}$

$(1 - \phi_1 L - \phi_2 L^2 - ... - \phi_r L^r)\xi_{1,k+1} = \varepsilon_{k+1}$

Из уравнения (2), уравнения наблюдений, имеем:

$y_k = (1 + \theta_1 L + \theta_2 L^2 + ... + \theta_{r-1} L^{r-1})\xi_{1,k}$

Умножим данное уравнение на $(1 - \phi_1 L - \phi_2 L^2 - ... - \phi_r L^r)$, получим

$(1-\phi_1 L-\phi_2 L^2-...-\phi_r L^r)y_k = (1+\theta_1 L+\theta_2 L^2+...+\theta_{r-1} L^{r-1})(1-\phi_1 L-\phi_2 L^2-...-\phi_r L^r)\xi_{1,k}$

$(1 - \phi_1 L - \phi_2 L^2 - ... - \phi_r L^r)y_k = (1 + \theta_1 L + \theta_2 L^2 + ... + \theta_{r-1} L^{r-1})\varepsilon_k$

Раскроем скобки и получим исходную ARMA модель:

$y_k - \phi_1 y_{k-1} - ... - \phi_r y_{k-r} = \varepsilon_k + \theta_1 \varepsilon_{k-1} + ... + \theta_{r-1}\varepsilon_{k-r+1}$

Размерность вектора состояния гарантированно (если p и/или q не равны нулю) уменьшилась, что делает в дальнейшем процедуру фильтрации более оптимальной. Отметим, что модель несложно расширяется до модели ARMAX, путём добавления значений в вектор состояния.

№5. Интегрированная модель авторегрессии — скользящего среднего ARIMA(p,d,q).

Рассмотрим ARMA (p,q), которую в общем виде можно записать следующим образом:

$y_k = \phi_1 y_{k-1} + \phi_2 y_{k-2} + ... + \phi_p y_{k-p} + \varepsilon_k + \sum_{i=1}^{q} \theta_i \varepsilon_{k-i}$

$V(\varepsilon_k) = \sigma^2; \qquad E[\varepsilon_k \varepsilon_t] = 0,\ k \neq t$

Причем не будем исключать возможности, что корни $\phi(B)$ лежат внутри единичной окружности или на ней, то есть не исключаем модели ARIMA из анализа. Ниже представлено еще одно представление моделей ARMA, которое можно распостранить на ARIMA.

Данный способ был предложен Pearlman(1980.) Пусть m=max[p,q] и вектор состояния X_k - это вектор состояния $(m \times 1)$. Примем при j>p $\phi_j = 0$ и $\theta_j = 0$ при j>q.

Уравнение (1): $X_{k+1} = \begin{bmatrix} \phi_1 & 1 & 0 & \dots & 0 \\ \phi_2 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{m-1} & 0 & \dots & 0 & 1 \\ \phi_m & 0 & \dots & 0 & 0_{m,m} \end{bmatrix} X_k + \begin{bmatrix} \theta_1 + \phi_1 \\ \theta_2 + \phi_2 \\ \vdots \\ \theta_{m-1} + \phi_{m-1} \\ \theta_m + \phi_m \end{bmatrix} \varepsilon_{k+1}$

Уравнение (2): $Y_k = [1, 0, \dots, 0] X_k + \varepsilon_k$

Аналогично, предыдущему способу представления, найдем чему равны элементы вектора состояний X_k и покажем, что данное представление сводится к исходной ARMA модели.

Пусть $x_{j,k}$- jый элемент вектора состояния X_k, из уравнения (2), уравнения наблюдений, получим, что

$x_{1,k} = y_k - \varepsilon_k$

Из уравнения (1):

$x_{j,k+1} = \phi_j(y_k - \varepsilon_k) + x_{j+1,k} + (\theta_j + \phi_j)\varepsilon_k$

$= \phi_j y_k + x_{j+1,k} + \theta_j \varepsilon_k$

Раскрывая, подобным образом, $x_{j+1,k}$ и так далее, получим:

$x_{j,k+1} = \phi_j y_k + ... + \phi_m y_{k-m+j} + \theta_j \varepsilon_k + ... + \theta_m \varepsilon_{k-m+j}$

Подставим j=1 и получим:

$x_{1,k} = \phi_1 y_{k-1} + ... + \phi_m y_{k-m} + \theta_1 \varepsilon_{k-1} + ... + \theta_m \varepsilon_{k-m}$

Из уравнения наблюдений: $x_{1,k} + \varepsilon_k = y_k$; подставляя $x_{1,k}$ получим:

$y_k = \phi_1 y_{k-1} + ... + \phi_m y_{k-m} + \theta_1 \varepsilon_{k-1} + ... + \theta_m \varepsilon_{k-m} + \varepsilon_k$

Это представление называется max(p,q). В случае есть max(p,q+1)=q+1, то размерность вектора состояния в данном случае, по сравнению с предыдущем меньше, а значит и матрицы A_k. Отметим, что здесь приведена форма модели в общем случае, так при небольших значениях d, вектор состояния принимает более простой вид (с частными примерами d=1 и d=2 можно ознакомиться в J. Durbin,S. Koopman (2012)).

№6. Сезонная модель с шумом.

Пусть s_k - циклическая компонента с периодом c, такая что:

$s_{k+c} = s_k$ и $\sum_{i=1}^{c} s_{k+j} = 0$ *для любого k.*

Имея c-1 значений $s_k : s_1, s_2, ..., s_{c-1}$, тогда зная, что

$s_c = -s_1 - s_2 - ... - s_{c-1}$ и $s_{k+c} = s_k \ \forall \ k = 1, 2, ...$

s_k может быть найдено следующим образом:

$s_k = -s_{k-c+1} - s_{k-c+2} - ... - s_{k-1}$ Добавим шум в данное уравнение.

$s_k = -s_{k-c+1} - s_{k-c+2} - ... - s_{k-1} + \eta_k$, где $\eta_k \ iid \ N(0, \sigma_\eta^2)$

Запишем модель следующим образом:

$y_k = s_k + \varepsilon_k$
$s_k = -s_{k-c+1} - s_{k-c+2} - ... - s_{k-1} + \eta_k,$

В матричном виде и в терминах уравнений (1) и (2) запишем модель в форме пространства состояний.

$$X_k = \begin{bmatrix} s_k \\ s_{k-1} \\ s_{k-2} \\ \vdots \\ s_{k-c+2} \end{bmatrix}$$

$$X_{k+1} = \begin{bmatrix} -1 & -1 & \dots & -1 & -1 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} X_k + \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \eta_k$$

$y_k = [1, 0, ..., 0]X_k + \varepsilon_k$

Ранее нами были рассмотрены модели в общем виде, для того чтобы показать, что они без ограничений могут быть представлены в виде модели пространства состояний. Ниже нами будет рассматриваться модели, которые включает в себя процессы AR(p) и ARMA(p,q) и TVP.Для большей лаконичности записей в данной главе в моделях нами будут взяты небольшие значений p и q, и всего несколько экзогенных переменных, но стоит понимать, что при желании они могут быть несложно расширены до интересующего исследователя порядка.

№7. Модель с ненаблюдаемыми компонентами.

Данная модель была введена в Clark(1987). Предположим, что имеется некоторая наблюдаемая величина, которая раскладывается на две ненаблюдаемые компоненты (большее количество компонент наблюдаемого значения существенно не усложнит модель, а лишь увеличит размерность вектора состояния X_k и матрицы A_k: стохастический тренд и циклическая компонента. Предложенная модель выглядит следующим образом:

$y_k = n_k + z_k$
$n_{k+1} = \rho n_k + g_k + \varepsilon_{k+1}$ (тренд)
$g_k = \gamma g_{k-1} + w_k$ (уровень роста)
$z_k = \phi_1 z_{k-1} + \phi_2 z_{k-2} + e_k$
$E[e_k] = E[\varepsilon_k] = E[w_k] = 0 \quad V[e_k] = \sigma_e^2 \; V[\varepsilon_k] = \sigma_{\varepsilon_k}^2 \; V[w_k] = \sigma_{w_k}^2$
$E[e_k \varepsilon_k] = E[e_k w_k] = E[\varepsilon_k w_k] = 0$

Представим её в виде модели пространства состояний.

Пусть $X_k = \begin{bmatrix} n_k \\ z_k \\ z_{k-1} \\ g_k \end{bmatrix}$; $X_{k+1} = \begin{bmatrix} \rho & 0 & 0 & 1 \\ 0 & \phi_1 & \phi_2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \gamma \end{bmatrix} X_k + \begin{bmatrix} \varepsilon_{k+1} \\ e_{k+1} \\ 0 \\ w_{k+1} \end{bmatrix}$; $Y_k = [1, 1, 0, 0]X_k$

В данной модели вновь нет шумов в уравнении наблюдений, а есть только в уравнении состояний.

№8. Динамическая факторная модель.

Предположим, что у имеется две наблюдаемые переменные, которые зависят от общего ненаблюдаемого фактора, и каждая из переменных от своего набора экзогенных переменных, причем уравнения их динамики также известны. Тогда модель примет вид:

$y_{1,k} = \gamma_1 c_k + \beta_1 z_{1,k}$

$y_{2,k} = \gamma_2 c_k + \beta_2 z_{2,k}$

$c_k = \phi_1 c_{k-1} + \varepsilon_k$

$z_{i,k} = \alpha_1 z_{i,k-1} + w_{i,k}$; *i=1,2*

$E[w_{1,k}] = E[\varepsilon_k] = E[w_{2,k}] = 0$

$V[w_{1,k}] = \sigma^2_{w_{1,k}}\ V[\varepsilon_k] = \sigma^2_{\varepsilon_k}\ V[w_{2,k}] = \sigma^2_{w_{2,k}}$

$E[w_{1,k}\varepsilon_k] = E[w_{1,k}w_{2,k}] = E[\varepsilon_k w_{2,k}] = 0$

Зададим вектор состояния следующий образом:

$X_k = \begin{bmatrix} c_k \\ z_{1,k} \\ z_{2,k} \end{bmatrix}$ тогда, $X_{k+1} = \begin{bmatrix} \phi_1 & 0 & 0 \\ 0 & \alpha_1 & 0 \\ 0 & 0 & \alpha_2 \end{bmatrix} X_k + \begin{bmatrix} \varepsilon_{k+1} \\ w_{1,k+1} \\ w_{2,k+1} \end{bmatrix}$ и

$Y_k = \begin{bmatrix} y_{1,k} \\ y_{2,k} \end{bmatrix} = \begin{bmatrix} \gamma_1 & \beta_1 & 0 \\ \gamma_2 & 0 & \beta_2 \end{bmatrix} X_k$

№9. Динамическая факторная модель с общим стохастическим трендом.

Чтобы еще раз продемонстрировать универсальность данного представления, соединим модели №7 и №8, более того, предположим, что параметры AR(1) для переменной $z_{i,k}$ (модель №8), также меняются во времени, то есть объединим модели еще и с TVP.

Запишем уравнения модели в явном виде:

$y_{1,k} = c_k + \beta_1 z_{1,k}$

$y_{2,k} = c_k + \beta_2 z_{2,k}$

$c_k = c_{k-1} + w_k$

$z_{i,k} = \alpha_{i,k-1} z_{i,k-1} + \varepsilon_{i,k}$; *i=1,2*

$\alpha_{i,k} = \beta_i \alpha_{i,k-1} + e_{i,k}$; *i=1,2*

$E[w_k] = E[\varepsilon_{1,k}] = E[\varepsilon_{2,k}] = E[e_{1,k}] = E[e_{2,k}] = 0$

$V[w_k] = \sigma^2_{w_k}\ V[\varepsilon_{i,k}] = \sigma^2_{\varepsilon_{i,k}}\ V[e_{i,k}] = \sigma^2_{e_{i,k}}$; *i=1,2*

$E[w_k]\varepsilon_{i,k}] = E[e_{i,k}\varepsilon_{i,k}] = E[e_{i,k}w_k] = 0$; *i=1,2*

Пусть вектор состояний имеет вид: $X_k = \begin{bmatrix} c_k \\ z_{1,k} \\ z_{2,k} \\ \alpha_{1,k} \\ \alpha_{2,k} \end{bmatrix}$

Уравнение (1):

$$X_{k+1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \alpha_{1,k} & 0 & 0 & 0 \\ 0 & 0 & \alpha_{2,k} & 0 & 0 \\ 0 & 0 & 0 & \beta_1 & 0 \\ 0 & 0 & 0 & 0 & \beta_2 \end{bmatrix} X_k + \begin{bmatrix} w_{k+1} \\ \varepsilon_{1,k+1} \\ \varepsilon_{2,k+1} \\ e_{1,k+1} \\ e_{2,k+1} \end{bmatrix}$$

Уравнение (2) примет вид: $Y_k = \begin{bmatrix} y_{1,k} \\ y_{2,k} \end{bmatrix} = \begin{bmatrix} 1 & \beta_1 & 0 & 0 & 0 \\ 1 & 0 & \beta_2 & 0 & 0 \end{bmatrix} X_k$

Как можно заметить, в данном примере A_k непостоянная и меняется с изменением k, в отличие от предыдущих моделей.

Очевидно, что примеры моделей сводимых к форме пространства состояний, не ограничиваются приведенными выше. Но даже из них можно заключит, что довольно большая часть динамических моделей (например распространённые модели MA, AR, ARMA, ARIMA) могут быть представлены в виде модели пространства состояний, что делает её универсальной. Значит алгоритм Фильтра Калмана, применимый к моделям пространства состояний, также является универсальным и может быть применен к широкому классу моделей. Следовательно всё то, что будет описано в ходе данной работы, может быть применено ко всем моделям приведенным выше и к их расширениям.

К моделям, описанным выше, можно добавить, что фильтр Калмана также может быть использован, например, при оценивании авторегрессионных моделей условной гетероскедастичности (ARCH). Алгоритм данного фильтра используется при оценивании ненаблюдаемой переменной в моделях с марковскими переключениями (Markov Switching Model), как часть фильтра Кима (Kim's Filter), что еще на целый класс моделей расширяет применимость Фильтра Калмана. Все описанные выше факты, делают изучение Фильтра Калмана актуальным и полезным.

3 Фильтр Калмана.

Напомним уравнения (1^*) и (2^*) представленные выше:

(1^*) $X_{k+1}{=}A_kX_k{+}V_{k+1} \quad \in R^m$, k=0,1,... (уравнение состояний)

(2^*) $Y_k{=}C_kX_k{+}W_k \quad \in R^d$, k=0,1,... (наблюдения)

В данной главе будем предполагать соблюдение всех предположений сделанных ранее, и то что известны истинные вектора параметров модели в каждой момент времени *k=0,1,2,...*; θ_k, которые состоят из:

θ_k= $[\mu, \Sigma, A_k, C_k, Q_k, R_k, S_k]$

Основная цель данной модели, это оценить ненаблюдаемый вектор состояний X_k в каждый момент времени на основе имеющих наблюдаемых переменных, и ошибку возникающую при данном оценивании. То есть для каждого k=0,1,2,..., необходимо найти:

$\hat{X}_{k|k}$=$E[X_k|\mathcal{Y}_k]$

$\Sigma_{k|k}$= $E[(X_k - \hat{X}_{k|k})(X_k - \hat{X}_{k|k})^*]$

Основная проблема оценивания в том, что, несмотря на то, что зависимость между вектором состояний и наблюдаемым вектором предполагается известной (уравнение наблюдений), уравнение данной зависимости тоже содержит шумы.

Для получения данных оценок используется рекурсивная процедура, известная как Фильтр Калмана (Kalman Filter). С подробным алгоритмом вывода Фильтра Калмана можно ознакомиться практически в любом учебнике, где упоминается данный алгоритм, поэтому ниже приведены только основные формулы алгоритма и общие идеи их получения.

Рекурсивная процедура.

Основная идея Фильтра Калмана (ФК) - это разделить процедуру, оценивания неизвестного вектора состояния на два этапа: предсказание (prediction) и коррекция (correction).

Начинается процедура с оценивания $\hat{X}_{0|0}$ и $\Sigma_{0|0}$, исходя из следующего:

$X_0 \sim N(\mu, \Sigma)$; $Y_0 = C_0 X_0 + W_0$; $\hat{X}_{0|0}$=$E[X_0|Y_0]$

Нетрудно получить:

$\hat{X}_{0|0}$=$\mu + \Sigma C_0{}^*(C_0 \Sigma C_0{}^* + R_0)^{-1}(Y_0 - C_0\mu)$

$\Sigma_{0|0} = \Sigma - \Sigma C_0{}^*(C_0 \Sigma C_0{}^* + R_0)^{-1} C_0 \Sigma$

Затем, предположим, что на шаге k, нам известны,

$\hat{X}_{k|k}$=$E[X_k|\mathcal{Y}_k]$

$\Sigma_{k|k}$= $E[(X_k - \hat{X}_{k|k})(X_k - \hat{X}_{k|k})^*|\mathcal{Y}_k]$=$E[(X_k - \hat{X}_{k|k})(X_k - \hat{X}_{k|k})^*]$

Тогда, до того как поступило наблюдение Y_{k+1}, на основе $\mathcal{Y}_k$ "прогнозируем"значения вектора состояний в момент времени *k+1*,на основе уравнений (1*) и (2*) то есть:

$\hat{X}_{k+1|k}$=$A_k \hat{X}_{k|k}$

$\Sigma_{k+1|k}$=$E[(X_{k+1} - \hat{X}_{k+1|k})(X_{k+1} - \hat{X}_{k+1|k})^*|\mathcal{Y}_k]$=

$E[(A_k(X_k - \hat{X}_{k|k}) + V_{k+1})(A_k(X_k - \hat{X}_{k|k}) + V_{k+1})^*] = A_k \Sigma_{k|k} A_k{}^* + Q_{k+1}$

Затем поступает новое наблюдение Y_{k+1} и начинается "корректировка"полученных раннее прогнозных значений.

Введем следующие обозначения:

$\theta = X_{k+1} - E[X_{k+1}|\mathcal{Y}_k] = X_{k+1} - \hat{X}_{k+1|k}$

$\xi = Y_{k+1} - E[Y_{k+1}|\mathcal{Y}_k] = Y_{k+1} - \hat{Y}_{k+1|k}$= ν_{k+1} - инновации (innovations).

(θ, ξ)- гауссовские случайные величины, независимые от $\mathcal{Y}_k$

$E[\theta|\xi] = E[\theta|\xi, \mathcal{Y}_k] = E[\theta|\mathcal{Y}_{k+1}] = \hat{X}_{k+1|k+1} - \hat{X}_{k+1|k}$

Тогда оценка вектора состояний в периоде *k+1* на основе информации $\mathcal{Y}_{k+1}$ следующим образом:

$\hat{X}_{k+1|k+1} = \hat{X}_{k+1|k} + E[\theta|\xi]$;

Выражая $E[\theta|\xi]$, как $E[\theta|\xi] = E[\theta] + V_{\theta\xi}(V_{\xi\xi})^{-1}(\xi - E[\xi])$, получим:

$\hat{X}_{k+1|k+1} = \hat{X}_{k+1|k} + G_{k+1}\nu_{k+1}$, где

$G_{k+1} = \Sigma_{k+1|k}{C_{k+1}}^* {H_{k+1|k}}^{-1}$ - усиление Калмана (Kalman Gain)

$H_{k+1|k} = C_{k+1}\Sigma_{k+1|k}{C_{k+1}}^* + R_{k+1} = E[\nu_{k+1}{\nu_{k+1}}^*]$ - ковариация инноваций.

На заключительном этапе *k+1* шага рассчитывается матрица ошибки оценивания, с учетом полученного наблюдения Y_{k+1} $(\Sigma_{k+1|k+1})$.

$cov(\theta, \theta|\xi) = E[(\theta - E[\theta|\xi])(\theta - E[\theta|\xi])^*|\xi] = E[(\theta - E[\theta|\xi])(\theta - E[\theta|\xi])^*] =$

$E[(X_{k+1} - \hat{X}_{k+1|k+1})(X_{k+1} - \hat{X}_{k+1|k+1})^*] = \Sigma_{k+1|k+1}$

Таким образом, зная оценки на шаге *k*,получаем следующие уравнения для вычисления оценок на *(k+1)* шаге:

Уравнения прогнозирования:

$\hat{X}_{k+1|k} = A_k\hat{X}_{k|k}$

$\Sigma_{k+1|k} = A_k\Sigma_{k|k}{A_k}^* + Q_{k+1}$

Уравнения инноваций:

$\nu_{k+1} = Y_{k+1} - C_{k+1}\hat{X}_{k+1|k}$

$H_{k+1|k} = C_{k+1}\Sigma_{k+1|k}{C_{k+1}}^* + R_{k+1}$

Усиление Калмана:

$G_{k+1} = \Sigma_{k+1|k}{C_{k+1}}^* {H_{k+1|k}}^{-1}$

Уравнения корректировки:

$\hat{X}_{k+1|k+1} = \hat{X}_{k+1|k} + G_{k+1}\nu_{k+1}$

$\Sigma_{k+1|k+1} = (I - G_{k+1}C_{k+1})\Sigma_{k+1|k}$

Отметим, что уравнения приведенные выше имеют место для некоррелированных шумов в уравнениях (1*) и (2*). С подробными выводами данных формул (в том числе в случае коррелированных остатков) можно ознакомиться, например, в первоисточнике - Kalman (1960) или в работах Jonh van der Hoek (2009).

4 Сглаживание Калмана (Kalman Smoothers).

Если в рамках алгоритма Фильтра Калмана вычисляются оценки вектора состояний в периоде *k+1* на основе $\mathcal{Y}_l$, где l=k, то есть k>l, то в ходе сглаживаний Калмана цельвычислить:

$\hat{X}_{k|l} = E[X_k|\mathcal{Y}_l]$

$\Sigma_{k|l} = E[(X_k - \hat{X}_{k|l})(X_k - \hat{X}_{k|l})^*|\mathcal{Y}_k]$, при l>k

Интуитивно понятно, что идея данного алгоритма в получении новых оценок вектора состояний в каждый момент времени. Так если имеется N наблюдений, то оценка ФК для вектора состояния в момент k+1 условна на $\mathcal{Y}_{k+1}$, то в сглаживании Калмана, она условна на $\mathcal{Y}_N$, в то время как $\mathcal{Y}_N$ содержит в себе $\mathcal{Y}_{k+1}$.

Рекурсия для $\hat{X}_{k|l}$.

Обозначим,

$\theta = X_k - \hat{X}_{k|k}$

$\xi = X_{k+1} - \hat{X}_{k+1|k}$

θ и ξ - гауссовские случайные величины. Тогда:

$E[\theta|\xi] = E[\theta] + V_{\theta\xi}{V_{\xi\xi}}^{-1}[\xi - E[\xi]]$

Выражая неизвестные члены, через параметры модели, получим следующее выражение:

$X_k - \hat{X}_{k|k} = \Sigma_{k|k} A_k^* \Sigma_{k+1|k}^{-1}(X_{k+1} - \hat{X}_{k+1|k}) + Z_{k+1}$; где Z_{k+1} - не зависит от $X_{k+1} - \hat{X}_{k+1|k}$

Возьмем условное математическое ожидание $E[...|\mathcal{Y}_l]$ от левой и правой части.

$\hat{X}_{k|l} - \hat{X}_{k|k} = J_k(\hat{X}_{k+1|l} - \hat{X}_{k+1|k})$, где $J_k = \Sigma_{k|k} A_k^* \Sigma_{k+1|k}^{-1}$

$\hat{X}_{k|l} = \hat{X}_{k|k} + J_k(\hat{X}_{k+1|l} - \hat{X}_{k+1|k}) = (I - J_k A_k)\hat{X}_{k|k} + J_k \hat{X}_{k+1|l}$

Так как оценка $\hat{X}_{k|l}$ в выражении выше, зависит от $\hat{X}_{k+1|l}$, сглаживание производиться с предпоследнего значения. То есть, имея конечную выборку, N, сглаживание начинается со значения, соответствующего периоду N-1, и заканчивается нулевым периодом. Или же данную формулу можно преобразовать к иному виду, раскрывая$\hat{X}_{k+1|l}$ согласно этому же выражению:

$\hat{X}_{k|l} = J_k^{l-k} \hat{X}_{l|l} + (\sum_{r=0}^{l-k-1} J_k^r (I_k - J_k A_k))\hat{X}_{k|k}$

Как можно заметить, в данном выражении присутствуют только оценки полученные на этапе фильтрации. То есть до сглаживания необходимо применить ФК, и только затем находить сглаженные значения вектора состояний.

Рекурсия для $\Sigma_{k|l}$.

Для получения рекурсивной формулы оценки матрицы ошибки возникающей при оценивании $\Sigma_{k|l}$, из уравнения:

$X_k - \hat{X}_{k|l} = X_k - \hat{X}_{k|k} - J_k(X_{k+1|l} - A_k \hat{X}_{k|k})$, получим:

$X_k - \hat{X}_{k|l} + J_k X_{k+1|l} = X_k - \hat{X}_{k|k} + J_k A_k \hat{X}_{k|k}$. Найдем ковариационные матрицы выражения слева и справа, приравняв, получим:

$\Sigma_{k|l} = \Sigma_{k|k} + J_k[\Sigma_{k+1|l} - \Sigma_{k+1|k}]J_k^*$, где $J_k = \Sigma_{k|k} A_k^* \Sigma_{k+1|k}^{-1}$, как и ранее.

Заметим, что слагаемые $\Sigma_{k|k}$,$\Sigma_{k+1|k}$ и J_k известны из ФК, и только слагаемое $\Sigma_{k+1|l}$ неизвестно. То есть в данном случае значение матрицы ошибки в периоде k, зависит от её значения в *k+1* периоде, значит нахождение данных сглаженных оценок производится с конца.

Предположим имеется конечная выборка *N*, тогда формула для оценки данной матрица в *N-1* периоде выглядит следующим образом:

$\Sigma_{N-1|N} = \Sigma_{N-1|N-1} + J_{N-1}[\Sigma_{N|N} - \Sigma_{N|N-1}]J_{N-1}^*$, где уже все слагаемые известны из ФК. Тогда для N-2, получим:

$\Sigma_{N-2|N} = \Sigma_{N-2|N-2} + J_{N-2}[\Sigma_{N-1|N} - \Sigma_{N-1|N-2}]J_{N-2}^*$, где все слагаемые уже известны. Процедура продолжается до $\Sigma_{0|N}$.

Резюмируем. Сглаженные оценки векторов состояния и сглаженные оценки матрицы ошибок возникающих при данном оцениванию находятся следующим образом:

$J_k = \Sigma_{k|k} A_k^* \Sigma_{k+1|k}^{-1}$

$\hat{X}_{k|l} = \hat{X}_{k|k} + J_k(\hat{X}_{k+1|l} - \hat{X}_{k+1|k})$

$\Sigma_{k|l} = \Sigma_{k|k} + J_k[\Sigma_{k+1|l} - \Sigma_{k+1|k}]J_k^*$

Заметим еще раз, что для того, чтобы получить сглаженные оценки, необходимо найти оценки, после ФК. Потому как они используются, для получения сглаженных значений.

Некоторые полезные результаты из алгоритма сглаживания.

Ниже без доказательства будут представлен некоторые полезные результаты, которые будут необходимы в следующих главах при оценивании параметров модели. Их подробный вывод можно найти в Shumway and Stoffer (2000), или же в работах John Van der Hoek (2009).

Обозначим ковариационную матрицу между соседними оценками вектора состояний при сглаживании, следующим образом:

$\Sigma_{k,l|n} = E[(X_k - \hat{X}_{k|n})(X_l - \hat{X}_{l|n})^*]$, *где* $k,l \leq n$

Тогда

$\Sigma_{n,n-1|n} = (I - G_n C_n) A_{n-1} \Sigma_{n-1|n-1}$, используется как начально условие,(значение известно из ФК).

Тогда ковариационная матрица сглаживания для соседних оценок вектора состояния можно быть рассчитана как:

$\Sigma_{k-1,k-2|n} = \Sigma_{k-1|k-1} J_{k-2}^* + J_{k-1}[\Sigma_{k,k-1|n} - A_{k-1}\Sigma_{k-1|k-1}] J_{k-2}^*$

Запомним данный результат, он будем полезен далее.

Часть II

Оценивание параметров Фильтра Калмана

Ранее мы предполагали, что параметры фигурирующие в ФК, а именно

$\theta_k = [\mu, \Sigma, A_k, C_k, Q_k, R_k, S_k]$ в каждый момент k=1,2,... нам известны. Также были показаны идеи фильтрации и сглаживания, и то что их реализация несложна при всех известных параметрах модели θ_k. Однако, данное предположение нереалистично и истинные значения параметров модели неизвестны. Именно методам оценки параметров ФК и будет уделено внимание в данной главе.

5 Необходимые понятия.

Рассмотрим вероятностное пространство с фильтрацией $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$

Пусть с некоторой **вероятностью (reference probability)** $\bar{P}$, случайные величины X_k и Y_l подчиненны следующим законом распределения.

$X_k \sim \mathcal{N}(0, I_n)$; $Y_l \sim \mathcal{N}(0, I_m)$

И с **вероятностью** P:

$X_0 \sim N(\mu, \Sigma)$

$X_{k+1} = A_k X_k + V_{k+1}$

$Y_k = C_k X_k + W_k$

$\{V_k\}$ независимы, $V_k \sim N(0, Q_k)$

$\{W_k\}$ независимы, $W_k \sim N(0, R_k)$

$\{V_k\}$ и $\{W_k\}$ независимы.

$\dfrac{dP}{d\bar{P}}$=$\bar{\Lambda}_t$, где $\bar{\Lambda}_t$=$\prod_{k=0}^{t}\lambda_k$, и λ_k удовлетворяет следующим условиям:

1)λ_t $(\mathcal{F}_t)$-измерима

2) λ_t>0 с вероятностью 1.

3)E$[\lambda_{t+1}|\mathcal{F}_t] = 1$

4)E$[\lambda_0] = 1$

Тогда верно следующее соотношение для выражения математического ожидания некоторой случайной величины A:

$E^P[A] = E^{\bar{P}}[\bar{\Lambda}_t A]$

Будем использовать $\phi(X)$ как обозначение для функции плотности вероятности для $\mathcal{N}(0, I_n)$ случайной величины, и $\psi(Y)$ как функцию плотности вероятности для $\mathcal{N}(0, I_m)$ другой случайной величины.

Выразим первый член $\bar{\Lambda}_t$, $\bar{\lambda}_0$:

$$\bar{\lambda}_0 = \frac{|\Sigma|^{-1/2}\phi(\Sigma^{-1/2}(X_0-\mu))}{\phi(X_0)} \cdot \frac{|R_0|^{-1/2}\psi(R_0^{-1/2}(Y_0-C_0X_0))}{\psi(Y_0)}$$

$\bar{\Lambda}_0$=$\bar{\lambda}_0$

Несложно показать, что E$[\lambda_0]$ действительно равно 1.

Для $k \geq 1$

$$\bar{\lambda}_k = \frac{|Q_k|^{-1/2}\phi(Q_k^{-1/2}(X_k - A_{k-1}X_{k-1})}{\phi(X_0)} \cdot \frac{|R_k|^{-1/2}\psi(R_k^{-1/2}(Y_k - C_kX_k))}{\psi(Y_0)}$$

Таким образом, можно найти выражение $\bar{\lambda}_k$ для каждого k, а значит можно выразить $\bar{\Lambda}_t$=$\prod_{k=0}^{t}\lambda_k$.

Функция правдоподобия.

Предположим здесь и далее, что параметры модели не зависят от момента времени k, тогда вектор параметров, который надо оценить имеет вид: θ= $[\mu, \Sigma, A, C, Q, R, S]$

Необходимо выбрать критерий для выбора θ, основанный только на наблюдениях $Y_1, Y_2, ..., Y_N$. Фактически у на имеются только их наблюдения $y_1, y_2, ..., y_N$

Для произвольного $\varepsilon > 0$ пусть A_ε=$\{\omega \in \Omega | Y_i(\omega) \in (y_i - \varepsilon, y_i + \varepsilon)\} \in \mathcal{Y}_N$ для i=1,2,...,N. Тогда:

$P(A_\varepsilon) = E[I(A_\varepsilon)] = \bar{E}[\bar{\Lambda}_N I(A_\varepsilon)] = \bar{E}[\bar{E}[\bar{\Lambda}_N I(A_\varepsilon)]|\mathcal{Y}_N]$= $\bar{E}[I(A_\varepsilon)\bar{E}[\bar{\Lambda}_N|\mathcal{Y}_N]]$= $\bar{E}[I(A_\varepsilon)\bar{E}[\bar{\Lambda}(\theta)|\mathcal{Y}_N]]$

Выбирая параметр θ, чтобы максимизировать $P(A_\varepsilon)$ для любого $\varepsilon > 0$, он выбирает таким образом, чтобы максимизировать:

$\bar{E}[\bar{\Lambda}(\theta)|\mathcal{Y}_N]$=$\bar{E}[\bar{\Lambda}_N|\mathcal{Y}_N]$= $L(\theta)$, где $L(\theta)$ - функция правдоподобия, построенная на наблюдениях $y_1, y_2, ..., y_N$.

6 Оценка метода максимального правдоподобия (MLE).

Начинаем с некоторой оценки вектора θ_0. Затем пытаемся найти лучшие оценки. После нескольких шагов имеем оценку $\theta = \theta_j$ и ищем некоторую лучшую оценку $\hat{\theta} = \theta_{j+1}$ Зная оценку вектора параметров θ на предыдущем шаге, вычисляем инновации и их ковариационные матрицы:

$\nu_k(\theta) = y_k - C\hat{X}_{k|k-1}, k = 1, 2, ..., N$

$H_{k|k-1}(\theta) = C\Sigma_{k|k-1}C^* + R$

Функция правдоподобия имеет вид:

$L_Y(\theta) = \prod_{k=1}^{N} \frac{1}{(2\pi)^{n/2}}|H_{k|k-1}(\theta)|^{-1/2}exp(\frac{1}{2}\nu_k(\theta)^* H_{k|k-1}^{-1}\nu_k(\theta))$

Максимизируя данное выражение получаем новую оценку $\hat{\theta}$. Для получения данной оценки может быть использована процедура Ньютона-Рапсона. Данный алгоритм повторяется до сходимости, оценки полученные на последнем шаге, и отвечающие выбранному критерию и будут искомыми оценками.

7 EM-алгоритм. (Expectation-maximization (EM) algorithm)

Обозначим: $Q(\theta, \theta') = E_{\theta'}[log\frac{dP_\theta}{dP_{\theta'}}|\mathcal{Y}_N]$, где

Вектор : θ= $[\mu, \Sigma, A, C, Q, R, S]$ соответствует вероятности P_θ

И вектору : θ= $[\mu', \Sigma', A', C', Q', R', S']$ соответствует вероятность $P_{\theta'}$

Тогда исходя из неравенства Йенсена и из следующего выражения:

$\bar{E}[\frac{dP_\theta}{dP}|\mathcal{Y}_N] = \bar{E}[\bar{\Lambda}_N|\mathcal{Y}_N]$=$L(\theta)$, которое было получено ранее, имеем следующее неравенство:

$LogL(\theta) - LogL(\theta') \geq Q(\theta, \theta')$ или $LogL(\theta) \geq LogL(\theta') + Q(\theta, \theta')$

Основная цель, как и ранее, найти такой вектор θ, чтобы максимизировать $L(\theta)$. Начинаем алгорит с некоторой θ_0, тогда на шаге j+1, j≥0,зная предыдущую оценку θ_j

$\theta_{j+1} = argmax_\theta Q(\theta, \theta')$

Отметим, что согласно своему определению $Q(\theta, \theta') \geq 0$ (как логарифм единицы). Если получиться найти такое θ_{j+1}, что $Q(\theta, \theta') > 0$, тогда данное значение увеличивает значение функции правдоподобия по сравнению с предыдущем уровнем.

Сам EM-алгоритм, как понятно из названия, состоит из двух шагов, это формирования ожиданий (expectation) и их максимизация (maximization). Разберем более детально каждый из шагов.

E-шаг.

Цель данного шага вычислить следующее условное математическое ожидание

$Q(\theta, \theta') = E_{\theta'}[log\frac{dP_\theta}{dP_{\theta'}}|\mathcal{Y}_N]$, где θ'=θ_j

Так как $\frac{dP_\theta}{dP_{\theta'}}$=$\frac{\Lambda_\theta}{\Lambda_{\theta'}}$, то $log\frac{dP_\theta}{dP_{\theta'}}$=$log\Lambda_\theta - log\Lambda_{\theta'}$=$\sum_{l=0}^{N} log\bar{\lambda}_l + constant$

Вспоминая выражения для $\bar{\lambda}_l$, получаем:

$log\frac{dP_\theta}{dP_{\theta'}}$=$-\frac{1}{2}log|\Sigma| - \frac{1}{2}(X_0 - \mu)^*\Sigma^{-1}(X_0 - \mu)-$

$-\frac{N+1}{2}log|R| -\frac{1}{2}\sum_{k=0}^{N}(Y_k - CX_k)^*R^{-1}(Y_k - CX_k)-$

$-\frac{N}{2}log|Q|-\frac{1}{2}\sum_{k=1}^{N}(X_k-AX_{k-1})^*Q^{-1}(X_k-AX_{k-1})+constant$

Так как константа не зависит от θ, то её конкретный вид в данной задаче неважен.

Заметим, что

$(X_k-AX_{k-1})^*Q^{-1}(X_k-AX_{k-1})=trace[Q^{-1}(X_k-AX_{k-1})(X_k-AX_{k-1})^*]$

$(Y_k-CX_k)^*R^{-1}(Y_k-CX_k)=trace[R^{-1}(Y_k-CX_k)(Y_k-CX_k)^*]$, так как слева в обоих равенствах стоит скаляр.

Найдем условное математическое ожидание данной величины:

$E_{\theta'}[log\frac{dP_\theta}{dP_{\theta'}}|\mathcal{Y}_N]=-\frac{1}{2}log|\Sigma|-\frac{N+1}{2}log|R|-\frac{N}{2}log|Q|$

$-\frac{1}{2}trace[Q^{-1}E[\sum_{k=1}^{N}(X_k-AX_{k-1})(X_k-AX_{k-1})^*|\mathcal{Y}_N]]$

$-\frac{1}{2}trace[R^{-1}E[\sum_{k=1}^{N}(Y_k-CX_k)(Y_k-CX_k)^*|\mathcal{Y}_N]]$

$-\frac{1}{2}trace[\Sigma^{-1}E[(X_0-\mu)(X_0-\mu)^*|\mathcal{Y}_N]+\text{constant}$

На этом E-шаг закончен, и мы переходим к максимизации.

M-шаг.

Цель данного шага максимизировать функцию полученную на предыдущем шаге. Перед поиском оптимальных параметров введем следующую лемму:

Лемма 1.

Если $\bar{Z}=E[Z|\mathcal{Y}]$ и$V=E[(Z-\bar{Z})(Z-\bar{Z})^*|\mathcal{Y}]$, то справедливо следующее равенство:

$E[ZZ^*|\mathcal{Y}]=V+\bar{Z}\bar{Z}^*$

Лемма 2.

Пусть некоторая функция $f(S)=log|S|-trace[SU]$ определена на своей области определения $\mathcal{D}(f)$, содержащую положительно определенные симметричные матрицы $n\times n$, $\mathcal{D}(f)$- выпукло, f - вогнута на $\mathcal{D}(f)$. Тогда максимум f достигается при $S^{-1}=U$.

Данные леммы будут весьма полезны для нахождения оптимальных оценок параметров.

Оценка μ.

Заметим, что параметр μ присутствует только в одной части данной функции.

Согласно Лемме 1, можно выразить:

$E[(X_0-\mu)(X_0-\mu)^*|\mathcal{Y}_N]]=\Sigma_{0|N}+(\hat{X}_{0|N}-\mu)(\hat{X}_{0|N}-\mu)^*$, тогда

$trace[\Sigma^{-1}E[(X_0-\mu)(X_0-\mu)^*|\mathcal{Y}_N]=trace[\Sigma^{-1}\Sigma_{0|N}]+trace[\Sigma^{-1}(\hat{X}_{0|N}-\mu)(\hat{X}_{0|N}-\mu)^*]$.

Так как Σ^{-1} и $(\hat{X}_{0|N}-\mu)(\hat{X}_{0|N}-\mu)^*$ две симметричные, неотрицательно определенные матрицы, то справедливо следующее неравенство: $trace[\Sigma^{-1}(\hat{X}_{0|N}-\mu)(\hat{X}_{0|N}-\mu)^*]\geq 0$.

Но рассматриваемое выражение присутствует в исходной функции со знаком минус, поэтому чтобы максимизировать исходную функцию, необходимо минимизировать след данной матрицы. Согласно неравенству, минимальное значение, которое он принимает это

0, значит оптимальное $\mu = \hat{X}_{0|N}$, которое может быть получено из алгоритма сглаживания, на основе оценок θ'.

Оценка матрицы C.

Раскроем скобки в следующем выражении:

$E[(Y_k - CX_k)(Y_k - CX_k)^*|\mathcal{Y}_N]$, получим:

$E[Y_kY_k^*|\mathcal{Y}_N] - CE[X_kY_k^*|\mathcal{Y}_N] + CE[X_kX_k^*|\mathcal{Y}_N]C^* - E[Y_kX_k^*|\mathcal{Y}_N]C^*$, тогда заменив

$E[\sum\limits_{k=0}^{N} Y_kY_k^*|\mathcal{Y}_N]{=}K$; $-E[\sum\limits_{k=0}^{N} Y_kX_k^*|\mathcal{Y}_N]{=}N$; $E[\sum\limits_{k=0}^{N} X_kX_k^*|\mathcal{Y}_N]{=}M$; $\bar{C} = NM^{-1}$, тогда

$E[(Y_k - CX_k)(Y_k - CX_k)^*|\mathcal{Y}_N]{=}\ (C+\bar{C})M(C+\bar{C})^* + K - \bar{C}M\bar{C}^*$

И ту часть, где фигирирует матрица С, мы можем представить следующим образом:

$trace[R^{-1}E[\sum_{k=1}^{N}(Y_k - CX_k)(Y_k - CX_k)^*|\mathcal{Y}_N]]{=}\ trace[R^{-1}(C+\bar{C})M(C+\bar{C})^*] + rest$, где часть rest не зависит от С.

Как и ранее для двух симметричных, неотрицательно определенных матриц, R^{-1} и $(C+\bar{C})M(C+\bar{C})$, справедливо неравенство: $trace[R^{-1}(C+\bar{C})M(C+\bar{C})^*] \geq 0$, то есть минимум данного выражения равен нулю, что достигается при $C = -\bar{C}$. Значит оптимальный С:

$C = E[\sum\limits_{k=0}^{N} Y_kX_k^*|\mathcal{Y}_N]\ (E[\sum\limits_{k=0}^{N} X_kX_k^*|\mathcal{Y}_N])^{-1}$

Однако возникает закономерный вопрос, как найти данный параметр, если вектор состояния ненаблюдаем. Для ответа достаточно посмотреть, что в выражении для C, стоят условные на $\mathcal{Y}_N$ математические ожидания. Поэтому применяя *Лемму 1*, получим:

$E[\sum\limits_{k=0}^{N} Y_kX_k^*|\mathcal{Y}_N]{=}\sum\limits_{k=0}^{N} E[Y_kX_k^*|\mathcal{Y}_N]{=}\ \sum\limits_{k=0}^{N} Y_k\hat{X}_{k|N}^*$

$(E[\sum\limits_{k=0}^{N} X_kX_k^*|\mathcal{Y}_N]){=}\ \sum\limits_{k=0}^{N} E[X_kX_k^*|\mathcal{Y}_N]{=}\ \sum\limits_{k=0}^{N}[\Sigma_{k|N} + \hat{X}_{k|N}X_{k|N}^*]$. Таким образом,

$C = \sum\limits_{k=0}^{N} Y_k\hat{X}_{k|N}^*(\sum\limits_{k=0}^{N}[\Sigma_{k|N} + \hat{X}_{k|N}\hat{X}_{k|N}^*])^{-1}$

Все значения справа могут получены из алгоритма сглаживания при некоторой известной оценке параметра θ'.

Оценка матрицы А.

Для того, чтобы получить оценку теоретическое выражение для матрицы А, необходимо повторить те же операции, что делались ранее только для следующего выражения:

$E[\sum_{k=1}^{N}(X_k - AX_{k-1})(X_k - AX_{k-1})^*|\mathcal{Y}_N]$, но в предыдущем алгоритме необходимо заменить Y_k на X_k, C на A и X_k на X_{k-1}. Получим:

$\mathrm{A}{=}E[\sum\limits_{k=1}^{N} X_kX_k^*|\mathcal{Y}_N])(E[\sum\limits_{k=1}^{N} X_{k-1}X_{k-1}^*|\mathcal{Y}_N])^{-1}$

Аналогично, по Лемме 1, получим, что оптимальная оценка А, при некотором векторе параметров θ', равна:

$A = (\sum\limits_{k=1}^{N}[\Sigma_{k|N} + \hat{X}_{k|N}\hat{X}_{k|N}^*])(\sum\limits_{k=1}^{N}[\Sigma_{k-1|N} + \hat{X}_{k-1|N}\hat{X}_{k-1|N}^*])^{-1}$

Как и ранее, все оценки, которые стоят в выражении находятся из алгоритма сглаживания.

Оценка матрицы Σ.

Выпишем часть той части исходной функции, в которой присутствует искомый параметр Σ.

$-\dfrac{1}{2}log|\Sigma|\ -\dfrac{1}{2}trace[\Sigma^{-1}E[(X_0-\mu)(X_0-\mu)^*|\mathcal{Y}_N].$

В терминах Леммы 2, данное выражение можно представить, как функцию от S, где $S=\Sigma^{-1}$ и $U=E[(X_0-\mu)(X_0-\mu)^*|\mathcal{Y}_N]$, тогда:

$\dfrac{1}{2}f(S)=\dfrac{1}{2}[logS-trace[SU]].$

По Лемме 2, максимум данной функции достигается при $\Sigma=U$, получим:

$\Sigma=E[(X_0-\mu)(X_0-\mu)^*|\mathcal{Y}_N]$, и по Лемме 1:

$\Sigma=E[(X_0-\mu)(X_0-\mu)^*|\mathcal{Y}_N]=\Sigma_{0|N}+\hat{X}_{0|N}\hat{X}^*_{0|N}$

Оценка матрицы R.

Процесс нахождение оптимального значения R, аналогичен с рассмотреным ранее. Выражение исходной функции содержащее R:

$-\dfrac{N+1}{2}log|R|-\dfrac{1}{2}trace[R^{-1}E[\sum\limits_{k=1}^{N}(Y_k-CX_k)(Y_k-CX_k)^*|\mathcal{Y}_N]]$

В терминах Леммы 2, с $S=R^{-1}$ и $U=\dfrac{1}{N+1}E[\sum\limits_{k=1}^{N}(Y_k-CX_k)(Y_k-CX_k)^*|\mathcal{Y}_N]]$

$\dfrac{N+1}{2}f(S)=\dfrac{N+1}{2}[log|S|-trace[SU]$, и максимум функции f(S), достигается при R=U, то есть:

$R=\dfrac{1}{N+1}E[\sum\limits_{k=1}^{N}(Y_k-CX_k)(Y_k-CX_k)^*|\mathcal{Y}_N]]$

Если подставить полученную ранее теоретическое выражение для C,

$C=E[\sum\limits_{k=0}^{N}Y_kX_k^*|\mathcal{Y}_N]\ (E[\sum\limits_{k=0}^{N}X_kX_k^*|\mathcal{Y}_N])^{-1}$, получим:

$R=\dfrac{1}{N+1}[\sum\limits_{k=0}^{N}Y_kY_k^*-E[\sum\limits_{k=0}^{N}Y_kX_k^*|\mathcal{Y}_N](E[\sum\limits_{k=0}^{N}X_kX_k^*|\mathcal{Y}_N])^{-1}E[\sum\limits_{k=0}^{N}X_kY_k^*|\mathcal{Y}_N]]$, тогда

$R=\dfrac{1}{N+1}[\sum\limits_{k=0}^{N}Y_kY_k^*-\sum\limits_{k=0}^{N}Y_k\hat{X}^*_{k|N}(\sum\limits_{k=0}^{N}[\Sigma_{k|N}+\hat{X}_{k|N}X^*_{k|N}])^{-1}\hat{X}_{k|N}Y_k^*]$

Оценка матрицы Q.

Аналогично двум предыдущим случаям, выпишем часть функции, содержащую Q:

$-\frac{N}{2}log|Q|-\frac{1}{2}trace[Q^{-1}E[\sum_{k=1}^{N}(X_k-AX_{k-1})(X_k-AX_{k-1})^*|\mathcal{Y}_N]]$

Запишем данное выражение как f(S), где $S=Q^{-1}$ и

$U=\frac{1}{N}E[\sum_{k=1}^{N}(X_k-AX_{k-1})(X_k-AX_{k-1})^*|\mathcal{Y}_N]]$, тогда оптимальное $R=U$, то есть:

$Q=\dfrac{1}{N}E[\sum\limits_{k=1}^{N}(X_k-AX_{k-1})(X_k-AX_{k-1})^*|\mathcal{Y}_N]]$

Подставим полученное ранее значение A, получим

$Q=\dfrac{1}{N}[E[\sum\limits_{k=1}^{N}X_kX_k^*|\mathcal{Y}_N]-E[\sum\limits_{k=1}^{N}X_kX_{k-1}^*|\mathcal{Y}_N](E[\sum\limits_{k=1}^{N}X_{k-1}X_{k-1}^*|\mathcal{Y}_N])^{-1}E[\sum\limits_{k=1}^{N}X_{k-1}X_k^*|\mathcal{Y}_N]]$

До этого нами уже было показано, как посчитывается каждое из данных слагаемых, кроме $E[\sum_{k=1}^{N}X_{k-1}X_k^*|\mathcal{Y}_N]$

Согласно Лемме 1,

$E[\sum_{k=1}^{N} X_{k-1}X_k^*|\mathcal{Y}_N]$=$\sum_{k=1}^{N}[\Sigma_{k-1,k|N}+\hat{X}_{k-1|N}X_{k|N}^*]$, где фигурирует матрица ковариаций соседних векторов состояния, которая рассматривалась ранее. Таким образом, известно как считать каждый из членов по имеющейся выборке, имеем:

$Q=\frac{1}{N}[\sum_{k=0}^{N}[\Sigma_{k|N}+\hat{X}_{k|N}X_{k|N}^*-$

$\sum_{k=1}^{N}[\Sigma_{k-1,k|N}+\hat{X}_{k-1|N}X_{k|N}^*](\sum_{k=0}^{N}[\Sigma_{k-1|N}+\hat{X}_{k-1|N}X_{k-1|N}^*)^{-1}\sum_{k=1}^{N}[\Sigma_{k-1,k|N}+\hat{X}_{k|N}X_{k-1|N}^*]]$

Таким образом, у нас есть выражения для каждого из параметров нового вектора θ, подсчитанные на основе вектора θ'.

На этом второй шаг процедуру закончен. Все параметры определены. Таким образом, все значения для параметров вектора θ позволяют максимизировать фукнцию Q. И если данное значение положительно, значит получилось найти оценку лучшую, чем предыдущий векторов θ', лучшую в том смысле, что новая функция правдоподобия больше предыдущей.

8 Общая схема оценивания векторов состояния и параметров модели.

Подведем итог всему сказанному выше.

Предположим имеется модель, заданная двумя уравнениями, (1*) и (2*) со всеми предположениями введеными выше. Пусть имеется некоторый вектор параметров θ' θ=$[\mu',\Sigma',A',C',Q',R',S']$(полученный на предыдущем шаге, или, в случае первого вектора, заданный из вне) и имеется конечная выборка объема N.

Тогда, зная данные параметры, для каждого k=0,1,2...,N, посредством Фильтра Калмана (начиная с X_0 и $\Sigma_{0|0}$) найдем следующие переменные:

$\hat{X}_{k+1|k}$=$A\hat{X}_{k|k}$

$\Sigma_{k+1|k}$=$A\Sigma_{k|k}A^*+Q$

$\nu_{k+1}=Y_{k+1}-C\hat{X}_{k+1|k}$

$H_{k+1|k}=C\Sigma_{k+1|k}C^*+R$

$G_{k+1}=\Sigma_{k+1|k}C^*{H_{k+1|k}}^{-1}$

$\hat{X}_{k+1|k+1}=\hat{X}_{k+1|k}+G_{k+1}\nu_{k+1}$

$\Sigma_{k+1|k+1}$=$(I-G_{k+1}C)\Sigma_{k+1|k}$

Как результат, мы имеем отфильтрованные значения векторов состояния для каждого момента времени k=0,1,...,N и соответствующие, возникающим при их оценивании, ковариационные матрицы ошибок.На этом алгоритм Фильтра Калмана на данном этапе закончен. Затем переходим к Сглаживанию Калмана. Аналогично, используя предыдущие оценки для каждого k=0,1,...,N находим сглаженные значения вектора состояния для каждого рассматриваемого момента времени и соответствующие им матрицы ковариации ошибок. Отметим, что на данном шаге используются результаты ФК(что нетрудно видеть из формул):

$J_k=\Sigma_{k|k}A^*\Sigma_{k+1|k}^{-1}$

$\hat{X}_{k|N} = \hat{X}_{k|k} + J_k(\hat{X}_{k+1|N} - \hat{X}_{k+1|k})$

$\Sigma_{k|N} = \Sigma_{k|k} + J_k[\Sigma_{k+1|N} - \Sigma_{k+1|k}]J_k^*$

Также рассчитываем для каждых соседних периодов времени ковариационные матрицы, которые будут необходимы позже:

$\Sigma_{k-1,k-2|N} = \Sigma_{k-1|k-1}J_{k-2}^* + J_{k-1}[\Sigma_{k,k-1|N} - A\Sigma_{k-1|k-1}]J_{k-2}^*$

Имея отфильтрованные и сглаженные значения векторов состояния и ковариационных матриц ошибок оценивания данных векторов, переходим к поиску новых значений вектора парамеров θ, которые бы могли улучшить функцию максимального правдоподобия, если это возможно. Новые параметры рассчитываются следующим образом:

$\mu = \hat{X}_{0|N}$

$C = \sum_{k=0}^{N} Y_k \hat{X}_{k|N}^* (\sum_{k=0}^{N}[\Sigma_{k|N} + \hat{X}_{k|N}\hat{X}_{k|N}^*])^{-1}$

$A = (\sum_{k=1}^{N}[\Sigma_{k|N} + \hat{X}_{k|N}\hat{X}_{k|N}^*])(\sum_{k=1}^{N}[\Sigma_{k-1|N} + \hat{X}_{k-1|N}\hat{X}_{k-1|N}^*])^{-1}$

$\Sigma = E[(X_0 - \mu)(X_0 - \mu)^* | \mathcal{Y}_N] = \Sigma_{0|N} + \hat{X}_{0|N}\hat{X}_{0|N}^*$

$R = \frac{1}{N+1}[\sum_{k=0}^{N} Y_k Y_k^* - \sum_{k=0}^{N} Y_k \hat{X}_{k|N}^* (\sum_{k=0}^{N}[\Sigma_{k|N} + \hat{X}_{k|N} X_{k|N}^*])^{-1} \hat{X}_{k|N} Y_k^*]$

$Q = \frac{1}{N}[\sum_{k=0}^{N}[\Sigma_{k|N} + \hat{X}_{k|N} X_{k|N}^* -$

$\sum_{k=1}^{N}[\Sigma_{k-1,k|N} + \hat{X}_{k-1|N} X_{k|N}^*](\sum_{k=0}^{N}[\Sigma_{k-1|N} + \hat{X}_{k-1|N} X_{k-1|N}^*)^{-1} \sum_{k=1}^{N}[\Sigma_{k-1,k|N} + \hat{X}_{k|N} X_{k-1|N}^*]]$

То есть, имеем новый векторов параметров θ, и повторяем весь алгоритм описанный выше, но уже на основе новых значений параметров вектора θ. То есть пусть теперь θ=θ' и ищем новую оценку параметров θ. Процедура прододжается до сходимости.